Modelos Lineales

Modelos Lineales Generalizados

Fernando Massa; Bruno Bellagamba

20 de junio 2023

FCEA FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN
IESTA INSTITUTO DE ESTADÍSTICA
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Temario

Objetivos

- Realizaremos una introducción a la familia *GLM*.
- Reconoceremos sus componentes y cómo los modelos que hemos visto hasta ahora son casos particulares.
- Recordaremos algunas propiedades de la familia exponencial y su vínculo con estos modelos.
- Reconoceremos algunas distribuciones como miembros de la familia exponencial.
- Plantearemos nuestro primer *GLM* no normal.

Introducción

Hasta la semana pasada realizamos distintos análisis basados en modelos lineales asumiendo siempre que la variable de respuesta Y era cuantitativa continua.

Introducción

Hasta la semana pasada realizamos distintos análisis basados en modelos lineales asumiendo siempre que la variable de respuesta Y era cuantitativa continua.

Aún más específica fue nuestra elección de densidad de probabilidad para realizar inferencia.

Introducción

Hasta la semana pasada realizamos distintos análisis basados en modelos lineales asumiendo siempre que la variable de respuesta Y era cuantitativa continua.

Aún más específica fue nuestra elección de densidad de probabilidad para realizar inferencia.

En cualquiera de los casos, la especificación del modelo pasaba por introducir variables explicativas con el fin de modelar la media condicional de Y a través de una función lineal.

Introducción

Hasta la semana pasada realizamos distintos análisis basados en modelos lineales asumiendo siempre que la variable de respuesta Y era cuantitativa continua.

Aún más específica fue nuestra elección de densidad de probabilidad para realizar inferencia.

En cualquiera de los casos, la especificación del modelo pasaba por introducir variables explicativas con el fin de modelar la media condicional de Y a través de una función lineal.

$$\mathbb{E}(Y|X) = X\beta$$

En la literatura se suele referir al *Modelo Lineal General* como el caso más amplio de estos modelos donde básicamente se incluyen varias variables explicativas, pudiendo estas ser de naturaleza cuantitativa o cualitativa.

Introducción

Hasta la semana pasada realizamos distintos análisis basados en modelos lineales asumiendo siempre que la variable de respuesta Y era cuantitativa continua.

Aún más específica fue nuestra elección de densidad de probabilidad para realizar inferencia.

En cualquiera de los casos, la especificación del modelo pasaba por introducir variables explicativas con el fin de modelar la media condicional de Y a través de una función lineal.

$$\mathbb{E}(Y|X) = X\beta$$

En la literatura se suele referir al *Modelo Lineal General* como el caso más amplio de estos modelos donde básicamente se incluyen varias variables explicativas, pudiendo estas ser de naturaleza cuantitativa o cualitativa. No obstante, este modelo puede no resultar adecuado en otros casos...

La generalización

El modelo con el que hemos venido trabajando podría ser caracterizado por los siguientes componentes:

La generalización

El modelo con el que hemos venido trabajando podría ser caracterizado por los siguientes componentes:

Modelo Lineal General

- Las observaciones $Y_1, Y_2, \ldots, Y_n$ son realizaciones de una variable aleatoria con distribución Normal de varianza constante σ^2.
- $\mathbb{E}(Y) = \mu = X\beta$.

La generalización

El modelo con el que hemos venido trabajando podría ser caracterizado por los siguientes componentes:

Modelo Lineal General

- Las observaciones $Y_1, Y_2, \ldots, Y_n$ son realizaciones de una variable aleatoria con distribución Normal de varianza constante σ^2.
- $\mathbb{E}(Y) = \mu = X\beta$.

El primero de estos componentes refiere al componente **aleatorio** del modelo con el que hemos estado trabajando.

La generalización

El modelo con el que hemos venido trabajando podría ser caracterizado por los siguientes componentes:

Modelo Lineal General

- Las observaciones $Y_1, Y_2, \ldots, Y_n$ son realizaciones de una variable aleatoria con distribución Normal de varianza constante σ^2.
- $\mathbb{E}(Y) = \mu = X\beta$.

El primero de estos componentes refiere al componente **aleatorio** del modelo con el que hemos estado trabajando.

