Estadística Bayesiana

Modelos Lineales Generalizados

Lizbeth Naranjo Albarrán

Facultad de Ciencias, UNAM

December 25, 2020

Índice

Capítulo 1

Modelos lineales generalizados

Los modelos lineales generalizados (MLG) son modelos en los que una variable se explica por medio de la combinación lineal de otras variables. La variable explicada se conoce como "dependiente" o "respuesta", mientras que las variables que explican se conocen como "independientes", "explicativas" o "covariables" (cuando son categóricas se les llama "factores"). Así que los MLG modelan la relación que existe entre la variable respuesta y las explicativas.

1.1 La familia exponencial y los MLG

Sea Y una v.a. con distribución perteneciente a la familia exponencial, entonces su fdp es de la forma

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

donde θ es el parámetro canónico y ϕ es el parámetro de dispersión. Las funciones $a(\theta)$ y $c(y, \phi)$ determinan la fdp o fmp de Y.

En términos de $a(\theta)$ se tiene que $\mathbb{E}(Y) = a'(\theta)$ y $var(Y) = \phi a''(\theta)$, donde $a'(\theta)$ y $a''(\theta)$ denotan la primera y segunda derivadas de $a(\theta)$, respectivamente.

Las distribuciones más importantes de la familia exponencial se resumen en la tabla 1.1.

La función varianza $V(\mu)$ indica la relación entre la media y la varianza. Conforme la media varía, también varía la varianza a través de $V(\mu)$. De esta manera, un modelo que relaciona la media con las variables explicativas, también relaciona la varianza con las variables explicativas.

Para la distribución normal $V(\mu) = 1$, que significa que la varianza no cambia con la media,

es decir, la respuesta es homocedástica. Para la distribución Poisson $V(\mu) = \mu$, así que el cambio en la media impacta directamente en la varianza. Para la distribución Gamma $V(\mu) = \mu^2$ y por tanto la desviación estándar varía directamente con la media.

	fmp $\mathbb{P}(Y = y)$	
Binomial	$\binom{n}{y}\pi^y(1-\pi)^{n-y}$	$y = 0, 1, \ldots, n$, $n \in \mathbb{N}$, $\pi \in (0, 1)$
Poisson	$e^{-\mu}\mu^y/y!$	$y = 0, 1, \ldots, \infty$, $\mu > 0$
Binomial Negativa	$\frac{\Gamma(y+r)}{y!\Gamma(r)}\pi^r(1-\pi)^y$	$y = 0, 1, \ldots, \infty$, $r > 0$, $\pi \in (0, 1)$, $\mu = \frac{r(1-\pi)}{\pi}$, $\kappa = \frac{1}{r}$
	fdp $f_Y(y)$	
Normal	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$	$y \in \mathbb{R}$, $\mu \in \mathbb{R}$, $\sigma > 0$
Gamma	$\frac{y^{\nu-1}}{\Gamma(\nu)}\left(\frac{y\nu}{\mu}\right)^{\nu} e^{-y\nu/\mu}$	$y > 0$, $\mu > 0$, $\nu > 0$
Gaussiana Inversa	$\frac{1}{\sqrt{2\pi y^3}}\frac{1}{\sigma} \exp\left\{-\frac{1}{2y}\left(\frac{y-\mu}{\mu\sigma}\right)^2\right\}$	$y \in \mathbb{R}$, $\mu \in \mathbb{R}$, $\sigma > 0$

