USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Modelos de regresión para datos de recuento con ceros truncados, ceros inflados y ceros apartados

Saborido Muñiz, Martín

2023-2024

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRADO DE MATEMÁTICAS

Traballo Fin de Grao

Modelos de regresión para datos de recuento con ceros truncados, ceros inflados y ceros apartados

Saborido Muñiz, Martín

Julio, 2024

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Conocimiento: Estadística, Análisis Matemático y Optimización
Título: Modelos de regresión para datos de recuento con ceros truncados, ceros inflados y ceros apartados
Breve descripción del contenido
En los datos de recuento, que consisten en observaciones de cantidades enteras no negativas, como el número de individuos de una organización, el número de recaídas de una enfermedad o el número de solicitudes de un servicio, suele aparecer un comportamiento especial en la frecuencia de ceros, bien por ausencia de ceros (ceros truncados), por subpoblaciones con ceros estructurales (ceros inflados) o por una mezcla de ambas (ceros apartados). En este trabajo estudiaremos modelos de regresión cuya variable respuesta sea un recuento con alguna de estas características especiales en el valor cero.
Recomendaciones
Otras observaciones

Índice

Resumen viii

Introducción xi

1. Modelos de regresión para variables de recuento ordinarias **1**

1.1. El modelo de regresión de Poisson. Estimación de parámetros 1

1.2. Inferencia 5

1.3. Predicción 6

1.4. Diagnosis 6

1.5. El fenómeno de sobre-dispersión 8

1.5.1. Modelo de regresión Binomial Negativa 8

1.5.2. Estimación ad hoc de la sobre-dispersión 11

2. Modelos de regresión para variables de recuento con ceros truncados **13**

2.1. Introducción 13

2.2. La distribución de Poisson truncada en el cero 14

2.3. El modelo de Poisson truncado en el cero 16

2.4. El modelo Binomial Negativo truncado en el cero 20

3. Modelos de regresión para variables de recuento con ceros inflados **23**

3.1. Introducción 23

3.2. La distribución de Poisson inflada en cero 24
3.3. El modelo de Poisson inflado en el cero 27
3.4. El modelo Binomial Negativo inflado en el cero 33

4. Modelos de regresión para variables de recuento con ceros apartados 37

4.1. Introducción 37
4.2. La distribución de Poisson con ceros apartados 39
4.3. El modelo de Poisson con ceros apartados 41
4.4. El modelo Binomial Negativo con ceros apartados 46

I. Códigos de R 51

I.1. Capítulo 1 51
I.1.1. Distribución Binomial Negativa 51
I.1.2. Modelos usuales (ejemplo de las telas) 53
I.2. Capítulo 2 56
I.2.1. Distribución de Poisson truncada 56
I.2.2. Modelos con ceros truncados (ejemplo de las viviendas) 58
I.3. Capítulo 3 60
I.3.1. Distribución de Poisson inflada 60
I.3.2. Modelos con ceros inflados (ejemplo de los taxis) 64
I.4. Capítulo 4 67
I.4.1. Distribución de Poisson con ceros apartados 67
I.4.2. Modelos con ceros apartados (ejemplo de los fumadores) 72

Bibliografía 77

[

]

Resumen

En este trabajo se han estudiado los modelos de regresión en los cuales la variable respuesta es un recuento que presenta anomalías en el valor cero, ya sea por ausencia, defecto o exceso, que nos obligan a considerar otros tipos de distribuciones de recuento diferentes a la distribución de Poisson o la Binomial Negativa. Se han estudiado las distribuciones con ceros truncados, ceros inflados y ceros apartados, que proporcionan una manera de actuar ante esas anomalías, y se han construido modelos de regresión para variables de recuento que siguen una de esas distribuciones, ilustrando su aplicación con ejemplos.

Abstract

In this work we have studied regression models in which the response variable is a count that presents anomalies in the zero value, whether due to absence, defect or excess, that force us to consider different types of distributions other than the Poisson or the Negative Binomial distribution. We have studied the zero-truncated, zero-inflated and hurdle distributions, that provide us with a way to act in the event of those anomalies, and we have built regression models for each one of those distributions, illustrating their application with examples.

Introducción

Las variables de recuento son algo muy común de observar en el día a día. Por ejemplo, la cantidad de huevos que utilizamos para hacer una tortilla, el número de tetrabriks de leche que lleva un individuo en su carrito del supermercado en un momento dado o el número de países extranjeros que ha visitado una persona en los últimos dos años se corresponden con variables de recuento. Lo que tienen en común este tipo de variables es que todas toman exclusivamente valores enteros no negativos.

Cuando trabajamos con una variable de recuento y construimos un modelo de regresión lineal para hacer una predicción sobre ella, nos enfrentamos a un problema importante: la distribución de los errores no es la distribución normal. Es por ello que necesitamos la introducción de los modelos lineales generalizados, como el modelo de Poisson o el Binomial Negativo, que nos permitirán solventar este problema.

Además de esto, una variable de recuento puede presentar anomalías, sobre todo en el valor cero, que nos impidan la aplicación de estos modelos de una manera exitosa. Por ejemplo, puede ocurrir que una variable no pueda tomar el valor cero, como con la cantidad de huevos necesarios para hacer una tortilla (no se puede cocinar una tortilla sin huevos). La primera situación se solventará con la introducción de una nueva distribución, la distribución con ceros truncados, que tiene en cuenta la condición de que la variable no puede valer cero, y nos permite realizar regresión sobre ella y obtener mejores resultados. Pero también puede ocurrir al contrario, que tome el valor cero más o menos veces de lo esperado. Puede haber un exceso de ceros, como ocurre en el número de tetrabriks de leche observados en un carrito de la compra (si una persona no lleva leche, o bien no la va a comprar, o bien la va a comprar pero todavía no la ha metido en el carrito), o puede ser que los ceros sean todos estructurales, como ocurre en el número de países extranjeros visitados (si una persona ha viajado al extranjero, habrá visitado al menos un país, mientras que si no ha viajado, no habrá visitado ninguno). En esta situación, podemos estar ante un fenómeno de ceros inflados, en el caso de los tetrabriks, o ceros apartados, en el caso de los viajes al extranjero, dependiendo de cual sea el origen de esos ceros, y se introducen las distribuciones con ceros inflados y ceros apartados que nos resolverán el problema para que

podamos realizar regresión correctamente.

En el capítulo 1 de este trabajo introduciremos los modelos de regresión para variables de recuento ordinarias, así como el concepto de sobre-dispersión. En el capítulo 2 estudiaremos las distribuciones con ceros truncados, en el capítulo 3, las distribuciones con ceros inflados, y finalmente, en el capítulo 4, las distribuciones con ceros apartados. En todos ellos construiremos los modelos de regresión apropiados para variables de recuento que sigan esas distribuciones, y apoyaremos su aplicación con ejemplos.

Capítulo 1

Modelos de regresión para variables de recuento ordinarias

Si aplicamos un modelo de regresión lineal sobre unos datos de recuento, puede ocurrir, y ocurrirá con gran frecuencia, que no se cumplan las hipótesis básicas del propio modelo: linealidad, homocedasticidad y normalidad. Por ejemplo, podría ser que la recta de ajuste cruce el eje de abscisas, en cuyo caso habría predicciones negativas para la variable, y eso no es posible.

Una solución que puede ser útil es aplicar la raíz cuadrada a la variable respuesta, pero realmente no nos solucionaría demasiado, ya que los datos seguirían siendo discretos. Además, podría dar problemas en los ceros y los unos, ya que son invariantes por la raíz cuadrada.

El modelo de regresión de Poisson es un remedio que, en la gran mayoría de casos, soluciona el problema al que nos enfrentamos cuando la variable que se quiere predecir es de recuento. Durante este capítulo, introduciremos este modelo, así como el concepto de sobre-dispersión y una alternativa para solucionarlo.

1.1. El modelo de regresión de Poisson. Estimación de parámetros

La función de masa de probabilidad de Poisson tiene la siguiente expresión:

$$F(y) = \frac{\lambda^y}{k!}e^{-y},$$

donde y toma valores enteros no negativos. La distribución de Poisson tiene media λ, como ya es bien conocido.

Ahora bien, si queremos predecir el valor de una variable Y de recuento en función de otras variables explicativas X, entonces el modelo de regresión de Poisson, que supone que la variable

Y tiene distribución de Poisson condicionalmente al valor de X, nos ayudará en nuestra meta. En otras palabras, debemos construir un modelo para describir

$$\lambda(x) = E[Y/X = x].$$

Para empezar, debido a que la variable respuesta es de recuento y no toma valores negativos, necesitamos aplicar una función sobre esa variable, de forma que nos sirva de unión con un modelo lineal, en concreto,

$$g(\lambda(x, \beta)) = x'\beta,$$

siendo β el vector de coeficientes del modelo lineal.

Una función intuitiva para ello, dado el dominio del parámetro de Poisson, que es el intervalo $(0, \infty)$, es el logaritmo. Así, la función de regresión tendría la expresión

$$\log(\lambda(x, \beta)) = x'\beta,$$

o lo que es lo mismo,

$$\lambda(x, \beta) = e^{x'\beta}.$$

Se puede encontrar más información sobre este modelo, conocido como modelo log-lineal, en el Zuur [1]. En el caso particular de que solamente haya una variable explicativa, x, la función de regresión tiene la siguiente forma:

$$\lambda(x, \beta) = e^{\beta_0 + \beta_1 x} = e^{\beta_0} e^{\beta_1 x}.$$

Ahora debemos estimar el vector de parámetros β a partir de una muestra de datos de la forma $(X_1, Y_1), \ldots, (X_n, Y_n)$. Para ello, se utiliza el método de máxima verosimilitud, donde la función de verosimilitud es:

$$L(\beta) = \prod_{i=1}^{n} e^{-\lambda(x_i,\beta)} \frac{\lambda(x_i, \beta)^{y_i}}{y_i!}.$$

Procediendo de forma habitual, considerando el logaritmo, derivando e igualando a cero, se calcula el vector de parámetros β. Más concretamente, se resuelven las siguientes ecuaciones en β:

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^{n} x_i' \left(y_i - \lambda(x_i, \beta)\right) = 0.$$

En lo que a la matriz hessiana se refiere, basta con derivar de nuevo, y se tiene

$$\frac{\partial^2 \log L(\beta)}{\partial \beta \partial \beta'} = -\sum_{i=1}^{n} x_i x_i' \lambda(x_i, \beta).$$

Así, resolviendo las ecuaciones anteriores por un método iterativo, por ejemplo, Newton-Raphson, se calcula la estimación de los parámetros β.

Ejemplo 1.1. Ilustraremos los conceptos de este capítulo utilizando el dataset `cloth` de la librería `boot` [7, 8]. En él, se hace un recuento del número de defectos, tales como roturas, encontrados en 32 piezas de tela de longitudes distintas, medidas en cientos de metros. En la Figura 1.1 se pueden observar los fenómenos que ya hemos explicado: la tendencia de los datos no parece completamente lineal, y además la dispersión en longitudes grandes es mayor en general que en longitudes pequeñas de tela.

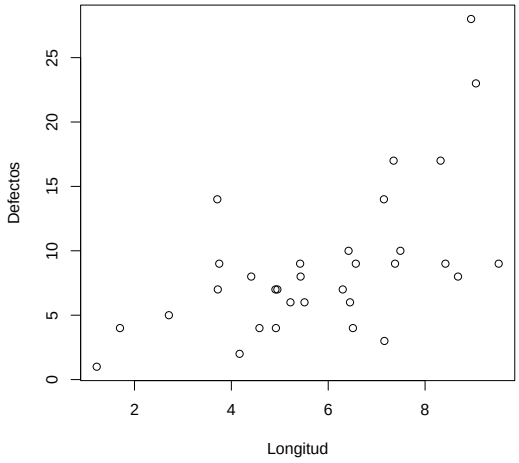

Figura 1.1: Diagrama de dispersión del número de defectos de una pieza de tela frente a su longitud.

El ajuste de los coeficientes del modelo de Poisson se puede llevar a cabo con la función `glm`, disponible en la instalación básica de R [2].

```
> modelo = glm(Flaws~Length, data = cloth, family = poisson(link = log))
> summary(modelo)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.97175    0.21247   4.574 4.79e-06 ***
Length       0.19297    0.03063   6.300 2.97e-10 ***
---
```

Observamos que las estimaciones de los coeficientes son 0.972 para el intercepto, y 0.183 para la pendiente. Sin embargo, debemos aplicar la exponencial de esos coeficientes para obtener una

interpretación de los mismos, situándolos en la escala del parámetro de Poisson.

Así, la exponencial del intercepto es 2.643. Esto quiere decir que, si la longitud de la tela fuese nula, el número medio de defectos esperado sería 2.643. Es cierto que el valor cero no se encuentra en el rango de la variable explicativa, pero se interpretaría como un punto de partida formal del modelo para una longitud nula.

En cuanto al estimador de la pendiente, ahora ya no se interpreta como una pendiente, sino como una tasa de crecimiento (o decrecimiento) de la variable respuesta, es decir, el número de defectos, por cada unidad que aumentemos la variable explicativa, es decir, la longitud de la pieza de tela. Así, por ejemplo, la cantidad media esperada de defectos de una tela de longitud $L+1$ será la esperada para L, multiplicada por la tasa, que vale 1.2129 (o es del 21.29 %).

$$\lambda(L+1,\beta) = e^{\beta_0} e^{\beta_1(L+1)} = e^{\beta_1}\left(e^{\beta_0} e^{\beta_1 L}\right) = e^{\beta_1}\lambda(L,\beta).$$

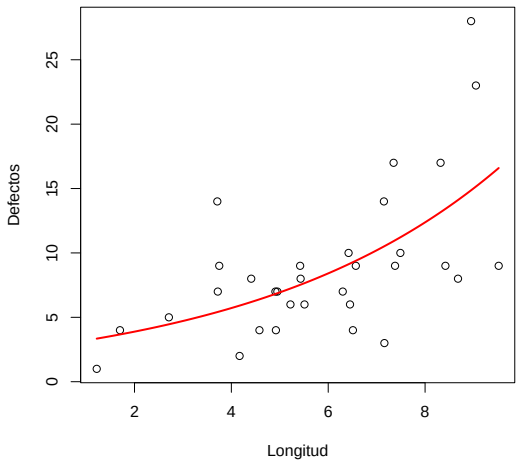

Figura 1.2: Ajuste del modelo de Poisson sobre el diagrama de dispersión de los datos.

En la Figura 1.2 podemos observar la recta de ajuste del modelo de Poisson sobre nuestros datos. Se aprecia que la predicción en el cero es en efecto 2.643, y la pendiente de la curva es cada vez mayor, de acuerdo a la tasa de crecimiento del 21.29 %.

1.2. Inferencia

Para construir intervalos de confianza o contrastar hipótesis de significación para los parámetros, se puede utilizar, o bien el perfil de verosimilitud para los intervalos, o bien la distribución asintótica.

El perfil de verosimilitud, o *profile likelihood* en inglés de un coeficiente β_i se define de la siguiente forma:

$$PL(\beta_i) = \max_{\beta_1,\ldots,\beta_{i-1},\beta_i,\beta_{i+1},\ldots,\beta_p} L(\beta_1,\ldots,\beta_{i-1},\beta_i,\beta_{i+1},\ldots,\beta_p),$$

donde L es la función de verosimilitud. Es decir, consiste en prefijar el coeficiente β_i en la función de verosimilitud, y maximizarla respecto de todos los demás coeficientes. Lo que ocurre es que al fijar este valor de β_i, la verosimilitud empeorará, es decir:

$$PL(\beta_i) < PL(\hat{\beta}_i) = L(\hat{\beta}).$$

Para construir el intervalo de confianza del parámetro β_i, consideraremos los valores que menos empeoran la verosimilitud, de forma que un el intervalo de confianza del parámetro estará compuesto por todos los β_i que cumplan

$$2\left(PL(\hat{\beta}_i) - PL(\beta_i)\right) < \chi^2_{1,\alpha},$$

siendo $\chi_{1,\alpha}$ el cuantil $(1-\alpha)$ de una distribución ji-cuadrado con un grado de libertad, y donde $(1-\alpha)$ es el nivel de confianza del intervalo.

En el caso de la distribución asintótica, el estimador de los parámetros β es asintóticamente insesgado, y la matriz de varianzas-covarianzas asintótica es $(X'VX)^{-1}$, donde V es la matriz diagonal tal que $v_{i,i} = \lambda(x_i, \beta)$. La matriz V se define así puesto que la varianza de la distribución de Poisson es la misma que la media, es decir, la varianza es el parámetro de Poisson, λ.

Ejemplo 1.2. Volviendo a nuestro ejemplo de las telas, cabe destacar que R utiliza el perfil de verosimilitud para calcular los intervalos de confianza de los parámetros. Por defecto, ya hemos visto que las estimaciones de los parámetros que nos proporciona R son las de β_0 y β_1. Sin embargo, en nuestro modelo, parece más intuitivo considerar las exponenciales de los mismos, como ya hemos explicado anteriormente. Por tanto, tendremos que aplicar la exponencial a la función `confint`, para así tener un intervalo de confianza para el intercepto y la tasa de crecimiento.

```
> exp(confint(modelo))
               2.5 %   97.5 %
(Intercept) 1.724917 3.968516
Length      1.142814 1.288669
```

Los intervalos de confianza están al nivel de significación del 95 %. Por tanto, estos intervalos ubican la exponencial del intercepto entre 1.72 y 3.97, y la tasa entre el 14.28 % y el 28.87 %.

1.3. Predicción

Hacer una predicción de un valor y_i en función de las variables explicativas x_i, es tan sencillo como utilizar la fórmula

$$\hat{y}_i = \lambda(x_i, \hat{\beta}) = e^{x_i \hat{\beta}}.$$

Ejemplo 1.3. Hagamos la predicción del número de defectos para piezas de tela de longitudes enteras de 2 a 5 hectómetros.

```
Length          3    4    5    6    7
Predicted flaws 4.71 5.72 6.94 8.41 10.2
```

Obsérvese que para una tela de 300 metros se esperan entre 4 y 5 defectos, mientras que para una de 700 metros, se esperan en torno a 10 defectos.

1.4. Diagnosis

En un modelo lineal usual se tiene que los errores son normales y son homocedásticos, y se espera un comportamiento similar en los residuos. Sin embargo, en un modelo de Poisson, esto no tiene porque ser así. De hecho, la distribución de los residuos ni siquiera es simétrica y centrada en cero. Debido a esto, vamos a enumerar varios tipos de residuos que se pueden tener en cuenta, y que nos ayudarán a la hora de hacer la diagnosis de nuestro modelo:

- **Residuos brutos:** Son la diferencia usual entre el valor observado y la predicción del modelo, es decir:
$$r_i := y_i - \hat{y}_i = y_i - e^{x_i' \hat{\beta}}.$$
- **Residuos de Pearson:** Consiste en estandarizar los residuos brutos teniendo en cuenta que, por la naturaleza de la distribución de Poisson, la varianza coincide con la media. Esto es:
$$r_i^P := \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i}} = \frac{y_i - e^{x_i' \hat{\beta}}}{e^{\frac{1}{2} x_i' \hat{\beta}}}.$$
- **Residuos de la deviance:** Se definen como:
$$r_i^D := sign\,(y_i - \hat{y}_i)\,\sqrt{d_i},$$

donde $d_i = 2\left(y_i \log(\frac{y_i}{\hat{y}_i}) - (y_i - \hat{y}_i)\right)$ son cada sumando de la deviance. La deviance del modelo de Poisson se puede probar que es:

$$D = 2\sum_{i=1}^{n}\left(y_i \log(\frac{y_i}{\hat{y}_i}) - (y_i - \hat{y}_i)\right).$$

Los residuos brutos es común que sean heterocedásticos. Sin embargo, los residuos de Pearson y de la deviance deberían seguir una pauta más estándar en caso de que fuese cierto el modelo de Poisson. La clave de la diagnosis está en observar los comportamientos de los diferentes tipos de residuos frente a las predicciones del modelo, y buscar patrones que se repitan en todos ellos y que nos hagan dudar de la homocedasticidad de los residuos.

