Leticia Hachuel
Gabriela Boggio
Guillermina Harvey
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística

MODELOS ALTERNATIVOS PARA EL ANÁLISIS DE DATOS DE CONTEO CON EXCESO DE CEROS

1. INTRODUCCIÓN

Para datos de conteo se suele utilizar la distribución Poisson como componente aleatorio en el proceso de ajuste de un modelo lineal generalizado. Esta distribución se caracteriza por la igualdad entre su media y su variancia, supuesto difícil de verificar ya que en la práctica las observaciones de conteos frecuentemente exhiben una variabilidad que excede la supuesta para una variable del tipo Poisson. El fenómeno por el cual un modelo lineal generalizado tiene mayor variabilidad que la presupuesta por el componente aleatorio del mismo se denomina sobredispersión.

Son muchas las posibles causas de la presencia de sobredispersión y en el proceso de modelar datos, varias de ellas pueden intervenir simultáneamente. A veces la sobredispersión puede corregirse mediante la incorporación de un predictor apropiado, la inclusión de interacciones significativas, o bien utilizando una función de enlace diferente. Otras veces la sobredispersión puede provenir de la presencia excesiva de ceros en las observaciones y en tal caso hay una variedad de métodos utilizados para manejarla (Hilbe, 2007).

Los métodos para tratar la sobredispersión producto del exceso de ceros se basan en modelos que complementan los métodos más convencionales que se concentran solamente en modelar correctamente la relación media-variancia.

En este trabajo se presentan, a través de una aplicación a un problema específico, diferentes alternativas para la detección y posible solución del fenómeno de sobredispersión.

2. ASPECTOS METODOLÓGICOS

La falla del supuesto de equidispersión en Poisson tiene consecuencias cualitativas similares a la falta de homocedasticidad en el modelo de regresión lineal, pero la magnitud del efecto sobre los errores estándares de los estimadores de los parámetros del modelo puede ser mucho mayor.

Los datos Poisson se dicen sobredispersos si la variancia excede la media, por lo que un indicador simple de la magnitud de la sobredispersión la da la comparación entre la media muestral y la variancia de la variable de conteo en estudio. Sin embargo, hay que tener en cuenta que al realizar un análisis de regresión con estos datos, si la variancia muestral es más del doble de la media, probablemente los datos permanezcan sobredispersos aún después de la inclusión de regresores.

Una práctica común para detectar sobredispersión en el modelo Poisson es ajustar dicho modelo y luego llevar a cabo diversos tests cuyas características hacen que sus distribuciones no sean estándares y requieran el ajuste del nivel de significación.

Si bien el modelo Poisson es el paradigma o el modelo básico para respuestas tipo conteo, la regresión Binomial Negativa es casi siempre pensada como el modelo alternativo al Poisson cuando hay sobredispersión en los datos. En el modelo Binomial Negativo se supone que la media condicional de Y_i está determinada no sólo por la heterogeneidad explicada por el vector de covariables X_i sino también por una heterogeneidad no observada independiente de X_i.

El modelo de regresión estándar Binomial Negativo generalmente se denomina, siguiendo a Cameron y Trivedi (1998) como NB2. En él la media condicional $E(y_i/x_i)$ sigue siendo μ_i, pero $Var(y_i/x_i)$ se transforma en $\mu_i+\alpha\mu_i^2$. Como $\mu_i>0$ y $\alpha>0$, la variancia debe exceder la media. Para interpretar la magnitud de la sobredispersión es útil reescribir la variancia tipo NB2 como $(1+\alpha\mu_i)\mu_i$. Un valor de considerable sobredispersión aparece si $\alpha\mu i>1$ ya que entonces $1+\alpha\mu_i>2$ multiplica a la media. Así un valor de α igual a 0,5 indica una modesta sobredispersión si la variable dependiente toma mayormente valores 0,1 y 2 pero, en cambio, indica una alta sobredispersión si los conteos observados son de 10 y más.

