UNIVERSIDAD NACIONAL DEL COMAHUE

Maestría en Estadística Aplicada

MODELACIÓN DE VARIABLES DE RESPUESTA DE CONTEO CON SOBREDISPERSIÓN MEDIANTE MODELOS LINEALES GENERALIZADOS MIXTOS

Trabajo de Tesis presentado por

Mabel Elizabeth Basanta Fernandez

Profesora en Matemáticas

(Universidad Nacional del Comahue 2.003)

Neuquén, mayo 2.023

COMISIÓN ASESORA DE TESIS

Director

Gustavo Néstor Giménez

Ingeniero Agrónomo

Magister Scienteae

Doctor en Ciencias Agrarias

Codirectora

Natalia Rubio

Profesora en Matemáticas

Magister en Estadística Aplicada

JURADO DE TESIS

Miembro 1: Dra. María Cristina Sosa.

Miembro 2: MSc. Edgardo Rodríguez.

Miembro 3: Mg. Silvina Pistonesi.

Fecha de aprobación de la tesis: 12 de mayo de 2.023

Director: Gustavo Néstor Giménez

Codirectora: Natalia Rubio

CERTIFICAN:

Que Mabel Elizabeth Basanta Fernandez ha realizado bajo nuestra dirección la tesis “MODELACIÓN DE VARIABLES DE RESPUESTA DE CONTEO CON SOBREDISPERSIÓN MEDIANTE MODELOS LINEALES GENERALIZADOS MIXTOS” la cual se encuentra finalizada y en condiciones de ser presentada.

Y para que así conste a todos los efectos del interesado se expide el presente certificado en Neuquén a los 13 días del mes de octubre del año 2.022

GUSTAVO NÉSTOR GIMÉNEZ NATALIA RUBIO

Agradecimientos

Para aquellos que me ayudaron a mostrar mi mejor versión:

En primer lugar, a Gustavo Giménez y Natalia Rubio, mi equipo de dirección, que confiaron en mi para hacer este trabajo de tesis. Estuvieron disponibles en todo momento con su paciencia para mis consultas y sobre todo nunca dejaron de alentarme, estando ahí para cuando me ganaba el desánimo.

Mis compañeros de trabajo del dpto. de Estadística de FaEA-UNCo que están siempre presentes para brindar el sostén necesario en lo que los necesite.

Mi familia y amigos que fueron y son incondicional apoyo en toda mi carrera.

Dedicatorias

A mis hijas Camila y Agustina que son el motor de todos mis sueños.

A Miguel, mi compañero de viaje en esta vida.

A mis padres Olga y “Poroto”, por mucho de lo que soy.

Resumen de la Tesis

MODELACIÓN DE VARIABLES DE RESPUESTA DE CONTEO CON SOBREDISPERSIÓN MEDIANTE MODELOS LINEALES GENERALIZADOS MIXTOS

Por
Mabel E. Basanta Fernandez
Octubre 2.022
Director: Dr. Gustavo N. Giménez
Codirectora: Mg. Natalia Rubio

La reciente aparición en el Alto Valle (Argentina) de la enfermedad producida por el hongo *Stemphylium vesicarium* ha motivado este estudio, cuya principal característica es que los datos resultan de conteos de lesiones (manchas en los frutos) y en particular, los analizados presentan sobredispersión. La variable de respuesta en ellos es de naturaleza discreta y no negativa. En el caso de los datos de conteo, el modelo de referencia comúnmente usado es el Modelo de regresión Poisson, que, en presencia de sobredispersión no resulta adecuado por no cumplirse los supuestos de igualdad media-varianza. Para ello, se contemplan una variedad de modelos que intentan dar cuenta de esa sobredispersión: el modelo binomial negativo, modelos inflados con ceros, modelos Hurdle y modelos Conway-Maxwell-Poisson. Esta tesis analiza la aplicabilidad de los Modelos Lineales Generalizados Mixtos (MLGM) a datos de conteo sobredispersos. Los MLGM son una extensión de los Modelos Lineales Generalizados (MLG) cuando se le adicionan efectos aleatorios que permiten considerar la correlación de las observaciones que fueron tomadas en la misma planta en diferentes momentos. La implementación del paquete "*glmmTMB*" del software estadístico *R* posibilitó ampliar la gama de modelos analizados con el paquete "lme4" de *R* permitiendo así ajustarlos a todos rápidamente con un único paquete.

Palabras claves: MLGM, modelo binomial negativo, modelo hurdle, modelo inflado con cero, glmmTMB, conteo de lesiones.

Abstract of Dissertation

MODELING RESPONSE VARIABLES OF COUNT WITH OVERDISPERSION USING GENERALIZED LINEAR MIXED MODELS

By
Mabel E. Basanta Fernandez
April 2.022
Director: Gustavo N. Giménez
Co-director: Natalia Rubio

Summary

The recent appearance in the Alto Valle (Argentina) of the disease caused by the fungus *Stemphylium vesicarium* has motivated the study of this phenomenon whose main characteristic is that the data result from counts of lesions (counts of spots on the fruits) and from the analyzed show overdispersion. The response variable in them is therefore of a discrete and non-negative nature. In the case of counting data, the commonly used reference model is the Poisson regression model, but in the presence of overdispersion it is no longer adequate because the assumptions of equality of means, and variances are not fulfilled. To do this, a variety of models are considered that try to account for this overdispersion: the negative binomial model, the zero-inflated models, the Hurdle models, and the Conway-Maxwell-Poisson models. This thesis analyzes the applicability of Generalized Linear Mixed Models (MLGM) to overdispersed count data. The GMLM are an extension of the Generalized Linear Models (MLG) when random effects are added that allow considering the correlation of the observations that were taken in the same plant at different times. The implementation of the "glmmTMB" package of the R statistical software allowed to expand the range of models analyzed with the R package "lme4", thus allowing them to be adjusted quickly with a single package.

Keywords: MLGM, negative binomial model, obstacle model, inflated zero model, glmmTMB, lesion count.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

“La forma en que tratamos de resolver un problema afecta al tipo de resultados que observamos” (Kuhn, 2019)

1 Introducción

1.1 Planteo del Problema

En algunas disciplinas como la ecología y la fitopatología, por ejemplo, se trabaja con datos generados a partir de conteos: abundancia de especies, cantidad de parásitos, cantidad de lesiones, frecuencias de hallazgos, emergencia de plántulas, germinación de semillas, entre otros. Cabe destacar que la principal característica idiosincrásica de estos datos es su naturaleza discreta y no negativa.

Los modelos lineales clásicos (regresión, ANOVA, ANCOVA), que se basan en los supuestos de distribución normal de los errores, varianza constante y linealidad entre la variable respuesta y las variables explicativas, no son adecuados para modelar este tipo de datos ya que nos encontramos con que uno o varios de estos supuestos no se cumplen. Esto se podría llegar a solucionar mediante la transformación de la variable respuesta (por ejemplo, tomando logaritmos). Sin embargo, no siempre se consigue corregir la falta de normalidad, la heterocedasticidad (varianza no constante) o la no linealidad de nuestros datos. Además, resulta muchas veces difícil interpretar los resultados obtenidos. (Cayuela, 2009)

Por otro lado, la distribución Normal es, por definición, continua. Por lo tanto, cuando la variable respuesta sigue una distribución que no es continua, los valores estimados por el modelo han de seguir el mismo tipo de distribución que los datos de partida. Cualquier otro tipo de valor estimado por el modelo no debería ser válido desde un punto de vista lógico, aunque en la práctica no se presta mucha atención a esto. Por ejemplo, los datos correspondientes a cantidad de lesiones en peras (datos que se analizarán en esta tesis) son conteos y, por lo tanto, corresponden a una variable discreta. En este caso, es razonable asumir que la variable respuesta seguirá una distribución de tipo Poisson y no una Normal (Agresti, 2015; Cayuela, 2009).

Un aspecto importante para tener en cuenta es que cuando la distribución de la variable respuesta pertenece a la familia de modelos exponenciales, la relación entre la media y la varianza caracterizan dicha distribución. Por ejemplo, en el caso en que la media sea igual a la varianza, necesariamente la variable respuesta tendrá una distribución Poisson (Agresti, 2015; Cayuela, 2009).

Una alternativa a la transformación de la variable respuesta y a la falta de normalidad en la estructura de los errores, es el uso de los Modelos Lineales Generalizados (MLG), (Nelder & Wedderburn, 1972), que son una extensión de los Modelos Lineales clásicos, los cuales permiten utilizar distribuciones no normales (Binomiales, Poisson, Gamma, Binomial Negativa, etc.) de los errores y varianzas no constantes.

Los MLG cuantifican cómo cambian los conteos esperados en función de las variables predictoras. Las mediciones repetidas en el mismo individuo, el mismo lugar o las observaciones tomadas en el mismo individuo, pero en distintos momentos a menudo están correlacionadas y esta correlación se puede explicar considerando efectos aleatorios en el modelo.

Una extensión a los MLG que permite adicionar efectos aleatorios son los Modelos Lineales Generalizados Mixtos (MLGM). Con éstos se pueden incorporar múltiples niveles de jerarquías, anidamientos, medidas repetidas, y se pueden analizar datos utilizando un amplio rango de modelos de distribución No Gaussiana (McCulloch & Searle, 2001).

Con estas potentes herramientas se logran modelizar adecuadamente la variación entre grupos, y además a través de espacio y tiempo, permitiendo una mayor precisión de los resultados estadísticos a datos no observados, así como a test estadísticos de componentes de varianza (Agresti, 2015; Bolker, 2017).

En este tipo de datos de conteo, el método para analizarlos es asumir que se aproximan a una distribución Poisson y especificar modelos estadísticos acordes a la situación. Sin embargo, un problema que se encuentra con frecuencia es que los datos muestran una marcada sobredispersión, es decir que la varianza es mayor a la esperanza y esto hace que el ajuste de los datos sea deficiente.

En la práctica, el modelo de Poisson es a menudo útil para describir la media, pero en general subestima la varianza de los datos.

Cuando los datos se dispersan en exceso, lo que significa que la varianza es mayor que la media, a menudo se modelan utilizando la distribución Binomial Negativa que tiene un parámetro que contempla la dispersión. Esta distribución se puede derivar como un tipo de mezcla de distribuciones Gamma y Poisson y tiende a esta última cuando el parámetro de dispersión tiende a cero (Bolker, y otros, 2009).

Otra forma de contrarrestar la sobredispersión de datos es utilizar modelos Quasi-Poisson. El problema es que estos modelos no poseen una función de verosimilitud especificada; y por lo tanto no se puede hacer la comparación entre modelos y se ve limitada la inferencia.

Para las distribuciones de Poisson y Binomial Negativa, el número esperado de ceros disminuye a medida que aumenta la media. Sin embargo, cuando múltiples procesos subyacen a los recuentos observados, que es casi omnipresente en biología, los recuentos pueden contener muchos ceros incluso si la media es mucho mayor que cero. A estos modelos se los conoce como “modelos inflados en cero”. Ocurre en estos casos que, una fracción sustancial de la población tiene necesariamente un resultado cero, y la fracción restante sigue alguna distribución que puede tener una pequeña probabilidad de un resultado cero. Esto conduce a un modelo que es una mezcla de un modelo de conteo ordinario como el de Poisson o Binomial Negativo con uno que coloca toda su masa en cero (Agresti, 2015; Bolker, y otros, 2009).

Un enfoque alternativo para modelar la inflación en ceros es utilizar un modelo de dos partes llamado modelo Hurdle (obstáculo). El modelo Hurdle puede manejar tanto una inflación cero como una deflación cero. Los modelos inflados en cero son más naturales que el modelo Hurdle cuando la población se considera naturalmente una mezcla, con un conjunto de sujetos que necesariamente tiene una respuesta cero. Sin embargo, el modelo Hurdle también es adecuado cuando, en algunos entornos, los datos tienen menos ceros de lo que se espera bajo supuestos de distribución estándar (Hilbe, 2017).

Los modelos que ignoran la inflación en ceros, o intentan manejarla de la misma manera que la sobredispersión simple, producen estimaciones de parámetros sesgadas (Harrison, 2014).

1.2 Caso de estudio

Los valles de Rio Negro y Neuquén constituyen la principal zona de producción de frutales de pepita de Argentina. La superficie implantada actualmente con peras (*Pyrus communis*) es de 18.367 hectáreas, siendo los principales cultivares Williams, Packham´s Triumph, Beurré D'Anjou, y Abate Fetel (SENASA, 2021). En el año 2015 se detectó en la región una nueva enfermedad que afecta a perales (Dobra & Garcia, 2015). La enfermedad conocida como "brown spot of pear" (BSP), "mancha marrón" o "estemfiliosis del peral" es causada por el hongo *Stemphylium vesicarium.* El patógeno tiene las fases de reproducción asexual y sexual en su ciclo biológico y produce ambos tipos de inóculo en su ciclo de vida, El inóculo sexual (ascosporas) corresponde al hongo *Ascomycete Pleospora allii* y el inóculo asexual (conidios) al *Deuteromycete S. vesicarium* (Rossi, Pattori, & Giosuè, 2006). Ambos estados producen síntomas en hojas y frutos. Si bien, al presente, la presencia de esta enfermedad es incipiente, resulta sumamente importante el estudio del comportamiento de este patógeno en la zona del Valle de Río Negro y Neuquén; ya que, en varias regiones mediterráneas de Europa, a pocos años de la detección de la enfermedad, se extendió progresivamente, convirtiéndose en un problema de importancia económica (Llorente, Moragrega, Ruz, & Montesinos, 2012).

En la EEA del INTA Alto Valle, se estudia la enfermedad con el objetivo principal de determinar la susceptibilidad de los principales cultivares de pera, presentes en la zona productiva del Alto Valle de Río Negro y Neuquén, a la infección por el hongo *Stemphylium vesicarium;* así como establecer la influencia de la edad del fruto en la incidencia y la severidad de la enfermedad en cada cultivar.

1.3 Estado del conocimiento del tema y/o antecedentes

Los Modelos Lineales Generalizados Mixtos (MLGM) (Breslow & Clayton, 1993) fueron obtenidos desde la generalización de los modelos lineales generalizados (MLG) (McCullagh & Nelder, 1989) por la incorporación de efectos aleatorios en

los predictores lineales. Estos modelos incluyen a los modelos lineales mixtos con respuesta Normal como un caso especial.

Existen numerosos antecedentes en modelos con efectos aleatorios tanto lineales como no lineales (Bolker, y otros, 2009; Rubio, 2016; Zuur, y otros, 2009; Zuñiga, 2019; Tassile, 2020; Godoy, y otros, 2008; Gimenez, 2020), entre otros.

En el caso de variables de conteo encontramos ejemplos de investigaciones aplicadas en disciplinas como demografía (Melkersson & Rooth, 2000), (Wang & Famoye, 1997), farmacología (Lindsey, Jones, & Jarvis, 2001), Psicología (Vives, Losilla, & Rodrigo, 2006), fitopatología (Azpilicueta, Aruani, & Morales, 2017), biología (Zuñiga, 2019), por citar algunos. Sin embargo, más numerosas resultan las investigaciones que contienen variables de conteo en medicina (Salina, Manrique, & Sosa, 2009), ciencias políticas y ciencias económicas. De hecho, estas tres disciplinas han hecho valiosas aportaciones en el tratamiento estadístico de este tipo de variables (Cameron & Trivedi, 2013; Winkelmann, 2008). En el caso de datos de conteo con sobredispersión encontramos trabajos realizados por Vives Brosa, 2002; Llorens, Palmer, & Losilla, 2004; Menni, 2011; Hilbe, 2017, entre otros.

El modelo de regresión inflado en ceros fue propuesto por Lambert (1992) con el fin de abordar el problema en los datos que muestran una cantidad de ceros mucho más alta, que la que es compatible con las distribuciones usualmente utilizadas. A los primeros trabajos en el tema los inicia Heilbron (1989).

En cuanto a la implementación práctica de estos modelos existen varios paquetes R disponibles para ajustar modelos inflados en ceros: pscl (Zeileis, Kleiber, & Jackman, 2008), MCMCglmm (Hadfield, 2010), glmmADMB (Skaug, Fournier, Nielsen, Magnusson, & Bolker, 2010), mgcv (Wood S. , 2017), brms (Bürkner, 2017), gamlss (Rigby & Stasinopoulos, 2005) y glmmTMB (Brooks, y otros, 2017). El paquete pscl, ampliamente utilizado, puede ajustar MLG inflados en cero y modelos Hurdle utilizando la estimación de máxima verosimilitud (Zeileis, Kleiber, & Jackman, 2008). Sin embargo, pscl no puede modelar la correlación dentro de las unidades de muestreo causada por muestras repetidas; que requiere efectos aleatorios. El paquete glmmADMB no puede ajustarse a modelos donde el grado de inflación con ceros varía entre las unidades de observación; por lo tanto, solo

es apropiado para modelos donde todas las unidades de observación tienen la misma probabilidad de producir un cero estructural. El paquete mgcv solo puede adaptarse a MLGM inflados en cero con predictores de inflación cero cuando se utiliza una distribución de Poisson (Wood, Pya, & Säfken, 2016). Los paquetes MCMCglmm y brms, que son utilizados para ajustar modelos bayesianos, son relativamente lentos porque se basan en el muestreo de la cadena de Markov Monte Carlo (MCMC) (Bürkner, 2017; Hadfield, 2010). El paquete gamlss es muy flexible y se ajusta a modelos aditivos generalizados con predictores en todos los parámetros de una distribución; su alcance incluye varias distribuciones infladas y modificadas en cero (Stasinopoulos, Rigby, Heller, Voudouris, & De Bastiani, 2017). Finalmente, glmmTMB equilibra la velocidad y la flexibilidad entre los paquetes para Modelos Lineales Generalizados Mixtos inflados en ceros mientras mantiene una interfaz similar al ampliamente difundido *lme4*.

1.4 Objetivos

1.4.1 Objetivo general

Analizar y evaluar la aplicabilidad de los Modelos Lineales Generalizados Mixtos, y las herramientas disponibles, en la modelación de una variable de respuesta de conteo con sobredispersión. Asimismo, aportar a los investigadores de fitopatología y de otras disciplinas biológicas afines, las herramientas más adecuadas para el análisis de este tipo de datos.

1.4.2 Objetivos específicos

- Plantear un procedimiento estadístico que permita analizar y resolver problemas cuyos datos de conteo presenten sobredispersión.
- Analizar los aportes de los Modelos Lineales Generalizados Mixtos en el análisis de datos de conteo con sobredispersión.
- Buscar y analizar softwares y paquetes disponibles que contemplen el problema de sobredispersión y el exceso de ceros.

2 Marco teórico

2.1 Modelos Lineales Generalizados (MLG)

Tal como destaca Vives Brosa (2002) en su tesis doctoral, una de las mayores contribuciones en el campo de la estadística, en el último cuarto del siglo pasado, ha sido la introducción por Nelder & Wedderburn (1972) de modelos lineales generalizados (MLG). Las dos principales contribuciones del MLG son:

- La introducción del concepto de modelado estadístico como procedimiento general para el análisis de datos.
- La integración de modelos de datos categóricos y cuantitativos en un mismo entorno.

Los MLG incluyen, además de modelos con un componente aleatorio normalmente distribuido, aquellos cuyo componente aleatorio pertenece a la familia exponencial de distribuciones. Al mismo tiempo, integran mecanismos para abordar la presencia de relaciones tanto lineales como no lineales entre variables explicativas y variables de respuesta. De esta manera, los MLG integran el modelo lineal general como un caso particular (Bolker, y otros, 2009; Correa Morales & Salazar Uribe, 2016; McCullagh & Nelder, 2019).

El modelo lineal general se caracteriza por los siguientes aspectos:

- Los valores observados y_i son independientes y siguen una distribución normal con media μ y varianza σ^2 constante.
- Las variables explicativas proporcionan un conjunto de predictores lineales $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij}. \quad j = 1,2,\dots,p$
- Los valores esperados y pronosticados están en la misma escala. Es decir, $\mu_i = \eta_i$

En general, en la bibliografía se destacan básicamente dos situaciones en las que el modelo lineal general puede ser insatisfactorio:

- cuando la distribución de la variable de respuesta no es normal,
- cuando los valores esperados no tienen una correspondencia directa con el predictor lineal, sino que son una función de este. Es decir, $g(\mu_i) = \eta_i$.

Ante estas situaciones es necesario aplicar modelos lineales generalizados basados en distribuciones no normales (McCullagh & Nelder, 2019).

Todos los modelos contienen supuestos, y es importante que las características físicas del problema coincidan tanto como sea posible con estos supuestos.

El proceso de ajuste del modelo estadístico consta de cuatro pasos (Dobson & Barnett, 2018) que se mencionan a continuación:

- **Especificación del modelo:** un modelo se especifica en dos partes: una ecuación que vincula la respuesta y las variables explicativas, y la distribución de probabilidad de la variable de respuesta.
- **Estimación de los parámetros del modelo.**
- **Comprobación de la idoneidad del modelo:** qué tan bien se ajusta o resume los datos.
- **Inferencia:** cálculo de intervalos de confianza y prueba de hipótesis sobre los parámetros en el modelo y la interpretación de los resultados.

2.1.1 Especificación del modelo bajo MLG

Los MLG amplían los modelos de regresión lineal estándar para abarcar distribuciones de respuesta no normales y posiblemente funciones no lineales de la media. Tienen tres componentes (Agresti, 2015):

- **Componente aleatorio**: especifica la variable de respuesta y y su distribución de probabilidad. Las observaciones $y = (y_1, y_2, \dots, y_n)^t$ en esa distribución se tratan como independientes.
- **Predictor lineal**: Para un vector de parámetros $\beta = \left(\beta_1, \beta_2, \dots, \beta_p\right)^t$ y una matriz X del modelo $n\ x\ p$ que contiene valores de p variables explicativas para las n observaciones, el predictor lineal es $X\beta$**.**
- **Función de enlace**: esta es una función g aplicada a cada componente de $E(y)$ que la relaciona con el predictor lineal, $g[E(y)] = X\beta$.

A continuación, presentamos más detalles sobre cada componente de un MLG.

2.1.1.1 Componente aleatorio

El componente aleatorio de un MLG consiste en una variable de respuesta y con observaciones *independientes e idénticamente distribuidas* $(y_1, y_2, \dots, y_n)$ que tienen densidad de probabilidad o función de masa para una distribución en la familia exponencial.

Recordemos que, si la distribución de y es un miembro de la familia exponencial, entonces su función de densidad se puede expresar como lo plantean Correa Morales & Salazar Uribe (2016), de la siguiente manera:

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

donde:

$f(*)$ denota la función de probabilidad en el caso que f sea una variable discreta, o la función de densidad en el caso que y sea una variable continua.

θ: Un parámetro de localización o canónico (no necesariamente la media de y).

ϕ: Un parámetro de dispersión (ϕ: solo aparece en distribuciones que tienen 2 parámetros, tal como la normal).

$a(\phi)$, $b(\theta)$ y $c(y,\phi)$ son funciones específicas de cada elemento de la familia. La función $a(\phi)$ es comúnmente escrita de la forma $a(\phi) = \phi/\omega$, donde ω es una ponderación para cada observación.

Si la variable tiene una distribución que depende solo de un parámetro, entonces:

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

puede demostrarse que la media y varianza de y están especificadas según Correa Morales & Salazar Uribe (2016), como:

$$E(y) = \mu = b'(\theta)$$

y

$$V(y) = a(\phi)b''(\theta) = a(\phi)b''\left(b'^{-1}(\mu)\right) = a(\phi)V(\mu)$$

$V(\mu)$ se denomina función de varianza. Esta función captura la relación entre $E(y)$ y $Var(y)$.

Al restringir los MLG a distribuciones de familias exponenciales, obtenemos expresiones generales para las ecuaciones de verosimilitud del modelo, las distribuciones asintóticas de los estimadores para los parámetros del modelo y un algoritmo para calcular los modelos. Las distribuciones más comúnmente utilizadas en Estadística, como la *Normal, Binomial y Poisson*, son distribuciones de la *familia exponencial* (Agresti, 2015).

En la Tabla 2-1 se resumen los elementos principales que caracterizan a algunas de las distribuciones más utilizadas de la familia exponencial (McCullagh & Nelder, 2019; Faraway, 2016).

Tabla 2-1. Elementos principales de distribuciones de la familia exponencial

Distribución	Rango de Y	θ	$a(\phi)$	$b(\theta)$	$V(\mu)$
Binomial $B(n,p)$	$[0,n]$	$ln\left(\frac{p}{1-p}\right)$	1	$n\ ln(1+exp\ (\theta)$	$np(1-p)$
Binomial Negativa $NB(\mu,\alpha)$	$Ent[0,\infty)$	$ln\ \left(\frac{\mu}{\alpha+\mu}\right)$	1	$\frac{-ln\ [1-ex\,p(\theta)]}{\alpha}$	$\mu+\alpha\mu^2$
Gamma $G(\mu,v)$	$(0,\infty)$	$\frac{-1}{\mu}$	$\frac{1}{v}$	$-ln(-\theta)$	μ^2
Normal $N(\mu,\sigma^2)$	$(-\infty,\infty)$	μ	σ^2	$\frac{\theta^2}{2}$	1
Poisson $P(\lambda)$	$Ent[0,\infty)$	$ln\ (\lambda)$	1	$ex\,p(\theta)$	λ

2.1.1.2 Predictor lineal

El predictor lineal o también denominado componente sistemática, de un MLG relaciona los parámetros $\{\eta_i\}$ pertenecientes a $\{E(y)\}$ con las variables explicativas $(x_1, x_2, \dots, x_p)$ usando una combinación lineal de ellos. Recoge la variabilidad de Y explicada por X a partir de una combinación lineal:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip}, \qquad i = 1, \dots, n; \quad j = 1, \dots, p$$

Notemos que esta expresión es lineal en los parámetros. Las variables explicativas en sí mismas pueden ser funciones no lineales de variables subyacentes, como un término de interacción (por ejemplo, $x_{i3} = x_{i1}x_{i2}$) o un término cuadrático (por ejemplo, $x_{i2} = x_{i1}^2$), aunque sigue siendo lineal en los parámetros.

En forma de matriz, expresamos el predictor lineal como:

$$\eta = X\beta,$$

Donde $\eta = (\eta_1, \eta_2, \dots, \eta_n)^t$, β es el vector de columna $p\ x\ 1$ de los parámetros del modelo, y X es la matriz $n\ x\ p$ de valores de variables explicativas $\{x_{ij}\}$. La matriz X se llama matriz modelo o también se le llama matriz de diseño.

Los MLG tratan a y_i como aleatorio y a x_i como fijo. Debido a esto, el predictor lineal, a veces se denomina componente sistemático.

2.1.1.3 Función de enlace

El tercer componente de un MLG es la función de enlace que conecta el componente aleatorio con el predictor lineal. La función de enlace escribe cómo la respuesta media, $E(y) = \mu$, está vinculada a las covariables a través del predictor lineal: $\eta = g(\mu)$

Así, $g(*)$vincula μ_i a variables explicativas a través de la fórmula:

$$\eta_i = g(\mu_i) = x_i^t \beta, i = 1, \dots, n$$

En principio, cualquier función monótona continua y diferenciable funcionará, pero existen algunas opciones más convenientes y comunes. No obstante, será el investigador quién deberá decidir la más apropiada para cada caso. Para simplificar el proceso, es habitual utilizar el parámetro canónico de la distribución como función de enlace, de forma que esta última recibe el nombre de *"función de enlace canónico"*.

En la Tabla 2-2 se presentan las funciones de enlace canónicas para las distribuciones más conocidas de la familia exponencial

Tabla 2-2. Enlaces canónicos.

Distribución	Función de enlace canónica $\theta_i = \theta(\mu_i) = \eta_i = g(\mu_i)$	
Binomial $B(n,p)$	Logit	$g(\mu_i) = ln\left(\frac{\mu_i}{1-\mu_i}\right)$
Binomial Negativa $NB(\mu,\alpha)$	Logit	$g(\mu_i) = ln\left(\frac{\mu_i}{\alpha+\mu_i}\right)$
Gamma $G(\mu,\nu)$	Recíproca	$g(\mu_i) = \frac{-1}{\mu_i}$
Normal $N(\mu,\sigma^2)$	Identidad	$g(\mu_i) = \mu_i$
Poisson $P(\lambda)$	Logarítmica	$g(\mu_i) = ln(\mu_i)$

Los modelos que tienen especial interés y que pueden formalizarse a través de la modelización lineal son los siguientes:

- Modelos para respuestas binarias (Bernoulli) o binomiales. Permiten considerar variables objetivos del tipo 0/1 (tener/no tener una característica, éxito/fracaso, etc.) muy útiles en la evaluación de incidencia de enfermedades tanto en medicina como fitopatología, análisis de riesgos, etc.

- Modelos para datos de conteo (Poisson, Binomial Negativo). Permiten considerar y analizar el comportamiento de variables de conteo (número de accidentes, nacimientos, individuos de una especie, etc.), frente a los valores del conjunto de variables explicativas.

- Modelos para variables respuesta continuas (Normal, Exponencial, Elíptica, etc.). Permiten considerar y analizar el comportamiento de variables continuas, cuantitativas (ganancias, niveles de un compuesto químico, tiempo de vida, etc.), frente a los valores del conjunto de variables explicativas.

2.1.2 Estimación de los parámetros del modelo

Luego de la especificación de uno o varios modelos, se estiman, para cada modelo especificado, los parámetros del predictor lineal y posteriormente se valora la precisión de esas estimaciones a través del cálculo de la discrepancia entre pares de modelos, con el objetivo de seleccionar el modelo óptimo.

