



UNIVERSIDAD NACIONAL DE COLOMBIA

Estudio en bloques completos vía regresión Poisson en presencia de sobredispersión

Ana María Torres Blanco

Universidad Nacional de Colombia
Facultad Ciencias , Departamento de Estadística
Bogotá, Colombia
2011

Estudio de bloques completos vía regresión Poisson en presencia de sobredispersión

Ana María Torres Blanco

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

Director:
LUIS ALBERTO LOPEZ Ph.D

Universidad Nacional de Colombia
Facultad Ciencias, Departamento de Estadística
Bogotá, Colombia
2011

A mi mamá María Cleofe Blanco y a mi esposo
Adalberto Talaigua.

Agradecimientos

A Dios por haber permitido todas las condiciones para poder llegar a este punto.

A mi Director Luis Alberto López Ph.D por su generosidad y valiosa colaboración en el desarrollo de este trabajo.

A todos mis compañeros de estudio en especial Adalberto Talaigua.

A la Universidad de Cartagena y la Universidad Nacional de Colombia por haber hecho realidad el convenio de la maestría en Ciencias Estadísticas.

Ana María Torres Blanco

Resumen

En este trabajo se lleva a cabo el modelamiento estadístico de datos provenientes de un ensayo experimental realizado en el Instituto Colombiano Agropecuario ICA en 1983. En el ensayo se considera un diseño en bloques completamente aleatorizado para estudiar el grado de pudrición de mazorcas en nueve variedades de maíz. Como la variable respuesta de interés eran conteos, se procedió a modelar estos datos inicialmente sin tener en cuenta la sobredispersión, pero como se encontró en algunos grados evidencia de este fenómeno se procedió a modelarlo alternativamente con el modelo Binomial Negativo para efectos de modelar esta sobredispersión. Adicionalmente se trabajo con el método GSK propuesto por Grizzle et al. (1969), el cual busca modelar datos de conteos vía ajuste ponderado de modelos lineales bajo normalidad, y por último se ajustaron los datos sobre los grados de pudrición en este diseño en bloques con el modelo multinomial.

Como resultado a destacar en el desarrollo del trabajo, se encontró que dependiendo el grado de pudrición, los modelos propuestos fueron distintos. En el caso concreto del grado de pudrición 3 (G2) no fue necesario ajustar con modelos de sobredispersión, basto con ajustar un modelo Poisson, en los otros grados de pudrición fue necesario ajustar un modelo Binomial Negativo debido a que se encontró que había presencia de sobredispersión. Cuando se ajustó por el método GSK, los resultados mostraron una subestimación o sobrestimación, por lo tanto este método no es recomendado para datos de conteo con o sin presencia de sobredispersión.

Palabras clave: Sobredispersión, modelo Binomial Negativo, regresión Poisson, modelo multinomial.

Abstract

This work is carried out statistical modeling of data from an experimental trial conducted in the Colombian Agricultural Institute (ICA) in 1983. The trial is considered a completely randomized block design to study the degree of rot in nine ears of corn varieties. As the response variable of interest were counts, we proceeded to model these data initially without considering overdispersion, but as found in some degree evidence of this phenomenon is alternatively proceed to model it with the Negative Binomial model for purposes of modeling this on overdispersion. Additionally, working with the method proposed by GSK Grizzle et al. (1969), which seeks to model count data adjustment via weighted linear models under normality, and finally adjusted the data on the levels of decay in this block design with the multinomial model. As a notable result in the development of work, it was found that depending on the degree of decay, the proposed models were different. In the case of the degree of rot 3 (G2) was not necessary to adjust for overdispersion models, coarse with a Poisson model fit in the other degrees of decay was necessary to adjust a Negative Binomial model because it was found that was present over- dispersion. When adjusted for GSK method, the results showed an underestimation or overestimation, so this method is not recommended for count data with or without the presence of overdispersion.

Keywords: overdispersion, negative binomial model, Poisson regression, multinomial model.

Contenido

Agradecimientos	vii
Resumen	ix
Abstract	x
1. Introducción	1
2. Revisión teórica	4
2.1. Familia exponencial	4
2.2. Modelos lineales generalizados (GLM)	5
2.2.1. Estimación e inferencia	5
2.3. Regresión Poisson	8
2.4. Sobredispersión	11
2.5. Modelos de sobredispersión	12
2.5.1. Modelos de sobredispersión para datos de conteo	14
2.5.2. Pruebas de sobredispersión	17
2.5.3. Selección de covariables	18
2.5.4. Diagnósticos	18
2.6. Modelo Multinomial	19
2.6.1. Regresión Logística Nominal	22
2.6.2. Regresión Logística Ordinal	22
2.7. Método GSK	23
3. Aplicación de los modelos	27
3.1. Análisis exploratorio	28
3.2. Ajuste de los modelos	29
3.2.1. Ajuste del modelo de regresión Poisson	29
3.2.2. Ajuste del modelo Binomial Negativo	35
3.2.3. Ajuste del modelo Multinomial	39
3.2.4. Ajuste del modelo GSK	41
4. Conclusiones y recomendaciones	43
4.1. Conclusiones	43

4.2. Recomendaciones	45
A. Anexo: Implementación de los modelos en R	46
Bibliografía	53

Lista de símbolos

Símbolos con letras latinas

Símbolo	Término
$a(\cdot)$	función
$b(\cdot)$	función
$c(\cdot)$	función
$d(\cdot)$	función
D	función desvío
$E()$	valor esperado
$g(\cdot)$	función de enlace
n	número de observaciones
o_i	valores observados
p	número de parámetros
Q_i^+	función de cuasi-verosimilitud extendida
Q_i	función de cuasi-verosimilitud
r_{p_i}	residual de Pearson
r_{A_i}	residual de Anscombe
r_{D_i}	residual de desvío
U	función score

Símbolos con letras griegas

Símbolo	Término
α	parámetro

Símbolo Término

β	parámetro
η	predictor lineal
θ	parámetro
$\hat{\theta}$	estimación del parámetro θ
μ	media de la variable aleatoria Y
π	probabilidades
τ	tratamiento
ϕ	parámetro de dispersión

Abreviaturas**Abreviatura Término**

<i>GLM</i>	Modelo Lineal Generalizado
<i>MBN</i>	Modelo Binomial Negativo
<i>MPR</i>	Modelo de regresión Poisson
<i>RR</i>	Riesgo Relativo
<i>GO</i>	Mazorcas clasificadas en el grado de pudrición cero
<i>G1</i>	Mazorcas clasificadas en el grado de pudrición uno
<i>G2</i>	Mazorcas clasificadas en el grado de pudrición dos

1. Introducción

En investigación experimental y no experimental es frecuente que se tenga interés en analizar datos cuya característica son conteos, proporciones con estructura binomial o proporciones continuas. Casos frecuentes de este tipo de datos los encontramos en áreas como finanzas, ecología o medio ambiente, áreas biológicas y médicas entre otras. Como ejemplos específicos de este tipo de datos se pueden mencionar: número de mazorcas clasificadas en cada grado de pudrición, número de insectos atrapados en trampas, número de accidentes en una aerolínea en un intervalo de tiempo, el número de visitas al médico en un año, el número de fármacos prescritos, entre otros.

Para el análisis de este tipo de datos, la teoría estadística ha venido proporcionando varios desarrollos teóricos, encontrando que hay una gran variedad de métodos estadísticos los cuales están basados en ajustes de modelos lineales clásicos como es el uso del método GSK propuesto por Grizzle et al. (1969), el cual se emplea para trabajar modelos lineales con datos categóricos, o los métodos recientemente propuestos en la literatura, los cuales están soportados en la teoría de los modelos lineales generalizados y los modelos de sobredispersión, en el caso concreto de datos de conteos estos se deben estudiar como datos que provienen de la distribución Poisson. Este trabajo se centra en el análisis de datos de conteos, los cuales deben modelarse teniendo como base la distribución Poisson. Como resultado importante de esta distribución se tiene que:

En 1837 se obtiene la distribución Poisson como caso límite de la binomial cuando existe un número de realizaciones observadas (n) grande y la probabilidad de éxito asociada al evento de interés es pequeña.

Posteriormente en 1898 Bortkiewicz realiza una de las primeras aplicaciones de esta distribución la cual consistió en modelar el número de muertes anuales causadas por patadas de mulas en el ejército de Prusia, los datos de estos eventos se pueden encontrar en el libro titulado *The Law of Small Numbers* y luego Greenwood & Yule (1920) obtuvieron una generalización de la distribución Poisson a la Binomial Negativa.

Otro punto importante en el desarrollo de los modelos de regresión de datos de conteos, fue la aparición de los “Modelos Lineales Generalizados” descritos por primera vez por Nelder & Wedderburn (1972) y desarrollados por McCullagh & Nelder (1989), quienes desarrollarán

la regresión Poisson como caso especial de estos modelos.

Sin embargo, el investigador al implementar el modelo de regresión Poisson se enfrenta con el supuesto de la igualdad de la media y la varianza, que en la vida real pocas veces se cumple, puesto que por lo general los datos Poisson son equidispersos y lo que realmente puede suceder es que la varianza que exhiban los datos sea mayor que la varianza teórica del modelo, es decir haya presencia de sobredispersión. McCullagh & Nelder (1989) señalan que cuando se trabaja con datos de conteos se debe prever este fenómeno.

En Hinde & Demetrio (2007) se encuentra un amplio desarrollo teórico para modelar datos en presencia de sobredispersión. Entre ellos se encuentran los modelos de media- varianza que asumen una forma general de la función de varianza incluyendo parámetros adicionales y los modelos en dos etapas o compuestos, que asumen que el parámetro asociado a la respuesta no es fijo sino que tiene alguna distribución de probabilidad conocida. Cox (1983) obtuvo el modelo Binomial Negativo a partir de la distribución Poisson tomando el parámetro de esta distribución como una variable aleatoria que se distribuye Gamma y, encontró que este modelo se puede usar como alternativa para corregir la sobredispersión. Teniendo en cuenta los resultados de Cox (1983), Morales & López (2008) implementaron y aplicaron el modelo en dos etapas y propusieron una prueba para la sobredispersión constante.

Considerando el amplio campo de aplicación de los modelos para datos de conteos, en este trabajo se exponen los resultados de un experimento realizado en el Centro Nacional de Investigaciones Agropecuaria ICA, cuyo objetivo fue evaluar el grado de pudrición de mazorcas, causados por el hongo *Fusarium*, S.P en la variedad de maíz sogamoseño, teniendo como ensayo experimental un diseño en bloques completos con seis repeticiones (bloques).

El abordaje inicial para el análisis de los datos provenientes de este ensayo, se hizo a través de la metodología GSK propuesta por Grizzle et al. (1969), como puede verse en el trabajo propuesto por López & Chavez (1984).

El método GSK analiza los datos como si fueran normales, esto por supuesto no es adecuado por tratarse de datos de conteos, por lo cual surge la necesidad de analizar este conjunto de datos a través de procedimientos alternativos como los modelos de regresión Poisson, la regresión basada en la distribución Binomial Negativa (por la presencia de sobredispersión) y la estimación basada en el modelo Multinomial.

En el desarrollo de este trabajo inicialmente se presenta una breve revisión teórica de los modelos lineales generalizados (MLG), enfatizando en los métodos de estimación, se estudia la regresión Poisson (MP), el modelo Multinomial y la metodología basada en la propuesta de Grizzle et al. (1969); posteriormente en el tercer capítulo se muestran los resultados de los

modelos de regresión Poisson, binomial negativo y multinomial, en donde se estudió el caso de la sobredispersión en el modelo de regresión Poisson. Finalmente en el cuarto capítulo se exponen las respectivas conclusiones y recomendaciones del trabajo.

2. Revisión teórica

En este capítulo se presenta una breve descripción teórica de los modelos que reporta la literatura para el análisis de datos donde la respuesta son conteos; se hace una revisión breve de la familia exponencial, modelos lineales generalizados (GLM), así como de la regresión Poisson, los modelos basados en la distribución Binomial Negativa (MBN) y modelo multinomial. Del mismo modo se hace una revisión del método GSK desarrollada por Grizzle et al. (1969) para los modelos con datos categóricos (de conteos) trabajados en forma normal.

2.1. Familia exponencial

De acuerdo con Dobson (2002) si se tiene una variable aleatoria cuya función de distribución de probabilidad depende de un sólo parámetro digamos θ , la distribución pertenece a la familia exponencial si se puede escribir en la forma:

$$f(y, \theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)] \quad (2-1)$$

donde las funciones $a(\cdot), b(\cdot), c(\cdot), d(\cdot)$ son funciones conocidas, si $a(y) = y$ se dice que la distribución está en la forma canónica natural y $b(\theta)$ se conoce como la distribución del parámetro natural.

En McCullagh & Nelder (1989), la ecuación (2-1) la escriben de la forma:

$$f(y, \theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right] \quad (2-2)$$

donde $a(\phi) > 0$, es una función del parámetro de dispersión que representa el parámetro de escala. θ es llamado el parámetro canónico, el cual se asocia a la localización.

Si en (2-2) se obtiene el logaritmo de la función de verosimilitud entonces la derivada con respecto a θ está dada por:

$$U = \frac{\partial(l(\theta, \phi; y))}{\partial\theta} = \frac{y - b'(\theta)}{a(\phi)} \quad (2-3)$$

(2-3) es llamada la función score, la cual cumple las siguientes propiedades: $E(U) = 0$ y $Var(U) = -E(U')$.

Si en (2-3) se obtiene el valor esperado y se iguala a cero, entonces se sigue que $E[Y] = b'(\theta)$ y al obtener en (2-3) la primera derivada con respecto a θ y por propiedades de la función score se tiene que $Var(U) = \frac{b''(\theta)}{a(\phi)}$. Los desarrollos formales de estos resultados pueden encontrarse en Dobson (2002), Hinde & Demetrio (2007) entre otras.

En la siguiente sección se hace un breve desarrollo sobre los modelos lineales generalizados, los cuales tienen gran interés en el desarrollo de este trabajo.

2.2. Modelos lineales generalizados (GLM)

Sean Y_1, \dots, Y_n variables aleatorias independientes, cada una con distribución de la familia exponencial, como en (2-1) y (2-2), las cuales satisfacen las siguientes tres condiciones características:

1. **Componente aleatoria:** especifica la distribución de probabilidad de la variable respuesta. Las observaciones independientes $Y_i, i = 1, \dots, n$ son idénticamente distribuidas pertenecientes a la familia exponencial en forma canónica.
2. **Componente sistemática:** asocia a cada valor de la variable respuesta $Y_i, i = 1, \dots, n$ un vector de covariables $x_i^t = (x_{i1}, \dots, x_{ip})$ de dimensión p , que multiplicado con otro vector fijo de parámetros desconocidos $\beta = (\beta_1, \dots, \beta_p)$ da como resultado el i -ésimo predictor lineal $\eta_i = x_i^t \beta$.

En notación matricial se puede escribir $\eta = X\beta$ donde η es un vector de orden $n \times 1$, X es la matriz de diseño del modelo de orden $n \times p$ y β es el vector de parámetros.

3. **Función de enlace :** función monótona y diferenciable $g(\cdot)$, que relaciona la media de la variable respuesta con el predictor lineal η_i de la siguiente forma: $g(\mu_i) = \eta_i = x_i^t \beta$ con $i = 1, \dots, n$

La función $g(\cdot)$ proporciona una escala lineal de relación entre la media y el vector de covariables.

La combinación de estos tres componentes define completamente un GLM, en el que se busca estimar β y el parámetro de dispersión ϕ (en algunos casos como en la distribución binomial y Poisson este parámetro es igual a uno).

