Una vez repasados los aspectos importantes del **modelo de regresión logística binomial**, vamos a dar un paso más y estudiar cómo se debe proceder cuando la variable dependiente es cualitativa, pero **el número de categorías (K) es superior a 2**. Esto es equivalente a asumir que la variable dependiente tiene una **distribución multinomial** cuyo primer parámetro es igual a 1 (pues cada individuo realiza una única elección). Adicionalmente,

- K debe ser finito y
- las categorías deben ser mutuamente excluyentes.

FUNDAMENTOS DEL MODELO DE REGRESIÓN LOGÍSTICA MULTINOMIAL

Siguiendo con el GLM, queremos modelizar la **esperanza de la distribución de Y** condicionada a los valores de las variables predictoras. En este caso, **la esperanza es un vector** cuyos componentes coinciden con las probabilidades de cada categoría.

Por ello, este modelo consta de varias ecuaciones (que se obtienen partiendo de las vistas en el caso binario), la última de las cuales se obtiene teniendo en cuenta que las **probabilidades deben sumar uno**:

$$P(Y = j|X_1, \ldots, X_p) = \frac{e^{(\beta_{0,j}+\beta_{1,j}X_1+\cdots+\beta_{p,j}X_p)}}{1 + \sum_{i=2}^{K} e^{(\beta_{0,i}+\beta_{1,i}X_1+\cdots+\beta_{p,i}X_p)}}, \quad \forall j = 2, \cdots, K$$

$$P(Y = 1|X_1, \ldots, X_p) = \frac{1}{1 + \sum_{i=2}^{K} e^{(\beta_{0,i}+\beta_{1,i}X_1+\cdots+\beta_{p,i}X_p)}}$$

Como se puede observar, esta formulación del modelo **proporciona $K-1$ conjuntos de parámetros β** (para todas las categorías excepto la primera).

Siguiendo con el concepto de odds-ratio, podemos plantear cocientes de las probabilides del modelo obteniendo que el modelo de regresión logística multinomial también se puede definir como:

$$\frac{P(Y=j|X_1,\dots,X_p)}{P(Y=1|X_1,\dots,X_p)} = e^{(\beta_{0,j}+\beta_{1,j}X_1+\cdots+\beta_{p,j}X_p)}, \quad \forall j=2,\cdots,K,$$

lo que permitirá una **mejor interpretación de los parámetros** (y, por ende, del efecto de las variables explicativas sobre la dependiente).

Nótese que los parámetros $\{\beta_{i,j}\}_{i=1,\dots,p}$ van asociados a la **comparación de la probabilidad** de la categoría j con la probabilidad de la categoría 1, que recibe el nombre de categoría de referencia.

ESTIMACIÓN DE LOS PARÁMETROS

Como ocurre con la regr. logística binaria, la estimación de los parámetros se basa en el **método de máxima verosimilitud**. Por tanto, la función de verosimilitud a maximizar viene dada por:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_{1i}^{y_{1i}} p_{2i}^{y_{2i}} \cdots p_{Ki}^{y_{Ki}}$$

con $p_{ji} = P(Y = j | x_{1i}, \ldots, x_{pi}),\ \forall i = 1, \ldots, n;\ j = 1, \ldots, K$ y

$$y_{ji} = \begin{cases} 1 & \text{si } y_i = j \\ 0 & \text{en otro caso} \end{cases},\ \forall i = 1, \ldots, n;\ j = 1, \ldots, K$$

De nuevo, no existe una fórmula explicita para la obtención de los parámetros, por lo que será necesario recurrir a **métodos iterativos de optimización**, los cuales serán mas lentos que el caso binario debido al aumento sustancial de parámetros a estimar.

Una vez obtenidas las probabilidades de cada categoría, la estrategia de clasificación consiste en asignar las observaciones a la categoría con mayor probabilidad predicha.

INTERPRETACIÓN DE LOS PARÁMETROS

Como en el caso binomial, la interpretación de los parámetros se basa en los ***odds ratio***:

- Si el parámetro está asociado a una **variable cuantitativa**, $e^{\beta_{i,j}}$ se interpreta como el efecto multiplicativo que tiene un **incremento unitario** de la variable i en las posibilidades de que se dé la categoría j frente a la de referencia (la denotada como 1).
- Si el parámetro está asociado a un nivel de una variable cualitativa, $e^{\beta_{i,j}}$ se interpreta como el aumento/dismunición que se observa en las posibilidades de que se dé la categoría j frente a la de referencia (la denotada como 1) cuando la **variable explicativa analizada pasa de su nivel de referencia al** i.