A veces se presenta un alto porcentaje de ceros en todos los niveles de los predictores y sus efectos no pueden ser capturados por las funciones de variancia de las distribuciones Poisson o Binomial Negativa. Para modelar este exceso de ceros puede ser apropiado un modelo llamado "zero-inflated". Éste asume que las observaciones pueden pertenecer a dos grupos. Un grupo es muy probable que tenga un conteo igual a cero, el otro grupo sigue una de las dos distribuciones tradicionales para conteos: Poisson o Binomial Negativa.

Otro modelo que maneja el exceso de ceros es el denominado modelo "Hurdle" que difiere del modelo anterior en cómo entienden el origen o generación de los ceros extras.

Las diferentes alternativas de análisis se presentan aplicadas a un conjunto de datos los cuales se describen a continuación.

3. LOS DATOS

El equipo de investigación a cargo de la Dra. Silvia Revelli en el Instituto de Inmunología de la Facultad de Ciencias Médicas ha realizado una serie de experiencias con el objeto de evaluar, mediante el modelo de la enfermedad de Chagas en ratas, los efectos de distintos Actinomycetales muertos por calor sobre la parasitemia (número de parásitos en sangre). A través de estas experiencias se intentó determinar si la teoría de la higiene se podía aplicar a las enfermedades parasitarias.

El objetivo específico del experimento que originó los datos que se utilizan en esta presentación fue determinar cuál o cuáles Actinomycetales provocan una mayor reducción en la parasitemia para utilizarlos en el tratamiento de la enfermedad de Chagas. El diseño de la experiencia se describe a continuación.

Al momento del nacimiento, a los 7 y a los 28 días de vida se inocularon ratas machos de la línea denominada "l" en la región posterior del cuello con los siguientes Actinomycetales: Gordonia bronchialis (Gb), Rhodococcus coprophilus (Rc) y Tsukamurella inchonensis (Ti) muertas por calor, o bien con Solución Fisiológica (SF) -placebo-.

A los 21 días de vida en todas las ratas se indujo la infección con Trypanosoma cruzi (Tc) - agente causal de la enfermedad de Chagas - por inoculación de tripomastigotes vivos. El objetivo es, entonces, determinar el Actinomycetal que produce una mayor reducción en el número de parásitos en sangre en comparación con el grupo placebo (SF).

La parasitemia es una variable de repercusión sistémica que muestra cómo varía la enfermedad. Desde el momento en que las ratas son inyectadas con Tc comienzan a reproducirse los parásitos en la sangre hasta que se produce un pico. Paralelamente la rata produce anticuerpos que matan a estos parásitos, por lo tanto, a medida que transcurren los días el número de parásitos en la sangre disminuye. La infección aguda se monitoreó mediante su evaluación a distintos días post infección y en esta oportunidad se analizan las mediciones registradas a los 10 días de ser infectadas.

La Tabla 1 presenta algunas medidas descriptivas del número de parásitos en sangre observado, parasitemia, para los cuatro grupos de ratas inoculadas con los diferentes tratamientos mencionados o el placebo.

Tabla 1 – Medidas descriptivas de la parasitemia en el décimo día luego de la infección con Tc para los distintos grupos.

Grupo	Media	Variancia	Mínimo	Máximo	Nº de ratas con 0 parásitos	Nº de ratas
SF	13,56	172,13	0	40	1	9
Gb	1,33	3,76	0	5	5	9
Rc	3,00	10,30	0	9	3	8
Ti	9,17	30,14	2	18	0	6
Total	**6,66**	**79,21**	**0**	**40**	**9**	**32**

Se puede observar la diferencia notoria entre la media y la variancia para los cuatro grupos, sobre todo para SF, así como también para la totalidad de los datos. Esto indica una potencial presencia de sobredispersión en los datos. Se observa además que el 28% de las ratas (9 de 32) no presentaron parásitos en sangre en esta evaluación.

Para la aplicación de las diferentes alternativas que se presentan en este trabajo se utilizan los procedimientos IML, GENMOD y NLMIXED del programa SAS.

