Contajes (Sesión 1)

Modelos Lineales Generalizados

Grado de Estadística

16/11/2018

Índice

Clasificación

Explicative Variables	Response Variable				
	Dicothomic or Binary	*Polythomic*	*Counts (discrete)*	*Continuous*	
				Normal	*Time between events*
Dicothomic	Contingency tables Logistic regression Log-linear models	Contingency tables Log-linear models	Log-linear models	Tests for 2 subpopulation means: t.test	Survival Analysis
Polythomic	Contingency tables Logistic regression Log-linear models	Contingency tables Log-linear models	Log-linear models	ONEWAY, ANOVA	Survival Analysis
Continuous (covariates)	Logistic regression	*	Log-linear models	Multiple regression	Survival Analysis
Factors and covariates	Logistic regression	*	Log-linear models	Covariance Analysis	Survival Analysis
Random Effects	Mixed models	Mixed models	Mixed models	Mixed models	Mixed models

Introducción

- En este bloque se va a abordar la problemática de los contajes:
  - **Con variable respuesta**: modelos log-lineales (poisson, binomial negativa,...)
  - **Relaciones entre variables**: modelos multinomiales (tablas de contingencia)
- Las observaciones:
  - Serán **no negativas**
  - En general, serán **no acotadas superiormente**
  - Tomarán el **valor cero en un porcentaje no despreciable** (excepto en los modelos cero truncados)

Ejemplos

- Número de denuncias diarias por robos en una determinada región.
- Horas de estudio semanales por parte de los estudiantes
  - Nota: esta variable se podría considerar como continua si se recoge con suficiente precisión
- Número de daños en una flota de barcos a lo largo de los años
  - Nota: se tiene 1 fuente adicional de variabilidad: el tipo de barco
- Número de artículos científicos publicados por estudiantes de doctorado

Modelos log-lineales

- Los modelos log-lineales poseen 2 características:
  - La dependencia de la esperanza de la respuesta (μ_i) condicionada a un vector de covariables y/o factores (x_i) es multiplicativa y generalmente se fórmula en forma logarítmica:

$$log(\mu_i) = x_i^T \beta \qquad i = 1, \ldots, n$$

  - Se asume que la varianza condicional de la respuesta es proporcional a su esperanza y constante en todo el conjunto de datos:

$$V(Y_i|X_i) = \phi \cdot \mu_i$$

Infradispersión	Equidispersión	Sobredispersión
$\phi<1$	$\phi=1$	$\phi>1$

- Se verá la relación entre los modelos log-lineales y multinomiales.

Repaso de distribuciones. Binomial

- **Definición**: Número de éxitos en la repetición de n pruebas de Bernoulli independientes con probabilidad constante p
- **Notación**: $X \sim B(n, p)$
- **Parámetros**: n (núm. de repeticiones), p (probabilidad de éxito)

- **Función de probabilidad**:

$$P(X = K) = \binom{n}{k} \cdot p^k \cdot q^{n-k} \ [q=1-p]$$

- **Indicadores**:
  - $E(X) = n \cdot p$
  - $V(X) = n \cdot p \cdot q$
- **R**: *dbinom, pbinom, qbinom*

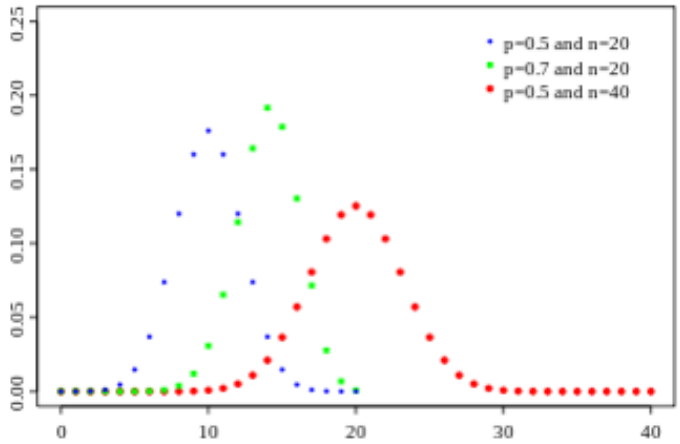


Figure 1: Binomial distribution

Repaso de distribuciones. Binomial negativa

- **Definición**: Número de repeticiones k de un experimento de Bernoulli hasta obtener r éxitos
- **Notación**: $X \sim BN(r, p)$
- **Parámetros**: r (núm. de éxitos), p (probabilidad de éxito)

- **Función de probabilidad**:

$$P(X = K) = \binom{k-1}{r-1} \cdot p^r \cdot q^{k-r} \quad k = r, \cdots, n$$

- **Indicadores**:
  - $E(X) = \frac{r \cdot q}{p}$
  - $V(X) = \frac{r \cdot q}{p^2}$
- **R**: *dnbinom, pnbinom, qnbinom*

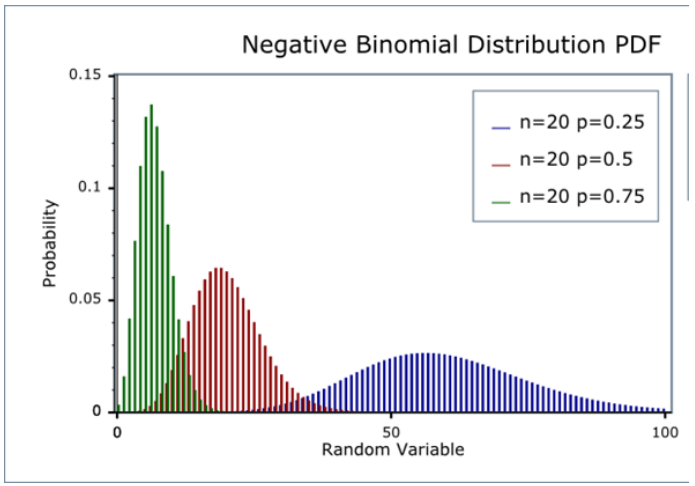


Figure 2: Negative Binomial distribution

Repaso de distribuciones. Multinomial

- **Definición**: Vector con la frecuencia de aparición de cada clase en n repeticiones de una experiencia aleatoria con probabilidades constantes para cada clase $p : p_1, p_2, \cdots, p_k$
- **Notación**: $X \sim MN(n, p_1, \ldots, p_k)$
- **Parámetros**: n (núm. de repeticiones), p_i (probabilidad de la clase i-ésima)

