Trabajo de Fin de Grado

Grado en Matemáticas

Título del trabajo:

Modelos Lineales Generalizados:
Modelos de Regresión Poisson cero-inflados

Nombre del estudiante:

Jose Maria Buades Rubio

Centro asociado al que pertenece: Illes Balears
Nombre del tutor del trabajo: Prof. Dr. Alfonso García Pérez

Curso académico 2024-2025

A Daniel, Ariadna y Sandra.

Índice general

Resumen

En este trabajo se estudian los modelos lineales generalizados (GLM) con especial énfasis en su aplicación al análisis de datos de conteo con exceso de ceros, utilizando como caso de estudio la relación entre la concentración de arsénico en pozos y la incidencia de muertes por cáncer. Se revisan los fundamentos teóricos de los modelos Poisson, binomial negativo y sus versiones cero infladas (ZIP y ZINB), así como las herramientas de comparación de modelos como AIC, BIC y el test de razón de verosimilitudes.

A través de un análisis estadístico aplicado con R, se explora el comportamiento de los modelos bajo distintas configuraciones y subconjuntos de datos. El modelo Poisson ajustado a todos los datos resulta inadecuado por sobredispersión y exceso de ceros. Al restringir el análisis a datos con exposición positiva y emplear modelos alternativos, se obtienen resultados más consistentes. En este contexto, los modelos cero inflados permiten distinguir entre ceros estructurales —observaciones en las que el evento no puede ocurrir— y ceros aleatorios —propios del proceso de conteo—. Esta distinción se implementa mediante una combinación de un modelo binomial para la ocurrencia de ceros y un modelo de conteo para el resto de valores, ofreciendo así una solución flexible y realista para el análisis de datos con acumulación de ceros.

Abstract

This work explores generalized linear models (GLMs), with a particular focus on their application to count data with excess zeros, using as a case study the relationship between arsenic concentration in water wells and cancer mortality. The theoretical foundations of the Poisson and negative binomial models are reviewed, as well as their zero-inflated extensions (ZIP and ZINB), along with model comparison tools such as AIC, BIC, and likelihood ratio tests.

Through an applied statistical analysis using R, the performance of different models is evaluated under various data configurations and subsets. The Poisson model fitted to the complete dataset proves inadequate due to overdispersion and an excessive number of zeros. When the analysis is restricted to exposed data and alternative models are employed, more consistent results are obtained. In this context, zero-inflated models allow for a distinction between structural zeros—observations in which the event cannot occur—and random zeros—arising naturally from the count process. This is achieved by combining a binomial model for the occurrence of zeros with a count model for the non-zero observations, providing a flexible and realistic solution for analyzing data with an accumulation of zeros.

Capítulo 1

Introducción

1.1. Evolución histórica

La regresión lineal es una de las técnicas estadísticas más utilizadas en la actualidad. Sus raíces se remontan al siglo XIX, y su aparición no fue el resultado de un descubrimiento aislado, sino de una evolución gradual motivada por problemas científicos concretos y la necesidad de establecer relaciones entre variables. Áreas de conocimiento dispares, como astronomía o antropología son el origen de la regresión lineal.

Su estudio se inició antes de que se le acuñara el nombre: cómo ajustar una función matemática a un conjunto de datos empíricos con errores de medida. Esta necesidad era particularmente crítica en la astronomía, donde los astrónomos trataban de describir y predecir el movimiento de los cuerpos celestes a partir de observaciones imperfectas.

En el siglo XVII, astrónomos como Tycho Brahe y Johannes Kepler registraban observaciones sistemáticas del firmamento para modelar el movimiento planeatrios. Kepler consiguió ajustar órbitas elípticas a los datos planetarios usando un enfoque por prueba y error. Aunque su método no era formalmente una regresión, anticipaba la idea de buscar la "mejor curva" que explique los datos observados.

En el siglo XVIII surgieron con más claridad los problemas relacionados con la estimación a partir de observaciones ruidosas. Matemáticos como Leonhard Euler, Joseph-Louis Lagrange y Daniel Bernoulli debatían cómo encontrar estimaciones óptimas de parámetros físicos cuando los datos contenían error. El objetivo era ajustar modelos matemáticos a los datos observados, típicamente líneas rectas o curvas suaves, a fin de predecir valores futuros o inferir cantidades físicas medibles.

El problema se solucionó eligiendo la estimación que minimizaba alguna función de error, la diferencia entre el valor observado y el predicho. Algunos proponían minimizar la suma de errores absolutos, otros la suma de los cuadrados de los errores. Sin embargo, no existía una justificación formal clara para ninguna de estas técnicas, siendo un problema abierto. El punto de inflexión llegó con la formalización del método de los mínimos cuadrados, estrechamente relacionado con el problema de determinar órbitas de cometas y planetas. Los astrónomos necesitaban estimar trayectorias a partir de observaciones dispersas y ruidosas de los objetos celestes. Fue el francés Adrien-Marie Legendre quien publicó en 1805 un breve trabajo donde

introdujo, de forma clara y sistemática, el método de los mínimos cuadrados. Este método consiste en encontrar los parámetros de una función (normalmente una recta o parábola) que minimizan la suma de los cuadrados de las diferencias entre los valores observados y los predichos por el modelo:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Legendre lo aplicó al problema de determinar la órbita de cometas como el de 1801, con datos imprecisos y mal distribuidos en el tiempo. La elección de la minimización cuadrática tenía ventajas matemáticas (facilidad de derivación) y era factible llevar a cabo los cálculos con los métodos manuales de la época. Unos años más tarde, Carl Friedrich Gauss afirmó haber usado ya ese método desde 1795, aunque no lo publicó hasta 1809. En su libro "Theoria motus corporum coelestium in sectionibus conicis solem ambientium", Gauss no solo describe el método, sino que ofrece una justificación probabilística: si los errores de las observaciones siguen una distribución normal, entonces el estimador de mínimos cuadrados es el más probable (máxima verosimilitud).

Este enfoque marcó un punto de partida clave: la estadística empezaba a independizarse de la matemática pura y de la física para convertirse en un campo propio, con un fundamento probabilístico que lo respaldaba. En el siglo XIX, el método de mínimos cuadrados se expandió más allá de la astronomía. La geodesia, que busca medir con precisión la forma y dimensiones de la Tierra, se benefició ampliamente de estas técnicas. La construcción de redes de triangulación requería ajustar modelos geométricos a medidas de distancias y ángulos, afectadas por errores instrumentales. Aquí también se adoptó el criterio de mínimos cuadrados como el más eficaz.

Otro campo temprano fue la economía política, donde comenzaron a surgir intentos de cuantificar relaciones entre variables económicas. Aunque estos trabajos no usaban aún la formalización matemática de la regresión, ya apuntaban al tipo de análisis que se haría habitual más tarde. Finalmente, el método se impuso en las diferentes áreas de conocimiento como un método fiable para obtener predicciones a partir de datos ruidosos.

Aunque el método de los mínimos cuadrados era conocido y aplicado en contextos físicos, el concepto de regresión como relación estadística entre variables surge posteriormente, en un contexto completamente diferente: la biología. El término "regresión" fue introducido por el científico británico Francis Galton en la década de 1880. Galton investigaba la relación entre la estatura de padres e hijos, y encontró que los hijos de padres muy altos tendían a ser más bajos que ellos, mientras que los hijos de padres muy bajos tendían a ser más altos. Es decir, "regresaban" hacia la media poblacional. Esta observación la denominó "regresión hacia la media" (regression towards mediocrity). En su artículo de 1886 "Regression Towards Mediocrity in Hereditary Stature", Galton propuso la relación lineal entre la estatura media de los padres y la de sus hijos. Utilizó gráficos con líneas rectas ajustadas visualmente, aunque no formuló explícitamente el modelo matemático como lo haríamos hoy.

Los modelos lineales generalizados (GLM) surgieron en 1972. Debido a la necesidad de extender la regresión lineal a situaciones donde la variable a predecir no seguía una distribución normal. Esto ocurre en el caso de datos binarios (puede tomar dos valores), de conteo o de proporciones. Hasta ese momento, los modelos para estos casos eran tratados de forma aislada, sin una estructura común que los unificara. Esta limitación motivó a los estadísticos británicos John Nelder y Robert Wedderburn a desarrollar en un marco estadístico general que permitiera aplicar principios comunes a una gran variedad de situaciones, ofreciendo mayor coherencia y flexibilidad al análisis de datos en diversas disciplinas como medicina, biología, economía y ciencias sociales. Su trabajo de abstracción en un marco común basándose en la familia de funciones exponencial Nelder and Wedderburn (1972).

El modelo GLM no encaja bien en problemas de conteo donde se tiene un número no explicable de ceros. En 1996 los estadísticos Lambert y Mullahy propusieron los primeros modelos cero inflados, combinando un componente binario que modela la probabilidad de generar ceros "estructurales" con un componente de conteo (como Poisson o binomial negativa) que se encarga del resto de la distribución. Esta formulación permitió capturar tanto la frecuencia excesiva de ceros como la variabilidad de los conteos positivos, mejorando sustancialmente el ajuste y la interpretación de este tipo de datos. En dicho artículo utilizó como caso de estudio la aparición de defectos en procesos de manufactura, donde muchos artículos no presentaban ningún defecto, y los modelos clásicos (como Poisson estándar) no lograban ajustarse correctamente. El modelo propuesto combinaba una distribución binaria para modelar la presencia o ausencia estructural del evento con una distribución de Poisson para el conteo condicional, marcando un hito en el análisis estadístico de datos dispersos. Lambert, Diane (1992) Lambert (1992).

Hoy en día, la regresión sigue siendo una herramienta muy utilizada. El auge de la inteligencia artificial, acontecido por los avances hardware en calculo paralelo homogéneo, ha provocado el uso de modelos extremadamente complejos, pero dichos modelos son no explicables. La regresión es un parte clave en la explicabilidad en IA (XAI). Cuando se necesita comprender cómo influye cada variable en una predicción, se usan modelos lineales locales como LIME Ribeiro et al. (2016) o SHAP Lundberg and Lee (2017) para aproximar la salida de modelos complejos de forma interpretable.

En la era del aprendizaje automático, donde muchos modelos son "cajas negras", la regresión sigue destacando por su transparencia. Saber que un aumento de una unidad en X aumenta entre 3.4 y 3.6 unidades la respuesta Y, con un 95 % de confianza, es algo que ningún árbol de decisión o red neuronal profunda puede proporcionar de forma tan directa y fiable. Esto hace que la regresión siga siendo preferida en sectores regulados o donde la interpretación es crítica, como en salud, justicia, educación o banca.

1.2. Objetivos

El objetivo de este trabajo es explorar en profundidad los modelos de regresión, centrándose especialmente en el estudio de los modelos lineales generalizados (GLM) y, en particular, en sus extensiones para datos de conteo con exceso de ceros, como los modelos cero inflados. Como aplicación práctica de los conceptos desarrollados, se abordará un problema real relacionado con la evaluación del impacto de la concentración de arsénico en pozos sobre el número de casos de cáncer detectados (Chen et al. (1985)), utilizando técnicas estadísticas adecuadas para este tipo de datos.

1.3. Estructura del documento

Este Trabajo de Fin de Grado (TFG) se organiza en varios capítulos que siguen una progresión lógica desde los fundamentos teóricos hasta el análisis aplicado y la interpretación de resultados. El capitulo 2 presenta el modelo de regresión lineal. El capítulo 3 introduce el marco teórico de los modelos lineales generalizados (GLM), abordando su formulación, propiedades y motivación desde la familia exponencial. A continuación, en el capítulo 4 se presenta la necesidad de extender los GLM en presencia de sobredispersión, especialmente cuando ésta se manifiesta como un exceso de ceros, dando lugar al estudio de los modelos de Poisson y binomial negativo cero inflados.

El capítulo 5 presenta el problema práctico a resolver, y el capítulo 6 se dedica a describir el entorno de trabajo utilizado, incluyendo el lenguaje R y las principales librerías empleadas para el ajuste, diagnóstico y comparación de modelos.

El capítulo 7 inicia con una presentación del conjunto de datos, analizado y realiza una inspección preliminar de las variables y su distribución. Posteriormente, se desarrolla el análisis estadístico completo, incluyendo el ajuste de diversos modelos (Poisson, ZIP, binomial negativo y ZINB), su diagnóstico, la comparación mediante criterios como AIC, BIC y `lrtest`, y la selección del modelo más adecuado.

Finalmente, el capítulo 8 recoge las conclusiones.

Capítulo 2

Modelo de regresión lineal

La regresión lineal constituye uno de los métodos fundamentales del análisis estadístico y sirve como punto de partida para el modelo lineal generalizado (GLM). En este capítulo se presentarán en detalle dos de sus formas más utilizadas: la regresión lineal simple (RLS), que permite estudiar la relación lineal entre una variable dependiente y una sola variable independiente, y la regresión lineal múltiple (RLM), que generaliza esta idea para incorporar varias variables explicativas. Ambas técnicas son ampliamente aplicadas en contextos científicos, sociales y económicos, tanto para realizar predicciones como para analizar el impacto de distintos factores sobre un fenómeno de interés. Se abordarán sus formulaciones matemáticas, los supuestos sobre los que se sustentan, y los métodos de estimación de los parámetros.

2.1. Modelo de regresión lineal simple

2.1.1. Hipótesis

Llamaremos y a la variable respuesta o dependiente, x a la variable explicativa o independiente que se supone conocida al observar y, la observación se ve influenciada por una perturbación u. El modelo es el siguiente:
$y_i = \beta_0 + \beta_1 x_i + u_i$, donde y_i y u_i son variables aleatorias, x_i es una variable con valores conocidos y β_0 y β_1 son parámetros desconocidos que se desea estimar. Y las hipótesis son:

1) $\mathrm{E}[u_i] = 0$ la perturbación tiene esperanza nula.
2) $\mathrm{Var}(u_i) = \sigma^2$ la perturbación es homocedástica.
3) La perturbación tiene distribución normal.
4) $\mathrm{E}[u_i u_j] = 0$ las perturbaciones son independientes entre sí.

El objetivo es estimar β_0, β_1 y σ^2, y obtener la recta de regresión:

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_0 x_i$$

2.1.2. Estimación de los parámetros

Para estimar los parámetros se puede utilizar al método de máxima verosimilitud. Siendo la función de densidad de una observación

$$f(y_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2\sigma^2}(y_i-\beta_0-\beta_1 x_i)^2}$$

Debido a la independiencia de las observaciones, y tomando logaritmos la función soporte para la función se obtiene

$$L(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum (y_i - \beta_0 - \beta_1 x_i)^2$$

derivando respecto a β_0 e igualando a cero se obtiene

$$\sum e_i = 0$$

donde $e_i = y_i - \hat{\beta}_0 - \hat{\beta}_0 x_i$. La suma de los e_i debe ser cero, estos valores se llaman residuos. Derivando respecto a β_1 e igualando a a cero obtenemos $\frac{\partial L}{\partial \beta_1} = \sum (y_i - \beta_0 - \beta_1 x_i) x_i = 0$, que puede expresarse como

$$\sum e_i x_i = 0$$

que se interpreta que los residuos deben estar incorrelados con la variable x. De lo contrario, se podría usar esa información para obtener una mejor estimación.
Así pues, la estimación de los parámetros es la siguiente:

$$\hat{\beta}_1 = \frac{\mathrm{Cov}(x, y)}{s_x^2}$$
$$\hat{\beta}_0 = \hat{\beta}_1 \overline{x} - \overline{y}$$
$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n}$$

donde $\overline{x} = \sum x_i/n$ y $\overline{y} = \sum y_i/n$.