2.2.1. Estimación e inferencia

El procedimiento de estimación más conocido en GLM está basado en el método de la máxima verosimilitud, el cual consiste en solucionar para β el logaritmo que maximiza la

verosimilitud que según McCullagh & Nelder (1989) está dada por:

$$l(\mu(\beta), \phi; y) = \sum_{i=1}^n \left[\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y, \phi) \right] = \sum_{i=1}^n l_i(\mu_i(\beta), \phi; y_i) \quad (2-4)$$

Para obtener la estimación de máxima verosimilitud de los parámetros β_j se derivan las funciones de verosimilitud de la forma siguiente:

$$U_j = \frac{\partial l(\mu(\beta), \phi; y)}{\partial \beta_j} = 0 \quad j=1, \dots, p \text{ con } \beta_j \text{ el } j\text{-ésimo elemento de } \beta$$

U_j se define como la j -ésima componente de la función score. Al derivar la función U parcialmente con respecto a β_j , aplicando la regla de la cadena se tiene el siguiente sistema de ecuaciones:

$$U_j = \frac{\partial l(\mu(\beta), \phi; y)}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial l_i(\mu(\beta), \phi; y_i)}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}$$

para $j = 1, \dots, p$ donde cada término del lado derecho de la expresión anterior queda determinado por:

$$\frac{\partial l_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)}$$

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\frac{\partial \mu_i}{\partial \theta_i}} = \frac{1}{b''(\theta_i)}$$

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{\frac{\partial \eta_i}{\partial \mu_i}} = \frac{1}{g'(\mu_i)}$$

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$$

con estos resultados se tiene que la j -ésima componente de la función score omitiendo el parámetro de dispersión está dada por

$$U_j = \sum_{i=1}^n \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \sum_{i=1}^n \frac{(Y_i - \mu_i)}{a(\phi)} \frac{1}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij}; j = 1, \dots, p \quad (2-5)$$

La estimación de máxima verosimilitud de β denotada por $\hat{\beta}$ se obtiene de la solución de (2-5) cuando $U_j = 0$. El proceso de estimación se desarrolla por medio de métodos numéricos iterativos como el de Newton Raphson o una variante del mismo, conocida como el algoritmo de Fisher -Scoring, (ver detalles en McCullagh & Nelder (1989)). Cuando el modelo incluye un parámetro ϕ desconocido, se acostumbra a no incluirlo en las ecuaciones dadas en (2-5), en lugar de ello la estimación se hace como función de los residuales de Pearson generalizado, los cuales según McCullagh & Nelder (1989) están definidos como:

$$a(\hat{\phi}) = \frac{1}{n - p - 1} \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{Var(\hat{\mu}_i)}$$

La inferencia en un GLM está basada en el estadístico de desvío. Este estadístico se fundamenta en la prueba de razón de verosimilitud, el cual es dado por:

$$D = 2\{l[\hat{\beta}_{max}; y] - l[\hat{\beta}; y]\}$$

donde $\hat{\beta}_{max}$ se define como el estimador de máxima verosimilitud para β en un modelo saturado (con tantos parámetros como vectores de covariables x_i^t distintos tenga el modelo) con m parámetros y $\hat{\beta}$ estimador de máxima verosimilitud para el modelo de interés con p parámetros. El estadístico desvío tiene una distribución asintótica chi-cuadrada con $m - p$ grados de libertad.

Este estadístico a su vez proporciona una forma de juzgar la bondad de ajuste del modelo; ya que valores muy grandes en el desvío conlleva a que haya una o más fuentes de variabilidad no explicadas por el modelo.

En el caso de una respuesta de conteo, típicamente se plantea un GLM cuya componente aleatoria es una variable de Poisson. Este modelo distribucional posee la restricción de igualdad entre la media y la varianza; sin embargo otra opción (aunque menos frecuente) es modelar la respuesta de conteo por medio de una distribución Binomial Negativa.

Residuales

Según McCullagh & Nelder (1989) para modelos lineales normales se puede expresar la variable dependiente en la forma

$$Y = \hat{\mu} + (Y - \hat{\mu})$$

es decir, dato = valor ajustado + residual.

Los residuales pueden ser usados para explorar la adecuación del ajuste de un modelo con respecto de la escogencia de la función de varianza, función de enlace y términos en el predictor lineal. Los residuales pueden también indicar la presencia de valores atípicos los cuales requieren una mayor investigación.

Los modelos lineales generalizados requieren una definición más amplia de los residuales, ya que la distribución de estos van a depender de la distribución que sigue la variable dependiente del modelo de interés. Además es adecuado que estos residuales puedan ser usados para el mismo propósito que los residuales normales estándar.

A continuación se presentan algunos tipos de residuales ampliamente estudiados en los modelos lineales generalizados, se presentan específicamente los residuales de Pearson, residuales de Anscombe y los residuales de componentes de desvío.

a. Residuales de Pearson

Los residuales de Pearson están definidos por:

$$r_{p_i} = \frac{Y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} \quad (2-6)$$

donde los Y_i son los valores observados y los $\hat{\mu}_i$ son los valores ajustados. El nombre de estos residuales se toma del hecho que para la distribución Poisson, los residuales de Pearson son precisamente la raíz cuadrada de la componente de la estadística de bondad de ajuste X^2 de Pearson, los cuales satisfacen que $\sum r_{p_i}^2 = X^2$

b. Residuales de Anscombe

Para MLG con variable dependiente no normal, la distribución de los residuales de Pearson es a menudo asimétrica. Con el fin de superar esta dificultad Anscombe (1953) propuso un residual que utilizaba una función $A(Y)$ en lugar de Y en la derivación de residuales McCullagh & Nelder (1989). La función $A()$ es elegida para hacer la distribución de Y lo más normal posible y viene dada por:

$$A(Y) = \int \frac{du}{\sqrt[3]{Var(u)}}$$

donde el residual esta dado por:

$$r_{A_i} = \frac{A(Y_i) - A(\hat{\mu}_i)}{A'(\hat{\mu}_i)\sqrt{Var(\hat{\mu}_i)}} \quad (2-7)$$

c. Residuales de Desvio

El residual más utilizado en MLG es el residual de componente de desvio, el cual se basa en la contribución de desvio global aportada por cada observación, de forma similar a los residuales de los modelos normales. Así el desvio juega un papel clave en las derivaciones del MLG y en las inferencias de los resultados. Estos se definen como:

$$r_{D_i} = sign(Y_i - \hat{\mu}_i)\sqrt{d_i} \quad (2-8)$$

donde los Y_i son los valores observados, los $\hat{\mu}_i$ son los valores ajustados y d_i es la contribución individual al desvio.

2.3. Regresión Poisson

En la siguiente sección se presenta un resumen teórico de la regresión Poisson, ya que esta fue la base para ajustar el modelo a los datos provenientes del ensayo experimental de referencia.

Los modelos de regresión Poisson son ampliamente utilizados en el análisis de datos de conteos, para los cuales se considera como variable dependiente por ejemplo los conteos que se presentan en un intervalo de tiempo, en un determinado espacio o región. Algunos ejemplos de este tipo de variables son el número de insectos en una planta, el número de accidentes de tránsito reportados en una ciudad en un período de tiempo, número de mazorcas podridas luego de haber sido sometidas a un tratamiento, etc. Como la variable dependiente Y representa conteos, entonces esta sigue una distribución Poisson, es decir, $Y \sim Poisson(\mu)$ y su función de densidad de probabilidad esta dada por

$$f_Y(y) = \begin{cases} e^{-\mu} \frac{\mu^y}{y!}, & y = 0, 1, \dots \\ 0, & e.o.c \end{cases}$$

Además la variable aleatoria Y con distribución Poisson de parámetro μ se caracteriza por tener valor esperado y varianza igual, es decir $E(Y) = \mu$ y $Var(Y) = \mu$. Una propiedad importante de esta distribución, es que si se considera Y_1, \dots, Y_n variables aleatorias independientes con distribución Poisson, entonces la suma de las Y_1, \dots, Y_n variables aleatoria también se distribuye Poisson, es decir, si $Y_i \sim Poisson(\mu_i)$ entonces

$$Y_1 + \dots + Y_n \sim Poisson(\mu_1 + \dots + \mu_n)$$

La distribución de una variable aleatoria Poisson pertenece a la familia exponencial, ya que se puede escribir en la forma (2-2) de la siguiente manera $f_Y(y, \mu) = \exp\{y \ln \mu - \mu - \ln y!\}$ donde se observa que el parámetro natural es $\theta = \ln(\mu)$, el parámetro de escala $\phi = 1$ la función de varianza es $Var(Y) = \mu$ y la función de enlace natural es la función logaritmo natural.

Si se supone ahora que las variables aleatoria Y_i , $i = 1, \dots, n$ representan conteos con medias μ_i , entonces el modelo Poisson estándar asume que

$$Y_i \sim Poisson(\mu_i) \quad \text{con} \quad Var(Y_i) = \mu_i \quad (2-9)$$

Luego en forma general en estudios donde se busca modelar datos con la regresión Poisson se pretende representar la media de una variable de tipo conteo teniendo en cuenta un conjunto de covariables o factores de interés, los cuales van a estar relacionados como

$$\mu_i = n_i \theta_i$$

donde θ_i depende de las variables explicativas $(x_{i1}, x_{i2}, \dots, x_{ip})$. La cual es usualmente modelada por $\theta_i = e^{x_i^t \beta}$ por tanto el modelo lineal generalizado es

$$E(Y_i) = \mu_i = n_i e^{x_i^t \beta} \quad (2-10)$$

la función de enlace natural es la función logaritmo natural la cual se expresa de la siguiente manera

$$\ln(\mu_i) = \ln(n_i) + x_i^t \beta \quad (2-11)$$

en la ecuación (2-11) el término $\ln(n_i)$ se conoce como una variable *offset*, el cual hace que la especificación de la componente lineal del modelo Poisson estándar.

La estimación de los parámetros β en el modelo (2-10) se hace vía estimación de máxima verosimilitud. Si se utiliza el enlace canónico o natural en este modelo, las ecuaciones de score y el estadístico de desvío, según Dobson (2002) para el modelo de referencia toman la forma:

$$U_j = \sum_{i=1}^n \left[\frac{y_i - \exp(x_i^t \beta)}{\exp(x_i^t \beta)} \right] x_{ij} \exp(x_i^t \beta) \quad j = 0, 1, \dots, p,$$

y

$$D = 2 \sum_{i=1}^n \left\{ y_i \ln \left(\frac{y_i}{\exp(x_i^t \beta)} \right) - (y_i - \exp(x_i^t \beta)) \right\} \quad (2-12)$$

Una vez obtenidos los estimadores, se busca una mejor interpretación de los resultado. Dichos resultados van encaminados a observar el impacto que tienen las variables explicativas al momento de explicar la media en la variable respuesta, en la aplicación estas variables explicativas pueden ser:

a. Variables indicadoras.

b. Variables continuas.

Para el caso **a**, una manera de obtener una interpretación es por medio del Riesgo Relativo (RR), está estadística es dada por:

$$RR = \frac{E(Y_i | presencia)}{E(Y_i | ausencia)} = \frac{n_i e^{\beta_0 + \dots + x_i \beta_i + \dots + x_k \beta_k}}{n_i e^{\beta_0 + \dots + x_{i-1} \beta_{i-1} + x_i \beta_i + \dots + x_k \beta_k}} = e^{\beta_i}$$

Como su nombre lo indica, la estadística da a conocer cuánto riesgo se va a tener si se expone al contexto creado en presencia de la variable indicadora.

Para el caso **b**, la metodología para encontrar una interpretación es similar. Por medio del RR aplicado en el modelo de referencia va a dar la información de lo que ocurre con la media de los datos si aumenta en una unidad la variable explicativa, su forma matemática es dada por:

$$RR = \frac{E(Y_i|x_i + 1)}{E(Y_i|x_i)} = \frac{n_i e^{\beta_0 + \dots + (x_i+1)\beta_1 + \dots + x_k\beta_k}}{n_i e^{\beta_0 + \dots + x_{i-1}\beta_{i-1} + x_i\beta_i + \dots + x_k\beta_k}} = e^{\beta_i}$$

Residuales

En el modelo de regresión Poisson los residuales (diferencia entre los valores estimados y los valores observados) más utilizados son:

a. Desvío Residual

En el modelo Poisson la función de desvío dada en (2-12) se puede escribir como

$$D = 2 \sum o_i \ln\left(\frac{o_i}{e_i}\right)$$

donde o_i y e_i denotan los valores observados y estimados respectivamente. Así en el modelo Poisson el valor de D puede ser calculado directamente de los datos, y ser comparado con la distribución X_{n-p}^2 donde p es el número de parámetros a estimar.

Luego la i -ésima componente del desvío residual para el modelo Poisson es

$$r_{D_i} = \text{sig}(o_i - e_i) \sqrt{2 \left[o_i \log\left(\frac{o_i}{e_i}\right) - (o_i - e_i) \right]}, i = 1, 2, \dots, N$$

de modo que $D = \sum r_{D_i}^2$

b. Residuales de Pearson

Los residuales de Pearson para el modelo de regresión Poisson son definidos por $r_{pi} = \frac{o_i - e_i}{\sqrt{e_i}}$ donde o_i y e_i denotan los valores observados y estimados respectivamente, de modo que

$$X^2 = \sum r_{pi}^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

La estadística X^2 al igual que D puede ser comparada con la distribución X_{n-p}^2 donde p es el número de parámetros a estimar.

2.4. Sobredispersión

En muchas situaciones prácticas, cuando se ajusta un modelo lineal generalizado a un conjunto de datos, se puede observar un desvío mucho más grande que el esperado por el modelo, es decir que la variación de los datos sea mucho más grande que la predicha por el modelo. A este fenómeno de acuerdo con Hinde & Demetrio (2007) se le conoce como sobredispersión.

En experimentos controlados o no controlados, casi que la sobredispersión de los datos es la regla, y particularmente en datos de conteos esta es más bien la excepción, pues pocos conjuntos de datos Poisson son realmente equidispersos. Por tanto se puede decir que la sobredispersión en cierta medida es inherente a la gran mayoría de los datos Poisson.

La sobredispersión puede ser originada por varias causas entre ellas tenemos:

- i. La correlación positiva de la variable respuesta y las explicativas .
- ii. Cuando hay violación en los supuestos de la distribución de los datos, por ejemplo, cuando los datos se agrupan y las probabilidades violan el supuesto de independencia de las observaciones.
- iii. El modelo omite predictores, los datos son atípicos, la función de enlace es mal especificada. El modelo no puede incluir un número suficiente de términos de interacción.

La sobredispersión puede causar ciertos problemas tales como: errores estándar de las estimaciones de los parámetros que se sobrestima o subestima, es decir, una variable puede aparecer como un predictor significativo cuando de hecho no lo es. Lo que conlleva a que haya una mala interpretación del modelo ajustado.

2.5. Modelos de sobredispersión

Para ajustar datos en presencia de sobre dispersión, el modelamiento estadístico presenta varias alternativas que surgen de los supuestos que se asumen para explicar este fenómeno. Según Hinde & Demetrio (2007), los modelos se pueden clasificar en dos grandes grupos:

1. Modelos de media varianza: asumen una forma más general de la función de varianza, incluidos parámetros adicionales. Estos posiblemente no correspondan a una distribución de probabilidad específica para la respuesta, pero se puede ver como una extensión del modelo básico; cuando esto sucede los parámetros pueden estimarse usando métodos de cuasi verosimilitud.

2. Modelos en dos etapas: asumen que el parámetro asociado a la respuesta no es fijo; sino que tiene alguna distribución de probabilidad conocida. Estos modelos conducen a un modelo de probabilidad compuesto; en principio todos los parámetros pueden estimarse usando máxima verosimilitud. Sin embargo, en general la distribución resultante no toma una forma simple.

Entre los métodos de estimación utilizados en los modelos de sobredispersión se pueden mencionar: el de máxima verosimilitud, máxima cuasi-verosimilitud, cuasi-verosimilitud extendida, y la pseudo verosimilitud.

Estimación de máxima cuasi-verosimilitud: este método introducido por Wedderburn (1974), proporciona una forma de estimación cuando se cuenta con funciones de varianza y hay una función de enlace pero no se tiene conocimiento de la distribución de la variable respuesta.

Una exploración del proceso de ajuste de los modelos lineales generalizados revela que las ecuaciones de estimación de máxima verosimilitud sólo dependen de la distribución de la variable respuesta Y_i a través de μ_i y $Var(Y_i) = Var(\mu_i)$. La estimación por cuasi-verosimilitud asume sólo una relación entre la media y la varianza en lugar de una distribución específica para la respuesta Y_i . Se tiene una función de enlace y un predictor lineal en la forma usual de los modelos lineales generalizados pero en lugar de asumir una distribución para Y_i sólo se supone que $Var(Y_i) = \phi Var(\mu_i)$ para alguna función conocida $Var(\cdot)$ Agresti (2002).

El cálculo del vector de parámetros β y los errores estándar a menudo no es suficiente si se requiere alguna forma de inferencia. Para calcular un desvío se requiere una verosimilitud y para calcular una verosimilitud se requiere una distribución. Se necesita entonces una alternativa para la verosimilitud que pueda calcularse sin asumir una distribución de probabilidad.

Sean Y_i para $i = 1, \dots, n$ variables aleatorias con media μ_i y varianza $Var(\mu_i)$, donde se asume que las Y_i son independientes. Wedderburn (1974) define la función score como:

$$U_i = \frac{Y_i - \mu_i}{\phi Var(\mu_i)}$$

se verifica que

$$E(U_i) = 0 \quad y \quad Var(U_i) = \frac{1}{\phi Var(\mu_i)} = -E\left(\frac{\partial U_i}{\partial \mu_i}\right)$$

propiedades que también tiene la derivada del logaritmo de verosimilitud. Esto sugiere que se puede usar U en lugar del logaritmo de la verosimilitud. Así se define según Wedderburn (1974)

$$Q_i(y_i, \mu_i) = \int^{\mu_i} \frac{y_i - t}{Var(t)} dt$$

lo que se busca es que Q_i se comporte como la función de log-verosimilitud.