Ejemplo 1.4. En la Figura 1.3 observamos los distintos tipos de residuos graficados frente al número esperado de defectos predicho por el modelo para los trozos de tela de longitudes varias. Se aprecia que, en el caso de los residuos brutos, estos tienen una mayor varianza al aumentar la longitud de la tela, pero esto se corrige en cierta medida al considerar los residuos de Pearson o los de la deviance, y ya no se aprecia ese patrón de crecimiento en los mismos. Sin embargo, se observa que, para predicciones menores que 8.5, la desviación típica de los residuos de Pearson es de 1.265, mientras que para los mayores, es de 1.617. Ambas son distintas, y mucho mayores que 1. Esto nos hace dudar de la homocedasticidad de los residuos, y además, también nos hace ver que el modelo de Poisson presenta mayor dispersión de lo esperado. Este fenómeno se conoce como sobre-dispersión, del que hablaremos a continuación, y se resuelve con el modelo Binomial Negativo.

Figura 1.3: Diagrama de dispersión de los distintos tipos de residuos frente al número de defectos esperados en las piezas de tela.

1.5. El fenómeno de sobre-dispersión

La mayor peculiaridad que posee la distribución de Poisson es que su media es igual a su varianza. Sin embargo, es muy posible que, tras construir un modelo de regresión de Poisson sobre unos datos, se observe que la varianza es visiblemente mayor que la media. Intuitivamente, si pensamos en poblaciones de animales repartidas en un terreno, por ejemplo, es natural que los individuos tiendan a agruparse y moverse juntos, de forma que al hacer una observación en una zona de ese terreno, o bien no se observa ningún animal, o bien se observan varios juntos.

Este fenómeno se conoce como sobre-dispersión, y nos indica que el modelo de Poisson no es correcto del todo, y necesitamos encontrar alguna medida para mitigar este fenómeno.

La sobre-dispersión se puede resolver utilizando el modelo de regresión Binomial Negativa, que trataremos a continuación.

1.5.1. Modelo de regresión Binomial Negativa

La distribución Binomial Negativa mide la probabilidad de obtener k éxitos antes de obtener el fracaso n-ésimo de un determinado experimento de Bernoulli, que tiene una probabilidad de éxito $p \in (0,1)$ en cada repetición del mismo. Consideramos el intervalo abierto, ya que en los extremos, o bien nunca habría éxitos, o bien nunca habría fracasos, y en esos casos ya no hay nada que investigar.

La masa de probabilidad de una variable $Y \in BN(n,p)$ tiene la siguiente expresión:

$$P(Y=k) = \binom{k+n-1}{k}(1-p)^n p^k, \quad k = 0,1,2,3\ldots$$

La media de la distribución Binomial Negativa es

$$E(Y) = \frac{pn}{1-p},$$

mientras que su varianza, que ya no es igual a la media como ocurría en Poisson (de hecho es siempre mayor) es

$$Var(Y) = \frac{pn}{(1-p)^2}.$$

Ahora bien, el número n es entero, pero se puede extender sin mayor problema al intervalo $(0,\infty)$ sin más que utilizar la función Gamma, que se define como:

$$\Gamma(n) = \int_0^\infty x^{n-1}e^{-x}dx, \quad n \in (0,\infty).$$

Teniendo esto en cuenta, se tiene que:

$$\binom{k+n-1}{k} = \frac{(k+n-1)!}{k!n!} = \frac{\Gamma(k+n)}{k!\Gamma(n)}.$$

Además, podemos realizar una nueva parametrización de la función de probabilidad, de forma que:

$$\mu := \frac{pn}{1-p}, \qquad \theta := n.$$

Utilizando esta reparametrización junto con la función Gamma, la función de probabilidad se expresa como:

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

donde su media es

$$E(Y) = \mu$$

y su varianza es

$$Var(Y) = \mu + \frac{\mu^2}{\theta}.$$

Expresada así, tenemos que el parámetro μ representa la media de la Binomial Negativa, mientras que el parámetro θ indica la sobre-dispersión.

A la vista de la fórmula de la varianza, notamos que, a medida que θ crece, el cociente $\frac{\mu^2}{\theta}$ se hace más pequeño, de forma que la $Var(Y)$ se acerca cada vez más a $\mu = E(Y)$. Es decir, cuanto más aumente el parámetro de sobre-dispersión, menor será la sobre-dispersión de los datos. En el Cuadro 1.1 observamos esa convergencia de la varianza a la media cuando aumentamos el parámetro de sobre-dispersión.

	θ	1	2	5	10	20	50	100	300	500	1000
$\mu = 4$	$Var(Y)$	20.00	12.00	7.20	5.60	4.80	4.32	4.16	4.08	4.03	4.02
$\mu = 8$	$Var(Y)$	72.00	40.00	20.80	14.40	11.20	9.28	8.64	8.32	8.13	8.06

Cuadro 1.1: Varianza de una distribución Binomial Negativa de media $\mu = 4$ o $\mu = 8$ para distintos valores del parámetro de sobre-dispersión θ.

En la Figura 1.4 se se puede observar cómo la distribución Binomial Negativa converge a la de Poisson al aumentar el parámetro θ, es decir, al disminuir la sobre-dispersión, como hemos comentado.

Consideremos ahora una variable respuesta discreta, Y, que queremos predecir utilizando un conjunto de variables explicativas, X. Entonces, procediendo de manera similar al modelo de Poisson, si suponemos que $Y \in BN(\mu(x,\beta),\theta)$, la función de regresión la tomaríamos como sigue:

$$\mu(x,\beta) = e^{x'\beta}.$$

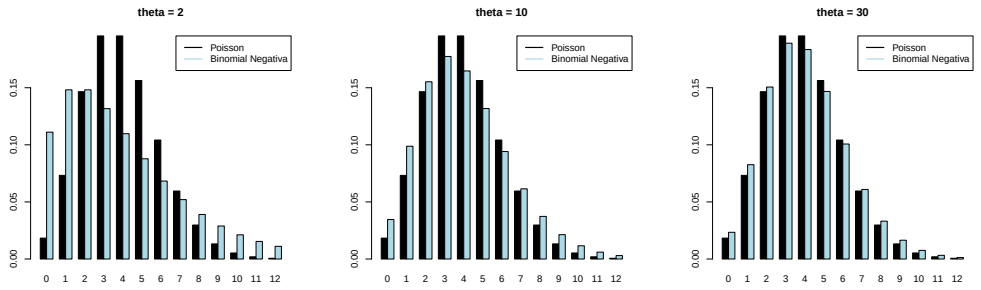

Figura 1.4: Diagramas de barras de la masa de probabilidad para las distribuciones de Poisson (en negro) y Binomial Negativa (en azul), ambas de media $\lambda = \mu = 4$, para diferentes valores de θ.

La estimación de los coeficientes β se realiza por máxima verosimilitud de la misma forma que en el modelo de Poisson, por lo que omitiremos los detalles. La función de verosimilitud se puede encontrar en el Zuur [1] (p.234).

Ejemplo 1.5. Volviendo a nuestro ejemplo de las telas, podemos construir un modelo Binomial Negativo utilizando la función `glm.nb` del paquete `MASS` [3].

```
> mod_NB = glm.nb(Flaws~Length, data = cloth, link = log)
> summary(mod_NB)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.00147    0.27837   3.598 0.000321 ***
Length       0.18827    0.04214   4.468  7.9e-06 ***
---
(Dispersion parameter for Negative Binomial(9.5697) family taken to be 1)

               Theta:  9.57
```

Las estimaciones de los coeficientes son algo diferentes a los que obteníamos en el modelo de Poisson. La del intercepto ahora es $\beta_0 = 1.001$, mientras que la de la pendiente es $\hat{\beta}_1 = 0.188$. Obsérvese que la sobre-dispersión, θ, es estimada por R, y en este caso es de 9.57.

Sobre la Figura 1.5, podemos apreciar que los residuos de Pearson, y sobre todo los de la deviance, ahora no presentan sobre-dispersión. Al considerar este modelo, la desviación típica de los residuos para predicciones menores que 8.5 ahora es 1.000, mientras que para los mayores

ahora es 1.065. Son ya prácticamente 1, y además, no hay diferencia entre ellas. Por tanto el modelo Binomial Negativo es el correcto sobre nuestros datos.

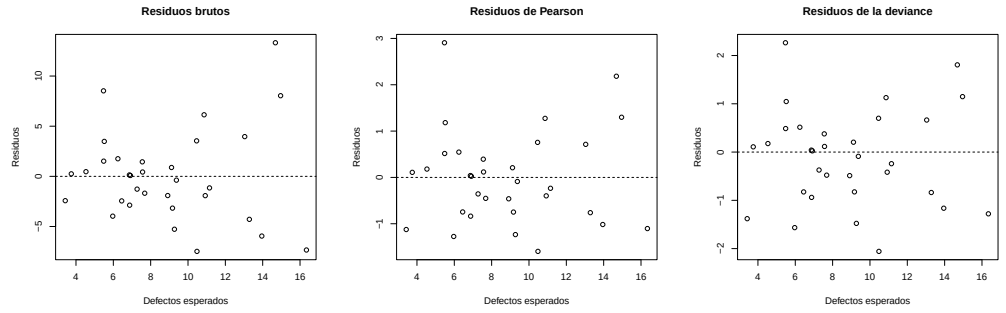

Figura 1.5: Diagrama de dispersión de los distintos tipos de residuos del modelo Binomial Negativo frente al número de defectos esperados en las piezas de tela.

Para confirmar lo anterior, nos podemos plantear un contraste de razón de verosimilitudes que de hecho cuestiona sobre la sobre-dispersión. Para ello utilizamos el paquete `lmtest` [9] de R:

```
> lmtest::lrtest(modelo,mod_NB)

Likelihood ratio test

  #Df  LogLik Df  Chisq Pr(>Chisq)
1   2 -92.528
2   3 -87.277  1 10.502   0.001192 **
```

Como el nivel crítico es pequeño, de 0.001, se rechaza que el modelo de Poisson sea mejor, y por tanto, nos quedamos con el modelo Binomial Negativo, como ya esperábamos.

1.5.2. Estimación ad hoc de la sobre-dispersión

La sobre-dispersión en el modelo de Poisson se puede estimar sin necesidad de considerar un nuevo modelo, utilizando la siguiente fórmula:

$$\hat{\Phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i-\hat{y_i})^2}{\hat{y_i}},$$

donde n es el número observaciones de la muestra y p el número de parámetros del modelo. Si $\hat{\Phi} > 1$, se habla de sobre-dispersión, mientras que si $\hat{\Phi} < 1$, se habla de infra-dispersión.

Al contrario de lo que ocurre al construir un modelo Binomial Negativo, las estimaciones de los coeficientes β no se ven afectadas por la estimación de la sobre-dispersión con la fórmula anterior. Sin embargo, sus errores típicos se multiplican por $\sqrt{\hat{\Phi}}$, por lo que si la dispersión es muy grande, las estimaciones de β serán mucho menos significativas.

Debido a que este procedimiento no afecta a estimaciones ajenas, sino que solo concierne a la sobre-dispersión, se denomina estimación *ad hoc* (que en latín significa «para esto»).

Ejemplo 1.6. Para hacer una estimación *ad hoc* de la sobre-dispersión en un modelo de regresión de Poisson con R, indicaremos como modelo quasipoisson.

```
> mod_qp = glm(Flaws~Length, data = cloth, family = quasipoisson(link = log))
> summary(mod_qp)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.97175    0.30950   3.140 0.003781 **
Length       0.19297    0.04462   4.325 0.000155 ***
---
(Dispersion parameter for quasipoisson family taken to be 2.121965)
```

Como se ha comentando, las estimaciones de los coeficientes β no se ven alteradas por la estimación de la dispersión en el modelo de Poisson, sino que solo cambian los errores típicos de los mismos, y en consecuencia, su significación.

En este caso, la dispersión se estima como $\hat{\Phi} = 2.12$, que es mayor que 1 (sobre-dispersión). Es por eso que la significación de los parámetros es menor.

Capítulo 2

Modelos de regresión para variables de recuento con ceros truncados

2.1. Introducción

Supongamos ahora que estamos observando una variable de recuento que no puede tomar el valor cero. Por ejemplo, si contamos el número de miembros de una organización o un grupo de personas, las observaciones que hagamos van a ser todas estrictamente positivas, ya que, con cero personas, el supuesto grupo o organización sería inexistente.

Ilustremos esta situación utilizando el conjunto de datos `house` [12], que recoge observaciones de variables relacionadas con una persona, como pueden ser su edad, su estado civil, la renta anual total de todas las personas de su vivienda (en dólares estadounidenses) o el propio número de habitantes que viven en su casa, contándose a sí mismo. Vamos a quedarnos con esta última variable, aunque tomaremos solo un subconjunto menor. El código para obtenerlo se encuentra en el Anexo I, junto con el código de los ejemplos de este capítulo. Claramente, esta variable es de recuento y, además, no puede tomar el valor cero, ya que en la casa de una persona vive como mínimo el propio individuo. En la Figura 2.1 podemos apreciar el fenómeno de truncamiento del cero que sufre esta variable, como acabamos de describir.

Para casos como el anterior, donde una variable de recuento no puede tomar el valor cero, las distribuciones de Poisson y Binomial Negativa que hemos estudiado en el capítulo anterior no serían apropiadas, ya que siempre estaríamos asignando una cierta probabilidad al valor cero, cuando realmente esta debería ser nula. Por este motivo, nos vemos obligados a introducir las distribuciones con ceros truncados, para poder así solucionar el problema al que nos enfrentamos.

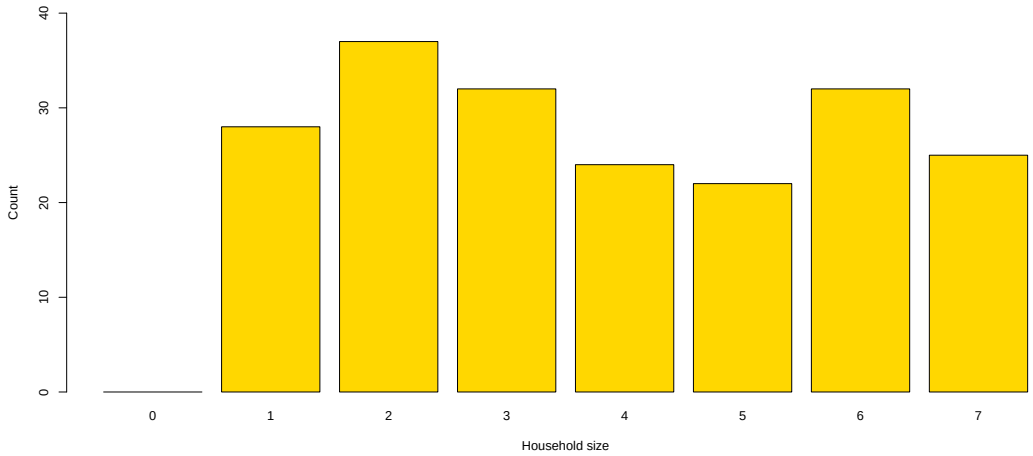

Figura 2.1: Diagrama de frecuencias absolutas del número de personas que habitan en una misma vivienda en la muestra del conjunto de datos.

Durante este capítulo, explicaremos la distribución de Poisson truncada y la Binomial Negativa truncada, así como los correspondientes modelos de regresión para variables de recuento con esas distribuciones.

2.2. La distribución de Poisson truncada en el cero

Sea Y una variable de recuento de Poisson con ceros truncados. Podemos interpretar esta situación como que existe una variable $Z \in Poisson(\lambda)$, la cual solamente es observada cuando es estrictamente positiva. La masa de probabilidad de una distribución de Poisson truncada tiene la siguiente expresión:

$$P(Y = y) = P(Z = y/Z > 0) = \frac{P(Z = y)}{P(Z > 0)} = \frac{e^{-\lambda}\lambda^y}{(1 - e^{-\lambda})y!}, \quad y = 1, 2, 3, \ldots$$

El problema es que ahora la media de esta distribución no es el parámetro λ, a diferencia de lo que ocurre en una distribución de Poisson usual. En efecto:

$$E(Y) = \sum_{y=1}^{\infty} y \frac{e^{-\lambda}\lambda^y}{(1 - e^{-\lambda})y!} = \frac{\lambda}{1 - e^{-\lambda}} \sum_{y=1}^{\infty} \frac{e^{-\lambda}\lambda^{y-1}}{(y-1)!}.$$

Realizando el cambio $z = y - 1$, concluimos que

$$E(Y) = \frac{\lambda}{1 - e^{-\lambda}} \sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^y}{y!} = \frac{\lambda}{1 - e^{-\lambda}},$$

donde hemos usado que $\sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^y}{y!} = 1$ por definición de función de masa de probabilidad (en este caso, de Poisson). La varianza de Y, por su parte, se puede calcular como

$$Var(Y) = E(Y^2) - E(Y)^2,$$

teniendo en cuenta que, por las propiedades de la media,

$$E(Y^2) = E(Y(Y-1) + Y) = E(Y(Y-1)) + E(Y).$$

Entonces, tenemos lo siguiente:

$$E(Y(Y-1)) = \sum_{y=2}^{\infty} = y(y-1)\frac{e^{-\lambda}\lambda^y}{(1-e^{-\lambda})y!} = \frac{\lambda^2}{1-e^{-\lambda}} \sum_{y=2}^{\infty} \frac{e^{-\lambda}\lambda^{y-2}}{(y-2)!}.$$

Realizando el cambio $z = y - 2$, obtenemos:

$$E(Y(Y-1)) = \frac{\lambda^2}{1-e^{-\lambda}} \sum_{z=0}^{\infty} \frac{e^{-\lambda}\lambda^z}{z!} = \frac{\lambda^2}{1-e^{-\lambda}} = \lambda E(Y).$$

Por lo tanto, deducimos que

$$E(Y^2) = \lambda E(Y) + E(Y),$$

y, en consecuencia,

$$Var(Y) = \lambda E(Y) + E(Y) - E(Y)^2 = E(Y)(1 + \lambda - E(Y)).$$

En resumen, tenemos las siguientes expresiones para la media y la varianza:

$$\begin{aligned} E(Y) &= \frac{\lambda}{1-e^{-\lambda}}, \\ Var(Y) &= \frac{\lambda}{1-e^{-\lambda}} \left(1 + \lambda - \frac{\lambda}{1-e^{-\lambda}}\right). \end{aligned}$$

Como podemos ver, la varianza de la distribución de Poisson truncada ya no coincide con su media, a diferencia de lo que ocurría en la distribución de Poisson usual. Además, en concreto, dicha varianza siempre será menor que la media, lo cual se puede probar fácilmente. Se tiene:

$$\frac{Var(Y)}{E(Y)} = \frac{E(Y)(1 + \lambda - E(Y))}{E(Y)} = 1 + \lambda - \frac{\lambda}{1-e^{-\lambda}}.$$

Teniendo en cuenta que $0 < 1 - e^{-\lambda} < 1$, entonces $\frac{\lambda}{1-e^{-\lambda}} > \lambda$ y, en consecuencia,

$$\frac{Var(Y)}{E(Y)} < 1 + \lambda - \lambda = 1.$$

Es decir, efectivamente $Var(Y) < E(Y)$.