Hay que destacar que debido a que los parámetros β_0 y β_1 aparecen únicamente en el exponente en la función de verosimilitud, este método coindice con el método de mínimos cuadrados:

$$M = \sum (y_i - \beta_0 - \beta_1 x_i)^2 = \sum e_i^2$$

No hay que confundir los residuos e_i, que son las diferencias entre las observaciones y las estimaciones, con u_i que son las perturbaciones habidas en la medición.

2.1.3. Diagnosis del modelo

Una vez estimado los parámetros del modelo es necesario estudiar que se cumplen las hipótesis: linealidad, homocedasticidad, normalidad e independencia. Las dos primeras hipótesis se pueden comprobar previa a la construcción del modelo, mientras que las dos últimas pueden ser estudiadas a partir de los residuos obtenidos. Los residuos también aportan información respecto a la linealidad y la homocedasticidad. Así pues, el estudio de los residuos es clave para validar los resultados obtenidos.
Una forma de analizar estos resultados es graficando los residuos respecto a la variable dependiente x.

2.2. Modelo de regresión lineal múltiple

2.2.1. Hipótesis

En el modelo de regresión lineal múltiple se dispone de varias covariables independientes:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$$

donde y_i es el valor de la respuesta en el elemento i, $x_{1i}, ..., x_{ki}$, los valores de las variables explicativas y cada coeficiente β_j mide el efecto marginal sobre la respuesta de un aumento unitario en x_j cuando el resto de las variables explicativas permanecen constantes. El término u_i es el efecto de todas las variables que afectan a la dependiente y no están incluidas en el modelo. Sobre u_i se supone:

1) Su esperanza es cero.
2) Su varianza es constante σ^2.
3) Las perturbaciones son independientes entre sí.
4) Su distribución es normal.

Estas condiciones son idénticas al caso de regresión lineal simple, además, se imponen las siguientes condiciones:

5) El número de datos disponibles es mayor que $k + 1$.
6) Ninguna de las variables explicativas es combinación lineal exacta de las demás. Es decir, las variables x_j son linealmente independientes.

2.2.2. Estimación de los parámetros

Nuevamente tenemos que el método de máxima verosimilitud equivale al método de mínimos cuadrados, esto es debido a que la distribución de y es una variable aletoria normal debido a las suposiciones realizadas sobre u_i. Tenemos

$$M = \sum (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_k x_{ki})^2$$

Derivando respecto a β_0 e igualando a cero se obtiene

$$\sum e_i = 0, \text{ donde } e_i = y_i - \hat{y}_i$$

y derivando respecto a β_j se obtiene

$$\sum e_i x_{ji} = 0, j = 1, ..., k$$

ecuaciones análogas al caso de RLS. A partir de estos resultados se obtiene, en forma matricial, la ecuación

$$\mathbf{X'Y} = \mathbf{X'X}\hat{\beta}$$

despejando se obtiene

$$\hat{\beta} = \left(\mathbf{X'X}\right)^{-1} \mathbf{X'Y}$$

2.2.3. Correlación

Una medida de ajuste del modelo es el coeficiente de determinación que viene dado por el cociente entre la variabilidad explicada por la regresión y la variabilidad total

$$R^2 = \frac{\text{Variablilidad Explicada (VE)}}{\text{Variabilidad Total (VT)}} = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2}$$

El valor R se le denomina coeficiente de correlación múltiple, y cumple $0 \leq |R| \leq 1$.

El coeficiente de determinación R^2 se emplea con frecuencia para evaluar la calidad de los modelos de regresión, sin embargo, esta práctica puede inducir a errores y está sujeto a críticas. En primer lugar, R^2 tiende a aumentar con la incorporación de nuevas variables explicativas, independientemente de si estas contribuyen significativamente al modelo. Esto implica que su valor puede inflarse de manera artificial al incluir predictores irrelevantes. En segundo lugar, R^2 es altamente dependiente de cómo se estructura el modelo y de la variable dependiente seleccionada. Es posible obtener modelos distintos que predigan de manera similar pero presenten valores diferentes de R^2, lo que pone en duda su fiabilidad como medida comparativa. Como herramienta descriptiva, la utilidad de R^2 está vinculada a la proporción entre k, el número de covariables independientes, y n, el tamaño de la muestra. En ausencia de una relación real entre las covariables independientes y la dependiente, se ha demostrado que el valor esperado de R^2 es $k/(n-1)$. Por esta razón, no es aconsejable utilizar un número elevado de predictores en un modelo cuando el cociente k/n es alto, ya que el incremento de R^2 puede deberse al azar y no a una verdadera asociación entre las variables. En consecuencia, un alto valor de R^2 no garantiza necesariamente una relación significativa entre las variables del modelo.

En base a esto se han propuesto otros valores como el coeficiente de determinación corregido por grados de libertad $\overline{R}^2$

$$\overline{R}^2 = 1 - \frac{\text{Varianza residual}}{\text{Varianza de } y} = 1 - \frac{n-1}{n-k-1} \cdot \frac{\sum e_i^2}{\sum (y_i - \overline{y})^2}$$

2.2.4. Predicción

Supongamos que se desea predecir el valor medio de la variable respuesta Y para un nuevo vector de predictores $\boldsymbol{x}_h = (1, x_{1k}, x_{2k}, \ldots, x_{kh})$. La predicción puntual del valor medio de Y para este nuevo punto es:

$$\mathrm{E}[\hat{y}_h] = \boldsymbol{x}_h' \hat{\beta} = m_h$$

y su varianza es

$$\mathrm{Var}(\hat{y}_h) = \sigma^2 \boldsymbol{x}_h' (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}_h = \sigma^2 h_{hh}$$

El término h_{hh} puede escribirse como $h_{hh} = 1/\hat{n}_h$ que indica el número equivalente de observaciones para estimar m_h. Este valor es mayor en la zona central de los datos y puede ser inferior a 1 en caso de realizar estimaciones fuera del rango de valores usados para estimar el modelo.

2.2.5. Robustez del modelo

El efecto palanca (robustez a priori) es la capacidad de una observación para atraer la ecuación de regresión a dicha observación, este efecto puede medirse como

$$h_{ii} = \frac{1}{n}(1 + \tilde{x_i}' \boldsymbol{S_{xx}^{-1}} \tilde{x_i})$$

donde $\tilde{x}_h = (\tilde{x}_{1h}, ..., \tilde{x}_{kh})'$ no incluye el uno inicial correspondiente a β_0, y S_{xx} es la matriz de varianzas y covarianzas muestral entre las x. Es habitual calcular el número equivalente de observaciones $\hat{n}_i = h_{ii}^{-1}$. Si sucede que $\hat{n}_i < n/(2k+2)$ e implica que tenemos pocos datos para estimar la relación en un punto si el número equivalente de observaciones es menor de la mitad del número de datos por variable.

Pero que una observación sea influyente a priori, no implica que realmente esto suceda. Otra medida de influencia (robustez a posteriori) es

$$D(i) = \frac{r_i^2}{k+1} \left(\frac{h_{ii}}{1 - h_{ii}} \right)$$

donde $r_i = e_i/\hat{s}_R\sqrt{1-h_{ii}}$ es el residuo estandarizado. Este estadístico se conoce como estadístico de Cook.

2.3. Selección de modelos de regresión

En determinadas ocasiones se plantea la situación de incluir o no en el modelo algunas variables, es decir, determinar que conjunto de las covariables independientes utilizar. Mediante la inclusión de todas las variables tendremos un coeficiente de determinación más alto, con el mejor ajuste, pero a cambio, tendremos un modelo con una peor predicción. Es por ello que se utilizan otros criterios para la selección del modelo. Los más utilizados son:

El **criterio de información de Akaike (AIC)** es una herramienta para comparar modelos estadísticos que equilibra el ajuste del modelo con su complejidad. Se define como

$$\text{AIC} = n\log(\hat{\sigma_p}^2) + 2p,$$

donde n es el número de muestras, $\hat{\sigma_p}^2$ es el estimador de MV del modelo y p es el número de parámetros estimados. El AIC penaliza los modelos con mayor número de parámetros para evitar sobreajuste, pero su penalización es moderada, por lo que puede favorecer modelos algo más complejos. Se utiliza para seleccionar el modelo que minimiza el AIC, ya que un valor más bajo indica un mejor equilibrio entre ajuste y parsimonia[1].

El **criterio bayesiano de información (BIC)** es similar al AIC, pero impone una penalización más fuerte por complejidad. Su fórmula es

$$\text{BIC} = n\log(\hat{\sigma_p}^2) + p\log(n),$$

Esta penalización más severa hace que BIC favorezca modelos más simples cuando el tamaño de muestra es grande. Desde una perspectiva bayesiana, el BIC se interpreta como una aproximación al logaritmo del valor predictivo marginal del modelo, por lo que se considera más conservador que el AIC en la selección de modelos.

El C_p **de Mallows** es un criterio diseñado específicamente para modelos de regresión lineal. Se basa en comparar la suma de cuadrados de los errores del modelo reducido con la del modelo completo, considerando el número de parámetros utilizados. Se define como

$$C_p = \frac{\text{SSE}_p}{\hat{\sigma}^2} - n + 2(p+1),$$

donde SSE_p es la suma de errores cuadráticos del modelo con p parámetros, $\hat{\sigma}^2$ es una estimación del error del modelo completo, y n es el tamaño muestral. Un modelo adecuado debe tener un valor de C_p cercano a p, lo que indica que no hay sesgo y la varianza se mantiene baja. Este criterio es útil en la selección entre subconjuntos de variables explicativas en regresión.

[1] Principio de utilizar el modelo más simple y conciso que represente adecuadamente los datos, minimizando la cantidad de variables o parámetros.

Capítulo 3

Modelos Lineales Generalizados

Un modelo lineal generalizado (GLM) se define especificando dos componentes. Primero, la variable respuesta debe pertenecer a la familia de distribuciones exponenciales, y segundo, la función de enlace (*link*) que describe cómo se relacionan la media de la respuesta y una combinación lineal de los predictores. A continuación se detalla la familia de exponenciales, posteriormente las funciones *link*. Una vez definido el modelo GLM se detalla como se obtiene los parámetros del modelo. Con la estimación de los parámetros se obtiene los residuos para estudiar el modelo, como valorar la sobre dispersión y finalmente las medidas de ajuste usadas.

3.1. Familia de distribuciones exponenciales

Una variable aleatoria Y tiene una distribución perteneciente a la **familia exponencial** si su función de densidad se puede expresar en la forma:

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

donde las variables son:

- θ: es el **parámetro canónico** (o natural),
- ϕ: es un **parámetro de escala**,

y las tres funciones son:

- $a(\phi)$: función para escalar la varianza,
- $b(\theta)$: es una función convexa que asegura la normalización,
- $c(y,\phi)$: es la función base (que depende de y y, opcionalmente, de ϕ).

Las siguientes distribuciones son ejemplos comunes que pertenecen a esta familia:

- Distribución normal $\mathcal{N}(\mu,\sigma^2)$
- Distribución binomial $\text{Bin}(n,p)$
- Distribución de Poisson $\text{Poi}(\lambda)$
- Distribución gamma $\text{Gamma}(\nu,\lambda)$
- Distribución exponencial $\text{Exp}(\lambda)$
- Distribución Bernoulli

Otras distribuciones como la binomial negativa y Weibull no forman parte de la familia exponencial (Faraway, 2016, p. 152), pero son suficientemente próximas como para poder ser usadas en un modelo GLM con algunas modificaciones.

3.2. Modelos posibles

A partir de las distintas distribuciones que pertenecen a la familia exponencial, es posible construir diferentes modelos lineales generalizados (GLM) adaptados a la naturaleza de los datos. Cada distribución de esta familia—como la normal, binomial, Poisson, gamma, exponencial o Bernoulli—da lugar a un modelo con una estructura particular, que se diferencia tanto por la forma de la variable respuesta como por la función de enlace que relaciona su media con el predictor lineal. La elección de la distribución apropiada depende del tipo de variable que se desea modelar (continua, discreta, de conteo, de proporciones, etc.), mientras que la función de enlace permite ajustar la relación funcional entre la media esperada de la respuesta y los predictores. A continuación, se detallan los principales modelos derivados de cada una de estas distribuciones.

3.2.1. Normal

El caso de una distribución normal corresponde al modelo de regresión lineal múltiple tratado en el capítulo anterior. A continuación, se expresa la densidad de probabilidad de la normal estándar en términos compatibles con la forma canónica de la familia exponencial y se deduce la correspondencia con los parámetros estructurales del modelo lineal generalizado.

La función de densidad de una variable aleatoria Y con distribución normal de media μ y varianza σ^2 es:

$$f(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

Para identificar esta distribución con la forma general de la familia exponencial:

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

debemos reescribir la densidad normal en esta forma. Comenzamos expandiendo el cuadrado en el exponente:

$$-\frac{(y-\mu)^2}{2\sigma^2} = -\frac{y^2 - 2y\mu + \mu^2}{2\sigma^2} = \frac{y\mu - \mu^2/2 - y^2/2}{\sigma^2}$$

Entonces, podemos reescribir la densidad como:

$$\begin{aligned} f(y \mid \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2}\right) \\ &= \exp\left(\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right) \end{aligned}$$

Comparando con la forma canónica, identificamos:

- $\theta = \mu$
- $\phi = \sigma^2$

- $a(\phi) = \phi$
- $b(\theta) = \theta^2/2$, ya que $b(\mu) = \mu^2/2$
- $c(y, \phi) = -\dfrac{y^2}{2\phi} - \dfrac{1}{2}\log(2\pi\phi)$

Por tanto, la distribución normal puede expresarse como miembro de la familia exponencial con media $\theta = \mu$, parámetro de dispersión $\phi = \sigma^2$, y el resto de funciones estructurales especificadas arriba.