- **Función de probabilidad**:

$$P(X = (x_1, \ldots, x_k)) = \frac{n!}{x_1! \cdots x_k!} \cdot p_1^{x_1} \cdots p_k^{x_k}$$

- **Indicadores**:
  - $E(X_i) = n \cdot p_i$
  - $V(X_i) = n \cdot p_i \cdot q_i$
- **R**: *dmultinom*

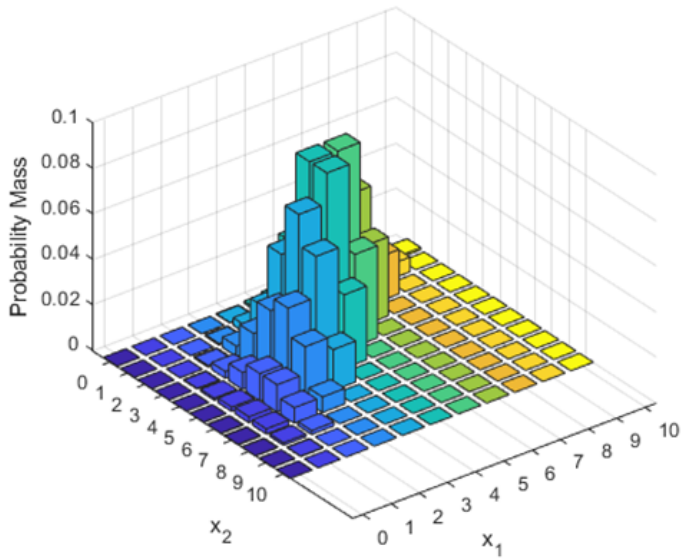


Figure 3: Multinomial distribution

Repaso de distribuciones. Poisson

- **Definición**: Número de eventos en un determinado intervalo de tiempo y/o espacio
- **Notación**: $X \sim P(\lambda)$
- **Parámetros**: λ (tasa de aparición) con $\lambda > 0$

- **Función de probabilidad**:

$$P(X = K) = \frac{e^{-\lambda} \cdot \lambda^k}{k!} \text{ con } k = 0, 1, \cdots$$

- **Indicadores**:
  - $E(X) = \lambda$
  - $V(X) = \lambda$
- **R**: *dpois, ppois, qpois*

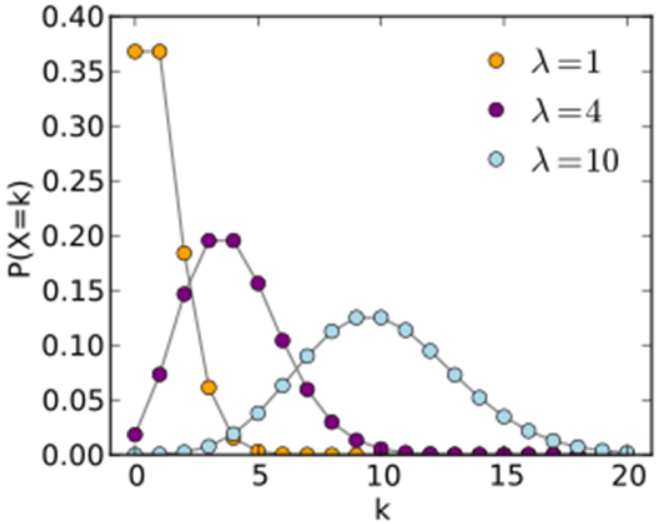


Figure 4: Poisson distribution

Repaso de distribuciones. Gamma

- **Definición**: Sirve para modelar el tiempo de la k-ésima llegada a un servicio con tiempo entre llegadas distribuido exponencialmente
- **Notación**: $X \sim G(\alpha, \beta)$
- **Parámetros**: α (parámetro de forma), β (parámetro de escala)

- **Función de probabilidad**:

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot x^{\alpha-1} \cdot e^{-\beta x}$$

- **Indicadores**:
  - $E(X) = \frac{\alpha}{\beta}$
  - $V(X) = \frac{\alpha}{\beta^2}$
- **R**: *dgamma, pgamma, qgamma*

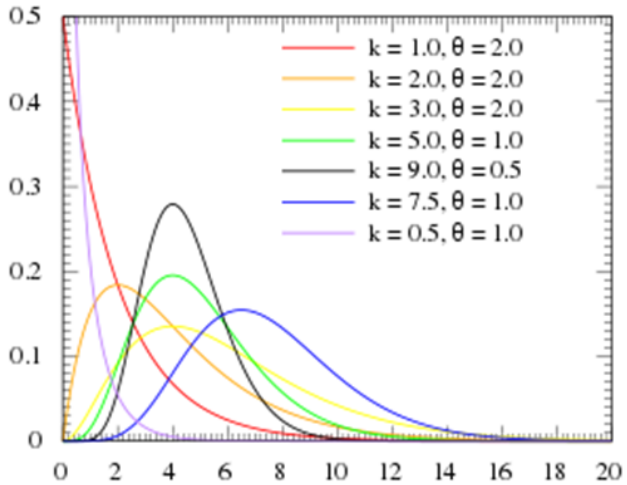


Figure 5: Gamma distribution

Modelo de Poisson. Introducción

- La distribución Poisson viene determinada completamente por su media. Esto es, se pueden calcular todas las probabilidades y los momentos de cualquier orden.
- En el modelo de respuesta de Poisson se modelará el logaritmo de la esperanza en función de un conjunto de variables explicativas o predictoras.

