Seamáiz

XI Congreso Nacional de Maíz

PROTECCIÓN VEGETAL

EVALUACIÓN DEL CARBÓN DE LA ESPIGA DEL MAÍZ (Ustilago maydis) CON MODELOS ESTADÍSTICOS INFLADOS EN CERO

Videla M. E.[1,2]; Kistner, B.[1,3]; Iglesias J.[3] y Bruno, C.[1,2]
[1]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
[2] Cátedra de Estadística y Biometría. Facultad de Ciencias Agropecuarias (FCA). Universidad Nacional de Córdoba (UNC) 3Mejoramiento Genético de Maíz. EEA INTA Pergamino.
brunocecilia@gmail.com

USING ZERO INFLATED MODELS TO ANALYZE CORN SMUT (USTILAGO MAYDIS)

ABSTRACT

Plant disease incidence is measured as the number of diseased plants relative to the total number of assessed plants (discrete variable). Consequently they are usually analyzed with mixed generalized linear models (MLGM) with Poisson or Negative Binomial distribution. However these count variables generally have overdispersion and a large proportion of zero values. An alternative for the modeling of this type of variables are the zero-inflated models. These models consider a binomial process to identify false zeros and a counting process to evaluate the probability of obtaining diseased plants (the rest of the counts). In this work, four count models are compared to model the distribution of corn smut in 79 genotypes of the INTA Pergamino maize breeding program phenotyped in two locations (Tucumán and Pergamino) with two replications in each location during the agricultural season 2017-2018. The models evaluated were MLGM Poisson, MLGM Binomial Negative, Model inflated in zero Poisson (ZIP) and Inflated Model in zero Negative Binomial (ZINB). The results indicated a better performance of the inflated models in zero because they present a lower value of AIC and lower mean squared error of prediction.

Palabras Clave
Modelos lineales generalizados mixtos – Variables discretas de conteo – Enfermedades de maíz – Distribución Poisson – Distribución Binomial Negativa.

Key Words
Generalized linear mixed models - Discrete count variable – Maize disease – Poisson Distribution – Negative Binomial Distribution

INTRODUCCIÓN

Las variables epidemiológicas medidas en cultivos agrícolas, suelen registrarse con variables discretas. Para modelizar este tipo de variables se utilizan Modelos Lineales Generalizados (MLG). Para el carbón de la espiga del maíz (*Ustilago maydis* (Persoon) Roussel), el efecto del patógeno sobre la planta suele registrarse como cantidad de plantas enfermas de una parcela (variable discreta). En estudios de selección genómica para identificar genotipos de buen comportamiento al carbón de la espiga, la opción clásica es ajustar Modelos Lineales Generalizados Mixtos (MLGM) con distribución Poisson y función de enlace canónico (log). El parámetro de la distribución Poisson (λ) supone que la media y la varianza son iguales. Este supuesto raramente se cumple en epidemiología y frecuentemente a mayor cantidad de plantas enfermas, aumenta la media y como consecuencia hay mayor varianza. Cuando la varianza observada es mayor que la varianza esperada, para el modelo propuesto, se dice que los datos presentan sobredispersión. Los modelos Poisson son vulnerables a la sobredispersión debido a la relación que existe entre media y varianza ($E(Y)=\lambda=Var(Y)$). Cuando hay sobredispersión los errores estándares son subestimados, aumenta la probabilidad de cometer error de tipo I y los intervalos de confianza son poco precisos. En este contexto, los MLGM basados en la distribución Binomial Negativa incorporan un componente de variabilidad para datos con sobredispersión.

Otro desafío en este tipo de variables (de conteo) utilizadas en epidemiología es que presentan la particularidad de contener una gran proporción de valores iguales a cero (plantas sin síntoma en toda la parcela de estudio) lo cual también generan sobredispersión. Ajustar modelos con distribuciones que no contemplen el exceso de ceros, como es el caso de la Poisson y Binomial Negativa, puede resultar en inferencias estadísticas ineficientes. Entonces, es neceserio identificar cuidadosamente cómo surgen estos ceros y cuál es la mejor forma de modelarlos. Estos ceros pueden clasificarse en "*falsos ceros*", causados por parcelas que no estuvieron expuestas a la enfermedad (escapes) o por errores del observador, y en "*ceros auténticos*" que son aquellos generados por parcelas que si han estado expuestas al hongo pero por diversos motivos (resistencia del genotipo, condiciones ambientales, etc.) ninguna planta ha sido infectada. Distinguir unos ceros de otros (falsos ceros de auténticos ceros) es fundamental no sólo para reducir el número de falsos ceros (por ejemplo, entrenar al observador o evitar áreas en las que no está el patógeno) y como consecuencia disminuir la sobredispersión, sino también para identificar genotipos que, habiendo estado expuestos a la enfermedad, no sufrieron infecciones (resistentes). Una alternativa para estos casos son los MLGM de tipo "Inflados en Cero" (Zero Inflated) o también llamados "Modelos Mezcla" (Mixture Models). Estos modelos asumen que la variable respuesta (Y) se comporta como una mezcla de dos distribuciones: una binaria para modelizar la probabilidad de medir falsos ceros (binomial) y una de conteo para modelizar la probabilidad de obtener el resto de los valores, incluyendo los ceros auténticos. Cuando se utiliza una distribución Poisson para el proceso de conteo, estos modelos son llamados "ZIP" y cuando se utiliza una distribución Binomial Negativa "ZINB".