- Tanto el contraste de los parámetros, como el análisis de tipo II, se pueden aplicar al caso multinomial para evaluar la **significatividad de los parámetros/variables** (con las mismas distribuciones subyacentes). Lo mismo ocurre con el **índice Kappa**.
- Los métodos de **selección de variables** son los mismos que en el caso binario pero adaptando el cálculo de la verosimilitud para la obtención de los estadísticos AIC/BIC.
- Los **métodos de remuestreo**, como la partición entrenamiento/prueba y la validación cruzada repetida, también son aplicables (y de nuevo son deseables para obtener medidas de evaluación más fiables).

EVALUACIÓN DEL MODELO DE REGRESIÓN LOGÍSTICA MULTINOMIAL

Como en el caso binomial, a la hora de evaluar un modelo de regresión logística multinomial **existen distintas medidas** que se pueden aplicar:

- **Matriz de confusión**: En este caso, la matriz de confusión deberá contar con tantas filas y columnas como categorías tenga la variable dependiente y, de nuevo, sus celdas n_{ij} permitirán mostrar **cuántas observaciones pertenecen a la categoría i pero han sido predichas como j**, por lo que todos los elementos que se encuentren en la diagonal principal serán aciertos y las celdas restantes corresponderán con fallos.

  Es importante recordar que, de cara a la clasificación, en el caso multinomial las observaciones son asignadas a aquella categoría con **mayor probabilidad predicha** (lo que puede resultar no ser la mejor estrategia posible en el caso de variables desbalanceadas pero no existen alternativas extendidas en la comunidad científica).

  Por otro lado, dado que no existen "eventos", los conceptos de **sensibilidad y especificidad** han de ser adaptados y pasan a calcularse **de manera individual** para cada categoría, representando la capacidad que tiene el modelo de capturar la categoría en estudio y la ausencia de la misma:

$$acc = \frac{\sum_{i=1}^{k} n_{ii}}{n} \qquad sens_i = \frac{n_{ii}}{n_{i.}} \qquad especif_i = \frac{n - \sum_{\substack{j=1 \\ j \neq i}}^{k} n_{ji}}{n - n_{i.}}$$

EVALUACIÓN DEL MODELO DE REGRESIÓN LOGÍSTICA MULTINOMIAL

- **Curva ROC/AUC**: Teniendo en cuenta que la curva ROC se basa en la sensibilidad y la especificidad no resultará extraño que **esta herramienta deba adaptarse** para el caso multinomial. En este caso, existen dos alternativas diferentes:
  - ***One vs. all***: esta alternativa consiste en obtener **tantas curvas ROC como categorías** tenga la variable dependiente para lo cuál se utilizan las probabilidades predichas agregadas en dos únicos niveles: categoría sí/categoría no, lo que permite trabajar de la manera "clásica". **Para cada una de estas curvas se obtiene el AUC y se calcula la media** como medida de evaluación general (de nuevo cuánto más próximo esté a uno, mejor será el modelo). Así mismo, se pueden representar gráficamente las K curvas, lo que permite determinar cómo de bien/mal detecta el modelo cada categoría.
  - ***Pairwise***: esta alternativa consiste en obtener **tantas curvas ROC como pares de categorías** existan, lo que permite hacerse una idea de la capacidad que tiene el modelo de **distinguir entre las categorías del par**. En este caso, la representación gráfica resulta compleja pero el AUC es relativamente sencillo de obtener a partir de ciertas propiedades estadísticas (en las que no vamos a adentrarnos) por lo que, de nuevo, se puede obtener la media de los valores obtenidos y utilizarlo como medida de evaluación general.

El último tipo de modelo de regresión logística que vamos a estudiar es el denominado **ordinal** (u ordenado). Este tipo de modelo se aplica a variables cualitativas cuyos **niveles pueden ordenarse de manera natural** (por ejemplo, el acuerdo/desacuerdo sobre alguna afirmación en una escala 1-5). De nuevo, como en el caso multinomial, el número de categorías K debe ser finito y las categorías deben ser mutuamente excluyentes.