3.2.2. Binomial

La distribución binomial es la base para el desarrollo de la regresión logística. Sea Y una variable aleatoria con distribución binomial de parámetros n y p, $Y \sim \text{Bin}(n, p)$, con función de masa de probabilidad:

$$f(y \mid n, p) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 0, 1, \ldots, n$$

El objetivo es expresar esta distribución en la forma canónica de la familia exponencial:

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

donde θ es el parámetro natural, $\mu = \text{E}(Y)$, y ϕ el parámetro de dispersión. Empezamos por reescribir la densidad binomial tomando logaritmos:

$$\begin{aligned} f(y \mid p) &= \binom{n}{y} \exp\left(y\log(p) + (n-y)\log(1-p)\right) \\ &= \exp\left(y\log\left(\frac{p}{1-p}\right) + n\log(1-p) + \log\binom{n}{y}\right) \end{aligned}$$

Si definimos la media poblacional como $\mu = np$, entonces la proporción de éxito es $p = \mu/n$, y el logit (la función de enlace canónica para la binomial) se define como:

$$\theta = \log\left(\frac{p}{1-p}\right)$$

$$\Rightarrow p = \frac{e^\theta}{1+e^\theta} \quad \text{y} \quad \mu = n \cdot \frac{e^\theta}{1+e^\theta}$$

Ahora reescribimos la función de probabilidad en términos de θ:

$$f(y \mid \theta) = \exp\left(y\theta - n\log(1+e^\theta) + \log\binom{n}{y}\right)$$

Esta expresión se ajusta perfectamente a la forma canónica de la familia exponencial, donde:

- $\theta = \log\left(\dfrac{p}{1-p}\right)$, el logit de la proporción de éxito
- $\phi = 1$, la binomial no tiene parámetro de dispersión libre
- $a(\phi) = 1$
- $b(\theta) = n\log(1+e^\theta)$
- $c(y, \phi) = \log\binom{n}{y}$

Por tanto, la distribución binomial puede representarse como:

$$f(y \mid \theta) = \exp\left(y\theta - n\log(1+e^{\theta}) + \log\binom{n}{y}\right)$$

lo que confirma que pertenece a la familia exponencial con función de enlace canónica logit. Esta formulación es la base de la regresión logística, en la cual la media $\mu = np$ se modela mediante una función logística del predictor lineal (Peña, 2010, p. 643).

3.2.3. Poisson

La distribución de Poisson es de especial interés, ya que es la distribución utilizada para resolver el problema propuesto. Sea $Y \sim \text{Poisson}(\mu)$ una variable aleatoria con media $\mu > 0$, su función de probabilidad es:

$$f(y \mid \mu) = \frac{e^{-\mu}\mu^{y}}{y!}, \quad y = 0, 1, 2, \ldots$$

Como en todos los casos, el objetivo es expresar esta función de probabilidad en la forma canónica de la familia exponencial:

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

Puede reescribirse como

$$f(y \mid \mu) = \exp\left(y \cdot \log(\mu) - \mu + \log\left(\frac{1}{y!}\right)\right)$$

Tomando logaritmos de la función de probabilidad original:

$$\log f(y \mid \mu) = -\mu + y\log(\mu) - \log(y!)$$

En el caso de la Poisson, no existe parámetro de dispersión libre, por lo que $\phi = 1$ y $a(\phi) = 1$. Definiendo el parámetro natural como $\theta = \log(\mu)$, se tiene que $\mu = e^{\theta}$, y sustituyendo en la expresión anterior:

$$\log f(y \mid \theta) = -e^{\theta} + y\theta - \log(y!)$$

por tanto,

$$f(y \mid \theta) = \exp\left(y\theta - e^{\theta} - \log(y!)\right)$$

lo que se ajusta exactamente a la forma canónica con:

- $\theta = \log(\mu)$, función de enlace logarítmica
- $\phi = 1$, sin parámetro de dispersión
- $a(\phi) = 1$
- $b(\theta) = e^{\theta}$
- $c(y, \phi) = -\log(y!)$

3.2.4. Gamma

Aunque la distribución Gamma y exponencial no se emplea en el análisis de este trabajo, es importante mencionar su inclusión dentro del marco general de los GLM, es por ello que se indica la correspondencia de los parámetros sin detallar la deducción de los mismos.

Sea $Y \sim \text{Gamma}(\nu, \lambda)$ con ν parámetro de forma y $\lambda = \nu/\mu$ el parámetro de escala reparametrizado en función de la media μ. Su función de densidad es:

$$f(y \mid \mu, \nu) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1} \exp\left(-\frac{\nu y}{\mu}\right)$$

Al expresarla en la forma de la familia exponencial:

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

puede deducirse la siguiente correspondencia:

- $\theta = -\dfrac{1}{\mu}$ (parámetro natural)
- $\phi = \dfrac{1}{\nu}$ (dispersión)
- $a(\phi) = \phi$
- $b(\theta) = -\log(-\theta)$
- $c(y, \phi) = \log\left(\dfrac{y^{1/\phi - 1}}{\Gamma(1/\phi)} \cdot \phi^{-1/\phi}\right)$

Esta parametrización es útil en modelos GLM con respuesta positiva y varianza proporcional al cuadrado de la media.

3.2.5. Exponencial

La distribución exponencial es un caso particular de la distribución Gamma con parámetro de forma $\nu = 1$. Su densidad se expresa como:

$$f(y \mid \mu) = \frac{1}{\mu} \exp\left(-\frac{y}{\mu}\right)$$

y puede reformularse dentro de la familia exponencial como:

$$f(y \mid \theta) = \exp\left(y\theta - \log(-\theta)\right)$$

con los componentes siguientes:

- $\theta = -\dfrac{1}{\mu}$ (parámetro natural)
- $\phi = 1$ (no hay dispersión libre)
- $a(\phi) = 1$
- $b(\theta) = -\log(-\theta)$
- $c(y, \phi) = 0$

3.2.6. Bernoulli

La distribución Bernoulli es un caso de la de distribución binomial, donde $\text{Bin}(1, p)$ es una distribución de Bernoulli. Donde el valor de y solo puede ser 0 o 1.

3.3. Funciones *link*

La función *link* conecta la media de la variable respuesta con la parte lineal del modelo. En el modelo GLM se amplía el modelo $Y = X\beta + u$, permitiendo que la variable Y siga cualquier distribución de la familia exponencial, y que la media de Y, denotada como $\mu = \mathrm{E}[Y]$, esté relacionada con una combinación lineal de predictores $\eta = X\beta$ a través de una función que no necesariamente tiene que ser lineal.

Aquí es donde entra en juego la función *link* $g(\cdot)$, es la función que relaciona la media μ de la variable respuesta con el predictor lineal $\eta = X\beta$:

$$g(\mu) = \eta \text{ o bien } \mu = g^{-1}(\eta)$$

esto permite trabajar con variables cuya media está restringida (por ejemplo, $\mu \in (0, 1)$ para proporciones), utilizando predictores $\eta \in \mathbb{R}$.

Es deseable que la función *link* cumpla:

- Función monótona (para que haya correspondencia unívoca).
- Mapea el rango de μ al espacio real $\mathbb{R}$ (dominio de η).
- A veces se elige la función *link* natural, que aparece directamente en la forma canónica de la familia exponencial (lo que facilita la estimación), pero en otros casos puede no usarse.

En la siguiente tabla aparece la función *link* canónica, aunque puede usarse otras (Pérez, 2021, p. 170).

Distribución	**Variable respuesta** Y	***link*** $g(\mu)$	**Rango de** μ
Normal	Continua en $\mathbb{R}$	Identidad: $g(\mu) = \mu$	$(-\infty, \infty)$
Binomial	Proporción $Y/n \in (0, 1)$	Logit: $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$	$(0, 1)$
Poisson	Cuenta $Y \in \mathbb{N}$	Log: $g(\mu) = \log(\mu)$	$(0, \infty)$
Gamma	Continua positiva $Y > 0$	Inversa: $g(\mu) = \frac{1}{\mu}$	$(0, \infty)$
Exponencial	Continua positiva $Y > 0$	Inversa negativa: $g(\mu) = -\frac{1}{\mu}$	$(0, \infty)$
Bernoulli	$Y \in \{0, 1\}$	Logit: $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$	$(0, 1)$

3.4. Relación entre la distribución y la función *link*

A continuación se justifica formalmente la elección de la función *link* para cada una de las distribuciones de la familia exponencial.

Veamos que si Y es una distribución de la familia exponencial entonces es obtiene:

$$\mathrm{E}[Y] = \mu = b'(\theta)$$
$$\mathrm{Var}(Y) = a(\phi) \cdot b''(\theta)$$

Se demuestra para el caso continuo, siendo el caso discreto demostrable mediante el mismo procedimiento teniendo en cuenta que son sumatorios en lugar de integrales.

Sea Y es una distribución de la familia exponencial, y sea f su función de densidad y M el espacio muestral. Por ser f función de densidad tenemos la siguiente igualdad

$$1 = \int_M f(y|\theta,\phi)dy = \int_M \exp\left(y\theta/a(\phi)\right)\exp\left(-b(\theta)/a(\phi)\right)\exp\left(c(y,\phi)\right)dy$$

por lo que podemos escribir

$$\frac{1}{\exp\left(-b(\theta)/a(\phi)\right)} = \int_M \exp\left(y\theta/a(\phi)\right)\exp\left(c(y,\phi)\right)dy$$

resultando que

$$b(\theta) = a(\phi)\log\left(\int_M \exp\left(y\theta/a(\phi) + c(y,\phi)\right)dy\right)$$

para obtener la derivada de $b(\theta)$ aplicamos el teorema de la convergencia dominada y la regla de la cadena, obteniendo:

$$\begin{aligned}
\frac{db(\theta)}{d\theta} &= a(\phi)\cdot\frac{1}{\int_M \exp\left(y\theta/a(\phi)+c(y,\phi)\right)dy}\cdot\int_M \frac{d}{d\theta}\left(\exp(y\theta/a(\phi)+c(y,\phi))\right)dy \\
&= a(\phi)\cdot\frac{\int_M \left(y/a(\phi)\right)\cdot\exp\left(y\theta/a(\phi)+c(y,\phi)\right)dy}{\exp(b(\theta)/a(\phi))} \\
&= \int_M y\cdot\exp\left(\left(y\theta/a(\phi)+c(y,\theta)\right)\exp\left(-b(\theta)/a(\phi)dy,\right.\right.
\end{aligned}$$

y como $\exp\left(\left(y\theta/a(\phi)+c(y,\theta)\right)\exp\left(-b(\theta)/a(\phi)\right)\right.$ es $f(y|\theta,\phi)$, llegamos a la expresión

$$\frac{db(\theta)}{d\theta} = \int_M y\cdot f(y|\theta,\phi)dy = \mathrm{E}[Y] = \mu$$

demostrando que

$$\mathrm{E}[Y] = \mu = b'(\theta)$$

Ahora calculamos $b''(\theta)$ mediante el mismo razonamiento

$$\begin{aligned}
\frac{d^2(b(\theta))}{d\theta^2} &= \frac{d}{d\theta}\left(\int_M y\cdot f(y|\theta,\phi)dy\right) \\
&= \int_M \frac{d}{d\theta}\left(y\cdot f(y|\theta,\phi)\right)dy \\
&= \int_M \left(y/a(\phi)\right)\cdot\exp\left(c(y,\phi)\cdot\exp\left(\left(y\theta-b(\phi)\right)/a(\phi)\right)\cdot\left(y-b'(\theta)\right)dy\right. \\
&= \frac{1}{a(\phi)}\cdot\left(\int_M y^2\cdot f(y|\theta,\phi)dy - b'(\theta)\cdot\int_M y\cdot f(y|\theta,\phi)dy\right) \\
&= \frac{1}{a(\phi)}\left(\mathrm{E}[Y^2] - E[Y]^2\right) \\
&= \frac{1}{a(\phi)}\mathrm{Var}(Y)
\end{aligned}$$

Por lo que efectivamente se obtiene

$$\mathrm{Var}(Y) = a(\phi)\cdot b''(\theta)$$

Dado este resultado, se deduce que la relación entre μ y θ viene dado por la función b', por lo que es razonable determinar el parámetro θ en función de μ. Teniendo

$$\theta = (b')^{-1}(\mu)$$

teniendo como función *link* la función que cumple

$$g(\mu) = (b')^{-1}(\mu)$$

la función que cumple dicha restricción se le denomina función ***link* canónica**.

3.5. Estimación de los parámetros

Ahora veremos como estimar los parámetros β_i $(0 \leq i \leq k)$ a partir de un conjunto de n observaciones en un modelo lineal generalizado.

Supongamos que tenemos un conjunto de n observaciones y_i sobre un conjunto de covariables independientes $x_1, x_2, ..., x_k$. Donde estas observaciones corresponden Y_i variables aleatorias independientes e idénticamente distribuidas, además con distribución de la familia exponencial.

Definimos la función de verosimilitud $L(\beta, \phi)$ como la probabilidad de observar los datos $y_1, ..., y_n$ como función de los parámetros β y ϕ. Siendo la función de distribución de las Y_i

$$f(y_i|\phi_i, \theta) = \exp\left(\frac{y_i \cdot \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right)$$

se tiene

$$\begin{aligned} L(\beta, \phi) &= \prod_{i=1}^{n} f(y_i|\phi_i, \theta) \\ &= \prod_{i=1}^{n} \exp\left(\frac{y_i \cdot \theta_i - b(\theta_i)}{a(\phi} + c(y_i, \phi)\right) \end{aligned}$$

al ser la función logarítmico monótona, y por facilidad de cálculo, podemos obtener el máximo en la función:

$$\ell(\beta, \phi) = \log L(\beta, \phi) = \sum_{i=1}^{n} \left(\frac{y_i \cdot \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right)$$

ahora se usa la función *link* definida previamente, susituyendo

$$\theta_i = (b')^{-1}(g^{-1}(\boldsymbol{x}_i'\beta))$$

y de ésta forma la función $\log L(\beta, \phi)$ depende únicamente de β y ϕ si es necesario estimarlo.

La máxima verosimilitud se produce cuando

$$\frac{\partial \log L(\beta, \phi)}{\partial \beta} = 0$$

Esta ecuación es un sistema de ecuaciones de $k+1$ derivadas con $k+1$ incógnitas $(\beta_0, \beta_1, \beta_2, ..., \beta_k)$. Se puede probar que el parámetro ϕ no influye en la estimación de los parámetros β.

Si se trata de una distribución normal estamos en el caso del modelo de regresión lineal múltiple, y puede resolverse mediante mínimos cuadrados, pero en cualquier otra caso se tiene una expresión complicada, y debe resolverse mediante un método iterativo. Uno de los métodos utilizados para determinar el valor del vector β es el método de Newton-Raphson.

3.5.1. Ajuste a la población: *offset*

En determinados problemas se desea modelar el número de eventos relativos a una unidad de exposición (como puede ser la población). Esto se conoce como ajuste del modelo a una covariable independiente que representa la población, introduciendo un término conocido como *offset* en el predictor lineal. Este término no se estima a partir de los datos, sino que se incorpora directamente con una pendiente fija igual a uno. Supongamos que se desea modelar el número de casos Y_i registrados en una población de tamaño n_i. En lugar de modelar directamente Y_i, se modela la tasa de eventos por habitante, es decir, $\mu_i = \mathrm{E}[Y_i]/n_i$. Al aplicar una función de enlace logarítmica, se obtiene:

$$\log(\mathrm{E}[Y_i]) = \mathbf{x}_i'\beta + \log(n_i)$$

donde $\log(n_i)$ se especifica como *offset* en el modelo. Esta formulación garantiza que las predicciones se ajusten proporcionalmente al tamaño poblacional.

El uso de `offset` es esencial en contextos donde las observaciones provienen de poblaciones con diferentes niveles de exposición (e.g., diferentes tamaños poblacionales, tiempos de observación, longitudes de tramos de carretera), y se desea que el modelo compare tasas ajustadas en lugar de frecuencias absolutas.