Modelo de Poisson. Ejemplo introductorio

Desperfectos	Casas (Obs.)	%	Pr. Poisson	Casas (Esp.)	Chi-square
0	18	0.041	0.047	20.89	0.4
1	53	0.120	0.145	63.65	1.78
2	103	0.234	0.220	97.00	0.37
3	107	0.243	0.224	98.54	0.73
4	82	0.186	0.171	75.08	0.64
5	46	0.105	0.104	45.77	0
6	18	0.041	0.053	23.25	1.18
7	10	0.023	0.023	10.12	0
8	2	0.005	0.009	3.86	0.89
9	1	0.002	0.003	1.31	0.07

- Número de desperfectos en una revisión rutinaria de viviendas:

$$\hat{\lambda} = \frac{no.desperfectos}{no.casas} = \frac{0\cdot 18+1\cdot 53+\cdots+9\cdot 1}{18+\cdots+1} = \frac{1341}{440} = 3.05$$

- El estadístico de Pearson permite comparar los desperfectos observados con los esperados según una Poisson.
- En este caso el ajuste es bueno ya que el p-valor es inferior a 0.05:

$$\chi^2 = 0.4 + \cdots + 0.07 = 6.07 \; ; \; df = 9 - 1 = 8 \rightarrow pvalue = 0.63$$

Modelo de Poisson. Componentes

- En el modelo de Poisson, la respuesta es función del parámetro μ que representa la esperanza para dicha distribución.
- **Componente aleatoria**: vector aleatorio Y de n componentes estadísticamente independientes y distribuidas poissonianamente con esperanza:

$$\mu^T = (\mu_1, \mu_2, \cdots, \mu_n)$$

- **Componente sistemática**: especificación de un vector η (predictor lineal) a partir de un número reducido de parámetros a estimar y un conjunto de variables explicativas.
  - Parámetros: $\beta^T = (\beta_1, \beta_2, \cdots, \beta_p)$
  - Predictores: $X^T = (X_1, X_2, \cdots, X_n)$
- **Relación entre ambas**: El vector μ se relaciona habitualmente con el predictor lineal η a través de la función de link logarítmica.

$$log(\mu_i) = \eta = x_i^T \beta \quad i = 1, \cdots, n$$

Modelo de Poisson. Familia exponencial

- Ya se había visto que la distribución de Poisson pertenecía a la familia exponencial.
  - **Familia exponencial**: $f_Y(y, \theta, \phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}$
  - **Poisson**: $f_Y(y, \theta, \phi) = \frac{\mu^y}{y!} \cdot exp\{-\mu\} = exp\left\{\frac{ylog(\mu) - \mu}{1} - log(y!)\right\}$

- $\theta = log(\mu) \rightarrow$ Representa el link canónico
- $a(\phi) = 1 \rightarrow$ El parámetro de sobre-dispersión es 1, lo que implica que:

$$V(Y) = a(\phi) \cdot E[Y] = E[Y]$$

Modelo de Poisson. Estimación de parámetros

- La estimación de los parámetros del modelo se realiza por máxima verosimilitud.
- Los estimadores máximo-verosímiles son asintóticamente consistentes y normales.
- La matriz de varianzas y covarianzas asintótica es $\hat{\phi} \cdot I_{\beta}^{-1}$ donde I_{β} es la matriz de información de Fisher.

Modelo de Poisson. Devianza

- La devianza sin término constante en el modelo viene dada por la siguiente expresión:

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

- La devianza con término constante en el modelo viene dada por la siguiente expresión (el término constante "absorbe" las diferencias entre lo observado y lo estimado):

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

- En el caso de μ grande, la devianza puede aproximarse por el estadístico de Pearson:

$$D(y,\hat{\mu}) = \chi^2 = 2\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{\hat{\mu}_i}$$

Modelo de Poisson. Validación (I)

- **Opción 1**: mediante el **estadístico de Pearson**:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \sim \chi^2_{J-p-1}$$

[J es el número de intervalos para calcular el estadístico y p es el número de parámetros]

- **Opción 2**: mediante el **test de la devianza** que también sigue una distribución χ^2 bajo la hipótesis nula

$$L^2 = 2 \sum_{cells} \left\{ y_i log\left(\frac{y_i}{\hat{\mu}_i}\right) \right\} \sim \chi^2_{J-p-1}$$

- Es preferible usar este último test al estadístico de Pearson (sobre todo si los efectivos esperados por celda son inferiores a 5). Ambas pruebas comparan nuestro modelo con el modelo saturado que explica toda la variabilidad de los datos
- Ambas pruebas deberían coincidir razonablemente en casos de tamaños muestrales grandes, ya que son asintóticamente equivalentes

Modelo de Poisson. Residuos. Validación (II)

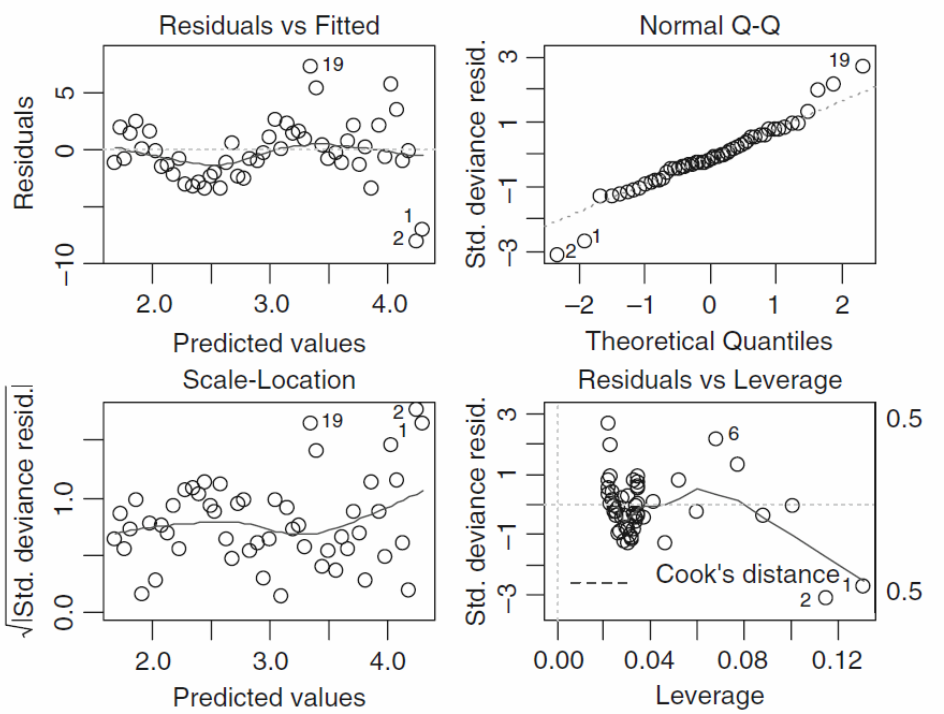
- Existen distintos tipos de residuos:
  - **Crudos**: $y_i - \hat{\mu}_i$
  - **Pearson**: $\frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$.Si se tiene sobredispersión, el denominador será $\sqrt{\phi \hat{\mu}_i}$
  - **Pearson estandarizados**: $\frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i \cdot (1 - \frac{\hat{\mu}_i}{n})}}$
  - **Devianza**: $sign(y_i - \mu_i)\sqrt{(d_i)}$ donde d_i es la contribución de la observación i-ésima a la devianza
- Estos últimos (*Pearson estandarizados* o *Devianza*) son los preferibles a la hora de hacer un análisis formal. En al caso de modelos no poissonianos, los de la devianza son preferibles.
- Tipos de gráficos de residuos (no se deben observar patrones):
  - Residuos vs. Valores predichos. Si heteroscedasticidad, ¿Sobredispersión?
  - Residuos vs. Variables explicativas del modelo. ¿Se deben transformar?
  - Residuos vs. Variables no incluidas en el modelo. ¿Se deben incluir?
  - Residuos vs. Eje temporal. ¿Se cumple la premisa de independencia?