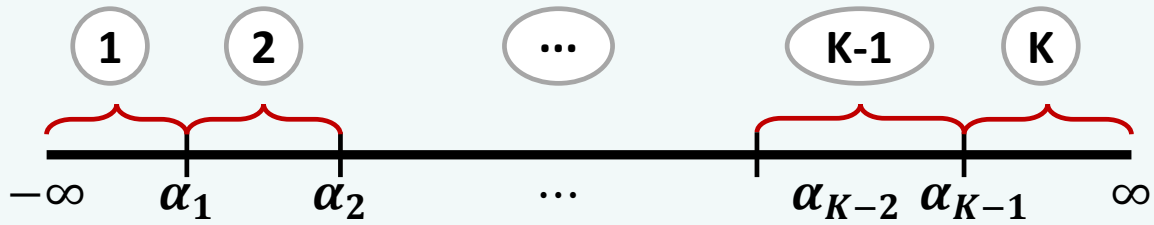
Este tipo de variables pueden modelizarse a partir de los modelos de **regresión logística multinomial** ya estudiados pero eso implicaría ignorar la componente ordinal de los datos, lo que en ocasiones no es recomendable.

Así mismo, también sería posible asignar un valor numérico a cada una de las categorías de manera creciente siguiendo el orden natural y ajustar un modelo de **regresión lineal**. No obstante, esta estrategia tiene dos inconvenientes: 1) la predicción puede dar lugar a **valores con decimales**, por lo que habría que decidir cómo obtener valores enteros; y 2) hay que tener cuidado con la **distancia numérica entre categorías** de manera que represente de manera adecuada las diferencias entre las mismas.

FUNDAMENTOS DEL MODELO DE REGRESIÓN LOGÍSTICA ORDINAL

En este caso, y basándonos en la idea de recurrir a valores numéricos para la predicción, lo que se hace es asumir que los "decisores" disponen de una **variable latente cuantitativa** (que es desconocida pero se puede estimar) a partir de la cual se selecciona el valor de la variable dependiente, a partir de una discretización de la latente.

Esta variable latente se construye a partir de una **combinación lineal** de las m características de los decisores (con la previa transformación de las variables cualitativas a *dummies*):
$\beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$



Lo que matemáticamente equivale a que **la variable dependiente Y** viene dada por:

$$Y = j, \quad \text{si } \alpha_{j-1} < \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \leq \alpha_j, \quad j = 1, \ldots, K,$$

con $\alpha_0 = -\infty$ y $\alpha_K = \infty$.

Teniendo en cuenta los fundamentos anteriores, bajo el modelo ordinal, la **probabilidad de cada alternativa** se puede obtener, usando la función logística, como:

$$
\begin{aligned}
p_{ji} &= P(Y = j | x_{1i}, \ldots, x_{pi}) \\
&= P(\alpha_{j-1} < \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i \leq \alpha_j) \\
&= P(\alpha_{j-1} - (\beta_1 x_{1i} + \cdots + \beta_p x_{pi}) < \epsilon_i \leq \alpha_j - (\beta_1 x_{1i} + \cdots + \beta_p x_{pi})) \\
&= F_\epsilon(\alpha_j - (\beta_1 x_{1i} + \cdots + \beta_p x_{pi})) - F_\epsilon(\alpha_{j-1} - (\beta_1 x_{1i} + \cdots + \beta_p x_{pi})) \\
&= \frac{e^{\alpha_j - (\beta_1 x_{1i} + \cdots + \beta_p x_{pi})}}{1 + e^{\alpha_j - (\beta_1 x_{1i} + \cdots + \beta_p x_{pi})}} - \frac{e^{\alpha_{j-1} - (\beta_1 x_{1i} + \cdots + \beta_p x_{pi})}}{1 + e^{\alpha_{j-1} - (\beta_1 x_{1i} + \cdots + \beta_p x_{pi})}}
\end{aligned}
$$

- Debido a las buenas propiedades que ofrece de cara a la interpretación de los parámetros, se asume la **función de distribución logística** como $F(\cdot)$ pero se podría aplicar cualquier otra función de enlace con un **soporte en el intervalo** $(0, 1)$, como la función de distribución normal, que daría lugar al modelo de regresión *probit* ordinal.
- Tal y como ocurre con el modelo multinomial, una vez obtenidas las probabilidades, las observaciones se clasifican en la categoría con la **probabilidad predicha máxima** (aunque también se pueden aplicar estrategias alternativas).