Distribución	Binomial	Poisson	Normal	Gamma	Gaussiana Inversa	Binomial Negativa
Notación	$\mathrm{B}(n, \pi)$	$\mathrm{P}(\mu)$	$\mathrm{Normal}(\mu, \sigma^2)$	$\mathrm{G}(\mu, \nu)$	$\mathrm{IG}(\mu, \sigma^2)$	$\mathrm{NB}(\mu, \kappa)$
θ	$\log\left(\frac{\pi}{1-\pi}\right)$	$\log(\mu)$	μ	$-\frac{1}{\mu}$	$-\frac{1}{2\mu^2}$	$\log\left(\frac{\kappa\mu}{1+\kappa\mu}\right)$
$a(\theta)$	$n\log(1+e^\theta)$	e^θ	$\frac{1}{2}\theta^2$	$-\log(-\theta)$	$-\sqrt{-2\theta}$	$-\frac{1}{\kappa}\log(1-\kappa e^\theta)$
ϕ	1	1	σ^2	$\frac{1}{\nu}$	σ^2	1
$\mathbb{E}(y)$	$n\pi$	μ	μ	μ	μ	μ
$V(\mu) = var(y)/\phi$	$n\pi(1-\pi)$	μ	1	μ^2	μ^3	$\mu(1+\kappa\mu)$
$c(y, \phi)$	$\binom{n}{y}$					

Table 1.1: Características de algunas distribuciones univariadas de la familia exponencial.

Los modelos lineales generalizados (MLG) son una extensión de los modelos lineales y se caracterizan por tres componentes:

- *Componenete aleatorio*: Un vector de observaciones $\mathbf{y}$ tiene n componentes, $\mathbf{y} = (y_1, \ldots, y_n)$, que son una realización del v.a. $\mathbf{Y} = (Y_1, \ldots, Y_n)$. Los componentes de $\mathbf{Y}$ son independientes con distribución perteneciente a la familia exponencial y con $\mathbb{E}[\mathbf{Y}] = \boldsymbol{\mu}$.

- *Componenete sistemática*: Un conjunto de covariables $\mathbf{x}_1, \ldots, \mathbf{x}_p$ forma un predictor lineal dado por

$$\boldsymbol{\eta} = \sum_{j=1}^{p} \mathbf{x}_j\beta_j = \mathbf{X}\boldsymbol{\beta}$$

  donde $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})$, x_{ij} es el valor de la j-ésima covariable para la observación i, $\boldsymbol{\eta}$ es un vector $n \times 1$, $\mathbf{X}$ es la matriz diseño $n \times p$, y $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ es un vector $p \times 1$ de parámetros desconocidos.

- *Liga o función de enlace*: La función de enlace relaciona los componentes aletorio y sistemático:

$$\eta_i = g(\mu_i)$$

1.2 Modelo lineal

El modelo de regresión se encarga de explicar el comportamiento de una variable por medio de otras variables. El modelo lineal clásico, o modelo lineal normal, es es el más importante y utilizado de los MLG, y es crucial para para el estudio de los demás modelos pertenecientes a la clase de MLG.

1.3 MLG para respuestas categóricas

Las variables categóricas toman los valores de un número posible de categorías. El ejemplo más simple de una variable categórica es cuando la respuesta es binaria, codificadas como 1 o 0, denotando la ocurrencia o no ocurrencia del evento ("éxito" o "fracaso"). Existen dos tipos de variables categóricas: las variables cuyas categorías tienen un orden natural ("ordinal") y las que no lo tienen ("nominal").

1.3.1 Respuesta binaria

Considere una variable respuesta y binaria, $y = 0$ o $y = 1$. Si π es la probabilidad de que $y = 1$, entonces $y \sim \mathrm{B}(1, \pi)$. El MLG Bernoulli (Binomial con $n = 1$) es

$$Y \sim \mathrm{B}(1, \pi), \qquad g(\pi) = \mathbf{X}\boldsymbol{\beta}$$

La proporción $\pi/(1-\pi)$ se llama *odds* o *momios*, e indica proporcionalmente cuanto más probable es la ocurrencia del evento comparada con la no-ocurrencia.

Regresión logística

En la regresión logística el *logit* (log de los momios) se modela en términos de las variables explicativas:

$$g(\mu) = \log\left(\frac{\pi}{1-\pi}\right) = \mathbf{X}\boldsymbol{\beta} \qquad \Rightarrow \qquad \pi = \frac{e^{\mathbf{X}\boldsymbol{\beta}}}{1 + e^{\mathbf{X}\boldsymbol{\beta}}}$$

La función de enlace logit asegura que las predicciones de π estén en el intervalo $(0, 1)$ para todo $\boldsymbol{\beta}$ y $\mathbf{X}$.