Por la forma de la expresión de la media, donde el denominador aumenta cuando lo hace λ, es claro que $E(Y) > \lambda$. Además, por el mismo motivo, también es intuitivo que, cuanto mayor

sea el valor del parámetro λ, la media de la Poisson truncada estará más cerca del parámetro de Poisson. En efecto:

$$\lim_{\lambda\to\infty} E(Y) - \lambda = \lim_{\lambda\to\infty} \frac{\lambda}{1-e^{-\lambda}} - \lambda = \lim_{\lambda\to\infty} \frac{\lambda e^{-\lambda}}{1-e^{-\lambda}} = \lim_{\lambda\to\infty} \frac{\lambda}{e^{\lambda}-1} = 0.$$

Con la varianza ocurre algo parecido. Por su parte, se puede probar que $Var(Y) < \lambda$, y además, también se acercará al valor del parámetro cuando este último aumenta:

$$\lim_{\lambda\to\infty} Var(Y) - \lambda = \lim_{\lambda\to\infty} E(Y)(1+\lambda-E(Y)) - \lambda = \lim_{\lambda\to\infty} E(Y) - \lambda = 0.$$

En el Cuadro 2.1 se representan los valores de la media y la varianza de una distribución de Poisson truncada para valores del parámetro λ desde 1 hasta 10. Se puede observar que la varianza es en todos los casos más pequeña que la media, como habíamos probado. Además, la varianza está cada vez más cerca de la media cuando aumentamos el valor del parámetro λ, lo que nos indica que para valores de λ grandes, la distribución de Poisson truncada es prácticamente la misma que la de Poisson usual, por lo que el hecho de considerar o no que una variable es truncada perdería importancia, y no tendría graves consecuencias.

λ	1	2	3	4	5	6	7	8	9	10
$E(Y)$	1.582	2.313	3.157	4.075	5.034	6.015	7.006	8.003	9.001	10.00
$Var(Y)$	0.661	1.589	2.661	3.771	4.863	5.925	6.962	7.981	8.991	9.996

Cuadro 2.1: Media y varianza de una distribución de Poisson truncada en el cero para distintos valores del parámetro λ.

En la Figura 2.2 se representan tres diagramas de barras, en los que se comparan las masas de probabilidad de las distribuciones de Poisson y Poisson truncada, para los valores $\lambda = 2$, $\lambda = 5$ y $\lambda = 8$. Obsérvese que cuanto mayor es λ, más semejantes son las masas de probabilidad de ambas distribuciones, ya que la distribución de Poisson truncada converge a la de Poisson a medida que aumentamos el parámetro λ, como hemos comentado.

2.3. El modelo de Poisson truncado en el cero

Sea entonces $(X_1, Y_1), \ldots, (X_n, Y_n)$ una muestra de observaciones, donde X son las variables explicativas e Y es la variable respuesta de recuento, en la que se observa un truncamiento en el valor cero. Para estimar λ, utilizaremos de nuevo el modelo log-lineal, que tenía la expresión

$$\lambda(x_i, \beta) = e^{x_i\beta},$$

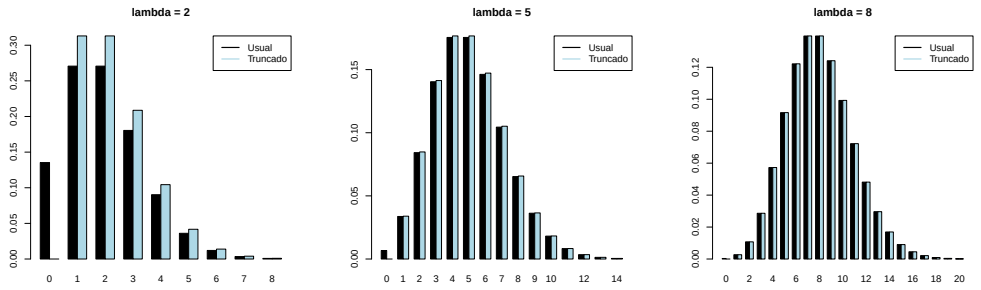

Figura 2.2: Diagramas de barras de la masa de probabilidad para las distribuciones de Poisson (en negro) y Poisson truncada (en azul), para diferentes valores de λ.

donde x_i son las variables explicativas que utilizaremos para predecir y_i, y β son los coeficientes de regresión. Estos últimos se estiman por máxima verosimilitud de la misma manera que en el capítulo anterior, pero utilizando la distribución de la variable Y. Es decir, la función de verosimilitud tiene el siguiente aspecto:

$$L(\beta) = \prod_{i=1}^{n} \frac{e^{-\lambda(x_i,\beta)}\lambda(x_i,\beta)^{y_i}}{(1-e^{-\lambda(x_i,\beta)})y_i!}.$$

De esta manera, a pesar de que la expresión para $\lambda(x,\beta)$ es la misma que para el modelo usual, las estimaciones de los coeficientes β serán diferentes.

En cuanto a la inferencia sobre los parámetros y la diagnosis del modelo, se utilizan los mismos métodos y fórmulas que en un modelo usual, aunque los resultados que obtengamos serán obviamente diferentes.

Para realizar una predicción correcta, debemos tomar la estimación de λ y efectuar una corrección. Dicha corrección la hemos introducido en la expresión de $E(Y)$. Así, la predicción correcta es:

$$\hat{y}_i = \frac{\lambda(x_i,\hat{\beta})}{1-e^{-\lambda(x_i,\hat{\beta})}}.$$

En cuanto a los contrastes y los intervalos de confianza asintóticos, ahora la matriz diagonal V será de tal forma que $v_{i,i} = \hat{y}_i(1+\lambda(x_i,\hat{\beta})-\hat{y}_i)$, ya que la expresión de la varianza de Y es diferente a la que teníamos en el modelo usual. Por este mismo motivo, la fórmula de los residuos de Pearson ahora también es diferente:

$$r_i^P := \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i(1+\lambda(x_i,\hat{\beta})-\hat{y}_i)}}.$$

Observación 2.1. Si ajustamos un modelo de Poisson usual a unos datos con ceros truncados, lo que ocurrirá es que los residuos de Pearson tendrán infra-dispersión, ya que la varianza de

la distribución de Poisson truncada en el cero es menor que su media. Para ilustrar esto, se ha realizado una simulación en R de la siguiente forma. Primero, se han generado 5 muestras aleatorias de 100 observaciones cada una, procedentes de una distribución de Poisson truncada en el cero de parámetro $\lambda = 2$. Después, para cada muestra, se ha ajustado sobre ella un modelo de Poisson y otro de Poisson truncado en el cero, ambos sin variables explicativas, y se ha calculado la desviación típica de los residuos de Pearson de cada modelo. Las desviaciones típicas obtenidas para cada modelo y cada muestra se pueden ver en el Cuadro 2.2.

	Muestra 1	Muestra 2	Muestra 3	Muestra 4	Muestra 5
Usual	0.801	0.761	0.869	0.835	0.896
Truncado	0.978	0.933	1.038	0.999	1.066

Cuadro 2.2: Desviación típica de los residuos de Pearson de un modelo de Poisson usual y un modelo de Poisson truncado en el cero para varias muestras de recuento aleatorias procedentes de una distribución de Poisson con ceros truncados de parámetro $\lambda = 2$.

Como vemos en el Cuadro 2.2, ajustar un modelo de Poisson sobre las muestras truncadas causa infra-dispersión en los residuos de Pearson, puesto que la desviación típica de los residuos es notablemente menor que 1 para las cinco muestras. Sin embargo, al ajustar un modelo de ceros truncados, la desviación típica es prácticamente 1, como se espera en los residuos de Pearson de un modelo.

Ejemplo 2.2. Vamos a construir un modelo de Poisson truncado sobre el subconjunto de los datos `house` que hemos introducido al principio del capítulo. Intentaremos predecir el número de habitantes de una vivienda (*size*) a partir de tres variables explicativas: la edad de la persona observada (*age*), la renta anual total en su vivienda (*income*) y su estado civil (*marital_status*).

Para construir el modelo, utilizamos el paquete `VGAM` [4] de R, que no viene incluido en la instalación básica.

```
> m = vglm(size~age + income + marital_status, data = house,
+ family = pospoisson)
> summary(m)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            1.274e+00  1.600e-01   7.964 1.66e-15 ***
age                    5.428e-03  2.540e-03   2.137   0.0326 *
income                -2.029e-08  2.031e-08  -0.999   0.3178
marital_statusMarried -2.621e-01  1.303e-01  -2.012   0.0443 *
```

```
marital_statusSingle  -1.210e-01  1.324e-01  -0.914   0.3607
---
```

Para el intercepto, la estimación de su coeficiente es $\hat{\beta}_0 = 1.274$. En el caso de las variables edad y renta, tenemos que sus coeficientes vienen estimados por los valores $\hat{\beta}_1 = 5.428 \cdot 10^{-3}$ y $\hat{\beta}_2 = -2.029 \cdot 10^{-8}$. En cuanto a la variable que determina el estado civil del individuo observado, hay que tener en cuenta que es de tipo factor, y toma tres valores diferentes: *Divorced* (Divorciado/a), *Married* (Casado/a) y *Single* (Soltero/a). Para cada una, tendremos un coeficiente $\beta_{3,j}$, para $j = 1, 2, 3$. El modelo considera que un individuo está divorciado como categoría de referencia, por lo que se tiene que $\hat{\beta}_{3,1} = 0$, y las estimaciones de los coeficientes para un individuo casado y uno soltero son $\hat{\beta}_{3,2} = -0.262$ y $\hat{\beta}_{3,3} = -0.121$, respectivamente.

Los intervalos de confianza de los coeficientes β al nivel de confianza del 95 % son los siguientes:

```
> confint(m)
                                2.5 %        97.5 %
(Intercept)             9.606997e-01  1.587914e+00
age                     4.502439e-04  1.040529e-02
income                 -6.008432e-08  1.951179e-08
marital_statusMarried -5.175354e-01 -6.722233e-03
marital_statusSingle  -3.805505e-01  1.384809e-01
```

Lo más destacable de estos intervalos es que, para las variables de la renta anual de cada vivienda, y el estado civil en caso de estar soltero, se tiene que el cero está contenido en sus intervalos. Esto ya nos lo anticipó el resumen del modelo, por el hecho de que los p-valores de las estimaciones de los coeficientes de la renta y el estado civil soltero son 0.318 y 0.361, respectivamente. Esto quiere decir que ambas variables no aportan demasiado al modelo, y que se podría considerar que su coeficiente es nulo. Aunque si hiciésemos eso con una de ellas, habría que comprobar si, en el modelo resultante tras suprimir esa variable, la otra variable sigue sin tener suficiente significación.

Hagamos ahora una predicción para tres individuos con diferentes condiciones de edad, renta anual y estado civil:

```
  Age Income Marital_Status Predicted_Size
1  20  60000         Single          3.634
2  35 140000        Married          3.443
3  50 100000       Divorced          4.725
```

El primer individuo tiene 20 años, tiene una renta total en su vivienda de 60000 dólares anuales y está soltero. El modelo estima que habrá una media de 3.634 personas en su vivienda, contando con él mismo. Es decir, predice que, en general, observaremos 3 ó 4 personas en una vivienda cuyo representante está en las condiciones del primer individuo. Para el segundo individuo, que tiene 35 años, tiene una renta de 140000 dólares y está casado, el modelo predice que habrá de media 3.443 personas en su vivienda, mientras que para el tercer individuo, de 50 años, con una renta de 100000 dólares y divorciado, se predice que conviven 4.725 personas en el mismo lugar de residencia.

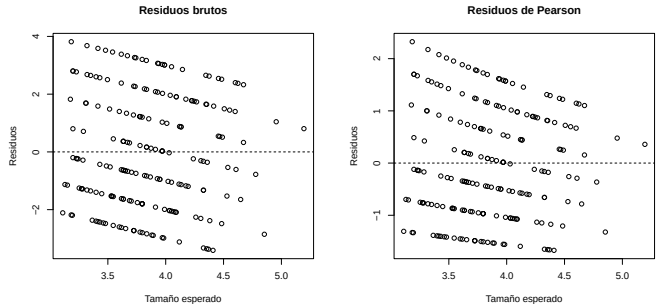

Figura 2.3: Diagrama de dispersión de los distintos tipos de residuos frente al número de habitantes esperados en una misma vivienda.

Si finalmente realizamos la diagnosis del modelo, apoyándonos en la Figura 2.3, observamos que los residuos de Pearson decrecen en cierta medida cuando el valor de la predicción aumenta. Sin embargo, debemos fijarnos en que hay más cantidad de datos para predicciones menores que para predicciones mayores. Esto implica que, para predicciones mayores, es normal que no haya tantos residuos que se encuentren lejos de su media, y no por eso su varianza será diferente. Por tanto, se puede aceptar que los residuos son homocedásticos. De hecho, si calculamos la desviación típica de los residuos de Pearson del modelo para las predicciones menores que 3.8, se obtiene un valor de 1.137, mientras que para las mayores, es de 0.985. No son valores extremadamente diferentes, ni tampoco muy lejanos a 1, por lo que no tenemos presencia de sobre-dispersión en nuestros datos.

2.4. El modelo Binomial Negativo truncado en el cero

Supongamos ahora que hemos observado demasiada dispersión en la variable de recuento Y con ausencia de ceros, y que sospechamos que puede pertenecer a una distribución Binomial Negativa, que también será truncada en el cero. Utilizando la expresión en función de la media,

μ, y la sobre-dispersión, θ, definamos $Z \in BN(\mu, \theta)$. De este modo, teniendo en cuenta que

$$P(Z = 0) = \left(\frac{\theta}{\mu + \theta}\right)^{\theta},$$

la masa de probabilidad de la distribución Binomial Negativa truncada será

$$P(Y = y) = \frac{P(Z = y)}{P(Z > 0)} = \frac{\Gamma(\theta + y)}{\Gamma(\theta) y!} \frac{\mu^y \theta^\theta}{(\mu + \theta)^{(\theta+y)}} \frac{1}{1 - \left(\frac{\theta}{\mu+\theta}\right)^{\theta}}.$$

La media de esta distribución será

$$E(Y) = \frac{E(Z)}{P(Z > 0)},$$

es decir,

$$E(Y) = \frac{\mu}{1 - \frac{\theta^\theta}{(\mu+\theta)^\theta}}$$

sin más que tener en cuenta propiedades de la media y que $\Gamma(y) = y\Gamma(y - 1)$. Utilizando las mismas propiedades, también se puede probar que la varianza es

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

Definiremos el modelo log-lineal para estimar el parámetro μ de la misma forma que en el capítulo anterior:

$$\mu(x, \beta) = e^{x'\beta},$$

donde x son las observaciones de las variables explicativas y β son los coeficientes del modelo, que estimamos por máxima verosimilitud.

La función de verosimilitud se calcula de manera análoga a como se ha realizado en la regresión de Poisson truncada:

$$L(\beta, \theta) = \prod_{i=1}^{n} \frac{\Gamma(\theta + y_i)}{\Gamma(\theta) y_i!} \frac{\mu(x_i, \beta)^{y_i} \theta^\theta}{(\mu(x_i, \beta) + \theta)^{(\theta+y_i)}} \frac{1}{1 - \left(\frac{\theta}{\mu(x_i,\beta)+\theta}\right)^{\theta}}.$$

De aquí obtendremos las estimaciones $\hat{\beta}$ y $\hat{\theta}$, las cuales serán diferentes a las que se obtenían en el modelo usual, debido a que la función de verosimilitud es distinta.

Para hacer una predicción de un valor y_i en función de las variables explicativas x_i, tendremos que aplicar una corrección como ocurría en el modelo de Poisson truncado. Así, por la expresión que tiene la media, la predicción se calcula como

$$\hat{y_i} = \frac{\mu(x_i, \hat{\beta})}{1 - \frac{\hat{\theta}^{\hat{\theta}}}{(\mu(x_i,\hat{\beta})+\hat{\theta})^{\hat{\theta}}}}.$$

Ejemplo 2.3. Vamos a construir un modelo Binomial Negativo truncado para los datos de `house`, y vamos a comprobar que no es mejor, ya que no hay sobre-dispersión en nuestros datos.

```
> m_nb = vglm(size~age + income + marital_status, data = house,
+ family = posnegbinomial)
> summary(m_nb)

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept):1           1.269e+00  1.727e-01   7.345 2.05e-13 ***
(Intercept):2           3.306e+00  9.368e-01   3.529 0.000418 ***
age                     5.456e-03  2.727e-03   2.001 0.045415 *
income                 -2.100e-08  2.174e-08  -0.966 0.334227
marital_statusMarried -2.650e-01  1.411e-01  -1.879 0.060280 .
marital_statusSingle  -1.226e-01  1.435e-01  -0.855 0.392764
---
```

Obsérvese que la salida es algo extraña, con dos interceptos diferentes, pero es fácil de aclarar: el valor de `(Intercept):1` es el valor de $\hat{\beta}_0$, el intercepto, mientras que el valor de `(Intercept):2` será el valor de $\log(\hat{\theta})$. Obtenemos unas estimaciones de los coeficientes bastante parecidas a las que obteníamos para el modelo de Poisson truncado. Si nos fijamos en valor de la estimación de $\log\theta$, vemos que vale 3.306, por lo que al hacer su exponencial, tenemos una estimación del parámetro de sobre-dispersión de $\hat{\theta} = 27.270$. Este valor es grande, lo que quiere decir que la sobre-dispersión es pequeña o incluso prácticamente nula, como ya nos habían anticipado los residuos del modelo de Poisson con ceros truncados. De hecho, calculando de nuevo la desviación típica de los residuos de Pearson, se obtiene que esta vale 1.083 para las predicciones menores de 3.8, y 0.928 para las mayores. Entonces, el modelo Binomial Negativo es muy parecido al de Poisson (por eso obtenemos unas estimaciones de los coeficientes muy semejantes) y no tenemos ningún motivo por el que usar el Binomial Negativo truncado en el cero en lugar del de Poisson con ceros truncados sobre nuestros datos.

Capítulo 3

Modelos de regresión para variables de recuento con ceros inflados

3.1. Introducción

En el capítulo anterior nos encontrábamos en una situación donde una variable de recuento no presentaba ceros. En este capítulo, nos pondremos en la situación contraria: la variable presenta un exceso de ceros. Los motivos de que esto ocurra pueden ser muy variados. A veces, están causados por errores en la medición o por dificultades en la misma. Pero en general, se deberá a que realmente la variable que estamos observando tiene una distribución mixta. Por un lado, está la distribución de recuento, de la que provienen todos los valores positivos y algunos de los ceros, que puede ser la distribución de Poisson o la Binomial Negativa. Pero por otro lado, el resto de ceros provendrán de otra distribución.