ESTIMACIÓN DE LOS PARÁMETROS

Dado que las variables ordinales pueden ser consideradas simplemente nominales, la estimación de los parámetros se vuelve a basar en el **método de máxima verosimilitud** asumiendo que la variable dependiente tiene una **distribución multinomial** cuyo primer parámetro es igual a 1. Por ello, los parámetros $\beta_i, i = 1, \ldots, m$ y $\alpha_i, i = 1, \ldots, K-1$ serán aquellos que **maximicen** la siguiente función de verosimilitud:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_{1i}^{y_{1i}} p_{2i}^{y_{2i}} \cdots p_{Ki}^{y_{Ki}}$$

con $p_{ji} = P(Y = j | x_{1i}, \ldots, x_{pi}),\ \forall i = 1, \ldots, n;\ j = 1, \ldots, K$ calculado según la fórmula de la página anterior y

$$y_{ji} = \begin{cases} 1 & \text{si } y_i = j \\ 0 & \text{en otro caso} \end{cases},\ \forall i = 1, \ldots, n;\ j = 1, \ldots, K$$

De nuevo, no existe una fórmula explicita para la obtención de los parámetros, por lo que será necesario recurrir a **métodos iterativos de optimización**.

INTERPRETACIÓN DE LOS PARÁMETROS

Para la interpretación de los parámetros $\beta_i, i = 1, \ldots, p$, de nuevo recurrimos al concepto de ***odds ratio***, aunque en este caso aprovechamos el hecho de que **las categorías están relacionadas por el orden**. Por ello, los *odds* que analizaremos serán los acumulados:

$$\boldsymbol{odds(Y > j|x_{1i}, \ldots, x_{pi})} = \frac{\boldsymbol{P(Y > j|x_{1i}, \ldots, x_{pi})}}{\boldsymbol{P(Y \leq j|x_{1i}, \ldots, x_{pi})}} = \frac{\frac{1}{1+e^{\alpha_j - \beta_1 x_{1i} - \cdots - \beta_p x_{pi}}}}{\frac{e^{\alpha_j - \beta_1 x_{1i} - \cdots - \beta_p x_{pi}}}{1+e^{\alpha_j - \beta_1 x_{1i} - \cdots - \beta_p x_{pi}}}}$$

$$= \frac{1}{e^{\alpha_j - \beta_1 x_{1i} - \cdots - \beta_p x_{pi}}} = e^{-\alpha_j + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}}$$

Así, el *odds ratio acumulado* asociado al **aumento unitario de una de las variables**, como por ejemplo la primera (pero manteniendo el resto invariables), obtenemos:

$$\frac{\boldsymbol{odds(Y > j|(x_{1i}+1), \ldots, x_{pi})}}{\boldsymbol{odds(Y > j|x_{1i}, \ldots, x_{pi})}} = \frac{e^{-\alpha_j + \beta_1 (x_{1i}+1) + \ldots + \beta_p x_{pi}}}{e^{-\alpha_j + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}}} = e^{\beta_1}$$

- Cabe recordar que, si la variable en cuestión es cualitativa, un incremento unitario se traduce en comparar con la categoría de referencia debido a las variables dummy.

INTERPRETACIÓN DE LOS PARÁMETROS

Es importante destacar que **los parámetros $\alpha_i, i = 1, \ldots, K-1$ no tienen interpretación**, pero son necesarios para obtener las probabilidades.

- Tanto el contraste de los parámetros, como el análisis de tipo II, se pueden aplicar al caso ordinal para evaluar la **significatividad de los parámetros/variables** (con las mismas distribuciones subyacentes).
- Los métodos de **selección de variables** son los mismos que en el caso multinomial (pues incluso la verosimilitud se calcula con la misma fórmula por lo que no varía la obtención de los estadísticos AIC/BIC). Los **modelos penalizados** son posibles teóricamente, pero no están implementados en R (ni en Python).
- Los **métodos de remuestreo**, como la partición entrenamiento/prueba y la validación cruzada repetida, también son aplicables y recomendables.
- Nótese que los odds ratio para cualquier categoría j coinciden, es decir, **el efecto de las variables explicativas es el mismo en todas las comparaciones**. Por ello, este tipo de regresión reciba en ocasiones el nombre de *proportional odds logistic regression (polr)*.

  Si está hipótesis no se cumple, los modelos ordinales pueden ofrecer resultados de **mala calidad** y debe optarse por uno multinomial, que es más flexible.

