4. Modelos lineales generalizados

Introducción a los modelos de regresión generalizados

- Los modelos de regresión en general, son las herramientas estadísticas más usadas en la práctica. Estudian la relación entre dos o más variables observables "y" y "x". En general estamos interesados en la distribución condicional de y dado x, parametrizada por $f(y|\beta,x)$, en donde se tienen acceso a *n* observaciones (x_i,y_i) condicionalmente independientes.

- Existen muchas formas de especificar la relación entre "y" y "x" (modelos lineales generalizados, regresión lineal y no lineal, paramétricos y no paramétricos).

- La variable de interés "y" se conoce como *variable respuesta* y puede ser de naturaleza continua, discreta o categórica, mientras que las variables $x'=(x_1,x_2,\ldots,x_{p-1})$ son llamadas *variables explicativas o predictoras* y también pueden ser discretas, continuas o categóricas.

- MODELO LINEAL NORMAL: Sean Y_i, $x_i'=(1,x_{i1},x_{i2},\ldots,x_{i,p-1})$, i=1,2,..,n un conjunto de variables aleatorias tal que

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip-1} + \varepsilon_i$$
$$= x_i'\beta + \varepsilon_i,$$

donde, $\beta'=(\beta_0,\beta_1,\ldots,\beta_{p-1})$ en un vector de p parámetros,

$\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n$ son v.a.i.i.d. tal que $\varepsilon_i \sim N(0,\tau)$ para i=1,2,..,n, con $\mathrm{Var}(\varepsilon_i)=1/\tau=\sigma^2$.

Esto implica que

$$E\left(Y_i \middle| \beta,\sigma^2,x_i\right)=x_i'\beta$$

$$\mathrm{Var}\left(Y_i \middle| \beta,\sigma^2,x_i\right)=1/\tau=\sigma^2$$

- Alternativamente, el modelo se puede escribir como:

$$Y_i|\beta,\tau,x_i \sim N(x_i'\beta,\tau),\ i=1,...,n$$

De manera conjunta, el modelo se puede expresar como:

$$Y|\beta,\tau,X \sim N(X\beta,\tau I)$$

donde, $Y'=(Y_1,\ldots,Y_n)$, $\beta'=(\beta_0,\beta_1,\ldots,\beta_{p-1})$,

$$X=\begin{pmatrix} 1 & x_{11} & \cdots & x_{1p-1} \\ 1 & x_{21} & \cdots & x_{2p-1} \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np-1} \end{pmatrix}.$$

- OBJETIVOS:
  1) Estimar el valor de los parámetros y determinar si el efecto particular de una variable explicativa es significativo o no.
  2) Predicción de observaciones futuras Y_F, basados en un valor x_F ya sea conocido o hipotético, i.e.,

$$Y_F|\beta,\tau,x_F \sim N(x_F'\beta,\tau).$$

- El modelo de regresión lineal normal es un ejemplo de un modelo de regresión paramétrico y las variables explicativas sólo tienen efecto sobre la media de la distribución de la variable Y.

- Sin embargo, es posible que las variables explicativas no nadamás tengan un efecto sobre la media sino también sobre la varianza, i.e.,

$$E\left(Y_i \mid \beta, \sigma^2, x_i\right) = x_i'\beta$$

$$\mathrm{Var}\left(Y_i \mid \beta, \sigma^2, x_i\right) = \sigma^2(x_i) \quad \text{e.g. } \sigma^2(x_i) = e^{x_i'\gamma}$$

En este caso tendríamos un modelo de regresión a la media y a la varianza.

- NOTA: Cuidar el espacio parametral

- ANÁLISIS DEL MODELO: El análisis del modelo desde un punto de vista Bayesiano requiere de especificar una distribución inicial sobre los parámetros (β,τ). recordemos que $\tau=1/\sigma^2$ es la precisión de los errores. Consideremos la función de verosimilitud,

$$f\left(y \mid \beta, \tau, X\right) \propto \tau^{n/2} \exp\left\{-\frac{\tau}{2}(y - X\beta)'(y - X\beta)\right\}.$$