Modelo de Poisson. Residuos. Validación (II)

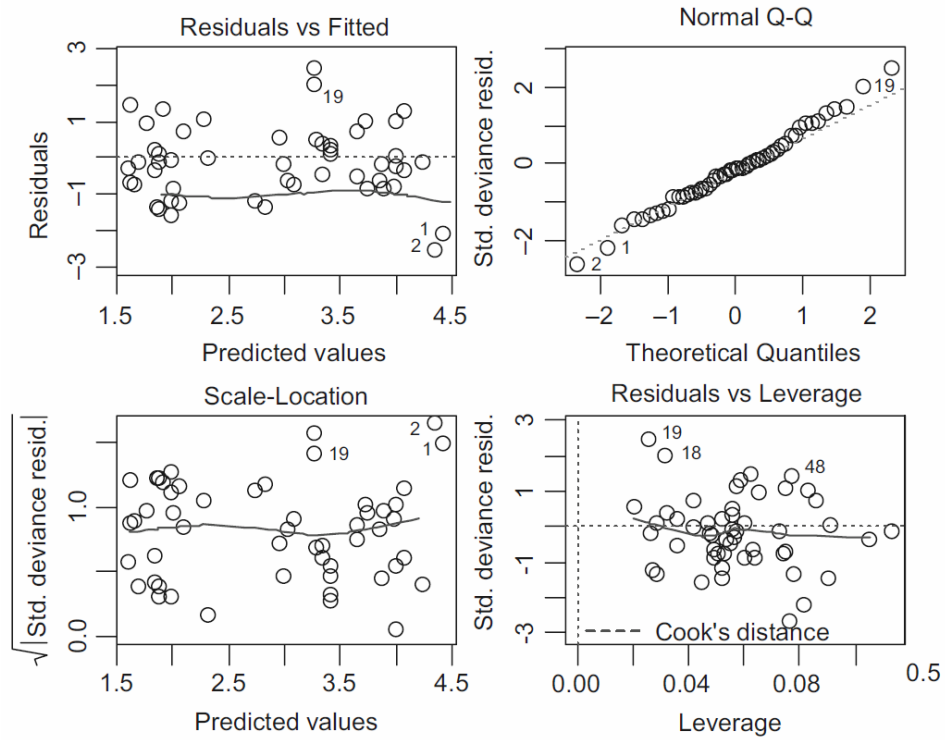
- Como siempre, los objetivos del análisis de los residuos son:
  - Descubrir patrones. P.ej, términos cuadráticos.
  - Identificar outliers. P.ej, distancias de Cook muy grandes.
  - Identificar "errores" en la especificación del modelo. P.ej., sobredispersión.
- Si los residuos no tienen ningún patrón, por lo general, validan el modelo.

Modelo de Poisson. Residuos. Validación (III)

Not validated model



Validated model



Modelo de Poisson. Interpretación (contraste basal)

- **Término constante**
  - En el modelo nulo, la exponencial del término se interpreta como la **media global**.
  - En un modelo con diversas variables explicativas, la exponencial del término se interpreta como la **media en el grupo con todas las categorías de referencia** y valores nulos en las covariables.
- **Coeficientes en un factor**
  - La exponencial del coeficiente se interpreta como el **incremento (o decremento) multiplicativo en la media** de la respuesta de la categoría pertinente respecto a la categoría basal.
- **Coeficientes en una variable continua**
  - La exponencial del coeficiente se interpreta como el **incremento (o decremento) multiplicativo en la media** de la respuesta por cada incremento unitario de la covariable.

Modelo de Poisson. Offset

- En ocasiones, es más relevante modelar las tasas en lugar de los contajes
- Esto ocurre, básicamente, en dos situaciones:
  - Cuando el tiempo de seguimiento para realizar el contaje es distinto
    - Ejemplo: número de brotes de psoriasis en un conjunto de pacientes con distinto tiempo de seguimiento
  - Cuando se disponen de datos agregados y los grupos están desbalanceados
    - Ejemplo: número de casos de legionela por regiones
- En ambos casos, el **offset** se introduce en el modelo habitualmente en forma logarítmica del tiempo de seguimiento (t) o tamaño del grupo (n) para poder obtener la tasa (λ)

$$log(y_i) = log(t_i) + x_i^T \beta \rightarrow log\left(\frac{y_i}{t_i}\right) = log(\lambda_i) = x_i^T \beta$$

Modelo de Poisson. R

- Para modelar una respuesta possioniana se usará la función *glm* con el parámetro *family=poisson*:

```
mod = glm(response ~ pred1 + . + predk, family=poisson)*
mod = glm(response~offset(log(time))+ pred1 + ... + predk, family=poisson)*
```

- La función *summary* nos da más información del modelo:

```
summary(mod)
```

- Las funciones *deviance* y *residuals* nos permiten obtener la devianza y los residuos de Pearson, respectivamente

```
deviance(mod)
residuals(mod,"pearson")
```

Modelo de Poisson. Sobredispersión. Problemática

- En ocasiones, la premisa de que la Esperanza condicionada es igual a la varianza condicionada ($\phi = 1$) puede no ser cierta
- En la mayoría de casos, el parámetro de dispersión puede ser mayor (más habitual) o menor (menos habitual) que 1.
- Los modelos de Poisson, a la práctica, son poco frecuentes porque existen fuentes adicionales de variabilidad que provocan sobredispersión. Por ejemplo:
  - El intervalo de tiempo es variable para cada observación (distinto tiempo de seguimiento). Podría solucionarse con el offset.
  - Los datos pueden ser producidos por un proceso de Poisson agrupado (cluster) donde cada evento contribuye una cantidad aleatoria al total (número de eventos de una enfermedad en una comunidad de animales que se agrupan en manadas)
  - En estudios de comportamiento y en estudios de propensión a accidentes donde hay variabilidad entre sujetos, el número de incidentes Y para un individuo determinado podría ser Poisson con una media de Z.