El otro compomente nos dice como es la relación entre la media de Y y la parte **sistemática** del modelo.

La generalización

En el Modelo Lineal **Generalizado** (GLM) podemos identificar estos componentes y algo más.

En el Modelo Lineal **Generalizado** (GLM) podemos identificar estos componentes y algo más.

Modelo Lineal Generalizado (GLM)

- Las observaciones $Y_1, Y_2, \ldots, Y_n$ tienen distribución $f_Y(y)$ perteneciente a la familia exponencial.
- $\eta = X\beta$.
- Siendo $\mu = \mathbb{E}(Y)$, la relaciíon entre μ y η es que $g(\mu) = \eta$.

En el Modelo Lineal **Generalizado** (GLM) podemos identificar estos componentes y algo más.

Modelo Lineal Generalizado (GLM)

- Las observaciones $Y_1, Y_2, \ldots, Y_n$ tienen distribución $f_Y(y)$ perteneciente a la familia exponencial.
- $\eta = X\beta$.
- Siendo $\mu = \mathbb{E}(Y)$, la relaciíon entre μ y η es que $g(\mu) = \eta$.

El primero de estos componentes refiere al componente **aleatorio**.

La generalización

En el Modelo Lineal **Generalizado** (GLM) podemos identificar estos componentes y algo más.

Modelo Lineal Generalizado (GLM)

- Las observaciones $Y_1, Y_2, \ldots, Y_n$ tienen distribución $f_Y(y)$ perteneciente a la familia exponencial.
- $\eta = X\beta$.
- Siendo $\mu = \mathbb{E}(Y)$, la relaciíon entre μ y η es que $g(\mu) = \eta$.

El primero de estos componentes refiere al componente **aleatorio**.

El segundo compomente refiere a la parte **sistemática** del modelo.

En el Modelo Lineal **Generalizado** (GLM) podemos identificar estos componentes y algo más.

Modelo Lineal Generalizado (GLM)

- Las observaciones $Y_1, Y_2, \ldots, Y_n$ tienen distribución $f_Y(y)$ perteneciente a la familia exponencial.
- $\eta = X\beta$.
- Siendo $\mu = \mathbb{E}(Y)$, la relaciíon entre μ y η es que $g(\mu) = \eta$.

El primero de estos componentes refiere al componente **aleatorio**.

El segundo compomente refiere a la parte **sistemática** del modelo.

El tercer componente establece la relación entre el componente sistemático y la media de Y a través una **función de enlace** $g()$.

Modelo Lineal Generalizado (GLM)

El marco de trabajo generado por los GLM permite unificar un conjunto de técnicas (más o menos dispares) bajo un gran abanico de posibilidades sin tener que recurrir a distintos nombres (regresión lineal simple o múltiple, ANOVA, ANCOVA u otros tantos):

Modelo Lineal Generalizado (GLM)

El marco de trabajo generado por los GLM permite unificar un conjunto de técnicas (más o menos dispares) bajo un gran abanico de posibilidades sin tener que recurrir a distintos nombres (regresión lineal simple o múltiple, ANOVA, ANCOVA u otros tantos):

La idea de esta familia de modelos permite acceder a estos modelos y otros más a partir de los 3 componentes nombrados previamente.

- Componente aleatorio (distribución).
- Componente sistemático (predictor lineal).
- Función de enlace.

Modelo Lineal Generalizado (GLM)

El marco de trabajo generado por los GLM permite unificar un conjunto de técnicas (más o menos dispares) bajo un gran abanico de posibilidades sin tener que recurrir a distintos nombres (regresión lineal simple o múltiple, ANOVA, ANCOVA u otros tantos):

La idea de esta familia de modelos permite acceder a estos modelos y otros más a partir de los 3 componentes nombrados previamente.

- Componente aleatorio (distribución).
- Componente sistemático (predictor lineal).
- Función de enlace.

En principio, el predictor lineal ni siquiera se altera con respecto a los modelos ya analizados, corresponde al investigador determinar una distribución de probabilidad y una función de enlace adecuadas al problema sobre el que trabaja.