Los métodos de estimación más utilizados son el de máxima verosimilitud y el de mínimos cuadrados. También se utilizan métodos numéricos y gráficos, cuando es apropiado, para complementar el cálculo y los métodos algebraicos de optimización (Dobson & Barnett, 2018).

El método de mínimos cuadrados no resulta adecuado cuando el componente aleatorio del modelo no es Normal, en cuyo caso, debe emplearse el de máxima verosimilitud.

2.1.2.1 Máxima Verosimilitud (ML)

Para obtener estimaciones máximo-verosímiles se necesita conocer la distribución de probabilidad conjunta de las *n* observaciones (Dobson & Barnett, 2018).

Las estimaciones de máxima verosimilitud de los parámetros son los valores de los argumentos que maximizan la función de verosimilitud (es decir, los valores de los parámetros que hacen que los valores observados de la variable dependiente sean más probables, dados los supuestos distributivos) (West, Welch, & Galecki, 2015).

Veamos el desarrollo del procedimiento (Dobson & Barnett, 2018):

Consideremos la muestra $y_1, y_2, \dots, y_i, \dots, y_n$ junto con las covariantes $x_1, x_2, \dots, x_i, \dots, x_n$, este método trata de maximizar la verosimilitud para obtener un estimador del vector de parámetros desconocidos β en el modelo:

$$E[y_i|X_i = x_i] = \mu_i = g(x_i^t \beta)$$

Asumiendo que cada componente de Y tiene una distribución proveniente de la familia exponencial de la forma denotada anteriormente, escribimos la función de verosimilitud como:

$$L(\theta; \phi, y) = f(y; \phi, \theta) = \prod_{i=1}^{n} f_i(y_i; \phi, \theta_i) \quad \text{con} \quad y = (y_1, \dots, y_n)^t$$

Como la transformación logarítmica es monótona, la maximización de la función de verosimilitud es equivalente a la maximización del logaritmo de la función verosimilitud ($\mathcal{L} = l\,n(L)$). En general, esta transformación simplifica los cálculos, al reemplazar los productos por sumas. Entonces, y dado que las observaciones son independientes, la función *log-verosimilitud* viene dada por:

$$\mathcal{L}(\theta; \phi, y) = \sum_{i=1}^{n} \mathcal{L}_i(\theta_i; \phi, y_i) = \sum_{i=1}^{n} \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}$$

Si la función $c(y_i, \phi)$ que no depende de θ_i se omite de la ecuación y se reemplaza la relación $\theta_i = \theta(\mu_i)$ entre el parámetro natural y la esperanza de la i-ésima observación tenemos,

$$\mathcal{L}(\mu_i; \phi, y) = \sum_{i=1}^{n} \mathcal{L}_i(\mu_i; \phi, y_i) = \sum_{i=1}^{n} \left\{ \frac{y_i \theta_i(\mu_i) - b(\theta_i(\mu_i))}{a(\phi)} \right\}$$

Dada la relación entre la esperanza y el vector de parámetros se tiene

$$\mathcal{L}(\beta; \phi, y) = \sum_{i=1}^{n} \mathcal{L}_i(\beta; \phi, y_i) = \sum_{i=1}^{n} \left\{ \frac{y_i \theta_i(g(x_i^t \beta)) - b\left(\theta_i(g(x_i^t \beta))\right)}{a(\phi)} \right\}$$

Su primera derivada es la denominada función score o función marcador:

$$s(\beta) = \frac{\partial \mathcal{L}}{\partial \beta} = \sum_i s_i(\beta)$$

Las contribuciones individuales a la función marcador son:

$$s_i(\beta) = x_i D_i(\beta) \sigma_i^{-2}(\beta)[y_i - \mu_i(\beta)]$$

donde

$$\begin{cases} \mu_i(\beta) = g\left(x_i^t \beta\right) \\ \sigma_i^{-2}(\beta) = a(\phi) v\left(g\left(x_i^t \beta\right)\right) \\ V(\mu) = \partial^2 b(\theta)/\partial \theta^2 \\ D_i(\beta) = \partial g\left(x_i^t \beta\right)/\partial \eta \quad \text{primera derivada de la función respuesta } g \text{ en } \eta_i = x_i^t \beta \end{cases}$$

La obtención de la estimación máximo-verosímil se plantea como las soluciones de las ecuaciones de verosimilitud:

$$s\left(\hat{\beta}\right) = 0$$

Estas ecuaciones no son, generalmente, lineales por lo que han de ser resueltas a través de métodos iterativos.

2.1.2.2 Máxima verosimilitud restringida (REML)

Un método relacionado con ML es la máxima verosimilitud restringida (o residual) (REML), que implica la idea de aplicar ML a funciones lineales de y, digamos $K^t y$, para las cuales K^t está diseñado específicamente para que $K^t y$ no contenga ninguno de los efectos fijos que son parte del modelo para y. Entonces, en ML, reemplazando y con $K^t y$ se tiene REML (McCulloch & Searle, 2001).

McCulloch & Searle (2001) destacan dos consecuencias valiosas del uso de REML:

i) Los componentes de la varianza se estiman sin verse afectados por los efectos fijos. Esto significa que las estimaciones de la varianza son invariantes a los valores de los efectos fijos.

ii) Al estimar los componentes de la varianza con REML, se tienen en cuenta implícitamente los grados de libertad de los efectos fijos, mientras que no son tomados en cuenta con ML.

2.1.2.3 ML vs REML

Ambos tienen el mérito de estar basados en el principio de máxima verosimilitud. Esto tiene el problema de que, si alguna de las soluciones maximizadoras es negativa, hay que ajustar esas soluciones para producir estimadores en el espacio de parámetros. Por otro lado, el principio de máxima verosimilitud produce varianzas de muestreo asintóticas de los estimadores de componentes de varianza; pero también tiene el inconveniente de su difícil computabilidad. ML proporciona una estimación de los efectos fijos, pero REML en sí mismo no lo hace.

Más allá de estas características, McCulloch & Searle (2001) señalan que existe una preferencia por REML, influenciada por sus siguientes méritos:

i) Es sensato para datos balanceados para los cuales las soluciones REML (no los estimadores) son los estimadores de ANOVA, y estos, a pesar de su capacidad de ser negativos, tienen el mérito sustancial de ser insesgados de mínima varianza, bajo normalidad e incluso insesgados de mínima varianza cuadrática. Sin embargo, no hay garantía de que los vínculos de propiedad de esta naturaleza se apliquen a las soluciones REML a partir de datos no balanceados.

ii) Los estimadores REML se basan en tener en cuenta los grados de libertad de los efectos fijos en el modelo. Esto es particularmente importante cuando el rango de X es grande en relación con el tamaño de la muestra.

iii) Debido a que β no está involucrado en REML, los estimadores resultantes (de componentes de varianza) son invariantes al valor de β.

iv) Los estimadores REML no son tan sensibles a los valores atípicos en los datos como lo son los estimadores ML.

2.1.3 Comprobación de la idoneidad del modelo (Bondad de ajuste)

Una vez que se estiman los parámetros se debe evaluar la magnitud de la discrepancia entre los datos observados y los esperados por el modelo. Se debe recordar que el objetivo del proceso de modelado es la obtención de un modelo que sea capaz de representar los datos y, al mismo tiempo, de reducir la complejidad. Es decir, se trata de atender a los criterios de bondad de ajuste, parsimonia y significación sustantiva (McCullagh & Nelder, 2019).

La evaluación del modelo se centra principalmente en la evaluación de posibles errores de especificación:

- De la distribución de probabilidad del componente aleatorio.
- De la función de enlace.
- Del componente sistemático, principalmente los errores de especificación relacionados con la selección de las variables explicativas.

De esta manera, si se admite una combinación satisfactoria de la distribución de la componente aleatoria y de la función enlace, el objetivo es determinar cuántos términos son necesarios en la estructura lineal para una descripción razonable de los datos. Un número grande de variables explicativas puede llevar a que un modelo explique bien los datos, pero con un aumento de complejidad en su interpretación. Por otro lado, un número pequeño de variables explicativas puede llevar a un modelo de fácil interpretación, pero que se ajuste pobremente a los datos. Lo que se desea en realidad es un modelo intermedio (McCullagh & Nelder, 2019).

Como se discutirá más adelante, es importante destacar que los errores de especificación mencionados en el último punto pueden provocar una violación del supuesto distribucional de la relación media-varianza de diversos MLG, como es el caso de la regresión Poisson (véase el apartado 2.4.2).

2.1.3.1 Test tradicionales de ajuste

En un MLG, la bondad del ajuste se puede evaluar de distintas formas, entre las que se destacan según Faraway (2016) las siguientes:

- La función o estadístico de desviación (*deviance*):

$$D = 2\sum_{i=1}^{n}\{\mathcal{L}(y_i; y_i) - \mathcal{L}(\mu_i; y_i)\}$$

donde

$\mathcal{L}(y_i; y_i)$ es el valor del logaritmo de la verosimilitud evaluado en $\mu = y$

$\mathcal{L}(\mu_i; y_i)$ es el valor del logaritmo de la verosimilitud para el modelo en consideración.

Que es la distancia entre el logaritmo de la función verosimilitud del modelo saturado (con n parámetros, uno por observación) y el modelo que se está investigando.

La *deviance* es un estadístico de comparación. La prueba se realiza utilizando una prueba χ^2 con el valor de la *deviance* y los grados de libertad como parámetros de χ^2 . Los grados de libertad de χ^2 son el número de predictores en el modelo, incluidas las interacciones. Si el valor p de la χ^2 resultante es inferior a 0,05, se considera que el modelo se ajusta bien (Agresti, 2015).

Un valor pequeño de la *deviance* indica que, para un número menor de parámetros, se obtiene un ajuste tan bueno, como cuando se ajusta el modelo saturado (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006).

La prueba de desviación se ha utilizado para comparar modelos de diferentes familias MLG y enlaces basados en una analogía con la varianza. La idea es que se prefiera un modelo con una desviación menor a un modelo con una desviación mayor. Sin embargo, este tipo de uso del estadístico actualmente es escaso en la práctica real. El problema con la prueba de *deviance* es que, de hecho, demasiados modelos mal ajustados parecen estar bien ajustados. McCullagh & Nelder (1989) y Hardin & Hilbe (2007) discuten esta prueba con más detalle.

- El coeficiente de determinación R^2 y pseudo-R^2:

R^2 es una prueba bien conocida asociada con la regresión de mínimos cuadrados ordinarios. R^2 es conocido como el coeficiente de determinación e indica el

porcentaje de variación en los datos explicados por el modelo. Este estadístico varía en valor de 0 a 1, donde los valores más altos representan un modelo mejor ajustado. Sin embargo, no es apropiada para modelos no lineales, por ejemplo, Poisson, binomial negativa y regresión logística. El estadístico que generalmente se muestra con las salidas de los modelos de conteo se denomina *pseudo-R²,* del cual se han formulado una gran variedad. La más común es la expresión:

$$R^2 = 1 - \frac{\mathcal{L}_F}{\mathcal{L}_I}$$

donde

$\mathcal{L}_F$: es la log-verosimilitud del modelo completo.

$\mathcal{L}_I$: es la log-verosimilitud del modelo con sólo intercepto.

- Razón de verosimilitud

La prueba de razón de verosimilitud (LR) es una prueba de ajuste comparativa de uso común. LR generalmente se usa para modelos anidados, pero también se ha usado para probar diferentes modelos, por ejemplo, si los datos se modelan mejor usando un binomial negativo o un Poisson. Algunos autores consideran que Poisson es un caso especial de la binomial negativa; por ello, los consideran anidados. Estos tipos de casos especiales se discutirán cuando se aborden los modelos mismos. La fórmula para la prueba de razón de verosimilitud, que se empleará con frecuencia en esta tesis, es:

$$LR = -2\{\mathcal{L}_I - \mathcal{L}_F\}$$

La prueba de razón de verosimilitud también se utiliza, a veces, como una prueba global de ajuste del modelo. La estadística de razón de verosimilitud proporcionada en la salida del modelo completo es la misma. Sin embargo, pocos estadísticos actualmente dan mucha credibilidad a esta estadística. La estadística simplemente nos dice que el modelo con predictores es superior a uno con sólo un intercepto. La mayoría de los modelos se ajustan mejor que un modelo de intercepción solamente, incluso si ninguno de los dos se ajusta a los datos. Sin embargo, la prueba LR es útil cuando se decide agregar un predictor a un modelo o agregar un conjunto de predictores, como una variable de factor multinivel. La

prueba LR es superior a la prueba de Wald para evaluar la importancia de los predictores individuales para un modelo.

2.1.3.2 Criterios de información

En cuanto a la selección del modelo, hacemos uso de los criterios de información para seleccionar entre diferentes modelos. Estos se basan en la comparación de log-verosimilitudes, pero penalizando a aquellos modelos con más variables explicativas, donde los valores más bajos indican un modelo mejor ajustado (Hilbe, 2011). Entre los más destacados podemos mencionar:

- El criterio de información Akaike (AIC):

El estadístico AIC se encuentra generalmente en dos formas:

$$AIC = \frac{-2\mathcal{L} + 2k}{n} = \frac{-2(\mathcal{L} - k)}{n}$$

y

$$AIC = -2\mathcal{L} + 2k = -2(\mathcal{L} - k)$$

donde $\mathcal{L}$ es la log-verosimilitud del modelo, k es el número de predictores, incluido el intercepto, y n representa el número de observaciones del modelo. En ambas parametrizaciones, se hace referencia a $2k$ como un término de penalización que se ajusta al tamaño y la complejidad del modelo. Ya que más parámetros hacen más probable lo que se observa, $-2\mathcal{L}$ se vuelve más pequeño a medida que aumenta k. Este sesgo se ajusta: sumando el término de penalización a $-2\mathcal{L}$.

Un n más grande también afecta el estadístico $-2\mathcal{L}$. La primera forma de la ecuación divide los términos principales por n, obteniendo así una contribución por observación al $-2\mathcal{L}$ ajustado. Todas las demás condiciones son iguales, un AIC más pequeño indica un modelo mejor ajustado.

Hilbe (2011) ideó una tabla basada en estudios de simulación que puede ayudar a decidir si la diferencia entre dos valores estadísticos de AIC es significativa. La Tabla 2-3 adaptada de Hilbe (2011) se basa en la segunda forma de la ecuación (es decir, AIC sin división por n). Esta tabla es sólo una guía general, y los casos intermedios generalmente se interpretan como indeterminados. Se deben usar otras pruebas, como BIC, para complementar.

Tabla 2-3. Criterio de selección de modelos según diferencia de AICs (adaptado de Hilbe (2011)).

Diferencia entre los modelos A y B	Resultado si A < B
(0 ; 2,5]	No hay diferencia entre los modelos
(2,5 ; 6]	Se prefiere A si n > 256
(6,0 ; 9,0]	Se prefiere A si n > 64
(10; ...]	Se prefiere A

- El criterio de información bayesiano (BIC):

Este estadístico ha sido objeto de una gran variedad de parametrizaciones. La mayoría de los softwares estadísticos usan:

$$BIC = -2\mathcal{L} + k\ ln(n)$$

donde k es el número de parámetros libres a ser estimados.

Está estrechamente relacionado con el AIC y al igual que él resuelve el problema de selección de modelos mediante la introducción de un término de penalización para el número de parámetros en el modelo, el término de penalización es mayor en el BIC que el en AIC.

Se selecciona aquel modelo con menor valor en el criterio que se utilice.

2.1.3.3 Residuos

En la práctica, puede ocurrir que, aun escogiendo cuidadosamente un modelo, al ajustarlo a un conjunto de datos, el resultado sea insatisfactorio. En este sentido pueden aparecer desviaciones sistemáticas producto de haber escogido mal la función de enlace o las variables explicativas incluidas en el modelo. También pueden ocurrir discrepancias aisladas debidas a puntos extremos.

Teniendo en cuenta esto, los residuos son los utilizados para verificar la adecuación del modelo, ya que expresan la discrepancia entre una observación y su valor ajustado, y también pueden indicar la presencia de valores anómalos o discordantes que puedan requerir de una investigación más detallada.

Los residuos más destacados son:

- **El residuo básico**: definido como la diferencia entre el valor observado, y_i, de la variable respuesta y el valor ajustado, $\hat{y}_i$, por el modelo.

$$r_i^b = y_i - \hat{y}_i \quad \text{con} \quad i = 1,2,\dots,n$$

- **El residuo de Pearson**: Es la contribución individual al estadístico χ^2 de Pearson, se define como:

$$r_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\phi} Var(\hat{\mu}_i)}} \quad \text{con} \quad i = 1,2,\dots,n$$

donde $\hat{\phi}$ es un estimador consistente del parámetro de escala ϕ.

Su versión estudentadizada es:

$$r_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\phi} Var(\hat{\mu}_i)(1 - h_i)}} \quad \text{con} \quad i = 1,2,\dots,n$$

donde h_i es el elemento diagonal de la matriz de proyección, H, donde

$$H = W^{1/2} X (X'WX)^{-1} X' W^{1/2}$$

con W una matriz diagonal, cuyos elementos de la diagonal principal vienen dados por:

$$w_i = \frac{1}{Var(\hat{\mu}_i)} \left(\frac{\partial \mu_i}{\partial \eta} \right)^2$$

La ventaja de usar este residuo estudentadizado frente al anterior es que capta mejor la variabilidad de los datos, debido a que usa el valor de h_i que resulta útil para medir la influencia de la i-ésima observación (Alcaide Delgado, 2015).

- **El residuo de desviación o deviance**: se define como

$$r_i^D = sign(y_i - \hat{y}_i)\sqrt{d_i} \quad \text{con} \quad i = 1,2,\dots,n$$

donde $d_i = 2\{\mathcal{L}(y_i; y_i) - \mathcal{L}(\hat{\mu}_i; y_i)\}$ es la llamada componente de desviación o deviance, ya definida en 2.1.3.1.

Y su versión estudentadizada es:

$$r_i^{D\prime} = \frac{r_i^D}{\sqrt{\hat{\phi}(1-h_i)}}$$

donde $\hat{\phi}$ y h_i son los definidos anteriormente en el residuo de Pearson.

2.2 Modelos Lineales Mixtos

Un modelo mixto tiene efectos tanto fijos como aleatorios en la parte sistemática del modelo.

Los conjuntos de datos complejos, que tienen una estructura jerárquica o anidada o incluyen elementos longitudinales o espaciales, comparten la característica común de correlación de observaciones dentro del mismo grupo y, por lo tanto, los análisis que asumen la independencia de las observaciones son inapropiados. El uso de efectos aleatorios es una forma común y conveniente de modelar dicha estructura de agrupamiento (Faraway, 2016; McCulloch & Searle, 2001).

Un efecto fijo es una constante desconocida que intentamos estimar a partir de los datos. Por el contrario, un efecto aleatorio es una variable aleatoria. No tiene sentido estimar un efecto aleatorio; en su lugar, tratamos de estimar los parámetros que describen la distribución de este efecto aleatorio. Un enfoque de efectos aleatorios para modelar efectos es más ambicioso ya que intenta decir algo sobre la población más amplia más allá de la muestra particular.

2.3 Modelos Lineales Generalizados Mixtos (MLGM)

Una extensión a los MLG que permite adicionar efectos aleatorios son los Modelos Lineales Generalizados Mixtos (MLGM). Con estos modelos se pueden incorporar múltiples niveles de jerarquías, anidamientos, medidas repetidas, y se pueden analizar datos utilizando un amplio rango de modelos de distribución No Gaussiana.

Estas potentes herramientas permiten modelizar adecuadamente la variación entre grupos, y además a través de espacio y tiempo, permitiendo una mayor

precisión de los resultados estadísticos a datos no observados, así como a test estadísticos de componentes de varianza (Agresti, 2015).

Los efectos aleatorios se incorporan en la misma parte del modelo que los efectos fijos, resultando el predictor lineal en términos matriciales:

$$\eta = X\beta + Zu$$

En un MLGM se considera que las observaciones del vector de respuesta y, , son independientes condicionadas a los efectos aleatorios, cada uno con función de densidad en la familia exponencial (McCulloch & Searle, 2001), es decir:

$$y_i|u \sim \text{indep. } f_{Y_i|u}(y_i|u)$$

donde $$f_{Y_i|u}(y_i|u) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}$$

Se vio en 2.1.1.1 que $E(y_i) = \mu_i = b'(\theta_i)$. Es una transformación de esta media que se desea modelar como un modelo lineal, tanto en los factores fijos como aleatorios:

$$E(y_i|u) = \mu_i$$

$$g(\mu_i) = x_i^t\beta + z_i^t u$$

Tal como en 2.1.1.3, a $g(*)$ se la conoce como la función de enlace, x_i^t y z_i^t son los vectores de covariables de los efectos fijos y aleatorios, respectivamente y u el vector de efectos aleatorios. Para completar la especificación se designa una distribución a los efectos aleatorios:

$$u \sim f_U(u)$$

La esperanza y varianza marginales para la observación i serán:

$$E(y_i) = E_u\big(E(y_i|u)\big) = E_u(\mu_i) = E\big[g^{-1}\big(x_i^t\beta + z_i^t u\big)\big]$$

$$var(y_i) = var\big(E_u(y_i|u)\big) + E[var_u(y_i|u)] = var(\mu_i) + E[\phi^2 v(\mu_i)]$$

Mediante una expresión análoga para la covarianza, se puede observar que, aunque condicionalmente las observaciones sean independientes, marginalmente están correlacionadas debido al efecto aleatorio.

2.4 Modelos de conteo

Se denominan variables de conteo a aquellas variables que determinan el número de sucesos o eventos que ocurren en una misma unidad de observación en un intervalo espacial o temporal definido (Lindsey, 1995).

El modelo de regresión de Poisson es el modelo de referencia en el campo de los estudios de variables de conteo. Sin embargo, es aplicable sólo en situaciones en las que se cumplen ciertos supuestos, que restringen considerablemente el alcance del modelo.

La restrictividad exhibida por el modelo de regresión de Poisson ha contribuido al desarrollo tanto de modelos específicos, a veces llamados extensiones, así como de modelos más generales y menos restrictivos. Entre estos últimos, destaca el modelo de regresión Binomial Negativo por su flexibilidad y frecuencia de uso (Vives Brosa, 2002).

Dado que el objetivo de esta tesis se centra en el estudio de la modelación de variable de respuesta de conteo, los modelos antes mencionados serán objeto de una exposición detallada en secciones posteriores.

2.4.1 Modelos de datos de conteo truncados

Cuando los datos excluyen estructuralmente los conteos cero, la distribución de probabilidad subyacente debe modificarse para evitar este resultado si los datos se modelan correctamente. Las rutinas de los modelos de regresión Binomial Negativo y de Poisson no deberían usarse para modelar dichos datos. Las funciones de probabilidad de los modelos Binomial Negativo y de Poisson, y sus respectivas funciones logarítmicas de verosimilitud, deben modificarse para excluir los ceros y, al mismo tiempo, permitir que todas las probabilidades en la distribución sumen 1 (Hilbe, 2017).

Los modelos basados en este tipo de modificación son comúnmente denominados modelos de conteo truncados en cero. Aunque el truncamiento puede ocurrir en cualquier valor del conteo, es el truncamiento en cero el que se ve con más frecuencia.

2.4.2 El modelo de regresión de Poisson (MRP)

La ***distribución de Poisson*** nos da la probabilidad de que se produzca un número y_i de ocurrencias de un evento. Su función de probabilidad es:

$$f(y_i, \lambda_i) = P(Y = y_i|\lambda_i) = \frac{exp\,(-\lambda_i)\lambda_i^{y_i}}{y_i!}, \qquad y_i = 0,1,2,\dots,\infty; \quad \lambda_i \in \mathbb{R}^+,$$

donde $\lambda_i = \mu_i$ es el valor esperado de Y, $E(Y)$

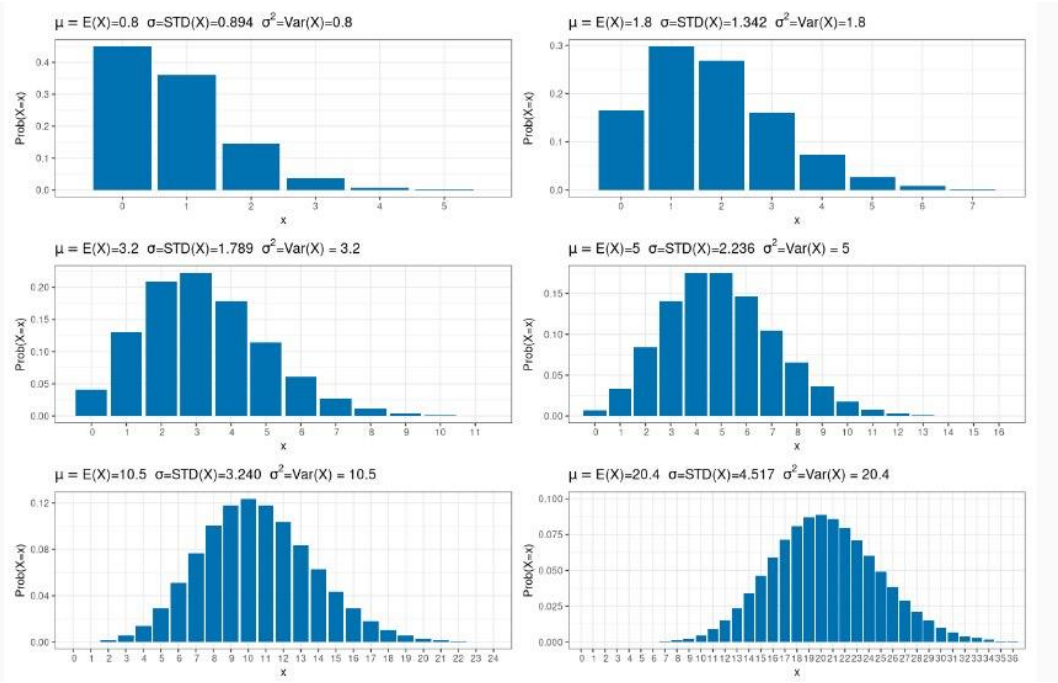

Figura 2.1. Distribución Poisson para distintas medias.

La Figura 2.1 muestra la distribución Poisson para distintos valores de μ, e ilustra alguna de las propiedades más importantes:

- A medida que aumenta μ, la masa de la distribución se desplaza hacia la derecha.
- La distribución se caracteriza por la *equidispersión*, es decir:

$$Var(y_i) = E(y_i) = \lambda_i = \mu_i$$

- Las probabilidades son estrictamente decrecientes para $0 < \mu < 1$ y el modo es 0. Para $\mu > 1$ las probabilidades se incrementan hasta $y \leq ent[\mu]$, y luego decrecen. Así, a medida que μ aumenta, la probabilidad de conteos con valor cero disminuye.
- La distribución tiende a la Normal a medida que aumenta su media μ.

La distribución Poisson es derivada a partir de un proceso estocástico que se conoce como proceso Poisson.

Un proceso de Poisson presenta las siguientes características:

- La probabilidad de al menos una ocurrencia de un evento en un intervalo temporal o espacial determinado es proporcional a la amplitud de dicho intervalo.
- La probabilidad de que se presenten dos o más ocurrencias de un evento en un intervalo muy reducido es aproximadamente 0.

Además, un proceso de Poisson se basa en dos supuestos críticos según Cameron & Trivedi (2013) y Winkelmann(2008) que se mencionan a continuación:

- Independencia entre eventos, esto es, la ocurrencia de un evento no afecta la ocurrencia futura de ese mismo evento.
- Estabilidad de los eventos, es decir, la probabilidad de un evento es constante en el espacio o el tiempo.

Si consideramos estos supuestos, es poco frecuente asignarle a un proceso la condición de Poisson puro. Es por ello, por lo que su aplicabilidad resulta restringida, lo que llevará a considerar no sólo la detección de procesos de conteo no poissonianos, sino también de técnicas y modelos capaces de dar cuenta de tales procesos.

Es a través de la construcción de un modelo, en que la variable independiente sigue una distribución Poisson, donde podemos especificar el parámetro μ_i como una forma funcional de las variables explicativas $X_1, X_2, \ldots, X_p$. La especificación

más habitual es una exponencial lineal con el fin de garantizar que $\mu_i > 0$, es decir, se hace uso de la función enlace canónica para la formulación del modelo.

$$\mu_i = exp\left(x_i^t \beta\right)$$

Se trata del modelo exponencial

$$E(Y_i/x_i) = exp\left(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}\right)$$

Por lo tanto, la distribución de Poisson condicionada a las variables explicativas viene dada por:

$$P(Y_i = y_i/x_i) = \frac{e^{-\mu_i(x_i)} \mu_i(x_i)^{y_i}}{y_i!} \qquad y_i = 0,1,2,\ldots$$

donde

$$E(Y_i/x_i) = \mu_i(x_i) = \mu\left(x_{i1}, \ldots, x_{ip}\right) = exp\left(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}\right)$$

Esta formulación se conoce como *Modelo de Regresión Poisson* (MRP) (Alcaide Delgado, 2015).