Modelo de Poisson. Sobredispersión. Estimación (I)

- Una alternativa para estimar el parámetro de sobredispersión es el estadístico de Pearson dividido por los grados de libertad:

$$\hat{\phi} = \frac{\chi^2}{n-p} = \frac{\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}}{n-p} \sim \chi^2_{J-p-1}$$

- Un valor del estadístico de Pearson χ^2 elevado puede indicar:
  - Una falta de ajuste del modelo (***sobredispersión aparente***). Se debe a covariables importantes o a interacciones ausentes, valores atípicos, efectos no lineales no considerados en la parte sistemática del modelo y/o a la elección de la función link incorrecta.
  - Un efecto de sobredispersión (***sobredispersión real***). La varianza de los datos es mayor que la media por un mayor número de ceros, por observaciones agrupadas o correlacionadas.

Modelo de Poisson. Sobredispersión. Estimación (II)

- Otra alternativa para estimar la sobredispersión es la propuesta por Cameron et al. (1985).
- Sea $\sigma^2 = V[Y|X]$ y $\mu = E[Y|X]$, se plantea el siguiente contraste de hipótesis

$$\begin{cases} H_0 : \sigma^2 = \mu \\ H_1 : \sigma^2 = \mu + \alpha \cdot g(\mu) \end{cases}$$

- $g(\mu)$ es cualquier función de μ, generalmente $g(\mu) = \mu$ o $g(\mu) = \mu^2$. Entonces el test es equivalente a:

$$\begin{cases} H_0 : \alpha = 0 \\ H_1 : \alpha > 0 \end{cases}$$

- La función *dispersiontest* del paquete *AER* realiza este test:

```
dispersiontest(model,trafo=1)   # Testea relacion lineal
dispersiontest(model,trafo=2)   # Testea relacion cuadratica
```

Modelo de Poisson. Sobredispersión. Soluciones.

- Usar los estimadores puntuales de la regresión de Poisson pero empleando el método **sandwich** para estimar los errores estándar (no se verá)
- Usar un modelo de **quasi-poisson** donde los coeficientes se estiman usando la distribución de Poisson pero los errores estándar se ajustan para $\phi \neq 1$
- Usar una distribución **binomial negativa** donde la varianza depende cuadráticamente de la media
- Si el problema es un exceso de ceros, se pueden usar modelos **cero-inflados**.
- Usar los modelos *Hurdle* (no los veremos)

[Nota: Los 2 primeros métodos no usan una estimación máximo verosímil plena en el sentido que combinan 2 métodos distintos para las estimaciones de las βs y de ϕ. Esto implica que no tiene sentido calcular su AIC]

Quasipoisson. Estimación

- Se realiza una estimación previa el parámetro de sobre-dispersión ϕ:

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

- Posteriormente, se estiman los errores estándar usando este parámetro de sobredispersión
  - Las estimaciones puntuales de los coeficientes no cambiarán respecto a la Poisson
  - Los errores estándar se verán inflados de forma constante por $\sqrt{\phi}$
- Con R se usará la función *glm* con el parámetro *family=quasipoisson*:

```
mod = glm(response ~ pred1 + . + predk, family=quasipoisson)
```

Quasipoisson. Comparación de modelos

- Al no ser estrictamente un modelo probabilístico no se puede calcular el AIC (ni el BIC) y, por tanto, no se pueden comparar modelos usando este estadístico.
- Sin embargo, se puede usar el test de la devianza modificado para comparar modelos quasipoisson anidados:

$$\frac{D_2 - D_1}{\phi(p_1 - p2)} \sim F_{p_1 - p_2, n - p_1}$$

- Para los modelos M_1 y M_2, D_1 y D_2 son sus devianzas residuales; y p_1 y p_2 son los grados de libertad. ϕ es el parámetro de sobredispersión estimado con el modelo con más parámetros.
- La función *anova* en R compara estos modelos pudiendo especificar el parámetro de sobredispersión:

```
anova(mod1,mod2,test='F')                    # Parametro de sobredispersion estimado
anova(mod1,mod2,test='F', dispersion=2)      # Parametro de sobredispersion igual a 2
```

- No es posible comparar modelos poisson con quassipoison con ningún test estadístico. Se usará la sobredispersión.

Binomial negativa. Introducción

- Otra opción para tratar la sobredispersión es partir de una distribución de Poisson suponiendo un efecto aleatorio multiplicativo Z para representar la heterogeneidad **no observable**. La distribución de la respuesta condicionada a un conjunto de variables no observables W es una Poisson de tasa μZ:

$$Y|W \sim P(\mu Z)$$

- Al no conocer las variables no observables, se debe hacer alguna premisa sobre Z:

$$Z \sim Gamma\left(\alpha = \tfrac{1}{\sigma^2}, \beta = \tfrac{1}{\sigma^2}\right) \rightarrow E(Z) = \tfrac{\alpha}{\beta} = 1 \text{ y } V(Z) = \tfrac{\alpha}{\beta^2} = \sigma^2$$

- Con esta premisa, se puede calcular la distribución incondicional de la respuesta que resulta ser una binomial negativa cuya función de probabilidad se puede describir mediante los parámetros de la distribución Gamma:

$$P(Y = y) = \tfrac{\Gamma(\alpha+y)}{y!\Gamma(\alpha)} \cdot \tfrac{\beta^\alpha \mu^y}{(\mu+\beta)^{\alpha+y}}$$

$$\text{si } \alpha = \beta = \tfrac{1}{\sigma^2} \rightarrow \begin{cases} E(Y) = \mu \\ V(Y) = \mu \cdot (1 + \sigma^2 \mu) = \mu + \tfrac{\mu^2}{\theta} \end{cases}$$

Binomial negativa. Introducción

- Si $\sigma^2 = 0$, no existe heterogeneidad y obtendríamos una distribución de Poisson
- Si $\sigma^2 > 0$, existe sobredispersión y tenemos una distribución Binomial negativa
- Dado que el modelo de Poisson es un caso particular de la Binomial Negativa, podemos comparar las verosimilitudes de ambos modelos con un test formal:
  - Sin embargo, como el modelo de Poisson se encuentra en el valor frontera, el ratio de verosimilitudes no converge a una distribución χ^2 con 1 grado de libertad
  - En consecuencia, este test es conservativo a la hora de ser rechazado.