Estimación por cuasi-verosimilitud extendida: el método de cuasi-verosimilitud original Wedderburn (1974), asume que el parámetro de dispersión es el mismo para todas las

observaciones. En ciertas aplicaciones puede ser necesario verificar ese supuesto o tal vez modelar ϕ como una función de variables conocidas. Según McCullagh & Nelder (1989), la idea básica de este método es construir una función $Q^+(\mu; y)$ que, para ϕ conocido, sea esencialmente la misma $Q(\mu; y)$ pero que exhiba las propiedades de una log-verosimilitud en lo referente a la derivada respecto a ϕ ; así, para una observación simple Y con media μ y varianza $\phi Var(\mu)$, se define la cuasi verosimilitud extendida como Nelder & Pregibon (1987)

$$Q_i^+(\mu_i, \phi; y_i) = -\frac{1}{2} \ln 2\pi(y_i) - \frac{1}{2\phi} D(y_i; \mu_i)$$

con $D(y; \mu) = -2 \int_y^\mu \frac{y_i - t}{Var(t)} dt$.

Estimación por pseudo-verosimilitud: una alternativa para extender el método de cuasi-verosimilitud es el enfoque de la pseudo-verosimilitud (PL) propuesto por Carroll and Ruppert (1988). Aquí la estimación de los β se obtiene por mínimos cuadrados generalizados, la cual es equivalente a la estimación de cuasi-verosimilitud para valores dados de ϕ .

La estimación de parámetros adicionales en la varianza se basa en la maximización de

$$P = -\frac{1}{2} \sum_{i=1}^n \left[\frac{(y_i - \mu_i)^2}{\phi_i Var(\mu_i)} + \ln(2\pi\phi_i Var(\mu_i)) \right]$$

Esto es de la misma forma que Q^+ , pero reemplazando los incrementos de la desviación por el cuadrado de los residuales de Pearson y $Var(y_i)$ por $V(\mu_i)$, que corresponde a una función de probabilidad normal para los residuales.

2.5.1. Modelos de sobredispersión para datos de conteo

En este trabajo se hace énfasis en los modelos de sobredispersión para datos de conteo, por consiguiente se hace una breve revisión teórica de estos modelos.

Supóngase que las variables aleatoria Y_i , $i = 1, \dots, n$ representan conteos con medias μ_i . El modelo Poisson estándar dado en (2-9) asume que $Var(Y_i) = \mu_i$. Cuando hay presencia de sobredispersión se deben tener en cuenta funciones que predicen mayor variabilidad. En este orden de idea un modelo simple de sobredispersión constante reemplaza $Var(Y_i) = \mu_i$ por $Var(Y_i) = \phi\mu_i$ y el proceso de estimación se hace vía cuasi-verosimilitud.

En el caso de datos de conteo, para respuestas provenientes de ensayos sin restricciones en la aleatorización de un experimento con $k > 2$ tratamientos y, con r_i repeticiones dentro del tratamiento i , la j -ésima observación del tratamiento i es y_{ij} , con $i = 1, 2, \dots, p$ y $j = 1, 2, \dots, r_i$ donde y_{ij} es el número de ocurrencias de un evento en una unidad de tiempo o espacio; que puede tomar los valores $0, 1, 2, \dots$ y se asume como una realización de la

variable aleatoria Y_{ij} tal que $Y_{ij}|\Theta_i \sim Poisson(\Theta_i)$. Donde se supone que Θ_i es una variable aleatoria con distribución Gamma de media μ_i , índice α y escala $\lambda_i = \frac{\alpha}{\mu_i}$ de tal forma que la distribución no condicional de Y_{ij} es binomial negativa la cual tiene función de distribución de probabilidad

$$f_{Y_{ij}}(y_{ij}) = \frac{\Gamma(\alpha + y_{ij})}{\Gamma(\alpha) y_{ij}!} \left(\frac{\mu_i}{\mu_i + \alpha}\right)^{y_{ij}} \left(\frac{\alpha}{\mu_i + \alpha}\right)^\alpha \quad (2-13)$$

con parámetros α y μ_i . En efecto como $Y|\Theta_i \sim Poisson(\Theta_i)$, entonces su función de densidad de probabilidad está dada por:

$$f_{Y_{ij}|\Theta_i}(y_{ij}|\theta_i) = \frac{\exp(-\theta_i) \theta_i^{y_{ij}}}{y_{ij}!} I_{0,1,\dots}(y_{ij}|\theta_i)$$

donde $E(Y_{ij}|\Theta_i) = \Theta_i$ y $Var(Y_{ij}|\Theta_i) = \Theta_i$. Análogamente, como Θ_i se distribuye gamma su función de densidad de probabilidad está dada por :

$$f_{\Theta_i}(\theta_i) = \frac{\lambda_i}{\Gamma(\alpha)} (\lambda_i \theta_i)^{\alpha-1} \exp(-\lambda_i \theta_i) I_{(0,\infty)}(\theta_i)$$

como $\lambda_i = \frac{\alpha}{\mu_i}$ se tiene que

$$f_{\Theta_i}(\theta_i) = \frac{\frac{\alpha}{\mu_i}}{\Gamma(\alpha)} \left(\frac{\alpha}{\mu_i} \theta_i\right)^{\alpha-1} \exp\left(-\frac{\alpha}{\mu_i} \theta_i\right) I_{(0,\infty)}(\theta_i)$$

donde

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} \exp(-t) dt$$

$$\begin{aligned} E(\Theta_i) &= \frac{\alpha}{\frac{\alpha}{\mu_i}} \\ &= \frac{\alpha \mu_i}{\alpha} \\ &= \mu_i \end{aligned}$$

$$\begin{aligned} Var(\Theta_i) &= \frac{\alpha}{\left(\frac{\alpha}{\mu_i}\right)^2} \\ &= \frac{\mu_i^2}{\alpha} \end{aligned}$$

Ahora por la regla de Bayes, la función de densidad de probabilidad conjunta de Y_{ij} y Θ_i está dada por:

$$f_{(Y_{ij}, \Theta_i)}(y_{ij}, \theta_i) = f_{(\Theta_i)}(\theta_i) f_{(Y_{ij}|\Theta_i)}(y_{ij}|\theta_i)$$

Luego la función de distribución de Y_i será

$$\begin{aligned}
 f_{(Y_{ij})}(y_{ij}) &= \int_0^{\infty} f_{(Y_{ij}, \theta_i)}(y_{ij}, \theta_i) d\theta_i \\
 &= \int_0^{\infty} \frac{\frac{\alpha}{\mu}}{\Gamma(\alpha)} \left(\frac{\alpha}{\mu_i} \theta_i\right)^{\alpha-1} \exp\left(-\frac{\alpha}{\mu} \theta_i\right) \frac{\exp(-\theta_i) \theta_i^{y_{ij}}}{y_{ij}!} d\theta_i \\
 &= \frac{\left(\frac{\alpha}{\mu_i}\right)^{\alpha}}{\Gamma(\alpha) y_{ij}!} \int_0^{\infty} \theta_i^{(\alpha+y_{ij})-1} \exp(-\theta_i) \left(\frac{\alpha}{\mu_i} + 1\right) d\theta_i \\
 &= \frac{\left(\frac{\alpha}{\mu_i}\right)^{\alpha}}{\Gamma(\alpha) y_{ij}! \left(\frac{\alpha}{\mu_i} + 1\right)^{(\alpha+y_{ij})-1}} \int_0^{\infty} \left(\frac{\alpha}{\mu_i} + 1\right)^{(\alpha+y_{ij})-1} \theta_i^{(\alpha+y_{ij})-1} \exp\left\{-\theta_i \left(\frac{\alpha}{\mu_i} + 1\right)\right\} d\theta_i
 \end{aligned}$$

Sea $z = \left(\frac{\alpha}{\mu_i} + 1\right) \theta_i$ y sea $dz = \left(\frac{\alpha}{\mu_i} + 1\right) d\theta_i$, entonces

$$\begin{aligned}
 f_{Y_{ij}}(y_{ij}) &= \frac{\left(\frac{\alpha}{\mu_i}\right)^{\alpha}}{\Gamma(\alpha) y_{ij}! \left(\frac{\alpha}{\mu_i} + 1\right)^{(\alpha+y_{ij})}} \int_0^{\infty} z^{(\alpha+y_{ij})-1} \exp\{-z\} dz \\
 &= \frac{\left(\frac{\alpha}{\mu_i}\right)^{\alpha}}{\Gamma(\alpha) y_{ij}! \left(\frac{\alpha}{\mu_i} + 1\right)^{(\alpha+y_{ij})}} \Gamma(\alpha + y_{ij})
 \end{aligned}$$

por tanto

$$f_{Y_{ij}}(y_{ij}) = \frac{\Gamma(\alpha+y_{ij})}{\Gamma(\alpha) y_{ij}!} \left(\frac{\mu_i}{\mu_i+\alpha}\right)^{y_{ij}} \left(\frac{\alpha}{\mu_i+\alpha}\right)^{\alpha}$$

El valor esperado de la variable aleatoria Y_{ij} está dado por:

$$\begin{aligned}
 E(Y_{ij}) &= E[E(Y_{ij}|\Theta_i)] \\
 &= E(\Theta_i) \\
 &= \mu_i
 \end{aligned}$$

La varianza de la variable aleatoria Y_{ij} está dada por:

$$\begin{aligned}
 Var(Y_{ij}) &= E[Var(Y_{ij}|\Theta_i)] + Var[E(Y_{ij}|\Theta_i)] \\
 &= E(\Theta_i) + Var(\Theta_i) \\
 &= \mu_i + \frac{\mu_i^2}{\alpha} \\
 &= \mu_i \left(1 + \frac{\mu_i}{\alpha}\right)
 \end{aligned}$$

Por tanto para definir el modelo alternativo se asume $\Theta_i \sim \Gamma(\alpha, \lambda_i)$ con $\lambda_i = \frac{\mu_i}{\alpha}$ de donde la variable respuesta Y_{ij} con $i = 1, \dots, p$, $j = 1, \dots, r_i$ tiene distribución binomial negativa con media y varianza dadas por

$$E(Y_{ij}) = \mu_i$$

$$\text{Var}(Y_{ij}) = E[\text{Var}(Y_{ij}|\Theta_i)] + \text{Var}[E(Y_{ij}|\Theta_i)] = \mu_i \left(1 + \frac{\mu_i}{\alpha}\right)$$

De este modo para valores fijos de α la función de distribución (2-13) pertenece a la familia exponencial; es decir se puede escribir en la forma (2-2) de la siguiente manera

$$f_{(Y_{ij})}(y_{ij}) = \exp \left\{ y_{ij} \ln \left(\frac{\mu_i}{\mu_i + \alpha} \right) + \ln \left(\frac{\Gamma(\alpha + y_{ij})}{\Gamma(\alpha) y_{ij}!} \right) + \alpha \ln \left(\frac{\alpha}{\mu_i + \alpha} \right) \right\}$$

donde $\theta = \ln \left(\frac{\mu_i}{\mu_i + \alpha} \right)$, $b(\theta) = \alpha \ln \left(\frac{\alpha}{\mu_i + \alpha} \right)$ y $a(\phi) = 1$. Como α no es estrictamente un parámetro de dispersión en el sentido de la familia exponencial, no puede ser estimado mediante el estimador de pearson generalizado. Morales y López (2009) proponen el siguiente estimador basado en el método de los momentos:

$$\tilde{\alpha} = \frac{n\hat{\mu}_i}{\sum_{i=1}^n y_i^2 - n\hat{\mu}_i - n\hat{\mu}_i^2}$$

donde $\hat{\mu}_i$ se obtiene ajustando un modelo Poisson estándar. Las ecuaciones de score y el estadístico de desvío vienen dados respectivamente por:

$$U_j = \sum_{i=1}^n \left[\frac{y_{ij} - \mu_i}{\mu_i \left(1 + \frac{\mu_i}{\alpha}\right)} \right] x_{ij} \mu_i; j = 1, \dots, p$$

$$D = 2 \sum_{i=1}^n \left\{ y_{ij} \ln \left(\frac{y_{ij}}{\hat{\mu}_i} \right) - (y_{ij} - \tilde{\alpha}) \ln \left(\frac{1 + \hat{\mu}_i}{1 + y_{ij}} \right) \right\}$$

las cuales son útiles para establecer la bondad de ajuste del modelo y hacer inferencia.

2.5.2. Pruebas de sobredispersión

Según Hinde & Demetrio (2007) la obtención de una prueba de bondad de ajuste para los modelos de sobredispersión no es tan simple como en el caso de un modelo binomial o Poisson, donde el desvío residual o Pearson X^2 se puede utilizar como una prueba aproximada. La justificación de esto se debe al hecho que los parámetros adicionales de sobredispersión actúan con frecuencia como una forma de parámetro de escala.

Sin embargo, a menudo es posible poner a prueba el modelo de sobredispersión en comparación con el ajuste del modelo estándar. Para un modelo completamente especificado en dos etapas, la prueba de sobredispersión frecuentemente se reduce a pruebas de un único parámetro adicional. Por ejemplo, la comparación con los modelos binomial negativo y Poisson se puede pensar en una prueba en la familia Binomial Negativa que compara

En Lawless (1987) señala que dado que esto implica pruebas de un valor de parámetro en el limite del espacio de parámetros, la adecuada distribución asintótica de este estadístico bajo

la hipótesis nula es aquella que tiene masa de probabilidad de 0 a $\frac{1}{2}$ y una distribución $\frac{1}{2}X^2$ por encima de cero.

En Dean (1992) se obtienen pruebas de sobredispersión respecto a una familia exponencial natural, da expresiones cerradas de las estadísticas (score test) para las pruebas de sobredispersión constante y modelos en dos etapas (beta-binomial, los modelos binomial negativo y modelos con efectos aleatorios) en contra de los modelos binomial y regresión Poisson. En tanto que Lambert & K.Roder (1995) introducen un plot de convexidad, o C plot, que detecta las pruebas de sobredispersión y varianza relativa que ayudan a comprender la naturaleza de la sobredispersión. Además afirman que los plots de convexidad a veces detectan mejor la sobredispersión que las pruebas score test, y las pruebas de curvas de varianza relativa en la mayoría de los casos distinguen el origen de la sobredispersión mejor que las pruebas score tests.

2.5.3. Selección de covariables

Teniendo en cuenta el caso de sobredispersión constante, si el modelo ajustado es correcto se tiene que $E(X^2) \approx (n-p)\phi$ lo que sugiere que X^2 tiene una distribución ϕX_{n-p}^2 , donde X_{n-p}^2 denota una variable aleatoria chi-cuadrado con $n-p$ grados de libertad. Con el resultado de que el desvío suele tener distribución aproximadamente igual a X^2 , se espera que el desvío también tenga distribución ϕX^2 . Este resultado puede verse en McCullagh & Nelder (1989).

Una vez que el parámetro de sobredispersión se ha estimado a partir del ajuste del modelo completo, diferentes sub-modelos se pueden ajustar fijando sus valores y pesos de las observaciones formuladas por $w_i = \frac{1}{\phi_i}$. Luego los modelos (anidados) alternativos se pueden comparar en la forma habitual. Es decir, la diferencia en el desvío para los dos modelos se compara con el percentil de la estadística X^2 ; un resultado no significativo quiere decir que los dos modelos son estadísticamente independientes. Este resultado también se usa para otros modelos de sobredispersión.

2.5.4. Diagnósticos

En la evaluación del ajuste de los modelos de sobredispersión cuando el parámetro de sobredispersión se estima, el desvío residual o Pearson son relativamente cercanos a los grados de libertad. Puesto que teóricamente $Var(\mu_i) = \mu_i$ el índice de sobredispersión debería ser igual a 1. De este modo se tiene que un índice de sobredispersión mayor que 1 indica la posible existencia de sobredispersión, en tanto que, un índice de sobredispersión menor que 1 indica la posible existencia de subdispersión en los datos.

Los plot de residuales estándar pueden ser usados para explorar lo adecuado del predictor lineal, la función de enlace y la identificación de datos atípicos. Un plot de residuales provee una revisión informal de la especificación de la función de varianza $Var(\mu)$, sin embargo este puede no ser útil en la elección de los modelos de sobredispersión que involucran los parámetros de escala ϕ .

Atkinson (1985) sugiere una técnica útil para el examen general de los residuales por medio del uso de un plot half - norm, con simulación de envelope, que tiene en cuenta la sobredispersión del modelo. Si el modelo ajustado es correcto se espera que el plot de los valores observados caigan dentro de los límites de confianza del envelope. Para modelos lineales generalizados estos plot proporcionan una herramienta útil para comprobar las hipótesis del modelo, en lo referente al ajuste del mismo.

2.6. Modelo Multinomial

La distribución multinomial es en muchos sentidos la distribución natural a considerar cuando la variable respuesta es categórica, con más de dos categorías.