EVALUACIÓN DEL MODELO DE REGRESIÓN LOGÍSTICA ORDINAL

Dado que, en esencia, las variables ordinales son variables nominales, es habitual recurrir a las **medidas de evaluación multinomiales** (como las medidas derivadas de la matriz de confusión, la curva ROC en sus dos versiones o el índice Kappa) para este tipo de modelos.

No obstante, esta estrategia no resulta idónea pues **no tiene en cuenta el carácter ordinal de la variable**. Este hecho se traduce en que, de aplicar dichas medidas, se estaría asumiendo que **todos los errores de predicción tienen la misma importancia** y, en este caso, un modelo funciona peor si clasifica de manera incorrecta las observaciones en categorías con dos niveles de diferencia que otro que lo hace con un único nivel de diferencia.

Por ello, se hace necesario definir **nuevas medidas** que permitan tener en cuenta este hecho:

EVALUACIÓN DEL MODELO DE REGRESIÓN LOGÍSTICA ORDINAL

- **Kappa ponderado**: Este indicador es una modificación del índice Kappa en el que se tienen en cuenta todas las celdas de la matriz de confusión pero ponderándolas para darle **mayor peso a los elementos más próximos a la diagonal principal** (los aciertos). Recordemos que este índice evalúa los aciertos del modelo, pero eliminando aquellos que se pueden producir por azar.

  Como en el índice Kappa clásico, partimos de la **matriz de confusión relativa** en la que las celdas p_{ij} representan la proporción de observaciones que pertenecen a la categoría i pero son clasificados en la j, y $p_{i\cdot}$ y $p_{\cdot i}$ son las proporciones de observaciones en la categoría i real y predicha, respectivamente. Así mismo, se definen los **pesos** w_{ij} para cada una de las celdas de la matriz de confusión, que tomarán valores entre 0 y 1, donde el 1 representa el mayor acierto y el 0, el mayor error.

  El **índice Kappa ponderado** viene dado por:

$$\kappa_p = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K} w_{ij}p_{ij} - \sum_{i=1}^{K}\sum_{j=1}^{K} w_{ij}p_{i\cdot}p_{\cdot j}}{1 - \sum_{i=1}^{K}\sum_{j=1}^{K} w_{ij}p_{i\cdot}p_{\cdot j}}$$

  La tabla de la página 14 también sirve como **guía para la interpretación** de este índice (donde el valor 0 representa que el modelo de predicción no es mejor que la predicción al azar y el 1, que el modelo es perfecto).

Para finalizar este tema, vamos a estudiar cómo construir modelos de predicción para otro tipo especial de datos: los de **conteo**.

DATOS DE CONTEO - DISTRIBUCIÓN DE POISSON

Los datos de conteo, como su propio nombre indica, están asociados a variables que "cuentan" y que, por ende, dan como resultado **variables discretas**.

En este tipo de datos, es muy habitual asumir que la distribución aleatoria subyacente es la ***Poisson***, que tiene las siguientes propiedades:

- Toma **valores enteros**, de cero a infinito.
- El parámetro de la distribución λ coincide con la **esperanza y la varianza** de la misma, por lo que toma valores estrictamente positivos.
- La moda (es decir, el valor más probable) es la parte entera del parámetro λ.
- La **función de probabilidad** es la siguiente:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad i = 0, 1, 2, \ldots$$

Las propiedades anteriores tienen **implicaciones a la hora de modelizar** variables dependientes con dicha distribución, como veremos enseguida.

FUNDAMENTOS DEL MODELO DE CONTEO - *Poisson*

Para poder estudiar cómo construir modelos predictivos para variables dependiente de conteo, empezamos por recordar los **fundamentos del GLM**:

- Se modeliza el **comportamiento esperado** de la Y.
- El efecto de las variables explicativas se puede resumir con una **combinación lineal**.
- $$E[Y|X_1, X_2, \ldots, X_p] = f(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$$

Dado que el parámetro debe tomar **valores positivos**, no es recomendable utilizar la función identidad como enlace, pues puede dar lugar a valores negativos.

Para asegurar la no negatividad, se opta por la **función exponencial**, de forma que:

$$\lambda|X_1, X_2, \ldots, X_p = e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}$$

Nótese que el valor predicho por esta ecuación es el conteo esperado (dado el valor de las variables explicativas), por lo que generalmente **no toma un valor entero**.

