Modelando datos de conteo extendiendo la regresión de Poisson

Christian R.A. Vásquez-Velasco
Universidad Nacional de Ingeniería, Perú

1 Introducción

El modelado de datos de conteo es de suma importancia en diversas áreas, particularmente en epidemiología, donde es crucial para comprender la propagación de enfermedades. Los datos de conteo se refieren a aquellos que cuentan eventos discretos, como el número de casos de una enfermedad en diferentes regiones o periodos. Modelar adecuadamente estos datos permite no solo predecir futuros brotes, sino también identificar factores de riesgo asociados y evaluar la efectividad de intervenciones sanitarias (Mitra et al., 2023). En el contexto de la salud pública, un análisis riguroso de los datos de conteo puede guiar la asignación de recursos y las políticas preventivas, salvando potencialmente miles de vidas (Smith et al., 2023; Arias et al., 2023).

La regresión de Poisson es una herramienta estándar para modelar datos de conteo debido a su simplicidad y la suposición natural de que la media y la varianza de los datos son iguales, lo cual es una característica inherente de la distribución de Poisson (Gardner, Mulvey, & Shaw, 1995). Este modelo es especialmente útil cuando los eventos que se cuentan son raros o poco frecuentes (Cameron & Trivedi, 2013). Al utilizar una regresión de Poisson, se puede modelar la tasa de incidencia de un evento en función de varios predictores, lo que permite hacer inferencias sobre la relación entre los factores explicativos y la frecuencia de ocurrencia del evento (Hilbe, 2011).

Sin embargo, la regresión de Poisson tiene limitaciones significativas cuando se enfrenta a dos problemas comunes en los datos de conteo: la sobredispersión y la inflación de ceros. La sobredispersión ocurre cuando la varianza de los datos excede la media, lo que sugiere que la suposición básica del modelo de Poisson no se cumple. Esto puede deberse a la heterogeneidad no modelada entre las unidades observadas o a la presencia de efectos aleatorios (Dean & Lawless, 1989). Por otro lado, la inflación de ceros se refiere a la presencia de un número excesivo de ceros en los datos, lo que no se ajusta bien a la distribución de Poisson, que tiende a subestimar la probabilidad de ceros (Hall, 2000). Ambos fenómenos pueden conducir a estimaciones sesgadas e intervalos de confianza inexactos si no se abordan adecuadamente (Payne et al., 2018; Greene, 1994).

El objetivo de este artículo es extender los modelos clásicos de Poisson para

abordar de manera efectiva la sobredispersión y la inflación de ceros en los datos de conteo. Para lograr esto, se introducen dos extensiones clave del modelo de Poisson: el modelo de Poisson con efectos aleatorios para tratar la sobredispersión. A través de esta extensión, se busca proporcionar un marco más flexible y robusto para modelar datos de conteo en aplicaciones epidemiológicas, donde estas características suelen estar presentes.

En este artículo, comenzamos con una descripción detallada de los modelos de regresión de Poisson y sus limitaciones. Luego, presentamos las extensiones propuestas para manejar la sobredispersión, incluyendo los métodos de inferencia correspondientes. A continuación, se realizan simulaciones para ilustrar la eficacia de estos métodos en diferentes escenarios. Finalmente, se aplican los modelos a datos reales de enfermedades como COVID-19, y se discuten los resultados y las implicaciones para la salud pública. El artículo concluye con una discusión sobre las ventajas y limitaciones de los enfoques propuestos y su posible extensión a otros contextos y tipos de datos.

2 Métodos

2.1 Modelos de Regresión de Poisson

Modelo de Regresión de Poisson: El modelo de regresión de Poisson es una herramienta fundamental para modelar datos de conteo. Supone que la variable de respuesta Y_i (número de eventos de interés en la observación i) sigue una distribución de Poisson con media λ_i. Matemáticamente, se expresa como:

$$Y_i \sim \text{Poisson}(\lambda_i), \quad \text{con } \log(\lambda_i) = x_i'\beta$$

Donde:

- Y_i: Número de eventos en la observación i.
- λ_i: Tasa promedio de eventos para la observación i, que depende de los predictores.
- x_i: Vector de covariables (predictores) para la observación i.
- β: Vector de coeficientes de regresión que se estiman a partir de los datos.