Considérese una variable aleatoria Y con J categorías de respuestas, si se denota por π_1, \dots, π_J sus respectivas probabilidades, con $\pi_1 + \dots + \pi_J = 1$, cuando se tienen n observaciones independientes de Y con resultado y_1 en la categoría 1, y_2 en la categoría 2 y así sucesivamente. Satisfaciendo que

$$\mathbf{y} = (y_1, \dots, y_J)^t \text{ con } \sum_{j=1}^J y_j = n$$

Con este arreglo la función de distribución multinomial es dada por:

$$f_{Y|n}(\mathbf{y}|n) = \frac{n!}{y_1! y_2! \dots y_J!} \pi_1^{y_1} \pi_2^{y_2} \dots \pi_J^{y_J} \quad (2-14)$$

Obsérvese que existe una relación entre la distribución multinomial y la binomial, ya que para el caso $J = 2$ la distribución multinomial coincide con la distribución binomial esto es:

$$\begin{aligned} \pi_1 + \pi_2 &= 1 \Rightarrow \pi_2 = 1 - \pi_1 \\ y_1 + y_2 &= n \Rightarrow y_2 = n - y_1 \end{aligned}$$

Remplazando los valores anteriores en (2-14) se obtiene:

$$\begin{aligned}
f_{Y|n}(y_1|n, y_2|n) &= P(y_1|n = y_1, y_2|n = y_2) \\
&= P(y_1|n = y_1) \\
&= f_{Y_1}(y_1|n)
\end{aligned}$$

donde Y_1 se distribuye binomial con parámetros n y π_1 .

Para el caso en que se tienen J categorías de respuestas es de interés la distribución marginal de Y_k . Para deducir la distribución marginal de Y_k , se parte del supuesto inicial que no se selecciona un elemento de la categoría y_k , esto significa que se selecciona un elemento del resto de categorías. Esto se hace con probabilidad $1 - \pi_k$, por lo que la distribución marginal de Y_k debería ser binomial con parámetros n y π_k , como se muestra a continuación

$$\begin{aligned}
f_{Y_k|n}(y_k|n) &= P(Y_k|n = y_k) \\
&= \sum_{i=1}^{y_1} \sum_{i=1}^{y_2} \dots \sum_{i=1}^{y_{k-1}} \sum_{i=1}^{y_{k+1}} \dots \sum_{i=1}^{y_J} P(Y_1/n = y_1, \dots, Y_J/n = y_J) \\
&= \sum_{i=1}^{y_1} \sum_{i=1}^{y_2} \dots \sum_{i=1}^{y_{k-1}} \sum_{i=1}^{y_{k+1}} \dots \sum_{i=1}^{y_J} \binom{n}{y_1, \dots, y_J} \pi_1^{y_1} \dots \pi_{k-1}^{y_{k-1}} \pi_k^{y_k} \pi_{k+1}^{y_{k+1}} \dots \pi_J^{y_J} \\
&= \pi_k^{y_k} \sum_{i=1}^{y_1} \sum_{i=1}^{y_2} \dots \sum_{i=1}^{y_{k-1}} \sum_{i=1}^{y_{k+1}} \dots \sum_{i=1}^{y_J} \frac{n!}{y_1! \dots y_{k-1}! y_k! y_{k+1}! \dots y_J!} \pi_1^{y_1} \pi_{k-1}^{y_{k-1}} \pi_{k+1}^{y_{k+1}} \dots \pi_J^{y_J} \\
&= \frac{n!}{(n - y_k)! y_k!} \pi_k^{y_k} \sum_{i=1}^{y_1} \sum_{i=2}^{y_2} \dots \sum_{i=1}^{y_{k-1}} \sum_{i=1}^{y_{k+1}} \dots \sum_{i=1}^{y_J} \frac{(n - y_k)!}{y_1! \dots y_J!} \pi_1^{y_1} \dots \pi_{k-1}^{y_{k-1}} \pi_{k+1}^{y_{k+1}} \dots \pi_J^{y_J} \\
&= \frac{n!}{(n - y_k)! y_k!} \pi_k^{y_k} (\pi_1 + \dots + \pi_{k-1} + \pi_{k+1} + \dots + \pi_J)^{n - y_k} \\
&= \frac{n!}{(n - y_k)! y_k!} \pi_k^{y_k} (1 - \pi_k)^{n - y_k} \\
&= \binom{n}{y_k} \pi_k^{y_k} (1 - \pi_k)^{n - y_k}
\end{aligned}$$

Por lo tanto la distribución marginal de Y_k tiene distribución binomial con parámetros n , $q_1 = \pi_k$, $q_2 = \sum_{i \neq k} \pi_k$. De este modo se tiene que $E(Y_j) = n\pi_j$ y $Var(Y_j) = n\pi_j(1 - \pi_j)$.

Además por resultado inmediatamente anterior $Y_i + Y_j \sim Binom(n, \pi_i + \pi_j)$ por lo que

$$\begin{aligned} Var(Y_i + Y_j) &= Var(Y_i) + Var(Y_j) + 2Cov(Y_i, Y_j) \\ n(\pi_i + \pi_j)(1 - (\pi_i + \pi_j)) &= n\pi_i(1 - \pi_i) + n\pi_j(1 - \pi_j) + 2Cov(\pi_i, \pi_j) \\ 2Cov(\pi_i, \pi_j) &= n(\pi_i + \pi_j) - n(\pi_i^2 + 2\pi_i\pi_j + \pi_j^2) - n\pi_i + n\pi_i^2 - n\pi_j + n\pi_j^2 \\ Cov(\pi_i, \pi_j) &= -n\pi_i\pi_j \end{aligned}$$

Así la covarianza de la distribución multinomial es $Cov(\pi_i, \pi_j) = -n\pi_i\pi_j$.

Un aspecto importante de la distribución multinomial (2-14) es que para $J = 2$ categorías la distribución coincide con la distribución Binomial la cual pertenece a la familia exponencial. Pero en general (2-14) no pertenece a la familia exponencial puesto que no se puede escribir en la forma (2-1) y (2-2).

De este modo y dentro del contexto de los modelos lineales generalizados la distribución multinomial es derivada a partir de variables aleatorias independientes con distribución Poisson, condicionada a la suma de estas variables aleatorias que se distribuye Poisson. Esto es, si Y_1, \dots, Y_J son variables aleatorias independientes Poisson con medias μ_1, \dots, μ_J , su función de distribución de probabilidad conjunta está dada por:

$$f(y_1, y_2, \dots, y_J) = \prod_{j=1}^J \frac{\mu_j^{y_j} e^{-\mu_j}}{y_j!}$$

sea $n = Y_1 + Y_2 + \dots + Y_J$ entonces la distribución condicional de Y dado n es

$$\begin{aligned} f_{Y|n}(y|n) &= \frac{f(y, n)}{f(n)} \\ &= \frac{\prod_{j=1}^J \frac{\mu_j^{y_j} e^{-\mu_j}}{y_j!}}{\frac{(\sum_{j=1}^J \mu_j)^{\sum_{j=1}^J y_j} e^{-\sum_{j=1}^J \mu_j}}{(\sum_{j=1}^J \mu_j)!}} \\ &= \frac{\prod_{j=1}^J \frac{\mu_j^{y_j} e^{-\mu_j}}{y_j!}}{\frac{(\sum_{j=1}^J \mu_j)^n e^{-(\mu_1 + \mu_2 + \dots + \mu_J)}}{n!}} \\ &= \frac{\mu_1^{y_1} \mu_2^{y_2} \dots \mu_J^{y_J}}{y_1! y_2! \dots y_J!} \\ &= \frac{(\sum_{j=1}^J \mu_j)^{y_1} (\sum_{j=1}^J \mu_j)^{y_2} \dots (\sum_{j=1}^J \mu_j)^{y_J}}{n!} \\ &= \frac{\mu_1^{y_1} \mu_2^{y_2} \dots \mu_J^{y_J}}{\left(\sum_{j=1}^J \mu_j\right)^{y_1} \left(\sum_{j=1}^J \mu_j\right)^{y_2} \dots \left(\sum_{j=1}^J \mu_j\right)^{y_J}} \cdot \frac{n!}{y_1! y_2! \dots y_J!} \\ &= \left(\frac{\mu_1}{\sum_{j=1}^J \mu_j}\right)^{y_1} \left(\frac{\mu_2}{\sum_{j=1}^J \mu_j}\right)^{y_2} \dots \left(\frac{\mu_J}{\sum_{j=1}^J \mu_j}\right)^{y_J} \cdot \frac{n!}{y_1! y_2! \dots y_J!} \end{aligned}$$

Luego la distribución multinomial puede ser observada como la distribución conjunta de variables aleatorias Poisson, condicionada a su suma n . La cual pertenece a la familia exponencial, de este modo se puede definir el modelo multinomial $E(Y_i) = n\theta_i$ donde $\sum_{i=1}^N \theta_i = 1$ y $\sum_{i=1}^N y_i = n$. Con este último resultado se ajustaron los datos al modelo multinomial.

2.6.1. Regresión Logística Nominal

Es usada cuando no existe un orden natural entre las respuestas categóricas. Una categoría arbitraria es escogida como variable categórica de referencia. Supongase que se escoge la primera categoría. Entonces la logística para las otras categorías son definidas por:

$$\text{logit}(\pi_j) = \log\left(\frac{\pi_j}{\pi_1}\right) = x_j^t \beta_j \quad \text{para } j = 2, 3, \dots, J \quad (2-15)$$

Las $(J - 1)$ ecuaciones logísticas son usadas simultáneamente para estimar los parámetros β_j . Una vez estimados los parámetros β_j , el predictor lineal $x_j^t \beta_j$ puede ser calculado. De (2-15) $\hat{\pi}_j = \hat{\pi}_1 \exp(x_j^t \hat{\beta}_j)$, para $j = 2, 3, \dots, J$. Pero $\hat{\pi}_1 + \hat{\pi}_2 \dots + \hat{\pi}_J = 1$ de modo que $\hat{\pi}_1 = \frac{1}{1 + \sum_{j=2}^J \exp(x_j^t \hat{\beta}_j)}$ y $\hat{\pi}_j = \frac{\exp(x_j^t \hat{\beta}_j)}{1 + \sum_{j=2}^J \exp(x_j^t \hat{\beta}_j)}$ para $j = 2, 3, \dots, J$.

Valores ajustados, o frecuencias esperadas, para cada covariable pueden ser calculados multiplicando las probabilidades estimadas por el total de la frecuencia de la covariable.

2.6.2. Regresión Logística Ordinal

En muchas situaciones, las categorías de una variable respuesta tienen alguna clase de ordenamiento. De acuerdo con Dobson (2002), los modelos que se usan comúnmente cuando se tienen categorías ordinales son el modelo logit acumulativo, el modelo de odds proporcionales, modelos logit categóricos adyacentes y modelos logit de continuación de razón.

Modelo Logit

El *odds* acumulativo para la j -ésima categoría es $\frac{P(Y \leq c_j)}{P(Y > c_j)} = \frac{\pi_1 + \pi_2 + \dots + \pi_j}{\pi_{j+1} + \dots + \pi_J}$. El modelo logit acumulativo es

$$\log\left(\frac{\pi_1 + \pi_2 + \dots + \pi_j}{\pi_{j+1} + \dots + \pi_J}\right) = x_j^t \beta_j$$

Modelo Odds Proporcional

Si el predictor lineal en (2-15) tiene como intercepto el término β_{0j} el cual depende de la categoría j , pero las otras variables explicativas no dependen de j , entonces el modelo esta

dado por

$$\log \left(\frac{\pi_1 + \pi_2 + \dots + \pi_j}{\pi_{j+1} + \dots + \pi_J} \right) = \beta_{0j} + \beta_1 x_1 + \dots + \beta_{p-1} x_{p-1}$$

El modelo de *odds* proporcional se basa en el supuesto que el efecto de las covariables $x_1 + x_2 \dots + x_p$ es igual para todas las categorías en la escala logarítmica.

Modelo Logit De Categorías Adyacentes

El modelo logit de categorías adyacentes está dado por $\log \left(\frac{\pi_j}{\pi_{j+1}} \right) = x_j^t \beta_j$. Si este es simplificado queda expresado como

$$\log \left(\frac{\pi_j}{\pi_{j+1}} \right) = \beta_{0j} + \beta_1 x_1 + \dots + \beta_{p-1} x_{p-1}$$

el efecto de cada variable explicativa se asume la misma para todos los pares adyacentes de categorías.

Modelo Logit De Continuación De Razón

Otra alternativa es el modelo de razón de probabilidad

$$\frac{\pi_1}{\pi_2}, \frac{\pi_1 + \pi_2}{\pi_3}, \dots, \frac{\pi_1 + \dots + \pi_{j-1}}{\pi_j}$$

o

$$\frac{\pi_1}{\pi_2 + \dots + \pi_j}, \frac{\pi_2}{\pi_3 + \dots + \pi_j}, \dots, \frac{\pi_{j-1}}{\pi_j}$$

La ecuación $\log \left(\frac{\pi_j}{\pi_{j+1} + \dots + \pi_j} \right) = x_j^t \beta_j$ Modela el *odds* de la respuesta que está en la categoría j esto es $C_{j-1} < Y \leq C_j$ condicional sobre $Y \geq C_{j-1}$.

2.7. Método GSK

La metodología GSK desarrollada por Grizzle et al. (1969), es empleada para trabajar modelos lineales con datos categóricos, en particular para modelos cuya variable respuesta tiene distribución multinomial. Una extensión del modelo GSK propuesta por Bonett, Woodward y Bentler (1985), considera modelos lineales cuya variable respuesta siguen una distribución condicional Poisson (con distribución multinomial y producto multinomial como caso especial). La metodología GSK consiste en que dado un conjunto de datos, dispuestos en un arreglo como el que se muestra en la tabla (2-1)

Categorías de Respuesta					
Tratamiento	1	2	...	r	Total
1	n_{11}	n_{12}	...	n_{1r}	$n_{1.}$
2	n_{21}	n_{22}	...	n_{2r}	$n_{2.}$
⋮	⋮	⋮	...	⋮	⋮
s	n_{s1}	n_{s2}	...	n_{sr}	$n_{s.}$

Tabla 2-1.: Frecuencias observadas asociada a la estructura multinomial.

donde los n_{ij} , $i = 1, \dots, s$; $j = 1, \dots, r$ son frecuencias observadas asociadas a las celdas de la tabla anterior. Si π_{ij} corresponde a la probabilidad que una observación pertenezca a la población i de la categoría j , entonces el vector definido por $\pi_i^t = [\pi_{i1}, \pi_{i2}, \dots, \pi_{ir}]$ proporciona las probabilidades asociadas a las r categorías de respuesta, como se muestra en la tabla de probabilidades (2-2).

Categorías de Respuesta					
Tratamiento	1	2	...	r	Total
1	π_{11}	π_{12}	...	π_{1r}	1
2	π_{21}	π_{22}	...	π_{2r}	1
⋮	⋮	⋮	...	⋮	⋮
s	π_{s1}	π_{s2}	...	π_{sr}	1

Tabla 2-2.: Probabilidades asociadas a cada observación de la estructura multinomial.

Si $P_{ij} = \frac{n_{ij}}{n_{i.}}$ es el estimador de máxima verosimilitud de π_{ij} , el vector de estimadores de π_i es de la forma $P_i^t = [P_{i1}, P_{i2}, \dots, P_{ir}]$, Johnson (1969). La varianza estimada asociada a una población específica en una repetición del experimento está dada por:

$$V_{r \times r}(P_i) = \frac{1}{n_i} (D_{P_i} - P_i P_i^t)$$

Con D_{P_i} una matriz diagonal con elementos P_i . La matriz de varianzas y covarianzas estimada de $V(\pi)$ para los s tratamientos es diagonal por bloques de la forma:

$$V_{rs \times rs}(P) = \begin{bmatrix} V(P_1) & & & 0 \\ & V(P_2) & & \\ & & \ddots & \\ 0 & & & V(P_s) \end{bmatrix}$$

El ajuste de datos categóricos al modelo se hace por el método de mínimos cuadrados generalizados, en donde la variable dependiente es una función lineal de las estimaciones de la probabilidad π_{ij} . El modelo propuesto es $Y = X\theta + e$, con

$$\theta^t = [\mu, \tau_1, \tau_2, \dots, \tau_{s-1}; \beta_1, \beta_2, \dots, \beta_{k-1}]$$

$$\tau_s = - \sum_{i=1}^{s-1} \tau_i$$

$$\beta_k = - \sum_{j=1}^{s-1} \beta_j$$

$$Y = F(\pi) = A\pi$$

A : matriz de orden (n, w) , en donde n es el número de observaciones.

$w = r \times s \times k$ siendo k el número de repeticiones.

X : la matriz de diseño reparametrizada de orden (n, π) asociada a una estructura de bloques al azar.

θ : es el vector de parámetros.

e : el error aleatorio no observable, el cual tiene distribución normal con media 0 y matriz de varianzas y covarianzas $AV(\pi)A^t$.