3.6. Residuos

En el contexto de los modelos estadísticos, los residuos representan la diferencia entre los valores observados de la variable respuesta y los valores predichos por el modelo. Constituyen una medida del error de predicción y permiten evaluar qué tan bien se ajusta el modelo a los datos. El análisis de los residuos es fundamental para detectar posibles desviaciones respecto a los supuestos del modelo, como la presencia de valores atípicos, heterocedasticidad o mala especificación funcional. En un modelo lineal generalizado, existen diferentes tipos de residuos, veremos: residuos deviance y residuos Pearson. Cada uno con propiedades específicas que los hacen útiles para distintos fines diagnósticos. Veremos la medida deviance como una medida de ajuste global.

3.6.1. Residuo de Pearson

Sea $\hat{\mu}_i = g^{-1}(x_i'\beta)$ el valor ajustado por el modelo con el vector de entrada $\boldsymbol{x}'$. Se define como residuo de Pearson:

$$r_{P,i} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

donde $V(\hat{\mu}_i)$ es la función varianza de la distribución. El residuo de Pearson es una medida estandarizada de la discrepancia entre los valores observados y los valores esperados bajo el modelo.

3.6.2. Residuo Deviance

Se define como residuo deviance:

$$r_{D,i} = \mathrm{sign}(y_i - \hat{\mu}_i) \cdot \sqrt{2\left(\ell(y_i) - \ell(\hat{\mu}_i)\right)}$$

donde ℓ es la log-verosimilitud. Este residuo es una medida que cuantifica la discrepancia entre el valor observado y el valor ajustado por el modelo basado en la contribución individual de cada observación.

3.6.3. Deviance

La deviance es una medida de calidad del ajuste del modelo, análoga a la suma de cuadrados de residuos en la regresión lineal clásica. Se define como una función que compara el modelo ajustado con el modelo saturado, es decir, el modelo que se ajusta perfectamente a los datos (predice exactamente cada observación). Se denota por D el valor

$$D = 2\sum_{i=1}^{n} (log(\ell(y_i) - \ell(\hat{\mu}_i))$$

El término $\ell(y_i)$ corresponde al modelo saturado que sería con un ajuste perfecto y $\ell(\hat{\mu}_i)$ corresponde a la estimación realizada sobre el vector $\boldsymbol{x}$.

3.6.4. Modelo de Poisson

Veamos el desarrollo específico para el caso de una distribución de Poisson. Asumimos que:

$$Y_i \sim \text{Poisson}(\mu_i), \quad \text{con} \quad \mu_i = \text{E}[Y_i]$$

y la función de enlace es:

$$\eta_i = \log(\mu_i) \quad \Rightarrow \quad \mu_i = e^{\eta_i}$$

La función de varianza para la distribución de Poisson es $V(\mu_i) = \mu_i$, al ser la media igual a la varianza. Sobre este aspecto incidiremos en el siguiente apartado. Por tanto, el residuo de Pearson se define como:

$$r_{P,i} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

Para el residuo de Deviance, tenemos que la log-verosimilitud para una observación y_i bajo la distribución de Poisson es:

$$\ell(y_i|\mu_i) = y_i \log(\mu_i) - \mu_i - \log(y_i!)$$

por lo que la deviance se define como:

$$D = 2\sum_{i=1}^{n} [\ell(y_i|y_i) - \ell(y_i|\hat{\mu}_i)]$$

Desarrollando para cada observación:

$$\begin{aligned} D_i &= 2\left[y_i \log(y_i) - y_i - \log(y_i!) - (y_i \log(\hat{\mu}_i) - \hat{\mu}_i - \log(y_i!))\right] \\ &= 2\left(y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right) \end{aligned}$$

Teniendo en cuenta que si $y_i = 0$, se define $y_i \log(y_i/\hat{\mu}_i) = 0$.
Entonces, el **residuo de deviance** se define como:

$$r_{D,i} = \text{sign}(y_i - \hat{\mu}_i) \cdot \sqrt{2\left(y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right)}$$

La deviance total del modelo es la suma de los residuos de deviance al cuadrado:

$$D = \sum_{i=1}^{n} r_{D,i}^2 = 2\sum_{i=1}^{n} \left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i) \right]$$

De esta expresión se deduce que si $y_i \approx \hat{\mu}_i$ para $i = 1...n$, entonces $\log(\frac{y_i}{\hat{\mu}_i}) \approx \log(1) = 0$ y además $y_i - \hat{\mu}_i \approx 0$, por lo que se obtiene

$$D \approx 0$$

lo que supone un buen ajuste del modelo.

3.7. Estimación del parámetro ϕ

En un modelo lineal generalizado (GLM), el parámetro de dispersión ϕ aparece en la forma general de la función de densidad de la familia exponencial:

$$f(y_i|\theta_i, \phi) = \exp\left(\frac{y_i \cdot \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right)$$

Para algunas distribuciones como la binomial y la de Poisson, el parámetro de dispersión está fijado: $\phi = 1$. Sin embargo, en modelos como el normal, gamma o inversa gaussiana, ϕ debe ser estimado.

Una vez estimados los coeficientes β del modelo, una estimación común de ϕ se obtiene a partir de los residuos, en particular de la **deviance total** D, como:

$$\hat{\phi} = \frac{D}{n - p}$$

donde:

- D es la deviance del modelo ajustado,
- n es el número total de observaciones,
- p es el número de parámetros estimados (grados de libertad del modelo).

Este estimador es el equivalente a la estimación del error cuadrático medio en la regresión lineal clásica y es insesgado bajo el supuesto de que el modelo está correctamente especificado.

Como alternativa, puede estimarse ϕ utilizando los *residuos de Pearson* $r_{P,i}$, mediante:

$$\hat{\phi} = \frac{1}{n - p} \sum_{i=1}^{n} \left(\frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} \right)^2$$

donde $V(\hat{\mu}_i)$ es la función de varianza evaluada en el valor ajustado $\hat{\mu}_i$.

Cabe destacar que ambos estimadores coinciden asintóticamente.

3.8. Sobredispersión

En los modelos lineales generalizados con distribución de Poisson, se asume que la varianza de la variable respuesta Y_i es igual a su media:

$$\text{Var}(Y_i) = \mu_i$$

Este supuesto puede no cumplirse en la práctica. Cuando los datos presentan una variabilidad mayor que la esperada bajo la distribución de Poisson, se dice que existe **sobredispersión**. Formalmente, hablamos de sobredispersión cuando $\text{Var}(Y_i) > \mu_i$, de forma análoga, si $\text{Var}(Y_i) < \mu_i$, se habla de **infradispersión**.

Una forma práctica de evaluar la presencia de sobredispersión en un modelo de Poisson es mediante el **estadístico de Pearson**:

$$X^2 = \sum_{i=1}^{n} \left(\frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}} \right)^2 = \sum_{i=1}^{n} r_{P,i}^2$$

donde $r_{P,i}$ es el residuo de Pearson.

Se estima el parámetro de dispersión empírico como:

$$\hat{\phi} = \frac{X^2}{n-p}$$

donde n es el número de observaciones y p es el número de parámetros estimados en el modelo. La interpretación más habitual es la siguiente:

- Si $\hat{\phi} \approx 1$, no hay evidencia de sobredispersión.
- Si $\hat{\phi} \gg 1$, hay sobredispersión.
- Si $\hat{\phi} \ll 1$, hay infradispersión.

La presencia de sobredispersión puede invalidar los errores estándar y las inferencias obtenidas bajo el modelo Poisson. En tales casos, se recomienda considerar modelos alternativos como el *modelo cuasi-Poisson*, el *modelo binomial negativo* o el *modelo de Poisson cero inflado* (Faraway, 2016, capítulo 5).

Una de las causas de la sobredispersión (McCullagh and Nelder, 1989, p. 198) sucede si observamos $Y = Z_1 + Z_2 + \cdots + Z_n$ donde Z_i son variables aleatorias independientes e identicamente distribuidas, y N es una variable aleatoria con distribución de Poisson e independiente de Z. En este caso se tiene que:

$$\text{E}[Y] = \text{E}[N]\,\text{E}[Z]$$

y

$$\text{Var}(Y) = \text{E}[N]\,\text{Var}(Z) + \text{Var}(N)\,\text{E}[Z]^2 = \text{E}[N]\,\text{E}[Z^2]$$

por lo que habrá sobredispersión si $\text{E}[Z^2] > \text{E}[Z]$.

Otro caso de sobredispersión puede suceder si hay variabilidad entre sujetos, en estos casos el modelo se aproxima mejor con una distribución binomial negativa. Pero esta distribución no pertenece a la familia de distribuciones exponencial.

3.9. Binomial Negativa

Tal y como se detalla en (Faraway, 2016, p. 92), dada una serie de ensayos independientes, cada uno con probabilidad de éxito p, llamamos Z al número de ensayos hasta que ocurre el k-ésimo éxito. Entonces:

$$P(Z = z) = \binom{z-1}{k-1} p^k (1-p)^{z-k}, \quad z = k, k+1, \ldots$$

Notemos que la distribución binomial negativa se obtiene como una generalización de la Poisson, donde el parámetro λ sigue una distribución gamma.

Si definimos $Y = Z - k$ y $p = (1+\alpha)^{-1}$, se obtiene una parametrización más adecuada:

$$P(Y = y) = \binom{y+k-1}{k-1} \frac{\alpha^y}{(1+\alpha)^{y+k}}, \quad y = 0, 1, 2, \ldots$$

cuya media y varianza son:

$$\mathrm{E}[Y] = \mu = k\alpha \quad \text{y} \quad \mathrm{Var}(Y) = k\alpha + k\alpha^2 = \mu + \frac{\mu^2}{k}$$

Así, la log-verosimilitud para una muestra de tamaño n es:

$$\sum_{i=1}^{n} \left(y_i \log \frac{\alpha}{1+\alpha} - k \log(1+\alpha) + \sum_{j=0}^{y_i-1} \log(j+k) - \log(y_i!) \right)$$

Entonces, la forma más conveniente de vincular la media de respuesta μ con una combinación lineal de los predictores X es mediante:

$$\eta = x'\beta = \log \frac{\alpha}{1+\alpha} = \log \frac{\mu}{\mu+k}$$

donde el parámetro k puede considerarse conocido, o determinarase mediante estimación.

Aunque la regresión de Poisson es un caso particular de los modelos lineales generalizados (GLM), la regresión binomial negativa no puede considerarse formalmente un GLM en sentido estricto, ya que la distribución binomial negativa no pertenece a la familia exponencial. Esto se debe a que su función de probabilidad no puede expresarse en la forma canónica de la familia exponencial, lo que impide aplicar directamente los fundamentos teóricos de los GLM.

Sin embargo, la regresión binomial negativa mantiene una estructura similar a un GLM: permite modelar la media de la variable respuesta mediante una función de enlace (generalmente el logaritmo), y presenta una relación entre la media y la varianza, dada por $\mathrm{Var}(Y) = \mu + \mu^2/\theta$, donde θ es un parámetro de dispersión. Por ello, en la práctica se utiliza una estimación por máxima verosimilitud ajustando un modelo de regresión extendido mediante funciones específicas.

3.10. Comparación de modelos

Es una situación normal disponer de dos o más modelos, cuya diferencia entre ellos es el uso de diferentes covariables independientes para estimar el valor de la variable respuesta Y. Existen diversos criterios de comparación y selección del modelo más adecuado. A continuación se presenta tres criterios de comparación validos para modelos GLM: el test de deviance, el AIC y el BIC.

3.10.1. Test de Deviance

La función del test de deviance es comparar dos modelos **anidados**, es decir, cuando un modelo reducido (menos parámetros) es un caso particular de otro completo. La deviance se define como:

$$D = 2 \sum_{i=1}^{n} [\ell(y_i|y_i) - \ell(y_i|\hat{\mu}_i)]$$

donde $\ell(y_i|y_i)$ es la log-verosimilitud del modelo saturado (el que predice perfectamente cada observación) y $\ell(y_i|\hat{\theta})$ es la log-verosimilitud del modelo ajustado.

Para comparar dos modelos anidados M_0 reducido y M_1 completo debe calcularse:

$$\Delta D = D_{M_0} - D_{M_1} = 2\sum_{i=1}^{n} [\ell(y|\hat{\mu}_1) - \ell(y|\hat{\mu}_0)]$$

Bajo la hipótesis nula de que el modelo reducido es correcto. Puede demostrarse (McCullagh and Nelder, 1989, p. 119) que ΔD sigue asintóticamente una distribución chi-cuadrado:

$$\Delta D \sim \chi^2_{gl}$$

donde $gl = p_1 - p_0$ representa la diferencia en el número de parámetros entre ambos modelos. Un valor grande de ΔD significa que el modelo completo mejora el ajuste. El contraste de hipótesis establece como hipótesis nula H_0 que los modelos son iguales, frente a H_1 que los modelos son diferentes.

3.10.2. Criterio de Información de Akaike (AIC)

El AIC es una medida de ajuste basada en la log-verosimilitud que penaliza la complejidad del modelo. Se define como:

$$\text{AIC} = -2\ell(y|\hat{\mu}) + 2p$$

donde p es el número de parámetros del modelo. Esta medida permite comparar modelos anidados o no anidados. Un modelo con menor AIC se considera preferible frente a otro con mayor AIC. La penalización por el número de parámetros trata de evita el sobreajuste motivado por la inclusión de parámetros que aportan escasa información.

3.10.3. Criterio de Información Bayesiano (BIC)

El BIC penaliza la complejidad del modelo de una manera más restrictiva. Se define como:

$$\text{BIC} = -2\ell(y|\hat{\mu}) + p\log(n)$$

donde n es el número de observaciones. Al igual que el AIC, el modelo con menor BIC es preferido. Debido a que la penalización depende del número de observaciones n, el BIC tiende a favorecer modelos más simples cuando el tamaño muestral es grande.

3.10.4. Comparativa de criterios

El test de deviance resulta adecuado cuando se desea comparar modelos anidados, ya que proporciona un contraste formal basado en hipótesis estadísticas. En cambio, los criterios de información AIC y BIC son aplicables de forma más general, permitiendo la comparación entre modelos GLM no necesariamente anidados. Estos criterios buscan un equilibrio entre la calidad del ajuste y la complejidad del modelo. Sin embargo, la decisión sobre qué modelo adoptar no debe basarse exclusivamente en estos valores numéricos, sino que también debe tenerse en cuenta el contexto específico del estudio y la claridad interpretativa del modelo. Es fundamental evitar la selección de modelos que contradigan la lógica y el conocimiento previo del problema a resolver.

Capítulo 4

Modelos de Conteo Cero Inflados

4.1. Motivación

En el análisis estadístico de variables de conteo (es decir, variables que toman valores enteros no negativos, como el número de eventos en un intervalo de tiempo o espacio), uno de los modelos más utilizados es el modelo lineal generalizado de Poisson. Este modelo asume que la variable respuesta Y sigue una distribución de Poisson con parámetro λ, y que la media y la varianza de Y coinciden ($\mathrm{E}[Y] = \mathrm{Var}(Y) = \lambda$).