Selección de $f_Y(y)$ y $g(\mu)$

A la hora de seleccionar el componente aleatorio es importante fijarse en el tipo de variable que es Y y en su recorrido. De esta manera, algunas posibilidades serían:

Selección de $f_Y(y)$ y $g(\mu)$

A la hora de seleccionar el componente aleatorio es importante fijarse en el tipo de variable que es Y y en su recorrido. De esta manera, algunas posibilidades serían:

- Si Y es cuantitativa continua $\Rightarrow N(\mu, \sigma)$ (Por ej: temperatura).

Selección de $f_Y(y)$ y $g(\mu)$

A la hora de seleccionar el componente aleatorio es importante fijarse en el tipo de variable que es Y y en su recorrido. De esta manera, algunas posibilidades serían:

- Si Y es cuantitativa continua $\Rightarrow N(\mu, \sigma)$ (Por ej: temperatura).
- Si Y es cuantitativa continua NO negativa $\Rightarrow Gamma(\alpha, \beta)$ (Por ej: tiempo hasta fallecimiento).

Selección de $f_Y(y)$ y $g(\mu)$

A la hora de seleccionar el componente aleatorio es importante fijarse en el tipo de variable que es Y y en su recorrido. De esta manera, algunas posibilidades serían:

- Si Y es cuantitativa continua $\Rightarrow N(\mu, \sigma)$ (Por ej: temperatura).
- Si Y es cuantitativa continua NO negativa $\Rightarrow Gamma(\alpha, \beta)$ (Por ej: tiempo hasta fallecimiento).
- Si Y es cuantitativa continua entre 0 y 1 $\Rightarrow Beta(\alpha, \beta)$ (Por ej: % de cobertura).

Selección de $f_Y(y)$ y $g(\mu)$

A la hora de seleccionar el componente aleatorio es importante fijarse en el tipo de variable que es Y y en su recorrido. De esta manera, algunas posibilidades serían:

- Si Y es cuantitativa continua $\Rightarrow N(\mu, \sigma)$ (Por ej: temperatura).
- Si Y es cuantitativa continua NO negativa $\Rightarrow Gamma(\alpha, \beta)$ (Por ej: tiempo hasta fallecimiento).
- Si Y es cuantitativa continua entre 0 y 1 $\Rightarrow Beta(\alpha, \beta)$ (Por ej: % de cobertura).
- Si Y es cuantitativa discreta $\Rightarrow Poisson(\lambda)$ (Por ej: Nro. de avistamientos).

Selección de $f_Y(y)$ y $g(\mu)$

A la hora de seleccionar el componente aleatorio es importante fijarse en el tipo de variable que es Y y en su recorrido. De esta manera, algunas posibilidades serían:

- Si Y es cuantitativa continua $\Rightarrow N(\mu, \sigma)$ (Por ej: temperatura).
- Si Y es cuantitativa continua NO negativa $\Rightarrow Gamma(\alpha, \beta)$ (Por ej: tiempo hasta fallecimiento).
- Si Y es cuantitativa continua entre 0 y 1 $\Rightarrow Beta(\alpha, \beta)$ (Por ej: % de cobertura).
- Si Y es cuantitativa discreta $\Rightarrow Poisson(\lambda)$ (Por ej: Nro. de avistamientos).
- Si Y es cuantitativa discreta con máximo $m \Rightarrow Binomial(m, p)$ (Por ej: Nro. de camas ocupadas).

Selección de $f_Y(y)$ y $g(\mu)$

A la hora de seleccionar el componente aleatorio es importante fijarse en el tipo de variable que es Y y en su recorrido. De esta manera, algunas posibilidades serían:

- Si Y es cuantitativa continua $\Rightarrow N(\mu, \sigma)$ (Por ej: temperatura).
- Si Y es cuantitativa continua NO negativa $\Rightarrow Gamma(\alpha, \beta)$ (Por ej: tiempo hasta fallecimiento).
- Si Y es cuantitativa continua entre 0 y 1 $\Rightarrow Beta(\alpha, \beta)$ (Por ej: % de cobertura).
- Si Y es cuantitativa discreta $\Rightarrow Poisson(\lambda)$ (Por ej: Nro. de avistamientos).
- Si Y es cuantitativa discreta con máximo $m \Rightarrow Binomial(m, p)$ (Por ej: Nro. de camas ocupadas).
- Si Y es cualitatitva con 2 categorías $\Rightarrow Bernoulli(p)$ (Por ej: enfermo/sano).