El resultado de AP , con P estimador de máxima verosimilitud de π es el porcentaje promedio ponderado sobre las r categorías. Las categorías se asignan en orden ascendente de acuerdo al número de categorías estudiadas. La estructura de A con r categorías de respuesta es:

$$A = \begin{bmatrix} 1 & 2 & 3 & \dots & r & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 & 2 & 3 & \dots & r & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & & & & & \dots & & & & & \dots & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 1 & 2 & 3 & \dots & r \end{bmatrix}$$

Como el interés del investigador es estimar los parámetros en el modelo y probar hipótesis acerca de la diferencia entre tratamientos, se tiene que el estimador de mínimos cuadrados generalizado para θ es: $\hat{\theta} = (X^t S^{-1} X)^{-1} X S^{-1} y$ con $S = AV(P)A^t$ matriz de varianzas y covarianzas estimadas $A\pi$. La hipótesis lineal general a probar es $H_0 : C\theta = 0$ en donde C es una matriz cuya estructura es de la forma:

$$C = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & & & \\ 0 & 0 & 0 & \cdots & 1 & 0 & \cdots & 0 \end{bmatrix}$$

La naturaleza de C , depende de la hipótesis que se plantea con respecto a contrastes entre tratamientos. La suma de cuadrados debida a la hipótesis $H_0 : C\theta = 0$ es:

$$SC(C\hat{\theta}) = (C\hat{\theta})^t [C(X^t S^{-1} X)^{-1} C^t]^{-1} (C\hat{\theta})$$

bajo $H_0 : SC(C\hat{\theta}) \sim \chi_{r(C)}^2$ en donde: $r(C)$ es el rango de C ; si $SC(C\hat{\theta}) > \chi_{r(C)}^2$ se rechaza H_0 . La suma de cuadrados debido al error, se define como:

$$SC(error) = Y^t [S^{-1} - \hat{\theta}^t (X^t S^{-1} \hat{\theta})^{-1}] Y$$

Bajo la H_0 : los datos se ajustan al modelo; se tiene que:

$$SC(error) \sim \chi_{(n-r(C))}^2$$

si

$$SC(error) > \chi_{(n-r(C))(\alpha)}^2$$

se rechaza la hipótesis nula.

Esta metodología fue la implementada por López & Chavez (1984) en el ensayo experimental considerado en este trabajo.

3. Aplicación de los modelos

Los modelos de regresión Poisson, binomial negativo y multinomial, se aplicaron al conjunto de datos presentados en la tabla(3-1). Estos datos fueron obtenidos a partir de un experimento realizado con el fin de evaluar el grado de pudrición de mazorcas en la variedad de maíz sogamoseño al ataque del hongo Fusarium, S.P. En el experimento se consideran 9 tratamientos correspondientes a las variedades de maíz, en un diseño en bloques completos, teniendo como variable respuesta el grado de pudrición y los bloques que representan las réplicas de los tratamientos.

Tratamientos									
b1	t1	t2	t3	t4	t5	t6	t7	t8	t9
GO	17	14	14	26	38	13	52	60	65
G1	20	20	29	34	35	47	48	45	37
G2	6	8	17	7	7	14	12	12	7
b2	t1	t2	t3	t4	t5	t6	t7	t8	t9
GO	9	16	19	18	27	21	67	65	63
G1	24	30	35	42	41	34	34	66	33
G2	6	3	2	9	8	9	4	6	5
b3	t1	t2	t3	t4	t5	t6	t7	t8	t9
GO	11	17	22	23	25	19	67	58	42
G1	21	40	30	32	55	54	34	27	51
G2	5	2	5	5	8	2	4	4	7
b4	t1	t2	t3	t4	t5	t6	t7	t8	t9
GO	12	22	9	12	47	15	42	53	42
G1	28	27	42	40	37	43	40	44	50
G2	5	0	4	6	2	3	8	6	7
b5	t1	t2	t3	t4	t5	t6	t7	t8	t9
GO	13	11	5	23	17	23	41	45	47
G1	18	0	29	26	46	28	68	47	40
G2	9	6	12	9	11	6	7	8	11
b6	t1	t2	t3	t4	t5	t6	t7	t8	t9
GO	15	5	4	11	14	55	48	45	5
G1	38	30	31	36	42	25	49	57	14
G2	5	9	2	7	5	5	7	2	3

Tabla 3-1.: Número de mazorcas clasificadas en cada grado de pudrición

Los tratamientos descritos en este ensayo fueron: Sogamoseño V.O variedad original t1 (**tratt1**), Ciclo I Sogamoseño (M.P.) I, selección masal por prolificidad t2 (**tratt2**), Ciclo II Sogamoseño (M.P.) II, selección masal por prolificidad t3 (**tratt3**), Ciclo III Sogamoseño (M.P.) III, selección masal por prolificidad t4 (**tratt4**), Cruzamiento de Sogamoseño V.O con MB510 (M.P.) VIII, material no prolífico t5 (**tratt5**), Cruzamiento de Sogamoseño V.O con MB513 (M.N.P) VIII, material no prolífico t6 (**tratt6**), MB510 (M.P) VIII (testigo) t7

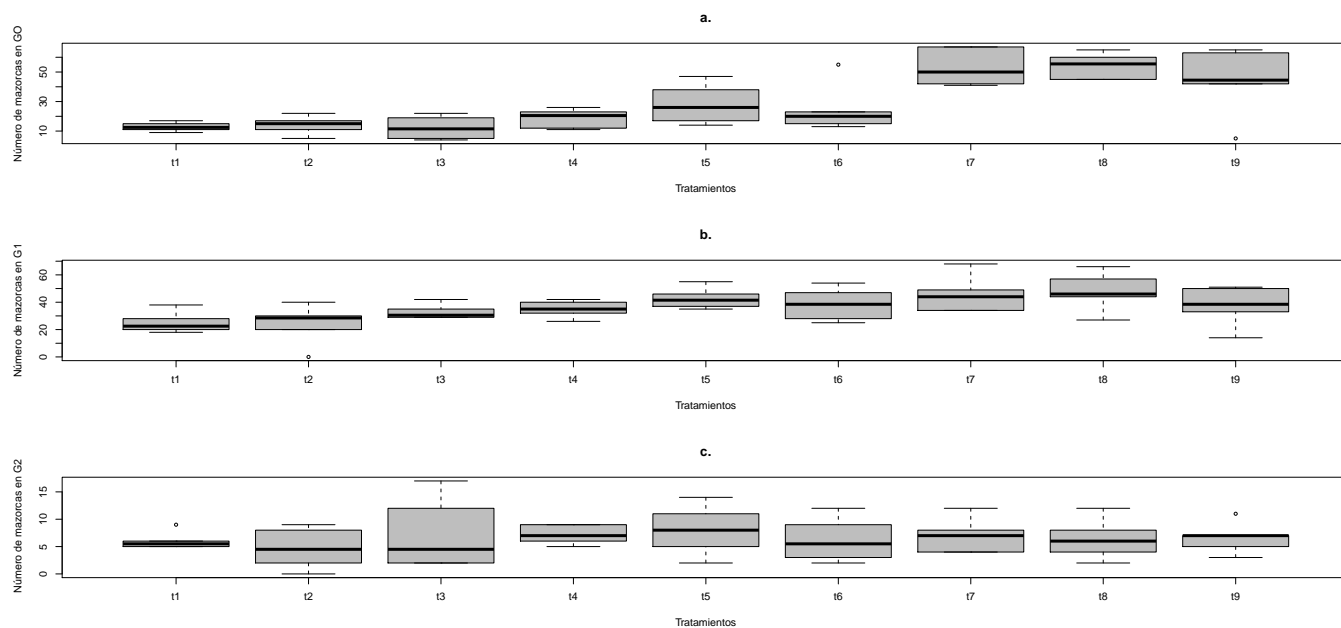


Figura 3-1.: Gráfico de box-plot para los diferentes grados de pudrición por tratamientos

(**tratt7**), MB513 (M.N .P) VIII (testigo) t8 (**tratt8**), ICAV506 y variedad comercial (testigo) t9 (**tratt9**).

El grado de pudrición (**cont**) es la variable respuesta y contiene tres categorías las cuales son grado de pudrición (**G0**)(mazorcas clasificadas en el grado de pudrición cero (mazorcas aparentemente sanas), grado de pudrición (**G1**)(mazorcas clasificadas en el grado de pudrición uno (mazorcas con un porcentaje de 1 %-25 % de tejido enfermo) y grado de pudrición (**G2**)(mazorcas clasificadas en el grado de pudrición dos(mazorcas con un porcentaje de 26 %-50 % de tejido enfermo).

3.1. Análisis exploratorio

En esta sección se realizara la descripción del comportamiento del número de mazorcas clasificadas en cada grado de pudrición por tratamientos. Los resultados se presentarán a través de gráficas de cajas.

En la figura (3-1 a.) parece formarse 3 grupos de tratamientos: los de material de selección masal y prolifíco (t1 a t5) el de material no prolifíco t6 y los testigos (t7 a t9). En el primer grupo la media y la variabilidad son similares, el segundo grupo tiene diferente variabilidad pero medias parecidas y el tercer grupo tiene variabilidad y media muy parecidas pero mayor

que los demás grupos. En conclusión el grupo que presenta mayor media y variabilidad con respecto a los demás es el grupo de variedades de maíz testigo.

La figura (3-1 b.) muestra un leve incremento en la media del número de mazorcas de un tratamiento a otro. Además los tratamientos t6 y t8 presentan mayor variabilidad con respecto a los demás.

En la figura (3-1 c.) se observa que la media del número de mazorcas clasificadas en el grado de pudrición G2 es bastante similar, sin embargo los tratamiento t2, t3 y t5 presentan mayor variabilidad que los demás y los tratamientos t1 y t9 presentan datos atípicos. El tratamiento t1 es el que menor variabilidad presenta con respecto a los demás.

3.2. Ajuste de los modelos

En esta sección se procede a ajustar los diferentes modelos propuestos en este trabajo, para evaluar el efecto de tratamiento según el grado de pudrición de las mazorcas. Se realizó un análisis por separado para cada uno de los grados de pudrición y en cada escenario se evaluó el efecto de la sobredispersión a través de el índice de sobredispersión y mediante el envelope simulado.

Los datos se ajustaron mediante el modelo de Poisson, binomial negativo y el modelo multinomial mediante el procedimiento GSK.

3.2.1. Ajuste del modelo de regresión Poisson

En esta sección se ajustan, el modelo de regresión Poisson con respecto al grado de pudrición GO, G1 y G2 . La caracterización de cada modelo seda por:

Componente aleatoria

Sea Y_{ij} la variable aleatoria correspondiente al número de mazorcas clasificadas en cada grado de pudrición GO, G1 y G2 donde $i = 1, \dots, 9$ hace referencia a los tratamiento y $j = 1, \dots, 6$ a los bloques. Se asume que la distribución de Y_{ij} es Poisson con parámetro μ .

Componente sistemática

El predictor lineal $\eta = X^t\beta$ donde X es la matriz de diseño (igual para cada grado de pudrición) y $\beta = (\beta_0, \beta_1, \dots, \beta_9, \beta_{10}, \dots, \beta_{15})^t$ es el vector de parámetros que toma valores de acuerdo a cada grado de pudrición, para el modelo lineal de generalizado tiene la forma

$$\eta_{ij} = \beta_0 + \beta_1 tratt1 + \dots + \beta_9 tratt9 + \beta_{10} bloquesb1 + \dots + \beta_{15} bloquesb6 \quad (3-1)$$

Función de enlace

En cada modelo se consideraron diversas funciones de enlaces, para los grados de pudrición GO y G2 se tomaron las funciones de enlace logaritmo natural e inversa y para el grado de pudrición G1 las funciones de enlace logaritmo natural y raíz cuadrada . Esto con el fin de determinar cuál es el modelo que mejor se ajusta a los datos.

Ajuste del modelo de regresión Poisson con respecto al grado de pudrición GO

Variable	MRP(link=log)			MRP(link=inversa)		
	Estimación	E.Standar	p.valor	Estimación	E.Standar	p.valor
Intercep	2.686	0.125	$< 2e^{-16}$	0.075	0.009	$i < 2e^{-16} ***$
bloquesb2	0.019	0.081	0.807	-0.001	0.002	0.491
bloquesb3	-0.051	0.083	0.534	0.001	0.002	0.567
bloquesb4	-0.163	0.085	0.056 .	0.004	0.002	0.053 .
bloquesb5	-0.284	0.088	0.001 **	0.007	0.002	0.004 **
bloquesb6	-0.392	0.091	$1.66e^{-05} ***$	0.011	0.003	$6.50e^{-05} ***$
tratt2	0.099	0.157	0.529	-0.009	0.012	0.464
tratt3	-0.053	0.163	0.744	0.002	0.013	0.876
tratt4	0.384	0.148	0.009 **	-0.026	0.010	0.012 *
tratt5	0.780	0.138	$1.44e^{-08} ***$	-0.043	0.009	$4.36e^{-06} ***$
tratt6	0.639	0.141	$5.55e^{-06} ***$	-0.035	0.009	0.000 ***
tratt7	1.415	0.127	$< 2e^{-16} ***$	-0.059	0.009	$6.45e^{-11} ***$
tratt8	1.443	0.127	$< 2e^{-16} ***$	-0.059	0.009	$4.51e^{-11} ***$
tratt9	1.232	0.129	$< 2e^{-16} ***$	-0.056	0.009	$5.50e^{-10} ***$
P. Dispersión	1			1		
Desvio nulo-gl	680.1 - 53			680.11- 53		
Desvio residual-gl	176.9 - 40			174.49 -40		
Índice de disppersión	4.43			4.36		
AIC	473.1			470.7		
Fisher Scoring	5			8		

Tabla 3-2.: Regresión Poisson asociado al grados de pudrición GO tomando como funciones de enlace logaritmo e inversa.

De acuerdo con los resultados mostrados en la tabla (3-2) al ajustar los datos al modelo de regresión Poisson respecto al número de mazorcas clasificadas en el grado de pudrición GO se obtuvo con relación a la función de enlace logaritmo natural un desvio residual de 176,92 con 40 grados de libertad, lo que advierte una posible sobredispersión en los datos con un índice de 4,43, esto sugiere falta de ajuste del modelo. El envelope simulado en (3-2 a y 3-2 b) confirma este resultado.

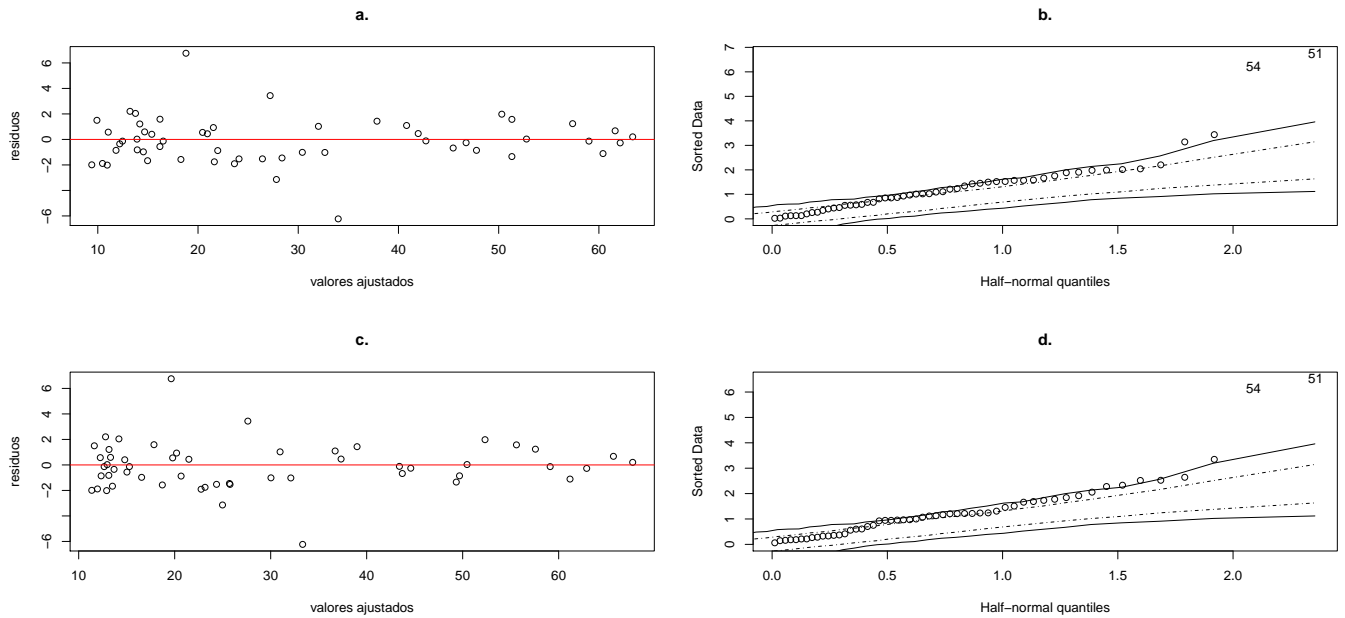


Figura 3-2.: Envelopes para el modelo de regresión Poisson tomando como función de enlace logaritmo e inversa.

Con la función de enlace inversa se obtuvo un desvío residual de 174,49 con 40 grados de libertad, e índice de dispersión de 4,36, lo que indica la presencia de sobredispersión en los datos, como se puede confirmar en el envelope simulado en (3-2 c y 3-2 d). Adicionalmente se observa que el AIC más bajo lo registra el MRP con la función de enlace inversa.