Sin embargo, en muchas aplicaciones reales se observa una proporción excesiva de ceros que no puede ser explicada adecuadamente por el modelo de Poisson clásico. Este fenómeno se conoce como "inflación de ceros". Por ejemplo, en estudios epidemiológicos donde se modela el número de casos de una enfermedad en pequeñas localidades, muchas de ellas pueden no presentar ningún caso (cero casos), pero no debido a la distribución natural de Poisson sino a procesos estructurales diferentes (falta de exposición, protección inmunológica, etc.).

4.1.1. Detección exceso de ceros

Es importante evaluar si el modelo de Poisson es capaz de explicar adecuadamente la cantidad de ceros observados en los datos. Una forma sencilla de hacerlo es comparar la proporción de ceros observados con la proporción de ceros esperados bajo el modelo ajustado. Sea $\hat{\lambda}_i$ la media estimada para la observación i bajo el modelo de Poisson. Entonces, la probabilidad de que $Y_i = 0$ bajo este modelo es $P(Y_i = 0) = e^{-\hat{\lambda}_i}$. El promedio de estas probabilidades a lo largo de todas las observaciones proporciona una estimación de la proporción esperada de ceros. Si la proporción de ceros observados en los datos es sustancialmente mayor que la esperada por el modelo, se puede concluir que hay evidencia de **inflación de ceros**, lo que justificaría el uso de modelos más complejos.

4.1.2. Modelos cero inflados

Para solucionar esta problemática se han desarrollado modelos de conteo cero inflados, en particular el modelo Poisson cero inflado (ZIP) y su extensión, el modelo binomial negativo cero inflado (ZINB). Estos modelos combinan un componente de conteo (Poisson o binomial

negativo) con un componente logístico que modela la probabilidad de que un cero provenga de un proceso estructural distinto. A continuación haremos una exposición formal de ambos modelos.

4.2. Modelo Cero Inflado de Poisson (ZIP)

4.2.1. Formulación

Sea $Y \in \mathbb{N}$ una variable aleatoria de conteo. El modelo ZIP asume que Y tiene una distribución compuesta de dos procesos:

- Con probabilidad $\pi \in [0, 1]$, se genera un cero estructural (proceso inflado)
- Con probabilidad $1 - \pi$, Y sigue una distribución de Poisson con media $\lambda > 0$

La función de masa de probabilidad del modelo ZIP es:

$$P(Y = y) = \begin{cases} \pi + (1 - \pi)e^{-\lambda}, & \text{si } y = 0 \\ (1 - \pi)\dfrac{\lambda^y e^{-\lambda}}{y!}, & \text{si } y > 0 \end{cases}$$

Puede producirse que $Y = 0$ por dos situaciones, que sea un cero estructural (con probabilidad π), o bien que la variable de conteo sea cero (con probabilidad $(1 - \pi)e^{-\lambda}$).

4.2.2. Justificación matemática

Sea Z la variable latente $Z \sim \text{Bernoulli}(1 - \pi)$, teniendo:

$$Y \mid Z = \begin{cases} 0, & Z = 0 \\ \text{Poisson}(\lambda), & Z = 1 \end{cases}$$

Por lo que la esperanza es:

$$\mathrm{E}[Y] = \mathrm{E}[\mathrm{E}[Y \mid Z]] = \pi \cdot 0 + (1 - \pi) \cdot \mu = (1 - \pi)\lambda$$

Para la varianza tenemos:

$$\mathrm{Var}(Y) = \mathrm{E}[\mathrm{Var}(Y \mid Z)] + \mathrm{Var}(\mathrm{E}[Y \mid Z])$$

Calculamos cada término:

$$\mathrm{E}[\mathrm{Var}(Y \mid Z)] = \pi \cdot 0 + (1 - \pi) \cdot \lambda = (1 - \pi)\lambda$$

$$\mathrm{Var}(\mathrm{E}[Y \mid Z]) = \mathrm{Var} \begin{cases} 0, & \text{con probabilidad } \pi \\ \lambda, & \text{con probabilidad } 1 - \pi \end{cases} = \pi(1 - \pi)\lambda^2$$

Sumando ambos términos, se obtiene:

$$\mathrm{Var}(Y) = (1 - \pi)\lambda + \pi(1 - \pi)\lambda^2 = (1 - \pi)\lambda(1 + \pi\lambda)$$

operando, se llega a la expresión

$$\mathrm{Var}(Y) = (1 - \pi)\lambda(1 + \pi\lambda)$$

Como vemos, la varianza del modelo ZIP excede a su media cuando $\pi > 0$, permitiendo modelar datos con sobredispersión e inflación de ceros aimultáneamente. La esperanza y la varianza de Y bajo el modelo ZIP son:

$$\begin{aligned} \mathrm{E}[Y] &= (1-\pi)\lambda \\ \mathrm{Var}(Y) &= (1-\pi)\lambda(1+\pi\lambda) \end{aligned}$$

El modelo ZIP se especifica habitualmente como un modelo de regresión doble:

- Componente de conteo: $\log(\lambda_i) = \mathbf{x}_i'\beta$
- Componente de inflación: $\mathrm{logit}(\pi_i) = \mathbf{z}_i'\gamma$

donde $\mathbf{x}_i$ y $\mathbf{z}_i$ son vectores de covariables para la unidad i, y β, γ son vectores de parámetros. Destacar que puede darse que algunas covariables independientes afecten sólo al conteo, sólo a la inflación o ambos.

4.3. Modelo Cero Inflado Binomial Negativo (ZINB)

4.3.1. Motivación y formulación

El modelo ZIP puede ser insuficiente cuando los datos presentan sobredispersión adicional a la que explica la inflación de ceros. En estos casos, se utiliza el modelo binomial negativo cero inflado (ZINB), que sustituye la componente de Poisson por una binomial negativa, que permite mayor varianza.

Recordemos que la distribución binomial negativa es una distribución discreta que modela el número de fracasos Y hasta observar r éxitos en una secuencia de ensayos de Bernoulli independientes, cada uno con probabilidad de éxito $p \in (0,1)$. Su función de probabilidad está dada por:

$$P(Y=y) = \binom{y+r-1}{y}(1-p)^r p^y, \quad y = 0,1,2,\ldots$$

Alternativamente, en parametrización en términos de la media μ y un parámetro de dispersión $\theta > 0$, se puede escribir como:

$$P(Y=y) = \frac{\Gamma(y+\theta)}{\Gamma(\theta)\,y!}\left(\frac{\mu}{\mu+\theta}\right)^y\left(\frac{\theta}{\mu+\theta}\right)^\theta$$

donde $\Gamma(\cdot)$ representa la función gamma. Esta parametrización es útil en contextos de regresión, ya que la media de la distribución es $\mathrm{E}[Y] = \mu$ y su varianza es $\mathrm{Var}(Y) = \mu + \mu^2/\theta$, lo que permite capturar sobredispersión en datos de conteo. El parámetro θ controla el grado de sobredispersión: a medida que $\theta \to \infty$, la distribución converge a una Poisson.

La función de probabilidad del modelo ZINB es similar a ZIP, pero se usa una distribución binomial negativa en lugar de una distribucion de Poisson:

$$P(Y=y) = \begin{cases} \pi + (1-\pi)\cdot BN(y=0|\mu,\theta), & y=0 \\ (1-\pi)\cdot BN(y|\mu,\theta), & y>0 \end{cases}$$

donde $BN(y|\mu,\theta)$ es la función de probabilidad de una binomial negativa con media μ y parámetro de dispersión θ.

4.3.2. Justificación matemática

Al igual que en el modelo ZIP, introducimos una variable latente $Z \sim \text{Bernoulli}(1-\pi)$ tal que:

$$Y \mid Z = \begin{cases} 0, & Z = 0 \\ \text{BN}(\mu, \theta), & Z = 1 \end{cases}$$

La esperanza es

$$\text{E}[Y] = \pi \cdot 0 + (1-\pi) \cdot \mu = (1-\pi)\mu$$

Usamos de nuevo la fórmula de la varianza total:

$$\text{Var}(Y) = \text{E}[\text{Var}(Y \mid Z)] + \text{Var}(\text{E}[Y \mid Z])$$

y sabiendo que

$$\text{Var}(Y \mid Z = 1) = \mu + \tfrac{\mu^2}{\theta}$$

$$\text{Var}(Y \mid Z = 0) = 0$$

entonces:

$$\text{E}[\text{Var}(Y \mid Z)] = (1-\pi)\left(\mu + \frac{\mu^2}{\theta}\right)$$
$$\text{Var}(\text{E}[Y \mid Z]) = \pi(1-\pi)\mu^2$$

Por tanto, la varianza total es:

$$\text{Var}(Y) = (1-\pi)\left(\mu + \frac{\mu^2}{\theta}\right) + \pi(1-\pi)\mu^2$$

operando, llegamos a

$$\text{Var}(Y) = (1-\pi)\mu\left(1 + \frac{\mu}{\theta} + \pi\mu\right)$$

Este resultado muestra que la varianza del modelo ZINB excede a la del modelo binomial negativo, incorporando tanto la sobre-dispersión propia de la distribución como el efecto adicional del exceso de ceros.

Y en el modelo ZINB:

$$\text{E}[Y] = (1-\pi)\mu$$
$$\text{Var}(Y) = (1-\pi)\mu\left(1 + \mu\left(\frac{1}{\theta} + \pi\right)\right)$$

Este modelo permite modelar simultáneamente inflación de ceros y sobredispersión.

La especificación como modelo de regresión es análoga al modelo ZIP:

- Componente de conteo: $\log(\mu_i) = \mathbf{x}_i'\beta$
- Componente de inflación: $\text{logit}(\pi_i) = \mathbf{z}_i'\gamma$

donde, igual que en el modelo ZIP, $\mathbf{x}_i$ y $\mathbf{z}_i$ son vectores de covariables para la unidad i, y β, γ son vectores de parámetros.

4.4. Estimación de la probabilidad de ceros estructurales π

En ambos modelos se asume que la variable respuesta Y proviene de una mezcla de dos procesos estocásticos distintos. Con probabilidad π_i, la observación Y_i se genera por un proceso estructural que produce exclusivamente ceros; y con probabilidad $1 - \pi_i$, la observación proviene de un modelo de conteo como la distribución de Poisson o binomial negativa.

Para modelar la probabilidad de inflación de ceros, π_i, se emplea una función de enlace logit que relaciona esta probabilidad con un conjunto de covariables independientes $\mathbf{z}_i$. Este conjunto de covariables independientes no tiene porque ser el mismo conjunto que $\mathbf{x}_i$, pudiendo ser un subconjunto de la misma o incluso diferente.

$$\pi_i = \frac{1}{1 + \exp(-\mathbf{z}_i'\gamma)} \quad \Leftrightarrow \quad \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \mathbf{z}_i'\gamma$$

donde γ es un vector de parámetros que controla cómo influyen las covariables independientes $\mathbf{z}_i$ en la probabilidad de que una observación sea un cero estructural.

La estimación de los parámetros del modelo se realiza mediante *máxima verosimilitud*. La función de verosimilitud del modelo refleja la probabilidad conjunta de observar los datos bajo la suposición de una mezcla de dos procesos: uno que genera ceros estructurales y otro que genera recuentos (que pueden incluir ceros "naturales"). La verosimilitud total tiene en cuenta tanto el componente de inflación de ceros como el de conteo, y los parámetros β (del modelo de conteo) y γ (del modelo de inflación) se estiman simultáneamente.

Es importante destacar que el parámetro π_i no representa la proporción observada de ceros, sino la probabilidad condicional de que una determinada observación Y_i pertenezca al componente de ceros estructurales, dadas las covariables independientes $\mathbf{z}_i$. Por tanto, π_i puede variar entre observaciones y depende de la especificación del modelo logístico para el componente inflado.

4.5. Comparativa

Los modelos de conteo cero inflados como el ZIP y ZINB son herramientas potentes para modelar datos de conteo con exceso de ceros. El modelo ZIP es adecuado cuando hay inflación de ceros pero varianza moderada, mientras que el ZINB es preferible cuando hay sobredispersión adicional. La formulación doble de regresión permite incorporar factores explicativos distintos para el proceso de generación de ceros y para el conteo propiamente dicho, dotando a estos modelos de una gran flexibilidad y aplicabilidad en contextos reales.

Capítulo 5

Problema

El enunciado del problema planteado es el siguiente:

(García-Pérez (2013)) El arsénico es una sustancia natural que es tóxica sólo si se ingiere en gran cantidad. Los límites tolerados en el agua potable dependen de los países y de los momentos aunque suelen ser valores de hasta 50 ppb (partes por billón americano, 10^{-9}).

En 1999, la Academia Nacional de Ciencias de Estados Unidos realizó un estudio en el que concluyó que el arsénico en el agua causaba cáncer de piel, de vejiga y de pulmón y que podía causar cáncer de riñón e hígado. El estudio también concluyó que el arsénico en el agua potable también causaba daños en el sistema nervioso periférico y central, así como en el corazón y vasos sanguíneos, además de diversos problemas en la piel.

En Chen et al. (1985) aparece un estudio llevado a cabo en áreas rurales del suroeste de Taiwan en donde se habían excavado pozos en el terreno para proporcionar agua fresca a la población pero con el consiguiente riesgo de contaminación medioambiental.

Parte de los datos de ese trabajo están en el fichero `arsenic` que muestra, como variable dependiente, el número de muertos por cáncer, `events`; en 43 pequeñas ciudades, `village`; apareciendo en el fichero como posibles covariables: los niveles medianos de arsénico específicos de la ciudad, `conc`; el número de personas en riesgo, `at_risk`; en cada edad, `age`; así como el género de la persona, `gender` (0 mujer, 1 hombre); y el tipo de cáncer, `type` (0 vejiga, 1 pulmón).

Analizar estos datos mediante una Regresión Poisson considerando como variable dependiente `events` y, primero, como posibles covariables independientes `conc`, `gender` y `type`. Analizar si se verifican las suposiciones necesarias para poder aplicar este modelo y luego interpretar los resultados.

Después, quitar del análisis los individuos que no recibieron arsénico, es decir aquellos para los que es `conc = 0` y volver a analizar los datos con las mismas covariables.

Además, de los análisis indicados en el enunciado se han explorado otras soluciones como el modelo binomial negativo y ZINB.

Capítulo 6

Herramientas del Análisis de Datos

6.1. Elección del entorno de análisis estadístico

Existen diversas alternativas ampliamente utilizadas, entre las que destacan R (R Core Team (2025)), SPSS, Stata, SAS, Python y MATLAB.

Para la resolución del presente problema se ha optado por utilizar el lenguaje R, debido a su potencia, flexibilidad, enfoque orientado al análisis de datos complejos y ejemplos hallados en la bibliografía consultad.

R presenta ventajas en el contexto de modelos de regresión como los modelos de conteo cero inflados, y la existencia de paquetes para dicho propósito como `pscl`, `MASS` o `brms`.

Para la resolución del problema se ha usado el entorno de desarollo "R Studio 2024.12.1 Build 563", y el motor de R "R version 4.4.2 (2024-10-31 ucrt)".