4. ALTERNATIVAS METODOLÓGICAS

4.1 Modelo Poisson

Los datos sobre el número de parásitos en sangre se modelan a través de la distribución de Poisson. Si se supone, entonces, que las variables aleatorias Y_i, i = 1, ... ,n con n = 32, se comportan siguiendo una distribución Poisson con media $\mu_{i,}$, $Y_i \sim P(\mu_i)$, es de esperar que $E(Y_i) = \mu_i$ y $Var(Y_i) = \mu_{i.}$

Se ajusta un modelo lineal generalizado (MLG) Poisson con la variable explicativa grupo que, por ser una variable categórica con cuatro niveles, implica la consideración de tres variables indicadoras D_i, i = 1, 2, 3 para cada uno de los tres tratamientos considerando al grupo SF como categoría de referencia.

El modelo se explicita:

$$\text{Log}(\boldsymbol{\mu}) = \beta_0+\beta_1 D_1+\beta_2 D_2+\beta_3 D_3 \qquad \textbf{(1)}$$

siendo **μ** el vector de promedios del número de parásitos en sangre para cada uno de los

cuatro grupos.

Los resultados obtenidos se presentan a continuación (el programa utilizado figura en el Anexo 1).

Tabla 2 - Parámetros estimados, errores estándares y probabilidades asociadas para el modelo Poisson.

Parámetro	Estimación	Error estándar	Prob. asociada
β_0	2,6068	0,0905	<,0001
β_1 (Gb)	-2,3191	0,3025	<,0001
β_2 (Rc)	-1,5082	0,2233	<,0001
β_3 (Ti)	-0,3912	0,1624	0,0219

El valor obtenido para la estadística de bondad de ajuste de Pearson fue igual a X^2 = 164,48 similar al valor de la Deviance, D = 170,88, ambas con 28 grados de libertad, lo cual indica falta de ajuste.

Una "señal" acerca de la presencia o no de sobredispersión en datos de conteo está dada por la relación entre la Deviance y sus grados de libertad (gl). Cuando se ajusta un MLG con un parámetro de escala conocido, como es el caso de la distribución Poisson para el cual φ = 1, sujeto a ciertas condiciones asintóticas y cuando el modelo ajusta bien se espera que D ≈ gl. Si por el contrario, la Deviance resulta mayor que los grados de libertad, se puede presumir la presencia de sobredispersión (Demetrio y Hinde, 1998).

Parece razonable, entonces, considerar como medida de sobredispersión el cociente entre el valor de D o de X^2 por los grados de libertad. Pero para evaluar la importancia de la sobredispersión hay que además tener en cuenta la cantidad de observaciones. Por ejemplo, para modelos ajustados a un gran número de observaciones se puede considerar que hay sobredispersión si ese cociente es sólo mayor a 1,05. En cambio para un número moderado de observaciones, el cociente debería ser mayor que 1,25 (Hilbe, 2007).

En este caso particular estos cocientes resultan D/gl = 6,10 y X^2/gl = 5,87, siendo claramente una señal positiva de sobredispersión.

Si bien el fenómeno de la sobredispersión no tiene influencia en la estimación de los coeficientes β, es importante tenerlo en cuenta ya que provoca que los errores estándares de las estimaciones obtenidos a partir del modelo sean incorrectos y puedan ser seriamente subestimados. Esta subestimación conduce a que resulten significativos coeficientes que en realidad pueden no serlo (Hilbe, 2007).

Antes de adoptar una conducta que corrija los efectos de la sobredispersión sobre las estimaciones, resulta útil poder confirmar su presencia a partir de tests estadísticos específicos.

4.2 Tests

Las pruebas estadísticas evalúan la presencia de sobredispersión contrastando la igualdad entre media y variancia impuesta por Poisson contra la alternativa de que la varian-

cia excede la media. Cameron y Trivedi (1998) plantearon un test de sobredispersión a partir de la estimación de un modelo Poisson teniendo en cuenta como hipótesis alternativa la forma Binomial Negativa tipo 2 (NB2) para la variancia, según la cual:

$$Var(Y_i) = \mu_i + \alpha\mu_i^2 .$$

De ella se deduce que:

$$\alpha = E\left[\frac{(y_i - \mu_i)^2 - \mu_i}{\mu_i^2}\right]$$

Esta expresión conduce a que el test se haya planteado como una regresión lineal clásica auxiliar sin la constante o *intercepto* utilizando los valores ajustados $\hat{\mu}_i = \exp(X_i\hat{\beta})$ $\hat{\mu}_i = exp\left(X_i\widehat{\beta}\right)$ $\hat{\mu}_i = \exp\left(X_i\widehat{\beta}\right)$ como variable independiente y la expresión $\frac{(y_i - \hat{\mu}_i)^2 - y_i}{\hat{\mu}_i}$ $\frac{(y_i - \hat{\mu}_i)^2 - y_i}{\hat{\mu}_i}$ como variable dependiente. Así, el modelo de regresión lineal resulta:

$$\frac{(y_i - \hat{\mu}_i)^2 - y_i}{\hat{\mu}_i} = \alpha\hat{\mu}_i + \varepsilon_i ,$$ $$\frac{(y_i - \hat{\mu}_i)^2 - y_i}{\hat{\mu}_i} = \alpha\hat{\mu}_i + u_i$$ donde ε_i $\hat{\mu}_i = \exp\left(X_i\widehat{\beta}\right)$ u_i es el término correspondiente al error.

La significación estadística del coeficiente α indica la existencia de sobredispersión. La estadística correspondiente es asintóticamente normal bajo la hipótesis nula de no sobredispersión contra la alternativa de la forma NB2 (Liu y Cela, 2008).

Hilbe (2007) presenta un test de características similares al recién descripto cuya estadística es:

$\sum_{i=1}^{n}\frac{Z_i}{n}$, donde $Z_i = \frac{(y_i - \hat{\mu}_i)^2 - y_i}{\hat{\mu}_i\sqrt{2}}$. La estadística Z tiene distribución t de student bajo la hipótesis nula.

Este autor presenta también el test del multiplicador de Lagrange, cuya estadística es:

$$\chi^2 = \frac{\left(\sum_i \mu_i^2 - n\bar{y}\right)^2}{2\sum_i \mu_i^2}$$ $\chi^2 = \frac{(\sum_{i=1}^{n}\hat{\mu}_i^2 - n\bar{y})^2}{2\sum_{i=1}^{n}\hat{\mu}_i^2}$, la cual bajo la hipótesis nula de no sobredispersión, tiene una distribución X^2 con un grado de libertad.

En el Anexo 2 se presentan los programas desarrollados para llevar a cabo estos tests.

Los tres tests son post-hoc, es decir se realizan luego del ajuste de los datos con el modelo Poisson. Para los datos en estudio ellos conducen a la misma conclusión confirmando la existencia de sobredispersión (Tabla 3).

Tabla 3 - Tests de sobredispersión

Test		Prob. asociada
Cameron y Trivedi	$\hat{\alpha} = 0,63$	0,0013
Score (Hilbe)	Z = 446,41	<0,0001
Multiplicador de Lagrange (Hilbe)	$X^2 = 920,07$	<0,0001

Resulta importante, entonces, corregir sus efectos. La forma más simple de hacerlo es a través de la corrección de los errores estándares, tal como se describe a continuación.

4.3 Modelo Poisson corregido por sobredispersión

Una alternativa simple de tener en cuenta la sobredispersión es asumir una forma diferente para la función de variancia. Este método consiste en reemplazar la función media-variancia del modelo Poisson original, $Var(Y_i)=\mu_i$, por uno más general involucrando parámetros adicionales $Var(Y_i)=\varphi \mu_i$. En este caso, la distribución de la variable deja de ser Poisson y no se pueden estimar los parámetros por máxima verosimilitud. Weddeburn (1974) propuso el denominado enfoque cuasi-verosímil. Este método consiste en utilizar directamente en las ecuaciones de verosimilitud una función de variancia particular aún cuando no se trate de la específica para la distribución, en este caso Poisson. Esta función de variancia diferente afecta los desvíos estándares de las estimaciones de los parámetros del modelo y no a las estimaciones en sí mismas ya que φ sólo interviene en la definición de la variancia. Esta alternativa se traduce en ajustar el modelo ordinario Poisson y luego corregir las estimaciones de los errores estándares de los parámetros estimados multiplicándolos por una estimación de $\varphi^{1/2}$. Para estimar φ se pueden usar los cocientes D/gl ó X^2/gl, pero cabe aclarar que en este caso ya no se dispone de una medida de bondad de ajuste.