Ajuste del modelo de regresión Poisson con respecto al grado de pudrición G1

Variable	MRP(link=log)			MRP(link=raíz cuadrada)		
	Estimación	E.Standar	p.valor	Estimación	E.Standar	p.valor
Intercep	3.16921	0.09685	$< 2e^{-16}$ ***	4.83631	0.25459	$< 2e^{-16}$ ***
bloquesb2	0.07343	0.07826	0.34811	0.24459	0.23570	0.29940
bloquesb3	0.08807	0.07798	0.25877	0.32684	0.23570	0.16555
bloquesb4	0.10821	0.07761	0.16323	0.37356	0.23570	0.11300
bloquesb5	-0.04215	0.08053	0.60075	-0.20753	0.23570	0.37860
bloquesb6	0.02198	0.07925	0.78152	0.11259	0.23570	0.63289
tratt2	-0.01351	0.11625	0.90746	-0.07077	0.28868	0.80634
tratt3	0.27417	0.10869	0.01165 *	0.73144	0.28868	0.01128 *
tratt4	0.34316	0.10711	0.00136 **	0.93027	0.28868	0.00127 **
tratt5	0.54123	0.10304	$1.50e^{-07}$ ***	1.55580	0.28868	$7.07e^{-08}$ ***
tratt6	0.43847	0.10507	$3.01e^{-05}$ ***	1.21632	0.28868	$2.51e^{-05}$ ***
tratt7	0.60553	0.10185	$76e^{-09}$ ***	1.79446	0.28868	$09e^{-10}$ ***
tratt8	0.65205	0.10103	$1.09e^{-10}$ ***	1.93315	0.28868	$2.13e^{-11}$ ***
tratt9	0.41215	0.10562	$9.53e^{-05}$ ***	1.13887	0.28868	$97e^{-05}$ ***
P. Dispersion	1			1		
Desvio nulo-gl	259.33 - 53			259.33 - 53		
Desvio residual-gl	161.22 - 40			158.16 - 40		
Indice de disppersion	4.03			3.95		
AIC	475.98			472.92		
Fisher Scoring	5			6		

Tabla 3-3.: Regresión Poisson asociado al grados de pudrición G1 tomando como funciones de enlace logaritmo y raíz cuadrada.

Los resultados del ajuste de los datos al modelo de regresión Poisson respecto al grado de pudrición G1 en la tabla (3-3) muestran que con relación a la función de enlace logaritmo natural se obtuvo un desvio residual de 161,22 con 40 grados de libertad, con índice de dispersión de 4,03 indicando la presencia de sobredispersión en los datos, lo cual muestra falta de ajuste en el modelo. El envelope simulado en (3-3 a y b) confirma este resultado.

El envelope simulado en la figura (3-3 c. y 3-3 d.) y el índice de dispersión 3,95 para el modelo Poisson de este grado de pudrición con función de enlace raíz cuadrada confirman la presencia de sobredispersión indicando la falta de ajuste del modelo. Además el AIC de este modelo es el más bajo entre todas las funciones de enlace tomadas.

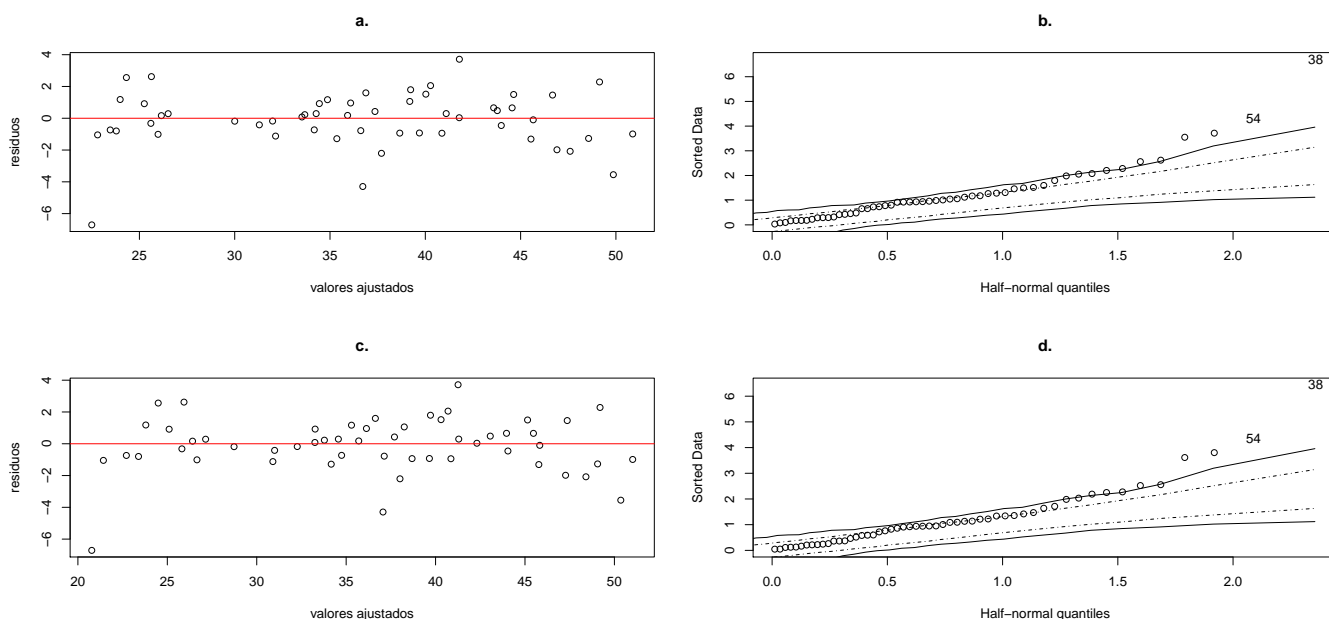


Figura 3-3.: Envelopes para el modelo de regresión Poisson tomando como función de enlace logaritmo y raíz cuadrada.

Ajuste del modelo de regresión Poisson con respecto al grado de pudrición G2

Variable	MRP(link=log)			MRP(link=inversa)		
	Estimación	E.Standar	p.valor	Estimación	E.Standar	p.valor
Intercep	2.26810	0.18836	$< 2e^{-16}$ ***	0.11414	0.02701	$2,37e^{-05}$ ***
bloquesb2	-0.60263	0.17250	0.000477 ***	0.07876	0.02551	0.002020 **
bloquesb3	-0.81621	0.18530	$1,06e^{-05}$ ***	0.11545	0.03328	0.000522 ***
bloquesb4	-0.84030	0.18686	$6,89e^{-06}$ ***	0.12267	0.03491	0.000442 ***
bloquesb5	-0.18443	0.15226	0.225804	0.01964	0.01529	0.199185
bloquesb6	-0.74721	0.18097	$3.64e^{-05}$ ***	0.11091	0.03227	0.000588 ***
tratt2	-0.25131	0.25198	0.318583	0.03153	0.04391	0.472721
tratt3	0.15415	0.22713	0.497332	-0.04162	0.02995	0.164659
tratt4	0.17768	0.22591	0.431559	-0.01993	0.03312	0.547315
tratt5	0.28768	0.22048	0.191960	-0.04362	0.02971	0.142041
tratt6	0.02740	0.23410	0.906830	-0.01494	0.03398	0.660304
tratt7	0.15415	0.22713	0.497332	-0.02431	0.03240	0.453052
tratt8	0.05407	0.23258	0.816175	-0.01898	0.03328	0.568587
tratt9	0.10536	0.22973	0.646508	-0.01539	0.03390	0.649904
P. Dispersion	1			1		
Desvio nulo-gl	97.859 - 53			97.859 - 53		
Desvio residual-gl	50.549 - 40			49.509 - 40		
Indice de dispersion	1.263			1.237		
AIC	271.35			270.31		
Fisher Scoring	5			7		

Tabla 3-4.: Regresión Poisson asociado al grados de pudrición G2 tomando como funciones de enlace logaritmo e inversa.

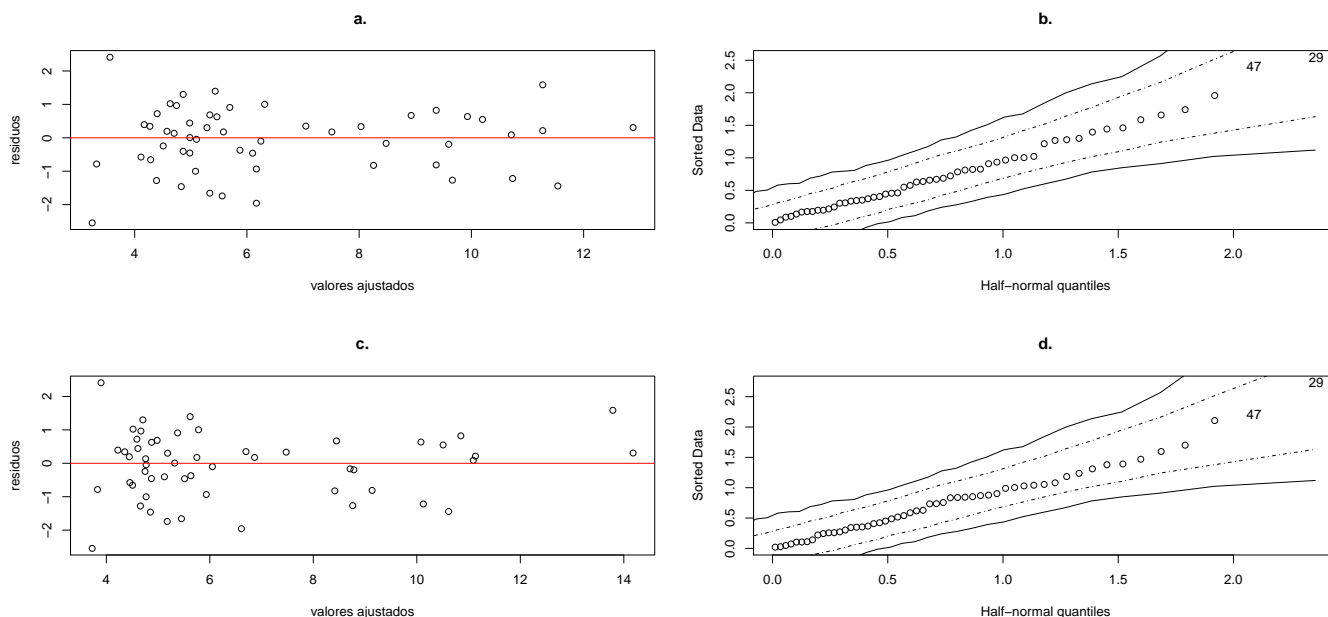


Figura 3-4.: Envelopes para el modelo de regresión Poisson tomando como función de enlace logaritmo e inversa.

De acuerdo con los resultados mostrados en la tabla (3-4) al ajustar los datos al modelo de regresión Poisson respecto al número de mazorcas clasificadas en el grado de pudrición G2 se obtuvo con relación a la función de enlace logaritmo natural un desvío residual de 50,549 con 40 grados de libertad e índice de dispersión de 1,263, lo que advierte una posible sobredispersión en los datos y por ende una posible falta de ajuste del modelo. Pero el envelope simulado en (3-4 a. y 3-4 b.) sugiere un buen ajuste del modelo a los datos, esto se puede deber al hecho que el índice de dispersión es bastante cercano a 1 y por tanto el supuesto de igualdad de la varianza y la media se está cumpliendo.

Con la función de enlace inversa se obtuvo un desvío residual de 49,509 con 40 grados de libertad con índice de dispersión de 1,237 lo que advierte también posible sobredispersión, y esto puede conducir también a una posible falta de ajuste del modelo. Pero el envelope simulado en la figura (3-4 c. y 3-4 d.) indica que el modelo se ajusta bien a los datos esto es debido a que se tiene un índice de dispersión cercano a uno de igual forma que el caso anterior se puede estar cumpliendo el supuesto de equidispersión de los datos Poisson. Adicionalmente se observa en este caso que el AIC más bajo lo registra el MRP con la función de enlace logaritmo natural.

3.2.2. Ajuste del modelo Binomial Negativo

En esta sección se ajusto el modelo binomial negativo a los datos respecto a los grados de pudrición GO, G1 y G2 como modelo alternativo para corregir la sobredispersión presente en el modelo de regresión Poisson. Para efectos del ajuste de los modelos se toma el mismo predictor lineal dado en (3-1) y la función de enlace logaritmo natural.

Ajuste del modelo Binomial Negativo con respecto al grado de pudrición GO

BN(link=log)			
Variable	Estimación	E.Standar	p.valor
Intercep	2.70585	0.20241	$1.2e^{-16}$ ***
bloquesb2	-0.01503	0.17220	0.930458
bloquesb3	-0.03913	0.17249	0.820541
bloquesb4	0.15059	0.17387	0.386431
bloquesb5	-0.29367	0.17585	0.094918 .
bloquesb6	-0.39075	0.17734	0.027566 *
tratt2	0.07281	0.23834	0.760003
tratt3	-0.09743	0.24295	0.688405
tratt4	0.36534	0.23189	0.115134
tratt5	0.75256	0.22561	0.000851 ***
tratt6	0.68699	0.22652	0.002423 **
tratt7	1.40027	0.21908	$1.64e^{-10}$ ***
tratt8	1.42837	0.21887	$6.75e^{-11}$ ***
tratt9	1.18710	0.22080	$7.60e^{-08}$ ***
P. Dispersion	10.4641		
E.Standar	3.04		
Desvio nulo-gl	191.95-53		
Desvio residual-gl	57.09 -40		
AIC	421.26		
Fisher Scoring	1		
Test de dispersión	Valor Critico alpha= 0.05	level: 2.7055 Chi-Square Test Statistic = 53.8657	p.valor = $1.073e^{-13}$

Tabla 3-5.: Modelo Binomial Negativo asociado al grados de pudrición GO

De acuerdo con los resultados mostrados en la tabla 3-5 al ajustar los datos al modelo binomial negativo con respecto al número de mazorcas clasificadas en el grado de pudrición GO tomando como función de enlace logaritmo natural se obtuvo un desvio residual de 50,09 con 40 grados de libertad asociados y un parámetro de dispersión de 10,46. Al observar el envelope simulado en (3-5 b.) este nos indica que es mejor el ajuste de los datos con el modelo binomial negativo ya que el plot de los valores observados caen dentro de los limites

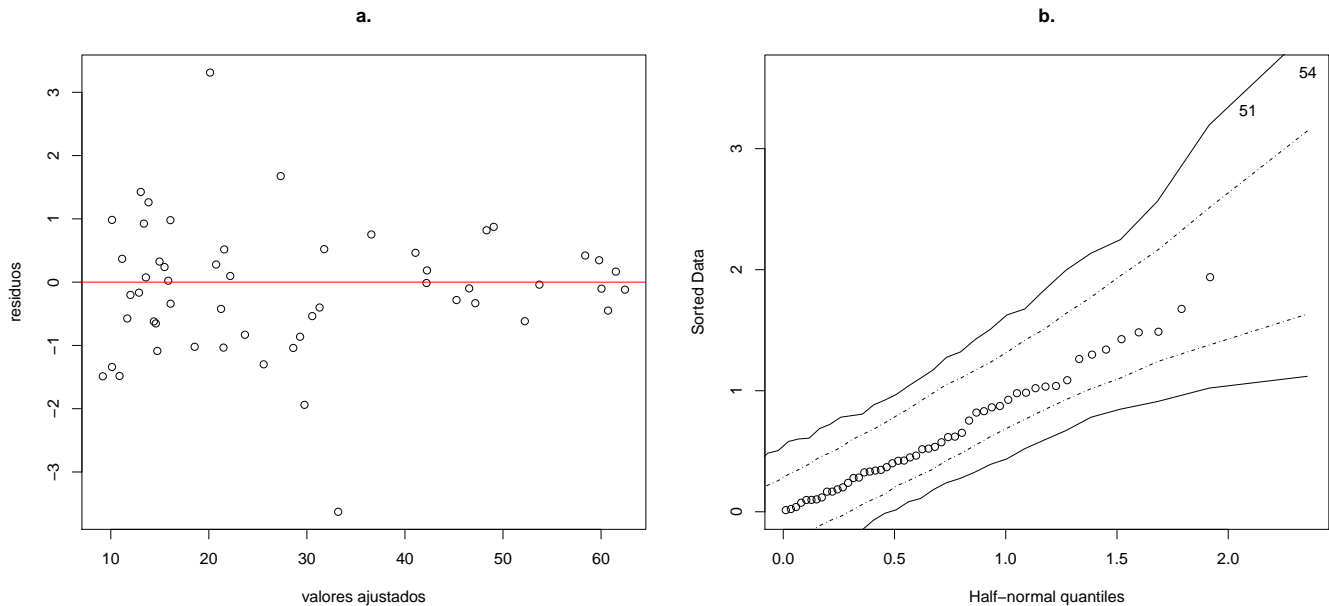


Figura 3-5.: a. Gráfico de residuales versus valores ajustado del modelo binomial negativo
 b. Envelope para el modelo binomial negativo asociado al grado de pudrición GO.

de confianza del mismo.