Selección de $f_Y(y)$ y $g(\mu)$

A la hora de seleccionar el componente aleatorio es importante fijarse en el tipo de variable que es Y y en su recorrido. De esta manera, algunas posibilidades serían:

- Si Y es cuantitativa continua $\Rightarrow N(\mu, \sigma)$ (Por ej: temperatura).
- Si Y es cuantitativa continua NO negativa $\Rightarrow Gamma(\alpha, \beta)$ (Por ej: tiempo hasta fallecimiento).
- Si Y es cuantitativa continua entre 0 y 1 $\Rightarrow Beta(\alpha, \beta)$ (Por ej: % de cobertura).
- Si Y es cuantitativa discreta $\Rightarrow Poisson(\lambda)$ (Por ej: Nro. de avistamientos).
- Si Y es cuantitativa discreta con máximo $m \Rightarrow Binomial(m, p)$ (Por ej: Nro. de camas ocupadas).
- Si Y es cualitatitva con 2 categorías $\Rightarrow Bernoulli(p)$ (Por ej: enfermo/sano).

Existen más situaciones donde otras distribuciones son adecuadas e incluso otras distribuciones pueden resultar adecuadas para estos ejemplos.

Familia exponencial

Al momento de seleccionar una función de enlace, también existen diversas alternativas para cada caso. En el caso que vimos hasta esta parte del curso, usamos la función identidad para el caso del modelo lineal general. Los componentes de ese GLM son:

Familia exponencial

Al momento de seleccionar una función de enlace, también existen diversas alternativas para cada caso. En el caso que vimos hasta esta parte del curso, usamos la función identidad para el caso del modelo lineal general. Los componentes de ese GLM son:

GLM normal

- Las observaciones $Y_1, Y_2, \ldots, Y_n$ tienen distribución $N(\mu_i, \sigma^2)$.
- $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$.
- $g(\mu_i) = \eta_i \Rightarrow \mu_i = \eta_i$

Familia exponencial

Al momento de seleccionar una función de enlace, también existen diversas alternativas para cada caso. En el caso que vimos hasta esta parte del curso, usamos la función identidad para el caso del modelo lineal general. Los componentes de ese GLM son:

GLM normal

- Las observaciones $Y_1, Y_2, \ldots, Y_n$ tienen distribución $N(\mu_i, \sigma^2)$.
- $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$.
- $g(\mu_i) = \eta_i \Rightarrow \mu_i = \eta_i$

Sin embargo, es posible utilizar otras funciones de enlace ya que pueden ofrecer un mejor ajuste.

Familia exponencial

Al momento de seleccionar una función de enlace, también existen diversas alternativas para cada caso. En el caso que vimos hasta esta parte del curso, usamos la función identidad para el caso del modelo lineal general. Los componentes de ese GLM son:

GLM normal

- Las observaciones $Y_1, Y_2, \ldots, Y_n$ tienen distribución $N(\mu_i, \sigma^2)$.
- $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$.
- $g(\mu_i) = \eta_i \Rightarrow \mu_i = \eta_i$

Sin embargo, es posible utilizar otras funciones de enlace ya que pueden ofrecer un mejor ajuste.

En el caso normal, la función identidad es la que llamamos **función de enlace canónica**. Estas pueden ser obtenidas planteando la densidad seleccionada como componente aleatorio como miembro de la *familia exponencial*.

Familia exponencial

Si bien existen varias parametrizaciones, la utilizada por McCullagh y Nelder es:

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

Cada densidad perteneciente a esta familia, tiene ciertas funciones $a()$, $b()$ y $c()$. Al parámetro θ se lo suele conocer como parámetro *canónico* y ϕ es un parámetro de dispersión. Veamos si la distribución normal pertenece a esta familia.