El programa en SAS para la estimación del parámetro de dispersión y su utilización para la corrección de los desvíos estándares figura en el Anexo 1.

Los resultados hallados se presentan en la siguiente tabla.

Tabla 4 - Parámetros estimados, errores estándares y probabilidades asociadas para el modelo Poisson corregido por sobredispersión.

Parámetro	Estimación	Error estándar	Prob. asociada
β_0	2,6068	0,2194	<,0001
β_1 (Gb)	-2,3191	0,7333	0,0016
β_2 (Rc)	-1,5082	0,5412	0,0053
β_3 (Ti)	-0,3912	0,3936	0,3203
$\sqrt{\varphi}$	2,4237	0,0000	-

De acuerdo a lo esperado, el valor de las estimaciones no se ve afectado por el hecho de haber considerado φ ≠ 1 y por el contrario, los errores estándares de dichas estimaciones sí se han modificado. Como resultado de este ajuste, el grupo tratado con Ti no resulta significativo a diferencia de lo observado en la Tabla 2.

4.4 Modelo Binomial Negativo

Una alternativa que puede resultar más realista para tratar la sobredispersión, es no suponer que hay un mismo factor afectando a la media de todos los tratamientos, sino suponer que:

$$Var(Y_i) = \mu_i + \alpha\mu_i^2 .$$

Para ello el modelo se formaliza de igual manera que el modelo **(1)** pero ahora se asume que *Y* tiene distribución Binomial Negativa.

El programa utilizando el procedimiento GENMOD de SAS figura en el Anexo 1.

El ajuste del modelo da como resultado un valor de la estadística de Pearson X^2 igual a 26,81 y un valor de la Deviance igual a 37,64, ambos con 28 gl. El ajuste satisfactorio obtenido con el MLG con distribución Binomial Negativa indica que existe una variabilidad intrínseca a la variable en estudio que hace que no sea correcto aplicar el primer enfoque.

Los resultados encontrados para las estimaciones se presentan en la Tabla 5.

Tabla 5 - Parámetros estimados, errores estándares y probabilidades asociadas para el modelo Binomial Negativo.

Parámetro	Estimación	Error estándar	Prob. asociada
β_0	2,6068	0,3240	<,0001
β_1 (Gb)	-2,3191	0,5340	<,0001
β_2 (Rc)	-1,5082	0,5055	0,0029
β_3 (Ti)	-0,3912	0,5181	0,4501
α^{-1}	0,8712	0,3206	-

Los valores estimados de los coeficientes de regresión asociados a los grupos son iguales a los obtenidos para el modelo Poisson. Los desvíos estándares se modifican en relación al modelo Poisson inicial si bien en diferente magnitud que en lo observado en el modelo Poisson corregido. En esta oportunidad tampoco resulta significativo el efecto del tratamiento Ti con respecto al grupo SF. Es decir, los grupos tratados con Gb y Rc muestran una disminución significativa en comparación con el grupo tratado con placebo.

4.5 Comparación entre modelos

Es interesante notar que la regresión Binomial Negativa es una extensión de la regresión Poisson con un supuesto de variancia más liberal y podría colapsar en la regresión Poisson con el parámetro α igual a cero. WenSui Liu y Jimmy Cela (2008) y Hilbe (2007) se basan en este hecho para proponer comparar ambos modelos mediante sus log-verosimilitudes a través de un test de razón de verosimilitud.

Sin embargo, en la literatura sobre modelos lineales generalizados, es considerado un obstáculo para la comparación de dos modelos el hecho en que difieran en el componente aleatorio y se propone utilizar el criterio de información de Akaike, AIC, o el criterio de información bayesiano, BIC, para la selección del modelo más adecuado.

Estos coeficientes resultan ser para el modelo Poisson: AIC = 263,98 y BIC = 269,85 y para el modelo NB2: AIC = 179,32 y BIC = 186,65, mostrando que el modelo más adecuado es el que utiliza la distribución Binomial Negativa, por lo que el parámetro α de esta distribución parece captar el exceso de variación.