Como la distribución Poisson pertenece a la familia exponencial cuya expresión general es:

$$f(y; \theta, \phi) = ex\,p\left\{\frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}$$

donde, $\theta_i = ln\,(\lambda_i) = ln\,(\mu_i)$

y las funciones son

$b(\theta_i) = \lambda_i = \mu_i = exp\,(\theta_i)$

$a_i(\phi) = 1$

$c(y_i, \phi) = -ln\,(y_i!)$,

resulta su distribución:

$$f(y;\mu) = exp\{y_i ln\,(\mu_i)\ - \mu_i - ln\,(y_i!)\}$$

Luego, podemos identificar a las componentes del MRP como sigue:

- Componente sistemático. El predictor lineal que expresa la combinación lineal de las variables explicativas y proporciona el valor predicho es:

$$\eta_i = ln\,(\mu_i)\ = x_i^t \beta$$

- Componente aleatorio. La variabilidad de Y no explicada por η sigue una distribución de Poisson:

$$Y \sim Poisson(\mu)$$

- Función de enlace. La función que relaciona η con μ es:

$$g(\mu_i) = ln\,(\mu_i)$$

2.4.2.1 Estimación de parámetros

Para estimar el vector de parámetros β_j el método más utilizado es el de máxima verosimilitud. Para el MRP la función de verosimilitud se puede expresar como lo plantean Dobson & Barnett (2018):

$$L(\mu;y) = \prod_{i=1}^{n} exp\{y_i ln\,(\mu_i)\ - \mu_i - ln\,(y_i!)\}$$

El parámetro de interés en el modelo Poisson es μ, por eso, el cambio en la estimación de *y* a μ en la fórmula anterior.

Para simplificar los cálculos se toma el logaritmo de la función L, resultando la función log-verosímil $\mathcal{L}$:

$$lnL(\mu;y) = \mathcal{L}(\mu;y) = \sum_{i=1}^{n} \{y_i ln\,(\mu_i)\ - \mu_i - ln\,(y_i!)\}$$

teniendo en cuenta que:

$log\ (\mu_i)\ = x_i^{\mathrm{t}}\beta$

$$\mathcal{L}(\beta; y) = \sum_{i=1}^{n} \{y_i(x_i^t\beta)\ - exp(x_i^t\beta) - ln\ (y_i!)\}$$

Se deriva respecto a β y se iguala a cero para encontrar el máximo de la función. De esa ecuación se despejan los valores que serán los estimadores de los β_j.

2.4.2.2 Ajuste y selección del modelo

Dado que el modelo de regresión de Poisson pertenece a la familia de modelos lineales generalizados, haciendo uso de lo desarrollado en la sección 2.1.3 para analizar la bondad de ajuste del modelo, se obtienen las siguientes expresiones:

- La función *deviance*:

$$D = 2\sum_{i=1}^{n} \{\mathcal{L}(y_i; y_i) - \mathcal{L}(\mu_i; y_i)\}$$

Sustituyendo los términos de Poisson tenemos:

$$D = 2\sum_{i=1}^{n} \{y_i ln(y_i) - y_i - y_i ln(\mu_i) + \mu_i\} = 2\sum_{i=1}^{n} \left\{y_i ln\left(\frac{y_i}{\mu_i}\right) - (y_i - \mu_i)\right\}$$

- El coeficiente de determinación R^2:

$$R^2 = 1 - \frac{D(y; \hat{\mu})}{D(y; \hat{\mu}_0)} = \frac{\sum_{i=1}^{n}\{y_i ln(y_i/\mu_i) - (y_i - \mu_i)\}}{\sum_{i=1}^{n}\{y_i ln(y_i/\bar{y}_i)\}}$$

2.4.2.3 Interpretación

Luego de estimar los parámetros β_j, Vives Brosa (2002) indica que la información que podemos extraer de ellos puede clasificarse según Long (1997) en dos tipos:

- Cálculo de los recuentos condicionales esperados: $\mu_i|x = x_i$ que a su vez se pueden evaluar de distintas formas:
  - Factor de cambio
  - Porcentaje de cambio

- Cambio discreto

- Cálculo de la probabilidad de un recuento y_i en función de los valores que tomen las variables explicativas, es decir, $P(y = y_i|x = x_i)$

2.5 La sobredispersión en los modelos de conteo

Como se señaló en 2.4.2, el modelo de regresión Poisson tiene un ámbito de aplicación restringido debido a la propia restrictividad del modelo. Tal restrictividad es el resultado de los supuestos de la distribución de Poisson, a saber:

- Independencia: la ocurrencia de un evento no afecta la ocurrencia futura de ese mismo evento.
- Estabilidad: la probabilidad de un evento es constante en el espacio o el tiempo.

La violación de alguna de estas asunciones deriva en una incorrecta especificación del modelo.

La relación entre la media y la variancia en la distribución Poisson, se caracteriza por la equidispersión. Cuando esta relación no se cumple se tiene alguna de las dos desigualdades siguientes:

- $Var(y_i|x_i) > E(y_i|x_i)$, situación que define la sobredispersión.
- $Var(y_i|x_i) < E(y_i|x_i)$, esto es, infradispersión.

La sobredispersión es común en el modelado de conteos. Cuando el modelo para la media es correcto, pero la distribución verdadera no es Poisson, las estimaciones máximo verosímil de los parámetros del modelo siguen siendo consistentes, pero los errores estándar son incorrectos (Agresti, 2015).

En la sección 2.5.2 se presentan las pruebas para detectar la sobredispersión.

Varios autores destacan, sin embargo, que la ausencia de equidispersión no está originada únicamente por el incumplimiento de las asunciones anteriores, sino que puede ser el resultado de otras situaciones o procesos que se describen en la siguiente sección.

2.5.1 Causas

No toda dispersión excesiva es real. Hilbe (2011) destaca que a veces se puede identificar una sobredispersión aparente y modificar el modelo para eliminarla.

Las causas y posibles soluciones que se pueden aplicar a un modelo cuando se enfrenta a una sobredispersión aparente pueden resumirse en la Tabla 2-4.

Si la sobredispersión persiste, hay una variedad de métodos que los estadísticos han utilizado para tratarla, cada uno basado en abordar una razón que da lugar a la sobredispersión.

Tabla 2-4. Sobredispersión aparente en el modelo: causas y posibles soluciones.

Causas	Posibles soluciones
El modelo omite predictores explicativos importantes	Agregar el predictor apropiado
El modelo no incluye un número suficiente de términos de interacción	Construir las interacciones requeridas
Un predictor debe transformarse a otra escala	Transformar el(los) predictor(es)
	Transformar la respuesta
Los datos incluyen valores atípicos	Ajustar para valores atípicos
La función de enlace está mal especificada	Usar la función de enlace correcta

2.5.2 Diagnóstico

Existen diversas pruebas para la detección de sobredispersión. A continuación, enumeramos alguna de ellas:

- Prueba de razón de verosimilitud.
- Prueba de Wald.
- Prueba de multiplicadores de Lagrange
- Prueba de Vuong

En general no hay un consenso acerca de cuál es la prueba más adecuada. Las tres primeras pruebas mencionadas se aplican a modelos anidados dentro de un modelo más general, por ejemplo, el MRP está anidado en el modelo de regresión Binomial Negativo (MRBN). Mientras que la prueba de Vuong se aplica a modelos

no anidados, por ejemplo, MRBN y modelos inflados en ceros Binomial Negativo (ZINB) o MRP y modelos inflados en ceros Poisson (ZIP).

En todos los casos, se sugiere escoger en función del software disponible y de la complejidad de su cálculo.

2.6 Modelos para datos de conteo con sobredispersión

Al elegir el modelo más apropiado para los datos, se está seleccionando una distribución de probabilidad o mezcla de distribuciones que mejor describen los datos de la población de los cuales se extrajo la muestra a ser modelada.

Antes de pasar a detallar los principales modelos que pretenden tomar cuenta de la sobredispersión, es importante señalar que también existe la posibilidad, como señalan, entre otros Hilbe (2011) de corregir directamente el error estándar de los coeficientes del MRP mediante su producto por diferentes índices:

- $EE\sqrt{\frac{\chi^2}{gl}}$
- $EE\sqrt{\phi}$ donde ϕ es el parámetro de dispersión
- $EE\sqrt{\frac{D}{gl}}$ donde D es la discrepancia o *deviance*

Asimismo, la corrección del error estándar es posible a través de técnicas sandwich, jacknife y bootstrap que se relacionan con la simulación como lo plantean Cameron & Trivedi (2013).

Los modelos que se listan a continuación pretenden modelar directamente la causa de la sobredispersión, en lugar de sólo corregir la infraestimación de los errores estándar de los coeficientes en el MRP.

- Modelo de regresión Binomial Negativo (MRBN)
- Modelo Conway-Maxwell-Poisson
- Modelos con varianza generalizada
    - Regresión de Poisson generalizada
    - Regresión de Poisson robusta
- Modelos de datos de conteo truncados

    - MRP de ceros truncados
    - MRBN de ceros truncados
- Modelos de datos de conteo con ceros modificados
    - Modelos de datos de conteo de umbral
    - Modelo de datos de conteo con ceros aumentados

El modelo de regresión Binomial Negativo es el más utilizado del listado precedente y parte del hecho de que la causa más frecuente de sobredispersión es que el mecanismo generador de datos no es Poisson. Este es el motivo del tratamiento detallado en la sección siguiente.

2.6.1 Modelo de regresión Binomial Negativo (MRBN)

El MRBN puede ser derivado de diversas formas y con diferentes objetivos, pero como destaca Hilbe (2011) el modelo no se basa en una derivación. Puede derivarse como una mezcla de Poisson-Gamma, como una serie de ensayos de Bernoulli, como un modelo de urna de Polya-Eggenberger o como un tipo de distribución binomial inversa, e incluso Boswell & Patil (1970) mostraron doce formas específicas de derivar la distribución del modelo Binomial Negativo. Sin embargo, cuando la función de distribución de probabilidad (FDP) binomial negativa se entiende dentro del contexto de MLG, debe pensarse que se caracteriza como la probabilidad de éxitos y fracasos en una serie de ensayos de Bernoulli independientes e idénticamente distribuidos (iid). Específicamente, la distribución de probabilidad que subyace al modelo Binomial Negativo basado en MLG se basa en la distribución de "*y*" fallas antes del r-ésimo éxito en una serie de pruebas de Bernoulli 1/0 y su expresión de acuerdo con Hilbe (2011) es:

$$f(y;p,r) = \binom{y_i + r - 1}{r - 1} p_i^r (1 - p_i)^{y_i}$$

donde p es la probabilidad de r éxitos que, expresada en forma de la familia exponencial, resulta:

$$f(y;p,r) = exp\left\{y_i ln(1 - p_i) + r ln(p_i) + ln\binom{y_i + r - 1}{r - 1}\right\}$$

donde

$$\theta_i = ln(1 - p_i) \rightarrow p_i = 1 - exp(\theta_i)$$

$$b(\theta_i) = -rln(p_i) \rightarrow b(\theta_i) = -rln\big(1 - exp(\theta_i)\big)$$

$$a_i(\phi)(scale) = 1$$

Las derivadas primera y segunda, con respecto a θ, dan respectivamente las funciones de media y varianza.

Media: $b'(\theta_i) = \frac{r(1-p_i)}{p_i}$ Varianza: $b''(\theta_i) = \frac{r(1-p_i)}{p_i^2}$

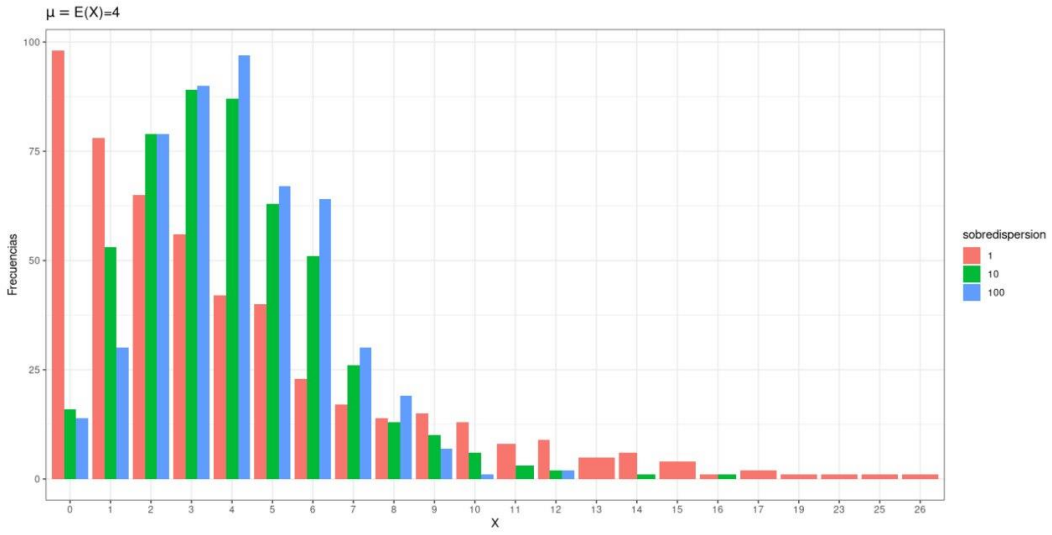

Figura 2.2. Distribución binomial negativa con media 4 según diferente dispersión

La Figura 2.2 muestra cómo cambia la distribución del modelo Binomial Negativo para una media $\mu = 4$ y distintos valores de dispersión (varianza).

La distribución del modelo Binomial Negativo es miembro de la familia exponencial de distribuciones, y es por ello que forma parte también de la familia de MLG, donde se pueden identificar los tres **componentes del MRBN** como:

- Componente sistemático. El predictor lineal que expresa la combinación lineal de las variables explicativas y proporciona el valor predicho es:

$$\eta_i = \theta_i = ln\,(1 - p_i)\; = x_i^t \beta$$

- Componente aleatorio. La variabilidad de Y no explicada por η sigue una distribución Binomial Negativa:

$$Y \sim BN\left(\frac{1}{r}, p\right)$$

- Función de enlace. La función que relaciona η con μ es:

$$g(\mu_i) = ln\,(1 - p_i)$$

Debe enfatizarse que un MRBN tiene un vínculo insignificante con la forma en que se deriva la FDP subyacente. Cuando se utiliza un modelo de este tipo para adaptarse a la sobredispersión de Poisson, o para estimar recuentos predichos, poco importa cuántos errores se hayan producido antes de un número específico de éxitos (Hilbe, 2011).

Por lo tanto, si se considera en la FDP a $\alpha = \frac{1}{r}$ y $p = \frac{1}{1+\alpha\mu}$ se tiene:

$$f(y; \alpha, \mu) = exp\left\{ y_i ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) + \frac{1}{\alpha} ln\left(\frac{1}{1+\alpha\mu_i}\right) + ln\begin{pmatrix} y_i + \frac{1}{\alpha} - 1 \\ \frac{1}{\alpha} - 1 \end{pmatrix} \right\}$$

donde α se conoce como el parámetro de dispersión o heterogeneidad.

Este parámetro de heterogeneidad está relacionado con la cantidad de sobredispersión de Poisson en los datos.

Resultando a su vez, en términos de μ y α:

Media: $b'(\theta_i) = \frac{r(1-p_i)}{p_i} = \mu$

Varianza: $b''(\theta_i) = \frac{r(1-p_i)}{p_i^2} = \mu + \alpha\mu^2$

La función de log-verosimilitud de FDP es:

$$\mathcal{L}(\mu; y, \alpha) = \sum_{i=1}^{n} \left\{ y_i ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) + \frac{1}{\alpha} ln\left(\frac{1}{1+\alpha\mu_i}\right) + ln\begin{pmatrix} y_i + \frac{1}{\alpha} - 1 \\ \frac{1}{\alpha} - 1 \end{pmatrix} \right\}$$

La función de desviación (deviance) MLG se deriva de las funciones de verosimilitud logarítmica saturada y ajustada. La función saturada consiste en reemplazar el valor de *y* por cada valor de μ.

$$D = 2\sum_{i=1}^{n}\{\mathcal{L}(y_i; y_i) - \mathcal{L}(\mu_i; y_i)\}$$

Sustituyendo la función de verosimilitud logarítmica tenemos,

$$D = 2\sum_{i=1}^{n}\left\{y_i ln\left(\frac{y_i}{\mu_i}\right) - \left(\frac{1}{\alpha} + y_i\right) ln\left(\frac{1+\alpha y_i}{1+\alpha\mu_i}\right)\right\}$$

El modelo derivado en esta sección se conoce como Binomial Negativo 2 (BN2), debido a la naturaleza cuadrática de su varianza $\mu + \alpha\mu^2$.

En ese sentido Cameron & Trivedi (2013) fueron los primeros en hacer la distinción entre los modelos binomial negativo lineal (BN1) y cuadrático (BN2). El modelo BN1 se denomina parametrización lineal, debido a su forma en la varianza: $\mu + \alpha\mu$.

La función de log-verosimilitud queda expresada como:

$$\mathcal{L}(\mu; y) = \sum_{i=1}^{n}\big\{ln\big(\Gamma(y_i + \mu_i)\big) - ln\big(\Gamma(y_i + 1)\big) - ln\big(\Gamma(\mu_i)\big) + y_i ln(\alpha) - (y_i + \mu_i) ln(1 + \alpha)\big\}$$

donde $\Gamma(*)$es la función Gamma.

Además de los modelos BN1 y BN2 considerados anteriormente, algunos autores como Cameron & Trivedi (2013) y Winkelmann (2008) proponen un modelo Binomial Negativo más general, el denominado hipermodelo Binomial Negativo K(BN k), en el cual $Var(Y_i|x_i) = \mu_i + \alpha\mu_i^{2-k}$. Esta extensión queda fuera del alcance de esta tesis.

2.6.2 Modelo Conway-Maxwell-Poisson (CMP)

Las distribuciones de Conway-Maxwell-Poisson (CMP), desarrolladas originalmente por Conway y Maxwell en 1962, tienen la capacidad de manejar tanto la sobredispersión como la subdispersión, que contienen la distribución de Poisson clásica como un caso especial y que, como describe Huang (2017), son un puente continuo entre otras distribuciones clásicas, como la geométrica y de Bernoulli. Las distribuciones de CMP también son modelos de probabilidad completa, lo que los hace particularmente útiles para las predicciones y la estimación de probabilidades de eventos.

La distribución CMP con parámetros de tasa λ y dispersión ν tiene una función de probabilidad (FDP) dada por:

$$P(Y = y|\lambda, \nu) = \frac{\lambda^y}{(y!)^\nu} \frac{1}{Z(\lambda, \nu)}\ , \quad y = 0,1,2, \ldots$$

donde $Z(\lambda, \nu) = \sum_{y=0}^{\infty} \lambda^y/(y!)^\nu$ es una constante de normalización. Cuando $\nu \to 0$ y $0 < \lambda < 1$ la distribución se reduce a una geométrica y cuando $\nu \to \infty$ a una Bernoulli.

Considerando $\mu = E(Y)$, FDP queda:

$$P(Y = y|\mu, \nu) = \frac{\big(\lambda(\mu, \nu)\big)^y}{(y!)^\nu} \frac{1}{Z(\lambda(\mu, \nu), \nu)}\ , \quad y = 0,1,2, \ldots$$

donde $\lambda(\mu, \nu)$ es una función de μ y de ν dada por la solución de

$$0 = \sum_{y=0}^{\infty} (y - \mu) \frac{\lambda^y}{(y!)^\nu}$$

La función de probabilidad (FDP) se puede escribir también como:

$$P(Y = y|\mu, \nu) = exp\{y\ ln[\lambda(\mu, \nu)] - \nu\ ln(y!) - lnZ(\lambda(\mu, \nu), \nu)\}$$

de manera que la distribución CMP pertenece a la familia exponencial, luego puede ser considerada miembro de MLG.

En este marco la variable de respuesta Y puede ser especificada como

$$Y|X \sim CMP(\mu(X^t\beta), \nu)$$

donde $\mu(*)$ es algún modelo de media y β es un vector de coeficientes de regresión. Luego

$$E(Y|X) = \mu(X^t\beta) = exp(X^t\beta)$$

La función de log-verosimilitud se expresa como:

$$\mathcal{L}(\mu, \nu|Y) = Y\ ln\big(\lambda(\mu, \nu)\big) - \nu\ ln(Y!) - lnZ(\lambda(\mu, \nu), \nu)$$

de donde se puede obtener

$$\frac{\partial \mathcal{L}}{\partial \mu} = \frac{\partial \mathcal{L}}{\partial \lambda}\frac{\partial \lambda}{\partial \mu} = \frac{Y - \mu}{V(\mu, \nu)}$$

donde la variancia de Y resulta ser

$$V(\mu, \nu) = \sum_{i=0}^{\infty} \frac{(y - \mu)^2\big(\lambda(\mu, \nu)\big)^y}{(y!)^\nu Z(\lambda(\mu, \nu), \nu)}$$

2.6.3 Modelos de datos de conteo con ceros modificados

Los datos de conteo con ceros modificados son aquellos que presentan un exceso o bien una falta de ceros. Por ser el exceso de ceros el fenómeno que se observa en los datos analizados en esta tesis, sólo se describen los modelos que lo tratan.

2.6.3.1 Modelos de datos de conteo de umbral (Hurdle)

Los modelos Hurdle se basan en la suposición de que los conteos cero se generan a partir de un proceso diferente al de los conteos positivos en una situación de datos determinada. Asume la presencia de una mezcla de distribuciones que tiene una interpretación en dos partes, por ello se conocen también como modelos en dos partes. La primera, que se refiere a un modelo con variable de respuesta binaria, modela la probabilidad de que el umbral sea cruzado, y la segunda, a un modelo de datos de conteo truncados, que modela el valor esperado de los valores positivos.

Lo más frecuente es que se fije ese umbral en cero y en la literatura se encuentran dos acepciones:

- al sobrepasar el valor del umbral se obtiene una distribución truncada en cero (Hilbe, 2017; Winkelmann, 2008; Yen, 1999).
- la distribución resultante al cruzar el umbral es una distribución Poisson o Binomial Negativa, de manera que el valor cero es posible y su probabilidad viene determinada por la distribución de referencia (Cameron & Trivedi, 2013; Mullahy, 1997).

Mientras que los modelos inflados en cero (que se detallan en la sección 2.6.3.2) requieren conteos de cero excesivos en los datos, los modelos de umbral pueden ajustarse para modelos con muy pocos conteos de cero. Sin embargo, algunos conteos cero son necesarios.

El tipo de componentes binarios y de conteo que se utilizan para estimar los modelos de umbral son:

BINARIO: binomial, Poisson, geométrico, BN2

CONTEO: Poisson, geométrico, Binomial Negativo

Los enlaces binomiales para el componente binomial incluyen: logit, probit, complemento loglog, Cauchit y log.

Siguiendo a Hilbe (2011), primera acepción, hasta que se cruza el umbral, el proceso genera una respuesta binaria (1/0). La naturaleza del umbral no se especifica, pero numéricamente puede considerarse simplemente como que los datos tienen un recuento positivo. En este sentido, se cruza el umbral si una cuenta es mayor que cero. En cualquier caso, los dos procesos se unen utilizando la siguiente función de log-verosimilitud:

$$\mathcal{L} = \ln\big(f(0)\big) + \ln[1 - f(0)] + lnP(t)$$

donde $f(0)$ representa la probabilidad de la parte binaria del modelo y $P(t)$ representa la probabilidad de un conteo positivo. En el caso de un modelo logit, la probabilidad de cero es:

$$f(0) = P(y = 0; x) = \frac{1}{1 + \exp\left(x_i^{\mathrm{t}}\beta_b\right)}$$

con β_b indicando las estimaciones de parámetros del componente binario del modelo de obstáculos. De ello se deduce que $1 - f(0)$ viene dado por

$$\frac{\exp\left(x_i^{\mathrm{t}}\beta_b\right)}{1 + \exp\left(x_i^{\mathrm{t}}\beta_b\right)}$$

Y sabiendo que la función logarítmica de verosimilitud para un BN2 truncado en cero es

$$P(y|x > 0) = y_i ln\left(\frac{\exp\left(x_i^{\mathrm{t}}\beta\right)}{1 + \exp\left(x_i^{\mathrm{t}}\beta\right)}\right) - \frac{ln\left(1 + \exp\left(x_i^{\mathrm{t}}\beta\right)\right)}{\alpha} + ln\Gamma\left(y_i + \frac{1}{\alpha}\right) - ln\Gamma(y_i + 1) - \Gamma\left(\frac{1}{\alpha}\right) - ln\left(1 - (1 + \exp\left(x_i^{\mathrm{t}}\beta\right)\right)^{-1/\alpha}$$

se obtiene finalmente la probabilidad de las dos partes del modelo Binomial Negativo Hurdle:

$si\ (y = 0)$: $\quad \frac{1}{1 + ex\,p\left(x_i^t\beta_b\right)}$

$si\ (y > 0)$: $\quad y_i ln\left(\frac{ex\,p\left(x_i^t\beta\right)}{1 + ex\,p\left(x_i^t\beta\right)}\right) - \frac{ln\left(1 + exp\left(x_i^t\beta\right)\right)}{\alpha} + ln\Gamma\left(y_i + \frac{1}{\alpha}\right) - ln\Gamma(y_i + 1) - \Gamma\left(\frac{1}{\alpha}\right) - ln\left(1 - (1 + exp\left(x_i^t\beta\right)\right)^{-1/\alpha}$

La log-verosimilitud se da como:

$si\ (y = 0)$: $\quad - exp\left(x_i^t\beta_b\right)$

$si\ (y > 0)$: $\quad ln\ (1 - exp\ (- exp\left(x_i^t\beta_b\right))) + y_i ln\left(\frac{exp\left(x_i^t\beta\right)}{1 + exp\left(x_i^t\beta\right)}\right) - \frac{ln\left(1 + exp\left(x_i^t\beta\right)\right)}{\alpha} + ln\Gamma\left(y_i + \frac{1}{\alpha}\right) - ln\Gamma(y_i + 1) - \Gamma\left(\frac{1}{\alpha}\right) - ln\left(1 - (1\,exp\left(x_i^t\beta\right)\right)^{-1/\alpha}$

2.6.3.2 Modelo de datos de conteo con ceros aumentados

Los modelos inflados en ceros son modelos de dos partes, siendo el componente binario una estimación de $P(y = 0)$, y el componente de conteo una estimación del rango completo de conteos, incluido el cero. Se diferencia del modelo de umbral en que este último separa el modelado de ceros del modelado de conteos, lo que implica que sólo un proceso genera ceros. Además, los dos componentes de un modelo inflado en cero no se pueden separar como se puede hacer con los modelos Hurdle.

De manera similar a cómo el modelo de Poisson está anidado en BN2, el modelo de Poisson inflado en cero, o ZIP (por sus siglas en inglés), está anidado dentro del Binomial Negativo inflado en ceros, o ZINB (por sus siglas en inglés). Esto permite la comparación de los dos modelos utilizando una prueba de razón de verosimilitud. También se puede utilizar una prueba de Wald para evaluar el valor del modelo. Además, el uso de las estadísticas de ajuste AIC y BIC se pueden utilizar para determinar el modelo de inflado en cero mejor ajustado entre varios tipos de modelo (Hilbe, 2011).

Las funciones logarítmicas de verosimilitud de los modelos NB-logit y NB-probit se enumeran a continuación. Tener en cuenta que β_1 indica el predictor lineal de componente binario y β indica predictor del componente de conteo.

- Modelo NB-logit

$$si\ (y == 0): \quad \sum_{i=1}^{n} \left\{ ln\left(\frac{1}{1+exp(-x_i^t\beta_1)}\right) + \frac{1}{1+exp(x_i^t\beta_1)}\left(\frac{1}{1+\alpha exp(-x_i^t\beta)}\right)^{1/\alpha} \right\}$$

$$\text{si } (y > 0): \quad \sum_{i=1}^{n} \left\{ ln\left(\frac{1}{1+exp(-x_i^t\beta_1)}\right) + ln\Gamma\left(\frac{1}{\alpha}+y_i\right) - ln\Gamma(y_i+1) - ln\Gamma\left(\frac{1}{\alpha}\right) + \left(\frac{1}{\alpha}\right) ln\left(\frac{1}{1+\alpha\ exp(x_i^t\beta)}\right) + y_i ln\left[1 - \frac{1}{1+\alpha exp(x_i^t\beta)}\right] \right\}$$

- Modelo NB-probit

$$si\ (y == 0): \quad \sum_{i=1}^{n} \left\{ ln\left(\Phi x_i^t \beta_1\right) + \left(1 - \Phi x_i^t \beta_1\right) + \left(\frac{1}{1 + \alpha\ exp\left(-x_i^t \beta\right)}\right)^{1/\alpha} \right\}$$

$$si\ (y > 0): \quad \sum_{i=1}^{n} \left\{ ln\left(1 - \Phi x_i^t \beta_1\right) + ln\Gamma\left(\frac{1}{\alpha} + y_i\right) - ln\Gamma(y_i + 1) - ln\Gamma\left(\frac{1}{\alpha}\right) + \left(\frac{1}{\alpha}\right) ln\left(\frac{1}{1 + \alpha\ exp\left(x_i^t \beta\right)}\right) + y_i ln\left[1 - \frac{1}{1 + \alpha\ exp\left(x_i^t \beta\right)}\right] \right\}$$

donde $exp(x\beta_1)$ es el ajuste, o μ, del proceso binario, y $exp(x\beta)$ es lo mismo con respecto al proceso de conteo. Φ representa la función de distribución acumulativa normal o gaussiana. Estos términos a veces reciben símbolos completamente diferentes, por ejemplo, $x\beta_1$ se representa comúnmente como ζ para la covariable binaria y γ (gamma) para sus coeficientes asociados. Dados estos términos, la función logística de probabilidad acumulada se expresa como:

$$\Lambda = \frac{exp(\zeta_i \gamma)}{1 + exp(\zeta_i \gamma)}$$

La función acumulativa Gaussiana está dada por:

$$\Phi(\zeta_i \gamma) = \int_0^{\zeta_i \gamma} \frac{1}{\sqrt{2\pi}} exp \frac{-\mu^2}{2} d\mu$$

A diferencia de los modelos Hurdle, el proceso binario normalmente tiene diferentes predictores que en el proceso de conteo. De modo que el modelo sirve para determinar qué variables o elementos de los datos tienen una relación directa con los conteos cero. Esta es la razón por la que el modelo inflado en cero, a diferencia de los modelos Hurdle, tiene su proceso de conteo que predice ceros (Hilbe, 2011).