Escribiendo $y - X\beta = y - X\hat{\beta} + X\left(\hat{\beta} - \beta\right)$, la verosimilitud toma la forma,

$$f\left(y \mid \beta, \tau, X\right) \propto \tau^{n/2} \exp\left[-\frac{\tau}{2}\left\{(n-p)s^2 + \left(\beta - \hat{\beta}\right)' X'X\left(\beta - \hat{\beta}\right)\right\}\right],$$

donde $\hat{\beta} = (X'X)^{-1}X'y$,

$$s^2 = \frac{1}{n-p}\left(y - X\hat{\beta}\right)'\left(y - X\hat{\beta}\right).$$

A continuación consideraremos dos opciones de distribuciones iniciales y sus correspondientes distribuciones finales:

- *Análisis de referencia*:

<u>Inicial</u>: $\pi(\beta,\tau) \propto \tau^{-1}$

Final: $\pi(\beta,\tau \mid y,X)=N\left(\beta \mid \hat{\beta},(X'X)\tau\right)Ga\left(\tau \mid (n-p)/2,(n-p)s^2/2\right)$,

Marginalmente:

$$\pi(\beta \mid y,X)=St\left(\beta \mid \hat{\beta},(X'X)\left(s^2\right)^{-1},n-p\right)$$

$$\pi(\tau \mid y,X)=Ga\left(\tau \mid (n-p)/2,(n-p)s^2/2\right)$$

Predictiva final: Dado un vector de covariables x_F,

$$\pi(y_F \mid y)=St\left(y_F \mid x_F'\hat{\beta},\left(1+x_F'(X'X)^{-1}x_F\right)^{-1}\left(s^2\right)^{-1},n-p\right).$$

- *Análisis conjugado*:

  Inicial: $f(\beta,\tau)=N\left(\beta \mid b_0,T_0\tau\right)Ga\left(\tau \mid a_0/2,c_0/2\right)$

  Final: $f(\beta,\tau \mid y,X)=N\left(\beta \mid b_1,T_1\tau\right)Ga\left(\tau \mid a_1/2,c_1/2\right)$

  donde, $b_1=\left(X'X+T_0\right)^{-1}\left(X'X\hat{\beta}+T_0b_0\right)$

  $T_1=X'X+T_0$

  $a_1=a_0+n$

  $c_1=c_0+\left(y-Xb_1\right)'\left(y-Xb_1\right)+\left(b_1-b_0\right)'T_0\left(b_1-b_0\right)$

  Nota: T_0 puede ser usada para romper multicolinealidad.

  Marginalmente:

  $$f(\beta \mid y,X)=St\left(\beta \mid b_1,T_1,a_1\right)$$

  $$f(\tau \mid y,X)=Ga\left(\tau \mid a_1/2,c_1/2\right)$$

  Predictiva final: Dado un vector de covariables x_F,

  $$f(y_F \mid y)=St\left(y_F \mid x_F'b_1,\frac{a_1}{c_1}\left(1+x_F'T_1x_F\right)^{-1},n-p\right).$$

- INFERENCIA: Recordemos que la inferencia se realiza utilizando la distribución final de los parámetros y la predictiva final.
  - *Estimación / Predicción puntual*. La estimación o predicción puntual puede realizarse utilizando alguna de las funciones de pérdida mencionadas anteriormente (cuadrática, absoluta o vecindad), obteniéndose la media, mediana o moda como estimadores o predictores puntuales.
  - *Pruebas de hipótesis*. La prueba de hipótesis más utilizada en regresión es sobre la significancia de las variables explicativas del modelo, i.e.,

    $$H_0: \beta_j=0 \text{ vs. } H_1: \beta_j\neq 0$$

    La forma común de tomar la decisión es obteniendo un intervalo de credibilidad de β_j al $(1-\alpha)$ de probabilidad, y si el intervalo contiene al cero se dice que la variable explicativa x_j no es significativa para al modelo, en caso contrario se dice que sí es significativa al nivel α.

- AJUSTE DEL MODELO. En estadística frecuentista, una forma de medir el ajuste del modelo es mediante el coeficiente de determinación. Este se construye de la siguiente manera:

  Sea $\hat{y}_i = x_i'\hat{\beta}$ el predictor puntual máximo verosímil de y_i. Se definen

  $S_{YY} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ la variabilidad total de las observaciones

  $SCR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ la variabilidad explicada por la regresión

  $SCE = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$ la variabilidad no explicada por la regresión

  Se puede demostrar que

$S_{YY} = SCR + SCE$.