Ilustremos la situación a la que nos enfrentamos usando el conjunto de datos `taxi` [11], que recoge observaciones de ciertas variables relacionadas con el mundo de los taxis en Nueva York, entre ellas, la propina que recibe un taxista de parte del cliente al finalizar el trayecto. Como tiene demasiadas observaciones (varias decenas de miles), nos quedaremos con un subconjunto más pequeño, cuyo código para reproducirlo se encuentra en el Anexo I, dentro del código de los ejemplos de modelos de este capítulo.

En la Figura 3.1 se representan las frecuencias absolutas de la cantidad de propina (o la ausencia de ella), en dólares estadounidenses, que los clientes de nuestra muestra han dado al taxista. Como observamos, la cantidad de ceros presentes es muy grande, mucho más de lo que cabría esperar para una distribución de Poisson o una Binomial Negativa.

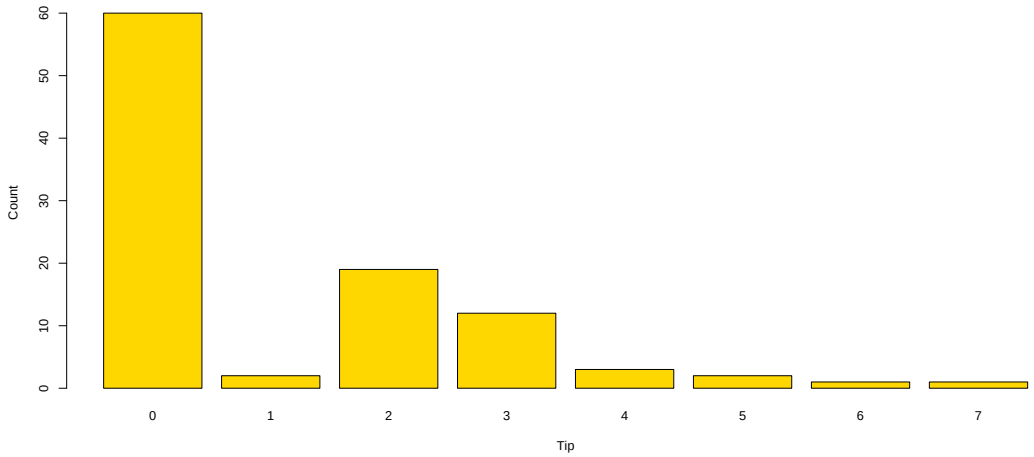

Figura 3.1: Diagrama de frecuencias absolutas de la cantidad de propina, en dólares estadounidenses, pagada al taxista en la muestra del conjunto de datos.

Si un cliente da propina al taxista, entonces esta provendrá de la distribución de recuento, ya que será un valor estrictamente positivo. Sin embargo, si un cliente no da propina, entonces puede ser por dos motivos distintos. El primero es que algo ha ocurrido durante el viaje que ha hecho decidir al cliente que no dará propina al taxista al finalizar el trayecto. Por ejemplo, si el taxista es demasiado hablador, o demasiado lento, o si el cliente en ese momento no puede dar propina porque tiene problemas económicos y quiere ahorrar lo máximo posible. Este sería un cero auténtico, proveniente de la distribución discreta. La otra razón por la que un cliente no dé propina puede ser porque es una persona que jamás da propina, independientemente de si disfruta o no del trayecto o de si tiene dinero o no. Este sería un falso cero, que proviene de una distribución diferente a la de las demás observaciones.

Es por todo esto que necesitamos las distribuciones con ceros inflados, para solventar los problemas de exceso de ceros que se dan a menudo en observaciones de variables de tipo recuento. En concreto, estudiaremos las distribuciones de Poisson y Binomial Negativa infladas en el cero, así como los modelos de regresión para variables de recuento que sigan estas distribuciones.

3.2. La distribución de Poisson inflada en cero

Sea Y una variable de recuento, la cual contiene ceros, y sospechamos que no todos proceden de la misma distribución debido a que presenta una cantidad de ceros mucho más grande de lo esperado. La situación es la siguiente: la variable Y será un falso cero con una cierta probabilidad

π, mientras que con probabilidad $1-\pi$, esta tomará el mismo valor, sea cero o no, que una variable Z que sigue una distribución de Poisson.

Por tanto, sea $Z \in Poisson(\lambda)$. Tenemos las siguientes expresiones para la distribución de la variable Y:

$$\begin{aligned} P(Y=0) &= \pi + (1-\pi)P(Z=0) \\ P(Y=y) &= (1-\pi)P(Z=y), \quad y = 1,2,3,\ldots \end{aligned}$$

Sustituyendo la expresión de la distribución de Poisson en ambas expresiones, tenemos que:

$$\begin{aligned} P(Y=0) &= \pi + (1-\pi)e^{-\lambda} \\ P(Y=y) &= (1-\pi)e^{-\lambda}\frac{\lambda^y}{y!}, \quad y = 1,2,3,\ldots \end{aligned}$$

La media de la variable Y ahora ya no es λ, como ocurría en el modelo usual, sino que se calcula de la siguiente forma:

$$E(Y) = \sum_{i=0}^{\infty} yP(Y=y) = \sum_{i=1}^{\infty}(1-\pi)y\frac{e^{-\lambda}\lambda^y}{y!} = (1-\pi)\lambda\sum_{i=1}^{\infty}\frac{e^{-\lambda}\lambda^{y-1}}{(y-1)!.}$$

Realizando el cambio $z = y-1$, obtenemos:

$$E(Y) = (1-\pi)\lambda\sum_{i=0}^{\infty}\frac{e^{-\lambda}\lambda^z}{z!} = (1-\pi)\lambda.$$

Es intuitivo que, como $0 < (1-\pi) < 1$, entonces $E(Y) < \lambda$. Para calcular la varianza de Y, usaremos de nuevo el método utilizado en el capítulo anterior. Por tanto:

$$E(Y(Y-1)) = \sum_{i=2}^{\infty} y(y-1)(1-\pi)\frac{e^{-\lambda}\lambda^y}{y!} = (1-\pi)\lambda^2\sum_{i=2}^{\infty}\frac{e^{-\lambda}\lambda^{y-2}}{(y-2)!}.$$

Considerando el cambio de variable $z = y-2$, se sigue que:

$$E(Y(Y-1)) = (1-\pi)\lambda^2\sum_{i=0}^{\infty}\frac{e^{-\lambda}\lambda^z}{z!} = (1-\pi)\lambda^2.$$

Entonces, utilizando ahora que $E(Y^2) = E(Y(Y-1)) + E(Y)$, se tiene que:

$$Var(Y) = E(Y^2) - E(Y)^2 = (1-\pi)(\lambda^2+\lambda) - (1-\pi)^2\lambda^2 = (1-\pi)(\lambda + \pi\lambda^2).$$

En resumen, tenemos que:

$$\begin{aligned} E(Y) &= (1-\pi)\lambda, \\ Var(Y) &= (1-\pi)(\lambda + \pi\lambda^2). \end{aligned}$$

Al igual que ocurría en la distribución de Poisson truncada, la media y la varianza de la distribución de Poisson inflada tampoco coinciden, y en particular, en este caso la varianza es más grande que la media. Esto se puede ver a través del cociente

$$\frac{Var(Y)}{E(Y)} = \frac{(1-\pi)\lambda(1+\pi\lambda)}{(1-\pi)\lambda} = 1 + \pi\lambda > 1,$$

lo cual equivale a decir que $Var(Y) > E(Y)$.

En la distribución con ceros truncados, la media se acercaba al parámetro de Poisson a medida que este aumentaba, y con la varianza ocurría lo mismo. Sin embargo, esto no ocurre en la distribución con ceros inflados. Por un lado:

$$\lim_{\lambda\to\infty} \lambda - E(Y) = \lim_{\lambda\to\infty} \lambda - (1-\pi)\lambda = \lim_{\lambda\to\infty} \pi\lambda = \infty.$$

Por otro lado, también tenemos que:

$$\lim_{\lambda\to\infty} Var(Y) - \lambda = \lim_{\lambda\to\infty} (1-\pi)(\lambda + \pi\lambda^2) - \lambda = \lim_{\lambda\to\infty} (\pi - \pi^2)\lambda^2 - \pi\lambda = \infty,$$

puesto que, al ser $0 < \pi < 1$, entonces $\pi - \pi^2 > 0$.

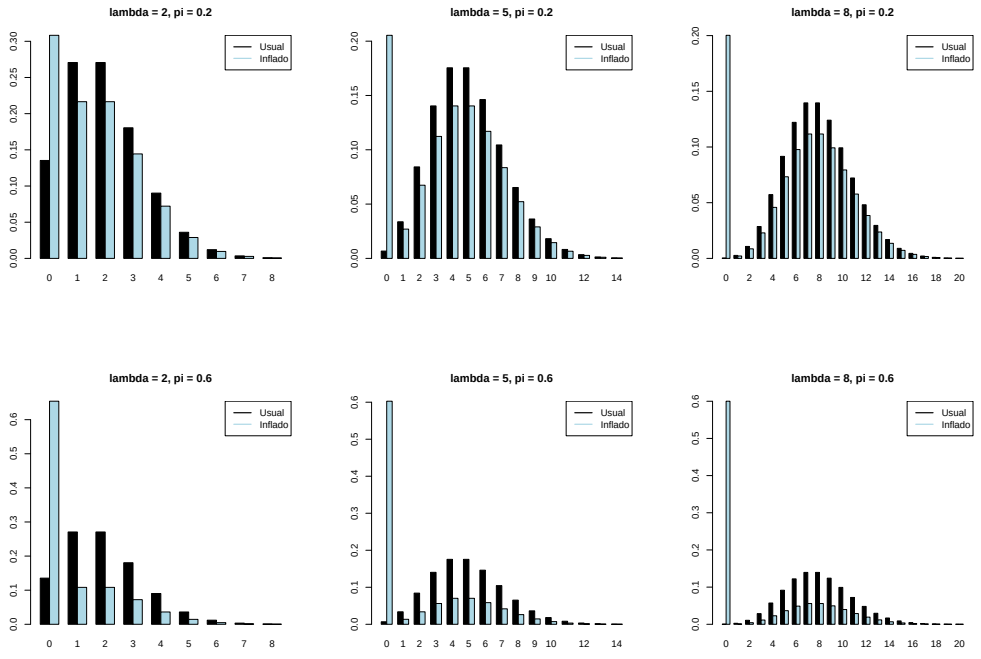

Figura 3.2: Diagramas de barras de la masa de probabilidad de las distribuciones de Poisson (en negro) y Poisson inflada en el cero (en azul), para diferentes valores de λ y diferentes valores de π.

El hecho de que estos límites sean no finitos tiene bastantes consecuencias si tratamos con distribuciones de Poisson de parámetros grandes. Si calculamos la media de una muestra, y suponemos que proviene de una Poisson usual, estaremos estimando una distribución de parámetro $(1-\pi)\lambda$, en lugar del propio λ. Para un λ pequeño, esto no influye demasiado, pero para un λ más grande, puede haber demasiada diferencia, sobre todo para valores de π cercanos a uno.

En el Cuadro 3.1 se puede observar cómo efectivamente la media de la distribución inflada en el cero es cada vez más diferente al parámetro de Poisson cuanto más grande sea este, y cómo esta diferencia es mayor al considerar una probabilidad π más alta. Además, también se aprecia cómo con la varianza ocurre algo semejante. Al aumentar λ, la varianza crece con mucha rapidez, y lo hará todavía más rápido para valores mayores de π.

	λ	1	2	3	4	5	6	7	8	9	10
$\pi = 0.2$	$E(Y)$	0.80	1.60	2.40	3.20	4.00	4.80	5.60	6.40	7.20	8.00
	$Var(Y)$	0.96	2.24	3.84	5.76	8.00	10.56	13.44	16.64	20.16	24.00
$\pi = 0.6$	$E(Y)$	0.40	0.80	1.20	1.60	2.00	2.40	2.80	3.20	3.60	4.00
	$Var(Y)$	0.64	1.76	3.36	5.44	8.00	11.04	14.56	18.56	23.04	28.00

Cuadro 3.1: Media y varianza de una distribución de Poisson inflada en el cero para distintos valores del parámetro λ y para $\pi = 0.2$ y $\pi = 0.6$.

En la Figura 3.2 se pueden observar los diagramas de barras de la masa de probabilidad de las distribuciones de Poisson y Poisson inflada en el cero para los valores $\lambda = 2$, $\lambda = 5$ y $\lambda = 8$, para probabilidades de falso cero $\pi = 0.2$ y $\pi = 0.6$. Como se observa, cuanto más grandes sean λ y π, más afecta la inflación del cero, como hemos comentado previamente.

3.3. El modelo de Poisson inflado en el cero

Sea $(X_1, Y_1), \ldots, (X_n, Y_n)$ una muestra, donde X representa las variables explicativas e Y representa la variable de recuento que queremos predecir, la cual presenta muchos ceros, más de los esperados. Vamos a suponer que la variable Y proviene de una distribución de Poisson, a excepción de una cierta cantidad de ceros que proceden de otra distribución. Para estimar el parámetro λ de Poisson, utilizaremos el modelo log-lineal, que tiene la expresión

$$\lambda(x, \beta) = e^{x'\beta},$$

donde x son las variables explicativas de una observación, y β son los coeficientes del modelo.

La novedad que se nos presenta en el modelo de ceros inflados es la necesidad de estimar la probabilidad de obtener un falso cero, es decir, un cero proveniente de la otra distribución que

no es la de Poisson. Hay varias formas de llevar a cabo esta estimación, pero la más utilizada es construir un modelo logístico para π. El modelo que vamos a utilizar para este fin tiene la siguiente expresión:

$$\pi(x, \alpha) = \frac{e^{x'\alpha}}{1 + e^{x'\alpha}},$$

donde x son los valores de un conjunto de variables explicativas y α son los coeficientes del modelo. Se pueden ver más detalles sobre regresión logística en el Hosmer [13].

En cuanto a la estimación los coeficientes α y β, esta se calcula por el método de máxima verosimilitud, utilizando la distribución de Y que hemos explicado antes. De esta forma, la función de verosimilitud tiene el siguiente aspecto:

$$L(\alpha, \beta) = \prod_{i=1}^{n} \Bigg((\pi(x_i, \alpha) + (1 - \pi(x_i, \alpha))e^{-\lambda(x_i, \beta)}) I_{\{y_i=0\}} + \\ + (1 - \pi(x_i, \alpha))e^{-\lambda(x_i, \beta)} \frac{\lambda(x_i, \beta)^{y_i}}{y_i!} I_{\{y_i>0\}} \Bigg).$$

Obsérvese el uso de los indicadores

$$I_{\{y_i=0\}} = \begin{cases} 1, & \text{si } y_i = 0, \\ 0, & \text{en otro caso,} \end{cases} \qquad I_{\{y_i>0\}} = \begin{cases} 1, & \text{si } y_i > 0, \\ 0, & \text{en otro caso.} \end{cases}$$

En cuanto a la inferencia sobre los parámetros y la diagnosis del modelo con ceros inflados, estos procesos son análogos a los vistos para los modelos usual y truncado en el cero.

Para obtener la predicción de la variable Y a partir del modelo de Poisson inflado en el cero, basta con tomar la predicción hecha para λ y aplicar una corrección. Esta corrección la hemos visto ya al inicio de la sección, puesto que habíamos afirmado que $E(Y) = (1 - \pi)\lambda$. Entonces, para una observación y_i, su predicción será:

$$\hat{y}_i = (1 - \pi(x_i, \hat{\alpha}))\lambda(x_i, \hat{\beta}).$$

En lo relativo a los contrastes e intervalos de confianza asintóticos, los elementos de matriz diagonal V ahora son de la forma $v_{i,i} = (1 - \hat{\pi}_i)(\hat{\lambda}_i + \hat{\pi}_i \cdot \hat{\lambda}_i^2)$, teniendo en cuenta la expresión de $Var(Y)$. Asimismo, la fórmula para los residuos de Pearson del modelo de Poisson inflado en el cero será:

$$r_i^P := \frac{y_i - \hat{y}_i}{\sqrt{(1 - \hat{\pi}_i)(\hat{\lambda}_i + \hat{\pi}_i \cdot \hat{\lambda}_i^2)}}, \quad \hat{\pi}_i = \pi(x_i, \hat{\alpha}), \quad \hat{\lambda}_i = \lambda(x_i, \hat{\beta}).$$

Observación 3.1. Si ajustamos un modelo de Poisson usual a unos datos con ceros inflados, notaremos que los residuos de Pearson presentarán sobre-dispersión, ya que la varianza de la distribución de Poisson inflada en el cero es mayor que su media. Para comprobarlo, hemos hecho una simulación en R de la siguiente manera. Primero, hemos generado 5 muestras aleatorias

de 100 observaciones cada una, procedentes de una distribución de Poisson inflada en el cero de parámetro $\lambda = 8$ y probabilidad de falso cero $\pi = 0.2$. A continuación, hemos ajustado sobre cada muestra un modelo de Poisson y otro de Poisson con ceros inflados, ambos sin variables explicativas, y hemos calculado la desviación típica de los residuos de Pearson de cada modelo. Finalmente, se ha repetido la misma simulación, pero considerando ahora $\pi = 0.6$. Las desviaciones típicas obtenidas para cada modelo y cada muestra se pueden ver en el Cuadro 3.2.

		Muestra 1	Muestra 2	Muestra 3	Muestra 4	Muestra 5
$\pi = 0.2$	Usual	1.500	1.554	1.639	1.640	1.480
	Inflado	0.968	0.982	0.967	1.024	1.023
$\pi = 0.6$	Usual	2.516	2.473	2.471	2.382	2.465
	Inflado	0.999	1.010	1.003	1.020	1.040

Cuadro 3.2: Desviación típica de los residuos de Pearson de un modelo de Poisson usual y un modelo de Poisson con ceros inflados para varias muestras de recuento aleatorias procedentes de una distribución de Poisson con ceros inflados de parámetro $\lambda = 8$, para $\pi = 0.2$ y $\pi = 0.6$.

Como vemos en el Cuadro 3.2, al ajustar un modelo de Poisson sobre las muestras infladas en el cero aparece sobre-dispersión en los residuos de Pearson, ya que la desviación típica de los residuos es mayor que 1 para las cinco muestras. Además, este exceso de dispersión es mayor cuando consideramos valores de π mayores. No obstante, al ajustar un modelo de ceros inflados, la desviación típica es prácticamente 1 en ambos casos, que es lo ideal para un modelo.

Ejemplo 3.2. Vamos a utilizar el conjunto `taxi` introducido previamente en este capítulo, e intentaremos predecir la cantidad de propina, en dólares estadounidenses, que recibirá un taxista utilizando la distancia del trayecto, medida en millas, como variable explicativa. Construyamos en primer lugar un modelo de Poisson inflado en el cero utilizando la variable distancia en el modelo de recuento, pero no en el modelo logístico. Para ello utilizaremos el paquete `pscl` [5, 6], no disponible en la instalación básica de R.

```
> m0 = zeroinfl(tip~distance | 1, data = taxi, dist = 'poisson',
    link = 'logit')
> summary(m0)

Count model coefficients (poisson with log link):

            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.28438    0.19414   1.465    0.143
```

```
distance     0.23304    0.05206   4.476  7.6e-06 ***

Zero-inflation model coefficients (binomial with logit link):

            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.1294     0.1900   0.681    0.496
---
```

Comencemos analizando la salida de R. Se puede apreciar que el comando `zeroinfl` genera dos tablas con coeficientes. Una de ellas, la primera, corresponde al modelo de Poisson para λ, mientras que la otra le corresponde con el modelo logístico construido para estimar π.