Una ventaja de este modelo es que, a partir del parámetro predicho, se puede **calcular cualquier estadístico de interés** para cada observación (como la probabilidad de que tome el valor 0 o que supere determinado valor crítico).

Así, si se desea una predicción "con sentido", tal y como se hace en los modelos de regresión logística vistos previamente, se puede facilitar el **valor con mayor probabilidad** (la moda).

ESTIMACIÓN DE LOS PARÁMETROS

De nuevo, los parámetros se estiman mediante el **método de máxima verosimilitud** asumiendo que la variable dependiente tiene una **distribución *Poisson***. Por ello, los parámetros $\beta_i, i = 1, \ldots, p$ y serán aquellos que **maximicen** la siguiente función de verosimilitud:

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

con $\lambda_i = \lambda | x_{1i}, \ldots, x_{pi}$, $\forall i = 1, \ldots, n$, calculado según la fórmula de la página anterior.

De nuevo, no existe una fórmula explicita para la obtención de los parámetros, por lo que será necesario recurrir a **métodos iterativos de optimización**.

INTERPRETACIÓN DE LOS PARÁMETROS

Para la interpretación de los **parámetros** $\beta_i, i = 1, \ldots, p$, debemos estudiar cómo cambia la variable dependiente con cambios en las variables explicativas.

A diferencia de lo que ocurre con el modelo de regresión lineal, los cambios considerados serán de tipo **multiplicativo**:

- Variable dicotómica:

$$\frac{\lambda|x=1}{\lambda|x=0} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

  Por lo tanto, la exponencial del parámetro representa **cuántas veces aumenta el conteo** si se comparan individuos para los que $x = 1$ frente a $x = 0$.

- Variable categórica: Se realiza igual que el caso anterior pero comparando cada categoría considerada como *dummy* **frente a la de referencia**.

- Variable cuantitativa:

$$\frac{\lambda|x=a+1}{\lambda|x=a} = \frac{e^{\beta_0+\beta_1(a+1)}}{e^{\beta_0+\beta_1 a}} = e^{\beta_1}$$

  De nuevo, la exponencial del parámetro representa cuántas veces aumenta el conteo por un **incremento unitario** en la variable explicativa.

EVALUACIÓN DEL MODELO

- Tanto el contraste de los parámetros, como el análisis de tipo II, se pueden aplicar al modelo de *Poisson* para evaluar la **significatividad de los parámetros/variables** (con las mismas distribuciones subyacentes).
- Los métodos de **selección de variables** son los mismos que se han usado hasta el momento (únicamente teniendo en cuenta el cambio de fórmula de la verosimilitud para la obtención de los estadísticos AIC/BIC).
- Los **métodos de remuestreo**, como la partición entrenamiento/prueba y la validación cruzada repetida, también son aplicables (y de nuevo son deseables para obtener medidas de evaluación más fiables).
- Es importante recordar que los modelos de *Poisson* asumen que **la esperanza y la varianza de la distribución coinciden**. Esta hipótesis se tiene en cuenta a la hora de calcular los errores estandar de los parámetros por lo que una **infraestimación** de estos puede implicar un aumento de la significatividad de los parámetros artificial (encontrando parámetros significativos que no lo son).

  Por tanto, habrá que verificar si se cumple dicha hipótesis antes de utilizar los correspondientes contrastes pues, de no ser así, se deberá recurrir a otro modelo en el que se asuma una **varianza mayor** (se dice que hay sobredispersión).

DATOS DE CONTEO - DISTRIBUCIÓN BINOMIAL NEGATIVA

Como ya se ha visto, la principal limitación de la distribución de *Poisson* es que asume que **esperanza y varianza coinciden**, lo que no es frecuente en datos reales.

Una alternativa a dicha distribución, útil en presencia de **sobredispersión**, es la distribución binomial negativa, que tiene las siguientes propiedades:

- Toma **valores enteros**, de cero a infinito.
- Depende de **dos parámetros**, que en el contexto de los modelos GLM son $NB(\theta, \frac{\theta}{\theta+\lambda})$, lo que implica que:
  - $E(X) = \lambda$.
  - $Var(X) = \lambda + \frac{\lambda^2}{\theta}$.
  - La **función de probabilidad** es la siguiente:

$$P(X = x) = \frac{\lambda^x}{x!} \cdot \frac{\Gamma(\theta + x)}{\Gamma(\theta)\ (\theta + \lambda)^x} \cdot \frac{1}{\left(1 + \frac{\lambda}{\theta}\right)^\theta}, \quad i = 0, 1, 2, \ldots$$

- Por lo tanto, si $\theta \to \infty$, $Poisson(\lambda) \equiv NB(\theta, \frac{\theta}{\theta+\lambda})$.