Los parámetros en la regresión logística se interpretan de la siguiente manera.

Considere una variable explicativa continua x y sea $\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x$, equivalentemente $\frac{\pi}{1-\pi} = e^{\alpha+\beta x}$. Suponga que x se incrementa en una unidad, entonces $\frac{\pi}{1-\pi} = e^{\alpha+\beta(x+1)} = e^{\alpha+\beta x}e^{\beta}$, es decir, los momios se multiplicarían por e^{β}. Si β es pequeña entonces $(e^{\beta} - 1) \approx \beta$ y en este caso 100β es el cambio porcentual aproximado en los momios cuando la variable explicativa se incrementa en una unidad. Si $\beta < 0$ el efecto de un incremento en x implicará que decrezcan los momios, pero si $\beta > 0$ entonces los momios se incrementarán. Finalmente, si $\beta = 0$ entonces $e^{\beta} = 1$ y por tanto no existe efecto en los momios.

Considere una variable explicativa categórica x con r niveles. Suponga que el nivel r es el nivel base, entonce el modelo es $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{r-1}x_{r-1}$, donde las x_j's son variables indicadores con $x_j = 1$ si x se encuentra en el nivel j o $x_j = 0$ en otro caso, para $j = 1, \ldots, r-1$. Cuando x está en el nivel r, los momios de ocurrencia son $\frac{\pi}{1-\pi} = e^{\beta_0}$. Cuando x está en el nivel $j \neq r$, los momios son $e^{\beta_0+\beta_j}$. De tal manera que, el efecto de x cuando está en el nivel j, comparado con el nivel r, es multiplicar a los momios por el factor e^{β_j}. Si $\beta_j < 0$ entonces $x = j$ decrementa los momios, en comparación con $x = r$, mientras que si $\beta_j > 0$ implica que $x = j$ incrementa los momios.

En el modelo de regresión con respuestas binarias, las funciones de enlace *probit* y *log-log complementario* se definen como:

$$g(\pi) = \Phi^{-1}(\pi), \qquad g(\pi) = \log\left\{-\log(1-\pi)\right\},$$

respectivamente, donde Φ^{-1} es la inversa de la fda normal estándar. Por tanto,

$$\pi = \Phi(\mathbf{X}\boldsymbol{\beta}), \qquad \pi = 1 - \exp\left(-e^{\mathbf{X}\boldsymbol{\beta}}\right)$$

Modelos Probit

El modelo de regresión *probit* se modela en términos de las variables explicativas:

$$g(\pi) = \Phi^{-1}(\pi),$$

donde Φ^{-1} es la inversa de la fda normal estándar. Por tanto,

$$\pi = \Phi(\mathbf{X}\boldsymbol{\beta})$$

Es común usar **variables latentes** en la definición de los modelos Probit.

Datos binarios agrupados

Cuando todas las variables explicativas son categóricas es posible expresar un conjunto de datos en forma agrupada. Un grupo consiste de todos los casos con los mismos valores de las variables explicativas y puede corresponder a un conjunto de riesgos homogéneo.

En el caso de una respuesta binaria, una vez que los datos están agrupados, la respuesta observada es el número de eventos que ocurren en cada grupo. La notación es la siguiente:

$$
\begin{array}{rcl}
m & = & \text{número de grupos} \\
n_i & = & \text{número de casos en el grupo } i \\
y_i & = & \text{número de eventos ocurridos en el grupo } i \\
\pi_i & = & \text{probabilidad de que el evento ocurra para un caso en el grupo } i \\
n & = & \text{tamaño de la muestra, } n = \sum_{i=1}^{m} n_i
\end{array}
$$

Por lo tanto, y_i es el número de ocurrencias del evento, de un máximo de n_i, donde la probabilidad de ocurrencia del evento es π_i. Así que la respuesta observada tiene una distribución binomial, $y_i \sim \mathrm{B}(n_i, \pi_i)$, donde la probabilidad π_i se modela como una función de las variables explicativas.