El objetivo de este trabajo es comparar modelos de conteo que no contemplan los excesos de ceros (MLGM Poisson (MP) y MLGM (MNB) Binomial Negativa) respecto a modelos de conteo para datos inflados en cero (ZIP y ZINB) para modelar la distribución del carbón de la espiga del maíz con la finalidad de seleccionar genotipos resistentes.

MATERIALES Y MÉTODOS

Se evaluaron 79 líneas endocriadas de maíz, pertenecientes al grupo de mejoramiento de maíz de INTA Pergamino, bajo infección natural para el carbón de la espiga del maíz (*Ustilago maydis*). El fenotipado se realizó contabilizando la cantidad de plantas con síntomas sobre parcelas con un total de 30 plantas. Los datos fueron recolectados en dos localidades (Tucumán y Pergamino) con dos repeticiones cada uno durante la campaña agrícola 2017/2018.

Se ajustaron cuatro modelos para datos de conteo: MLGM Poisson (MP), MLGM Binomial Negativa (MNB), Modeo Poisson Inflado en Cero (ZIP) y Modelo Binomial Negativa Inflado en Cero (ZINB). Los modelos contemplaron el efecto de localidad y de repetición dentro de localidad como fijo y el efecto de genotipo como aleatorio para estimar el mejor predictor lineal insesgado (BLUP) y posteriormente realizar un ranking de los genotipos de acuerdo a su comportamiento de resistencia al carbón de la espiga. La interacción GenotipoxLocalidad no fue estimable debido a que algunos genotipos fueron evaluados en una sola localidad o en ambas localidades pero con sólo una repetición. El desempeño de los modelos se comparó con criterio de información de Akaike (AIC) y el error cuadrático medio de predicción (ECMP). La validación de los modelos se realizó sobre 100 conjuntos de datos simulados a partir de los datos reales usando AIC. Para la simulación se mantuvo la proporción de ceros (60%) y los conteos de plantas enfermas se simularon para un parámetro λ igual al observado para cada genotipo en cada localidad en cada repeticón.

RESULTADOS Y DISCUSIÓN

Los modelos inflados en cero (ZIP y ZINB) tuvieron mejor ajuste (menor AIC y menor ECMP) que los modelos de conteo tradicionales que no contemplaron el exceso de ceros (MP y MNB) (Tabla 1). Para los modelos inflados en cero, el modelo ZINB tuvo mejor performance que el ZIP (AIC=844.39 vs AIC=847.26, respectivamente).

En la validación de los modelos, a través de los 100 conjuntos de datos simulados, también presentaron menores valores de AIC los modelos inflados en cero (ZIP y ZINB vs MP y MNB). En promedio se obtuvieron menores valores de AIC para el modelo ZINB respecto al modelo ZIP (844,4 y 851,2 respectivamente) y menor coeficiente de variación (CVZINB=0,35 vs CVZIP=0,4 respectivamente)

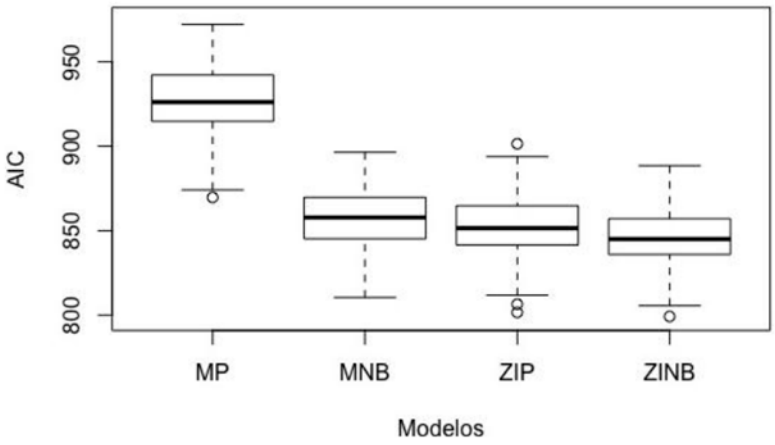

Figura 1. *Gráficos de cajas de criterio de AIC de cuatro modelos para 100 sets de datos simulados. MP=MLGM Poisson; MNB=MLGM Binomial Negativo; ZIP= Modelo Inflado en cero Poisson; ZINB=Modelo Inflado en cero Binomial Negativo.*

Modelo	df	AIC	ECMP
ZINB	8	844.39	**2.26**
ZIP	7	847.26	2.79
MNB	5	851.28	3.05
MP	4	889.01	3.33

Tabla 1. *Criterio de Akaike (AIC), Error cuadrático medio de predicción (ECMP) y grados de libertad (df) para modelos ajustados a carbón de la espiga del maíz para 79 genotipos (menor es mejor).*

Dado el mejor comportamiento del modelo ZINB respecto a los otros modelos, se obtuvieron los BLUPs de los genotipos para ordenarlos de mayor a menor susceptibilidad.

CONCLUSIÓN

Para modelar variables de conteo con gran proporción de ceros, como los observados en enfermedades de cultivos agrícolas, los modelos inflados en cero son más precisos respecto a modelos tradicionales como MLGM Poisson y Binomial Negativa.

Apoyo financiero: PNCYO 1127043-INTA. Concejo Nacional de Investigaciones Científicas y Técnicas (CONICET).

Agradecimientos

Los autores agradecen a los técnicos auxiliares del grupo de Mejoramiento de Maíz de la EEA INTA Pergamino y IIACS Santa Rosa de Leales por su participación en actividades de campo y a la Cátedra de Fitopatología de la UNNOBA por su colaboración en la realización del presente trabajo.