Entonces, una medida de la variabilidad explicada por el modelo es:

$$R^2 = \frac{SCR}{S_{YY}} \text{ (coeficiente de determinación).}$$

- *Nota 1*. La partición anterior de suma de cuadrados no es válida si en lugar del predictor máximo verosímil se usa algún otro predictor Bayesiano para y_i. Sin embargo es posible calcular $\{Corr(y_i, \hat{y}_i)\}^2$ como medida de ajuste, que en el caso de que $\hat{y}_i$ sea el predictor máximo verosímil, coincidirá con R^2.
- *Nota 2*. Para determinar el ajuste del modelo es mejor usar medidas de comparación de modelos como: LPML, medida-L y DIC.

- EJERCICIO 3. A continuación se presenta una base de datos de calificaciones de 20 empresas financieras hechas por las dos compañías calificadores más importantes S&P y Moody's. Realiza un análisis Bayesiano completo de los datos, ajustando un modelo de regresión lineal, tomando como variable respuesta las calificaciones de S&P y como variable explicativa las calificaciones de Moody's.

- EJERCICIO 4. Un investigador desea evaluar la relación entre el salario anual de trabajadores de una compañía de nivel medio y alto (Y, en miles de dólares) y el índice de calidad de trabajo (X_1), número de años de experiencia (X_2) y el índice de éxito en publicaciones (X_3). La muestra consiste de 24 trabajadores. Realiza un análisis Bayesiano completo de los datos y obtén las predicciones de salarios para 3 nuevos empleados con variables explicativas:

$$x_{1F}'=(5.4,17,6.0),\ x_{2F}'=(6.2,12,5.8) \text{ y } x_{1F}'=(6.4,21,6.1).$$

Modelos lineales generalizados

- Recordemos que el modelo de regresión lineal es una forma de describir la relación entre una variable respuesta "y" y un conjunto de variables explicativas $x'=(x_1,\ldots,x_p)$. Una forma más general de describir la distribución condicional de Y dado X, f(y|β,x), es mediante la clase de modelos lineales generalizados.

- La idea general de los modelos lineales generalizados consiste en modelar el valor esperado de Y, digamos $\mu(x)=E(Y|x)$, a través de una función paramétrica simple de las variables explicativas, digamos $\varphi(\beta,x)$, i.e.,
$$\mu(x)=\varphi(\beta,x).$$

- Al considerar distintas distribuciones para la variable respuesta Y y distintas formas para la función φ(·), este modelo produce una clase muy amplia de modelos de regresión generalizados.

- Los modelos lineales generalizados suponen que la función de densidad de la variable respuesta Y (discreta o continua) es un miembro de la familia exponencial, i.e.,
$$f(y|\theta,\phi)=b(y,\phi)\exp[\phi\{y\theta-a(\theta)\}]$$

donde a(·) y b(·) son funciones monótonas. El parámetro ϕ es un parámetro de dispersión y cuando es conocido describe una familia exponencial natural. La media y la varianza de Y están dadas por:

$$\mu = E(Y|\theta) = a'(\theta) \quad \text{y} \quad Var(Y|\theta) = \frac{a''(\theta)}{\phi}.$$

- MODELO: Los modelos lineales generalizados se describen a través de una componente aleatoria, una componente sistemática y una función liga:
  - *Componente aleatoria*: Sean $Y_1,Y_2,...,Y_n$ v.a.i. de la función de densidad
  $$f(y_i|\theta_i,\phi_i) = b(y_i,\phi_i)\exp[\phi_i\{y_i\theta_i - a(\theta_i)\}]$$
  - *Componente sistemática*: Para cada respuesta Y_i se tiene un vector de covariables o variables explicativas $x_i' = (1, x_{i1}, \ldots x_{ip-1})$ las cuales producen el predictor lineal
  $$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots \beta_{p-1} x_{ip-1} = x_i'\beta$$
  - *Función liga*: Las componentes aleatoria y sistemática se relacionan a través de una función liga g(·), tal que
  $$\eta_i = g(\mu_i).$$
  Un caso particular importante se obtiene cuando $g^{-1}(\cdot) = a'(\cdot)$. En este caso, $\eta_i = \theta_i$ y g(·) es llamada liga canónica.