En cuanto al modelo de recuento, obtenemos las estimaciones $\hat{\beta}_0 = 0.284$ para el intercepto y $\hat{\beta}_1 = 0.233$ para la pendiente. Por su parte, en el modelo binomial obtenemos una estimación $\hat{\alpha}_0 = 0.129$. Esto dará lugar a una probabilidad de $\hat{\pi} = 0.532$, que permanecerá constante independientemente de la distancia del recorrido.

Ahora bien, podemos creer que la probabilidad de obtener un falso cero se puede ver también afectada por la distancia del trayecto en taxi, lo cual nos sugiere construir un segundo modelo en donde también incluyamos esa variable como explicativa del modelo logístico. Así, este modelo sería el siguiente:

```
> m = zeroinfl(tip~distance | distance, data = taxi, dist = 'poisson',
    link = 'logit')
> summary(m)

Count model coefficients (poisson with log link):

            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3320     0.1937   1.714   0.0865 .
distance      0.2214     0.0525   4.218 2.47e-05 ***

Zero-inflation model coefficients (binomial with logit link):

            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.5006     0.3878   1.291    0.197
distance     -0.1377     0.1279  -1.076    0.282
---
```

La novedad más llamativa es que ahora tenemos dos coeficientes en el modelo logístico, en lugar

de solamente uno, ya que estamos incluyendo la variable `distance` en el modelo. Al hacer esto, no solo cambian las estimaciones de los coeficientes α, sino que este cambio también afecta al modelo de Poisson, alterando las estimaciones de los coeficientes β. Esto se debe a que ahora, en la función de verosimilitud, estamos cambiando $\pi(\alpha)$ por $\pi(\texttt{distance}, \alpha)$, y esto afecta a los estimadores resultantes por máxima verosimilitud, incluidos los de β.

Las estimaciones de β ahora son $\hat{\beta}_0 = 0.332$ para el intercepto y $\hat{\beta}_1 = 0.221$ para la pendiente. En cuanto a la estimación de π, ahora ya no tenemos un único valor, si no que tenemos una función no constante que dependerá de la distancia del trayecto. Las estimaciones de α son $\hat{\alpha}_0 = 0.501$ para el intercepto y $\hat{\alpha}_1 = -0.137$ para la pendiente. Así, para las observaciones de nuestra muestra, las probabilidades de falso cero correspondientes forman un vector con el siguiente resumen:

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.4126  0.5076  0.5492  0.5410  0.5787  0.6223
```

Observamos que la probabilidad de obtener un falso cero aumenta con la distancia, lo cual tiene sentido, ya que la cantidad de propina también aumenta con la distancia. Esto hace que hace que aumente la media de esta última y, en consecuencia, la probabilidad de obtener un cero proveniente de la distribución de Poisson de parámetro esa media es cada vez menor. Es decir, la probabilidad de observar un falso cero es cada vez mayor.

En la Figura 3.3 se representa el ajuste de ambos modelos sobre los datos. Debido a los valores que toma π en cada uno de los modelos, observamos que el segundo modelo estima propinas más bajas que el primero para distancias pequeñas, pero sin embargo estima propinas más altas para distancias mayores.

Una vez tenemos los dos modelos construidos, surge la natural duda de cuál es mejor. Realizando un test de razón de verosimilitudes sobre ambos modelos, se obtiene lo siguiente:

```
> lmtest::lrtest(m0,m)

Likelihood ratio test

Model 1: tip ~ distance | 1
Model 2: tip ~ distance | distance
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   3 -192.53
2   4 -191.96  1 1.1497     0.2836
```

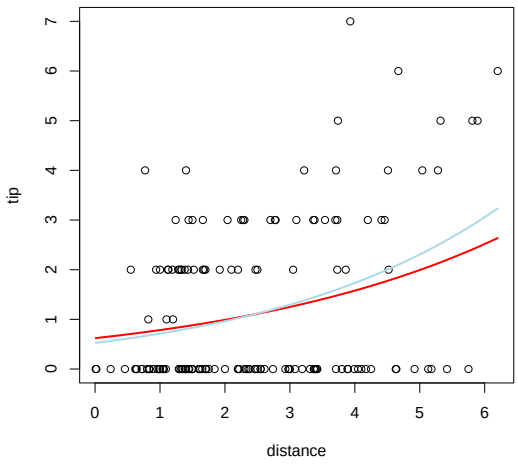

Figura 3.3: Diagrama de dispersión de la propina dada por el cliente al taxista frente a la distancia del trayecto, junto con las curvas de ajuste del primer modelo, en rojo, y del segundo modelo, en azul.

El nivel crítico del test es 0.284, el cual es suficientemente grande como para aceptar que el modelo donde la probabilidad π no depende de la distancia es mejor. Continuaremos el ejemplo utilizando este modelo.

Los intervalos de confianza para los coeficientes β y α al nivel de confianza del 95 % son los siguientes:

```
> confint(m0)
                          2.5 %    97.5 %
count_(Intercept) -0.09613411 0.6648970
count_distance     0.13099873 0.3350878
zero_(Intercept)  -0.24303428 0.5018366
```

Lo más destacable de los intervalos de confianza es que, en el caso de α_0, se tiene que el valor cero está dentro de su intervalo de confianza. Esto quiere decir que se podría suprimir el intercepto del modelo logístico sin que haya grandes consecuencias para el modelo inflado. Si lo suprimiésemos, la probabilidad de observar un falso cero estaría siendo estimada por 0.5.

Ahora veamos cuánta propina se espera que den los clientes, en media, para varias distancias:

```
        1         2         3         4         5
0.7846522 0.9905733 1.2505355 1.5787212 1.9930347
```

Como vemos, las medias de las propinas recibidas por el taxista para distancias enteras de entre 1 y 5 millas se moverán entre 0.785 y 1.993 dólares estadounidenses.

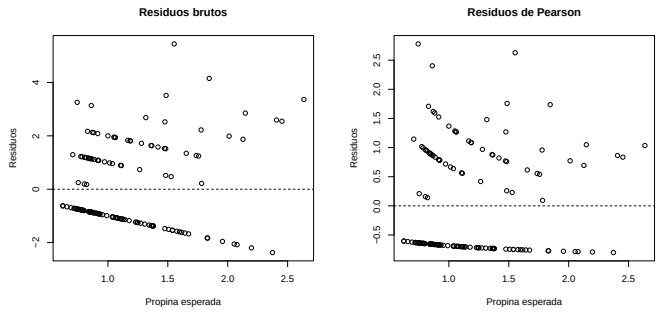

Figura 3.4: Diagrama de dispersión de los distintos tipos de residuos frente a la cantidad de propina esperada recibida por un taxista, para un modelo de Poisson con ceros inflados.

En cuanto a la diagnosis del modelo, debemos graficar los residuos del modelo contra los valores ajustados. A la vista de la Figura 3.4, se observa que los residuos de Pearson podrían tener una pequeña tendencia a decrecer cuanto mayor sea la propina estimada. Sin embargo, para predicciones menores que 1, la desviación típica de los residuos es de 0.909, mientras que para las mayores, es de 0.890. Estos valores son prácticamente iguales entre sí y no muy lejanos a 1. Tampoco se observa sobre-dispersión, ya que esas desviaciones no son mayores que 1, por lo que el modelo de Poisson es homocedástico y parece correcto.

3.4. El modelo Binomial Negativo inflado en el cero

Supongamos ahora que tenemos una variable de recuento Y con ceros inflados, pero ahora creemos que la distribución de recuento es una Binomial Negativa, puesto que hemos observado demasiada dispersión. Las expresiones para la distribución de Y en un caso de ceros inflados eran las siguientes:

$$P(Y = 0) = \pi + (1 - \pi)P(Z = 0)$$
$$P(Y = y) = (1 - \pi)P(Z = y), \quad y = 1, 2, 3, \ldots$$

Como ahora $Z \in BN(\mu, \theta)$, en lugar de ser una Poisson de parámetro λ, tenemos que la masa de probabilidad de la Binomial Negativa con ceros inflados es:

$$P(Y = 0) = \pi + (1 - \pi)\left(\frac{\theta}{\mu + \theta}\right)^{\theta}$$
$$P(Y = y) = (1 - \pi)\frac{\Gamma(\theta + y)}{\Gamma(\theta)y!}\frac{\mu^y \theta^{\theta}}{(\mu + \theta)^{\theta + y}}, \quad y = 1, 2, 3, \ldots$$

La media de Y tendrá la misma expresión que en la distribución de Poisson inflada en el cero,

$$E(Y) = (1-\pi)E(Z),$$

aunque ahora $Z \in BN(\mu, \theta)$, es decir,

$$E(Y) = (1-\pi)\mu.$$

La varianza de Y en este caso cambia su expresión, y se puede probar que es

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

Para el modelo Binomial Negativa inflado utilizamos los mismos dos modelos: el log-lineal para estimar μ, la media de la distribución Binomial Negativa, y el logístico para estimar π, la probabilidad de observar un falso cero. El primer modelo tiene la siguiente forma:

$$\mu(x, \beta) = e^{x'\beta},$$

siendo β los coeficientes del modelo, que en general tomarían valores diferentes a los de Poisson. El segundo no cambia:

$$\pi(x, \alpha) = \frac{e^{x'\alpha}}{1 + e^{x'\alpha}},$$

siendo también α los coeficientes del modelo, que en este caso tampoco tomarían los mismos valores que los del modelo de Poisson inflado.

La función de máxima verosimilitud que nos permite estimar α, β y la sobre-dispersión θ ahora tiene la siguiente apariencia:

$$L(\alpha, \beta, \theta) = \prod_{i=1}^{n}\Bigg(\left(\pi(x_i, \alpha) + (1-\pi(x_i, \alpha))\left(\frac{\theta}{\mu+\theta}\right)^{\theta}\right) I_{\{y_i=0\}} + \\ + (1-\pi(x_i, \alpha))\frac{\Gamma(\theta+y)}{\Gamma(\theta)y!}\frac{\mu^y\theta^\theta}{(\mu+\theta)^{\theta+y}} I_{\{y_i>0\}}\Bigg).$$

La inferencia, predicción y diagnosis se realizan siguiendo los mismos criterios que para el modelo de Poisson inflado en el cero, y las expresiones de $\hat{y}_i$, V y r_i^P son análogas, por lo que omitimos los detalles.

Ejemplo 3.3. Construyamos ahora un modelo Binomial Negativa inflado para los datos de los taxis, suponiendo que la probabilidad de observar un falso cero no depende de la distancia del trayecto.

```
> m0_nb = zeroinfl(tip~distance | 1, data = taxi, dist = 'negbin',
```

```
+ link = 'logit')
> summary(m0_nb)

Count model coefficients (negbin with log link):

             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.28438    0.19414   1.465    0.143
distance      0.23304    0.05206   4.476  7.6e-06 ***
Log(theta)   13.52174  103.23548   0.131    0.896

Zero-inflation model coefficients (binomial with logit link):

            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.1294     0.1900   0.681    0.496
---

Theta = 745447.174
```

Algo bastante llamativo del modelo Binomial Negativo con ceros inflados para estos datos es que las estimaciones de los coeficientes α y β son las mismas que teníamos para el modelo de Poisson inflado en el cero. Pero quizás lo realmente más llamativo sea el valor de la estimación del parámetro de sobre-dispersión, que es $\hat{\theta} = 7.45 \cdot 10^5$, el cual es inmensamente grande. Esto quiere decir que no hay sobre-dispersión, como ya habíamos visto en los residuos de Pearson del modelo de Poisson, y que además, por ser tan grande, la varianza del modelo Binomial Negativo inflado será prácticamente igual a su media. Es decir, el modelo es, a efectos prácticos, el mismo modelo que el modelo de Poisson con ceros inflados, y es por eso que tiene los mismos coeficientes.

De hecho, si calculamos la desviación típica de los residuos de Pearson para predicciones menores que 1, obtenemos un valor de 0.909, mientras que para mayores, obtenemos 0.890, que son exactamente las mismas que calculamos en el modelo de Poisson con ceros inflados.

Como consecuencia de lo anterior, es intuitivo que, si hacemos un test de razón de verosimilitudes sobre ambos modelos, vamos a llegar a la conclusión de que no hay absolutamente ningún motivo para escoger el modelo Binomial Negativo antes que el modelo de Poisson.

```
> lmtest::lrtest(m0,m0_nb)