Además, al contrastar el modelo de regresión Poisson con el modelo binomial negativo, se observó que el parámetro de dispersión es estadísticamente mayor a uno, ya que su p -valor fue de $1,073e^{-13}$, por lo que se rechaza el modelo de regresión Poisson a favor del modelo alternativo binomial negativo. Por tanto el modelo adecuado para el ajuste de los datos es el modelo binomial negativo el cual arroja como resultado que los tratamientos testigo (t7 a t9), son significativos al 0,1% como también los tratamientos no prolífico t5 y t6 al 0,1% y al 1%, lo que muestra evidencia estadística que estas variedades de maíz contribuyen en la explicación de la variable respuesta número de mazorcas clasificadas en el grado de pudrición GO. En resumen las variedades de maíz testigo (t7 a t9), los tratamientos no prolíficos (t5 a t6) son las más resistentes al grado de pudrición al ataque del hongo *Fusarium*, S.P teniendo en cuenta la clasificación GO. Las variedades de maíz de selección masal por prolificidad son las menos resistentes.

Ajuste del modelo Binomial Negativo con respecto al grado de pudrición G1

BN(link=log)			
Variable	Estimacion	E.Standar	p.valor
Intercep	3.16004	0.15242	1.2×10^{-16} ***
bloquesb2	0.08376	0.13451	0.533489
bloquesb3	0.11550	0.13416	0.389284
bloquesb4	0.12880	0.13401	0.336496
bloquesb5	-0.07675	0.13644	0.573732
bloquesb6	0.04130	0.13499	0.759644
tratt2	-0.02789	0.17687	0.874698
tratt3	0.27399	0.17169	0.110536
tratt4	0.34222	0.17071	0.044990 *
tratt5	0.54486	0.16812	0.001191 **
tratt6	0.43616	0.16945	0.010053 *
tratt7	0.62190	0.16726	0.000201 ***
tratt8	0.65866	0.16687	7.9×10^{-5} ***
tratt9	0.41161	0.16977	0.015326 *
P. Dispersion	18.9194		
E.Standar	6.23		
Desvio nulo-gl	104.683 -53		
Desvio residual-gl	68.881- 40		
AIC	442.5		
Fisher Scoring	1		
Test de dispersión	Valor Critico alpha= 0.05	level: 2.7055 Chi-Square Test Statistic = 35.4839	p.valor = 1.286×10^{-09}

Tabla 3-6.: Modelo Binomial Negativo asociado al grados de pudrición G1

Con base en los resultados mostrados en la tabla (3-6) al ajustar los datos al modelo binomial negativo con respecto al número de mazorcas clasificadas en el grado de pudrición G1 tomando como función de enlace logaritmo natural se obtuvo un desvío residual de 68,881 con 40 grados de libertad asociados y el parámetro de dispersión de 18,9194. Lo que señala un mejor ajuste de los datos al modelo binomial negativo, como puede ver en el envelope simulado de la figura (3-6 b.).

Además al contrastar el modelo de regresión Poisson con el modelo binomial negativo, se observo que el parámetro de dispersión es estadísticamente mayor a uno, ya que su p - valor fue de 1.286×10^{-09} , por lo que se rechaza el modelo de regresión Poisson a favor del modelo binomial negativo.

Por tanto del ajuste del modelo binomial negativo se tiene que los tratamientos testigo (t7 - t8) y t9 son significativos al 0,1 % y 10 % respectivamente, los tratamiento no prolífico t5 y t6 son significativos al 1 % y el tratamiento de material prolífico al 10 %. De donde se tiene evidencia estadística que estas variedades de maíz contribuyen a la explicación de la variable respuesta número de mazorcas clasificadas en el grado de pudrición G1. De este modo las variedades menos resistentes de acuerdo al grado de pudrición G1 son las variedades de selección masal por prolificidad.

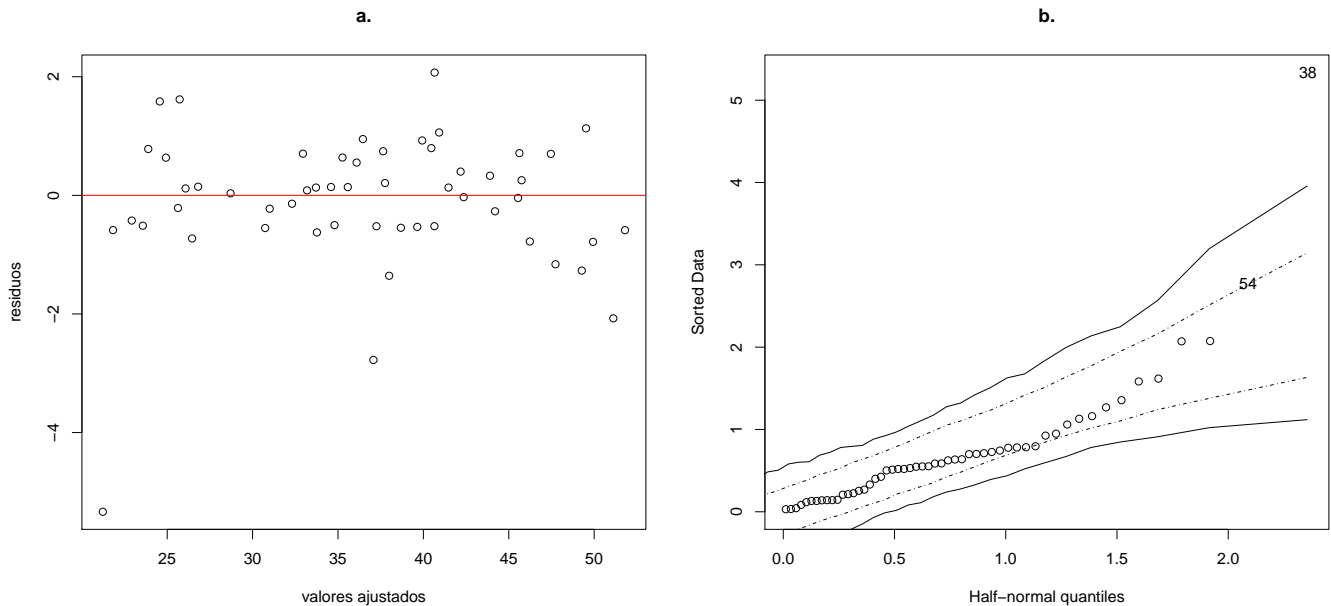


Figura 3-6.: a. Gráfico de residuales versus valores ajustado del modelo binomial negativo
 b. Envelope para el modelo binomial negativo asociado al grado de pudrición G1.

Ajuste del modelo Binomial Negativo con respecto al grado de pudrición G2

BN(link=log)			
Variable	Estimacion	E.Standar	p.valor
Intercep	2.26810	0.18837	1.2×10^{-16} ***
bloquesb2	-0.60263	0.17251	0.000477 ***
bloquesb3	-0.81621	0.18531	1.06×10^{-5} ***
bloquesb4	-0.84031	0.18687	6.90×10^{-6} ***
bloquesb5	-0.18443	0.15227	0.225829
bloquesb6	-0.74721	0.18097	3.65×10^{-5} ***
tratt2	-0.25132	0.25198	0.318589
tratt3	0.15413	0.22714	0.497394
tratt4	0.17768	0.22592	0.431571
tratt5	0.28767	0.22049	0.191990
tratt6	0.02739	0.23411	0.906857
tratt7	0.15415	0.22714	0.497357
tratt8	0.05406	0.23259	0.816206
tratt9	0.10536	0.22974	0.646522
P. Dispersion	88912.24		
E.Standar	1726311		
Desvio nulo-gl	97.852 - 53		
Desvio residual-gl	50.546 - 40		
AIC	273.35		
Fisher Scoring	1		
Test de dispersión	Valor Critico alpha= 0.05	level: 2.7055 Chi-Square Test Statistic = $-8e^{-04}$	p.valor =0.5

Tabla 3-7.: Modelo Binomial Negativo asociado al grados de pudrición G2

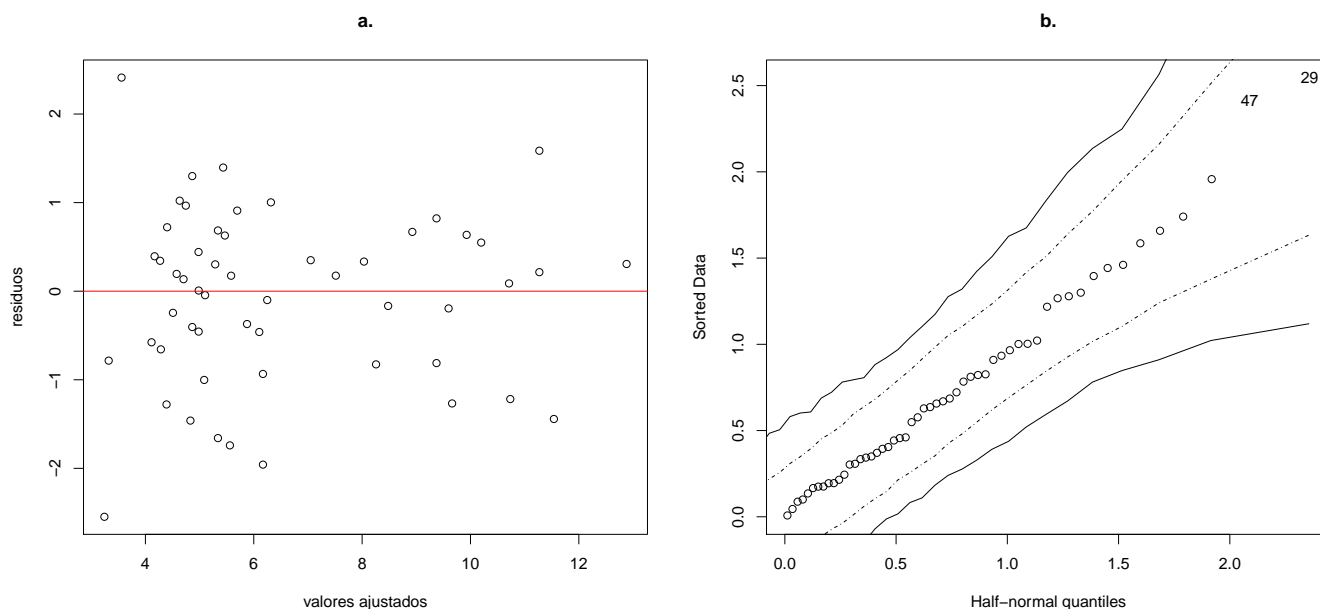


Figura 3-7.: a. Gráfico de residuales versus valores ajustado del modelo binomial negativo
 b. Envelope para el modelo binomial negativo asociado al grado de pudricion G2.

Al ajustar los datos al modelo binomial negativo con respecto al número de mazorcas clasificadas en el grado de pudrición G2 (tabla 3-7) tomando como función de enlace logaritmo natural se obtuvo un desvío residual de 50,546 con 40 grados de libertad asociados y el parámetro de dispersión de 88912,24, el cual es grande y al observar el envelope simulado de la figura (3-7 b.) se observa un buen ajuste del modelo.

Además al contrastar el modelo de regresión Poisson con el modelo Binomial negativo, se observó que el parámetro de dispersión es estadísticamente uno, ya que su p-valor fue de 0,5, por lo que el modelo binomial negativo se reduce al modelo de regresión Poisson. Por tanto del ajuste del modelo de regresión Poisson respecto al grado de pudrición G2 se tiene que los tratamientos testigo (t7 a t9), el tratamiento no prolifírico y los prolifírico no son significativo dado que en el intercepto y en los bloques se acumula toda la información del modelo.

3.2.3. Ajuste del modelo Multinomial

En esta sección se discute el ajuste del modelo multinomial a los datos del diseño experimental de la tabla (3-1) obtenidos con el fin de evaluar el grado de pudrición de mazorcas en la

variedad de maíz sogamoseño al ataque del hongo Fusarium S.P. Se consideran 9 tratamientos correspondientes a las variedades de maíz y la variable respuesta grado de pudrición que tiene tres modalidades: G0 grado de pudrición cero, G1 grado de pudrición uno y G2 grado de pudrición dos.

Frecuencias			
trat	G0	G1	G2
t1	77	149	36
t2	85	147	28
t3	73	196	42
t4	73	210	43
t5	168	256	48
t6	146	231	37
t7	317	273	42
t8	326	286	38
t9	264	225	40

Tabla 3-8.: Resumen del número de mazorcas según el grado de pudrición

Para efecto de ajuste del modelo Multinomial se compacta la tabla (3-1) en la tabla (3-8) la cual se obtuvo a partir de la suma de los conteos correspondientes a los grados de pudrición y tratamientos en una sola tabla. Se tomó como categoría de referencia G0 para el grado de pudrición y para los tratamientos la categoría de referencia es el tratamiento t9. El ajuste de los parámetros del modelo se pueden ver en la tabla (3-9) y los intervalos de confianzas correspondiente a los riesgos relativos en la tabla (3-10).

	Modelo multinomial		nominal	
variables	G1	E.Standar	G2	E.Standar
Intercepto1	0.6601341	0.1403509	-0.7602933	0.2019029
t2	-0.1123516	0.1956165	-0.3501388	0.2970578
t3	0.3275167	0.1962115	0.2074799	0.2797736
t4	0.3965072	0.1953427	0.2309917	0.2787828
t5	-0.2389191	0.1719214	-0.4924962	0.2599058
t6	-0.2013171	0.1757177	-0.6123816	0.2732049
t7	-0.8095622	0.1628372	-1.2609407	0.2602479
t8	-0.7910388	0.1620567	-1.3890130	0.2648546
t9	-0.8199838	0.1671248	-1.1267708	0.2637281
Desvio	residual:	6969.809	AIC:	7005.809

Tabla 3-9.: Modelo multinomial nominal

Para calcular la probabilidad que una frecuencia se encuentre en el grado de pudrición $k = 1, 2$ donde $k = 1$ indica el grado de pudrición G1 y $k = 2$ indica el grado de pudrición G2 y el tratamiento correspondiente, se emplea la fórmula

$$P(Y = k) = \frac{\exp[\beta_k^t X]}{1 + \sum_{i=1}^2 [\beta_i^t X]}; k = 1, 2$$

donde $\exp[\beta_k^t X] = \exp[\text{intercepto}_i + \beta_{i1}^t X]$. De este modo la probabilidad que una frecuencia se encuentre en el grado de pudrición G1 y provenga del tratamiento t_6 es

$$\begin{aligned} P(Y = 1) &= \frac{\exp[\text{intercepto}_1 \beta_1 + \text{tratt}_6 \beta_2]}{1 + \exp[\text{intercepto}_1 \beta_1 + \text{tratt}_2 \beta_2] + \exp[\text{intercepto}_2 \beta_1 + \text{tratt}_6 \beta_2]} \\ &= \frac{\exp[0,660 - 0,201]}{1 + \exp[0,660 - 0,201] + \exp[-0,760 - 0,612]} \\ &= 0,557 \end{aligned}$$

es decir, las mazorcas sometidas al tratamiento t_6 tienen una probabilidad de 0,557 de pertenecer al grado de pudrición G1. Las demás probabilidades se obtienen de manera análoga y se encuentran en la tabla del anexo (A-1)

Modelo	multinomial	nominal	intervalos de	confianza 95 %
variables	G1		G2	
	Coef/E.Standar	RR(IC 95 %)	Coef/E.Standar	RR(IC 95 %)
Intercepto1	0.66/0.14***		-0.76/0.202***	
t2	-0.11/0.196	0.89(0.61,1.31)	-0.35/0.297	0.7(0.39,1.26)
t3	0.33/0.196	1.39(0.94,2.04)	0.21/0.28	1.23(0.71,2.13)
t4	0.4/0.195*	1.49(1.01,2.18)	0.23/0.279	1.26(0.73,2.18)
t5	-0.24/0.172	0.79(0.56,1.1)	-0.49/0.26	0.61(0.37,1.02)
t6	-0.2/0.176	0.82(0.58,1.15)	-0.61/0.273*	0.54(0.32,0.93)
t7	-0.81/0.163***	0.45(0.32,0.61)	-1.26/0.26***	0.28(0.17,0.47)
t8	-0.79/0.162***	0.45(0.33,0.62)	-1.39/0.265***	0.25(0.15,0.42)
t9	-0.82/0.167***	0.44(0.32,0.61)	-1.13/0.264***	0.32(0.19,0.54)
Desvio	residual:	6969.809	AIC:	7005.809

Tabla 3-10.: Intervalos de confianza del 95 % del modelo multinomial nominal

3.2.4. Ajuste del modelo GSK

El análisis del arreglo experimental en la tabla (3-1) desarrollado por López & Chavez (1984), por medio de la aplicación del método GSK presento los siguientes resultados:

El valor del estadístico Chi cuadrado para probar la hipótesis de la igualdad entre materiales evaluados (tratamientos) fue de 171,27 con $\hat{\alpha}=0,0001$, lo cual muestra suficiente evidencia de diferencia al grado de pudrición entre los materiales.