- CASOS PARTICULARES: Algunos de los modelos lineales generalizados más usados en la práctica son el Normal, el Poisson y el Bernoulli.
  Sean $Y_1,Y_2,...,Y_n$ v.a.i. de la función de densidad
  - <u>*Normal*</u>: $Y_i|\mu_i \sim N(\mu_i,\tau_i)$, $\mu_i \in \Re$

$$f(y_i|\mu_i,\tau_i)=(2\pi/\tau_i)^{-1/2}\exp\left\{-\frac{\tau_i}{2}(y_i-\mu_i)^2\right\}I_{(-\infty,\infty)}(y_i)\text{ , donde}$$

$$\begin{cases}\phi_i=\tau_i\\ b(y_i,\phi_i)=(2\pi/\phi_i)^{-1/2}\exp\left\{-\frac{\phi_i}{2}y_i^2\right\}\\ \theta_i=\mu_i\\ a(\theta_i)=\frac{\theta_i^2}{2}\Rightarrow a'(\theta_i)=\theta_i\end{cases}$$

Liga canónica: $g(\mu_i)=\mu_i$

$\therefore\ \mu_i=\eta_i=x_i'\beta$ (modelo lineal)

- *Poisson*: $Y_i|\mu_i\sim Po(\mu_i)$, $\mu_i\in\Re^+$

$$f(y_i|\mu_i)=e^{-\mu_i}\frac{\mu_i^{y_i}}{y_i!}I_{\{0,1,\ldots\}}(y_i)\text{, donde}$$

$$\begin{cases}\phi_i=1\\ b(y_i,\phi_i)=b(y_i)=\frac{1}{y_i!}I_{\{0,1,\ldots\}}(y_i)\\ \theta_i=\log\mu_i\\ a(\theta_i)=e^{\theta_i}\Rightarrow a'(\theta_i)=e^{\theta_i}\end{cases}$$

Liga canónica: $g(\mu_i)=\log\mu_i$

$\therefore\ \log\mu_i=\eta_i=x_i'\beta$ (modelo log-lineal)

- *Bernoulli*: $Y_i|\mu_i\sim Ber(\mu_i)$, $\mu_i\in(0,1)$

$$f(y_i|\mu_i)=\mu_i^{y_i}(1-\mu_i)^{1-y_i}I_{\{0,1\}}(y_i)\text{, donde}$$

$$\begin{cases} \phi_i = 1 \\ b(y_i, \phi_i) = b(y_i) = I_{\{0,1\}}(y_i) \\ \theta_i = \log\left(\dfrac{\mu_i}{1-\mu_i}\right) \\ a(\theta_i) = \log\left(1+e^{\theta_i}\right) \Rightarrow a'(\theta_i) = e^{\theta_i} \big/ \left(1+e^{\theta_i}\right) \end{cases}$$

Liga canónica: $g(\mu_i) = \log\left(\dfrac{\mu_i}{1-\mu_i}\right)$

$$\therefore\ \log\left(\frac{\mu_i}{1-\mu_i}\right) = \eta_i = x_i'\beta \text{ (modelo logístico)}$$

Otras ligas: $g(\mu_i) = \Phi^{-1}(\mu_i)$, donde $\Phi(\cdot)$ es la función de distribución normal estándar $\therefore\ \mu_i = \Phi(\eta_i) = \Phi(x_i'\beta)$ (modelo probit)

- ANÁLISIS DEL MODELO: Consideremos la forma general del modelo lineal generalizado, en donde las observaciones $Y_1, Y_2, \ldots, Y_n$ son un conjunto de v.a.i. de la función de densidad

$$f(y_i | \theta_i, \phi_i) = b(y_i, \phi_i) \exp[\phi_i \{y_i\theta_i - a(\theta_i)\}], \text{ i=1,...,n.}$$

Por simplicidad supongamos que ϕ_i es conocido y que las componentes aleatoria y sistemática se relacionan a través dela liga canónica, i.e., $\theta_i = \eta_i = x_i'\beta$. En este caso, la función de verosimilitud es de la forma,

$$f(y|\beta) \propto \exp\left[\sum_{i=1}^{n} \phi_i \{y_i(x_i'\beta) - a(x_i'\beta)\}\right].$$

Inicial: $f(\beta) = N(\beta | b_0, T_0)$

Final: $f(\beta | y, X) \propto \exp\left[\sum_{i=1}^{n} \phi_i \{y_i(x_i'\beta) - a(x_i'\beta)\} - \frac{1}{2}(\beta - b_0)'T_0(\beta - b_0)\right]$