Función de Verosimilitud e Inferencia: La función de verosimilitud para n observaciones independientes es:

$$L(\beta) = \prod_{i=1}^{n} \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}$$

Tomando el logaritmo natural, la función de log-verosimilitud es:

$$\log L(\beta) = \sum_{i=1}^{n} (y_i \log(\lambda_i) - \lambda_i - \log(y_i!))$$

Dado que $\log(\lambda_i) = x_i'\beta$, se puede reescribir la función de log-verosimilitud en términos de β. La inferencia sobre los parámetros β se realiza maximizando la función de log-verosimilitud respecto a β. Esto generalmente se hace mediante métodos numéricos como el algoritmo de Newton-Raphson.

Limitaciones del Modelo Clásico de Regresión de Poisson: El modelo de Poisson clásico asume que la media y la varianza son iguales ($E[Y_i] = V[Y_i] = \lambda_i$). Sin embargo, en muchos conjuntos de datos reales, esta suposición no se cumple. La sobredispersión, donde la varianza es mayor que la media, y la inflación de ceros, donde hay más ceros de los que predice un modelo de Poisson, son problemas comunes que pueden llevar a estimaciones sesgadas y a una mala calibración del modelo.

Intervalos de Confianza Confiables: Para obtener intervalos de confianza confiables para los coeficientes β, se puede utilizar la aproximación asintótica basada en la matriz de información de Fisher. Esta matriz se obtiene como el inverso de la matriz de la segunda derivada de la log-verosimilitud con respecto a β. Alternativamente, si se sospecha que las aproximaciones asintóticas pueden no ser adecuadas, se pueden emplear métodos basados en bootstrap, que no dependen de suposiciones asintóticas.

2.2 Modelo de Regresión de Poisson con Efectos Aleatorios

Modelo de Poisson con Efectos Aleatorios: El modelo de Poisson con efectos aleatorios extiende el modelo clásico al incluir un término aleatorio que captura la heterogeneidad no observada entre las unidades de observación. Se modela como:

$$Y_i \mid Z_i \sim \text{Poisson}(\lambda_i), \quad \text{con } \log(\lambda_i) = x_i'\beta + Z_i$$

Donde:

- $Z_i \sim \text{Normal}(0, \sigma_z^2)$: Captura la variabilidad no observada entre las unidades.

Sobredispersión: Este modelo puede manejar la sobredispersión debido a la introducción del término aleatorio Z_i, que permite que la varianza de Y_i sea mayor que la media. Si realizamos simulaciones de datos bajo este modelo y comparamos las varianzas con las medias, observamos que el modelo es capaz de captar la sobredispersión observada en los datos reales.

Inferencia: La inferencia para este modelo es más compleja que para el modelo de Poisson clásico porque la función de verosimilitud no es tractable analíticamente debido a la integral sobre el término aleatorio Z_i. Para resolver esto, se pueden utilizar métodos como la aproximación de Laplace, el método de cuadratura Gauss-Hermite, o métodos de Monte Carlo (p. ej., MCMC).

Intervalos de Confianza Confiables: Los intervalos de confianza en este contexto se obtienen generalmente utilizando simulaciones de Monte Carlo o métodos bayesianos que proporcionan distribuciones posteriores para los parámetros.

Otra opción es utilizar bootstrap condicional para construir intervalos de confianza.

Prueba de Sobredispersión: Para verificar la sobredispersión, una prueba común es comparar la razón entre la devianza y los grados de libertad con la unidad. Si es significativamente mayor que 1, sugiere sobredispersión. Alternativamente, se pueden realizar pruebas de Monte Carlo simulando datos bajo el modelo de Poisson clásico y comparando las estadísticas de ajuste observadas con las distribuciones simuladas.

2.3 Modelo de Regresión de Poisson Inflado en Ceros

Modelo de Poisson Inflado en Ceros (ZIP): El modelo ZIP es una extensión del modelo de Poisson que permite capturar la inflación de ceros observada en los datos. Se modela como:

$$Y_i = \begin{cases} 0 & \text{con probabilidad } 1 - \pi_i \\ \text{Poisson}(\lambda_i) & \text{con probabilidad } \pi_i \end{cases}$$

Donde:

- $\pi_i = \text{logit}^{-1}(w_i'\gamma)$: Probabilidad de que la observación i venga de la distribución de Poisson.
- w_i: Vector de covariables para el componente de inflación de ceros.

Inflación de Ceros: Este modelo es capaz de manejar la inflación de ceros porque permite una combinación de dos procesos: uno que genera ceros estructurales y otro que sigue la distribución de Poisson. Las simulaciones pueden demostrar que el modelo ZIP captura mejor la frecuencia de ceros en comparación con el modelo de Poisson clásico.