4.6 Regresión "Zero-inflated"

Frecuentemente la sobredispersión se produce por la presencia de un número mayor de conteos nulos que el esperado bajo la distribución Poisson supuesta para las observaciones. En estos casos, una manera de modelar datos de conteo con excesivos ceros es la denominada Regresión "Zero-inflated" introducida por Lambert (1992). La misma asume que los conteos nulos pueden provenir de dos fuentes diferentes, por lo que considera una mezcla de dos procesos estadísticos, uno que genera sólo conteos iguales a cero y otro que genera tanto conteos ceros como distintos de cero. Más específicamente, se utiliza una variable aleatoria Bernoulli para determinar si un resultado de conteo individual surge del proceso que genera sólo ceros o no, a través de un modelo logit. Luego, se utiliza un modelo Poisson o Binomial Negativo para modelar los resultados generados por el otro proceso (Liu y Cela, 2008).
Cuando las covariables utilizadas en los modelos correspondientes a ambos procesos son las mismas, un modelo de Regresión "Zero-inflated" parsimonioso es aquél que supone que el vector de coeficientes del modelo logit es el producto entre el vector de coeficientes del sub-modelo Poisson y un escalar (Ridout et al, 1998). Esto es:

$$\log\left(\frac{\omega_i}{1-\omega_i}\right) = \tau X_i\beta \quad \text{y} \quad \log(\lambda_i) = X_i\beta \qquad \textbf{(2)}$$

suponiendo que la variable Y tiene una distribución Poisson "Zero-inflated" dada por:

$$\Pr(Y=y) = \begin{cases} \omega + (1-\omega)\exp(-\lambda), & y = 0 \\ (1-\omega)\exp(-\lambda)\lambda^y / y!, & y > 0 \end{cases}$$

El ajuste de este modelo se realiza mediante el procedimiento NLMIXED de SAS cuyo programa figura en el Anexo 1. A continuación se presentan los resultados cuando se utilizan las mismas covariables en ambas partes del modelo, en este caso, las variables de diseño asociadas al tratamiento (Tabla 6).

Tabla 6 - Parámetros estimados, errores estándares y probabilidades asociadas para el modelo "Zero-inflated".

Parámetro	Estimación	Error estándar	Prob. asociada
β_0	2,7250	0,0904	<0,0001
β_1 (Gb)	-1,8702	0,3297	<0,0001
β_2 (Rc)	-1,2036	0,2304	<0,0001
β_3 (Ti)	-0,4965	0,1612	0,0042
τ	-0,7343	0,2682	0,0101

Los coeficientes estimados mantienen el sentido ya observado en los otros modelos con alguna diferencia en magnitud. Es de notar que en este caso todos los coeficientes resultan significativos. Sin embargo el valor de AIC y BIC (227,70 y 235,10 respectivamente) no lo señalan como el mejor modelo en comparación con el NB2.

Esta discrepancia en los resultados favorece la idea de intentar comparar mediante otros métodos los modelos NB2 y "Zero-inflated" antes de concluir definitivamente que no existen dos procesos diferentes que den origen al exceso de ceros observados. Además resta la alternativa de probar el ajuste del denominado modelo "Hurdle", que también considera dos procesos diferentes para la generación de datos pero donde los ceros son sólo generados por uno de ellos.

5. CONSIDERACIONES FINALES

Frecuentemente ocurre que en los datos de conteos ajustados con la distribución Poisson se presenta sobredispersión. Es importante tener en cuenta este fenómeno, ya que de no hacerlo es probable arribar a conclusiones erróneas.

En este trabajo se presentan diferentes alternativas para el tratamiento de este fenómeno provocado por la presencia de una mayor cantidad de ceros que lo esperado en los conteos observados. Para ello se utilizan datos sobre parasitemia en ratas infectadas con Tc*,* consistentes en el recuento de parásitos en sangre después de haber inoculado ciertas bacterias (Actinomycetales) en tres grupos experimentales comparando los resultados obtenidos con un grupo control que fue inoculado con solución fisiológica.