Bondad de ajuste para la regresión logística

Existen distintos métodos y estadísticas para probar la bondad de ajuste de los modelos de regresión logística.

Tablas de clasificación. En las tablas de clasificación, se calculan las probabilidades ajustadas $\hat{\pi}_i$ y para cada caso i se obtienen las predicciones (o clasificaciones), "éxito" o "fallo" ("positivo" o "negativo"), dependiendo de si $\hat{\pi}_i$ es mayor o menor que cierto umbral. Con esto se obtiene una tabla de clasificación 2×2 que compara las ocurrencias observadas con las predicciones.

		Observaciones	
		Positivos	Negativos
Modelo	Positivos	VP	FP
(Predicciones)	Negativos	FN	VN

VP = verdaderos positivos
FN = falsos negativos
VN = verdaderos negativos
FP = falsos positivos

La utilidad del modelo se resume usando las siguientes medidas (lo ideal es que ambas sean cercanas a 1):

- *Sensibilidad*: Frecuencia relativa de predecir un evento como positivo cuando el evento observado es positivo. Es la fracción de verdaderos positivos.

- *Especificidad*: Frecuencia relativa de predecir un evento como negativo cuando el evento observado es negativo. Es la fracción de verdaderos negativos.

$$Sensibilidad = \frac{VP}{VP + FN}, \qquad Especificidad = \frac{VN}{VN + FP}$$

Curvas ROC. La curva ROC (*Receiver Operating Characteristic*) grafica la sensibilidad y especificidad para cada umbral. Tradicionalmente, en el eje horizontal se grafica $1 - especificidad$, y en el eje vertical se grafica *sensibilidad*. Con esta orientación de los ejes, un valor del eje x cercano a cero (alta especificidad) generalmente implica un valor del eje y bajo (baja sensibilidad), y viceversa.

Todas las curvas ROC comienzan en el punto $(0, 0)$, terminan en el punto $(1, 1)$ y son monótonas crecientes. Un modelo que predice adecuadamente resulta en una curva ROC que crece rápidamente a 1: cuanto más cercana esté la curva a la parte superior izquierda, mejor serán sus predicciones.

Generalmente se calcula el área debajo de la curva ROC, y ésta es una medidad de la capacidad predictiva del modelo.

1.3.2 Respuesta categórica con más de dos categorías

MLG multivariado

Considere una variable respuesta categórica con r categorías. Para la respuesta se definen $r - 1$ variables respuesta indicadoras y_j, $j = 1, \ldots, r - 1$, con $y_j = 1$ si la respuesta está en el nivel j y $y_j = 0$ si no lo está. De esta manera, la respuesta $\mathbf{y} = (y_1, \ldots, y_{r-1})'$ es multivariada. Los modelos con respuesta nominal y ordinal entran en la clase de los MLG multivariados, usando la familia exponencial multivariada.

Dadas n observaciones independientes de la respuesta $\mathbf{y}$, es de interés obtener el número de veces que la categoría j ocurre, $n_j = \sum_{i=1}^{n} y_{ij}$. La distribución conjunta de $n_1, \ldots, n_r$ es multinomial con fmp

$$\mathbb{P}(n_1, \ldots, n_r) = \frac{n!}{n_1! \cdots n_r!} \pi_1^{n_1} \cdots \pi_r^{n_r}$$

donde π_j es la probabilidad de que la respuesta sea la categoría j, $\sum_{j=1}^{r} \pi_j = 1$ y $n = \sum_{j=1}^{r} n_j$.