2.6.3.2.1 Ajuste del modelo

La prueba de ajuste estándar para los modelos ZINB es la prueba de Vuong. Esta prueba es una comparación de ZINB y ZIP. Esencialmente, la prueba de Vuong es una comparación de los valores de ajuste predichos de ZINB y ZIP, evaluando si existe una diferencia significativa entre los dos. Dado que $P_{ZIP}(y|x)$ es la probabilidad de observar *y* sobre la base de *x* en un modelo ZIP, y $P_{ZINB}(y|x)$ es la probabilidad de observar el mismo *y* sobre la base del mismo *x* usando el modelo ZINB, la fórmula de la prueba de Vuong puede expresarse como:

$$V = \frac{\sqrt{n}\bar{u}}{SD(u_i)}$$

donde

$$u_i = ln\left(\frac{\sum_i P_{ZIP}(y_i|x_i)}{\sum_i P_{ZINB}(y_i|x_i)}\right)$$

La prueba de Vuong utiliza una distribución normal para evaluar el valor comparativo. Con un nivel de confianza del 95 %, los valores de V superiores a +1,96 indican que ZIP es el modelo preferido. Valores inferiores a -1,96 indican que ZINB es el modelo preferido. Los valores entre estos y los puntos críticos indican que no se prefiere ningún modelo sobre el otro. La prueba de Vuong compara las probabilidades del numerador y el denominador, con valores superiores a 1,96 a favor de las probabilidades del numerador.

También se puede usar una prueba de razón de verosimilitud para comparar un ZINB con un ZIP. La lógica de la comparación se basa en la distancia entre el valor de α y 0. ¿Es α significativamente mayor que 0 de modo que el modelo sea BN2 en lugar de Poisson (α=0)? Nuevamente, la prueba se da como:

$$LR = -2(\mathcal{L}_P - \mathcal{L}_{NB})$$

Se utiliza una prueba χ^2 modificada para evaluar la significancia, con un grado de libertad y con el estadístico de prueba dividido por dos (Hilbe, 2011).

3 Software y paquetes disponibles para el ajuste de modelos

En este capítulo, se describen los paquetes de R disponibles para ajustar modelos para datos de conteo, sin hacer distinción entre métodos frecuentistas y/o bayesianos, y se presenta un nuevo paquete: glmmTMB (Brooks, y otros, 2017).

En esta tesis los datos se analizaron con R que es un entorno de software libre para el análisis estadístico y gráficos, y que tiene además la ventaja que se ejecuta en una amplia variedad de plataformas UNIX, Windows y MacOS (RCoreTeam, 2022).

Para ajustar modelos con ceros modificados, R cuenta con varios paquetes como: pscl (Zeileis, Kleiber, & Jackman, 2008), INLA, MCMCglmm (Hadfield, 2010), glmmADMB (Skaug, Fournier, Nielsen, Magnusson, & Bolker, 2010), mgcv (Wood, 2017), brms (Bürkner, 2017) y gamlss (Rigby & Stasinopoulos, 2005).

Los MLG se implementan en R mediante la función glm() en el paquete stats y la función glm.nb(), para el modelo binomial negativo, en el paquete MASS (Venables & Ripley, 2002). El modelo Hurdle y los inflados en cero se ajustan con las funciones hurde( ) y zeroinfl( ) en el paquete pscl.

El paquete pscl ajusta los MLG utilizando la estimación de máxima verosimilitud. Por ejemplo, pscl se puede utilizar para probar la hipótesis de que el conteo de manchas marrones en las peras depende de la edad y los ceros estructurales dependen del cultivar. Sin embargo, pscl no puede modelar la correlación dentro de las unidades de muestreo causada por muestras repetidas; esto requiere efectos aleatorios.

Otros paquetes tienen capacidades similares para ajustar MLG sin inflado flexmix, MXM (Lagani, Athineou, Farcomeni, Tsagris, & Tsamardinos, 2017), VGAM (Yee, 2020),sin embargo, en este capítulo nos enfocamos en paquetes que también pueden estimar efectos aleatorios porque es uno de los temas de esta tesis.

El paquete glmmADMB, que puede adaptarse a MLGM sin inflación en cero, no puede ajustarse a modelos donde el grado de inflación en cero varía entre las unidades de observación (Brooks, y otros, 2017); por lo tanto, sólo es apropiado para modelos donde todas las unidades de observación tienen la misma

probabilidad de producir un cero estructural. INLA tiene la misma limitación que glmmADMB (Rue, Martino, & Chopin, 2009).

El paquete mgcv solo puede adaptarse a MLGM inflado en cero con predictores de inflación cero cuando se utiliza una distribución de Poisson (Wood, Pya, & Säfken, 2016).

Los paquetes MCMCglmm y brms (que ajustan desde una óptica Bayesiana) pueden adaptarse a MLGM inflados en cero con predictores de inflación cero, pero, Brooks, y otros (2017) destacan que son relativamente lentos porque se basan en el muestreo de la cadena de Markov Monte Carlo (MCMC) (Bürkner, 2017; Hadfield, 2010).

El paquete gamlss es flexible, se ajusta a modelos aditivos generalizados con predictores en todos los parámetros de una distribución y su alcance incluye varias distribuciones infladas y modificadas por ceros (Stasinopoulos, Rigby, Heller, Voudouris, & De Bastiani, 2017).

Es importante aclarar que la lista de características descriptas aquí no es exhaustiva. En este sentido Brooks, y otros (2017) observan que debe tenerse en cuenta que brms, gamlss y MCMCglmm tienen características adicionales que van más allá del alcance de los MLGM con inflación cero.

3.1 Descripción del paquete glmmTMB

Brooks, y otros (2017) presentan un nuevo paquete de R, glmmTMB, construido en el Template Model Builder (motor de diferenciación automático), (Kristensen, Nielsen, Berg, Skaug, & Bell, 2015), que estima MLG, MLGM y extensiones de MLGM, incluidos MLGM sin inflación cero y con obstáculos mediante máxima verosimilitud (ML).

El objetivo de diseño del paquete glmmTMB es ampliar la flexibilidad de los MLGM en R manteniendo una interfaz familiar basada en el paquete lme4, uno de los más utilizados para ajustar MLGM (Bates, Mächler, Bolker, & Walker, 2014). Para maximizar la flexibilidad y la velocidad, la estimación del paquete glmmTMB se realiza utilizando el paquete TMB que evalúa y maximiza la aproximación de

Laplace de la verosimilitud marginal, donde los efectos aleatorios se integran automáticamente.

Brooks, y otros (2017) probaron que implementar el paquete glmmTMB otorga una ventaja de velocidad frente al paquete lme4 al estimar modelos no gaussianos y brinda una mayor flexibilidad en las clases de distribuciones que puede ajustar (Tabla 3-1)

Tabla 3-1. Familia de funciones de glmmTMB vs lme4.

Familia de funciones para glmmTMB	Familia de funciones para lme4
• nbinom2(link = "log") • nbinom1(link = "log") • compois(link = "log") • truncated_compois(link = "log") • genpois(link = "log") • truncated_genpois(link = "log") • truncated_poisson(link = "log") • truncated_nbinom2(link = "log") • truncated_nbinom1(link = "log") • beta_family(link = "logit") • betabinomial(link = "logit") • tweedie(link = "log") • ziGamma(link = "inverse")	• binomial(link = "logit") • gaussian(link = "identity") • Gamma(link = "inverse") • inverse.gaussian(link = 1/mu^2") • poisson(link = "log") • quasi(link = "identity", variance = "constant") • quasibinomial(link = "logit") • quasipoisson(link = "log")

Un modelo glmmTMB tiene cuatro componentes principales: una fórmula de modelo condicional, una distribución para el modelo condicional, una fórmula de modelo de dispersión y una fórmula de modelo de inflación en cero.

La función para ajustar los modelos con los valores por defecto es la siguiente:

- glmmTMB(**formula**, data = NULL, **family** = gaussian(), **ziformula** = ~0, **dispformula** = ~1, weights = NULL, offset = NULL, contrasts = NULL, na.action, se = TRUE, verbose = FALSE, doFit = TRUE, control = glmmTMBControl(), REML = FALSE, start = NULL, map = NULL, sparseX = NULL)

Los MLG y MLGM simples se pueden ajustar usando el modelo condicional, mientras se dejan las fórmulas de inflación cero y dispersión en sus valores predeterminados.

La media del modelo condicional se especifica mediante una fórmula de dos colas con la variable de respuesta a la izquierda y los predictores a la derecha, que

pueden incluir efectos aleatorios y compensaciones. Esta fórmula usa la misma sintaxis que lme4. Por ejemplo, si el conteo de manchas en las peras (Severidad) varían según el cultivar y la edad (Cultivar, Age) y varían aleatoriamente según la planta(Planta), entonces la fórmula para la dependencia del conteo medio de manchas podría ser:

ej: Severidad ~ Cultivar+Age + Cultivar:Age + (1|Planta)

La distribución alrededor de la media del modelo condicional se especifica utilizando el argumento “family”. Para los tipos de datos de conteo analizados en esta tesis, la distribución típicamente será de Poisson o binomial negativa.

Las distribuciones de Poisson y Binomial Negativa utilizan un enlace “log” de forma predeterminada, pero se pueden especificar otros enlaces ("log", "logit", "probit", "inverse", "cloglog", "identity", or "sqrt").

La fórmula del modelo de dispersión contiene solo efectos fijos: el valor predeterminado “~1” especifica la dispersión estándar dada cualquier familia. El argumento se ignora para las familias que no tienen un parámetro de dispersión. Alternativamente, el parámetro de dispersión puede variar con efectos fijos; en este caso, el modelo de dispersión utiliza un enlace “log”. El modelo de dispersión se puede utilizar para tener en cuenta la heteroscedasticidad. Por ejemplo, si la respuesta es más variable (en relación con la media) entre los distintos Cultivares, entonces en un modelo con una distribución binomial negativa podría usar la fórmula unilateral dispformula = ~ Cultivar.

En los modelos mixtos gaussianos, *dispformula=~0* fija la varianza residual en 0 (en realidad, un valor pequeño distinto de cero), forzando la varianza en los efectos aleatorios.

El valor exacto se puede controlar a través de *control = glmmTMBControl (zero_dispval=...)*; el valor predeterminado es qrt(.Machine$double.eps).

En glmmTMB, es posible incluir efectos aleatorios en los modelos condicional y de inflación en cero, pero no en el modelo de dispersión.

El resumen (summary) de glmmTMB en los MLGM simples tiene un formato de salida similar al del lme4. Esto se verá reflejado en el caso de estudio desarrollado en la sección 4.3.6.

3.2 Rootogramas

El rootograma es una herramienta gráfica asociada con el trabajo de JW Tukey que se usó originalmente para evaluar la bondad de ajuste de distribuciones univariadas. Kleiber & Zeileis (2016) extienden el uso del rootograma a los modelos de regresión y muestran que esto es particularmente útil para diagnosticar y tratar problemas como la dispersión excesiva y/o el exceso de ceros en los modelos de datos de conteo.

El rootograma compara gráficamente los valores observados y esperados trazando rectángulos o barras similares a histogramas para las frecuencias observadas y una curva para las frecuencias ajustadas, todo en una escala de raíz cuadrada. Las raíces cuadradas, en lugar de las observaciones no transformadas, se emplean para ajustar aproximadamente las diferencias de escala en los valores o intervalos de conteos y a ello debe su nombre por “raíz=root” en inglés. De lo contrario, las desviaciones sólo serían visibles para valores de conteos con grandes frecuencias observadas/esperadas.

Se han sugerido tres tipos diferentes de rootogramas:

- *de pie*: muestra las barras “de pie” sobre la línea cero para la raíz cuadrada de los valores observados y una curva para la raíz cuadrada de los valores esperados.
- *colgantes*: para alinear todas las desviaciones a lo largo del eje horizontal, las barras se dibujan desde la raíz cuadrada de los valores esperados hasta la diferencia entre la raíz cuadrada de los valores esperados y la raíz cuadrada de los valores observados, para que las barras “cuelguen” desde la curva que representa las frecuencias esperadas.
- *suspendidas*: Para enfatizar principalmente las desviaciones (más que las frecuencias observadas), una tercera alternativa es dibujar las barras para las diferencias entre las raíces cuadradas de las frecuencias esperadas y

observadas. Las barras cuelgan de la línea cero en lugar de la línea de los conteos esperados.

En la Figura 3.1 se muestran ejemplos de los distintos tipos de rootogramas para ajuste de modelo Poisson y para modelo Binomial Negativo. El modelo Poisson muestra un buen ajuste mientras que el modelo Binomial Negativo, no.

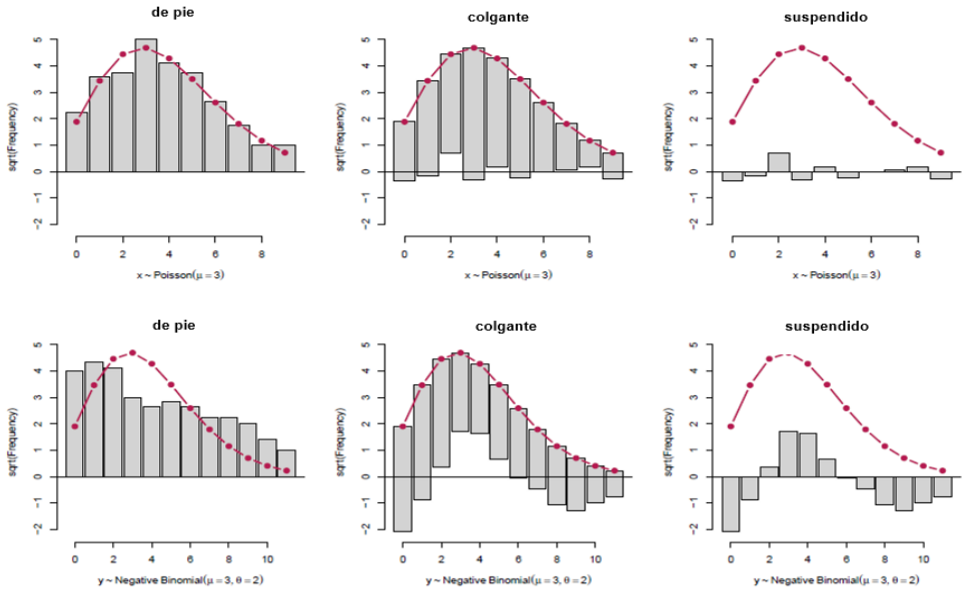

Figura 3.1. Distintos tipos de rootogramas

En un rootograma colgante, que son los que recomiendan Kleiber & Zeileis (2016), si una barra no alcanza la línea cero, entonces el modelo sobre predice un conteo en particular, y si la barra excede la línea cero, lo sub predice. Si se observa en la parte inferior de las barras un patrón ondulante con recorridos por encima o por debajo de la línea de referencia cero se podrá concluir en que hay una falta general de ajuste del modelo.

A diferencia de otros gráficos basados en modelos que a menudo se enfocan en los efectos sobre la media del ajuste de la distribución, los rootogramas capturan las desviaciones a través del soporte de toda la distribución y, por lo tanto, puede ayudar a diagnosticar el desajuste con respecto a la dispersión y/o forma. En este sentido, Kleiber & Zeileis (2016) destacan su relevancia para los modelos de datos de conteo, que a menudo se ven afectados por problemas como la sobredispersión y/o exceso de ceros. Esta herramienta gráfica está disponible en el paquete “countreg” de R (Zeileis & Kleiber, 2022).

“Modelar requiere un ir y volver entre lo que ya conocemos y lo que nuestros datos tienen que decirnos” (Lunneborg, 1994)

4 Aplicación

Para analizar y evaluar la aplicabilidad de los Modelos Lineales Generalizados Mixtos, y las herramientas disponibles, en la modelación de una variable de respuesta de conteo con sobredispersión, se utilizó un conjunto de datos que se enmarcan en la investigación de tesis doctoral de la Ing. Agr. Marisa Aluminé Tudela, con sede de trabajo en INTA Alto Valle, uno de cuyos objetivos fue determinar la susceptibilidad de los principales cultivares de pera presentes en la zona productiva del Alto Valle de Río Negro y Neuquén a la infección por el hongo *Stemphylium vesicarium,* así como establecer la influencia de la edad del fruto en la severidad de la enfermedad en cada cultivar.

4.1 Diseño de experimento y muestreo.

Para evaluar la susceptibilidad varietal, los frutos fueron obtenidos desde árboles frutales en buen estado nutricional y sanitario, ubicados en un cuadro perteneciente a la estación experimental agropecuaria de INTA Alto Valle. Los árboles se escogieron al azar, se utilizaron un total de 24 plantas, 6 de cada uno de los 4 cultivares. Los cultivares de pera (Pyrus communis) elegidos para el estudio fueron Williams Bon Chretien, Packham´s Thriump, Beurré D´Anjou y Abate Fetel. Como testigo no susceptible se utilizó manzana (*Malus sylvestris*) del cultivar Pink Lady. Se seleccionaron 5 frutos por planta en tres momentos (edades) diferentes. Las edades de los frutos se calcularon como días después de plena floración (ddpf), en base a las fechas de plena floración registradas. La primera selección se realizó a la edad de 55-60 ddpf, la segunda a la edad de 85-90 ddpf, y la tercera 5 a 7 días antes del sello de autorización de cosecha de cada cultivar. La fecha de plena floración del cultivar Williams fue el 27/09/2020, la de Packham´s el 25/09/2020, y las de D´Anjou y Abate Fetel coincidieron el 23/09/2020.

Para la producción del inóculo, desde el cepario del laboratorio de Fitopatología, se escogió una cepa altamente patogénica, identificada como S23, que fue obtenida desde una lesión en pera Beurré D´Anjou. La cepa fue cultivada en

placas de Petri sobre agar-agar e incubadas por 20 días a 20°C en oscuridad, condiciones que estimulan la esporulación. Transcurrido ese periodo, las placas se trasladaron a una cámara de flujo laminar. Las superficies de las placas se frotaron suavemente con un haza luego de añadir 9 ml d agua destilada estéril y una gota de Tween 20 (polisorbato 20, usado como detergente de baja toxicidad). Las suspensiones fueron ajustadas a una concentración de 5×10^{4} conidios/ml.

Con el fin de efectuar las inoculaciones, los frutos se cosecharon y se trasladaron al laboratorio de Fitopatología. Se desinfectaron por inmersión durante un minuto, en una solución de hipoclorito de sodio al 1 %. Se enjuagaron 3 veces con agua destilada estéril, en cámara de flujo laminar. La técnica de inoculación consistió en sumergir a los frutos en la suspensión conidial ajustada a 5×10^{4} conidios/ml, durante un minuto, mientras que los frutos empleados como testigo fueron sumergidos en una solución de agua destilada estéril. Luego de la inoculación, los frutos de una misma planta se dispusieron dentro de cámaras húmedas construidas por una bandeja con un soporte individual para cada fruto (5 frutos por cámara). En el fondo de cada bandeja se utilizó papel tisú mojado con agua destilada estéril para mantener las condiciones de humedad. Se incubó por un total de 7 días a temperatura ambiente (20-25°C). Al cabo de ese período de incubación, se contabilizó el número de manchas por fruto para determinar la severidad de la infección.

Los conteos (cantidad de manchas por fruto) se realizaron con el software *ImageJ* que es un programa de procesamiento digital de imágenes *(https://imagej.nih.gov/ij/)* para analizar las fotografías tomadas a cada fruto.

Los datos obtenidos se analizaron utilizando el software R versión 4.2.1

4.2 Análisis descriptivo de los datos

Se dispone de 360 observaciones de frutos en el archivo *"hsv"*. La variable respuesta de interés es la cantidad de manchas en el fruto, *Severidad*, que viene dada en función de las variables *Cultivar*, *Planta*, *Fruto* y *Age* (edad).

Las medidas resumen presentadas en la Tabla 4-1 muestran que la cantidad de manchas observadas tuvo un recorrido desde cero hasta 596 manchas por fruto,

Tabla 4-1. Medidas resumen de cantidad de manchas por fruto.

Mínimo	Mediana	Media	Máximo	Varianza
0.00	58.00	88.88	596	10329.9

La mediana de 58 indica que el 50% de los frutos tuvo como máximo 58 manchas y el 50% restante tuvo más de 58 manchas y el promedio de manchas por fruto fue de 88,88 manchas. La varianza muestral (10329.9) es bastante mayor que la media (88.88) lo que da cuenta de un problema de sobredispersión en los datos.

En la Figura 4.1 se representa la frecuencia de la severidad de la enfermedad, es decir, la cantidad de manchas en cada pera. Se puede notar en ella que los datos presentan una alta frecuencia de ceros (54) y una gran dispersión.

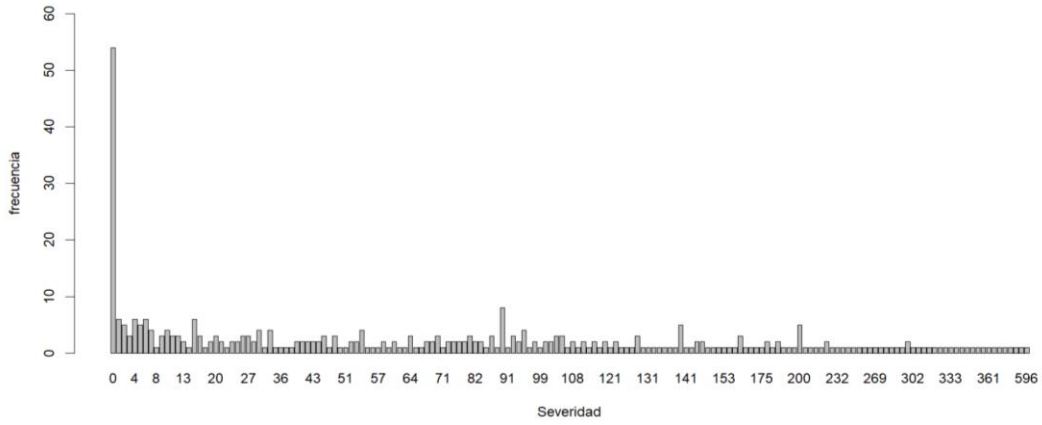

Figura 4.1. Distribución de frecuencias absolutas de la severidad de la infección por Stemphylium vesicarium, según los cuatro cultivares y las tres edades de observación.

En la Figura 4.2 se representa la frecuencia de la severidad de la enfermedad clasificada según los cuatro cultivares y las tres edades de observación. En ella se observa que las edades 1 y 2 presentan mayor variabilidad que en la edad 3 para los cultivares Abate Fetel, D'Anjou y Packhams. La mayor severidad se da en la edad 2 para todas los cultivares, excepto Williams. Vale recordar que la edad 1 se corresponde con 55-60 días después de plena floración (ddpl), la edad 2 con 85-90 ddpl y la edad 3 con 5 a 7 días antes del sello de autorización de cosecha de cada cultivar. En la misma figura se puede observar que el cultivar Williams es el menos afectado. Es este cultivar el que introduce los ceros al conjunto de datos, ya que el resto de los cultivares siempre registraron incidencia de la enfermedad con una severidad mayor a cero en todas sus edades.

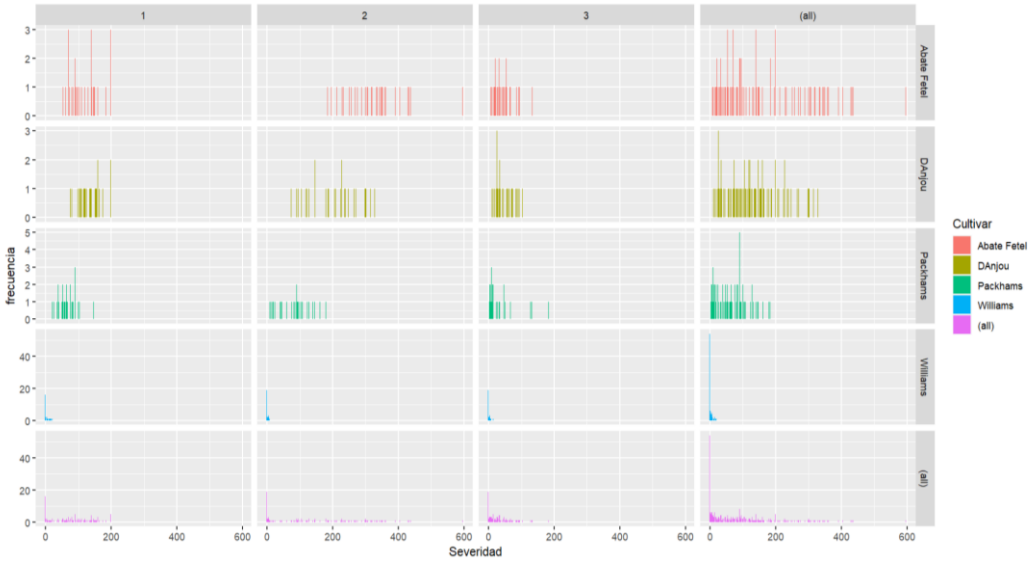

Figura 4.2. Distribución de frecuencias absolutas de la severidad de la infección por Stemphylium vesicarium según edad y cultivar.

La Figura 4.3 muestra otra forma de presentar la información, lo hace con diagramas de caja para cada edad según los cultivares. La información que suministra corrobora lo anteriormente dicho sobre la severidad de la enfermedad en la edad 2 de cada cultivar, salvo Williams que, con una media general de 2,54 manchas por fruto, tiene una leve mayor severidad en la edad 1.

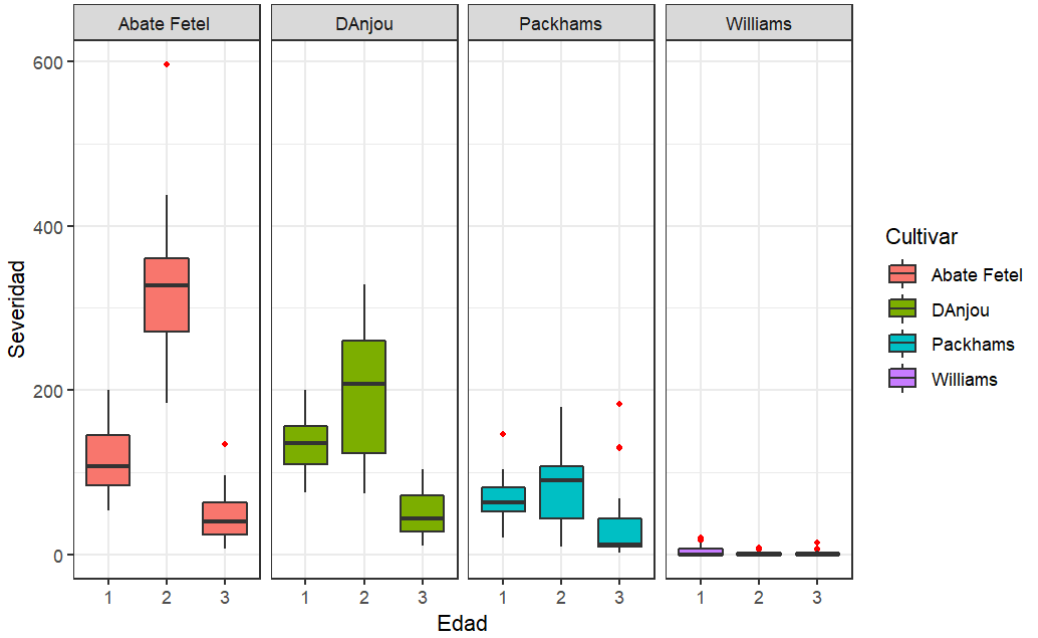

Figura 4.3. Box plot de la severidad de la enfermedad según edad y cultivar.

El cultivar más afectado es Abate Fetel con una media general de 164,62 manchas por fruto, le siguen D'Anjou con una media de 127,92 manchas por fruto y Packhams con una media de 60,44 manchas por fruto.

Comparando los conteos cero con el resto de los valores de la variable se observa en la Figura 4.4 que la proporción está en el límite (la línea roja indica un 15% de ceros, éste es un valor arbitrario de referencia) para considerar que la distribución de la variable tiene un exceso de ceros. Esto justificó la consideración de modelos inflados en cero y Hurdle para decidir al respecto, como así también modelar la sobredispersión.

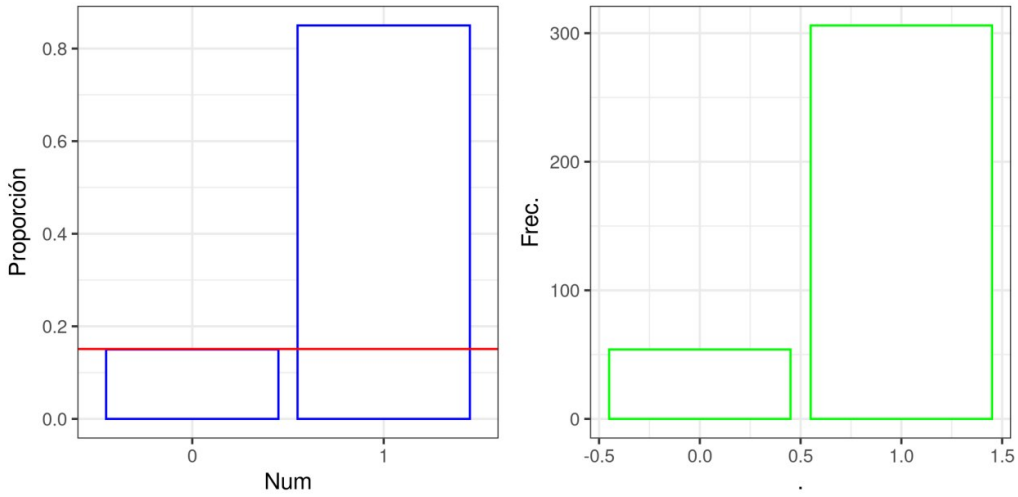

Figura 4.4. Proporción y frecuencia de ceros vs demás valores.