Binomial negativa. R

- Un primer paso es estimar el parámetro θ con la función *glm.nb {package: MASS}*

```
mod1 <- glm.nb(response ~ pred1 + ... + predk)
Theta <- mod1$theta
```

- Esta estimación se usará para formular el modelo completo con *glm* y el parámetro *family=negative.binomial(Theta)*

```
mod2 <- glm(response ~ pred1 + ... + predk, family=negative.binomial(Theta))
```

- Nota: las estimaciones puntuales de ambas funciones (*glm* y *glm.nb*) son las mismas

Binomial negativa. Consideraciones

- No se recomienda aplicar modelos con binomial negativa a muestras pequeñas.
- Una causa común de la dispersión excesiva es el exceso de ceros mediante un proceso de generación de datos adicional. En esta situación, debe considerarse el modelo de cero-inflado.
- Si el proceso de generación de datos no permite ningún valor de 0 (como el número de días de permanencia en el hospital), un modelo cero-truncado será más apropiado.
- Los contajes a menudo tienen una variable de exposición que indica la cantidad de veces que pudo haber ocurrido el evento. Esta variable debe incorporarse en el modelo como *offset*.

Modelos cero-inflados

- Otro problema común a los modelos de contaje se da cuando los datos empíricos muestran más ceros de los que deberian según ciertos modelos (p.ej., el de Poisson y Binomial negativa)
- Ejemplos:
  - Número de asesinatos cometidos por una persona en un año.
  - Número de conflictos bélicos donde ha intervenido un país durante un año.
- En general, se produce con eventos que son poco frecuentes para algunas observaciones y bastante frecuentes para otras. No obstante, es complicado prever de antemano si hemos de necesitar este modelo.

Modelos cero-inflados

- El modelo **Poisson cero inflado** (*ZIP*) postula que hay dos clases latentes de observaciones: un proporción π de ceros estructurales y una proporción (1-π) de población que sigue una Poisson:
  - $P(Y_i = 0) = \pi + (1 - \pi) \cdot e^{-\lambda}$
  - $P(Y_i = k) = (1 - \pi) \cdot \frac{\lambda^k e^{-\lambda}}{k!}$ para $k \geq 1$
- De la misma forma en el caso de la **Binomial negativa cero inflada** (*ZINB*):
  - $P(Y_i = 0) = \pi + (1 - \pi) \cdot \left(\frac{r}{\mu_i + r}\right)^r$ (donde r es uno de los parámetros de la BN)
  - $P(Y_i = k) = (1 - \pi) \cdot f_{NB}(y)$ para $k \geq 1$
- El modelo combina un modelo **logit** que predice a cuál de las dos clases latentes pertenece una observación con un modelo de **Poisson** o **Binomial negativa** que predice el resultado para aquellos en la segunda clase latente.
- Ignorar la excesiva presencia de ceros puede proporcionar resultados sesgados y/o sobredispersión.

Modelos cero-inflados

- En estos modelos hay dos tipos de ceros: estructurales y aleatorios
- Ejemplos:
  - En una muestra de personas escogidas al azar se podría preguntar el número de veces que les ha tocado la lotería a lo largo de su vida (aún siendo un premio menor). Los ceros estructurales serían aquella parte de la población que nunca juega, mientras que el resto (que también pueden ser 0's, en este caso aleatorios) serían personas que han jugado a la lotería alguna vez.
  - Otro ejemplo claro se podría ver con el número de cigarros fumados en un intervalo de tiempo. Los no fumadores conllevarían ceros estructurales.
- Se podría tener un tercer tipo de ceros de *diseño* no deseados: p.ej. búsqueda de nidos de un determinado tipo de ave fuera de su hábitat. Estos ceros no son deseables.

Modelos cero-inflados. Momentos

- En los modelos cero inflados cambian la esperanza y la varianza.
- Para el modelo de **Poisson cero inflado**:
  - Poisson: $E(Y_i|X_i) = \mu_i \rightarrow ZIP : E(Y_i|X_i) = \mu_i \cdot (1 - \pi_i)$
  - Poisson: $V(Y_i|X_i) = \mu_i \rightarrow ZIP : V(Y_i|X_i) = (1 - \pi_i) \cdot (\mu_i + \pi_i \cdot \mu_i^2)$
- Para el modelo de **Binomial negativa cero inflado**:
  - NB: $E(Y_i|X_i) = \mu_i \rightarrow ZINB : E(Y_i|X_i) = \mu_i \cdot (1 - \pi_i)$
  - NB: $V(Y_i|X_i) = \mu_i + \frac{\mu_i^2}{\theta} \rightarrow ZINB : V(Y_i|X_i) = (1 - \pi_i) \cdot (\mu_i + \frac{\mu_i^2}{\theta}) + \mu_i^2 \cdot (\pi_i^2 + \pi_i)$

Modelos cero-inflados. Estimadores de μ y π

- De forma no ajustada (cruda) se puede estimar la media de la distribución de Poisson y la probabilidad de cero en modelos cero inflados.
- Por el método de los **momentos**:

$$\hat{\mu}_{mo} = \frac{s^2+m^2}{m} - 1$$

$$\hat{\pi}_{mo} = \frac{s^2-m}{s^2+m^2-m}$$

- Por el método de **máxima verosimilitud** (para la μ se necesita resolver una ecuación numéricamente)

$$m \cdot (1 - e^{-\hat{\mu}_{ml}}) = \hat{\mu}_{ml} \left(1 - \frac{n_0}{n}\right)$$

$$\hat{\pi}_{ml} = 1 - \frac{\bar{x}}{\hat{\mu}_{ml}}$$

[m y s^2 son la media y varianza muestrales y $\frac{n_0}{n}$ es la proporción observada de ceros]

Modelos cero-inflados. R

- La función *zeroinfl {package: pscl}* es una opción para estimar este modelo un modelo Poisson cero-inflado

```
mod.poisson.zinfl <- zinfl(response ~ pred1 + ... + predk)
summary(mod.poisson.zinfl)
```

- La misma función serviría para ajustar un modelo cero-inflado de binomial negativa

```
mod.nb.zinfl <- zinfl(response ~ pred1 + ... + predk, dist = 'negbin')
summary(mod.nb.zinfl)
```

Modelos cero-truncados. Introducción

- Cero-truncado significa que la variable de respuesta no puede tener el valor 0. Ejemplo:
  - Días de ingreso de los pacientes en el hospital.
  - Minutos en que una ballena está en la superficie antes de volver a sumergirse
  - Rayas an las aletas de peces (identificación de poblaciones)
  - Edad de un animal (años o meses)
- Los datos cero-truncados no son necesariamente un problema.
- Sí que puede causar un problema el ajuste por modelos de Poisson o Binomial negativa, ya que estas distribuciones permiten ceros dentro de su rango de valores posibles. Con estos modelos, si la media es pequeña y no existen ceros, las estimaciones por GzLM pueden estar sesgadas.