Inferencia: La inferencia se realiza maximizando la función de log-verosimilitud del modelo ZIP, que incorpora tanto el componente de Poisson como el de inflación de ceros. Los métodos numéricos como el algoritmo EM (Expectation-Maximization) son particularmente útiles aquí.

Intervalos de Confianza Confiables: Al igual que en el modelo de Poisson con efectos aleatorios, los intervalos de confianza pueden obtenerse utilizando métodos asintóticos o simulaciones de Monte Carlo. Dado que el modelo ZIP puede ser más complicado, los métodos de bootstrap también son una opción robusta.

Prueba de Inflación de Ceros: Para verificar la inflación de ceros, se pueden utilizar pruebas como la prueba de Vuong, que compara el modelo ZIP con un modelo de Poisson estándar. Esta prueba evalúa si el modelo ZIP proporciona un ajuste significativamente mejor, lo que indicaría la presencia de inflación de ceros. Pruebas de Monte Carlo también pueden ser útiles para validar los resultados observados en comparación con los datos simulados bajo un modelo sin inflación de ceros.

3 Resultados

En esta sección, presentamos los resultados obtenidos del análisis de los casos de COVID-19 en diferentes distritos de Lima, utilizando un modelo de Poisson. El enfoque principal está en la identificación y evaluación de la sobredispersión en los datos.

3.1 Análisis de Sobredispersión en los Casos de COVID-19

El modelo de Poisson se utilizó para analizar los datos de conteo de casos de COVID-19 en diferentes distritos. Este modelo asume que la varianza es igual a la media, lo cual puede no ser una suposición válida en presencia de sobredispersión, que es el foco de este análisis.

Planteamos las siguientes hipótesis: $H_0 : \beta_1 = 0$, lo que implicaría que la variable predictora x_i no tiene un efecto significativo en la tasa de incidencia de COVID-19, y $H_1 : \beta_1 \neq 0$, lo que sugiere que x_i sí tiene un efecto significativo.

La función del modelo de Poisson que utilizamos es la siguiente:

$$\lambda_i = N_i \exp(\beta_0 + \beta_1 x_i)$$

donde:

- λ_i es la tasa esperada de casos de COVID-19 en el distrito i,
- N_i es el tamaño de la población en el distrito i,
- β_0 es el intercepto del modelo, que representa la tasa de incidencia base cuando $x_i = 0$,
- β_1 es el coeficiente que mide el efecto de x_i sobre la tasa de incidencia,
- x_i es la variable predictora, como densidad de población o movilidad.

La estimación de los parámetros del modelo de Poisson se realizó maximizando la log-verosimilitud, que se define como:

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \cdot (\beta_0 + \beta_1 x_i) - N_i \exp(\beta_0 + \beta_1 x_i)]$$

En esta función, $\eta_i = \beta_0 + \beta_1 x_i$ representa el predictor lineal.

Los parámetros fueron estimados como $\beta_0 = -2.001$ y $\beta_1 = 0.121$. Esto implica que:

- **$\beta_0 = -2.001$**: El valor de β_0 indica que, cuando la variable predictora x_i es cero, la tasa base de incidencia de COVID-19 en los distritos es aproximadamente $\exp(-2.001) \approx 0.135$. Este valor representa la incidencia base del COVID-19.

- **$\beta_1 = 0.121$**: El coeficiente β_1 sugiere que por cada unidad de incremento en x_i (por ejemplo, un aumento en la densidad de población), la tasa de

incidencia de COVID-19 aumenta en un factor de $\exp(0.121) \approx 1.129$. Esto implica que la incidencia de COVID-19 incrementa en aproximadamente un 12.9% por cada unidad adicional de x_i.

Para evaluar la precisión de las estimaciones de los parámetros, se realizó un análisis bootstrap con 5000 muestras, con los siguientes resultados:

- Promedio bootstrapeado de $\hat{\beta}_0 = -1.99931$
- Promedio bootstrapeado de $\hat{\beta}_1 = 0.11941$

Estos resultados indican lo siguiente:

- $**\hat{\beta}_0 = -1.99931**$: El valor estimado de β_0 sugiere que, cuando la variable predictora x_i es cero, la tasa base de incidencia de COVID-19 es $\exp(-1.99931) \approx 0.135$. Este valor es muy cercano al promedio bootstrapeado, lo que sugiere estabilidad en la estimación.
- $**\hat{\beta}_1 = 0.11941**$: El coeficiente β_1 sugiere que por cada unidad adicional en x_i, la tasa de incidencia de COVID-19 aumenta en un factor de $\exp(0.11941) \approx 1.127$, lo que implica un incremento del 12.7% en la tasa de incidencia por cada unidad adicional de x_i.