MODELO DE CONTEO - BINOMIAL NEGATIVA

- Se trata un modelo de tipo GLM, por lo que precisamos de una función de enlace para poder modelizar el **comportamiento esperado de Y**.
- Para asegurar la no negatividad, se opta por la **función exponencial**, de forma que:

$$\lambda|X_1, X_2, \ldots, X_p = e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}$$

- Como en el modelo *Poisson*, una vez estimados los parámetros β_i (y, en este caso, también θ), se puede **calcular cualquier estadístico de interés** para cada observación (como la probabilidad de que tome el valor 0 o que supere determinado valor crítico).
- De hecho, la principal ventaja de este modelo es precisamente la **mejora en la estimación de la varianza**, por lo que no siempre se observan grandes diferencias en las predicciones puntuales, pero sí en otros estadísticos de la distribución.
- Para estimar los parámetros, se vuelve a recurrir al método de máxima verosimilitud:

$$\beta_1, \ldots, \beta_p, \theta = \text{arg máx} \prod_{i=1}^{n} \frac{\lambda_i^{y_i}}{y_i!} \cdot \frac{\Gamma(\theta + y_i)}{\Gamma(\theta)\ (\theta + \lambda_i)^{y_i}} \cdot \frac{1}{\left(1 + \frac{\lambda_i}{\theta}\right)^{\theta}},$$

con $\lambda_i = \lambda|x_{1i}, \ldots, x_{pi},\ \forall i = 1, \ldots, n$.

EVALUACIÓN DEL MODELO

- Tanto el contraste de los parámetros, como el análisis de tipo II, se pueden aplicar al modelo Binomial Negativa para evaluar la **significatividad de los parámetros/variables** (con las mismas distribuciones subyacentes).
- A diferencia del modelo anterior, la sobredispersión ya se tiene en cuenta, por lo que los contrastes sobre los parámetros son siempre válidos.
- Los métodos de **selección de variables** son los mismos que se han usado hasta el momento (únicamente teniendo en cuenta el cambio de fórmula de la verosimilitud para la obtención de los estadísticos AIC/BIC).
- Los **métodos de remuestreo**, como la partición entrenamiento/prueba y la validación cruzada repetida, también son aplicables (y de nuevo son deseables para obtener medidas de evaluación más fiables).
- La **métrica de evaluación** más frecuente en datos de conteo es el R^2, pues se modeliza una variable cuantitativa.
- Para estudiar si es preferible un modelo de *Poisson* o un Binomial Negativo, basta con llevar a cabo un **contraste de razón de verosimilitudes** entre ambos (es el mismo fundamento que en el análisis de tipo II pero, en lugar de fijar en 0 algunos parámetros β_i, se fija $\theta = \infty$) siempre que cuenten con las mismas variables explicativas.

1. Determinar el tipo de la variable objetivo, así como el modelo más apropiado. Verificar que la variable está codificada de la manera requerida por *R*.
2. Realizar una partición entrenamiento-prueba.
3. Si el conjunto de datos no tiene problemas de multicolinealidad y el número de observaciones es superior al de variables:
    1. Crear un modelo con todas las variables input. Estudiar que se cumplen las posibles hipótesis del modelo.
    2. Analizar el aporte de las variables y crear un nuevo modelo más sencillo.
    3. Construir las 6 combinaciones de modelos automáticos de variables y determinar cuáles son diferentes.
    4. Comparar mediante validación cruzada repetida los modelos automáticos y el/los manuales y determinar el mejor.
4. Si el modelo sí sufre de alguno de estos problemas,
    1. Preparar los datos para poder aplicar modelos penalizados.
    2. Construir una gran batería de modelos ENET variando los parámetros α y λ y compararlos con validación cruzada.
    3. Determinar la combinación óptima de parámetros, construir el modelo final.
5. Analizar las variables/coeficientes del modelo.
6. Evaluar el modelo en entrenamiento y prueba y concluir cómo es su calidad.