Respuesta ordinal

Considere una respuesta ordinal y con r categorías ordenadas. Puede definirse una variable continua y^* y umbrales (puntos de corte) $\theta_0, \ldots, \theta_r$ tal que

$$y = j \qquad \text{si} \qquad \theta_{j-1} \leq y^* < \theta_j, \qquad j = 1, \ldots, r.$$

El modelo para la categoría ordinal y se define en términos de las probabilidades acumulativas:

$$\tau_j = \mathbb{P}(y \leq j) = \mathbb{P}(y^* < \theta_j), \qquad j = 1, \ldots, r.$$

El objetivo es relacionar las probabilidadeds acumulativas τ_j a las variables explicativas. Suponga que $y^* = -\mathbf{x}'\boldsymbol{\beta} + \epsilon$ con $\mathbb{E}[\epsilon] = 0$, implicando que $\mathbb{E}(y^*) = \mathbf{x}'\boldsymbol{\beta}$. Por tanto,

$$\tau_j = \mathbb{P}(\epsilon \leq \theta_j + \mathbf{x}'\boldsymbol{\beta}),$$

en donde la distribución de ϵ determina la forma exacta del modelo.

Modelo logístico acumulado o de momios proporcionales Suponga que ϵ tiene una distribución logística estándar:

$$\mathbb{P}(\epsilon \leq y) = \frac{1}{1 + e^{-x}}.$$

Entonces

$$\begin{aligned} \tau_j &= \mathbb{P}(\epsilon \leq \theta_j + \mathbf{x}'\boldsymbol{\beta}) = \frac{1}{1 + e^{-(\theta_j + \mathbf{x}'\boldsymbol{\beta})}} \\ \log\left(\frac{\tau_j}{1 - \tau_j}\right) &= \theta_j + \mathbf{x}'\boldsymbol{\beta}, \qquad j = 1, \ldots, r-1, \end{aligned}$$

donde los θ_j son términos de intersección que dependen sólo de j, $\mathbf{x}$ no contiene al 1 (no hay otro término de intersección), y los coeficientes $\boldsymbol{\beta}$ no dependen de j.

Modelo log-log complementario acumulado Suponga que a distribución de ϵ es la distribución de valores extremos mínimos, entonces el modelo es:

$$\log\{-\log(1 - \tau_j)\} = \theta_j + \mathbf{x}'\boldsymbol{\beta}, \qquad j = 1, \ldots, r-1.$$

Este modelo se conoce como modelo de riesgos proporcionales discreto (*discrete proportional hazards model*) o modelo de Cox agrupado, y es un modelo que está cercanamente relacionado al modelo de riesgos proporcionales de Cox en análisis de supervivencia.

Modelo probit acumulado Suponga que ϵ tiene una distribución normal, entonces el modelo es:

$$\Phi^{-1}(\tau_j) = \theta_j + \mathbf{x}'\boldsymbol{\beta}, \qquad j = 1, \ldots, r-1.$$

Respuesta nominal

Los modelos para respuestas nominales se conocen como regresión nominal, regresión politómica (*polytomous*), regresión policotómica (*polychotomous*) o regresión multinomial.

Considere una respuesta nominal y con r categorías nominales (no ordenadas). Las probabilidades multinomiales son $\pi_j = \mathbb{P}(y = j)$ con $\sum_{j=1}^{r} \pi_j = 1$. Los momios de y de la categoría j relativos a la categoría base se modelan como:

$$\log\left(\frac{\pi_j}{\pi_r}\right) = \theta_j + \mathbf{x}'\boldsymbol{\beta}_j, \qquad j = 1, \ldots, r-1,$$

donde se considera que el nivel de respuesta base es r. Por tanto,

$$\pi_r = \frac{1}{1 + \sum_{k=1}^{r-1} e^{\theta_k + \mathbf{x}'\boldsymbol{\beta}_k}}, \qquad \pi_j = \pi_r e^{\theta_j + \mathbf{x}'\boldsymbol{\beta}_j}, \qquad j = 1, \ldots, r-1.$$

Por tanto, para cada categoría j existe un modelo para los momios de la respuesta correspondiente a la categoría j relativo al nivel de respuesta base.