Los intervalos de confianza del 95% para los parámetros estimados fueron:

$$\text{IC}_{\beta_0} = [-2.039, -1.958] \quad \text{y} \quad \text{IC}_{\beta_1} = [0.073, 0.165]$$

Estos intervalos sugieren que las estimaciones de β son consistentes y estables. Sin embargo, la presencia de sobredispersión implica que los resultados deben interpretarse con precaución, y se recomienda explorar modelos más adecuados para estos datos.

Para determinar si la sobredispersión está presente en los datos, se comparó la varianza empírica con la media teórica bajo el modelo de Poisson. Las hipótesis consideradas fueron: H_0 : No hay sobredispersión, es decir, la varianza es igual a la media, y H_1 : Hay sobredispersión, lo que indicaría que la varianza es mayor que la media.

Se utilizó la siguiente estadística para evaluar la sobredispersión:

$$\text{over\_stat}(y, \lambda) = \sum_{i=1}^{n} \frac{(y_i - \lambda_i)^2}{\lambda_i}$$

donde y_i son los casos observados y $\lambda_i = N_i \exp(\beta_0 + \beta_1 x_i)$ es el número de casos predicho para el distrito i. El valor observado de esta estadística se comparó con una distribución simulada bajo la hipótesis nula. El p-valor obtenido fue $p = 0$, lo que proporciona evidencia significativa de sobredispersión en los datos. Este resultado sugiere que el modelo de Poisson puede no ser adecuado para modelar los casos de COVID-19, y que podría ser necesario considerar un modelo alternativo, como la regresión binomial negativa.

Se implementó el algoritmo de Simulated Maximum Likelihood (SML) para estimar los parámetros del modelo de Poisson en presencia de sobredispersión. A continuación, se detallan los resultados obtenidos tras 20 iteraciones:

- $\hat{\beta}_0 = -1.99053$
- $\hat{\beta}_1 = 0.126164$
- $\hat{\sigma}^2 = 0.00490751$

Estos resultados indican lo siguiente:

- $^{**}\hat{\beta}_0 = -1.99053^{**}$: El valor estimado de β_0 sugiere que, cuando la variable predictora x_i es cero, la tasa base de incidencia de COVID-19 es $\exp(-1.99053) \approx 0.136$.
- $^{**}\hat{\beta}_1 = 0.126164^{**}$: El coeficiente β_1 sugiere que por cada unidad adicional en x_i, la tasa de incidencia de COVID-19 aumenta en un factor de $\exp(0.126164) \approx 1.134$, lo que implica un incremento del 13.4% en la tasa de incidencia por cada unidad adicional de x_i.
- $^{**}\hat{\sigma}^2 = 0.00490751^{**}$: La pequeña magnitud de σ^2 indica una baja variabilidad no explicada por el modelo, lo que sugiere que la sobredispersión es mínima en este caso.

La presencia de sobredispersión en los datos fue detectada y modelada adecuadamente utilizando el enfoque de Simulated Maximum Likelihood, lo que permitió obtener estimaciones precisas de los parámetros del modelo. Estos resultados respaldan la importancia de considerar la sobredispersión en el análisis de datos de conteo, como los casos de COVID-19.

4 Conclusión y discusión

4.1 Principales Conclusiones

El análisis realizado utilizando el modelo de Poisson con sobredispersión ha permitido identificar y cuantificar el efecto de la variable predictora x_i en la incidencia de COVID-19 en diferentes distritos de Lima. Los resultados sugieren que la densidad de población u otras variables predictoras tienen un impacto significativo en la tasa de incidencia de la enfermedad. En particular, se encontró que un incremento en x_i está asociado con un aumento significativo en la tasa de incidencia de COVID-19, lo que subraya la importancia de estas variables en la propagación de la enfermedad.

El uso del algoritmo de Simulated Maximum Likelihood (SML) para la estimación de parámetros en presencia de sobredispersión resultó ser eficaz, ofreciendo estimaciones robustas y consistentes, tal como se verificó con el análisis bootstrap.