El ajuste de los distintos modelos que se describen en las secciones siguientes echa luz sobre las cuestiones antes mencionadas.

4.3 Ajuste de modelos

Inicialmente, se ajustaron modelos sin considerar efectos aleatorios. El primer modelo ajustado es el Poisson dado que es el modelo por excelencia para datos de conteo. El segundo modelo considerado es el Binomial Negativo (4.3.1 y 4.3.2).

En la sección 4.3.3 se consideran los modelos con ceros modificados para dar cuenta de la gran cantidad de ceros observados y analizar la conveniencia de un ajuste con modelos inflados en cero o con modelos Hurdle.

A continuación, en 4.3.4, se consideraron los efectos aleatorios y se ajustaron los modelos que propone la bibliografía consultada y referenciada en el marco teórico de este trabajo de tesis.

Por último, en 4.3.5 se ajustan diversos modelos usando el paquete “glmmTMB” donde se evidencian las ventajas de su implementación.

En el ANEXO se muestran los resúmenes de todos los modelos ajustados.

4.3.1 El modelo Poisson

Según lo visto en el apartado 2.1.1 sobre MLG y en particular 2.4.2 para el MRP

El modelo para analizar nuestro conjunto de datos es

$$ln(\mu_{ijk}) = \mu + \tau_i + \delta_j + (\tau * \delta)_{ij}\ ; \quad \boldsymbol{y_{ijk}} \sim \boldsymbol{Pois}(\boldsymbol{\mu_{ijk}})$$

donde

y_{ijk}: cantidad de manchas debido al i-ésimo cultivar en la j-ésima edad del k-ésimo fruto, $k = 1,2,\dots,5.$

μ: media general

τ_i: efecto del i-ésimo cultivar $i = 1,2,3,4.$ (Abate Fetel, D´Anjou, Packham, Williams)

δ_j: efecto de la j-ésima edad $j = 1,2,3.$

$(\tau * \delta)_{ij}$: efecto de la interacción del i-ésimo cultivar con la j-ésima edad

Para ajustar el Modelo 1 se utiliza la función “*glm*” del paquete base “*stats*”

Se observa en el resumen presentado en la Tabla 4-2 que la interacción Cultivar-edad es significativa pero este modelo presenta una dispersión importante. El cociente “devianza residual/grados de libertad” es mucho mayor a 1 (5936,1/348 = 17,06), lo mismo se verifica para el cociente χ^2 de Pearson /gl (6401,453/348 = 18,39).

Tabla 4-2. Modelo estimado Poisson MLG (Modelo 1) – deviance - χ^2 Pearson – AIC.

efecto	LR Chisq	grados de libertad (gl)	p-valor
Cultivar	20824.0	3	< 2.2 e-16 ***
edad	10239.2	2	< 2.2 e-16 ***
Cultivar:edad	1133.5	6	< 2.2 e-16 ***
deviance residual	5936.1	348	
χ^2 Pearson	6401.453	348	<0.001
AIC	7751.8		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Otra manera de visualizar la falta de ajuste es a través de los residuos de Pearson. En la Figura 4.5 se presentan los residuos de Pearson vs. predichos y en ella se puede observar que existe una gran cantidad de "ouliers".

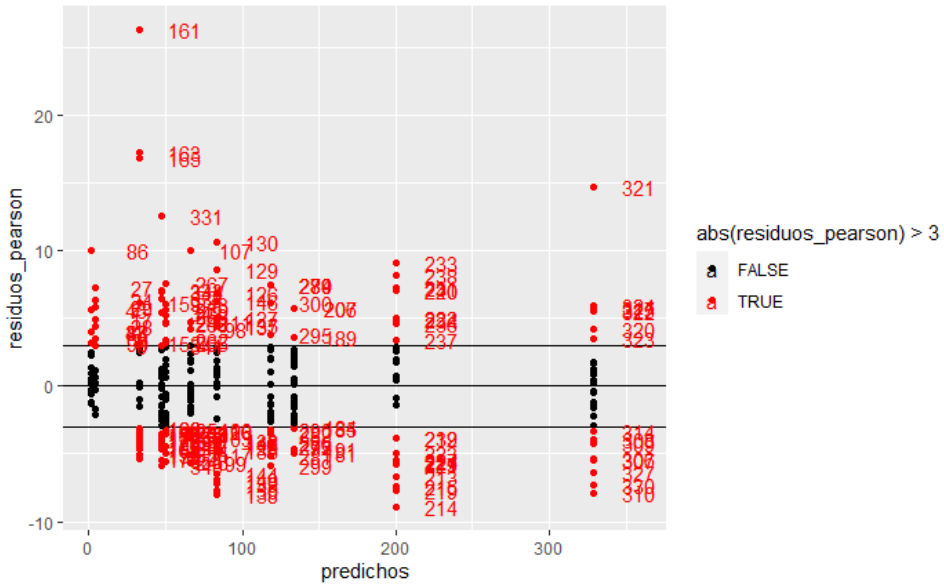

Figura 4.5. Residuos de Pearson vs predichos (Modelo 1).

También el rootograma del Modelo 1 representado en la Figura 4.6 muestra falta de ajuste. Se observa en particular un subajuste de ceros y el patrón ondulante de la parte inferior de las barras, por encima y por debajo de la línea de referencia cero, está indicando una falta general de ajuste del modelo.

Por todo esto se descarta que los datos sigan una distribución Poisson.

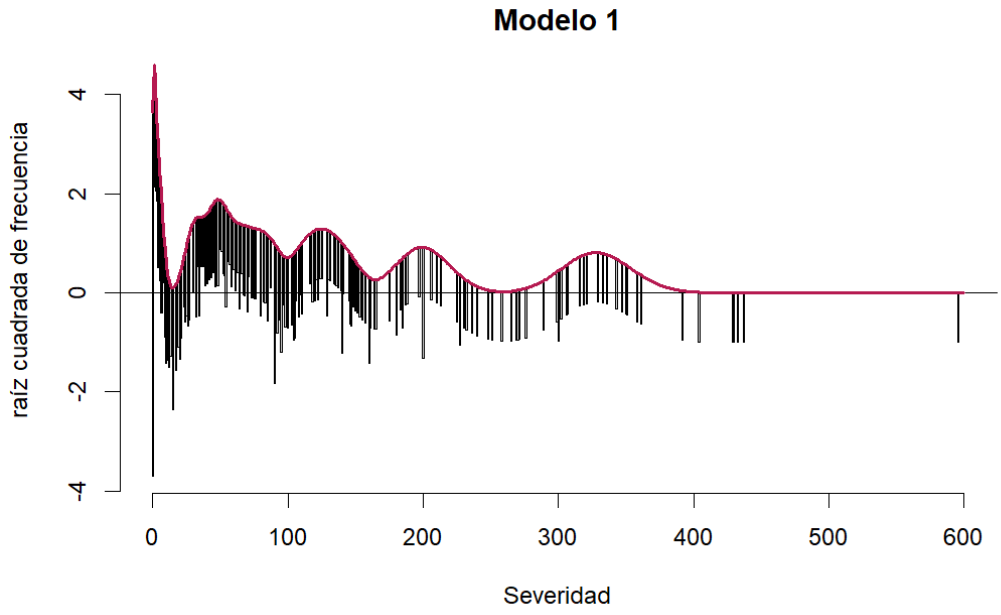

Figura 4.6. Rootograma del modelo Poisson (Modelo 1).

4.3.2 El modelo Binomial Negativo

Según lo visto en el apartado 2.6.1 el modelo se define igual que el MRP sólo que cambia su componente aleatoria.

Luego el modelo para analizar el conjunto de datos es:

$$ln(\mu_{ijk}) = \mu + \tau_i + \delta_j + (\tau * \delta)_{ij}\ ; \quad \boldsymbol{y_{ijk}} \sim \boldsymbol{NB(\mu_{ijk}, \alpha)}$$

media y varianza: $E(Y) = \mu$ $V(Y) = \mu + \alpha\mu^2$

Para ajustar un modelo Binomial Negativo se usó la función "glm.nb" del paquete “MASS” que estima el valor del parámetro de dispersión.

La Tabla 4-3 muestra la estimación de los coeficientes del modelo ajustado Binomial Negativo. Se observa en ella que este modelo logra modelar mejor la sobredispersión resultando el cociente entre la devianza residual y los grados de libertad más cercano a uno (472,72/348 = 1,36. Lo mismo ocurre con el cociente χ^2 Pearson/gl (501,114/348 = 1,44). Pero el test sigue indicando la existencia de sobredispersión.

Tabla 4-3. Modelo estimado Binomial Negativo MLG (Modelo 2) - deviance - χ^2 Pearson – AIC.

efecto	LR Chisq	grados de libertad (gl)	P(>Chisq)
Cultivar	1034.57	3	< 2.2 e-16 ***
edad	182.10	2	< 2.2 e-16 ***
Cultivar:edad	77.01	6	1.483 e-14 ***
deviance residual	472.72	348	
χ^2 Pearson	501.114	348	< 0.001 ***
AIC	3289.2		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Si se comparan las Figura 4.5 y Figura 4.7 se nota una notable reducción de los outliers respecto al modelo Poisson (Modelo 1).

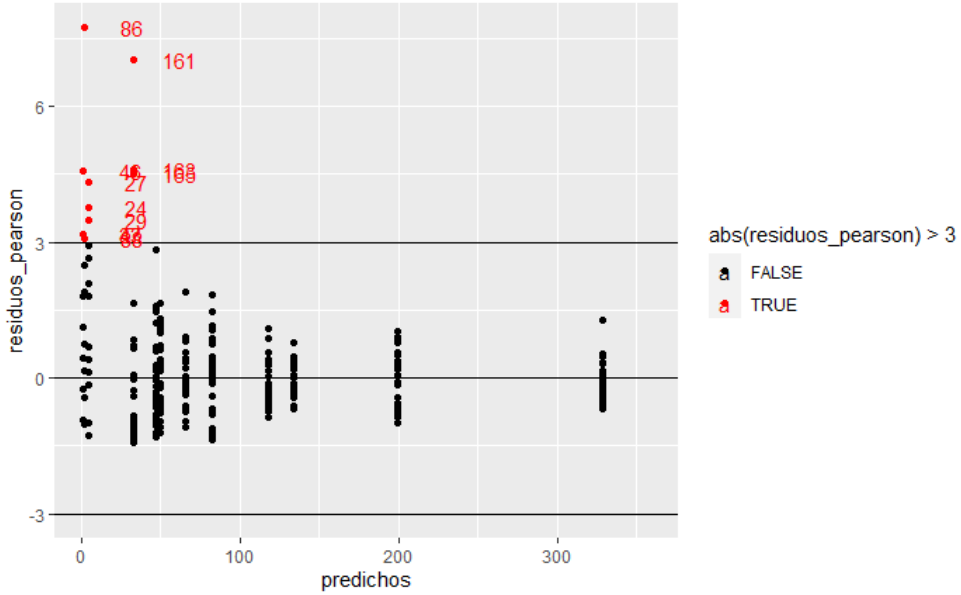

Figura 4.7. Residuos de pearson vs predichos (Modelo 2).

Sin embargo, el rootograma del Modelo 2 de la Figura 4.8, muestra los mismos problemas de ajuste que el modelo Poisson (Modelo 1). Aunque presenta una leve mejoría que se evidencia con un patrón menos ondulante de la parte inferior de las barras por encima y por debajo de la línea cero, sigue sin ajustar adecuadamente los ceros, los subajusta igual que a los valores de severidad mayores a 100.

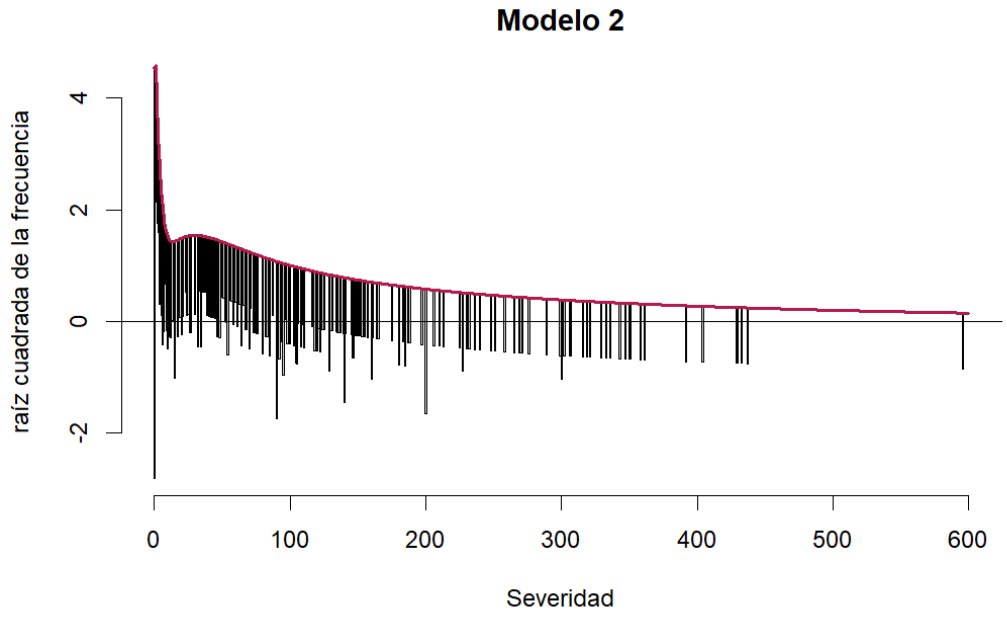

Figura 4.8. Rootograma modelo Binomial Negativo (Modelo 2).

Debido a esto último, en la sección siguiente se consideran modelos que tomen en cuenta el aparente exceso de ceros observados.

4.3.3 Modelos con ceros modificados

En el apartado 2.6.3 se explica que los datos de conteo con ceros modificados son aquellos que presentan un exceso de ceros o bien una falta de ellos.

Se ajustaron dos modelos inflados con ceros, el primero sin considerar los regresores para la parte del modelo inflado con ceros, Modelo 3 y el segundo considerando todos los regresores para esa parte del modelo, Modelo 4. En ninguno de los dos resultaron significativos los regresores de la parte del modelo inflado con ceros como se puede apreciar en las salidas publicadas.

Para ajustar un **modelo inflado en ceros** usamos la función “zeroinfl” del paquete “pscl” (pscl no ajusta BN1)

Los rootogramas de cada modelo representados en las figuras Figura 4.9 y Figura 4.10 evidencian la falta de ajuste de los modelos. En ambos rootogramas se puede observar que para valores de severidad mayores a 100 la parte inferior de las barras queda por debajo de la línea cero, lo cual indica un subajuste de modelo

para esos valores. De manera análoga, para valores inferiores a 100, varias barras quedan por encima de la línea cero, lo que indica un sobreajuste.

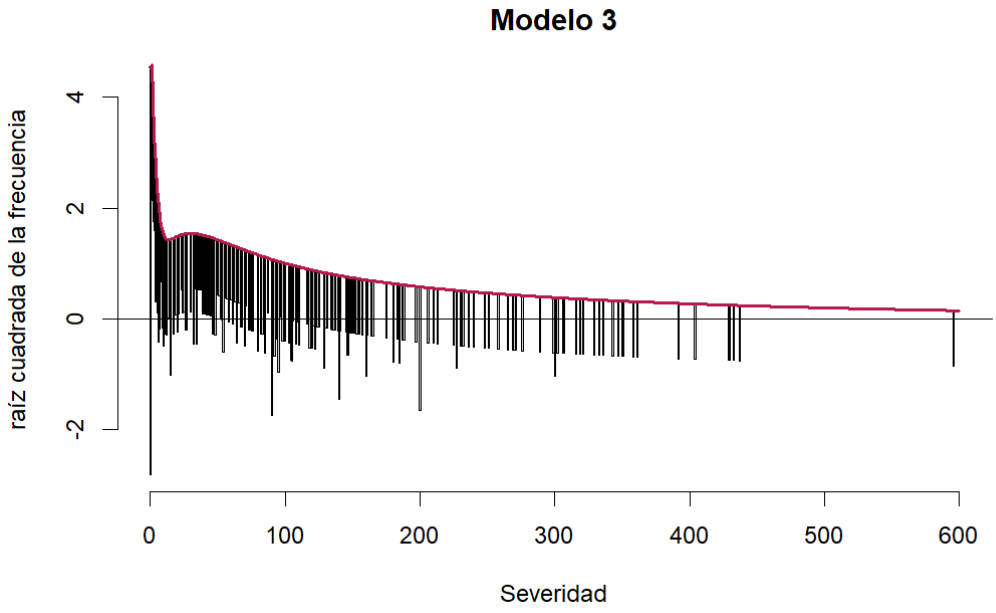

Figura 4.9. Rootograma modelo binomial negativo inflado en ceros, sólo en el intercepto (Modelo 3).

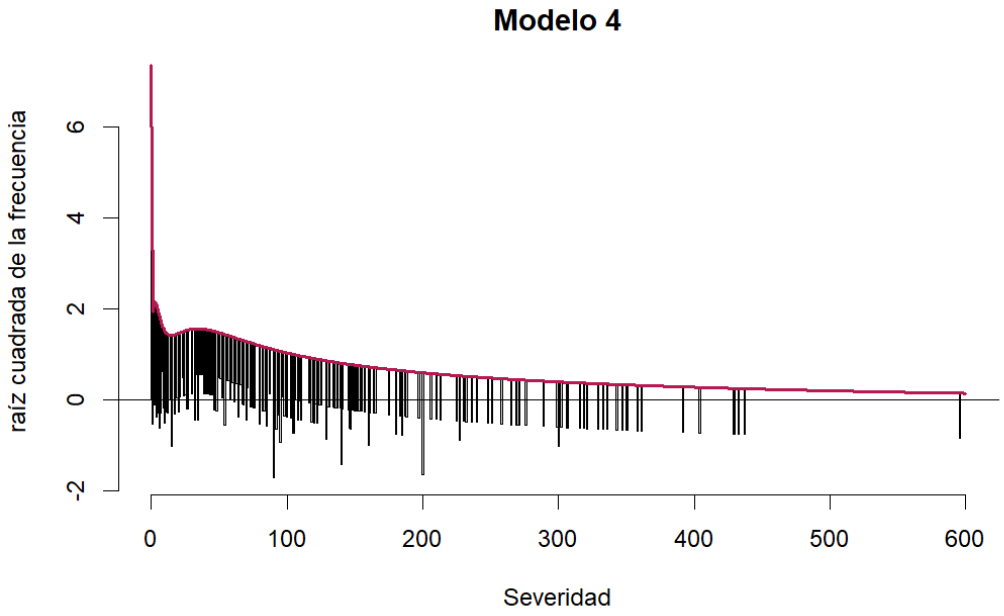

Figura 4.10. Rootograma modelo binomial negativo inflado en cero, en todos los regresores (Modelo 4).

Siguiendo con la consideración de los modelos con ceros modificados se ajustó un modelo Binomial Negativo Hurdle, que en el ANEXO se presenta la salida como Modelo 5.

Para ajustar un **modelo Hurdle** se usó la función "*hurdle*" del paquete "*pscl*"

No resultó significativo ningún regresor de la parte del modelo hurdle como se puede apreciar en las salidas publicadas.

Observando el rootograma del Modelo 5 (mostrado en el ANEXO) en la Figura 4.11 se nota un subajuste del modelo para la mayoría de los valores de severidad, algo similar a lo que ocurre cuando se ajustan los modelos inflados en cero (Modelo 3 y Modelo 4).

Luego, por todo lo analizado anteriormente no podemos considerar que la generación de ceros se deba a un proceso diferente al Binomial Negativo.

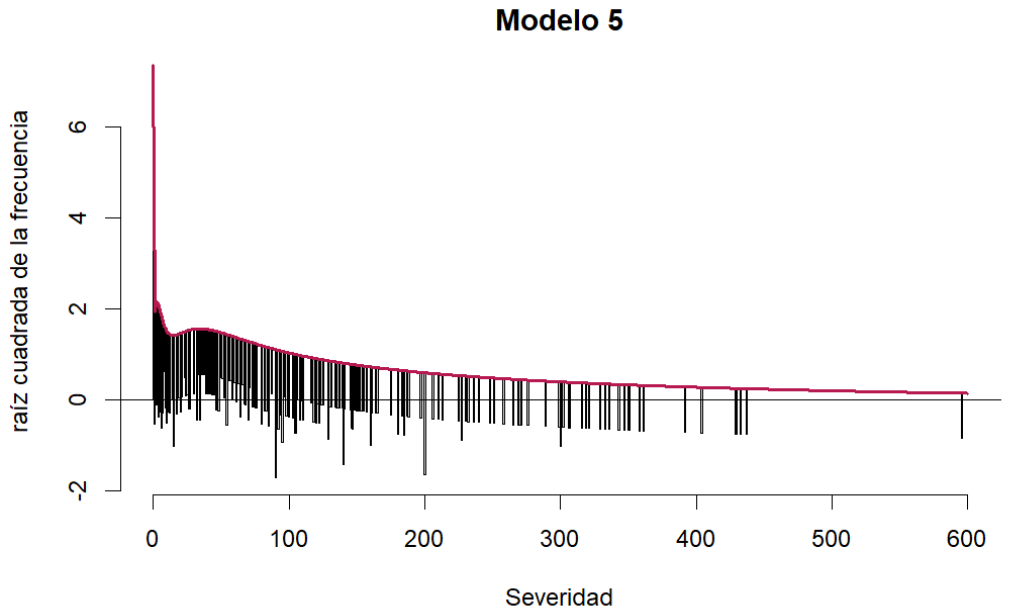

Figura 4.11. Rootograma modelo binomial negativo hurdle (Modelo 5).

A continuación, se consideran efectos aleatorios para ver si se logra modelar la correlación dado que se hizo una medición en el tiempo sobre la misma planta.

4.3.4 Efectos aleatorios: MLGM

Según lo visto en el apartado 2.3, se considera un nuevo modelo (Modelo 6) que incorpore el efecto "planta" como aleatorio.

$$ln(\mu_{ijkl}) = \mu + \tau_i + \delta_j + (\tau * \delta)_{ij} + p_k$$

$$Y_{ijkl}|p_k \sim BN(\mu_{ijkl}, \alpha); \qquad p_k \sim iid\ N(0, \sigma_p^2)$$

donde

y_{ijkl}: cantidad de manchas debido al i-ésimo cultivar, a la j-ésima edad y al efecto aleatorio de la k-ésima planta en el l-ésimo fruto, $l = 1,2,3,4,5$.

μ: media general

τ_i: efecto del j-ésimo cultivar $i = 1,2,3,4$. (Abate Fetel, D´Anjou, Packham, Williams)

δ_j: efecto de la j-ésima edad $j = 1,2,3$.

$(\tau * \delta)_{ij}$: efecto de la interacción del i-ésimo cultivar con la j-ésima edad

p_k: efecto de la k-ésima planta $k = 1,2,\dots,6$.

Para ajustar el modelo se usa la función "glmer.nb" del paquete "lme4".

En los MLGM se debe considerar la devianza condicional y no la marginal (que es la que da el "*summary"* por defecto en R). La distribución marginal no es Binomial Negativa, la que es Binomial Negativa es la condicional dados los efectos aleatorios.

En la Tabla 4-4 se muestra un resumen de la salida del Modelo 6. En ella se observa nuevamente el cociente devianza condicional/grados de libertad es más cercano a uno (447,232/346=1,29) que con el Modelo 2 sin efectos aleatorios. (1,29 vs 1,36). Lo mismo se puede observar con el cociente χ^2 Pearson/gl (452.736/346 = 1.308).

Tabla 4-4. Modelo binomial negativo con efectos aleatorios (Modelo 6) - deviance - χ^2 Pearson – AIC.

Efectos aleatorios:			
Groups Name	Varianza	Sts.Dev	
Planta (Intercepto)	0.1184	0.3441	
Efectos fijos:			
	Chisq	grados de libertad (gl)	P(>Chisq)
Cultivar	385.273	3	<2.2 e-16 ***
Edad	273.356	2	<2.2 e-16 ***
Cultivar:edad	71.382	6	2.128 e-13 ***
deviance condicional	447.23	346	
χ^2 Pearson	452.736	346	< 0.001
AIC	3257.2		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

La comparación de los AIC de los modelos con y sin efecto aleatorio arroja una disminución en el modelo con efecto aleatorio (3289,154 vs 3257,235) y siguiendo el criterio de (Hilbe, 2011) descripto en la sección 2.1.3.2, como la diferencia entre los AICs de ambos modelos es superior a 10 puntos, se prefiere este último modelo frente al que no se le adicionaron efectos aleatorios.

Tabla 4-5. Diferencia de AICs de modelos NB con y sin efecto aleatorio (Modelo 6 vs Modelo 2).

		diferencia AIC	grados de libertad
hsv_glmer.bn_p	Modelo 6	0.0	14
hsv_glm.bn	Modelo 2	31.9	13

Analizando los gráficos de residuos de deviance y de Pearson en las Figura 4.12 y Figura 4.13 podemos decir que el modelo Binomial Negativo en el marco de los MLGM (Modelo 6) presenta un buen ajuste, aunque se observan patrones (en ambas figuras) que muestran, para valores grandes y/o pequeños, una falta de ajuste.

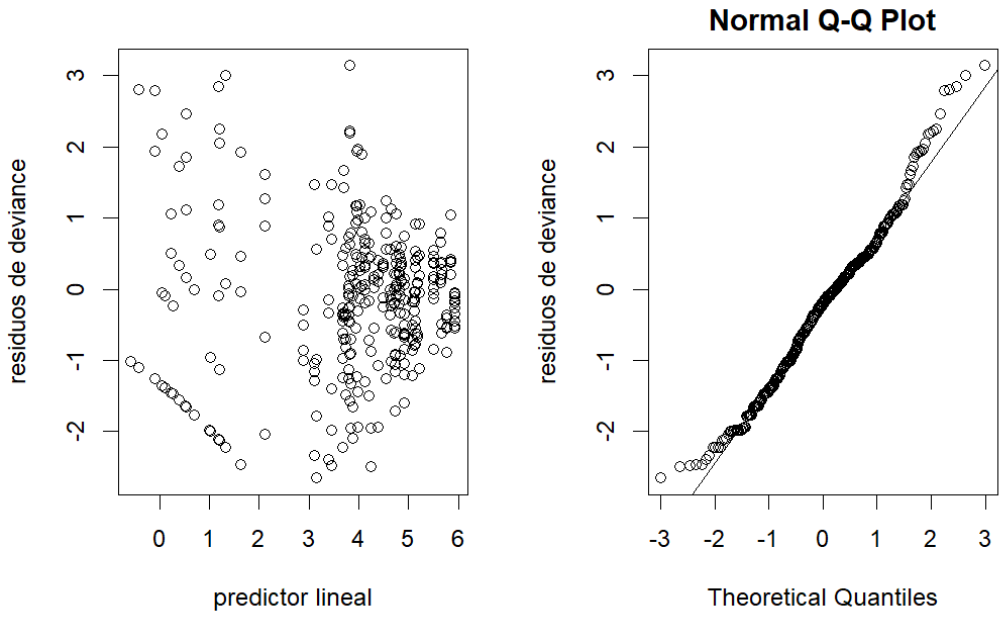

Figura 4.12. Residuos de deviance (Modelo 6).

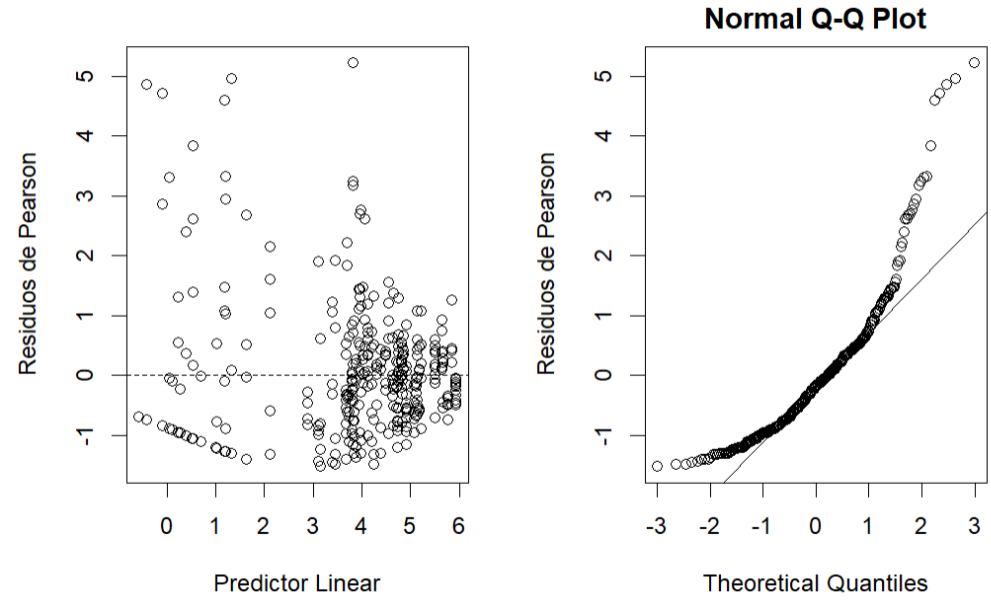

Figura 4.13. Residuos de Pearson (Modelo 6).

Se consideraron también otros dos modelos cuyas salidas están en el ANEXO.