6.2. Librerías utilizadas

Se enumeran las principales librería de R utilizadas:

- **dplyr**: Proporciona la manipulación y transformación de datos.
- **ggplot2**: Permite crear gráficos.
- **MASS**: Incluye funciones y conjuntos de datos clásicos para análisis estadísticos avanzados, como la regresión binomial negativa (`glm.nb()`).
- **pscl**: Implementa modelos de conteo extendidos como los modelos cero inflados (`zeroinfl()`).
- **lmtest**: Permite realizar pruebas estadísticas para modelos lineales y generalizados, incluyendo tests de razón de verosimilitud (`lrtest()`).
- **performance**: Ofrece métodos para calcular medidas sobre la calidad del modelo.

6.3. Funciones de R utilizadas más importantes

6.3.1. Función `glm()`

La función `glm()` en R (`Generalized Linear Model`) permite ajustar modelos lineales generalizados, que amplían el marco de los modelos lineales clásicos al permitir que la variable respuesta tenga una distribución perteneciente a la familia exponencial y que la media esperada esté relacionada con el predictor lineal mediante una función de enlace. Su sintaxis general es:

```
glm(formula, family = ..., data = ..., offset = ...)
```

El primer argumento, `formula`, especifica la estructura del modelo mediante una notación simbólica. Tiene la forma general:

$$\texttt{respuesta} \sim \texttt{predictor}_1 + \texttt{predictor}_2 + \ldots + \texttt{predictor}_k$$

donde el lado izquierdo de la fórmula representa la variable dependiente Y, y el lado derecho indica los predictores o covariables que forman parte del modelo.

El parámetro `family` especifica la distribución de probabilidad que se asume para la variable respuesta Y, así como la función de enlace. Algunas opciones son:

- `family = gaussian(link = "identity")`: modelo lineal clásico.
- `family = binomial(link = "logit")`: regresión logística.
- `family = poisson(link = "log")`: regresión de Poisson.
- `family = Gamma(link = "inverse")`: para datos positivos con varianza creciente.

El parámetro `offset` es importante en ciertos casos (ver 3.5.1). Si el modelo estándar tiene la forma

$$g(\mu_i) = \eta_i = \mathbf{x}_i'\beta$$

al incluir un offset o_i, el modelo se convierte en:

$$g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta} + o_i$$

Esto es especialmente útil cuando se desea modelar tasas en lugar de conteos brutos. Por ejemplo, en regresión de Poisson, si Y_i representa el número de eventos observados en una población de tamaño n_i, se desea modelar la tasa $\mu_i = \mathrm{E}[Y_i]/n_i$. Usando un enlace logarítmico, esto se traduce en:

$$\log(\mathrm{E}[Y_i]) = \log(n_i) + \mathbf{x}_i'\beta$$

donde $\log(n_i)$ se incorpora como *offset* en el modelo, asegurando que el efecto de la exposición esté representado de manera apropiada sin estimar un coeficiente adicional para él.

6.3.2. Función `glm.nb()`

La función `glm.nb()` permite ajustar modelos lineales generalizados en los que la variable respuesta sigue una distribución binomial negativa. La sintaxis de `glm.nb()` es muy similar a la de `glm()`:

```
glm.nb(formula, data = ..., )
```

La función `glm.nb()` puede entenderse como una extensión de `glm()` en la que se ajusta un modelo de la forma:

$$Y_i \sim \text{Binomial Negativa}(\mu_i, \theta)$$

donde $\mu_i = \mathrm{E}[Y_i]$ y θ es un parámetro adicional que representa la dispersión del modelo. La varianza bajo esta parametrización se define como:

$$\mathrm{Var}(Y_i) = \mu_i + \frac{\mu_i^2}{\theta}$$

A diferencia de `glm()`, que asume varianza igual a la media en el modelo Poisson, `glm.nb()` estima simultáneamente los coeficientes β del predictor lineal y el parámetro de dispersión θ mediante máxima verosimilitud marginal.

El `offset` se especifica en el parámetro `formula`, y cumple la misma función que en `glm()`. Si se desea modelar el número de eventos por unidad de exposición (por ejemplo, por individuo o por tiempo), se incorpora $\log(n_i)$ como *offset*. El modelo ajustado tendrá la forma:

$$\log(\mu_i) = \mathbf{x}_i'\beta + \log(n_i)$$

lo que equivale a modelar directamente la tasa de ocurrencia μ_i/n_i. En la práctica, esto se implementa con:

```
glm.nb(events ~ x1 + x2 + offset(log(n)), data = datos)
```

Esto asegura que la exposición o tamaño poblacional quede incorporado al modelo sin estimar un coeficiente adicional para él.

Por defecto, `glm.nb()` utiliza una función de enlace logarítmica (`link = "log"`), lo cual mantiene la estructura multiplicativa típica de los modelos de tasas. Sin embargo, también permite modificar la función de enlace si se requiere.

6.3.3. Función `zeroinfl()`

La función `zeroinfl()` del paquete `pscl` en R permite ajustar modelos de regresión de conteo con inflación de ceros, como el modelo Poisson cero inflado (ZIP) y el modelo binomial negativo cero inflado (ZINB).

La sintaxis general de la función es:

```
zeroinfl(formula, data = ..., offset = ..., dist = ...)
```

El parámetro `formula` especifica la estructura del modelo de regresión de forma extendida, ya que en el caso de los modelos cero inflados hay dos componentes: el modelo de conteo y el modelo logístico para la inflación de ceros. La fórmula tiene la forma:

$$\texttt{respuesta} \sim \texttt{predictores\_conteo} \mid \texttt{predictores\_inflado}$$

donde la parte a la izquierda del símbolo `~` indica la variable respuesta, la parte a la derecha (antes del `|`) contiene las covariables para el componente de conteo, y la parte después del `|` corresponde a las covariables del modelo logístico que estima la probabilidad de que una observación sea un cero estructural.

El parámetro `offset` tiene el mismo comportamiento que en la función `glm`. El offset afecta exclusivamente a la parte de conteo y se especifica fuera de la fórmula como un argumento adicional. Es fundamental para garantizar que las comparaciones entre unidades con diferentes niveles de exposición sean coherentes.

El parámetro `dist` indica la distribución que se utilizará para el componente de conteo del modelo. Puede tomar dos valores principales:

- `dist = "poisson"`: para un modelo Poisson cero inflado (ZIP).
- `dist = "negbin"`: para un modelo binomial negativo cero inflado (ZINB).

6.3.4. Comparación de modelos: `lrtest()`, `AIC()` y `BIC()`

Las funciones utilizadas para comparar modelos son `lrtest()`, `AIC()` y `BIC()`.

La función `lrtest()` implementa el contraste de razón de verosimilitudes (Likelihood Ratio Test, LRT) entre dos modelos anidados, es decir, donde uno (el restringido) es un caso particular del otro (el saturado o completo). La hipótesis nula establece que el modelo más simple es suficiente . El estadístico de contraste se define como:

$$D = 2\sum_{i=1}^{n} [\ell(\text{modelo completo}) - \ell(\text{modelo restringido})]$$

y sigue una distribución χ^2 bajo H_0, con grados de libertad igual al número de parámetros adicionales en el modelo completo (ver sección 3.10.1). Cabe decir que esta prueba solo es válida si ambos modelos han sido ajustados mediante máxima verosimilitud y sobre los mismos datos.

Este test también se podría llevar a cabo mediante un test ANOVA `anova(m1, m2, test = "Chisq")`, pero no es aplicable a modelos cero inflados por lo que se usa `lrtest()` para comparar modelos anidados.

La funciones `AIC()` y `BIC()` obtienen la medida descrita en los apartados 3.10.2 y 3.10.3. En ambos casos reciben como parámetros los diferentes modelos a evaluar.

Capítulo 7

Resultados

El código fuente, así como otros ficheros de interés, están accesibles públicamente en https://github.com/josebambu/TFG_Matematicas.

7.1. Estrategia del análisis de datos

Para estudiar la relación entre la concentración de arsénico en el agua y la incidencia de muertes por cáncer en distintas localidades, se ha seguido una metodología estructurada basada en el ajuste y la comparación de distintos modelos de regresión para datos de conteo. El análisis se ha desarrollado de forma progresiva, comenzando con una exploración y análisis visual de los datos, seguida del ajuste de modelos clásicos como la regresión de Poisson, y avanzando hacia modelos más complejos como los modelos cero inflados y de binomial negativa, que permiten capturar adecuadamente la sobredispersión y el exceso de ceros presentes en la variable respuesta.

En total se analizan 13 modelos diferentes:

Usando todos los datos

Modelo de Poisson

1. `Poisson_AllData: events ~ conc + gender + type`

Filtrado a valores de `conc` no nulos

Modelos de Poisson

2. `Poisson_CGT: events ~ conc + gender + type`
3. `Poisson_CT: events ~ conc + type`

Modelos de Poisson cero inflados

4. `ZIP_CGT_1: events ~ conc + gender + type | 1`
5. `ZIP_CGT_GT: events ~ conc + gender + type | gender + type`
6. `ZIP_C_1: events ~ conc | 1`
7. `ZIP_C_T: events ~ conc | type`

Modelos Binomial Negativa

8. `NB_CGT: events ~ conc + gender + type`
9. `NB_CT: events ~ conc + type`

Modelos Binomial Negativo Cero Inflado

10. `ZINB_CGT_1: events ~ conc + gender + type | 1`

11. `ZINB_CGT_GT: events ~ conc + gender + type | gender + type`
12. `ZINB_C_1: events ~ conc | 1`
13. `ZINB_C_T: events ~ conc | type`

7.2. Inspección y análisis visual de los datos

El conjunto de datos contiene un total de 2236 observaciones, estructuradas en siete variables: `village`, `conc`, `age`, `at_risk`, `events`, `gender` y `type`. La variable `village` identifica 43 aldeas distintas, cada una con exactamente 52 registros, lo que indica una estructura de datos completamente balanceada. Los 52 registros ($13 \times 2 \times 2$) corresponden a 13 rangos de edad (de 22.5 años a 82.5 años, en intervalos de 5 años), género (0: mujer, 1: hombre) y dos tipos de cáncer (0: vejiga, 1: pulmón).

Una característica notable es que únicamente la aldea número 1 presenta valores de concentración de arsénico (`conc`) iguales a cero en todas sus observaciones. Esta aldea puede considerarse como el total de población que no ha estado expuesta a contaminación por arsénico.

La variable `conc`, que representa la concentración de arsénico en el agua, presenta una alta variabilidad: los valores oscilan entre 0 y 934, con una media de aproximadamente 312.4 y una desviación estándar de 259.6. El primer cuartil se sitúa en 60, la mediana en 259 y el tercer cuartil en 529, lo que evidencia una distribución sesgada a la izquierda. En la figura7.1 se visualiza el histograma de aldeas según la concentración de arsénico `conc`. Puede verse una gran concentración alrededor del cero, y una distribución uniforme entre 200 y 800.

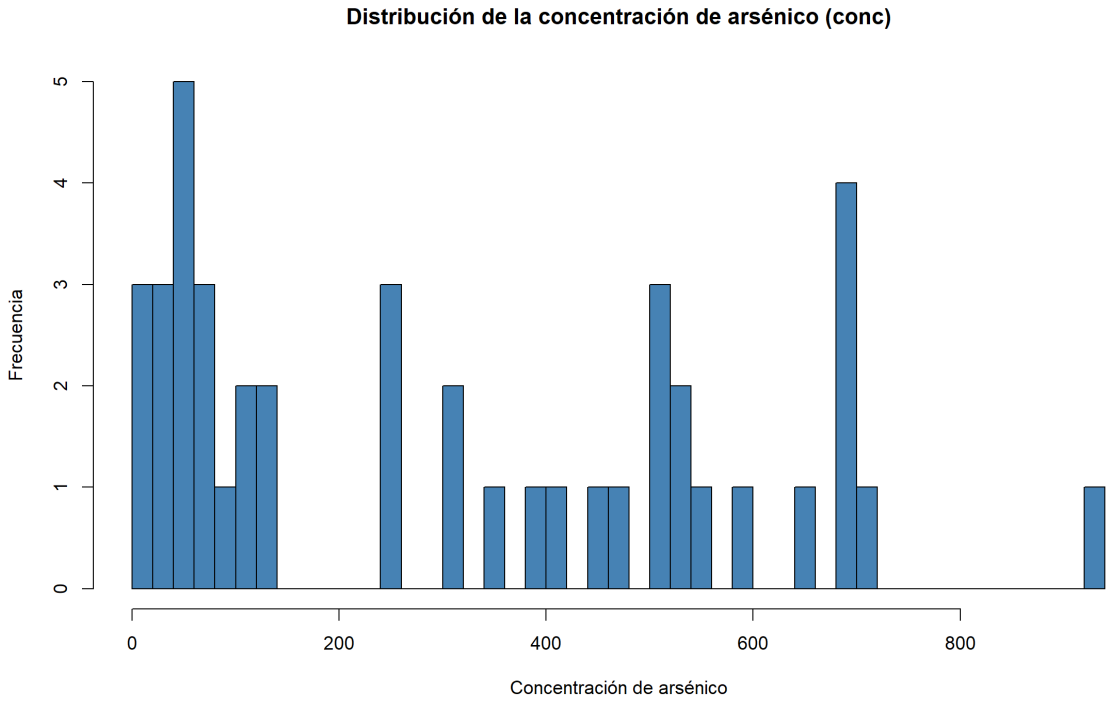

Figura 7.1: Número de aldeas para las diferentes concentraciones de arsénico.

Por otro lado, la variable respuesta `events`, que indica el número de muertes por cáncer observadas, presenta una gran cantidad de ceros. En total, hay **1838 observaciones con valor cero**, lo que representa aproximadamente el 82.2 % del total. Este exceso de ceros sugiere que

los modelos clásicos de Poisson pueden no ser adecuados, y motiva la exploración de modelos de conteo cero inflados. En la tabla 7.1 se muestra el número observaciones según los valores de casos de cáncer (`events`).

Casos de cáncer	Nº obs.
0	1838
1	282
2	51
3	16
4	3
≥ 5	46

Cuadro 7.1: Número de observaciones para los diferentes valores de `events`

La población total es 56 086 632 y hay 6109 casos de cáncer. La población no expuesta al arsénico es 55 104 170 y un total de 5671 casos de cáncer, y la población expuesta es 982 462, habiendo 438 casos de cáncer. Los porcentajes de casos de cáncer sobre la población total, no expuesta y expuesta a arsénico son 0.01089208 %, 0.01029142 % y 0.04458188 % respectivamente. El porcentaje de casos es cuatro veces mayor en las poblaciones expuestas a arsénico frente a la población no expuesta. Este dato confirma las evidencias científicas que la presencia de arsénico en el agua aumenta los casos de cáncer (Tseng et al. (1968) y Chen et al. (1985)).

En cuanto al tipo de cáncer, se han registrado 1225 casos de cáncer de vejiga y 4884 casos de cáncer de pulmón, haciendo un total de 6109 casos registrados.