Modelos cero-truncados. Introducción

- La clave es corregir la función de probabilidad de una Poisson estandar:

$$P(X = K) = \frac{e^{-\lambda} \cdot \lambda^k}{k!} \text{ con } k = 0, 1, \cdots$$

- Una Poisson sin ceros, se estandariza la probabilidad de los valores estrictamente positivos:

$$P(X = 0) = e^{-\lambda} \rightarrow P(X = K|K > 0) = \frac{e^{-\lambda} \cdot \lambda^k}{(1-e^{-\lambda})k!}$$

- Para la Binomial Negativa se podría realizar un proceso similar.
- El ajuste de un modelo sin considerar la NO presencia de ceros da lugar a estimaciones de los coeficientes del modelo sesgadas y errores estándar infraestimados.

Modelos cero-truncados. R

- La libreria *VGAM* contiene la función *vglm* para ajustar modelos cero-truncados

```
library(VGAM)
vglm(formula, family=pospoisson, data)       # poisson cero-truncada
vglm(formula, family=posnegbinomial, data)   # binomial negativa cero-truncada
```

Ejemplo. Artículos

- Variables
  - **art**: articles in last three years of Ph.D.
  - **fem**: coded one for females
  - **mar**: coded one if married
  - **kid5**: number of children under age six
  - **phd**: prestige of Ph.D. program
  - **ment**: articles by mentor in last three years
- Sample

art	fem	mar	kid5	phd	ment
0	Men	Married	0	2.52	7
0	Women	Single	0	2.05	6
0	Women	Single	0	3.75	6
0	Men	Married	1	1.18	3

Ejemplo. Artículos. Modelo Poisson (I)

```r
mp <- glm(art ~ fem + mar + kid5 + phd + ment, family=poisson, data=ab)
summary(mp)
```

```
##
## Call:
## glm(formula = art ~ fem + mar + kid5 + phd + ment, family = poisson,
##     data = ab)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5672  -1.5398  -0.3660   0.5722   5.4467
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.304617   0.102981   2.958   0.0031 **
## femWomen    -0.224594   0.054613  -4.112 3.92e-05 ***
## marMarried   0.155243   0.061374   2.529   0.0114 *
## kid5        -0.184883   0.040127  -4.607 4.08e-06 ***
## phd          0.012823   0.026397   0.486   0.6271
## ment         0.025543   0.002006  12.733  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1817.4  on 914  degrees of freedom
## Residual deviance: 1634.4  on 909  degrees of freedom
## AIC: 3314.1
##
## Number of Fisher Scoring iterations: 5
```

Ejemplo. Artículos. Modelo Poisson (II)

- El modelo es deficiente ya que la devianza residual (1634.4) es netamente superior al punto crítico ($\chi^2_{909} = 980.3$):

```
qchisq(0.95, df.residual(mp))
```

```
## [1] 980.2518
```

Ejemplo. Artículos. Sobredispersión (I)

```
library(AER)
dispersiontest(mp,trafo=1)
```

```
##
##  Overdispersion test
##
## data:  mp
## z = 5.7825, p-value = 3.681e-09
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##     alpha
## 0.8245398
```

```
dispersiontest(mp,trafo=2)
```

```
##
##  Overdispersion test
##
## data:  mp
## z = 6.5297, p-value = 3.295e-11
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##     alpha
## 0.5091216
```

Ejemplo. Artículos. Sobredispersión (II)

- Tanto el test para comprobar si la varianza depende linealmente de la esperanza (trafo=1) como para si depende cuadráticamente (trafo=2) dan significativos:
  - $trafo = 1 \rightarrow V[Y_i|X_i] = (1 + 0.82) \cdot E[Y_i|X_i]$
  - $trafo = 2 \rightarrow V[Y_i|X_i] = E[Y_i|X_i] + 0.51 \cdot E[Y_i|X_i]^2$

Ejemplo. Artículos. Quasi-Poisson (I)

```r
summary(mq <- glm(art~fem+mar+kid5+phd+ment, family=quasipoisson, data=ab))
```

```
##
## Call:
## glm(formula = art ~ fem + mar + kid5 + phd + ment, family = quasipoisson,
##     data = ab)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5672  -1.5398  -0.3660   0.5722   5.4467
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.304617   0.139273   2.187 0.028983 *
## femWomen    -0.224594   0.073860  -3.041 0.002427 **
## marMarried   0.155243   0.083003   1.870 0.061759 .
## kid5        -0.184883   0.054268  -3.407 0.000686 ***
## phd          0.012823   0.035700   0.359 0.719544
## ment         0.025543   0.002713   9.415  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 1.829006)
##
##     Null deviance: 1817.4  on 914  degrees of freedom
## Residual deviance: 1634.4  on 909  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

Ejemplo. Artículos. Quasi-Poisson (I)

- El AIC no puede calcularse al no existir un modelo probabilístico completo detrás.
- Estimación de ϕ. Esta estimación es la que utiliza el modelo quasi-poisson para recalcular los errores estándar:

```
pr <- residuals(mp,"pearson")
phi <- sum(pr^2)/df.residual(mp)
round(c(phi,sqrt(phi)),4)
```

```
## [1] 1.8290 1.3524
```

- Por tanto, los errores estándar quedarán "inflados" por 1.35

Ejemplo. Artículos. Quasi-Poisson vs. Poisson (I)