1.4 MLG para datos de conteo

En esta sección se estudian los MLG cuando la respuesta es una variable de conteo, por ejemplo, número de muertes, número de reclamaciones o número vehículos asegurados, y se desea explicar ésta en términos de otras variables.

1.4.1 Regresión Poisson

Cuando la variable respuesta es un conteo, frecuentemente se usa la distribución Poisson. En la regresión Poisson, la media μ se explica en términos de las variables explicativas x's, usando la función de enlace apropiada. El modelo de regresión Poisson se define como

$$Y \sim \mathrm{P}(\mu), \qquad g(\mu) = \mathbf{X}\boldsymbol{\beta},$$

donde la función de enlace generalmente es $g(\mu) = \log(\mu)$, aunque también podría usarse la identidad $g(\mu) = \mu$, pero ésta no garantiza que los conteos sean positivos.

Suponga que el modelo tiene una variable explicativa x_1, entonces $\mathbf{x} = (1, x_1)'$, $\boldsymbol{\beta} = (\beta_0, \beta_1)'$, $g(\mu) = \mathbf{x}'\boldsymbol{\beta} = \beta_0 + \beta_1 x_1$. Con la función de enlace log, el valor esperado de y es $\mu = e^{\beta_0 + \beta_1 x_1}$, así que si x_1 se incremente una unidad entonces existe un incremento multiplicativo en la media μ de e^{β_1}, ya que $e^{\beta_0 + \beta_1(x_1+1)} = e^{\beta_0 + \beta_1 x_1} e^{\beta_1}$.

Suponga que se tiene una variable explicativa categórica con r niveles, y que el nivel r es el nivel base, entonces la variable se reemplaza por $r-1$ variables indicadoras $x_1, \ldots, x_{r-1}$ y el modelo es $g(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{r-1} x_{r-1}$. Con la función de enlace log, el valor esperado de y es $\mu = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_{r-1} x_{r-1}}$. Cuando la variable explicativa está en el nivel base r se tiene que $\mu = e^{\beta_0}$, pero cuando está en el nivel j se tiene que $\mu = e^{\beta_0 + \beta_j}$, así que existe un efecto multiplicativo de e^{β_j} comparado con el nivel base. Si $\beta_j = 0$ la respuesta media para la categoría j es la misma que para el nivel base. Si $\beta_j > 0$ el efecto se incrementa ya que $e^{\beta_j} > 1$, pero si $\beta_j < 0$ el efecto sobre la respuesta media en la categoría j es decreciente.

1.4.2 Regresión binomial negativa

La distribución binomial negativa puede utilizarse para datos de conteo, en situaciones en donde la sobredispersión de los datos está explicada por la heterogeneidad de la media sobre la población. El modelo de regresión binomial negativa, usando la función de enlace log, es:

$$Y \sim \mathrm{NB}(\mu, \kappa), \qquad \log(\mu) = \log(n) + \mathbf{X}\boldsymbol{\beta}$$

1.5 MLG para respuestas continuas

Cuando las variables respuesta son continuas, no negativas y sesgadas a la derecha existen dos opciones que pueden utilizarse para modelar:

- Usar una transformación para normalidad, y usar el modelo lineal normal con la respuesta transformada.
- Usar MLG con una distribución para la variable respuesta que esté definida en los reales no negativos, por ejemplo la distribución gamma o Gaussiana inversa.

1.5.1 Regresión gamma

El MLG gamma es de la forma

$$y \sim \mathrm{G}(\mu, \nu), \qquad g(\mu) = x'\boldsymbol{\beta}.$$

La función de enlace canónica para la distribución gamma es la función inversa. Como los parámetros de un modelo con función de enlace inversa son difíciles de interpretar, es común utilizar la función de enlace log.

1.5.2 Regresión Gaussiana inversa

El MLG Gaussiana inversa se define como:

$$y \sim \text{IG}(\mu, \sigma^2), \qquad g(\mu) = \mathbf{x}'\boldsymbol{\beta}.$$

La función de enlace canónica es $g(\mu) = \mu^{-2}$. Sin embargo, es común usar la función de enlace log.