7.3. Modelo de Poisson con todos los datos

Siguiendo la propuesta del enunciado el primer modelo corresponde a un GLM de Poisson. Tomando como covariables independientes `conc`, `gender` y `type`, utilizando la covariable `at_risk` como *offset*.

```
modelo_all_data <- glm(events ~ conc + gender + type,
               offset = log(at.risk),
               family = poisson(link = "log"),
               data = datos)
```

Los resultados obtenidos son los siguientes:

```
Call:
glm(formula = events ~ conc + gender + type, family = poisson(link = "log"),
    data = datos, offset = log(at.risk))

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.040e+01  3.384e-02 -307.44   <2e-16 ***
conc          2.881e-03  9.212e-05   31.28   <2e-16 ***
genderHombre  5.418e-01  2.705e-02   20.03   <2e-16 ***
typePulmón    1.383e+00  3.195e-02   43.28   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 17399  on 2235  degrees of freedom
Residual deviance: 14072  on 2232  degrees of freedom
AIC: 15111

Number of Fisher Scoring iterations: 7
```

Los p-valores indican que las tres covariables son significativas. El modelo estimado es el siguiente:

$$\texttt{events} = \exp\left(-10.40 + 0.002881\texttt{conc} + 0.5418\texttt{gender} + 1.383\texttt{type}\right) \cdot \texttt{at\_risk}$$

El resultado, además de influir la concentración de arsénico, indica que suceden más casos en hombres y cáncer tipo pulmón.

No es suficiente que los p-valores del modelo sean aceptados, debemos comprobar que el modelo no sufre de sobredispersión.

```
# Diagnóstico de sobredispersión
print(check_overdispersion(modelo_all_data))
```

Evaluamos la sobredispersión teniendo un valor de 18.792, lo que indica que se está produciendo sobredispersión. El p-valor del test que contrasta que el modelo se ajusta correctamente a los datos es < 0.001, lo cual indica que debemos mejorar el modelo.

```
     dispersion ratio =    18.792
Pearson's Chi-Squared = 41944.716
            p-value =   < 0.001
```

Podemos comparar el número de ceros observados, con el número de ceros esperados por el modelo:

```
ceros_observados <- sum(datos$events == 0)
mu_hat <- predict(modelo_all_data, type = "response")
prob_cero_poisson <- exp(-mu_hat)
ceros_esperados <- sum(prob_cero_poisson)
cat("Ceros observados:", ceros_observados, "\n")
cat("Ceros esperados por modelo Poisson:", round(ceros_esperados, 2), "\n")

Ceros observados: 1838
Ceros esperados por modelo Poisson: 1928.48
```

teniendo que el modelo genera más ceros de los esperados.

7.4. Modelo de Poisson

En esta sección se explican el segundo modelo y tercer modelo. A partir de esta sección los datos tratados son datos filtrados donde se han eliminado las observaciones donde `conc` es nula. Llamamos `datos_filtrados` a este subconjunto de datos:

```
datos_filtrados <- filter(datos, conc > 0)
```

sobre estos datos estimamos el modelo de Poisson:

```
modelo_Poisson_CGT <- glm(events ~ conc + gender + type,
                          offset = log(at.risk),
                          family = poisson(link = "log"),
                          data = datos_filtrados)
```

cuyo modelo obtenido es

```
Call:
glm(formula = events ~ conc + gender + type, family = poisson(link = "log"),
    data = datos_filtrados, offset = log(at.risk))

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)  -8.4449824  0.1144285 -73.801  < 2e-16 ***
conc          0.0014153  0.0001673   8.457  < 2e-16 ***
genderHombre -0.0107188  0.0956452  -0.112    0.911
typePulmón    0.4073681  0.0975529   4.176 2.97e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2313.9  on 2183  degrees of freedom
Residual deviance: 2226.2  on 2180  degrees of freedom
AIC: 2979.2

Number of Fisher Scoring iterations: 7
```

El estadístico del género (`genderHombre`) indica que puede omitirse al ser su p-valor 0.911. Por lo que se opta por reformular el modelo:

```
modelo_Poisson_CT <- glm(events ~ conc + type,
                         offset = log(at.risk),
                         family = poisson(link = "log"),
                         data = datos_filtrados)
```

teniendo el siguiente modelo

```
Call:
glm(formula = events ~ conc + type, family = poisson(link = "log"),
    data = datos_filtrados, offset = log(at.risk))

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.4506071  0.1028977 -82.126  < 2e-16 ***
conc         0.0014154  0.0001674   8.457  < 2e-16 ***
typePulmón   0.4073681  0.0975529   4.176 2.97e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2313.9  on 2183  degrees of freedom
Residual deviance: 2226.2  on 2181  degrees of freedom
AIC: 2977.2

Number of Fisher Scoring iterations: 7
```

El modelo resultante es el siguiente:

$$\texttt{events} = \exp\left(-8.4506071 + 0.0014154\texttt{conc} + 0.4073681\texttt{type}\right) \cdot \texttt{at\_risk}$$

Para dar por válido el modelo debemos analizar la sobredispersión, ejecutamos el mismo código que en el apartado anterior pero aplicado al modelo `modelo_Poisson_CT`.

```
       dispersion ratio =    4.287
  Pearson's Chi-Squared = 9346.479
                p-value =  < 0.001
```

El resultado obtenido indica que hay sobredispersión con un valor 4.287, y un p-valor < 0.001. Existe sobredispersión puede ser mejorada por un modelo cero inflado que será el modelo utilizado en el siguiente apartado.

En cuanto al número de ceros observados y el número de ceros esperados por el modelo se obtiene un mejor ajuste.

```
Ceros observados: 1836
Ceros esperados por modelo Poisson: 1824.42
```

7.5. Modelo de Poisson Cero Inflado

En esta sección se muestra los resultados del modelo de Poisson cero inflado. Analizaremos hasta cuatro configuraciones diferentes. Las dos primeras corresponden a los modelos con todas las covariables para la regresión del conteo, y difieren en el modelo logístico para la predicción de un cero estructural, `ZIP_CGT_1` sólo se usa el término independiente y el modelo `ZIP_CGT_GT` usa las covariables `gender` y `type`.

Los modelos `ZIP_C_1` y `ZIP_C_T` son la simplificación de los modelos `ZIP_CGT_1` y `ZIP_CGT_GT`, eliminando las covariables que no son significativas. En `ZIP_CGT_1` y `ZIP_CGT_GT` las variables `gender` y `type` no son significativas en el conteo, y en `ZIP_CGT_GT` la covariable `gender` no es significativa en el predictor de ceros.

7.5.1. ZIP_CGT_1: events ∼ conc + gender + type | 1

```
Call:
zeroinfl(formula = events ~ conc + gender + type | 1,
    data = datos_filtrados, offset = log(at.risk),
    dist = "poisson", link = "logit")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.5893 -0.4271 -0.3541 -0.2135 14.4615
```

```
Count model coefficients (poisson with log link):
               Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.5796496  0.1973828 -33.334  < 2e-16 ***
conc          0.0009921  0.0002249   4.412 1.03e-05 ***
genderHombre  0.1555888  0.1194892   1.302    0.193
typePulmón    0.2055962  0.1303098   1.578    0.115

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.93290    0.09549    9.77   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 16
Log-likelihood: -1346 on 5 Df
```

7.5.2. ZIP_CGT_GT: events ∼ conc + gender + type | gender + type

```
Call:
zeroinfl(formula = events ~ conc + gender + type | gender + type,
    data = datos_filtrados, offset = log(at.risk),
    dist = "poisson")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.6495 -0.4244 -0.3447 -0.2142 14.3164

Count model coefficients (poisson with log link):
               Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.2999325  0.2000926 -31.485  < 2e-16 ***
conc          0.0009001  0.0002246   4.008 6.11e-05 ***
genderHombre  0.1554483  0.1446142   1.075    0.282
typePulmón   -0.1263887  0.1480922  -0.853    0.393

Zero-inflation model coefficients (binomial with logit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.295737   0.155374   8.339  < 2e-16 ***
genderHombre  0.003386   0.161653   0.021 0.983289
typePulmón   -0.587250   0.160389  -3.661 0.000251 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 16
Log-likelihood: -1340 on 7 Df
```

7.5.3. ZIP_C_1: events ∼ conc | 1

```
Call:
```

```
zeroinfl(formula = events ~ conc | 1,
    data = datos_filtrados, offset = log(at.risk),
    dist = "poisson")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.5723 -0.4251 -0.3571 -0.2187 13.9465

Count model coefficients (poisson with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.3264749  0.1496646  -42.27  < 2e-16 ***
conc         0.0009448  0.0002223    4.25 2.14e-05 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.97411    0.08993   10.83   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 13
Log-likelihood: -1348 on 3 Df
```

7.5.4. ZIP_C_T: events ∼ conc | type

```
Call:
zeroinfl(formula = events ~ conc | type,
    data = datos_filtrados, offset = log(at.risk),
    dist = "poisson")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.6421 -0.4257 -0.3441 -0.2104 15.4239

Count model coefficients (poisson with log link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.3206197  0.1493423 -42.323  < 2e-16 ***
conc         0.0009387  0.0002220   4.227 2.37e-05 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.2496     0.1151  10.860  < 2e-16 ***
typePulmón   -0.5156     0.1342  -3.843 0.000122 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 10
Log-likelihood: -1341 on 4 Df
```

7.6. Modelo Binomial Negativo

En esta sección se evalúan dos modelos binomiales negativos, se parte del modelo `NB_CGT` con todas la covariables como regresoras. Este modelo sufre de infradispersión, resultando que la variable `gender` no es significativa con un p-valor 0.21739. En el modelo `NB_CT` se elimina dicha variable del modelo, pero no mejora la infradispersión.

7.6.1. NB_CGT: events ∼ conc + gender + type

```
Call:
glm.nb(formula = events ~ conc + gender + type + offset(log(at.risk)),
    data = datos_filtrados, init.theta = 0.188734946, link = log)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -8.056129   0.152650 -52.775  < 2e-16 ***
conc          0.001595   0.000256   6.229 4.69e-10 ***
genderHombre  0.167659   0.135920   1.234  0.21739
typePulmón    0.408208   0.136278   2.995  0.00274 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.1887) family taken to be 1)

    Null deviance: 1096.8  on 2183  degrees of freedom
Residual deviance: 1056.8  on 2180  degrees of freedom
AIC: 2691.7

Number of Fisher Scoring iterations: 1


              Theta:  0.1887
          Std. Err.:  0.0201

 2 x log-likelihood:  -2681.7450
```

La dispersión obtenida con el comando `check_overdispersion`.

```
 dispersion ratio =   0.248
          p-value = < 0.001
```

7.6.2. NB_CT: events ∼ conc + type

```
Call:
glm.nb(formula = events ~ conc + type + offset(log(at.risk)),
    data = datos_filtrados, init.theta = 0.1892531954, link = log)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.9771938  0.1350567 -59.066  < 2e-16 ***
```

```
conc         0.0015987  0.0002557   6.251 4.07e-10 ***
typePulmón   0.4143452  0.1361483   3.043  0.00234 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.1893) family taken to be 1)

    Null deviance: 1098.1  on 2183  degrees of freedom
Residual deviance: 1059.3  on 2181  degrees of freedom
AIC: 2691

Number of Fisher Scoring iterations: 1


              Theta:  0.1893
          Std. Err.:  0.0202

 2 x log-likelihood:  -2682.9520
```

La infradispersión no mejora con la eliminación de la variable `gender`.

```
 dispersion ratio =   0.245
          p-value = < 0.001
```

7.7. Modelo Binomial Negativo Cero Inflado

Por último, se presentan los resultados usando un modelo binomial negativo cero inflado. Se presentan cuatro configuraciones diferentes, se parte de dos configuraciones con todas las covariables para la regresión de conteo. El modelo ZINB_CGT_1 usa únicamente el término independiente para la regresión logística de cero estructurales y ZINB_CGT_GT usa las variables `gender` y `type` para los ceros estructurales.

El modelo ZINB_C_1 es el ajuste de ZINB_CGT_1 donde se eliminan las covariables `gender` y `type` al no ser significativas. Por otra parte, el modelo ZINB_C_T es el ajuste de ZINB_CGT_GT, donde se elimina las mismas covariables en la regresión de conteo y se elimina `gender` en la regresión de ceros estructurales.

En los cuatro modelos $\log\theta$ no se considera significativa, recordemos que un modelo binomial negativa cero inflado modela la varianza como $\text{Var}(Y) = \mu + \frac{\mu^2}{\theta}$. Si dicha covariable no es significativa implica que no hay una gran sobredispersión respecto al modelo de Poisson cero inflado. Esto se tendrá en cuenta para elegir el modelo final.

7.7.1. ZINB_CGT_1: `events ~ conc + gender + type | 1`

```
Call:
zeroinfl(formula = events ~ conc + gender + type | 1,
    data = datos_filtrados, offset = log(at.risk),
    dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max
```

```
-0.4837 -0.3834 -0.3260 -0.2084 15.3400

Count model coefficients (negbin with log link):
               Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.1685249  0.3705774 -19.344  < 2e-16 ***
conc          0.0014534  0.0003104   4.682 2.84e-06 ***
genderHombre  0.1808055  0.1458653   1.240   0.2151
typePulmón    0.3477431  0.1553611   2.238   0.0252 *
Log(theta)   -0.1511033  0.7365956  -0.205   0.8375

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.2900     0.4809   0.603    0.546
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 0.8598
Number of iterations in BFGS optimization: 25
Log-likelihood: -1340 on 6 Df
```

7.7.2. ZINB_CGT_GT: events ∼ conc + gender + type | gender + type

```
Call:
zeroinfl(formula = events ~ conc + gender + type | gender + type,
    data = datos_filtrados, offset = log(at.risk),
    dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.5505 -0.3787 -0.3179 -0.2099 14.1788

Count model coefficients (negbin with log link):
               Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.7226897  0.3543558 -18.972  < 2e-16 ***
conc          0.0013250  0.0003225   4.108 3.98e-05 ***
genderHombre  0.2840401  0.2125381   1.336    0.181
typePulmón   -0.2061671  0.2192092  -0.941    0.347
Log(theta)    0.0847056  0.7131495   0.119    0.905

Zero-inflation model coefficients (binomial with logit link):
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.8160     0.3896   2.095  0.03621 *
genderHombre   0.1315     0.2380   0.553  0.58048
typePulmón    -0.7491     0.2591  -2.891  0.00383 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 1.0884
Number of iterations in BFGS optimization: 49
```

```
Log-likelihood: -1334 on 8 Df
```

7.7.3. ZINB_C_1: events ~ conc | 1

```
Call:
zeroinfl(formula = events ~ conc | 1,
    data = datos_filtrados, offset = log(at.risk),
    dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.4822 -0.3867 -0.3342 -0.2136 14.2356

Count model coefficients (negbin with log link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.7244347  0.3002726 -22.394  < 2e-16 ***
conc         0.0013636  0.0003227   4.226 2.38e-05 ***
Log(theta)   0.1082563  0.7223567   0.150    0.881

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.4935     0.3678   1.342     0.18
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 1.1143
Number of iterations in BFGS optimization: 16
Log-likelihood: -1343 on 4 Df
```

7.7.4. ZINB_C_T: events ~ conc | type

```
Call:
zeroinfl(formula = events ~ conc | type,
    data = datos_filtrados, offset = log(at.risk),
    dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.5360 -0.3853 -0.3204 -0.2091 15.7158

Count model coefficients (negbin with log link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.7065360  0.2946315 -22.762  < 2e-16 ***
conc         0.0013525  0.0003226   4.192 2.77e-05 ***
Log(theta)   0.1405026  0.7180773   0.196    0.845

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.8206     0.3329   2.465 0.013694 *
```

```
typePulmón   -0.5870     0.1709  -3.435 0.000593 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 1.1509
Number of iterations in BFGS optimization: 30
Log-likelihood: -1336 on 5 Df
```

7.8. Comparativa de modelos

Una vez ajustados los distintos modelos resulta esencial comparar su capacidad explicativa y su adecuación a los datos. Para ello, se han empleado dos criterios de selección de modelos. En primer lugar, el *Criterio de Información de Akaike* (AIC) y el *Criterio de Información Bayesiano* (BIC) permiten evaluar el compromiso entre ajuste y complejidad del modelo, penalizando el número de parámetros estimados. Modelos con valores más bajos de AIC o BIC son preferibles, aunque BIC impone una penalización más severa, especialmente en muestras grandes. Además, cuando los modelos son anidados, se ha aplicado el test de razón de verosimilitudes (`lrtest`), que permite contrastar formalmente si el modelo más complejo proporciona una mejora estadísticamente significativa del ajuste frente al modelo más simple.