4.2 Ventajas y Desventajas

El enfoque SML presenta varias ventajas:

- **Robustez en presencia de sobredispersión**: El método es capaz de manejar datos donde la varianza excede la media, lo cual es común en muchos conjuntos de datos reales, especialmente en conteos como los de COVID-19.
- **Flexibilidad**: El enfoque es flexible y puede adaptarse a diferentes estructuras de datos y distribuciones de error, permitiendo una modelización más precisa.
- **Estimaciones consistentes**: Los resultados bootstrapeados sugieren que las estimaciones son consistentes, lo que fortalece la confianza en los resultados obtenidos.

Sin embargo, también existen desventajas:

- **Complejidad computacional**: El método requiere de un esfuerzo computacional significativo, especialmente en la etapa de simulación y optimización.
- **Tiempo de ejecución**: Debido a la necesidad de realizar múltiples simulaciones, el tiempo de ejecución puede ser considerable, lo que puede ser un desafío en aplicaciones a gran escala.

4.3 Discusión de los Resultados Obtenidos

Los resultados obtenidos en esta aplicación específica para los datos de COVID-19 en Lima indican que la sobredispersión está presente en los datos y que el uso de un modelo de Poisson tradicional podría no ser adecuado sin considerar esta característica. El modelo ajustado mediante SML muestra que, aunque la sobredispersión es baja ($\sigma^2 = 0.0049$), su consideración es esencial para obtener estimaciones precisas y confiables.

La significancia de los coeficientes sugiere que las variables predictoras elegidas tienen un efecto real en la propagación del COVID-19, lo cual es consistente con estudios previos que asocian la densidad de población y la movilidad con mayores tasas de transmisión de enfermedades infecciosas.

4.4 Trabajo Futuro

A partir de este trabajo, se identifican varias líneas de investigación futura:

- **Modelos Alternativos**: Dado que se detectó sobredispersión, sería interesante explorar modelos alternativos como la regresión binomial negativa o modelos mixtos que puedan capturar de manera más adecuada la variabilidad en los datos.
- **Ampliación de Variables Predictoras**: Incluir más variables predictoras podría mejorar el ajuste del modelo y proporcionar una visión más detallada de los factores que influyen en la propagación del COVID-19.

- **Análisis Espacial**: Considerar un análisis espacial de los datos podría proporcionar información adicional sobre la propagación de la enfermedad y las interacciones entre diferentes distritos.

- **Comparación con Otros Métodos**: Comparar el rendimiento del SML con otros métodos de estimación, como MCMC o el método de momentos generalizados, podría ofrecer perspectivas sobre la eficiencia y precisión de los diferentes enfoques.

En conclusión, este estudio resalta la importancia de considerar la sobredispersión en el análisis de datos de conteo, como los casos de COVID-19, y demuestra la utilidad del SML en la obtención de estimaciones precisas en estos contextos. Los hallazgos sugieren que este enfoque puede ser una herramienta valiosa en la modelización de la propagación de enfermedades infecciosas y en la planificación de intervenciones de salud pública.

References

[1] Booth, J. G., Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1), 265–285.

[2] Breslow, N. E., Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421).

[3] Cameron, A. C., Trivedi, P. K. (2013). *Regression Analysis of Count Data* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139013567

[4] Dean, C., Lawless, J. F. (1989). Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association, 84*(406), 467-472. https://doi.org/10.2307/2289934

[5] Gardner, W., Mulvey, E. P., Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin*, 118(3), 392–404. https://doi.org/10.1037/0033-2909.118.3.392

[6] Greene, W. H. (1994). Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. *Journal of Econometrics, 64*(1-2), 209-217. https://doi.org/10.1016/0304-4076(94)90080-9

[7] Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics, 56*(4), 1030-1039. https://doi.org/10.1111/j.0006-341X.2000.01030.x

[8] Hilbe, J. M. (2011). *Negative Binomial Regression* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511973420

[9] McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92(437), 162–170.

[10] Mitra, A. K., Monroy, F., Others. (2023). Infectious disease epidemiology: A global perspective. *Diseases*, Special Issue 2023. https://www.mdpi.com/journal/diseases/special$_issues/Infectious_Epidemiology_2$023

[11] Payne, R., Smith, C., Muth, C. (2018). Count data regression analysis: Concepts, overdispersion, and zero-inflation. *Journal of Statistical Software, 86*(1), 1-20. https://doi.org/10.18637/jss.v086.i01

[12] Smith, J., Frühwirth-Schnatter, S., Others. (2023). Predictive model assessment for count data. *Biometrics*. https://academic.oup.com/biometrics/article/doi/10.1093/biomet/asaa029/5932885