La a prueba de bondad de ajuste al modelo resulto altamente significativa, el valor del estadístico Chi cuadrado fue 158,82.

Para dar una recomendación de la variedad más resistente a la pudrición, se plantearon las siguientes comparaciones:

1. $H_0 : 3(\tau_2 + \tau_3 + \tau_4 + \tau_5) - 4(\tau_7 + \tau_8 + \tau_9) = 0$. Este contraste compara la resistencia a la pudrición del material proflífico contra los materiales testigos, el valor del estadístico X^2 fue de 133.35 con un p - valor de 0.0001, lo cual indica que los materiales testigos son más resistentes a la pudrición que el material proflífico.
2. $H_0 : \tau_1 - \tau_6 = 0$. Compara la variedad regional contra el material proflífico. No se encontró diferencia significativa en la resistencia a la pudrición; el valor del estadístico Chi- cuadrado fue de 2,59 con $\hat{\alpha}=0,1075$.
3. $H_0 : 3\tau_1 - \tau_7 - \tau_6 - \tau_9 = 0$ contrasta la variedad regional contra los materiales testigos. El valor del estadístico $X^2=40.78$ y $\hat{\alpha}=0.0001$ conducen claramente a establecer diferencias.

En este caso los materiales testigos se mostraron mas resistentes a la pudrición que la variedad regional.

4. $H_0 : \tau_1 - \tau_9 = 0$ compara la variedad regional sogamoseño contra la variedad comercial ICAV506. Se encontró significativo el grado de pudrición, siendo más resistente la variedad comercial, como lo muestra el valor de la $X^2 = 27,07$ con $\hat{\alpha} = 0,0001$.

4. Conclusiones y recomendaciones

4.1. Conclusiones

1. Al modelar los datos con la regresión Poisson teniendo en cuenta los diferentes grados de pudrición G0, G1 y G2 se advierte presencia de sobredispersión con índice de dispersión 4,43 , 4,03 y 1,237 respectivamente, con lo cual se concluye que este modelo no es apropiada para el análisis de esta información a pesar de tener ventajas frente al método GSK (Grizzle et al. (1969)). Debido a la presencia de sobredispersión en este conjunto de datos, se procedió a modelar los diferentes grados de pudrición con el modelo alternativo binomial negativo el cual lleva en consideración este hecho. De los diferentes ajustes propuestos, se encontró que en general al modelar en forma independiente los diferentes grados de pudrición los datos se ajustan mejor principalmente con los grados G0 y G1 donde se presentan los mayores índices de dispersión, en el caso del grado G2 estos datos podrían ajustarse con el modelo de regresión Poisson, pues de los resultados se observa que el parámetro de dispersión en el modelo binomial negativo es estadísticamente igual a uno.
2. Las variedades de maíz de material testigo y no prolífico resultaron ser resistentes al hongo Fusarium S.P con respecto a los grados de pudrición G0, G1, estos resultados confirman los obtenidos por López & Chavez (1984) a través del método GSK, donde afirman que las variedades de maíz testigo son más resistentes que las de material prolífico, pero además hay evidencia estadística que también las variedades de maíz no prolífico son bastante resistentes. La razón es que al hacer uso del método GSK, se evidencia una clara falta de ajuste de este modelo, puesto que al no considerar la presencia de sobredispersión, se va a presentar sobrestimación o subestimación de los parámetro así como también se va a llevar a conclusiones erróneas sobre las hipótesis lineales de efectos de tratamientos. Se concluye que no es recomendable el método GSK para el análisis de datos de conteos, por lo menos dentro del marco presentado por los datos en este trabajo.
3. Las variedades de maíz de selección masal son las menos resistentes a los grados de pudrición G0 y G1 como se mostró en el modelo binomial negativo y confirmado en modelo multinomial, donde la probabilidad que una mazorca se encuentre clasificada en el grado de pudrición G2 y provenga de las variedades t1, t2 y t3 son respectivamente

0,137, 0,108 y 0,1350. Lo que lleva a pensar que en estas variedades de maíz el hongo *Fusarium*, S.P. no va a severar la pudrición.

4.2. Recomendaciones

Se recomienda llevar a cabo desarrollos teóricos donde se pueda considerar la interdependencia de los grados de pudrición G0, G1 y G2 , procurando ajustar modelos Poisson multivariados mixtos con efectos correlacionados de dependencia, pero considerando el efecto de bloque como aleatorio y no fijo como fue el estudio realizado en este trabajo.

Otro escenario de trabajo futuro sería llevar a cabo desarrollos teóricos que permitan hacer pruebas de comparación múltiple entre efectos de tratamientos considerando el efecto de bloque tanto fijo como aleatorio.

A. Anexo: Implementación de los modelos en R

```
# Paquetes requeridos para correr los modelos
library(MASS)
library(car)
library(dispmod)
library(epicalc)
library(nnet)
library(reshape)
library(aod)
library(boot)
library(faraway)
library(pscl)
#Introduccion de la base de datos
cont<-matrix(c(17,20,6,14,20,8,14,29,17,26,34,7,38,35,14,13,47,12,52,48,12,60,45,12,65,37,
cont0<-matrix(c(17,20,6,14,20,8,14,29,17,26,34,7,38,35,14,13,47,12,52,48,12,60,45,12,65,37
cont1<-matrix(c(17,20,6,14,20,8,14,29,17,26,34,7,38,35,14,13,47,12,52,48,12,60,45,12,65,37
cont2<-matrix(c(17,20,6,14,20,8,14,29,17,26,34,7,38,35,14,13,47,12,52,48,12,60,45,12,65,37
cont
colnames(cont)<-c("G0", "G1", "G2")
bloques<-rep(c("b1", "b2", "b3", "b4", "b5", "b6"), c(9,9,9,9,9,9))
bloques
trat<-rep(c("t1", "t2", "t3", "t4", "t5", "t6", "t7", "t8", "t9"), 6)
trat
datos<-data.frame(bloques, trat, cont)
datos0<-datos[,1:3]
par(mfrow=c(3,1))
#boxplot grado de pudrici\ 'o}n Vs n\ 'u}meros de mazorcas clasificadas en cada grado de p
boxplot(datos0[,3]~trat,xlab="Tratamientos",ylab="nÃºmero de mazorcas en G0",main="a.")
datos1<-datos[,c(1,2,4)]
boxplot(datos1[,3]~trat,xlab="Tratamientos",ylab="nÃºmero de mazorcas en G1",main="b.")
datos2<-datos[,c(1,2,5)]
boxplot(datos2[,3]~trat,xlab="Tratamientos",ylab="nÃºmero de mazorcas en G2",main="c.")
```

```
-----  
# Modelo de regresion Poisson con respecto al grado de pudrici\ '{o}n G0  
#Poisson simple (busqueda del mejor modelo alicando diversas funciones de enlace)  
modelo1<-glm(G0~bloques+trat,data=datos0,family=poisson(link="log"))  
modelo1  
anova(modelo1)  
summary(modelo1)  
modelo1.1<-glm(G0~bloques+trat,data=datos0,family=poisson(link="inverse"))  
modelo1.1  
anova(modelo1.1)  
summary(modelo1.1)
```

```
#Simulacion de envelope con los mejores modelos  
#Envelope aplicado al modelo de regresion Poisson  
#Respecto al grado de pudricion G0  
par(mfrow=c(2,2))  
res<-residuals(modelo1,type="deviance")  
plot(fitted(modelo1),res,xlab="valores ajustados", ylab="residuos", main="a." )  
abline(h=0,col="red")  
dffits(modelo1)  
go=residuals(modelo1)  
halfnorm(go,main="b." )  
lines(go.qq$x,go.env$point[1,],lty=4)  
lines(go.qq$x,go.env$point[2,],lty=4)  
lines(go.qq$x,go.env$overall[1,],lty=1)  
lines(go.qq$x,go.env$overall[2,],lty=1)
```

```
-----  
res<-residuals(modelo1.1,type="deviance")  
plot(fitted(modelo1.1),res,xlab="valores ajustados", ylab="residuos", main="c." )  
abline(h=0,col="red")  
dffits(modelo1.1)  
go=residuals(modelo1.1)  
halfnorm(go,main="d." )  
lines(go.qq$x,go.env$point[1,],lty=4)  
lines(go.qq$x,go.env$point[2,],lty=4)  
lines(go.qq$x,go.env$overall[1,],lty=1)  
lines(go.qq$x,go.env$overall[2,],lty=1)
```

```

# Modelo de regresion Poisson con respecto al grado de pudrici\`{o}n G1
#Poisson simple (busqueda del mejor modelo alicando diversas funciones de enlace)

modelo2<-glm(G1~bloques+trat,data=datos1,family=poisson(link="log"))
modelo2
anova(modelo2)
summary(modelo2)

modelo2.2<-glm(G1~bloques+trat,data=datos1,family=poisson(link="sqrt"))
modelo2.2
anova(modelo2.2)
summary(modelo2.2)

#Simulacion de envelope con los mejores modelos
#Envelope aplicado al modelo de regresion Poisson
#Respecto al grado de pudricion G1
par(mfrow=c(2,2))
res<-residuals(modelo2,type="deviance")
plot(fitted(modelo2),res,xlab="valores ajustados", ylab="residuos", main="a." )
abline(h=0,col="red")
dffits(modelo2)
go=residuals(modelo2)
halfnorm(go,main="b.")
lines(go.qq$x,go.env$point[1,],lty=4)
lines(go.qq$x,go.env$point[2,],lty=4)
lines(go.qq$x,go.env$overall[1,],lty=1)
lines(go.qq$x,go.env$overall[2,],lty=1)
-----
res<-residuals(modelo2.2,type="deviance")
plot(fitted(modelo2.2),res,xlab="valores ajustados", ylab="residuos", main="c." )
abline(h=0,col="red")
dffits(modelo2.2)
go=residuals(modelo2.2)
halfnorm(go,main="d.")
lines(go.qq$x,go.env$point[1,],lty=4)
lines(go.qq$x,go.env$point[2,],lty=4)
lines(go.qq$x,go.env$overall[1,],lty=1)
lines(go.qq$x,go.env$overall[2,],lty=1)
-----

```

```
#Modelo de regresion Poisson con respecto al grado de pudrici\'}n G2
#Poisson simple (busqueda del mejor modelo alicando diversas funciones de enlace)
modelo3<-glm(G2~bloques+trat,data=datos2,family=poisson(link="log"))
modelo3
anova(modelo3)
summary(modelo3)
modelo3.1<-glm(G2~bloques+trat,data=datos2,family=poisson(link="inverse"))
modelo3.1
anova(modelo3.1)

#Simulacion de envelope con los mejores modelos
#Envelope aplicado al modelo de regresion Poisson
#Respecto al grado de pudricion G2

par(mfrow=c(2,2))
res<-residuals(modelo3,type="deviance")
plot(fitted(modelo3),res,xlab="valores ajustados", ylab="residuos", main="a." )
abline(h=0,col="red")
dffits(modelo3)
go=residuals(modelo3)
halfnorm(go,main="b." )
lines(go.qq$x,go.env$point[1,],lty=4)
lines(go.qq$x,go.env$point[2,],lty=4)
lines(go.qq$x,go.env$overall[1,],lty=1)
lines(go.qq$x,go.env$overall[2,],lty=1)
-----
res<-residuals(modelo3.1,type="deviance")
plot(fitted(modelo3.1),res,xlab="valores ajustados", ylab="residuos", main="c." )
abline(h=0,col="red")
dffits(modelo3.1)
go=residuals(modelo3.1)
halfnorm(go, main="d." )
lines(go.qq$x,go.env$point[1,],lty=4)
lines(go.qq$x,go.env$point[2,],lty=4)
lines(go.qq$x,go.env$overall[1,],lty=1)
lines(go.qq$x,go.env$overall[2,],lty=1)

-----
#binomial negativa
```

```
#Grado de pudricion G0
par(mfrow=c(2,2))
mod1<-glm.nb(G0~bloques+trat,data=datos0,link="log")
mod1
summary(mod1)
##Envelope aplicado al modelo de regresion binomial negativo
#Respecto al grado de pudricion G0
res<-residuals(mod1,type="deviance")
plot(fitted(mod1),res,xlab="valores ajustados", ylab="residuos", main="a." )
abline(h=0,col="red")
dffits(mod1)
go=residuals(mod1)
halfnorm(go,main="b.")
lines(go.qq$x,go.env$point[1,],lty=4)
lines(go.qq$x,go.env$point[2,],lty=4)
lines(go.qq$x,go.env$overall[1,],lty=1)
lines(go.qq$x,go.env$overall[2,],lty=1)
odTest(mod1)

#binomial negativa
#Grado de pudricion G1
mod2<-glm.nb(G1~bloques+trat,data=datos1,link="log")
mod2
summary(mod2)
#Envelope aplicado al modelo de regresion binomial negativo
#Respecto al grado de pudricion G1
par(mfrow=c(1,2))
res<-residuals(mod2,type="deviance")
plot(fitted(mod2),res,xlab="valores ajustados", ylab="residuos", main="a." )
abline(h=0,col="red")
dffits(mod2)
go=residuals(mod2)
halfnorm(go,main="b.")
lines(go.qq$x,go.env$point[1,],lty=4)
lines(go.qq$x,go.env$point[2,],lty=4)
lines(go.qq$x,go.env$overall[1,],lty=1)
lines(go.qq$x,go.env$overall[2,],lty=1)
odTest(mod2)
#binomial negativa
#Grado de pudricion G2
```



```
trat<-relevel(trat,ref="t9")
pudricion<-factor(rep(c("G0","G1","G2"),9))
pudricion
#se define pudricion como un factor ordenado
#"G0"<"G1"<"G2"
pudricion<-ordered(pudricion,levels=c("G0","G1","G2"))
#organizamos los datos en un data.frame
datos.ordi<-data.frame(cont=cont,trat=trat,pudricion=pudricion)
datos.ordi
#se ajusta el modelo
poli2<-polr(pudricion~trat,weights=cont,data=datos.ordi)
summary(poli2)
#probabilidades estimadas
unique(data.frame(trat,fitted.values(poli2)))
```

Frecuencias			
trat	GO	G1	G2
t1	0.2938933	0.5687031	0.13740368
t2	0.3269266	0.5653826	0.10769080
t3	0.2347292	0.6302245	0.13504636
t4	0.2239293	0.6441713	0.13189945
t5	0.3559297	0.5423705	0.10169983
t6	0.3526532	0.5579665	0.08938031
t7	0.5015854	0.4319616	0.06645300
t8	0.5015399	0.4399948	0.05846524
t9	0.4990465	0.4253402	0.07561332

Tabla A-1.: Probabilidades estimadas

Bibliografía

- Agresti, A. (2002). Categorical data analysis. *John Wiley and Sons Hobken, New Jersey*.
- Anscombe, F. (1953). The statistical analysis of insects counts based on the negative binomial distribution. *Biometrics*, 15(15):165–73.
- Atkinson, A. (1985). Plots transformations and regression. *Oxford :Clarendon Press*.
- Carroll, R. & Ruppert, D. (1988). *Transformations and Weighting in Regression*.
- Cox, D. (1983). Some remark overdispersion. *Biometrika trust*, 70:268–274.
- Dean, C. (1992). Testing overdispersion in poisson and binomial regression models. *Journal of the American Statistical Associations*, 87:451–457.
- Dobson, A. J. (2002). *An Introduction to Generalized linear models*. 2 nd.
- Greenwood, M. & Yule, G. U. (1920). An inquiry into the nature of frequency distributions of multiple happening, whith particular reference to the ocurrence of multiple attacks of disease or repeated accidents,. *Journal of the Royal Statistical Society A*, 83:255 – 279.
- Grizzle, J., Starmer, C. F., & Koch, G. (1969). Analysis of categorical data by linear models. *Biometrics*, 25:489–504.
- Hinde, J. & Demetrio (2007). *Overdispersion: Models y Estimation*.
- Johnson, N.I. Kotz, S. (1969). Distribution in statistics. discrete distribution. *Houghton Milflin company- Boston*.
- Lambert, D. & K.Roder (1995). Overdispersion diagnostics for generalized linear models. *Journal of the American Statistical Associations*, 95:1225–1237.
- Lawless, J. (1987). Negative binomial and mixed poisson regression. *The Canadian Journal of Statistics*, 15:209–225.
- López, P. & Chavez, C. (1984). El modelo lineal en experimentos con variables categorizadas estudio de un caso. *Revista Colombiana de Estadística*, 9:87–101.
- McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models*. 2 nd.

-
- Morales, M. & López, L. (2008). Estudio de homogeneidad de la dispersion en diseños a una via de clasificación para datos de proporciones y conteo. *Revista Colombiana de Estadística*, 32:59–71.
- Nelder, J. & Pregibon, D. (1987). An extension quasi-likelihood function. *Biometrika*, 74:221–23.
- Nelder, J. A. & Wedderburn, R. (1972). Generalized linear models. *Journal of the American Statistical Society*, A, 135(3):370 – 384.
- Wedderburn, R. (1974). Quasi-likelihood funtion generalized linear models and the gauss-newton method. *Biometrika*, 64:439–47.