7.8.1. AIC y BIC

Los resultados AIC son los siguientes

```
> AIC(modelo_all_data,
+          modelo_Poisson_CGT, modelo_Poisson_CT,
+          modelo_ZIP_CGT_1, modelo_ZIP_CGT_GT, modelo_ZIP_C_1, modelo_ZIP_C_T,
+          modelo_NB_CGT, modelo_NB_CT,
+          modelo_ZINB_CGT_1, modelo_ZINB_CGT_GT, modelo_ZINB_C_1, modelo_ZINB_C_T)

                   df       AIC
modelo_all_data     4 15110.624
modelo_Poisson_CGT  4  2979.230
modelo_Poisson_CT   3  2977.243
modelo_ZIP_CGT_1    5  2702.751
modelo_ZIP_CGT_GT   7  2693.813
modelo_ZIP_C_1      3  2702.942
modelo_ZIP_C_T      4  2690.216
modelo_NB_CGT       5  2691.745
modelo_NB_CT        4  2690.952
modelo_ZINB_CGT_1   6  2691.172
modelo_ZINB_CGT_GT  8  2684.545
modelo_ZINB_C_1     4  2694.060
modelo_ZINB_C_T     5  2681.533
```

y los resultados BIC no difieren mucho, siendo `ZINB_C_T` el mejor modelo en ambos casos.

```
> BIC(modelo_all_data,
+     modelo_Poisson_CGT, modelo_Poisson_CT,
+     modelo_ZIP_CGT_1, modelo_ZIP_CGT_GT, modelo_ZIP_C_1, modelo_ZIP_C_T,
```

```
+     modelo_NB_CGT, modelo_NB_CT,
+     modelo_ZINB_CGT_1, modelo_ZINB_CGT_GT, modelo_ZINB_C_1, modelo_ZINB_C_T)

                    df       BIC
modelo_all_data      4 15133.474
modelo_Poisson_CGT   4  3001.986
modelo_Poisson_CT    3  2994.310
modelo_ZIP_CGT_1     5  2731.196
modelo_ZIP_CGT_GT    7  2733.635
modelo_ZIP_C_1       3  2720.008
modelo_ZIP_C_T       4  2712.972
modelo_NB_CGT        5  2720.190
modelo_NB_CT         4  2713.708
modelo_ZINB_CGT_1    6  2725.305
modelo_ZINB_CGT_GT   8  2730.057
modelo_ZINB_C_1      4  2716.816
modelo_ZINB_C_T      5  2709.978
```

Los resultados de los criterios AIC y BIC muestran diferencias sustanciales entre los modelos considerados. El modelo ajustado con todos los datos sin filtrar (`modelo_all_data`) presenta los valores más elevados tanto en AIC (15 110.6) como en BIC (15 133.5), lo que indica un mal ajuste relativo en comparación con el resto. Tras filtrar los datos para excluir la aldea no expuesta y considerar únicamente los casos con concentración positiva, los modelos mejoran considerablemente su rendimiento. Entre los modelos Poisson, el `Poisson_CT` presenta el menor AIC (2977.2) y BIC (2994.3), siendo preferible dentro de esta familia.

Sin embargo, los modelos cero inflados muestran un ajuste notablemente superior. Dentro de los modelos ZIP, el `ZIP_C_T` presenta los mejores valores (AIC = 2690.2, BIC = 2713.0), aunque modelos binomial negativo (NB) y binomial negativo cero inflado (ZINB) ofrecen también valores competitivos.

En particular, el modelo `ZINB_C_T` alcanza el menor AIC (2681.5) y el menor BIC (2709.9), lo que sugiere que, pese a su mayor complejidad, proporciona el mejor equilibrio entre ajuste y penalización. Por tanto, desde el punto de vista de la selección de modelos basada en estos criterios de información, el modelo `ZINB_C_T` se perfila como el más adecuado para describir la relación entre la concentración de arsénico y los eventos de cáncer.

Mediante el test de razón de verosimilitud evaluemos si los modelos completos `ZINB_CGT_GT` y `ZIP_CGT_GT` son más adecuados frente a `ZINB_C_T` y `ZIP_C_T` respectivamente. La hipótesis nula H_0 es que ambos modelos son similares, por lo que nos quedamos con el modelo más sencillo, frente a H_1 que existen diferencias significativas entre los modelos, y elegiremos el modelo más completo.

Para el caso ZINB el p-valor 0.3915 indica que las diferencias no son significativas, por lo que podemos usar el modelo más simple.

```
> lrtest(modelo_ZINB_C_T, modelo_ZINB_CGT_GT)
Likelihood ratio test

Model 1: events ~ conc | type
Model 2: events ~ conc + gender + type | gender + type
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   5 -1335.8
2   8 -1334.3  3 2.9881     0.3935
```

En el caso de ZIP el p-valor es 0.493 por lo que también es más correcto usar el modelo más simple al no existir diferencias significativas.

```
> lrtest(modelo_ZIP_C_T, modelo_ZIP_CGT_GT)
Likelihood ratio test

Model 1: events ~ conc | type
Model 2: events ~ conc + gender + type | gender + type
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   4 -1341.1
2   7 -1339.9  3 2.4034      0.493
```

Una vez elegidos estos modelos más simples, `ZIP_C_T` y `ZINB_C_T`, no es posible aplicar el test de ratio de verosimilitud para determinar si hay diferencias significativas. No es posible debido que no son modelos anidados en sentido estricto.

7.8.2. Modelo seleccionado

Dos motivos conducen a elegir el modelo `ZIP_C_T`. Primero, recordemos que $\log\theta$ no se considera significativa en los modelos `ZINB`. Segundo, los valores AIC y BIC son próximos, `ZIP_C_T` (AIC = 2690.216 y BIC = 2712.972) y `ZINB_C_T` (AIC = 2681.533 y BIC = 2709.978) y no es posible aplicar el test de ratio de verosimilitudes. Aunque podemos considerar cualquiera de los dos modelos como candidatos válidos, vamos a profundizar en el modelo de Poisson cero inflado .

El modelo `ZIP_C_T` corresponde a un modelo de Poisson cero inflado, con función de enlace logarítmica para la componente de conteo y enlace logit para la componente de inflación de ceros. El modelo ajustado es el siguiente:

$$\begin{aligned} \text{Conteo:} \quad & \mu = \exp\left(-6.321 + 0.0009387 \cdot \texttt{conc}\right) \cdot \texttt{at\_risk} \\ \text{Inflación de ceros:} \quad & \text{logit}(\pi) = 1.250 - 0.5156 \cdot \texttt{type} \end{aligned}$$

donde:

- μ es la media del número de eventos para la observación i (tasa ajustada por población).
- π es la probabilidad de que la observación pertenezca al componente estructural de ceros.
- `conc` es la concentración de arsénico en la observación.
- `type` es una variable indicadora para el tipo de cáncer (0 vejiga, 1 pulmón).
- `at_risk` representa la población.

En cuanto a los ceros estructurales, partiendo del modelo ajustado para la inflación de ceros, se tiene la expresión:

$$\begin{aligned} \text{logit}(\pi) &= 1.250 - 0.5156 \cdot \texttt{type} \\ \pi &= \frac{1}{1 + e^{-(1.250 - 0.5156 \cdot \texttt{type})}} \end{aligned}$$

- Para `type = 0` (cáncer de vejiga) $\pi = \frac{1}{1+e^{-1.250}} \approx 0.7772$
- Para `type = 1` (cáncer de pulmón) $\pi = \frac{1}{1+e^{-0.744}} \approx 0.6758$

Por tanto, la probabilidad estimada de que una observación pertenezca a la componente de ceros estructurales es mayor en los casos de cáncer de vejiga ($\pi \approx 0.7772$) que en los de cáncer de pulmón ($\pi \approx 0.6758$).

En cuanto al conteo, la figura 7.2 muestra el valor medio de casos esperado por 10 000 habitantes.

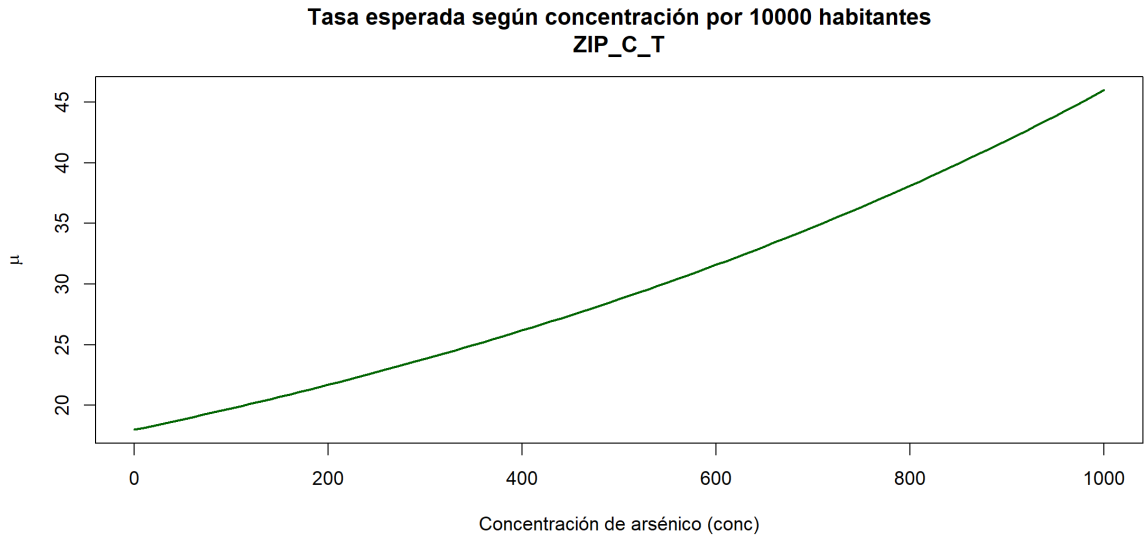

Figura 7.2: Número de casos esperados por 10 000 habitantes $\mu = 10000 \cdot \exp(-6.32062 + 0.0009386554 \cdot \texttt{conc})$ en el modelo `ZIP_C_T`.

Capítulo 8

Conclusiones

El estudio parte de los modelos de regresión lineal simple y modelos de regresión lineal múltiple, para avanzar a los modelos lineales generalizados (GLM), realizando un estudio riguroso sobre los mismos. Se ha profundizado en modelos existentes para el análisis de datos de conteo con exceso de ceros, como los modelos de Poisson cero inflados (ZIP) y los modelos binomiales negativos cero inflados (ZINB). A través de un recorrido teórico y aplicado, se ha demostrado cómo la flexibilidad de estos modelos permite abordar situaciones en las que las suposiciones clásicas de Poisson, como la igualdad entre media y varianza o la ausencia de ceros estructurales, no se cumplen.

El análisis empírico basado en datos reales sobre la incidencia de cáncer en función de la concentración de arsénico ha evidenciado la importancia de considerar la sobredispersión y la inflación de ceros. El modelo de Poisson ajustado a la totalidad de los datos ofrecía un ajuste claramente inadecuado, con una sobredispersión elevada y una gran proporción de ceros no explicada por el modelo. La exclusión de observaciones no expuestas y la consideración de modelos alternativos mejoraron sensiblemente los resultados.

Entre los modelos comparados, el modelo `ZINB_C_T` se posiciona como el más adecuado, al ofrecer el mejor compromiso entre calidad del ajuste (medido mediante AIC y BIC). Este modelo incorpora de manera tanto la relación positiva entre la concentración de arsénico y los eventos de cáncer como la probabilidad diferencial de aparición de ceros según el tipo de cáncer. Además, el contraste de razón de verosimilitudes permitió concluir que no es necesario recurrir a modelos más complejos que incluyan otras covariables no significativas.

Como alternativa se presenta como solución el modelo `ZIP_C_T`, que también ofrece un ajuste excelente con un número reducido de parámetros y una interpretación sencilla. Este modelo incluye la variable `conc` en el componente de conteo, y la variable `type` en el componente de inflación de ceros, permitiendo distinguir entre ceros estructurales y ceros aleatorios. El resultado evidencia una relación positiva entre la concentración de arsénico y la tasa de incidencia de cáncer, así como una menor probabilidad de pertenecer al componente de ceros estructurales en los casos de cáncer de pulmón. El valor del AIC obtenido, junto con la significación estadística de los coeficientes, posiciona al modelo `ZIP_C_T` como una alternativa factible como modelo.

Bibliografía

Chen, C. J., Y. C. Chuang, T. M. Lin, and H. Y. Wu (1985, November). Malignant neoplasms among residents of a blackfoot disease-endemic area in taiwan: high-arsenic artesian well water and cancers. *Cancer Res*@(11 Pt 2), 5895–5899.

Faraway, J. J. (2016). *Extending the Linear Model with R* (2nd ed.). Chapman and Hall/CRC.

García-Pérez, A. (2013). *Cuadernos de Estadística Aplicada: Biología y Ciencias Ambientales*, Chapter Problema 4.11. Editorial UNED.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*@(1), 1–14.

Lundberg, S. M. and S.-I. Lee (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, Red Hook, NY, USA, pp. 4768–4777. Curran Associates Inc.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models* (2nd ed.). London: Chapman and Hall.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*@(3), 370–384.

Peña, D. (2010). *Regresión y diseño de experimentos* (2 ed.). Madrid: Alianza Editorial.

Pérez, A. G. (2021). *Estadística aplicada avanzada con R*. Madrid: UNED.

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ribeiro, M. T., S. Singh, and C. Guestrin (2016). "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, New York, NY, USA, pp. 1135–1144. Association for Computing Machinery.

Tseng, W. P., H. M. Chu, S. W. How, J. M. Fong, C. S. Lin, and S. Yeh (1968, 03). Prevalence of skin cancer in an endemic area of chronic arsenicism in taiwan2. *JNCI: Journal of the National Cancer Institute*@(3), 453–463.