Likelihood ratio test
```

```
  #Df  LogLik Df Chisq Pr(>Chisq)
1   3 -192.53
2   4 -192.53  1 1e-04     0.9911
```

En efecto, el nivel crítico del test es 0.991, es decir, casi 1, por lo que aceptamos que el modelo de Poisson con ceros inflados es mejor que el modelo Binomial Negativo inflado en el cero, como esperábamos.

Capítulo 4

Modelos de regresión para variables de recuento con ceros apartados

4.1. Introducción

También llamado *hurdle model*, en inglés, el modelo con ceros apartados consiste en suponer que los ceros observados en una variable de recuento proceden todos de una distribución diferente a la de los demás valores, a diferencia de lo que ocurría con ceros inflados, donde solo una parte de ellos procedía de otra distribución diferente. Entonces, estamos de nuevo en una situación donde los datos proceden de dos modelos distintos. En el caso de las observaciones positivas, estas procederán de una distribución de recuento, como podrá ser una Poisson o una Binomial Negativa. Pero hay que notar que, al ser estas observaciones positivas, tal distribución de recuento estará truncada en el cero, cosa que no ocurría en una situación de ceros inflados. En el caso de las observaciones nulas, estas provendrán de una distribución distinta con una cierta probabilidad. En el capítulo de ceros inflados, hablábamos de falsos ceros y ceros verdaderos, ya que había dos fuentes para esos ceros. Sin embargo, ahora ya solamente se habla de ceros, ya que consideramos que todos provienen de la misma fuente, que será diferente a la distribución de las observaciones positivas.

Una distribución con ceros apartados es otra forma de actuar ante una situación como la que se nos presentaba en ceros inflados, donde la cantidad de ceros observados era notablemente mayor de lo esperado. Pero a diferencia de una distribución inflada, las distribuciones con ceros apartados también son utilizables en un caso donde la cantidad de ceros es menor de lo esperado.

Para ilustrar una situación de ceros apartados, consideraremos el conjunto de datos `smoking` de la librería `openintro` [10], no incluida en la instalación básica de R. En él se recogen ob-

servaciones de varias características en individuos del Reino Unido, entre ellas, la cantidad de cigarrillos que fuma cada persona un día de fin de semana. En la Figura 4.1 se representa el diagrama de frecuencias de esta variable, y observamos que la cantidad de ceros que hay es muy grande, mucho mayor de lo que se esperaría en el caso en el que los recuentos procediesen de una distribución de Poisson o Binomial Negativa.

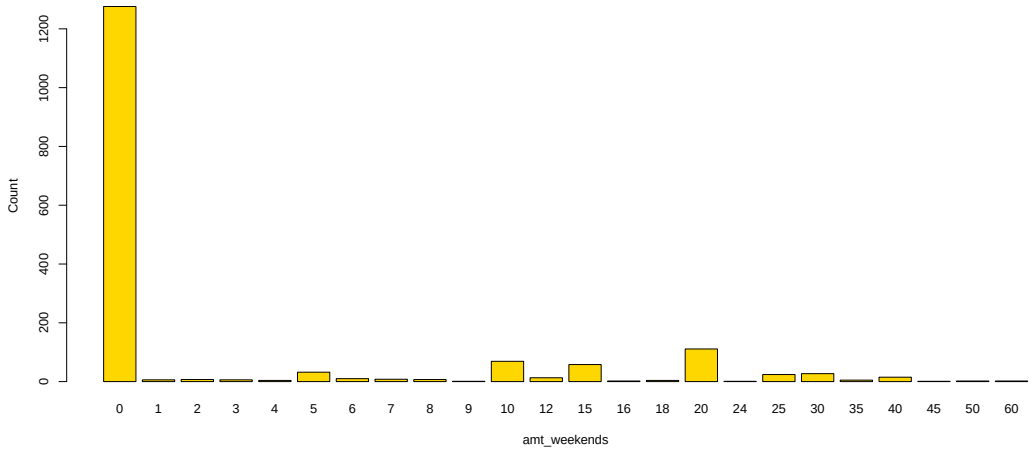

Figura 4.1: Diagrama de frecuencias del número de cigarrillos fumados por un individuo un día de fin de semana.

La razón de que haya una gran cantidad de ceros es la consecuencia de que, al hacer el recuento, también se ha incluido a individuos no fumadores, y estos representan la mayor parte de las observaciones. Si el individuo observado es no fumador, entonces el recuento para ese individuo será nulo, ya que si no lo fuese, contradiría al hecho de que es no fumador. Pero si el individuo fuma, entonces observaremos que el recuento de cigarrillos fumados un día de fin de semana será un número entero estrictamente positivo, que provendrá de una distribución truncada en el cero.

Las distribuciones con ceros apartados nos permiten proceder en una situación como la anterior, dándole un origen a todos esos ceros, y posibilitando el hecho de construir modelos de regresión para datos de recuento con exceso o escasez de ceros con respecto de los que se esperarían. En este capítulo estudiaremos la distribución de Poisson con ceros apartados y la Binomial Negativa con ceros apartados, y construiremos modelos de regresión para variables con esas distribuciones.

4.2. La distribución de Poisson con ceros apartados

Sea Y una variable de recuento, en la que se observa una cantidad sospechosamente anómala de ceros, ya sean demasiados, o demasiado pocos. Llamaremos π a la probabilidad de que esa variable sea cero, siendo $(1-\pi)$ la probabilidad de que tome el valor de una variable Z que sigue una distribución de Poisson truncada en el cero. La masa de probabilidad de la distribución de la variable Y, que es una distribución con ceros apartados, es la siguiente:

$$\begin{aligned} P(Y=0) &= \pi \\ P(Y=y) &= (1-\pi)P(Z=y), \quad y=1,2,3,\ldots \end{aligned}$$

Sustituyendo en la expresión de la distribución de Z, que recordemos que es una Poisson truncada en el cero, tenemos que:

$$\begin{aligned} P(Y=0) &= \pi \\ P(Y=y) &= (1-\pi)\frac{e^{-\lambda}}{1-e^{-\lambda}}\frac{\lambda^y}{y!}, \quad y=1,2,3,\ldots \end{aligned}$$

La media de la distribución de Y será

$$E(Y) = \sum_{y=0}^{\infty} yP(Y=y) = \sum_{y=1}^{\infty} y(1-\pi)\frac{e^{-\lambda}\lambda^y}{(1-e^{-\lambda})y!} = \frac{1-\pi}{1-e^{-\lambda}}\lambda\sum_{y=1}^{\infty}\frac{e^{-\lambda}\lambda^{y-1}}{(y-1)!}.$$

Tomando el cambio de variable $z=y-1$, obtenemos

$$E(Y) = \frac{1-\pi}{1-e^{-\lambda}}\lambda\sum_{z=0}^{\infty}\frac{e^{-\lambda}\lambda^z}{z!} = (1-\pi)\frac{\lambda}{1-e^{-\lambda}}.$$

Para la varianza, volvemos a tomar la descomposición $E(Y^2) = E(Y(Y-1)) + E(Y)$. Por un lado, tenemos

$$E(Y(Y-1)) = \sum_{y=0}^{\infty} y(y-1)(1-\pi)\frac{e^{-\lambda}\lambda^y}{(1-e^{-\lambda})y!} = \frac{1-\pi}{1-e^{-\lambda}}\lambda^2\sum_{y=2}^{\infty}\frac{e^{-\lambda}\lambda^{y-2}}{(y-2)!},$$

y haciendo el cambio $z=y-2$, resulta que

$$E(Y(Y-1)) = \frac{1-\pi}{1-e^{-\lambda}}\lambda^2\sum_{z=0}^{\infty}\frac{e^{-\lambda}\lambda^z}{z!} = \frac{1-\pi}{1-e^{-\lambda}}\lambda^2.$$

Utilizando esto, ahora tenemos

$$Var(Y) = E(Y^2) - E(Y)^2 = \frac{1-\pi}{1-e^{-\lambda}}(\lambda^2+\lambda) - \left((1-\pi)\frac{\lambda}{1-e^{-\lambda}}\right)^2.$$

En resumen, tenemos las siguientes expresiones para la media y la varianza:

$$\begin{aligned} E(Y) &= (1-\pi)\frac{\lambda}{1-e^{-\lambda}}, \\ Var(Y) &= \frac{1-\pi}{1-e^{-\lambda}}(\lambda+\lambda^2) - \left(\frac{1-\pi}{1-e^{-\lambda}}\cdot\lambda\right)^2. \end{aligned}$$

La varianza de una distribución de Poisson con ceros apartados no tiene un comportamiento tan sencillo como ocurría en la distribución truncada o la inflada. Esto se debe a que, por un lado, al aumentar λ, el hecho de que la distribución de los valores positivos sea truncada hace que la varianza disminuya con respecto a la de una distribución de Poisson usual. Sin embargo, por el fenómeno de la inflación del cero, la varianza aumenta con respecto a la de una distribución de Poisson usual. Entonces, no hay una fuerza que se imponga más en todos los casos, y tendremos tanto $E(Y) > Var(Y)$ como $E(Y) < Var(Y)$, dependiendo de los valores de λ y de π.

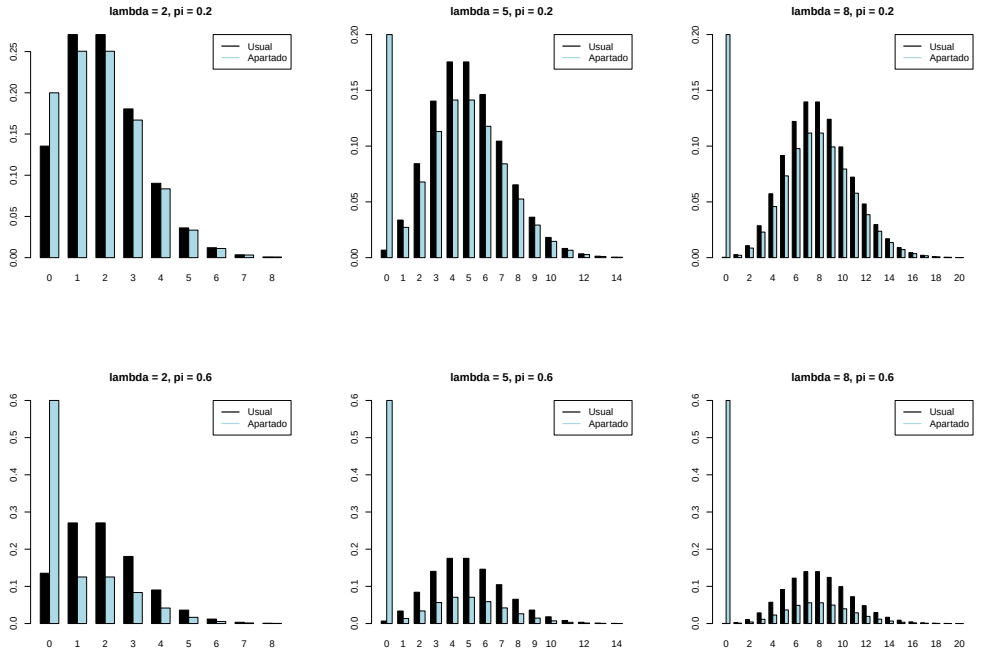

Figura 4.2: Diagramas de barras de la masa de probabilidad de las distribuciones de Poisson (en negro) y Poisson con ceros apartados (en azul), para diferentes valores de λ y diferentes valores de π.

En el Cuadro 4.1 tenemos una tabla donde podemos apreciar el fenómenos que acabamos de describir. Para $\pi = 0.2$, se tiene que la varianza es más pequeña que la media para $\lambda = 1$, pero a partir de $\lambda = 2$ ya ocurre al revés. En el caso de $\pi = 0.6$, ya se tiene que la varianza es mayor que la media a partir de $\lambda = 1$. Por último, también se observa que para valores grandes de λ, la media y la varianza de una distribución de Poisson con ceros apartados son prácticamente las mismas que para una distribución inflada en el cero, ya que la probabilidad de observar un cero verdadero en una distribución inflada es casi nula cuando λ es grande, y la distribución converge a una distribución con ceros apartados.

	λ	1	2	3	4	5	6	7	8	9	10
$\pi = 0.2$	$E(Y)$	1.27	1.85	2.53	3.26	4.03	4.81	5.61	6.40	7.20	8.00
	$Var(Y)$	0.93	2.13	3.72	5.67	7.95	10.53	13.42	16.63	20.16	24.00
$\pi = 0.6$	$E(Y)$	0.63	0.93	1.26	1.63	2.01	2.41	2.80	3.20	3.60	4.00
	$Var(Y)$	0.87	1.92	3.46	5.49	8.03	11.05	14.57	18.56	23.04	28.00

Cuadro 4.1: Media y varianza de una distribución de Poisson con ceros apartados para distintos valores del parámetro λ y para $\pi = 0.2$ y $\pi = 0.6$.

En la Figura 4.2 observamos una comparación de la distribución de Poisson usual y la distribución de Poisson con ceros apartados. Se puede apreciar cómo decrece la masa de probabilidad en los valores positivos cuando la probabilidad del cero es mayor, con respecto a la distribución de Poisson usual.

4.3. El modelo de Poisson con ceros apartados

Consideramos una muestra $(X_1, Y_1), \ldots, (X_n, Y_n)$, donde X son variables explicativas, e Y es una variable de recuento con exceso de ceros. Para estimar las observaciones positivas de la variable Y, consideraremos el modelo log-lineal de Poisson truncado en el cero. El parámetro λ vendrá dado por la expresión

$$\lambda(x, \beta) = e^{x'\beta},$$

donde x son las variables explicativas de una observación, y β son los coeficientes del modelo.

Una particularidad del modelo *hurdle* es que la parte logística modela la probabilidad de la presencia de un cero contra su ausencia, mientras que lo que se hacía en el modelo inflado era modelar la probabilidad de un falso cero contra otro tipo de dato. La consecuencia de esto es que los coeficientes α tendrán el signo contrario al que deberían en el modelo logístico presentado en el capítulo anterior. Para solventar esto, definiremos el modelo logístico para π con un signo negativo en los exponentes. Es decir:

$$\pi(x, \alpha) = \frac{e^{-x'\alpha}}{1 + e^{-x'\alpha}},$$

donde x son las variables explicativas de una observación, y α son los coeficientes del modelo logístico.

Los coeficientes α y β se estiman por máxima verosimilitud, donde la función de verosimilitud tiene la siguiente forma:

$$L(\alpha, \beta) = \prod_{i=1}^{n} \left(\pi(x_i, \alpha) I_{\{y_i = 0\}} + (1 - \pi(x_i, \alpha)) \frac{e^{-\lambda(x_i, \beta)}}{1 - e^{-\lambda}} \frac{\lambda(x_i, \beta)^{y_i}}{y_i!} I_{\{y_i > 0\}} \right).$$

La función de verosimilitud de un modelo con ceros apartados se puede expresar como una suma de dos funciones, de las cuáles una depende solamente de α, y la otra solamente de β. Por tanto, la función de verosimilitud se puede maximizar independientemente en α y en β, de forma que los valores de α que maximizan su función no influirán en los valores de β que maximizan su función. Como consecuencia, esto hace que las estimaciones de los coeficientes α no se vean alteradas al, por ejemplo, considerar otro tipo de distribución en el modelo de recuento.

En cuanto a la inferencia de parámetros y la diagnosis del modelo, no hay mucha diferencia, en lo que a procedimientos se refiere, con respecto al modelo usual, truncado o inflado en el cero.

Para hacer una predicción de la variable Y, debemos tomar la predicción de Z, y hacer la corrección adecuada para este modelo, como hemos indicado en la expresión de $E(Y)$. Entonces, la fórmula es la siguiente:

$$\hat{y}_i = (1 - \pi(x_i, \hat{\alpha}))\frac{\lambda(x_i, \hat{\beta})}{1 - e^{-\lambda(x_i,\hat{\beta})}}.$$

En lo que concierne a los contrastes e intervalos de confianza asintóticos, los elementos de matriz diagonal V tienen la forma $v_{i,i} = \frac{1-\hat{\pi}_i}{1-e^{-\hat{\lambda}_i}}(\hat{\lambda}_i + \hat{\lambda}_i^2) - \left(\frac{1-\hat{\pi}_i}{1-e^{-\hat{\lambda}_i}} \cdot \hat{\lambda}_i\right)^2$, teniendo en cuenta la expresión de $Var(Y)$. Entonces, la expresión para los residuos de Pearson del modelo de Poisson con ceros apartados será:

$$r_i^P := \frac{y_i - \hat{y}_i}{\sqrt{\frac{1-\hat{\pi}_i}{1-e^{-\hat{\lambda}_i}}(\hat{\lambda}_i + \hat{\lambda}_i^2) - \left(\frac{1-\hat{\pi}_i}{1-e^{-\hat{\lambda}_i}} \cdot \hat{\lambda}_i\right)^2}}, \quad \hat{\pi}_i = \pi(x_i, \hat{\alpha}), \quad \hat{\lambda}_i = \lambda(x_i, \hat{\beta}).$$

Observación 4.1. Al ajustar un modelo de Poisson usual a unos datos con ceros apartados, notaremos que los residuos de Pearson presentarán en algunos casos infra-dispersión y en otros casos sobre-dispersión, ya que, para λ pequeño, la varianza de la distribución de Poisson con ceros apartados puede ser menor que su media, mientras que para λ grande, será mayor. Para ilustrar este fenómeno, hemos realizado una simulación en R que consiste en lo siguiente. Para empezar, hemos generado 5 muestras aleatorias de 100 observaciones cada una, procedentes de una distribución de Poisson con ceros apartados de parámetro $\lambda = 1$ y probabilidad de cero $\pi = 0.3$. Luego, hemos realizado el ajuste sobre cada muestra de un modelo de Poisson y uno de Poisson con ceros apartados, ambos sin ninguna variable explicativa, y a continuación hemos calculado la desviación típica de los residuos de Pearson de cada modelo. Finalmente, se ha repetido la misma simulación, pero considerando ahora el parámetro $\lambda = 8$. Las desviaciones típicas obtenidas para cada modelo y cada muestra se pueden observar en el Cuadro 4.2.

Como vemos en el Cuadro 4.2, al ajustar un modelo de Poisson usual sobre las muestras obtenidas de una distribución de Poisson con ceros apartados de parámetro $\lambda = 1$, los residuos de Pearson presentan algo de infra-dispersión, mientras que los residuos del modelo con ceros apartados son muy próximos a 1. Para, $\lambda = 8$, ocurre lo contrario, los residuos del modelo

		Muestra 1	Muestra 2	Muestra 3	Muestra 4	Muestra 5
$\lambda = 1$	Usual	0.809	0.976	0.955	0.982	0.971
	Apartado	0.938	0.982	1.004	1.014	1.004
$\lambda = 8$	Usual	1.683	1.872	1.745	1.849	1.792
	Apartado	0.982	0.988	0.980	1.010	1.015

Cuadro 4.2: Desviación típica de los residuos de Pearson de un modelo de Poisson usual y un modelo de Poisson con ceros apartados para varias muestras de recuento aleatorias obtenidas de una distribución de Poisson con ceros apartados, tanto con $\lambda = 1$ como con $\lambda = 8$, y para $\pi = 0.3$.

de Poisson presentan sobre-dispersión, pero los del modelo con ceros apartados siguen siendo prácticamente 1.

Ejemplo 4.2. Vamos a construir un modelo de Poisson con ceros apartados para los datos del conjunto `smoking` introducido al inicio de este capítulo. Utilizando de nuevo el paquete `pscl`, en este caso la función `hurdle`, el modelo es:

```
> m = hurdle(amt_weekends~gender + marital_status| gender + marital_status,
+ data = smoking, dist = 'poisson', link = 'logit')
> summary(m)

Count model coefficients (truncated poisson with log link):

                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             2.81180    0.03213  87.513  < 2e-16 ***
genderMale              0.21810    0.02439   8.942  < 2e-16 ***
marital_statusMarried  -0.12075    0.03785  -3.191 0.001420 **
marital_statusSeparated -0.15003    0.06238  -2.405 0.016173 *
marital_statusSingle   -0.13001    0.03731  -3.485 0.000493 ***
marital_statusWidowed  -0.04729    0.04956  -0.954 0.339937
Zero hurdle model coefficients (binomial with logit link):

                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            -0.598033   0.168366  -3.552 0.000382 ***
genderMale              0.075917   0.118318   0.642 0.521111
marital_statusMarried  -1.007423   0.189807  -5.308 1.11e-07 ***
marital_statusSeparated -0.164348   0.306876  -0.536 0.592268
marital_statusSingle    0.009129   0.193564   0.047 0.962382
marital_statusWidowed  -0.974137   0.240951  -4.043 5.28e-05 ***
```

En la columna `Estimate` se nos proporcionan las estimaciones de los parámetros, tanto los $\hat{\beta}$ del modelo de Poisson truncado para λ como los $\hat{\alpha}$ del modelo logístico para π. Es importante hacer notar que ambas variables explicativas son de tipo factor. En el caso del género del individuo, este solamente toma dos valores, *Female* (mujer) y *Male* (hombre), mientras que para el estado civil, este puede tomar cinco valores, que son *Divorced* (Divorciado/a), *Married* (Casado/a), *Separated* (Separado/a), *Single* (Soltero/a) y *Widowed* (Viudo/a). Para cada variable explicativa i, con $i = 1, 2$, tenemos varios coeficientes $\beta_{i,j}$, siendo $i = 1$, $j = 1, 2$ para el género e $i = 2$, $j = 1, 2, 3, 4, 5$ para el estado civil, y ocurre lo mismo con los coeficientes α. Por ejemplo, $\hat{\beta}_{1,2} = 0.218$ sería la estimación del coeficiente del género para un hombre y $\hat{\beta}_{2,2} = -0.121$ la del estado civil para un individuo casado en el modelo de Poisson truncado. Por su parte $\hat{\alpha}_{1,2} = 0.076$ sería la estimación del género para un hombre y $\hat{\alpha}_{2,2} = -1.007$ la del estado civil para un individuo casado en el modelo logístico. Los casos de $j = 1$ correspondientes al género masculino y al estado civil divorciado/a, tienen coeficientes nulos, ya que se toman como categorías de referencia.

En el Cuadro 4.3 tenemos representado el número de cigarrillos que se espera que fume un individuo según su estado civil y su género. La media esperada para un hombre es en todos los casos mayor que para una mujer, y es máxima si está divorciado. En este caso, se espera que la media de los hombres divorciados sea 5.904 cigarrillos cada día de fin de semana. De la misma forma, para las mujeres casadas, que conformar el caso con la predicción mínima, la media se espera que sea de 2.466 cigarrillos.

En el Cuadro 4.3 también se refleja la relevancia de los coeficientes β del modelo. Por ejemplo, la estimación de $\beta_{2,5}$ tiene un nivel de significación de 0.340, que es bastante alto, lo cual nos indica que tal coeficiente podría ser prescindible sin demasiadas consecuencias para el modelo, y que la media para un individuo viudo es semejante a la de un individuo con otro estado civil. En concreto, se parecerá a la media de un individuo casado: para las mujeres viudas se esperan 2.728 cigarrillos, frente a los 2.466 de las casadas, mientras que para los hombres viudos se esperan 3.612 cigarrillos, frente a los 3.266 de los casados.

	Divorced	Married	Separated	Single	Widowed
Female	5.904	2.466	4.556	5.215	2.728
Male	7.706	3.266	5.964	6.805	3.612

Cuadro 4.3: Media de cigarrillos esperados para cada tipo de individuo en función de su género y su estado civil.

En el Cuadro 4.4 se facilita una tabla con las estimaciones de π para cada tipo de individuo

según su género y su estado civil. Por ejemplo, para un hombre casado, la probabilidad de ser no fumador es de 0.833, mientras que para una mujer divorciada es de 0.628. Observamos que el género no altera demasiado la probabilidad de ser no fumador. Esto ya lo podíamos sospechar al ver que la significación del coeficiente correspondiente al género en el modelo logístico es 0.521.

	Divorced	Married	Separated	Single	Widowed
Female	0.645	0.833	0.682	0.643	0.828
Male	0.628	0.822	0.665	0.626	0.817

Cuadro 4.4: Estimación de la probabilidad de que un individuo sea no fumador en función de su género y su estado civil.

En cuanto a la inferencia de los parámetros, los intervalos de confianza del nivel 95 % son:

```
> confint(m)
                                    2.5 %      97.5 %
count_(Intercept)              2.7488310  2.87477896
count_genderMale               0.1702950  0.26589799
count_marital_statusMarried   -0.1949308 -0.04657279
count_marital_statusSeparated -0.2722966 -0.02776114
count_marital_statusSingle    -0.2031273 -0.05688433
count_marital_statusWidowed   -0.1444238  0.04983888
zero_(Intercept)              -0.9280237 -0.26804206
zero_genderMale               -0.1559825  0.30781644
zero_marital_statusMarried    -1.3794374 -0.63540945
zero_marital_statusSeparated  -0.7658149  0.43711838
zero_marital_statusSingle     -0.3702486  0.38850728
zero_marital_statusWidowed    -1.4463913 -0.50188213
```

Las 6 primeras filas se corresponden con los coeficientes β, mientras que las 6 últimas se corresponden con los coeficientes α. Los intervalos de confianza de los coeficientes β son notoriamente más pequeños en longitud que los de α, ya que los errores típicos de los primeros coeficientes son menor que los de los segundos, como se ve en la columna `Std. Error` del resumen del modelo de ceros apartados.

En cuanto a la diagnosis del modelo, basta con observar el comportamiento de los residuos en la Figura 4.3. Los residuos de Pearson no parecen tener a simple vista ninguna tendencia clara independientemente del número de cigarrillos esperados. Además, se tiene que para predicciones menores que 4, la desviación típica de los residuos es de 1.120, mientras que para las mayores,

es de 1.192. No son tan diferentes como para dudar razonablemente de la homocedasticidad del modelo, pero sí son ambas mayores que 1, por lo que el modelo presenta sobre-dispersión.

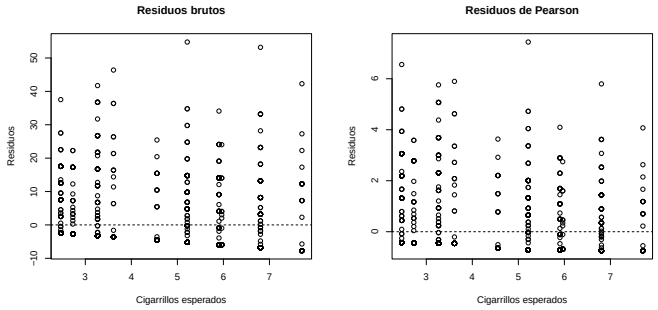

Figura 4.3: Diagrama de dispersión de los diferentes tipos de residuos frente al número de cigarrillos esperados en un día de fin de semana, para un modelo de Poisson con ceros apartados.

4.4. El modelo Binomial Negativo con ceros apartados

Asumamos que ahora la variable Z, en lugar de seguir una distribución de Poisson truncada, sigue una Binomial Negativa truncada. Teníamos las siguientes expresiones para la distribución de Y:

$$P(Y=0) = \pi + (1-\pi)P(Z=0)$$
$$P(Y=y) = (1-\pi)P(Z=y), \quad y = 1,2,3,\ldots$$

Por tanto, sustituyendo la distribución de Z en ambas, se tiene:

$$P(Y=0) = \pi,$$
$$P(Y=y) = (1-\pi)\frac{\Gamma(\theta+y)}{\Gamma(\theta)y!}\frac{\mu^y\theta^\theta}{(\mu+\theta)^{\theta+y}}\frac{1}{1-\left(\frac{\theta}{\mu+\theta}\right)^\theta}, \quad y = 1,2,3,\ldots$$

La expresión de la media de Y ahora cambia, ya que, como Z es ahora una Binomial Negativa truncada, al sustituir $E(Z)$ en

$$E(Y) = (1-\pi)E(Z),$$

obtenemos

$$E(Y) = (1-\pi)\frac{\mu}{\left(\frac{\theta}{\mu+\theta}\right)^\theta}.$$

La expresión para la varianza es

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

Entonces, el modelo para μ viene a ser el mismo que teníamos para λ, de la forma

$$\mu(x,\beta) = e^{x'\beta},$$

donde los coeficientes del modelo, β, serán distintos de los que se obtenían en el modelo de Poisson con ceros apartados. Sin embargo, el modelo logístico es el mismo,

$$\pi(x,\alpha) = \frac{e^{-x'\alpha}}{1+e^{-x'\alpha}},$$

y los coeficientes α del modelo sí serán iguales a los del modelo de Poisson con ceros apartados, como hemos explicado anteriormente.

La función de verosimilitud toma la siguiente forma:

$$L(\alpha,\beta,\theta) = \prod_{i=1}^{n}\left(\pi(x_i,\alpha)I_{\{y_i=0\}} + \right.$$
$$\left. + (1-\pi(x_i,\alpha))\frac{\Gamma(\theta+y_i)}{\Gamma(\theta)y_i!}\frac{\mu(x_i,\beta)^y\theta^{\theta}}{(\mu(x_i,\beta)+\theta)^{\theta+y_i}}\frac{1}{1-\left(\frac{\theta}{\mu(x_i,\beta)+\theta}\right)^{\theta}}I_{\{y_i>0\}}\right).$$

En lo que respecta a la inferencia de los parámetros y la diagnosis del modelo, se siguen aplicando los mismos procedimientos, y las expresiones para $\hat{y}_i$, V y r_i^P son análogas al modelo de Poisson con ceros apartados.

Ejemplo 4.3. Para terminar este estudio, construyamos un modelo Binomial Negativo con ceros apartados para nuestros datos del conjunto `smoking`:

```
> m_nb = hurdle(amt_weekends~gender + marital_status| gender + marital_status,
+ data = smoking, dist = 'negbin', link = 'logit')
> summary(m_nb)

Count model coefficients (truncated negbin with log link):

                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            2.80989    0.07932  35.425  < 2e-16 ***
genderMale             0.21725    0.05937   3.659 0.000253 ***
```

```
marital_statusMarried   -0.12374    0.09332  -1.326 0.184823
marital_statusSeparated -0.13508    0.14909  -0.906 0.364925
marital_statusSingle    -0.12562    0.09168  -1.370 0.170610
marital_statusWidowed   -0.06298    0.12287  -0.513 0.608252
Log(theta)               1.23724    0.08795  14.067  < 2e-16 ***
Zero hurdle model coefficients (binomial with logit link):

                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             -0.598033   0.168366  -3.552 0.000382 ***
genderMale               0.075917   0.118318   0.642 0.521111
marital_statusMarried   -1.007423   0.189807  -5.308 1.11e-07 ***
marital_statusSeparated -0.164348   0.306876  -0.536 0.592268
marital_statusSingle     0.009129   0.193564   0.047 0.962382
marital_statusWidowed   -0.974137   0.240951  -4.043 5.28e-05 ***
---
Theta: count = 3.4461
```

Las estimaciones $\hat{\alpha}$ no cambian con respecto al modelo de Poisson con ceros apartados, por el motivo que se ha explicado anteriormente. Sin embargo, las estimaciones $\hat{\beta}$ sí cambian ligeramente, ya que ahora estamos suponiendo que la distribución de nuestra variable de recuento es una Binomial Negativa truncada en el cero. El parámetro de sobre-dispersión es estimado por $\hat{\theta}$ = 3.446. Es bastante pequeño, lo cual quiere decir que la sobre-dispersión presente en nuestros datos es alta, como ya habíamos visto en los residuos del modelo de Poisson con ceros apartados, y lo más probable es que el modelo Binomial Negativo con ceros apartados sea más correcto para nuestros datos que el modelo de Poisson con ceros apartados. De hecho, si calculamos la desviación típica de los residuos de Pearson para predicciones menores que 4, se obtiene 0.978, mientras que para las mayores, se obtiene 0.999. Es decir, ambas son prácticamente iguales y muy próximas a 1. Esto se puede apreciar en la Figura 4.4, donde vemos que los residuos de Pearson ahora tiene una menor varianza que la de los del modelo de Poisson con ceros apartados.

```
> lmtest::lrtest(m,m_nb)
Likelihood ratio test

  #Df  LogLik Df  Chisq Pr(>Chisq)
1  12 -2950.1
2  13 -2393.9  1 1112.3  < 2.2e-16 ***
---
```

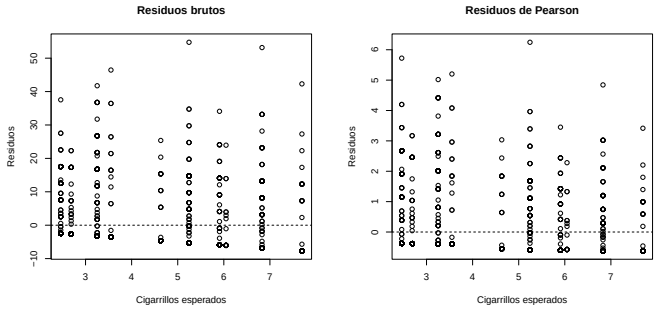

Figura 4.4: Diagrama de dispersión de los diferentes tipos de residuos frente al número de cigarrillos esperados en un día de fin de semana, para un modelo Binomial Negativo con ceros apartados.

Realizando un test de razón de verosimilitudes sobre ambos modelos, vemos que efectivamente el nivel crítico de test es prácticamente nulo, lo cual quiere decir que rechazamos el modelo de Poisson con ceros apartados en favor del Binomial Negativo con ceros apartados, y este último es mejor, como ya habíamos anticipado.

Anexo I

Códigos de R

I.1. Capítulo 1

I.1.1. Distribución Binomial Negativa

```
rm(list=ls(all=TRUE))

# Tabla de sobre-dispersión y varianza

mu = 4
theta = c(1,2,5,10,20,50,100,200,500,1000)
s2 = mu + mu^2/theta
tab = rbind(theta,s2)
colnames(tab) = 1:length(theta)
rownames(tab) = c('\u03b8', 'Var(Y)')
round(tab, 2)

mu = 8
theta = c(1,2,5,10,20,50,100,200,500,1000)
s2 = mu + mu^2/theta
tab = rbind(theta,s2)
colnames(tab) = 1:length(theta)
rownames(tab) = c('\u03b8', 'Var(Y)')
round(tab, 2)
```

```
# Gráficas

# Theta = 2

lambda = 4
mu = lambda
theta = 2
x = 0:12
pois = dpois(x, lambda = lambda)
nb = dnbinom(x, mu = mu, size = theta)
col = c('black', 'lightblue')
barplot(rbind(pois, nb), beside = TRUE, col = col, names.arg = x,
main = 'theta = 2')
legend(x = 'topright', legend = c('Poisson', 'Binomial Negativa'), col = col,
lty = 1, lwd = 2)

# Theta = 10

lambda = 4
mu = lambda
theta = 10
x = 0:12
pois = dpois(x, lambda = lambda)
nb = dnbinom(x, mu = mu, size = theta)
col = c('black', 'lightblue')
barplot(rbind(pois, nb), beside = TRUE, col = col, names.arg = x,
main = 'theta = 10')
legend(x = 'topright', legend = c('Poisson', 'Binomial Negativa'), col = col,
lty = 1, lwd = 2)

# Theta = 30

lambda = 4
mu = lambda
theta = 30
x = 0:12
pois = dpois(x, lambda = lambda)
nb = dnbinom(x, mu = mu, size = theta)
```

```
col = c('black', 'lightblue')
barplot(rbind(pois, nb), beside = TRUE, col = col, names.arg = x,
main = 'theta = 30')
legend(x = 'topright', legend = c('Poisson', 'Binomial Negativa'), col = col,
lty = 1, lwd = 2)
```

I.1.2. Modelos usuales (ejemplo de las telas)

```
rm(list=ls(all=TRUE))

library(boot)
data(cloth)
colnames(cloth) = c('Length','Flaws')

summary(cloth)
dim(cloth)

# Modelo de Poisson

modelo = glm(Flaws~Length, data = cloth, family = poisson(link = log))
summary(modelo)
beta = coef(modelo)
exp(coef(modelo))
plot(Flaws~Length, data = cloth, xlab = 'Longitud', ylab = 'Defectos')
curve(exp(beta[1] + beta[2]*x), add = TRUE, col = 'red', lwd = 2)

# Inferencia

exp(confint(modelo))

# Diagnosis

flaws = exp(predict(modelo))

r1 = residuals(modelo, type = 'response')
plot(r1~flaws, xlab = 'Defectos esperados', ylab = 'Residuos',
main = 'Residuos brutos')
abline(h=0,lty = 2)
```

```
r2 = residuals(modelo, type = 'pearson')
plot(r2~flaws, xlab = 'Defectos esperados', ylab = 'Residuos',
main = 'Residuos de Pearson')
abline(h=0,lty = 2)

r3 = residuals(modelo, type = 'deviance')
plot(r3~flaws, xlab = 'Defectos esperados', ylab = 'Residuos',
main = 'Residuos de la deviance')
abline(h=0,lty = 2)

sd(r2[flaws<=8.5])
sd(r2[flaws>8.5])

# Predicción

x.new = data.frame(Length = c(3,4,5,6,7))
fit = predict(modelo, newdata = x.new, type = 'response')
round(rbind(t(x.new),'Predicted flaws' = fit), 2)



# Estimación de los errores típicos de beta a mano

V = diag(modelo$fit)
X = model.matrix(modelo)
S2 = solve(t(X)%*%V%*%X)
dt = sqrt(diag(S2)); dt

# Contrastes a mano

beta/dt
2*pnorm(beta/dt, lower.tail = FALSE)

# Intervalos de confianza a mano

z = qnorm(0.025, lower.tail = FALSE)
exp(cbind('2.5%' = beta-z*dt, '97.5%' = beta+z*dt))
```

```
########################################

# Binomial Negativa

library(MASS)

mod_NB = glm.nb(Flaws~Length, data = cloth, link = log)
summary(mod_NB)
beta_NB = coef(mod_NB)
exp(coef(mod_NB))
plot(Flaws~Length, data = cloth, xlab = 'Longitud', ylab = 'Defectos')
curve(exp(beta_NB[1] + beta_NB[2]*x), add = TRUE, col = 'red', lwd = 2)

flaws_NB = exp(predict(mod_NB))

r1_NB = residuals(mod_NB, type = 'response')
plot(r1_NB~flaws_NB, xlab = 'Defectos esperados', ylab = 'Residuos',
main = 'Residuos brutos')
abline(h=0,lty = 2)

r2_NB = residuals(mod_NB, type = 'pearson')
plot(r2_NB~flaws_NB, xlab = 'Defectos esperados', ylab = 'Residuos',
main = 'Residuos de Pearson')
abline(h=0,lty = 2)

r3_NB = residuals(mod_NB, type = 'deviance')
plot(r3_NB~flaws_NB, xlab = 'Defectos esperados', ylab = 'Residuos',
main = 'Residuos de la deviance')
abline(h=0,lty = 2)

sd(r2_NB[flaws_NB<=8.5])
sd(r2_NB[flaws_NB>8.5])

library(lmtest)
lmtest::lrtest(modelo,mod_NB)

# Estimación ad hoc
```

```
mod_qp = glm(Flaws~Length, data = cloth, family = quasipoisson(link = log))
summary(mod_qp)
beta_qp = coef(mod_qp)
exp(coef(mod_qp))
plot(Flaws~Length, data = cloth, xlab = 'Longitud', ylab = 'Defectos')
curve(exp(beta_qp[1] + beta_qp[2]*x), add = TRUE, col = 'red', lwd = 2)
```

I.2. Capítulo 2

I.2.1. Distribución de Poisson truncada

```
rm(list=ls(all=TRUE))

# Tabla de media y varianza

lambda = 1:10
mu = lambda/(1-exp(-lambda))
s2 = mu*(1+lambda-mu)
tab = rbind(lambda,mu,s2)
colnames(tab) = lambda
rownames(tab) = c('\u03bb', 'E(Y)', 'Var(Y)')
round(tab, 3)



# Gráficas

col = c('black','lightblue')

L1 = 2
pois = dpois(0:8, lambda = L1)
poist = c(0,dpois(1:8, lambda = L1)/(1-dpois(0,lambda = L1)))
barplot(rbind(pois,poist), col = col, beside = TRUE, main = 'lambda = 2',
names.arg=0:8)
legend(x = "topright", legend = c('Usual','Truncado'), lty = 1, col = col,
lwd = 2)

L1 = 5
```

```
pois = dpois(0:14, lambda = L1)
poist = c(0,dpois(1:14, lambda = L1)/(1-dpois(0,lambda = L1)))
barplot(rbind(pois,poist), col = col, beside = TRUE, main = 'lambda = 5',
names.arg=0:14)
legend(x = "topright", legend = c('Usual','Truncado'), lty = 1, col = col,
lwd = 2)

L1 = 8
pois = dpois(0:20, lambda = L1)
poist = c(0,dpois(1:20, lambda = L1)/(1-dpois(0,lambda = L1)))
barplot(rbind(pois,poist), col = col, beside = TRUE, main = 'lambda = 8',
names.arg=0:20)
legend(x = "topright", legend = c('Usual','Truncado'), lty = 1, col = col,
lwd = 2)


# Simulación Poisson vs Poisson truncada en el cero

library(VGAM)
set.seed(27032023)
M = 5
n = 100
sdp = numeric(M)
sdt = numeric(M)

for (j in 1:M){
y = numeric(n)
for (i in 1:n){
yi = 0
while (yi==0){
yi = rpois(1,lambda = 2)
}
y[i] = yi
}
mp = glm(y~1, family = poisson(link = 'log'))
mt = vglm(y~1, family = pospoisson)
rp = residuals(mp, type = 'pearson')
rt = residuals(mt, type = 'pearson')
```

```
sdp[j] = sd(rp)
sdt[j] = sd(rt)
}

tab = rbind(sdp,sdt)
colnames(tab) = paste('Muestra',1:M)
rownames(tab) = c('Usual','Truncado')
round(tab,3)
```

I.2.2. Modelos con ceros truncados (ejemplo de las viviendas)

```
# https://www.kaggle.com/datasets/stealthtechnologies/
# regression-dataset-for-household-income-analysis

rm(list=ls(all=TRUE))

house = read.csv('house.csv', header = TRUE, strings = TRUE)
colnames(house) = c('age','education','occupation','dependents','location',
  'workexp','marital_status','employment_status','size',
  'homeown','type','gender','transport','income')

# Nos quedamos con un subconjunto más pequeño

set.seed(27032002)
house = house[sample(1:nrow(house),200),]
house = house[,c('size', 'age', 'income', 'marital_status')]
summary(house)
barplot(c(0,table(house$size)), col = 'gold', xlab = 'Household size',
ylab = 'Count', names.arg = 0:7, ylim = c(0,40))

# Modelo de Poisson truncado en el cero

library(VGAM)

m = vglm(size~age + income + marital_status, data = house,
family = pospoisson)
summary(m)
```

```
# Inferencia

confint(m)

# Predicción

x.new = data.frame(age = c(20,35,50),
   income = c(60000,140000,100000),
   marital_status = c('Single','Married','Divorced'))
new.fit = round(predict(m, newdata = x.new, type = 'response'),3)
tab = data.frame(Age = x.new[,1],
     Income = x.new[,2],
     Marital_Status = x.new[,3],
     Predicted_Size = new.fit)
tab

# Diagnosis

r1 = residuals(m, type = 'response')
plot(r1~fitted(m), main = 'Residuos brutos',
xlab = 'Tamaño esperado', ylab = 'Residuos')
abline(h=0, lty = 2)

r2 = residuals(m, type = 'pearson')
plot(r2~fitted(m), main = 'Residuos de Pearson',
xlab = 'Tamaño esperado', ylab = 'Residuos')
abline(h=0, lty = 2)

sd(r2[fitted(m)<=3.8])
sd(r2[fitted(m)>3.8])

# Modelo Binomial Negativo truncado en el cero

m_nb = vglm(size~age + income + marital_status,
data = house, family = posnegbinomial)
summary(m_nb)

exp(coef(m_nb)[2])
```

```
r2_nb = residuals(m_nb, type = 'pearson')[,1]
plot(r2_nb~fitted(m_nb), main = 'Residuos de Pearson',
xlab = 'Tamaño esperado', ylab = 'Residuos')
abline(h=0, lty = 2)

sd(r2_nb[fitted(m_nb)<=3.8])
sd(r2_nb[fitted(m_nb)>3.8])
```

I.3. Capítulo 3

I.3.1. Distribución de Poisson inflada

```
rm(list=ls(all=TRUE))

# Tabla de medias y varianzas

lambda = 1:10
p = 0.2
mu = (1-p)*lambda
s2 = (1-p)*(lambda+p*lambda^2)
tab = rbind(lambda,mu,s2)
colnames(tab) = lambda
rownames(tab) = c('\u03bb', 'E(Y)', 'Var(Y)')
round(tab, 3)

lambda = 1:10
p = 0.6
mu = (1-p)*lambda
s2 = (1-p)*(lambda+p*lambda^2)
tab = rbind(lambda,mu,s2)
colnames(tab) = lambda
rownames(tab) = c('\u03bb', 'E(Y)', 'Var(Y)')
round(tab, 3)


# Gráficas
```

```
col = c('black','lightblue')

# L = 2 | p = 0.2

L = 2
p = 0.2
x = 0:8
pois = dpois(x, lambda = L)
p0 = p + (1-p)*dpois(0, lambda = L)
py = (1-p)*dpois(x[x>0], lambda = L)
pois_zi = c(p0,py)
barplot(rbind(pois,pois_zi), beside = TRUE, col = col,names.arg = x,
main = 'lambda = 2, pi = 0.2')
legend(x = "topright", legend = c('Usual','Inflado'), lty = 1, col = col,
lwd = 2)