Se obtuvo un ajuste satisfactorio bajo el modelo Binomial Negativo lo que permitió responder al interrogante planteado sobre cuál de los Actinomycetales resulta más efectivo para reducir la parasitemia en las ratas, que fue el objetivo del estudio que dio origen a los datos. La gran variabilidad detectada caracteriza al fenómeno en estudio; es una variabilidad intrínseca a la variable estudiada, número de parásitos en sangre, y por lo tanto no es sencillo adjudicarla a una causa en particular. Hubo evidencia de disminución significativa en el número de parásitos en sangre cuando se inoculó con las bacterias Gb o con Rc, en comparación con los animales que no fueron inyectados con las bacterias.

Las diferentes metodologías presentadas, si bien no abarcan todos los procedimientos factibles en la actualidad, mostraron en la resolución del problema sus alcances y limitaciones para responder en forma eficiente el interrogante planteado.

6. REFERENCIAS BIBLIOGRÁFICAS

Cameron, A. C.; Trivedi, P. K. (1998). *Regression Analysis of Count Data*. Cambridge University Press.

García Borges Demetrio, C.; Hinde, C. J. (1998). *Overdispersion: Models and Estimation*. Associação Brasileira de Estadistica.

Hilbe, H. (2007). *Negative Binomial Regression*, Arizona State University, Cambridge University Press.

Lambert, D. (1992). *Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing*. *Technometrics,* Vol. 34, No. 1, 1 – 14.

Liu, W.; Cela, J. (2008). Count Data Models in SAS. SAS Global Forum 2008, *Statistics and Data Analysis.*

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

Weddeburn, W. M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, Vol. 61, 439 - 447.

ANEXOS

Anexo 1: Ajuste de modelos

Modelo Poisson

```
title "Modelo Poisson - PROC GENMOD";
proc genmod data = parasitos order=data;
    class grupo;
    model parasit10 = grupo / dist=poi link=log type3;
run;
```

Modelo Poisson corregido

```
title "Modelo Poisson corregido por sobredispersion";
proc genmod data=parasitos order=data;
    class grupo;
    model parasit10 = grupo / dist=poi link=log scale=pearson type3;
run;
```

Modelo Binomial Negativo

```
title "Modelo Binomial negativa - PROC GENMOD";
proc genmod order=data; *para que tome =0 al param corresp al grupo control;
    class grupo;
    model parasit10 = grupo / dist=nb link=log type3;
run;
```

Modelo “Zero inflated”

```
title "Zero-inflated Poisson Regression (ZIP) - PROC NLMIXED";
proc NLMIXED data = paras_gd tech = dbldog;
    parms a0 = 0 a1=0 a2=0 a3 = 0
                b0 = 0 b1=0 b2=0 b3 = 0;
    eta0 = a0 + a1*dam1 + a2*dam2 + a3*dam3;
    exp_eta0 = exp(eta0);
    p0 = exp_eta0 / (1 + exp_eta0);
    etap = b0 + b1*dam1 + b2*dam2 + b3*dam3;
    exp_etap = exp(etap);
    if parasit10 = 0 then ll = log(p0 + (1 - p0) * exp(-exp_etap));
    else ll = log(1 - p0) + parasit10*etap - exp_etap - lgamma(parasit10 + 1);
    model parasit10 ~ general (ll);
    predict exp_etap out= zip_out1 (keep = pred parasit10 rename = ( pred = Yhat));
    predict p0 out = zip_out2 (keep = pred rename = ( pred = p0));
run;
```

Anexo 2: Tests

```
/* 1º Test (Cameron y Trivedi)*/
data testCyT;
    set poi_out;
    dep = ((parasit10 - Yhat) ** 2 - parasit10)/ Yhat;
run;
proc reg data = testCyT;
    model dep = Yhat / noint;
run;

/* 2º Test (Hilbe)*/
data test3;
    set poi_out;
    z = ((parasit10 - Yhat) ** 2 - parasit10)/ (Yhat * sqrt(2));
run;
proc reg data = test3;
    model z = ;
run;

/* 3º Test (Multiplicador de Lagrange)*/
proc iml;
    use poi_out;
    read all var {parasit10} into y;
    read all var {Yhat} into yhat;
    close poisson_out;
    n = nrow(y);
    ybar = y`[, :];
    chi2 = (yhat` * yhat - n * ybar) ** 2 / (2 * yhat` * yhat);
    pvalue = 1 - probchi(chi2, 1);
    print chi2 Pvalue;
    quit;
```