- Modelo 7 con efecto aleatorio planta más interacción planta-fruto
- Modelo 8 con efecto aleatorio en el cultivar, es decir tanto la pendiente como la ordenada aleatorios

Ambos modelos se descartan por resultar un ajuste “singular” es decir, matrices de varianza-covarianza estimadas con un rango inferior al completo. Menos técnicamente, esto significa que algunas "dimensiones" de la matriz de varianza-covarianza se han estimado como exactamente cero.

Bates, Mächler, Bolker, & Walker (2014) observan que si bien los modelos singulares están bien definidos estadísticamente (es teóricamente sensato que la verdadera estimación de máxima verosimilitud corresponda a un ajuste singular), existen preocupaciones reales de que (1) los ajustes singulares correspondan a modelos sobreajustados que pueden tener poca potencia; (2) las posibilidades de problemas numéricos y falta de convergencia son mayores para modelos singulares (por ejemplo, puede ser computacionalmente difícil calcular los intervalos de confianza del perfil para tales modelos); (3) los procedimientos de inferencia estándar, como las estadísticas de Wald y las pruebas de razón de verosimilitud, pueden ser inapropiados.

En la próxima sección, el uso de un nuevo paquete permite evaluar el ajuste de otros modelos.

4.3.5 Ajuste de modelos con el paquete glmmTMB

Tal como se detalló en el apartado 3.1, el paquete *glmmTMB* otorga una ventaja de velocidad frente a *lme4* al estimar modelos no gaussianos y brinda una mayor flexibilidad en las clases de distribuciones que puede ajustar. Esta flexibilidad es la que permite ajustar y comparar muchas variedades de modelos con la seguridad de que los valores de verosimilitud se calculan de manera consistente. Otras de las ventajas de este paquete glmmTMB es que permite modelar tanto la dispersión como la inflación en cero para aquellos modelos como el Binomial Negativo que tienen parámetros de dispersión.

Se usó la función *glmmTMB* del paquete del mismo nombre para ajustar modelos de las familias Binomial Negativo 1 (*nbinom1*), Binomial Negativo 2 (*nbinom2*) y Conway-Maxwell-Poisson (*compois*). Para cada uno de estos modelos se consideraron distintos ajustes para la fórmula de dispersión y se descartó la inflación con ceros por lo analizado y concluido previamente en 4.3.3.

De acuerdo con la tabla sugerida por Hilbe, detallada en 2.1.3.2 y, teniendo en cuenta que los datos obtenidos suman 360, se observó que, según la tabla de diferencia de AICs, de los nueve modelos ajustados, no hay diferencia entre los dos primeros que se muestran en la Tabla 4-6:

Tabla 4-6. Diferencia de AICs de modelos ajustados con el paquete glmmTMB.

modelo		diferencia de AIC	grados de libertad
i.	Modelo 11	0.0	25
ii.	Modelo 14	0.3	25
iii.	Modelo 10	8.7	17
iv.	Modelo 9	13.3	14
v.	Modelo 17	37.1	25
vi.	Modelo 16	54.3	17
vii.	Modelo 13	56.1	17
viii.	Modelo 15	69.4	14
ix.	Modelo 12	193.6	14

A continuación, se muestran los detalles de la fórmula para cada modelo usando la función *glmmTMB*

i. Modelo 11: glmmTMB(Severidad ~ Cultivar + Age + Cultivar:Age + (1|Planta), ziformula = ~0, dispformula = ~ Cultivar*Age, family = nbinom1, data=hsv)

ii. Modelo 14: glmmTMB(Severidad ~ Cultivar+Age + Cultivar:Age + (1|Planta), ziformula = ~0, dispformula = ~ Cultivar*Age, family = nbinom2, data=hsv)

iii. Modelo 10: glmmTMB(Severidad ~ Cultivar+Age + Cultivar:Age + (1|Planta), ziformula = ~0, dispformula = ~ Cultivar, family = nbinom1, data=hsv)

iv. Modelo 9: glmmTMB(Severidad ~ Cultivar+Age + Cultivar:Age + (1|Planta), ziformula = ~ 0, dispformula = ~1, family = nbinom1, data=hsv)

v. Modelo 17: glmmTMB(Severidad ~ Cultivar+Age + Cultivar:Age + (1|Planta), ziformula= ~0, dispformula= ~Cultivar*Age, family=compois, data = hsv)

vi. Modelo 16: mglmmTMB(Severidad ~ Cultivar+Age + Cultivar:Age + (1|Planta), ziformula=~0, dispformula=~Cultivar, family=compois, data = hsv)

vii. Modelo 13: glmmTMB(Severidad ~ Cultivar+Age + Cultivar:Age + (1|Planta), ziformula=~0, dispformula=~Cultivar, family=nbinom2, data = hsv)

viii. Modelo 15: glmmTMB(Severidad ~ Cultivar+Age + Cultivar:Age + (1|Planta), ziformula=~0, dispformula=~1, family=compois, data = hsv)

ix. Modelo 12: glmmTMB(Severidad ~ Cultivar+Age + Cultivar:Age + (1|Planta), ziformula=~0, dispformula=~1, family=nbinom2, data = hsv)

Para tener en cuenta:

- *ziformula=~0,* especifica que no hay inflación cero.
- *dispformula=~1,* especifica la dispersión estándar dada cualquier familia.
- *dispformula=~Cultivar,* especifica que la dispersión se debe al Cultivar.
- *dispformula=~Cultivar*Age,* especifica que la dispersión se debe a la interacción Cultivar-Edad.

En los dos primeros modelos seleccionados (Modelo 11 y Modelo 14), como con mejor ajuste, la dispersión se debe a la interacción del cultivar con la edad.

4.3.6 Modelo final: descripción e inferencia

A continuación, se presenta el resumen del modelo elegido (recordar que las salidas completas se muestran en el ANEXO) que denominamos Modelo 18 puesto que se volvió a ajustar el Modelo 11 considerando REML para estimar los componentes de varianza del modelo sin que éstos se vean afectados por los efectos fijos, ya que de otra manera se producirían estimaciones sesgadas (usando LM).

Modelo 18: *glmmTMB(Severidad ~ Cultivar + Age + Cultivar:Age + (1|Planta), ziformula = ~0, dispformula = ~ Cultivar*Age, family = nbinom1, REML = TRUE, data=hsv)*

Este modelo permite que la media condicional dependa del cultivar, la edad, la interacción cultivar-edad y que varíe aleatoriamente según las plantas. Posibilita que el parámetro de dispersión dependa de la interacción cultivar-edad. Este modelo se puede representar mediante el siguiente conjunto de ecuaciones:

$$\mu = E(Severidad|p) = exp\big(\beta_0 + \beta_{CultivarDAnjou} + \beta_{CultivarPackhams} + \beta_{CultivarWilliams} + \beta_{Age2} + \beta_{Age3} + \beta_{CultivarDAnjou:Age2} + \beta_{CultivarPackhams:Age2} + \beta_{CultivarWilliams:Age2} + \beta_{CultivarDAnjou:Age3} + \beta_{CultivarPackhams:Age3} + \beta_{CultivarWilliams:Age3} + p\big) \quad \boldsymbol{ec.1}$$

$$p \sim N\big(0, \sigma_p^2\big) \quad \mathbf{ec.2}$$

$$\sigma^2 = Var(Severidad|p) = \mu + \mu/\theta \quad \boldsymbol{ec.3}$$

$$ln(\theta) = \beta_0^{(disp)} + \Big[\beta_{CultivarDAnjou}^{(disp)} + \beta_{CultivarPackhams}^{(disp)} + \beta_{CultivarWilliams}^{(disp)} + \beta_{Age2}^{(disp)} + \beta_{Age3}^{(disp)} + \beta_{CultivarDAnjou:Age2}^{(disp)} + \beta_{CultivarPackhams:Age2}^{(disp)} + \beta_{CultivarWilliams:Age2}^{(disp)} + \beta_{CultivarDAnjou:Age3}^{(disp)} + \beta_{CultivarPackhams:Age3}^{(disp)} + \beta_{CultivarWilliams:Age3}^{(disp)}\Big] \cdot Planta \quad \boldsymbol{ec.4}$$

donde p es el efecto aleatorio específico de las plantas y β_i son los coeficientes de regresión (β_0 denota la intersección).

El resumen del Modelo 18 que se presenta en la Figura 4.14 se puede dividir en cuatro secciones. La sección 1 es una descripción general que contiene una descripción de la especificación del modelo (Familia, Fórmula, Dispersión, Datos) y los criterios de información resultantes. En la línea superior del resumen se indica la función de enlace del modelo (log). La sección 2 describe la variabilidad de los efectos aleatorios. En este modelo, solo tuvimos efectos aleatorios en el modelo condicional (ec. 1). La varianza estimada 0,02811 es σ_p^2 en la ec. 2. La sección 3 describe los coeficientes del modelo condicional (los β_i en la ec. 1) incluyendo las estadísticas de Wald Z y los *valores p*. Aparte del intercepto, las estimaciones son todos contrastes como es estándar en los modelos de regresión. La sección 4 proporciona coeficientes estimados del modelo de dispersión (ec. 4), que utiliza un enlace logarítmico para mantener el parámetro de dispersión θ positivo (ec.4). Por el contrario, un modelo con el modelo de dispersión predeterminado (simple) (dispformula=~1) informaría el parámetro de dispersión

único (igual para todas las observaciones) en la escala natural (en lugar de logarítmica).

```
Family: nbinom1  ( log )

Formula:     Severidad ~ Cultivar + Age + Cultivar:Age + (1 | Planta)
Dispersion:                 ~Cultivar * Age
Data: hsv

     AIC      BIC   logLik deviance df.resid
  3092.1   3189.3  -1521.1   3042.1      347
```

sección 1

```
Random effects:

Conditional model:
 Groups Name        Variance Std.Dev.
 Planta (Intercept) 0.02811  0.1676
Number of obs: 360, groups:  Planta, 24
```

sección 2

```
Conditional model:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           4.76834    0.09942   47.96  < 2e-16 ***
CultivarDAnjou        0.11995    0.12634    0.95    0.342
CultivarPackhams     -0.57964    0.14013   -4.14 3.53e-05 ***
CultivarWilliams     -3.25881    0.39862   -8.18 2.95e-16 ***
Age2                  1.02071    0.07968   12.81  < 2e-16 ***
Age3                 -0.91440    0.13951   -6.55 5.58e-11 ***
CultivarDAnjou:Age2  -0.61947    0.10563   -5.86 4.51e-09 ***
CultivarPackhams:Age2 -0.78958    0.15526   -5.09 3.67e-07 ***
CultivarWilliams:Age2 -2.21945    0.54254   -4.09 4.30e-05 ***
CultivarDAnjou:Age3  -0.06462    0.17188   -0.38    0.707
CultivarPackhams:Age3  0.21334    0.24764    0.86    0.389
CultivarWilliams:Age3 -0.04742    0.57308   -0.08    0.934
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

sección 3

```
Dispersion model:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           2.85609    0.29863   9.564  < 2e-16 ***
CultivarDAnjou       -1.34332    0.45858  -2.929  0.00340 **
CultivarPackhams     -0.65771    0.44269  -1.486  0.13736
CultivarWilliams      0.07322    0.60309   0.121  0.90336
Age2                 -0.50150    0.52566  -0.954  0.34007
Age3                  0.10474    0.41640   0.252  0.80139
CultivarDAnjou:Age2   1.96414    0.69366   2.832  0.00463 **
CultivarPackhams:Age2  1.72877    0.69274   2.496  0.01258 *
CultivarWilliams:Age2 -0.83240    0.95445  -0.872  0.38314
CultivarDAnjou:Age3   0.87763    0.63034   1.392  0.16383
CultivarPackhams:Age3  1.25913    0.61949   2.033  0.04210 *
CultivarWilliams:Age3 -1.00533    0.89061  -1.129  0.25898
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

sección 4

Figura 4.14. Resumen del Modelo 18.

Se puede observar en este resumen que al modelar el parámetro de dispersión no solo en el intercepto sino también en la interacción, la varianza estimada del efecto aleatorio plantas ($\sigma_u^2 = 0{,}02811$), disminuye respecto de la del Modelo 6 ($\sigma_u^2 = 0{,}1184$).

En el gráfico de residuos de Pearson del Modelo 18 (Figura 4.15) se puede observar un mejor ajuste de este modelo respecto del ajustado sin usar el paquete glmmTMB (Modelo 18, Figura 4.15 vs Modelo 6, Figura 4.13).

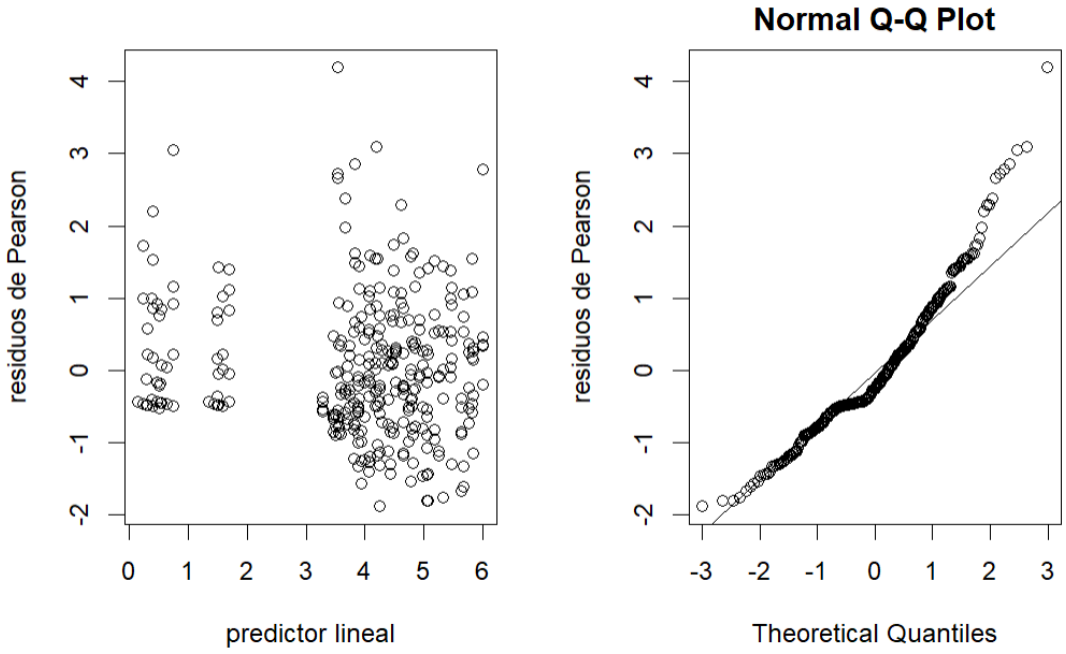

Figura 4.15. Análisis de residuos de Pearson (Modelo 18)

En la Tabla 4-7 se presentan las medias estimadas (en escala logarítmica) para cada cultivar según la edad, además de los desvíos estándar y los límites inferior y superior de los intervalos de confianza

Tabla 4-7. Medias estimadas - Errores estándar - Límites de intervalos de confianza - Test de medias - Escalas logarítmicas.

Cultivar	edad	medias	error estándar	lím inf	lím sup	grupos
Williams	2	0.311	0.3853	-0.798	1.42	a
Williams	3	0.548	0.4116	-0.636	1.73	a
Williams	1	1.510	0.3860	0.399	2.62	a
Packhams	3	3.488	0.2038	2.901	4.07	b
Abate Fetel	3	3.854	0.1378	3.458	4.25	bc
DAnjou	3	3.909	0.1158	3.576	4.24	bc
Packhams	1	4.189	0.0988	3.905	4.47	c
Packhams	2	4.420	0.1320	4.040	4.80	cd
Abate Fetel	1	4.768	0.0994	4.482	5.05	d
DAnjou	1	4.888	0.0780	4.664	5.11	d
DAnjou	2	5.290	0.0902	5.030	5.55	e
Abate Fetel	2	5.789	0.0767	5.569	6.01	f

Puede observarse que, debido a la interacción cultivar-edad, se formaron seis grupos diferentes. El cultivar William's en todas sus edades es la que presenta

menor severidad de la enfermedad y forma un grupo puro (a), el mismo comportamiento, pero en el otro extremo, se observa en los cultivares D'Anjou y Abate Fetel que en la edad 2 presentan la mayor severidad y forman cada uno de ellos grupos puros (e) y (f) respectivamente. D'Anjou y Abate Fetel en la edad 1 tienen comportamiento similar y son los que presentan mayor severidad en esa edad; lo mismo ocurre para ambos cultivares en la edad 3.

Lo concluido a partir de la Tabla 4-7 se puede observar también en la Figura 4.16.

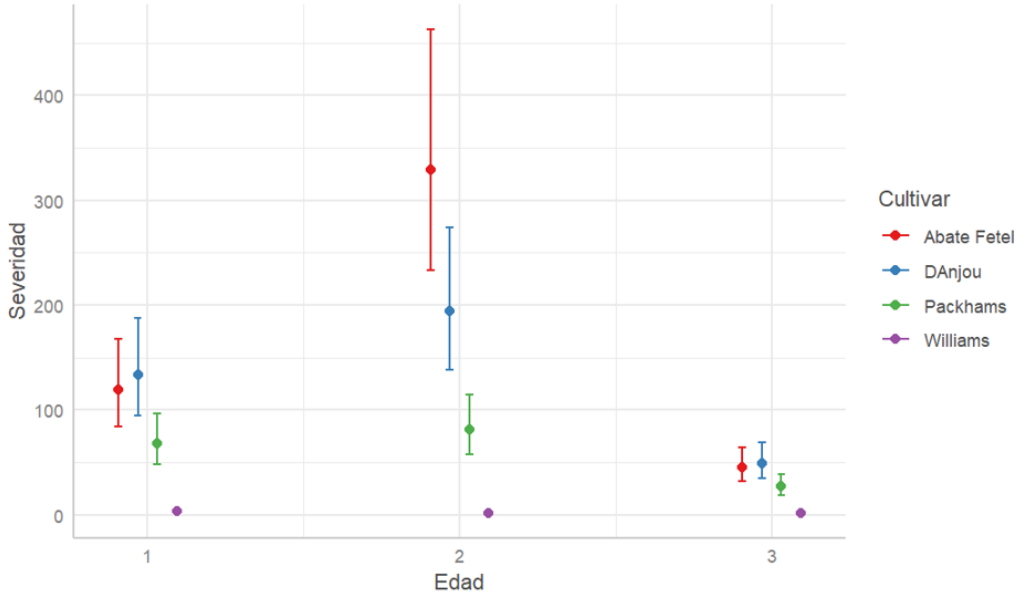

Figura 4.16. Intervalos para la respuesta media de conteos predichos según Modelo 18, estimados a partir de los errores estándar, según edad y cultivar.

Aplicando la prueba de Tukey para comparación de medias podemos observar en la Figura 4.17 que en la edad 1 no resultaron significativas las diferencias de medias de los cultivares D'Anjou y Abate Fetel, lo mismo ocurre en la edad 3 con las medias de los cultivares Packham's, D'Anjou y Abate Fetel. En cada edad, el comportamiento de la enfermedad no se diferenció en esos cultivares. Por el contrario, en la edad 2 todos los cultivares tuvieron comportamientos diferentes, aunque (salvo William's) valores medios estimados muy altos.

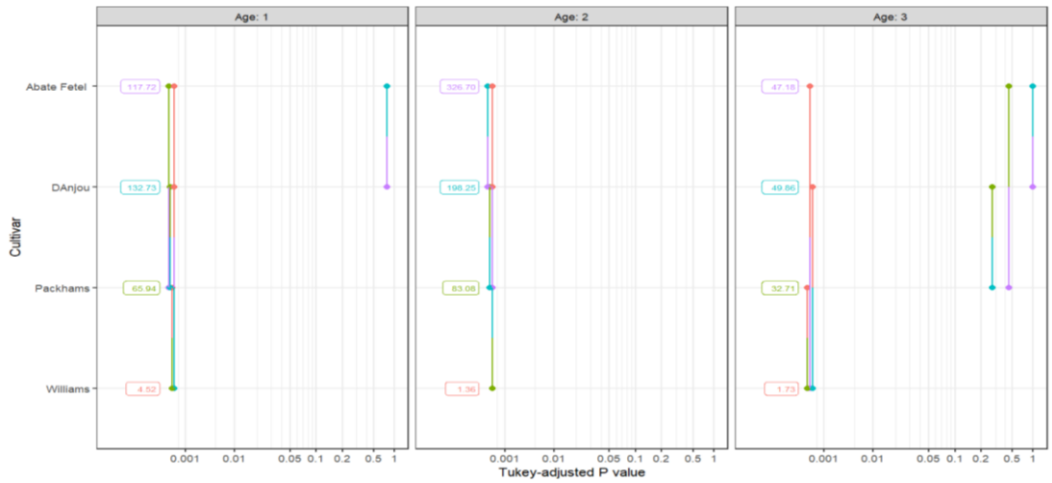

Figura 4.17. Comparación de pares de medias entre Cultivar según edad aplicando test de Tukey

La Figura 4.18 muestra el desvío que produce cada planta en la estimación del intercepto. La plantas que más desvío positivo aportan son Planta 5 del cultivar Abate Fetel, Planta 24 del cultivar Williams y Planta 18 del cultivar D'Anjou.

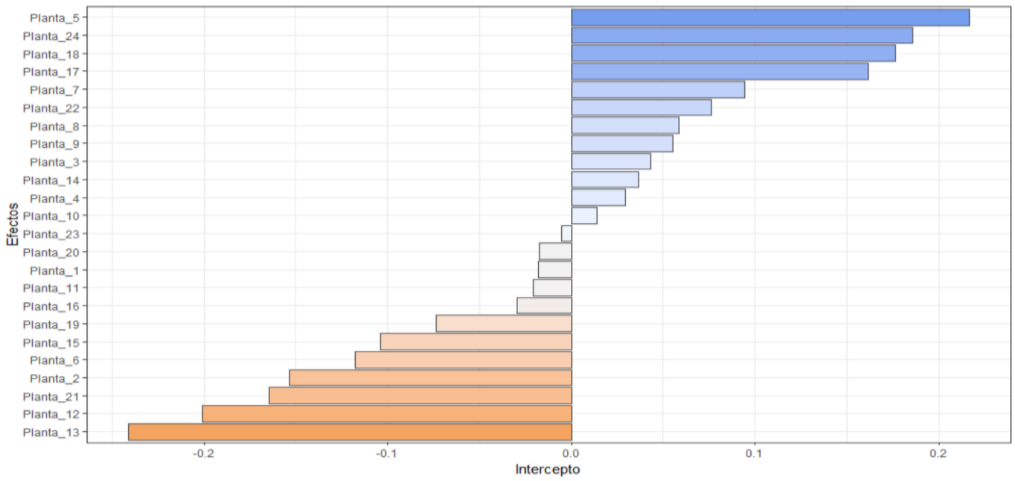

Figura 4.18. Digresión de los efectos aleatorios debido a la planta respecto al intercepto del modelo.

Mientras que las plantas que más desvío negativo aportan son Planta 13 del cultivar D'Anjou, Planta 12 del cultivar Packhams y planta 21 del cultivar Williams. Por su parte entre las plantas que menos desvío aportan se pueden mencionar a la Planta 1 del cultivar Abate Fetel, Planta 20 del cultivar Williams, Planta 10 del cultivar Packhams y Planta 23 del cultivar Williams.

Esta figura evidencia la variabilidad que introducen las plantas lo cual pone de manifiesto la necesidad de incorporar el efecto aleatorio de las mismas a la parte sistemática del modelo. Es por ello por lo que los MLGM se presentan como una herramienta muy apropiada para el análisis de este tipo de datos.

5 Discusión

La reciente aparición en la zona del Alto Valle de Río Negro y Neuquén (Argentina) de la enfermedad “mancha marrón del peral” producida por el hongo *Stemphylium vesicarium* ha motivado el estudio de la importancia de la enfermedad y el comportamiento varietal de los principales cultivares a la infección del patógeno. Como resultado de la cuantificación de la enfermedad, la principal característica es que los datos resultan de conteos. Estos conteos pueden ser número de frutos afectados o bien número de manchas por fruto. Al primer tipo de conteo, en fitopatología, se lo denomina incidencia y al segundo severidad. En esta tesis se analizó la severidad.

La posibilidad de realizar fotografías de alta calidad y el surgimiento de softwares para conteos (de eventos registrados en las imágenes) ha permitido recolectar grandes cantidades de datos y en poco tiempo. Los datos de conteo pueden encontrarse prácticamente en cualquier ámbito de aplicación, pero no existe una práctica generalizada en cuanto al tipo de modelo de regresión que deba utilizarse, aun cuando existe una gran variedad de modelos disponibles.

Es importante tomar en cuenta el hecho de que la media y la varianza para una variable de conteo no son casi nunca iguales y en la mayoría de los casos la varianza excede a la media ocasionando dispersión excesiva en los datos.

Cuando sólo se considera la distribución Poisson para ajustar datos de conteo, como es el caso del Modelo 1 en esta tesis, no debe perderse de vista que, aunque el modelo muestre ajustes significativos para todos sus parámetros, éstos están sobreajustados debido a la sobredispersión que no pudo modelarse. Este comportamiento ha sido observado por diversos autores, (Bolker, 2015; Cameron & Trivedi, 2013; Hilbe, 2011) entre otros, quienes recomiendan el uso de otros modelos para ajustar los datos que den cuenta de la sobredispersión, como por ejemplo el modelo Binomial Negativo.

En el caso en que se consideraron modelos con sólo efectos fijos, el modelo Binomial Negativo (NB), en el marco de los MLG, logra un mejor ajuste frente al modelo Poisson (Modelo 2 vs. Modelo 1), mostrando así que la distribución BN toma en cuenta parte de la sobredispersión que presentan los datos, lo cual

concuerda con estudios de similares características, (Menni, 2011; Hilbe, 2017; Alcaide Delgado, 2015; Basanta, Rubio, Giménez, & Tudela, 2021).

Una de las posibles causas de sobredispersión que se consideró fue el exceso de ceros dando lugar al estudio de los denominados *modelos con ceros modificados*. Son numerosos los trabajos publicados que dan cuenta de esta problemática, por ejemplo (Bolker, 2017; Heilbron, 1989; Hachuel, Boggio, & Harvey, 2010; Pol & Del Río, 2005; Llorens, 2005). Cada uno de estos autores en sus libros y/o publicaciones proponen el uso de modelos inflados en cero y modelos Hurdle con los que demostraron obtener mejores ajustes para los datos de conteo frente al modelo Poisson, al considerar la sobredispersión que generan los ceros presentes en los datos, ya sea en exceso o en defecto.

Cuando los modelos con ceros modificados tampoco logran dar cuenta de la sobredispersión observada, la adición de efectos aleatorios a la parte sistemática del modelo es la opción a seguir, tal y como lo sugieren diversos autores y publicaciones al respecto, por ejemplo (Bolker, y otros, 2009; Hilbe, 2017; McCulloch & Searle, 2001).

Existen numerosos trabajos publicados que se valen de los Modelos Mixtos o Modelos Mixtos no Lineales para incorporar efectos aleatorios que describan mejor el comportamiento de sus datos, por ejemplo (Rubio, 2016; Tassile, 2020). Sin embargo, la mayoría de las publicaciones encontradas y analizadas para este trabajo de tesis, que tratan sobre variables de respuesta de conteo, por ejemplo: (Azpilicueta, Aruani, & Morales, 2017; Salina, Manrique, & Sosa, 2009; Diaz & Jaramillo, 2018; Hachuel, Boggio, & Harvey, 2010), entre otros, tratan de modelar los datos, y la sobredispersión que estos puedan presentar, valiéndose de las herramientas que les proveen los MLG, sin tener en cuenta posibles efectos aleatorios. Vale destacar que el trabajo de Menni (2011), en el cual considera modelos con ceros modificados, menciona en su discusión que la contemplación de efectos aleatorios podría mejorar el ajuste de sus datos de conteo, pero no los considera por estar fuera del alcance de sus objetivos. A su vez los trabajos publicados por Basanta, Rubio, Giménez, & Tudela (2021) y Zuñiga (2019) contemplan la modelación de las variables de respuesta de conteos sobredispersos mediante un modelo Binomial Negativo en el marco de los MLGM, pero no estudian modelos inflados en cero o modelos Hurdle.

Los resultados de este trabajo de tesis muestran que el uso de los Modelos Lineales Generalizados Mixtos (MLGM) son una alternativa viable más apropiada para ajustar datos de conteo sobredispersos frente a la alternativa de los Modelos Lineales Generalizados (MLG). La incorporación de efectos aleatorios al modelo en el marco de los MLGM permite considerar la correlación de las observaciones, otra fuente de variación que provoca falta de ajuste de los datos.

La implementación del paquete "*glmmTMB*" del software estadístico *R* posibilitó, con este único paquete, estimar rápidamente una variedad de modelos, incluidos MLG, MLGM, MLGM inflados en cero y modelos Hurdle (más modelos de los disponibles en el paquete "*lme4*"). Otro beneficio es que los modelos estimados con un solo paquete se pueden comparar con la seguridad de que los valores de verosimilitud se calcularon de manera consistente incluidos los criterios de información.

6 Conclusiones

Del análisis descriptivo de los datos se puede destacar que: los datos presentan mucha dispersión, poseen elevada frecuencia de ceros, la severidad de la enfermedad es mayor en la edad 2, el cultivar de pera Williams tiene una severidad de la enfermedad mucho menor a la de los demás cultivares estudiados y también, a diferencia del resto, este cultivar mostró una gran proporción de frutos con cero manchas.