```
se <- function(model) sqrt(diag(vcov(model)))
kable(round(data.frame(p=coef(mp),  q=coef(mq),
                       se.p=se(mp), se.q=se(mq),
                       ratio=se(mq)/se(mp)), 4))
```

	p	q	se.p	se.q	ratio
(Intercept)	0.3046	0.3046	0.1030	0.1393	1.3524
femWomen	-0.2246	-0.2246	0.0546	0.0739	1.3524
marMarried	0.1552	0.1552	0.0614	0.0830	1.3524
kid5	-0.1849	-0.1849	0.0401	0.0543	1.3524
phd	0.0128	0.0128	0.0264	0.0357	1.3524
ment	0.0255	0.0255	0.0020	0.0027	1.3524

- Los coeficientes del modelo Poisson y Quasi-poisson son los mismos y los errores estándar están multiplicados por 1.35 que se obtiene de la raíz del estadístico de Pearson dividido por sus grados de libertad.

Ejemplo. Artículos. Binomial negativa (I)

```
summary(mnb <- glm.nb(art ~ fem + mar + kid5 + phd + ment, data=ab))
```

```
##
## Call:
## glm.nb(formula = art ~ fem + mar + kid5 + phd + ment, data = ab,
##     init.theta = 2.264387695, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1678  -1.3617  -0.2806   0.4476   3.4524
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.256144   0.137348   1.865 0.062191 .
## femWomen    -0.216418   0.072636  -2.979 0.002887 **
## marMarried   0.150489   0.082097   1.833 0.066791 .
## kid5        -0.176415   0.052813  -3.340 0.000837 ***
## phd          0.015271   0.035873   0.426 0.670326
## ment         0.029082   0.003214   9.048  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(2.2644) family taken to be 1)
##
##     Null deviance: 1109.0  on 914  degrees of freedom
## Residual deviance: 1004.3  on 909  degrees of freedom
## AIC: 3135.9
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  2.264
##           Std. Err.:  0.271
##
##  2 x log-likelihood:  -3121.917
```

Ejemplo. Artículos. Binomial negativa (II)

- Estimación de θ y aplicación en *glm*

```
1/mnb$theta
```

```
## [1] 0.4416205
```

```
summary(mnbg <- glm(art ~ fem + mar + kid5 + phd + ment,family=negative.binomial(mnb$theta), data=ab))
```

```
##
## Call:
## glm(formula = art ~ fem + mar + kid5 + phd + ment, family = negative.binomial(mnb$theta),
##     data = ab)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1678  -1.3617  -0.2806   0.4476   3.4524
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.256149   0.140008   1.830  0.06765 .
## femWomen    -0.216420   0.074043  -2.923  0.00355 **
## marMarried   0.150490   0.083686   1.798  0.07247 .
## kid5        -0.176416   0.053836  -3.277  0.00109 **
## phd          0.015271   0.036568   0.418  0.67633
## ment         0.029082   0.003277   8.876  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(2.2644) family taken to be 1.03911)
##
##     Null deviance: 1109.0  on 914  degrees of freedom
## Residual deviance: 1004.3  on 909  degrees of freedom
## AIC: 3133.9
##
```

Ejemplo. Artículos. Binomial negativa (III)

- Goodness of fit

```
deviance(mnbg)
```

```
## [1] 1004.281
```

```
qchisq(0.95,909)
```

```
## [1] 980.2518
```

- El modelo no ajusta bien los datos ya que la devianza residual se encuentra en la región de no aceptación de la H_0

Ejemplo. Artículos. Poisson zero inflated (I)

- En el modelo de Poisson cero-inflado, tenemos 2 modelos en uno:
  - Un modelo logit (output inferior de *zeroinfl*) para predecir la probabilidad de pertenecer a la clase de ceros estructurales
  - Un modelo de Poisson (output superior de *zeroinfl*) para los ceros aleatorios y el resto de valores

Ejemplo. Artículos. Poisson zero inflated (II). Ejercicio

```
summary(mzip <- zeroinfl(art ~ fem + mar + kid5 + phd + ment, data=ab))
```

```
##
## Call:
## zeroinfl(formula = art ~ fem + mar + kid5 + phd + ment, data = ab)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -2.3253 -0.8652 -0.2826  0.5404  7.2976
##
## Count model coefficients (poisson with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.640838   0.121307   5.283 1.27e-07 ***
## femWomen    -0.209145   0.063405  -3.299 0.000972 ***
## marMarried   0.103751   0.071111   1.459 0.144565
## kid5        -0.143320   0.047429  -3.022 0.002513 **
## phd         -0.006166   0.031008  -0.199 0.842378
## ment         0.018098   0.002294   7.888 3.07e-15 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.577059   0.509386  -1.133  0.25728
## femWomen     0.109746   0.280082   0.392  0.69518
## marMarried  -0.354014   0.317611  -1.115  0.26502
## kid5         0.217097   0.196482   1.105  0.26919
## phd          0.001274   0.145263   0.009  0.99300
## ment        -0.134114   0.045243  -2.964  0.00303 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 21
## Log-likelihood: -1605 on 12 Df
```

1. Cuál la probabilidad de cero estructural en una mujer soltera, sin hijos menores de 6 años en un programa con prestigio 3 y 1 articulo publicado por su mentor?
2. Cuál es la media esperada para esta mujer?
3. Cuánto más probable es que una mujer tenga un exceso de ceros respecto a un hombre?

Ejemplo. Artículos. Poisson zero inflated (II). Solucion

Solución 1:

```
pr1 <- predict(mzip,newdata = data.frame(fem='Women',mar='Single',kid5=0,phd=3,ment=1),type='zero')     # With function
co <- coef(mzip)                                                                                        # By hand
pl <- co[7] + co[8] + 3*co[11] + 1*co[12]
pr2 <- exp(pl)/(1+exp(pl))
round(as.numeric(c(pr1,pr2)),3)
```

```
## [1] 0.355 0.355
```

Solución 2:

```
mu1 <- predict(mzip,newdata = data.frame(fem='Women',mar='Single',kid5=0,phd=3,ment=1),type='response') # With function
pl0 <- co[1] + co[2] + 3*co[5] + 1*co[6]                                                                # By hand
mu2_0 <- exp(pl0)
mu2 <- (1-pr2)*mu2_0
round(as.numeric(c(mu1,mu2)),3)
```

```
## [1] 0.993 0.993
```

Solución 3:

```
round(as.numeric(exp(co[8])),3)   # Odds ratio
```

```
## [1] 1.116
```

References

1. Cameron et al. Regression-based tests for overdispersion in the Poisson model. Journal of Econometrics (1985)
2. Zuur, AF, Ieno, EN, Walker NJ et al. Mixed Effects Models and Extensions in Ecology with R. Springer (2009)