# L = 5 | p = 0.2

L = 5
p = 0.2
x = 0:14
pois = dpois(x, lambda = L)
p0 = p + (1-p)*dpois(0, lambda = L)
py = (1-p)*dpois(x[x>0], lambda = L)
pois_zi = c(p0,py)
barplot(rbind(pois,pois_zi), beside = TRUE, col = col,names.arg = x,
main = 'lambda = 5, pi = 0.2')
legend(x = "topright", legend = c('Usual','Inflado'), lty = 1, col = col,
lwd = 2)

# L = 8 | p = 0.2

L = 8
p = 0.2
x = 0:20
pois = dpois(x, lambda = L)
p0 = p + (1-p)*dpois(0, lambda = L)
```

```
py = (1-p)*dpois(x[x>0], lambda = L)
pois_zi = c(p0,py)
barplot(rbind(pois,pois_zi), beside = TRUE, col = col,names.arg = x,
main = 'lambda = 8, pi = 0.2')
legend(x = "topright", legend = c('Usual','Inflado'), lty = 1, col = col,
lwd = 2)


# L = 2 | p = 0.6

L = 2
p = 0.6
x = 0:8
pois = dpois(x, lambda = L)
p0 = p + (1-p)*dpois(0, lambda = L)
py = (1-p)*dpois(x[x>0], lambda = L)
pois_zi = c(p0,py)
barplot(rbind(pois,pois_zi), beside = TRUE, col = col,names.arg = x,
main = 'lambda = 2, pi = 0.6')
legend(x = "topright", legend = c('Usual','Inflado'), lty = 1, col = col,
lwd = 2)

# L = 5 | p = 0.6

L = 5
p = 0.6
x = 0:14
pois = dpois(x, lambda = L)
p0 = p + (1-p)*dpois(0, lambda = L)
py = (1-p)*dpois(x[x>0], lambda = L)
pois_zi = c(p0,py)
barplot(rbind(pois,pois_zi), beside = TRUE, col = col,names.arg = x,
main = 'lambda = 5, pi = 0.6')
legend(x = "topright", legend = c('Usual','Inflado'), lty = 1, col = col,
lwd = 2)

# L = 8 | p = 0.6
```

```
L = 8
p = 0.6
x = 0:20
pois = dpois(x, lambda = L)
p0 = p + (1-p)*dpois(0, lambda = L)
py = (1-p)*dpois(x[x>0], lambda = L)
pois_zi = c(p0,py)
barplot(rbind(pois,pois_zi), beside = TRUE, col = col,names.arg = x,
main = 'lambda = 8, pi = 0.6')
legend(x = "topright", legend = c('Usual','Inflado'), lty = 1, col = col,
lwd = 2)


# Simulación Poisson vs Poisson inflada en el cero

# p = 0.2

library(pscl)
set.seed(27032023)
M = 5
n = 100
p = 0.2
sdp = numeric(M)
sdi = numeric(M)

for (j in 1:M){
b = rbinom(n, size = 1, prob = p)
y = (1-b)*rpois(n, lambda = 8)
mp = glm(y~1, family = poisson(link = 'log'))
mi = zeroinfl(y~1 | 1)
rp = residuals(mp, type = 'pearson')
ri = residuals(mi, type = 'pearson')
sdp[j] = sd(rp)
sdi[j] = sd(ri)
}

tab = rbind(sdp,sdi)
colnames(tab) = paste('Muestra',1:M)
```

```
rownames(tab) = c('Usual','Inflado')
round(tab,3)

# p = 0.6

set.seed(27032023)
M = 5
n = 100
p = 0.6
sdp = numeric(M)
sdi = numeric(M)

for (j in 1:M){
b = rbinom(n, size = 1, prob = p)
y = (1-b)*rpois(n, lambda = 8)
mp = glm(y~1, family = poisson(link = 'log'))
mi = zeroinfl(y~1 | 1)
rp = residuals(mp, type = 'pearson')
ri = residuals(mi, type = 'pearson')
sdp[j] = sd(rp)
sdi[j] = sd(ri)
}

tab = rbind(sdp,sdi)
colnames(tab) = paste('Muestra',1:M)
rownames(tab) = c('Usual','Inflado')
round(tab,3)
```

I.3.2. Modelos con ceros inflados (ejemplo de los taxis)

```
# https://www.kaggle.com/datasets/anandaramg/taxi-trip-data-nyc

rm(list=ls(all=TRUE))

taxi = read.csv('taxi.csv', header = TRUE, strings = TRUE)
taxi = taxi[,c('tip_amount', 'fare_amount', 'trip_distance')]
taxi = taxi[taxi$trip_distance>0 & taxi$fare_amount>0
& taxi$trip_distance<=6.20 & taxi$fare<=26.83,]
```

```
# Nos quedamos con un subconjunto más pequeño y redondeamos las propinas

set.seed(27032002)
taxi = taxi[sample(1:nrow(taxi),150),c('tip_amount','trip_distance')]
taxi$tip_amount = round(taxi$tip_amount)

colnames(taxi) = c('tip','distance')
taxi = taxi[order(taxi$distance),]

plot(tip~distance, data = taxi)



summary(taxi)
barplot(table(taxi$tip), xlab = 'Tip', ylab = 'Count', col = 'gold')

# Modelo de Poisson inflado

library(pscl)

m0 = zeroinfl(tip~distance | 1, data = taxi, dist = 'poisson',
link = 'logit')
summary(m0)
p0 = predict(m0, type = 'zero')

m = zeroinfl(tip~distance | distance, data = taxi, dist = 'poisson',
link = 'logit')
summary(m)
p = predict(m, type = 'zero')
summary(p)

plot(tip~distance, data = taxi)
points(taxi$distance, m0$fit, col = 'red', pch = 19, type = 'l', lwd = 2)
points(taxi$distance, m$fit, col = 'lightblue', pch = 19, type = 'l', lwd = 2)

library(lmtest)
lmtest::lrtest(m0,m)
```

```
#--- Cálculo de E(Y) y Var(Y) a mano ---#

X = model.matrix(m0, model = 'count')
Z = model.matrix(m0, model = 'zero')
beta = as.matrix(m0$coef$count)
alfa = as.matrix(m0$coef$zero)
L = exp(X%*%beta)
p = exp(Z%*%alfa)/(1 + exp(Z%*%alfa))
mu = (1-p)*L
s2 = (1-p)*(L + p*L^2)

# Inferencia

confint(m0)

# Predicción

x.new = data.frame(distance = c(1,2,3,4,5))
predict(m0, newdata = x.new, type = 'response')

# Diagnosis

r1 = residuals(m0, type = 'response')
plot(r1~m0$fit, main = 'Residuos brutos',
xlab = 'Propina esperada', ylab = 'Residuos')
abline(h=0, lty = 2)

r2 = residuals(m0, type = 'pearson')
plot(r2~m0$fit, main = 'Residuos de Pearson',
xlab = 'Propina esperada', ylab = 'Residuos')
abline(h=0, lty = 2)

sd(r2[fitted(m0)<=1])
sd(r2[fitted(m0)>1])

# Modelo Binomial Negativo inflado

m0_nb = zeroinfl(tip~distance | 1, data = taxi, dist = 'negbin',
```

```
link = 'logit')
summary(m0_nb)

X_nb = model.matrix(m0_nb, model = 'count')
Z_nb = model.matrix(m0_nb, model = 'zero')
beta_nb = as.matrix(m0_nb$coef$count)
alfa_nb = as.matrix(m0_nb$coef$zero)
theta = m0_nb$theta
M = exp(X_nb%*%beta_nb)
p = exp(Z_nb%*%alfa_nb)/(1 + exp(Z_nb%*%alfa_nb))
mu.nb = (1-p)*M
s2.nb = (1-p)*(M + M^2/theta + p*M^2)

r2_nb = residuals(m0_nb, type = 'pearson')
sd(r2_nb[fitted(m0_nb)<=1])
sd(r2_nb[fitted(m0_nb)>1])

# Test de verosimilitud sobre ambos modelos

lmtest::lrtest(m0,m0_nb)
```

I.4. Capítulo 4

I.4.1. Distribución de Poisson con ceros apartados

```
rm(list=ls(all=TRUE))

# Tabla de medias y varianzas

lambda = 1:10
p = 0.2
mu = (1-p)*lambda/(1-exp(-lambda))
s2=(1-p)/(1-exp(-lambda))*(lambda+lambda^2)-((1-p)/(1-exp(-lambda))*lambda)^2
tab = rbind(lambda,mu,s2)
colnames(tab) = lambda
rownames(tab) = c('\u03bb', 'E(Y)', 'Var(Y)')
round(tab, 2)
```

```
lambda = 1:10
p = 0.6
mu = (1-p)*lambda/(1-exp(-lambda))
s2=(1-p)/(1-exp(-lambda))*(lambda+lambda^2)-((1-p)/(1-exp(-lambda))*lambda)^2
tab = rbind(lambda,mu,s2)
colnames(tab) = lambda
rownames(tab) = c('\u03bb', 'E(Y)', 'Var(Y)')
round(tab, 2)

# Gráficas

col = c('black','lightblue')

# L = 2 | p = 0.2

L = 2
p = 0.2
x = 0:8
pois = dpois(x, lambda = L)
p0 = p
py = (1-p)*dpois(x[x>0], lambda = L)/(1-dpois(0, lambda = L))
zap = c(p0,py)
barplot(rbind(pois,zap), beside = TRUE, col = col,names.arg = x,
main = 'lambda = 2, pi = 0.2')
legend(x = "topright", legend = c('Usual','Apartado'), lty = 1, col = col,
lwd = 2)

# L = 5 | p = 0.2

L = 5
p = 0.2
x = 0:14
pois = dpois(x, lambda = L)
p0 = p
py = (1-p)*dpois(x[x>0], lambda = L)/(1-dpois(0, lambda = L))
zap = c(p0,py)
barplot(rbind(pois,zap), beside = TRUE, col = col,names.arg = x,
```

```
main = 'lambda = 5, pi = 0.2')
legend(x = "topright", legend = c('Usual','Apartado'), lty = 1, col = col,
lwd = 2)

# L = 8 | p = 0.2

L = 8
p = 0.2
x = 0:20
pois = dpois(x, lambda = L)
p0 = p
py = (1-p)*dpois(x[x>0], lambda = L)/(1-dpois(0, lambda = L))
zap = c(p0,py)
barplot(rbind(pois,zap), beside = TRUE, col = col,names.arg = x,
main = 'lambda = 8, pi = 0.2')
legend(x = "topright", legend = c('Usual','Apartado'), lty = 1, col = col,
lwd = 2)

# L = 2 | p = 0.6

L = 2
p = 0.6
x = 0:8
pois = dpois(x, lambda = L)
p0 = p
py = (1-p)*dpois(x[x>0], lambda = L)/(1-dpois(0, lambda = L))
zap = c(p0,py)
barplot(rbind(pois,zap), beside = TRUE, col = col,names.arg = x,
main = 'lambda = 2, pi = 0.6')
legend(x = "topright", legend = c('Usual','Apartado'), lty = 1, col = col,
lwd = 2)

# L = 5 | p = 0.6

L = 5
p = 0.6
x = 0:14
pois = dpois(x, lambda = L)
```

```
p0 = p
py = (1-p)*dpois(x[x>0], lambda = L)/(1-dpois(0, lambda = L))
zap = c(p0,py)
barplot(rbind(pois,zap), beside = TRUE, col = col,names.arg = x,
main = 'lambda = 5, pi = 0.6')
legend(x = "topright", legend = c('Usual','Apartado'), lty = 1, col = col,
lwd = 2)

# L = 8 | p = 0.6

L = 8
p = 0.6
x = 0:20
pois = dpois(x, lambda = L)
p0 = p
py = (1-p)*dpois(x[x>0], lambda = L)/(1-dpois(0, lambda = L))
zap = c(p0,py)
barplot(rbind(pois,zap), beside = TRUE, col = col,names.arg = x,
main = 'lambda = 8, pi = 0.6')
legend(x = "topright", legend = c('Usual','Apartado'), lty = 1, col = col,
lwd = 2)


# Simulación Poisson vs Poisson con ceros apartados

# lambda = 1

library(pscl)
set.seed(27032023)
M = 5
n = 100
p = 0.3
sdp = numeric(M)
sdi = numeric(M)

for (j in 1:M){
y = numeric(n)
b = rbinom(n, size = 1, prob = p)
```

```
for (i in 1:n){
yi = 0
while (yi==0){
yi = rpois(1,lambda = 1)
}
y[i] = yi
}
y = (1-b)*y
mp = glm(y~1, family = poisson(link = 'log'))
mi = hurdle(y~1 | 1)
rp = residuals(mp, type = 'pearson')
ri = residuals(mi, type = 'pearson')
sdp[j] = sd(rp)
sdi[j] = sd(ri)
}

tab = rbind(sdp,sdi)
colnames(tab) = paste('Muestra',1:M)
rownames(tab) = c('Usual','Apartado')
round(tab,3)

# lambda = 8

set.seed(27032023)
M = 5
n = 100
p = 0.3
sdp = numeric(M)
sdi = numeric(M)

for (j in 1:M){
y = numeric(n)
b = rbinom(n, size = 1, prob = p)
for (i in 1:n){
yi = 0
while (yi==0){
yi = rpois(1,lambda = 8)
}
```

```
y[i] = yi
}
y = (1-b)*y
mp = glm(y~1, family = poisson(link = 'log'))
mi = hurdle(y~1 | 1)
rp = residuals(mp, type = 'pearson')
ri = residuals(mi, type = 'pearson')
sdp[j] = sd(rp)
sdi[j] = sd(ri)
}

tab = rbind(sdp,sdi)
colnames(tab) = paste('Muestra',1:M)
rownames(tab) = c('Usual','Apartado')
round(tab,3)
```

I.4.2. Modelos con ceros apartados (ejemplo de los fumadores)

```
rm(list=ls(all=TRUE))

library(openintro)

data(smoking)
smoking = data.frame(smoking)
smoking[is.na(smoking)] = 0   # Los NA son los ceros de los no fumadores
summary(smoking)

barplot(table(smoking$amt_weekends), xlab = 'amt_weekends',
ylab = 'Count', col = 'gold')

# Modelo de Poisson con ceros apartados

library(pscl)

m = hurdle(amt_weekends~gender + marital_status| gender + marital_status,
data = smoking, dist = 'poisson', link = 'logit')
summary(m)
```

```
X = model.matrix(m, model = 'count')
Z = model.matrix(m, model = 'zero')
beta = as.matrix(m$coef$count)
alfa = as.matrix(-m$coef$zero)
p = exp(Z%*%alfa)/(1+exp(Z%*%alfa))

# Predicción

x.new = data.frame(
gender = c(rep('Female',5),rep('Male',5)),
marital_status = rep(levels(smoking$marital_status),2)
)
tab = rbind(rep(NA,5),rep(NA,5))
tab[1,] = predict(m, newdata = x.new[x.new$gender=='Female',],
type = 'response')
tab[2,] = predict(m, newdata = x.new[x.new$gender=='Male',],
type = 'response')
colnames(tab) = levels(smoking$marital_status)
rownames(tab) = levels(smoking$gender)
round(tab, 3)

# Estimación de p

tabp = rbind(rep(NA,5),rep(NA,5))
tabp[1,] = 1-predict(m, newdata = x.new[x.new$gender=='Female',],
type = 'zero')
tabp[2,] = 1-predict(m, newdata = x.new[x.new$gender=='Male',],
type = 'zero')
colnames(tabp) = levels(smoking$marital_status)
rownames(tabp) = levels(smoking$gender)
round(tabp, 3)

# Inferencia

confint(m)
col = c(rep('red',6),rep('lightblue',6))
names = c('b0','b12','b22','b23','b24','b25',
'a0','a12','a22','a23','a24','a25')
```

```
boxplot(t(confint(m)), col = col, names = names)

# Diagnosis

r1 = residuals(m, type = 'response')
plot(r1~m$fit, data = smoking, main = 'Residuos brutos',
xlab = 'Cigarrillos esperados', ylab = 'Residuos')
abline(h=0, lty = 2)

r2 = residuals(m, type = 'pearson')
plot(r2~m$fit, data = smoking, main = 'Residuos de Pearson',
xlab = 'Cigarrillos esperados', ylab = 'Residuos')
abline(h=0, lty = 2)

sd(r2[fitted(m)<=4])
sd(r2[fitted(m)>4])



# Modelo Binomial Negativo con ceros apartados

m_nb = hurdle(amt_weekends~gender + marital_status| gender + marital_status,
data = smoking, dist = 'negbin', link = 'logit')
summary(m_nb)

r1_nb = residuals(m_nb, type = 'response')
plot(r1_nb~m_nb$fit, data = smoking, main = 'Residuos brutos',
xlab = 'Cigarrillos esperados', ylab = 'Residuos')
abline(h=0, lty = 2)
r2_nb = residuals(m_nb, type = 'pearson')
plot(r2_nb~m_nb$fit, data = smoking, main = 'Residuos de Pearson',
xlab = 'Cigarrillos esperados', ylab = 'Residuos')
abline(h=0, lty = 2)

sd(r2_nb[fitted(m_nb)<=4])
sd(r2_nb[fitted(m_nb)>4])

# Test de verosimilitud sobre ambos modelos
```

```
library(lmtest)
lmtest::lrtest(m,m_nb)
```

Bibliografía

[1] Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R* (Vol. 574, p. 574). New York: springer.

[2] R Core Team (2023). _*R: A Language and Environment for Statistical Computing*_. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

[3] Venables, W. N. & Ripley, B. D. (2002) *Modern Applied Statistics with S. Fourth Edition.* Springer, New York. ISBN 0-387-95457-0

[4] Thomas W. Yee (2024). *VGAM: Vector Generalized Linear and Additive Models.* R package version 1.1-10. URL https://CRAN.R-project.org/package=VGAM

[5] Simon Jackman (2020). *pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory.* United States Studies Centre, University of Sydney. Sydney, New South Wales, Australia. R package version 1.5.5.1. URL https://github.com/atahk/pscl/

[6] Achim Zeileis, Christian Kleiber, Simon Jackman (2008). *Regression Models for Count Data in R.* Journal of Statistical Software 27(8). URL http://www.jstatsoft.org/v27/i08/.

[7] Angelo Canty and Brian Ripley (2022). *boot: Bootstrap R (S-Plus) Functions.* R package version 1.3-28.1.

[8] Davison, A. C. & Hinkley, D. V. (1997) *Bootstrap Methods and Their Applications.* Cambridge University Press, Cambridge. ISBN 0-521-57391-2

[9] Achim Zeileis, Torsten Hothorn (2002). Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. URL https://CRAN.R-project.org/doc/Rnews/

[10] Çetinkaya-Rundel M, Diez D, Bray A, Kim A, Baumer B, Ismay C, Paterno N, Barr C (2022). _openintro: Data Sets and Supplemental Functions from 'OpenIntro' Textbooks and Labs_. R package version 2.4.0, <https://CRAN.R-project.org/package=openintro>.

[11] NYC Data. (2022). Taxi trip data NYC [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/3782900

[12] https://www.kaggle.com/datasets/stealthtechnologies/regression-dataset-for-household-income-analysis

[13] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.