Predictiva final: Dado un vector de covariables x_F y un valor ϕ_F,

$$f(y_F|y) = \int f(y_F|x_F'\beta,\phi_F)f(\beta|y,X)d\beta = E_{\beta|y,X}\{f(y_F|x_F'\beta,\phi_F)\}.$$

- AJUSTE DEL MODELO. Existen varias formas de medir la bondad del ajuste de un modelo. En modelos lineales generalizados la medida más común es la *devianza*, la cual se define como dos veces el logaritmo del cociente de verosimilitudes, es decir,

$$\text{Dev} = 2\log\frac{L(\tilde{\theta}\,|\,y)}{L(\hat{\theta}\,|\,y)},$$

donde $\hat{\theta}$ y $\tilde{\theta}$ son los valores maximizados de θ en dos modelos diferentes. En el caso de modelos lineales generalizados la *devianza* toma la forma:

$$\text{Dev} = 2\sum_{i=1}^{n}\phi_i\left\{y_i\left(\tilde{\theta}_i - \hat{\theta}_i\right) - a\left(\tilde{\theta}_i\right) + a\left(\hat{\theta}_i\right)\right\}.$$

Para los tres casos particulares de modelos lineales generalizados vistos anteriormente, la *devianza* es de la forma:

MODELO	DEVIANZA
Normal	$\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$
Poisson	$2\sum_{i=1}^{n}\left\{y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right\}$
Bernoulli	$2\sum_{i=1}^{n}\left\{y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right) + (1-y_i)\log\left(\frac{1-y_i}{1-\hat{\mu}_i}\right)\right\}$

donde $\hat{\mu}_i$ es el predictor puntual de y_i.

- NOTA. La Devianza es la medida de ajuste frecuentista usada para los modelos lineales generalizados, son embargo en el caso Bayesiano es posible usar la medida equivalente DIC definida anteriormente.

- EJERCICIO 7. Tasas de mortalidad (Congdon, 2001). Una compañía de seguros quiere lanzar un nuevo seguro médico para mineros. Para ello desea estimar la probabilidad de muerte (π_i), con base en el tiempo de exposición al mineral (x_i en horas). Se cuenta con información de las muertes registradas entre 1950 y 1959, junto con el tiempo de exposición al mineral y el número de mineros expuestos. Realiza un análisis Bayesiano de los datos y obtén la distribución predictiva del número de muertes suponiendo que hay 100 mineros con un tiempo de exposición de 200 horas. El modelo es el siguiente: Para i=1,...,N

$$Y_i | \pi_i \sim \text{Bin}(n_i, \pi_i)$$

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$$

con $\beta_0 \sim N(0,0.001)$ y $\beta_1 \sim N(0,0.001)$.

- EJERCICIO 8. En el mismo contexto del problema anterior, supongamos ahora que la compañía de seguros está interesada en modelar el número total de desastres (Y_t) que ocurren en la mina. Se cuenta con N=112 observaciones durante los años 1851 a 1962. Se proponen tres modelos:

a) Modelo con tasa variable en función del tiempo:

$$Y_t | \mu_t \sim \text{Po}(\mu_t)$$

$$\log(\mu_t) = \beta_0 + \beta_1 t$$

con $\beta_0 \sim N(0,0.001)$ y $\beta_1 \sim N(0,0.001)$.

b) Modelo con tasa constante en dos períodos: Se cree que la tasa promedio de desastres es constante, pero que en el siglo XX la tasa ha disminuido. Esto se traduce en el siguiente modelo:

$$Y_t | \mu_t \sim Po(\mu_t)$$

$$\log(\mu_t) = \beta_0 + \beta_1 I(t \geq \tau)$$

con $\beta_0 \sim N(0,0.001)$, $\beta_1 \sim N(0,0.001)$ y $\tau \sim U\{1,\ldots,N\}$.

c) Modelo Poisson dinámico:

Observación: $Y_t | \mu_t \sim Po(\mu_t)$, $\log \mu_t = \eta_t$

Evolución: $\eta_t = \eta_{t-1} + \omega_t$, $\omega_t \sim N(0, W^{-1}) \Rightarrow \eta_t | \eta_{t-1} \sim N(\eta_{t-1}, W^{-1})$

con $\eta_1 \sim N(0,0.0001)$ y $W^{-1} \sim Ga(0.01,0.01)$.