Se puede concluir que los cultivares D'Anjou y Abate Fetel se destacan por presentar mayor cantidad de manchas por fruto en todos los estadios de desarrollo evaluados. A pesar del tamaño considerablemente más pequeño que presentaron los frutos en la edad 1, la severidad fue mayor que en la edad 3, para todos los cultivares, lo que indica que los frutos jóvenes son más susceptibles a sufrir lesiones por la enfermedad. El cultivar Packham's, en todos sus estadios, muestra una severidad menos importante que los cultivares D'Anjou y Abate Fetel, a pesar de lo cual los resultados de severidad indican que este cultivar no está exento de sufrir infecciones graves.

En la etapa de modelización, al tener en cuenta la naturaleza discreta de la variable respuesta “conteo de manchas en los frutos” (severidad), se modeló adecuadamente la sobredispersión usando los MLGM. En particular el uso del paquete *glmmTMB* del software estadístico *R* posibilitó, con la ventaja que otorga el poder modelar la dispersión considerando los efectos fijos, ajustar un modelo BN1 considerando la dispersión variando debido a la interacción cultivar-edad (Modelo 18). El modelo Poisson y Binomial Negativo, en el marco de los MLG, fueron descartados por no lograr modelar adecuadamente la sobredispersión (Modelo 1 y Modelo 2). El análisis de los modelos inflados en cero y modelos Hurdle (Modelo 3, Modelo 4 y Modelo 5) permitió descartar que los ceros presentes en los datos analizados fuesen generados por un proceso diferente al Binomial Negativo. Esta alta proporción de cero no resultó ser la causante de la sobredispersión de los datos.

Conocer la susceptibilidad de los frutos de los distintos cultivares a lo largo de su desarrollo resulta de suma importancia a la hora de considerar las medidas necesarias para el control de la enfermedad. Es por ello por lo que los modelos

propuestos en el marco de los MLGM ofrecen una alternativa más conveniente que los que se han estado utilizando hasta ahora y que no permitían captar toda la variabilidad de los datos por lo cual hacían estimaciones insesgadas de los parámetros del modelo propuesto.

7 Perspectivas futuras

Se prevé continuar a futuro con el análisis de datos de conteo con sobredispersión en el contexto de aprendizaje estadístico para grandes bases de datos y, como alternativa conjunta, el análisis bayesiano.

En relación con el problema de la infección de los perales por el hongo *Stemphylium vesicarium*, y para problemáticas similares, se considera analizar la severidad de la infección considerando como variable respuesta el área de cada mancha, es decir considerar una variable de respuesta cuantitativa continua.

8 Bibliografía

Agresti, A. (2015). *Foundations of linear and generalized linear models.* John Wiley & Sons.

Alcaide Delgado, M. (2015). MODELO DE REGRESIÓN BINOMIAL NEGATIVA.

Azpilicueta, C., Aruani, M., & Morales, J. (2017). Efecto del tipo de cobertura vegetal en el espacio entre hileras de perales sobre la abundancia de nematodos fitófagos. *Revista de la Facultad de Agronomía, 116*(2), 249 - 257.

Basanta, M., Rubio, N., Giménez, G., & Tudela, A. (2021). Uso del Modelo Binomial Negativo Mixto para el análisis de datos con sobredispersión en susceptibilidad a la infección por el hongo Stemphylium Vesicarium en frutales del Alto Valle. *XIV Congreso Latinoamericano de Sociedades Estadísticas (CLATSE)*. Montevideo, Uruguay.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823.*

Bolker. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R journal,*(9), 378-400.

Bolker, B. M. (2015). Linear and generalized linear mixed models. En *Ecological Statistics: Contemporary Theory and Application* (págs. 309-333).

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H., & White, J. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution, 24*(3), 127-135.

Boswell, M., & G. P. Patil, G. (1970). *Chance mechanisms generating the negative binomial distribution, in Random Counts in Models and Structures.* Pennsylvania State University Press.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association, 88*(421), 9-25.

Brooks, M., Kristensen, K., van Benthem, K., Arni Magnusson, A., Berg, C., Nielsen, A., . . . Bolker, B. (2017). glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal, 9*(2), 378-400.

Bürkner, P. (2017). An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software, 80*(1), 1 - 28.

Bürkner, P. (2017). brms: An R package for Bayesian Multilevel Models Using Stan. *Journal of statistical software, 80*(1), 1-28.

Cameron, C., & Trivedi, P. (2013). *Regression analysis of count data* (segunda ed., Vol. 53). Cambridge university press.

Cayuela, L. (2009). Modelos lineales generalizados (GLM). *Materiales de un curso del R del IREC*.

Correa Morales, J. C., & Salazar Uribe, J. C. (2016). *Introducción a Los Modelos Mixtos.* Universidad Nacional de Colombia - Sede Medellín.

Di Massi, S., Lago, J., & Tudela, A. (2022). Presencia de la enfermedad "Mancha marrón del peral". *Boletín Sanitario N°7- Temporada 2021-2022- INTA Alto Valle*.

Diaz, E., G, & Jaramillo, A. (2018). Técnicas de regresión para datos de recuento. Colombia: IV Simposio de Estadística.

Dobra, A. C., & Garcia, L. (2015). Presencia de mancha negra del peral, Stemphylium vesicarium, en el Valle Medio del Río Negro, Patagonia Argentina. *XXXVIII Congreso Argentino de Horticultura.* Bahía Blanca, Buenos Aires.

Dobson, A., & Barnett, A. (2018). *An introduction to generalized linear models.* Chapman and Hall/CRC.

Faraway, J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models.* Chapman and Hall/CRC.

Gimenez, G. (2020). Análisis y aprovechamientode bases de datos agronómicas recurriendo al proceso "Knowledge Discovery in Data Bases" (KKD) y algoritmos de "Data Mining" (DM).

Godoy, C., Monterubbianesi, G., & Tognetti, J. (2008). Analysis of highbush blueberry (Vaccinium corymbosum L.) fruit growth with exponential mixed models. *Scientia Horticulturae, 115*(4), 368 - 376.

Hachuel, L. S., Boggio, G. S., & Harvey, G. B. (2010). *Modelos alternativos para el análisis de datos de conteo con exceso de ceros.*

Hadfield, J. (2010). MCMC Methods for Multi-Response Generalized Linear Mixed Models: The {MCMCglmm} {R} Package. *Journal of Statistical Software, 33*(2), 1-22.

Hardin, J., & Hilbe, J. (2007). *Generalized Linear Models and Extensions* (Segunda ed.). Stata Press.

Harrison, X. A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ, 2*, e616.

Heilbron, D. (1989). Generalized linear models for altered zero probabilities and overdispersion in count data. Unpublished Technical report, University of California, San Francisco, Department of Epidemiology and Biostatistics.

Hilbe, J. (2011). *Negative Binomial Regression.* Cambridge University Press.

Hilbe, J. (2017). El análisis estadístico de los datos de recuento. *Cultura y Educación, 29*(3), 434-460.

Huang, A. (2017). Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts. *Statistical Modelling, 17*(6), 359-380.

Kleiber, C., & Zeileis, A. (2016). Visualizing Count Data Regressions Using Rootograms. *The American Statistician, 70*(3), 296-303.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., & Bell, B. (2015). TMB: automatic differentiation and Laplace approximation. *arXiv preprint arXiv:1509.00660.* Obtenido de arXiv preprint arXiv:1509.00660.

Kuhn, T. S. (2019). *La estructura de las revoluciones científicas.* Fondo de cultura económica.

Lagani, V., Athineou, G., Farcomeni, A., Tsagris, M., & Tsamardinos, I. (2017). Feature Selection with the R Package MXM: Discovering Statistically Equivalent Feature Subsets. *Journal of Statistical Software, 80*(7), 1-25.

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

Lindsey, J. K., Jones, B., & Jarvis, P. (2001). Some statistical issues in modelling pharmacokinetic data. *Statistics in medicine, 20*(17 - 18), 2775 - 2783.

Littell, R., Milliken, G., Stroup, W., Wolfinger, R., & Schabenberger, O. (2006). *SAS for Mixed Models* (Segunda ed.). SAS Institute Inc.

Llorens A, N. (2005). Evaluación en el modelado de las respuestas de recuento. *Tesis doctoral*. Universitat de les Illes Balears.

Llorens, A., Palmer, P., & Losilla, V. (2004). Ajuste y estimación de los errores estándar de los parámetros del modelo de regresión de Poisson en presencia de sobredispersión. *Metodología de las ciencias del comportamiento,, 5*(1), 333 - 339.

Llorente, I., Moragrega, C., Ruz, L., & Montesinos, E. (2012). An update on control of brown spot of pear. *Trees, 26*, 239-245.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables.* Thousand Oaks, CA: Stage.

Lunneborg, C. (1994). Modeling experimental and observational data. *Duxbury Resource Center*.

Martinez Calcaterra, E., & Vaucher, P. (2017). Una revisión de los modelos de conteo con excesos de ceros. Montevideo, Uruguay: Udelar. FCEA.

McCullagh, P., & Nelder, J. (2019). *Generalized linear models.* Routledge.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models.* Springer US.

McCulloch, C., & Searle, S. (2001). *Generalized, Linear and Mixed Models.* John Wiley & Sons, Inc.

Melkersson, M., & Rooth, D. (2000). Modeling female fertility using inflated count data models. *Journal of Population Economics, 13*, 189 - 203.

Menni, M. F. (2011). *Estrategias de modelación para datos de conteo con exceso de ceros. Una ilustración asociada a ecología de poblaciones.* Tesis de maestría, Universidad Nacional del Comahue, Facultad de Economía.

Montesinos, E., Moragrega, C., Llorente, I., & Vilardell, P. (1995). Susceptibility of selected European pear cultivars to infection by Stemphylium vesicarium and influence of leaf and fruit age. *Plant disease (USA), 79*, 471 - 473.

Mullahy, J. (1997). Heterogeneity, excess zeros, and the structure of count data models. *Journal of applied econometrics, 12*(3), 337-350.

Nelder, J. A., & Wedderburn, R. W. (1972). Modelos lineales generalizados. *Revista de la Real Sociedad Estadística, 135*(3), 370-384.

Pol, A., & Del Río, M. (2005). Modelado del número de días de consumo de cannabis. *Psicothema*, 569 - 574.

RCoreTeam. (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Obtenido de URL https://www.R-project.org/

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics, 54*(3), 507-554.

Rigby, R., & Stasinopoulos, D. (2005). Generalized additive models for location, scale and shape,(with discussion). *Appl. Statist., 54*(3), 507 - 554.

Rossi, V., Pattori, E., & Giosuè, S. (2006). Temperature and humidity requirements for germination and infection by ascospores of Pleospora allii the teleomorph of Stemphylium vesicarium. *IOBC WPRS BULLETIN, 29*(1), 223.

Rubio, N. (2016). Modelación de respuestas ordinales longitudinales mediante modelos lineales generalizados mixtos.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology), 71*(2), 319-392.

Salina, R., Manrique, E., & Sosa, R. (2009). Análisis estadístico para datos de conteo: aplicaciones para el uso de los servicios de salud. *salud pública de méxico, 51*, 397- 406.

SENASA. (2021). *Anuario Estadístico.* Centro Regional Patagonia Norte: Servicio Nacional de Sanidad y Calidad Agroalimentaria.

Skaug, H., Fournier, D., Nielsen, A., Magnusson, A., & Bolker, B. (2010). glmmADMB: generalized linear mixed models using AD MOdel Builder. *R package versión 0.6, 5*, r143.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., & De Bastiani, F. (2017). *Flexible regression and smoothing: using GAMLSS in R.* CRC Press.

Tassile, V. (2020). Aplicación de modelos mixtos no lineales en el crecimiento de frutos de peral (pyrus communis) cv. William's.

Venables, W., & Ripley, B. (2002). *Modern Applied Statistics with S* (Cuarta ed.). Springer.

Vives Brosa, J. (2002). El diagnóstico de la sobredispersión en modelos de análisis de datos de recuento. *tesis doctoral*. Universitat Autònoma de Barcelona.

Vives, J., Losilla, J. M., & Rodrigo, M. F. (2006). Count data in psychological applied research. *Psychological Reports, 98*(3), 821 - 835.

Wang, W., & Famoye, F. (1997). Modeling household fertility decisions with generalized Poisson regression. *Journal of Population Economics, 10*, 273 - 283.

West, B., Welch, K., & Galecki, A. (2015). *Linear mixed models: a practical guide using statistical software.* A Chapman & Hall Book.

Winkelmann, R. (2008). *Econometric analysis of count data* (quinta ed.). Springer Science & Business Media.

Wood, S. (2017). *Generalized Additive Models: An Introduction with R* (Segunda ed.). Chapman and Hall/CRC.

Wood, S. (2017). *Generalized Aditive Models: An Introduction with R* (segunda ed.). Chapman and Hall/CRC.

Wood, S. N., Pya, N., & Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association, 111*(516), 1548-1563.

Yee, T. W. (2020). The VGAM package for negative binomial regression. *Australian & New Zealand Journal of Statistics, 61*(1), 116-131.

Yen, S. (1999). Gaussian versus count-data hurdle models: cigarette consumption by women in the US. *Applied Economics Letters, 6*(2), 73-76.

Zeileis, A., & Kleiber, C. (2022). _countreg: Count Data Regression_. R package version 0.2-1. Obtenido de https://R-Forge.R-project.org/projects/countreg/

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regresion Models for Count Data in R. *Journal of Statistical Software, 27*(8), 1-25.

Zuñiga, D. (2019). Regeneración de la comunidad vegetal en ambientes desertificados en aguada guzmán (departamento el cuy, río negro): diferencias según condiciones de pastoreo y precipitación . Tesis de Maestría en Intervención Ambiental - UNLP.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R* (Vol. 574). New York: Springer.

ANEXO

Códigos de funciones y salidas en R para ajuste de modelos.

Modelo 1

Modelo de regresión Poisson (MLG)

función *glm* (librería *stats*) – nombre del objeto *hsv_glm_pois*

```
glm(formula = Severidad ~ Cultivar + Age + Cultivar:Age, family =
poisson,
    data = hsv)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-10.223   -3.011   -1.039    2.015   18.174

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            4.77210    0.01680 284.130  < 2e-16
CultivarDAnjou         0.12475    0.02305   5.413 6.19e-08
CultivarPackhams      -0.58446    0.02807 -20.819  < 2e-16
CultivarWilliams      -3.26064    0.08738 -37.316  < 2e-16
Age2                   1.02224    0.01959  52.194  < 2e-16
Age3                  -0.91629    0.03142 -29.161  < 2e-16
CultivarDAnjou:Age2   -0.62210    0.02827 -22.002  < 2e-16
CultivarPackhams:Age2 -0.79304    0.03595 -22.062  < 2e-16
CultivarWilliams:Age2 -2.22132    0.17924 -12.393  < 2e-16
CultivarDAnjou:Age3   -0.06520    0.04360  -1.496    0.135
CultivarPackhams:Age3  0.21399    0.05015   4.267 1.98e-05
CultivarWilliams:Age3 -0.04512    0.16604  -0.272    0.786

(Intercept)           ***
CultivarDAnjou        ***
CultivarPackhams      ***
CultivarWilliams      ***
Age2                  ***
Age3                  ***
CultivarDAnjou:Age2   ***
CultivarPackhams:Age2 ***
CultivarWilliams:Age2 ***
CultivarDAnjou:Age3
CultivarPackhams:Age3 ***
CultivarWilliams:Age3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 38132.7  on 359  degrees of freedom
Residual deviance:  5936.1  on 348  degrees of freedom
AIC: 7751.8

Number of Fisher Scoring iterations: 6


Analysis of Deviance Table (Type II tests)

Response: Severidad
             LR Chisq Df Pr(>Chisq)
Cultivar      20824.0  3  < 2.2e-16 ***
Age           10239.2  2  < 2.2e-16 ***
Cultivar:Age   1133.5  6  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Modelo 2

Modelo de regresión Binomial Negativo (MLG)

función *glm.nb* (librería *MASS*) – nombre del objeto: *hsv_glm.bn*

```
glm.nb(formula = Severidad ~ Cultivar + Age + Cultivar:Age, data = hsv,
    init.theta = 2.503388879, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7083  -1.0914  -0.1934   0.3697   3.8832

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          4.77210    0.11661  40.924  < 2e-16 ***
CultivarDAnjou       0.12475    0.16481   0.757 0.449091
CultivarPackhams    -0.58446    0.16559  -3.530 0.000416 ***
CultivarWilliams    -3.26064    0.18511 -17.615  < 2e-16 ***
Age2                 1.02224    0.16436   6.220 4.99e-10 ***
Age3                -0.91629    0.16619  -5.514 3.51e-08 ***
CultivarDAnjou:Age2  -0.62210    0.23251  -2.676 0.007460 **
CultivarPackhams:Age2 -0.79304    0.23357  -3.395 0.000685 ***
CultivarWilliams:Age2 -2.22132    0.29221  -7.602 2.92e-14 ***
CultivarDAnjou:Age3  -0.06520    0.23487  -0.278 0.781306
CultivarPackhams:Age3  0.21399    0.23617   0.906 0.364884
CultivarWilliams:Age3 -0.04512    0.28431  -0.159 0.873904
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(2.5034) family taken to be 1)

    Null deviance: 1901.31  on 359  degrees of freedom
Residual deviance:  472.72  on 348  degrees of freedom
AIC: 3289.2

Number of Fisher Scoring iterations: 1


              Theta:  2.503
          Std. Err.:  0.244

 2 x log-likelihood:  -3263.154



Analysis of Deviance Table (Type II tests)

Response: Severidad
             LR Chisq Df Pr(>Chisq)
Cultivar      1034.57  3  < 2.2e-16 ***
Age            182.10  2  < 2.2e-16 ***
Cultivar:Age    77.01  6  1.483e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Modelo 3

Modelo de regresión Binomial Negativo inflado en cero (MLG).

Sin regresores para el componente cero.

función *zeroinfl* (librería *pscl*) - nombre del objeto: *hsv_zinb20*

```
zeroinfl(formula = Severidad ~ Cultivar + Age + Cultivar:Age | 1, data =
hsv, dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.4313 -0.8562 -0.1856  0.3992  7.7458

Count model coefficients (negbin with log link):
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          4.77211    0.11661  40.925  < 2e-16 ***
CultivarDAnjou       0.12477    0.16481   0.757 0.448995
CultivarPackhams     -0.58451    0.16558  -3.530 0.000416 ***
CultivarWilliams     -3.26052    0.18527 -17.599  < 2e-16 ***
Age2                 1.02222    0.16436   6.220 4.99e-10 ***
Age3                 -0.91629    0.16618  -5.514 3.51e-08 ***
CultivarDAnjou:Age2  -0.62209    0.23251  -2.676 0.007460 **
CultivarPackhams:Age2 -0.79295    0.23356  -3.395 0.000686 ***
CultivarWilliams:Age2 -2.22139    0.29227  -7.600 2.95e-14 ***
CultivarDAnjou:Age3  -0.06524    0.23486  -0.278 0.781192
CultivarPackhams:Age3  0.21404    0.23616   0.906 0.364763
CultivarWilliams:Age3 -0.04523    0.28436  -0.159 0.873627
Log(theta)           0.91769    0.09758   9.405  < 2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -10.79      49.44  -0.218    0.827
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 2.5035
Number of iterations in BFGS optimization: 34
Log-likelihood: -1632 on 14 Df
[1] "AIC= 3291.16"



Analysis of Deviance Table (Type II tests)

Response: Severidad
             Df    Chisq Pr(>Chisq)
Cultivar      3 1244.028  < 2.2e-16 ***
Age           2  201.507  < 2.2e-16 ***
Cultivar:Age  6   73.309  8.552e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Modelo 4

Modelo de regresión Binomial Negativo inflado en cero (MLG).

Con todos los regresores para el componente cero.

función *zeroinfl* (librería *pscl*) – nombre del objeto: *hsv_zinb21*

```
zeroinfl(formula = Severidad ~ Cultivar + Age + Cultivar:Age, data = hsv,
dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.6603 -0.6243 -0.1951  0.4278  8.1496

Count model coefficients (negbin with log link):
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         4.77210    0.09958  47.922  < 2e-16 ***
CultivarDAnjou      0.12475    0.14071   0.887  0.37532
CultivarPackhams   -0.58446    0.14162  -4.127 3.68e-05 ***
CultivarWilliams   -2.50860    0.19608 -12.794  < 2e-16 ***
Age2                1.02224    0.14019   7.292 3.05e-13 ***
Age3               -0.91629    0.14232  -6.438 1.21e-10 ***
CultivarDAnjou:Age2   -0.62210    0.19834  -3.137  0.00171 **
CultivarPackhams:Age2 -0.79304    0.19957  -3.974 7.08e-05 ***
CultivarWilliams:Age2 -2.06979    0.32695  -6.331 2.44e-10 ***
CultivarDAnjou:Age3   -0.06520    0.20109  -0.324  0.74575
CultivarPackhams:Age3  0.21399    0.20261   1.056  0.29090
CultivarWilliams:Age3  0.14743    0.31414   0.469  0.63885
Log(theta)             1.24122    0.08899  13.948  < 2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.058e+01  5.363e+03  -0.004    0.997
CultivarDAnjou      -1.268e-06  7.584e+03   0.000    1.000
CultivarPackhams    -1.268e-06  7.584e+03   0.000    1.000
CultivarWilliams     2.069e+01  5.363e+03   0.004    0.997
Age2                -7.167e-02  7.723e+03   0.000    1.000
Age3                -3.767e-02  7.656e+03   0.000    1.000
CultivarDAnjou:Age2  6.014e-07  1.092e+04   0.000    1.000
CultivarPackhams:Age2  6.014e-07  1.092e+04   0.000    1.000
CultivarWilliams:Age2  3.413e-01  7.723e+03   0.000    1.000
CultivarDAnjou:Age3    3.328e-07  1.083e+04   0.000    1.000
CultivarPackhams:Age3  3.328e-07  1.083e+04   0.000    1.000
CultivarWilliams:Age3  3.753e-01  7.656e+03   0.000    1.000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 3.4598
Number of iterations in BFGS optimization: 21
Log-likelihood: -1568 on 25 Df
[1] "AIC= 3186.99"



Analysis of Deviance Table (Type II tests)

Response: Severidad
            Df   Chisq Pr(>Chisq)
Cultivar     3 597.122  < 2.2e-16 ***
Age          2 285.248  < 2.2e-16 ***
Cultivar:Age  6  62.759  1.236e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Modelo 5

Modelo de regresión Binomial Negativo Hurdle (MLG)

función *hurdle* (librería *pscl*) – nombre del objeto: *hsv_hnb2*

```
hurdle(formula = Severidad ~ Cultivar + Age + Cultivar:Age, data = hsv,
dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.6608 -0.6242 -0.1951  0.4276  8.1520

Count model coefficients (truncated negbin with log link):
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           4.77209    0.09962  47.905  < 2e-16 ***
CultivarDAnjou        0.12475    0.14076   0.886  0.37548
CultivarPackhams      -0.58449    0.14167  -4.126 3.70e-05 ***
CultivarWilliams      -2.50861    0.19614 -12.790  < 2e-16 ***
Age2                  1.02224    0.14024   7.289 3.11e-13 ***
Age3                  -0.91638    0.14238  -6.436 1.23e-10 ***
CultivarDAnjou:Age2   -0.62210    0.19840  -3.136  0.00172 **
CultivarPackhams:Age2 -0.79303    0.19964  -3.972 7.12e-05 ***
CultivarWilliams:Age2 -2.06984    0.32702  -6.329 2.46e-10 ***
CultivarDAnjou:Age3   -0.06519    0.20117  -0.324  0.74591
CultivarPackhams:Age3  0.21381    0.20271   1.055  0.29153
CultivarWilliams:Age3  0.14749    0.31422   0.469  0.63880
Log(theta)            1.24052    0.08917  13.912  < 2e-16 ***

Zero hurdle model coefficients (binomial with logit link):
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           2.057e+01  5.337e+03   0.004    0.997
CultivarDAnjou        1.093e-06  7.548e+03   0.000    1.000
CultivarPackhams      1.090e-06  7.548e+03   0.000    1.000
CultivarWilliams      -2.070e+01  5.337e+03  -0.004    0.997
Age2                  1.059e-06  7.548e+03   0.000    1.000
Age3                  1.186e-06  7.548e+03   0.000    1.000
CultivarDAnjou:Age2   -1.072e-06  1.067e+04   0.000    1.000
CultivarPackhams:Age2 -1.054e-06  1.067e+04   0.000    1.000
CultivarWilliams:Age2 -4.130e-01  7.548e+03   0.000    1.000
CultivarDAnjou:Age3   -1.183e-06  1.067e+04   0.000    1.000
CultivarPackhams:Age3 -1.177e-06  1.067e+04   0.000    1.000
CultivarWilliams:Age3 -4.130e-01  7.548e+03   0.000    1.000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta: count = 3.4574
Number of iterations in BFGS optimization: 21
Log-likelihood: -1568 on 25 Df
[1] "AIC= 3186.96"



Analysis of Deviance Table (Type II tests)

Response: Severidad
             Df   Chisq Pr(>Chisq)
Cultivar      3 596.839  < 2.2e-16 ***
Age           2 285.067  < 2.2e-16 ***
Cultivar:Age  6  62.717   1.26e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Modelo 6

Modelo de regresión Binomial Negativo con efectos aleatorios (MLGM)

Efecto aleatorio *planta*

función *glmer.nb* (librería *lme4*) – nombre del objeto: *hsv_glmer.bn_p*

```
Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) ['glmerMod']
 Family: Negative Binomial(3.1515)  ( log )
Formula: Severidad ~ Cultivar + Age + Cultivar:Age + (1 | Planta)
   Data: hsv

     AIC      BIC   logLik deviance df.resid
  3257.2   3311.6  -1614.6   3229.2      346

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.5236 -0.8121 -0.1795  0.4128  5.2200

Random effects:
 Groups Name        Variance Std.Dev.
 Planta (Intercept) 0.1184   0.3441
Number of obs: 360, groups:  Planta, 24

Fixed effects:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           4.77885    0.17535  27.254  < 2e-16 ***
CultivarDAnjou        0.11292    0.24770   0.456   0.6485
CultivarPackhams     -0.55830    0.24890  -2.243   0.0249 *
CultivarWilliams     -3.48911    0.26817 -13.011  < 2e-16 ***
Age2                  1.01653    0.14788   6.874 6.24e-12 ***
Age3                 -0.95740    0.15156  -6.317 2.67e-10 ***
CultivarDAnjou:Age2  -0.63888    0.20867  -3.062   0.0022 **
CultivarPackhams:Age2 -0.84157    0.21106  -3.987 6.68e-05 ***
CultivarWilliams:Age2 -2.10708    0.28052  -7.511 5.86e-14 ***
CultivarDAnjou:Age3  -0.04790    0.21304  -0.225   0.8221
CultivarPackhams:Age3  0.03864    0.21952   0.176   0.8603
CultivarWilliams:Age3  0.02152    0.27322   0.079   0.9372
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) CltvDA CltvrP CltvrW Age2   Age3   CDA:A2 ClP:A2
ClW:A2 CDA:A3 ClP:A3
CultivrDAnj -0.708
CltvrPckhms -0.704  0.499
CultvrWllms -0.654  0.463  0.459
Age2        -0.425  0.301  0.300  0.278
Age3        -0.420  0.297  0.296  0.277  0.493
CltvrDAn:A2  0.301 -0.423 -0.212 -0.196 -0.709 -0.349
CltvrPck:A2  0.298 -0.211 -0.429 -0.193 -0.701 -0.345  0.497
CltvrWll:A2  0.224 -0.159 -0.158 -0.411 -0.527 -0.260  0.373  0.369
CltvrDAn:A3  0.299 -0.419 -0.210 -0.196 -0.350 -0.711  0.492  0.245
0.185
CltvrPck:A3  0.290 -0.205 -0.423 -0.183 -0.340 -0.689  0.241  0.493
0.178  0.491
CltvrWll:A3  0.233 -0.165 -0.164 -0.417 -0.273 -0.555  0.194  0.191
0.394  0.395  0.382

[1] "deviance condicional= 447.23"

Analysis of Deviance Table (Type II Wald chisquare tests)

Response: Severidad
              Chisq Df Pr(>Chisq)
Cultivar     385.273  3  < 2.2e-16 ***
Age          273.356  2  < 2.2e-16 ***
Cultivar:Age  71.382  6  2.128e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Modelo 7

Modelo de regresión Binomial Negativo con efectos aleatorios (MLGM)

Efecto aleatorio *planta* más interacción *planta:fruto*

función *glmer.nb* (librería *lme4*) – nombre del objeto: *hsv_glmer.bn_pf*

```
boundary (singular) fit: see help('isSingular')

Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) ['glmerMod']
 Family: Negative Binomial(3.1515)  ( log )
Formula: Severidad ~ Cultivar + Age + Cultivar:Age + (1 | Planta) + (1 |
Planta:Fruto)
   Data: hsv

     AIC      BIC   logLik deviance df.resid
  3259.2   3317.5  -1614.6   3229.2      345

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.5236 -0.8121 -0.1795  0.4128  5.2200

Random effects:
 Groups       Name        Variance  Std.Dev.
 Planta:Fruto (Intercept) 1.196e-10 1.094e-05
 Planta       (Intercept) 1.184e-01 3.441e-01
Number of obs: 360, groups:  Planta:Fruto, 120; Planta, 24

Fixed effects:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            4.77885    0.17534  27.254  < 2e-16 ***
CultivarDAnjou         0.11292    0.24769   0.456   0.6485
CultivarPackhams      -0.55831    0.24890  -2.243   0.0249 *
CultivarWilliams      -3.48912    0.26817 -13.011  < 2e-16 ***
Age2                   1.01653    0.14788   6.874 6.24e-12 ***
Age3                  -0.95739    0.15156  -6.317 2.67e-10 ***
CultivarDAnjou:Age2   -0.63888    0.20867  -3.062   0.0022 **
CultivarPackhams:Age2 -0.84157    0.21106  -3.987 6.68e-05 ***
CultivarWilliams:Age2 -2.10708    0.28052  -7.511 5.86e-14 ***
CultivarDAnjou:Age3   -0.04790    0.21304  -0.225   0.8221
CultivarPackhams:Age3  0.03864    0.21952   0.176   0.8603
CultivarWilliams:Age3  0.02152    0.27322   0.079   0.9372
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) CltvDA CltvrP CltvrW Age2   Age3   CDA:A2 ClP:A2
ClW:A2 CDA:A3 ClP:A3
CultivrDAnj -0.708
CltvrPckhms -0.704  0.499
CultvrWllms -0.654  0.463  0.459
Age2        -0.425  0.301  0.300  0.278
Age3        -0.420  0.297  0.296  0.277  0.493
CltvrDAn:A2  0.301 -0.423 -0.212 -0.196 -0.709 -0.349
CltvrPck:A2  0.298 -0.211 -0.429 -0.193 -0.701 -0.345  0.497
CltvrWll:A2  0.224 -0.159 -0.158 -0.411 -0.527 -0.260  0.373  0.369
CltvrDAn:A3  0.299 -0.419 -0.210 -0.196 -0.350 -0.711  0.492  0.245
0.185
CltvrPck:A3  0.290 -0.205 -0.423 -0.183 -0.340 -0.689  0.241  0.493
0.178  0.491
CltvrWll:A3  0.233 -0.165 -0.164 -0.417 -0.273 -0.555  0.194  0.191
0.394  0.395  0.382
optimizer (Nelder_Mead) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')

[1] "deviance condicional= 447.23"
```

Modelo 8

Modelo de regresión Binomial Negativo con efectos aleatorios (MLGM)

Efecto aleatorio *cultivar* | *planta*

función *glmer.nb* (librería *lme4*) – nombre del objeto: *hsv_glmer.bn_cp*

```
Warning: unable to evaluate scaled gradientWarning: Model failed to converge:
degenerate  Hessian with 2 negative eigenvaluesWarning: unable to evaluate scaled
gradientWarning: Model failed to converge: degenerate  Hessian with 3 negative
eigenvaluesboundary (singular) fit: see help('isSingular')

Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) [
glmerMod]
 Family: Negative Binomial(3.1923)  ( log )
Formula: Severidad ~ Cultivar + Age + Cultivar:Age + (Cultivar | Planta)
   Data: hsv

     AIC      BIC   logLik deviance df.resid
  3258.0   3347.4  -1606.0   3212.0      337

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.5341 -0.7907 -0.1915  0.4321  7.3752

Random effects:
 Groups Name             Variance  Std.Dev.  Corr
 Planta (Intercept)      4.741e-14 2.177e-07
        CultivarDAnjou   1.992e-02 1.411e-01  0.88
        CultivarPackhams 1.010e-01 3.178e-01  0.13 -0.07
        CultivarWilliams 6.244e-01 7.902e-01 -0.74 -0.63  0.10
Number of obs: 360, groups:  Planta, 24

Fixed effects:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           4.772096   0.103554  46.083  < 2e-16 ***
CultivarDAnjou        0.118668   0.157337   0.754  0.45071
CultivarPackhams     -0.553247   0.197289  -2.804  0.00504 **
CultivarWilliams     -3.587630   0.372125  -9.641  < 2e-16 ***
Age2                  1.022238   0.145829   7.010 2.39e-12 ***
Age3                 -0.916291   0.147886  -6.196 5.79e-10 ***
CultivarDAnjou:Age2  -0.635217   0.206593  -3.075  0.00211 **
CultivarPackhams:Age2 -0.845961   0.209001  -4.048 5.17e-05 ***
CultivarWilliams:Age2 -2.101785   0.283497  -7.414 1.23e-13 ***
CultivarDAnjou:Age3  -0.079432   0.209562  -0.379  0.70466
CultivarPackhams:Age3  0.003215   0.217608   0.015  0.98821
CultivarWilliams:Age3 -0.019178   0.275942  -0.070  0.94459
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) CltvDA CltvrP CltvrW Age2   Age3   CDA:A2 ClP:A2
ClW:A2 CDA:A3 ClP:A3
CultivrDAnj -0.658
CltvrPckhms -0.525  0.345
CultvrWllms -0.278  0.183  0.146
Age2        -0.710  0.467  0.373  0.198
Age3        -0.700  0.461  0.368  0.195  0.497
CltvrDAn:A2  0.501 -0.659 -0.263 -0.139 -0.706 -0.351
CltvrPck:A2  0.495 -0.326 -0.536 -0.138 -0.698 -0.347  0.493
CltvrWll:A2  0.365 -0.240 -0.192 -0.296 -0.514 -0.256  0.363  0.359
CltvrDAn:A3  0.494 -0.651 -0.259 -0.138 -0.351 -0.706  0.497  0.245
0.180
CltvrPck:A3  0.476 -0.313 -0.529 -0.132 -0.338 -0.680  0.239  0.495
0.174  0.480
CltvrWll:A3  0.375 -0.247 -0.197 -0.300 -0.266 -0.536  0.188  0.186
0.392  0.378  0.364
optimizer (Nelder_Mead) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')

[1] "deviance condicional= 451.11"
```

Modelos ajustados con el paquete *glmmTMB*

Modelo 9

Modelo de regresión Binomial Negativo BN1

ziformula=~0; dispformula=~1

función *glmmTMB* (librería *glmmTMB)* – nombre del objeto: *hsv_nbinom1_0_1*

```
Family: nbinom1  ( log )
Formula:          Severidad ~ Cultivar + Age + Cultivar:Age + (1 |
Planta)
Data: hsv

     AIC      BIC   logLik deviance df.resid
  3076.9   3131.3  -1524.4   3048.9      346

Random effects:

Conditional model:
 Groups Name        Variance Std.Dev.
 Planta (Intercept) 0.02796  0.1672
Number of obs: 360, groups:  Planta, 24

Dispersion parameter for nbinom1 family (): 16.2

Conditional model:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           4.76511    0.09650   49.38  < 2e-16 ***
CultivarDAnjou        0.15255    0.13374    1.14 0.254004
CultivarPackhams     -0.55534    0.14720   -3.77 0.000161 ***
CultivarWilliams     -3.34434    0.28753  -11.63  < 2e-16 ***
Age2                  1.02665    0.07956   12.90  < 2e-16 ***
Age3                 -0.93558    0.12367   -7.56 3.88e-14 ***
CultivarDAnjou:Age2  -0.66813    0.11409   -5.86 4.74e-09 ***
CultivarPackhams:Age2 -0.89285    0.14430   -6.19 6.12e-10 ***
CultivarWilliams:Age2 -1.53726    0.40388   -3.81 0.000141 ***
CultivarDAnjou:Age3  -0.04114    0.16932   -0.24 0.808014
CultivarPackhams:Age3 -0.02377    0.20037   -0.12 0.905585
CultivarWilliams:Age3  0.45669    0.41355    1.10 0.269458
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



Analysis of Deviance Table (Type II Wald chisquare tests)

Response: Severidad
              Chisq Df Pr(>Chisq)
Cultivar     409.184  3  < 2.2e-16 ***
Age          509.976  2  < 2.2e-16 ***
Cultivar:Age  78.704  6  6.618e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Modelo 10

Modelo de regresión Binomial Negativo BN1

ziformula=~0; dispformula=~Cultivar

función *glmmTMB* (librería *glmmTMB)* – nombre del objeto: *hsv_nbinom1_0_C*

```
Family: nbinom1  ( log )
Formula:          Severidad ~ Cultivar + Age + Cultivar:Age + (1 |
Planta)
Dispersion:                 ~Cultivar
Data: hsv

     AIC      BIC   logLik deviance df.resid
  3072.3   3138.3  -1519.1   3038.3      343

Random effects:

Conditional model:
 Groups Name        Variance Std.Dev.
 Planta (Intercept) 0.02911  0.1706
Number of obs: 360, groups:  Planta, 24

Conditional model:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           4.76246    0.09710   49.05  < 2e-16 ***
CultivarDAnjou        0.13658    0.13142    1.04  0.29866
CultivarPackhams     -0.50448    0.15758   -3.20  0.00137 **
CultivarWilliams     -3.52274    0.31418  -11.21  < 2e-16 ***
Age2                  1.02850    0.07848   13.10  < 2e-16 ***
Age3                 -0.93895    0.12258   -7.66 1.86e-14 ***
CultivarDAnjou:Age2  -0.66195    0.10538   -6.28 3.36e-10 ***
CultivarPackhams:Age2 -0.90743    0.15798   -5.74 9.24e-09 ***
CultivarWilliams:Age2 -1.58601    0.40330   -3.93 8.40e-05 ***
CultivarDAnjou:Age3  -0.06222    0.15954   -0.39  0.69654
CultivarPackhams:Age3  0.02949    0.21295    0.14  0.88986
CultivarWilliams:Age3  0.41982    0.41211    1.02  0.30834
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Dispersion model:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)       2.7453     0.1663  16.506   <2e-16 ***
CultivarDAnjou   -0.3211     0.2392  -1.342   0.1795
CultivarPackhams  0.4038     0.2366   1.707   0.0878 .
CultivarWilliams -0.3507     0.3683  -0.952   0.3410
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



Analysis of Deviance Table (Type II Wald chisquare tests)

Response: Severidad
              Chisq Df Pr(>Chisq)
Cultivar    289.180  3  < 2.2e-16 ***
Age         559.229  2  < 2.2e-16 ***
Cultivar:Age  79.581  6  4.362e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Modelo 11

Modelo de regresión Binomial Negativo BN1

ziformula=~0; dispformula=~Cultivar*Age

función *glmmTMB* (librería *glmmTMB)* – nombre del objeto: *hsv_nbinom1_0_i*

```
Family: nbinom1  ( log )
Formula:          Severidad ~ Cultivar + Age + Cultivar:Age + (1 |
Planta)
Dispersion:                 ~Cultivar * Age
Data: hsv

     AIC      BIC   logLik deviance df.resid
  3063.6   3160.7  -1506.8   3013.6      335

Random effects:

Conditional model:
 Groups Name        Variance Std.Dev.
 Planta (Intercept) 0.02014  0.1419
Number of obs: 360, groups:  Planta, 24

Conditional model:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           4.76375    0.09122   52.22  < 2e-16 ***
CultivarDAnjou        0.12171    0.11416    1.07    0.286
CultivarPackhams     -0.58418    0.12898   -4.53 5.92e-06 ***
CultivarWilliams     -3.26463    0.39762   -8.21  < 2e-16 ***
Age2                  1.02223    0.07851   13.02  < 2e-16 ***
Age3                 -0.91630    0.13726   -6.68 2.46e-11 ***
CultivarDAnjou:Age2  -0.62210    0.10455   -5.95 2.68e-09 ***
CultivarPackhams:Age2 -0.79304    0.15403   -5.15 2.62e-07 ***
CultivarWilliams:Age2 -2.22130    0.54467   -4.08 4.54e-05 ***
CultivarDAnjou:Age3  -0.06520    0.16945   -0.38    0.700
CultivarPackhams:Age3  0.21400    0.24610    0.87    0.385
CultivarWilliams:Age3 -0.04511    0.57569   -0.08    0.938
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Dispersion model:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)            2.8056     0.2957   9.487  < 2e-16 ***
CultivarDAnjou        -1.3240     0.4529  -2.923  0.00346 **
CultivarPackhams      -0.6343     0.4354  -1.457  0.14516
CultivarWilliams       0.1337     0.6018   0.222  0.82414
Age2                  -0.4095     0.5253  -0.779  0.43572
Age3                   0.1229     0.4132   0.297  0.76613
CultivarDAnjou:Age2    1.8989     0.6849   2.773  0.00556 **
CultivarPackhams:Age2  1.6511     0.6849   2.411  0.01593 *
CultivarWilliams:Age2 -0.9386     0.9548  -0.983  0.32557
CultivarDAnjou:Age3    0.8667     0.6228   1.392  0.16405
CultivarPackhams:Age3  1.2613     0.6116   2.062  0.03917 *
CultivarWilliams:Age3 -1.0284     0.8903  -1.155  0.24801
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



Analysis of Deviance Table (Type II Wald chisquare tests)

Response: Severidad
             Chisq Df Pr(>Chisq)
Cultivar   355.172  3  < 2.2e-16 ***
Age        486.457  2  < 2.2e-16 ***
Cultivar:Age  68.736  6  7.426e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Modelo 12

Modelo de regresión Binomial Negativo BN2

ziformula=~0; dispformula=~1

función *glmmTMB* (librería *glmmTMB)* – nombre del objeto: *hsv_nbinom2_0_1*

```
Family: nbinom2  ( log )
Formula:          Severidad ~ Cultivar + Age + Cultivar:Age + (1 |
Planta)
Data: hsv

     AIC      BIC   logLik deviance df.resid
  3257.2   3311.6  -1614.6   3229.2       346

Random effects:

Conditional model:
 Groups Name        Variance Std.Dev.
 Planta (Intercept) 0.1185   0.3442
Number of obs: 360, groups:  Planta, 24

Dispersion parameter for nbinom2 family (): 3.15

Conditional model:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          4.78785    0.17546  27.287  < 2e-16 ***
CultivarDAnjou       0.11296    0.24785   0.456   0.6486
CultivarPackhams    -0.55827    0.24907  -2.241   0.0250 *
CultivarWilliams    -3.48778    0.26847 -12.991  < 2e-16 ***
Age2                 1.01668    0.14792   6.873 6.28e-12 ***
Age3                -0.95771    0.15164  -6.316 2.69e-10 ***
CultivarDAnjou:Age2 -0.63896    0.20873  -3.061   0.0022 **
CultivarPackhams:Age2 -0.84163    0.21115  -3.986 6.72e-05 ***
CultivarWilliams:Age2 -2.11225    0.28137  -7.507 6.05e-14 ***
CultivarDAnjou:Age3  -0.04793    0.21315  -0.225   0.8221
CultivarPackhams:Age3  0.03856    0.21967   0.176   0.8607
CultivarWilliams:Age3  0.01771    0.27404   0.065   0.9485
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



Analysis of Deviance Table (Type II Wald chisquare tests)

Response: Severidad
             Chisq Df Pr(>Chisq)
Cultivar    385.00  3  < 2.2e-16 ***
Age         273.29  2  < 2.2e-16 ***
Cultivar:Age  71.25  6  2.265e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Modelo 13

Modelo de regresión Binomial Negativo BN2

ziformula=~0; dispformula=~Cultivar

función *glmmTMB* (librería *glmmTMB)* – nombre del objeto: *hsv_nbinom2_0_C*

```
Family: nbinom2  ( log )
Formula:          Severidad ~ Cultivar + Age + Cultivar:Age + (1 |
Planta)
Dispersion:                 ~Cultivar
Data: hsv

     AIC      BIC   logLik deviance df.resid
  3143.0   3209.1  -1554.5   3109.0      355

Random effects:

Conditional model:
 Groups Name        Variance Std.Dev.
 Planta (Intercept) 0.04473  0.2115
Number of obs: 360, groups:  Planta, 24

Conditional model:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           4.77765    0.11628   41.09  < 2e-16 ***
CultivarDAnjou        0.11366    0.15865    0.72 0.473723
CultivarPackhams     -0.57360    0.19528   -2.94 0.003311 **
CultivarWilliams     -3.32033    0.42346   -7.84 4.47e-15 ***
Age2                  1.01672    0.10943    9.29  < 2e-16 ***
Age3                 -0.95111    0.11326   -8.40  < 2e-16 ***
CultivarDAnjou:Age2  -0.63859    0.14244   -4.48 7.36e-06 ***
CultivarPackhams:Age2 -0.82584    0.21496   -3.84 0.000122 ***
CultivarWilliams:Age2 -2.18935    0.58729   -3.73 0.000193 ***
CultivarDAnjou:Age3  -0.05122    0.14736   -0.35 0.728168
CultivarPackhams:Age3  0.09669    0.22515    0.43 0.667602
CultivarWilliams:Age3 -0.01860    0.57988   -0.03 0.974410
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Dispersion model:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.7640     0.1664  10.602  < 2e-16 ***
CultivarDAnjou      0.3748     0.2407   1.557    0.119
CultivarPackhams   -1.0502     0.2293  -4.579 4.67e-06 ***
CultivarWilliams   -3.2423     0.2940 -11.027  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



Analysis of Deviance Table (Type II Wald chisquare tests)

Response: Severidad
               Chisq Df Pr(>Chisq)
Cultivar     245.284  3  < 2.2e-16 ***
Age          525.380  2  < 2.2e-16 ***
Cultivar:Age  45.448  6  3.812e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Modelo 14

Modelo de regresión Binomial Negativo BN2

ziformula=~0; dispformula=~Cultivar*Age

función *glmmTMB* (librería *glmmTMB)* – nombre del objeto: *hsv_nbinom2_0_i*

```
Family: nbinom2  ( log )
Formula:          Severidad ~ Cultivar + Age + Cultivar:Age + (1 |
Planta)
Dispersion:                 ~Cultivar * Age
Data: hsv

     AIC      BIC   logLik deviance df.resid
  3063.9   3161.1  -1507.0   3013.9      335

Random effects:

Conditional model:
 Groups Name        Variance Std.Dev.
 Planta (Intercept) 0.02011  0.1418
Number of obs: 360, groups:  Planta, 24

Conditional model:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           4.78451    0.09252   51.71  < 2e-16 ***
CultivarDAnjou        0.10386    0.11499    0.90    0.366
CultivarPackhams     -0.60357    0.12926   -4.67 3.02e-06 ***
CultivarWilliams     -3.29186    0.40607   -8.11 5.20e-16 ***
Age2                  0.99576    0.08066   12.34  < 2e-16 ***
Age3                 -0.92091    0.13760   -6.69 2.19e-11 ***
CultivarDAnjou:Age2  -0.61197    0.10656   -5.74 9.31e-09 ***
CultivarPackhams:Age2 -0.76610    0.15696   -4.88 1.06e-06 ***
CultivarWilliams:Age2 -2.18531    0.55369   -3.95 7.92e-05 ***
CultivarDAnjou:Age3   -0.06253    0.17089   -0.37    0.714
CultivarPackhams:Age3  0.20129    0.24912    0.81    0.419
CultivarWilliams:Age3 -0.04569    0.58240   -0.08    0.937
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Dispersion model:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)            1.9299     0.2963   6.513 7.35e-11 ***
CultivarDAnjou         1.5267     0.4506   3.388 0.000704 ***
CultivarPackhams       0.1388     0.4355   0.319 0.749873
CultivarWilliams      -3.3882     0.4622  -7.331 2.28e-13 ***
Age2                   1.4675     0.5388   2.724 0.006458 **
Age3                  -0.9781     0.3978  -2.459 0.013936 *
CultivarDAnjou:Age2   -2.6319     0.6894  -3.818 0.000135 ***
CultivarPackhams:Age2  -2.5728     0.6851  -3.755 0.000173 ***
CultivarWilliams:Age2  -1.2930     0.7956  -1.625 0.104091
CultivarDAnjou:Age3   -1.0776     0.5989  -1.799 0.071996 .
CultivarPackhams:Age3  -1.1554     0.5714  -2.022 0.043149 *
CultivarWilliams:Age3   0.9461     0.6827   1.386 0.165804
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



Analysis of Deviance Table (Type II Wald chisquare tests)

Response: Severidad
              Chisq Df Pr(>Chisq)
Cultivar     350.463  3  < 2.2e-16 ***
Age          461.296  2  < 2.2e-16 ***
Cultivar:Age  63.322  6  9.489e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Modelo 15

Modelo de regresión Conway-Maxwell-Poisson

ziformula=~0; dispformula=~1

función *glmmTMB* (librería *glmmTMB)* – nombre del objeto: *hsv_compois_0_1*

```
Family: compois  ( log )
Formula:          Severidad ~ Cultivar + Age + Cultivar:Age + (1 |
Planta)
Data: hsv

     AIC      BIC   logLik deviance df.resid
  3133.0   3187.4  -1552.5   3105.0      346

Random effects:

Conditional model:
 Groups Name        Variance Std.Dev.
 Planta (Intercept) 0.05462  0.2337
Number of obs: 360, groups:  Planta, 24

Dispersion parameter for compois family (): 20.6

Conditional model:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           4.76514    0.11991   39.74  < 2e-16 ***
CultivarDAnjou        0.11347    0.16802    0.68 0.499451
CultivarPackhams     -0.59432    0.17895   -3.32 0.000896 ***
CultivarWilliams     -3.37343    0.24637  -13.69  < 2e-16 ***
Age2                  1.01994    0.08538   11.95  < 2e-16 ***
Age3                 -0.91523    0.12770   -7.17 7.68e-13 ***
CultivarDAnjou:Age2  -0.62186    0.12354   -5.03 4.81e-07 ***
CultivarPackhams:Age2 -0.80026    0.15155   -5.28 1.29e-07 ***
CultivarWilliams:Age2 -2.15128    0.31533   -6.82 8.95e-12 ***
CultivarDAnjou:Age3  -0.06574    0.17781   -0.37 0.711595
CultivarPackhams:Age3  0.16487    0.19952    0.83 0.408615
CultivarWilliams:Age3 -0.04642    0.31967   -0.15 0.884541
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



Analysis of Deviance Table (Type II Wald chisquare tests)

Response: Severidad
              Chisq Df Pr(>Chisq)
Cultivar     526.260  3  < 2.2e-16 ***
Age          469.197  2  < 2.2e-16 ***
Cultivar:Age  90.678  6  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Modelo 16

Modelo de regresión Conway-Maxwell-Poisson

ziformula=~0; dispformula=~Cultivar

función *glmmTMB* (librería *glmmTMB)* – nombre del objeto: *hsv_compois_0_C*

```
Family: compois  ( log )
Formula:          Severidad ~ Cultivar + Age + Cultivar:Age + (1 |
Planta)
Dispersion:                 ~Cultivar
Data: hsv

     AIC      BIC   logLik deviance df.resid
  3117.9   3184.0  -1541.9   3083.9      343

Random effects:

Conditional model:
 Groups Name        Variance Std.Dev.
 Planta (Intercept) 0.04752  0.218
Number of obs: 360, groups:  Planta, 24

Conditional model:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            4.76509    0.11228   42.44  < 2e-16 ***
CultivarDAnjou         0.11249    0.15332    0.73  0.46312
CultivarPackhams      -0.58574    0.18182   -3.22  0.00127 **
CultivarWilliams      -3.36705    0.25522  -13.19  < 2e-16 ***
Age2                   1.02036    0.08036   12.70  < 2e-16 ***
Age3                  -0.91524    0.12137   -7.54 4.67e-14 ***
CultivarDAnjou:Age2   -0.62134    0.10684   -5.82 6.04e-09 ***
CultivarPackhams:Age2 -0.80557    0.17299   -4.66 3.21e-06 ***
CultivarWilliams:Age2 -2.15181    0.33250   -6.47 9.70e-11 ***
CultivarDAnjou:Age3   -0.06561    0.15766   -0.42  0.67730
CultivarPackhams:Age3  0.14838    0.21784    0.68  0.49579
CultivarWilliams:Age3 -0.04258    0.33471   -0.13  0.89876
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Dispersion model:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)            2.8906     0.1725  16.756  < 2e-16 ***
CultivarDAnjou        -0.3733     0.2393  -1.560  0.1187
CultivarPackhams       0.6871     0.3059   2.246  0.0247  *
CultivarWilliams      16.8376  2527.8574   0.007  0.9947
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



Analysis of Deviance Table (Type II Wald chisquare tests)

Response: Severidad
             Chisq Df Pr(>Chisq)
Cultivar   532.270  3  < 2.2e-16 ***
Age        569.978  2  < 2.2e-16 ***
Cultivar:Age  89.323  6  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Modelo 17

Modelo de regresión Conway-Maxwell-Poisson

ziformula=~0; dispformula=~Cultivar*Age

función *glmmTMB* (librería *glmmTMB)* – nombre del objeto: *hsv_compois_0_i*

```
Family: compois  ( log )
Formula:          Severidad ~ Cultivar + Age + Cultivar:Age + (1 |
Planta)
Dispersion:                ~Cultivar * Age
Data: hsv

     AIC      BIC   logLik deviance df.resid
  3100.7   3197.9  -1525.4   3050.7      335

Random effects:

Conditional model:
 Groups Name        Variance Std.Dev.
 Planta (Intercept) 0.03077  0.1754
Number of obs: 360, groups:  Planta, 24

Conditional model:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           4.76441    0.10117   47.09  < 2e-16 ***
CultivarDAnjou        0.11906    0.12922    0.92    0.357
CultivarPackhams     -0.58851    0.14125   -4.17 3.09e-05 ***
CultivarWilliams     -3.33838    0.24280  -13.75  < 2e-16 ***
Age2                  1.01945    0.07889   12.92  < 2e-16 ***
Age3                 -0.91137    0.13457   -6.77 1.27e-11 ***
CultivarDAnjou:Age2  -0.62012    0.10309   -6.02 1.80e-09 ***
CultivarPackhams:Age2 -0.78987    0.14496   -5.45 5.06e-08 ***
CultivarWilliams:Age2 -2.16572    0.32974   -6.57 5.10e-11 ***
CultivarDAnjou:Age3  -0.06963    0.16553   -0.42    0.674
CultivarPackhams:Age3  0.19037    0.24182    0.79    0.431
CultivarWilliams:Age3 -0.04874    0.33763   -0.14    0.885
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Dispersion model:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           2.988e+00  3.057e-01   9.777  < 2e-16 ***
CultivarDAnjou       -1.307e+00  4.156e-01  -3.145  0.00166 **
CultivarPackhams     -7.191e-01  4.376e-01  -1.643  0.10034
CultivarWilliams      1.578e+01  1.966e+03   0.008  0.99360
Age2                 -5.687e-01  5.071e-01  -1.122  0.26203
Age3                  2.932e-01  5.185e-01   0.566  0.57171
CultivarDAnjou:Age2   1.853e+00  6.444e-01   2.875  0.00404 **
CultivarPackhams:Age2  1.780e+00  7.134e-01   2.495  0.01261 *
CultivarWilliams:Age2  1.133e+00  5.535e+03   0.000  0.99984
CultivarDAnjou:Age3   6.695e-01  6.795e-01   0.985  0.32448
CultivarPackhams:Age3  2.891e+01  5.690e+05   0.000  0.99996
CultivarWilliams:Age3  2.625e+00  1.264e+04   0.000  0.99983
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Analysis of Deviance Table (Type II Wald chisquare tests)

Response: Severidad
             Chisq Df Pr(>Chisq)
Cultivar     644.975  3  < 2.2e-16 ***
Age          501.473  2  < 2.2e-16 ***
Cultivar:Age  92.103  6  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Modelo 18

Modelo de regresión Binomial Negativo BN1

ziformula=~0; dispformula=~Cultivar*Age; REML=TRUE

función *glmmTMB* (librería *glmmTMB)* – nombre del objeto: *hsv_nbinom1f_0_i*

```
Family: nbinom1  ( log )
Formula:          Severidad ~ Cultivar + Age + Cultivar:Age + (1 |
Planta)
Dispersion:                 ~Cultivar * Age
Data: hsv

     AIC      BIC   logLik deviance df.resid
  3092.1   3189.3  -1521.1   3042.1      347

Random effects:

Conditional model:
 Groups Name        Variance Std.Dev.
 Planta (Intercept) 0.02811  0.1676
Number of obs: 360, groups:  Planta, 24

Conditional model:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           4.76834    0.09942   47.96  < 2e-16 ***
CultivarDAnjou        0.11995    0.12634    0.95    0.342
CultivarPackhams     -0.57964    0.14013   -4.14 3.53e-05 ***
CultivarWilliams     -3.25881    0.39862   -8.18 2.95e-16 ***
Age2                  1.02071    0.07968   12.81  < 2e-16 ***
Age3                 -0.91440    0.13951   -6.55 5.58e-11 ***
CultivarDAnjou:Age2  -0.61947    0.10563   -5.86 4.51e-09 ***
CultivarPackhams:Age2 -0.78958    0.15526   -5.09 3.67e-07 ***
CultivarWilliams:Age2 -2.21945    0.54254   -4.09 4.30e-05 ***
CultivarDAnjou:Age3  -0.06462    0.17188   -0.38    0.707
CultivarPackhams:Age3  0.21334    0.24764    0.86    0.389
CultivarWilliams:Age3 -0.04742    0.57308   -0.08    0.934
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Dispersion model:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           2.85609    0.29863   9.564  < 2e-16 ***
CultivarDAnjou       -1.34332    0.45858  -2.929  0.00340 **
CultivarPackhams     -0.65771    0.44269  -1.486  0.13736
CultivarWilliams      0.07322    0.60309   0.121  0.90336
Age2                 -0.50150    0.52566  -0.954  0.34007
Age3                  0.10474    0.41640   0.252  0.80139
CultivarDAnjou:Age2   1.96414    0.69366   2.832  0.00463 **
CultivarPackhams:Age2  1.72877    0.69274   2.496  0.01258 *
CultivarWilliams:Age2 -0.83240    0.95445  -0.872  0.38314
CultivarDAnjou:Age3   0.87763    0.63034   1.392  0.16383
CultivarPackhams:Age3  1.25913    0.61949   2.033  0.04210 *
CultivarWilliams:Age3 -1.00533    0.89061  -1.129  0.25898
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1



Analysis of Deviance Table (Type II Wald chisquare tests)

Response: Severidad
              Chisq Df Pr(>Chisq)
Cultivar     329.933  3  < 2.2e-16 ***
Age          477.090  2  < 2.2e-16 ***
Cultivar:Age  67.493  6  1.334e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```