Si bien existen varias parametrizaciones, la utilizada por McCullagh y Nelder es:

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

Cada densidad perteneciente a esta familia, tiene ciertas funciones $a()$, $b()$ y $c()$. Al parámetro θ se lo suele conocer como parámetro *canónico* y ϕ es un parámetro de dispersión. Veamos si la distribución normal pertenece a esta familia.

$$\begin{aligned} f_Y(y|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-(y-\mu)^2/2\sigma^2\right\} \\ &= exp\left\{\frac{(y\mu - \mu^2/2)}{\sigma^2} - \frac{1}{2}\left[y^2/\sigma^2 + \log(2\pi\sigma^2)\right]\right\} \end{aligned}$$

Si bien existen varias parametrizaciones, la utilizada por McCullagh y Nelder es:

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

Cada densidad perteneciente a esta familia, tiene ciertas funciones $a()$, $b()$ y $c()$. Al parámetro θ se lo suele conocer como parámetro *canónico* y ϕ es un parámetro de dispersión. Veamos si la distribución normal pertenece a esta familia.

$$\begin{aligned} f_Y(y|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-(y-\mu)^2/2\sigma^2\right\} \\ &= exp\left\{\frac{(y\mu - \mu^2/2)}{\sigma^2} - \frac{1}{2}\left[y^2/\sigma^2 + \log(2\pi\sigma^2)\right]\right\} \end{aligned}$$

Entonces:

Si bien existen varias parametrizaciones, la utilizada por McCullagh y Nelder es:

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

Cada densidad perteneciente a esta familia, tiene ciertas funciones $a()$, $b()$ y $c()$. Al parámetro θ se lo suele conocer como parámetro *canónico* y ϕ es un parámetro de dispersión. Veamos si la distribución normal pertenece a esta familia.

$$\begin{aligned} f_Y(y|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-(y-\mu)^2/2\sigma^2\right\} \\ &= exp\left\{\frac{(y\mu - \mu^2/2)}{\sigma^2} - \frac{1}{2}\left[y^2/\sigma^2 + \log(2\pi\sigma^2)\right]\right\} \end{aligned}$$

Entonces: $\mu = \theta$,

Si bien existen varias parametrizaciones, la utilizada por McCullagh y Nelder es:

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

Cada densidad perteneciente a esta familia, tiene ciertas funciones $a()$, $b()$ y $c()$. Al parámetro θ se lo suele conocer como parámetro *canónico* y ϕ es un parámetro de dispersión. Veamos si la distribución normal pertenece a esta familia.

$$\begin{aligned} f_Y(y|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-(y-\mu)^2/2\sigma^2\right\} \\ &= exp\left\{\frac{(y\mu - \mu^2/2)}{\sigma^2} - \frac{1}{2}\left[y^2/\sigma^2 + \log(2\pi\sigma^2)\right]\right\} \end{aligned}$$

Entonces: $\mu = \theta$, $\phi = \sigma^2$,

Si bien existen varias parametrizaciones, la utilizada por McCullagh y Nelder es:

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

Cada densidad perteneciente a esta familia, tiene ciertas funciones $a()$, $b()$ y $c()$. Al parámetro θ se lo suele conocer como parámetro *canónico* y ϕ es un parámetro de dispersión. Veamos si la distribución normal pertenece a esta familia.

$$\begin{aligned} f_Y(y|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-(y-\mu)^2/2\sigma^2\right\} \\ &= exp\left\{\frac{(y\mu - \mu^2/2)}{\sigma^2} - \frac{1}{2}\left[y^2/\sigma^2 + \log(2\pi\sigma^2)\right]\right\} \end{aligned}$$

Entonces: $\mu = \theta$, $\phi = \sigma^2$, $a(\phi) = \phi$,

Si bien existen varias parametrizaciones, la utilizada por McCullagh y Nelder es:

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

Cada densidad perteneciente a esta familia, tiene ciertas funciones $a()$, $b()$ y $c()$. Al parámetro θ se lo suele conocer como parámetro *canónico* y ϕ es un parámetro de dispersión. Veamos si la distribución normal pertenece a esta familia.

$$\begin{aligned} f_Y(y|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-(y-\mu)^2/2\sigma^2\right\} \\ &= exp\left\{\frac{(y\mu - \mu^2/2)}{\sigma^2} - \frac{1}{2}\left[y^2/\sigma^2 + \log(2\pi\sigma^2)\right]\right\} \end{aligned}$$

Entonces: $\mu = \theta$, $\phi = \sigma^2$, $a(\phi) = \phi$, $b(\theta) = \theta^2/2$

Si bien existen varias parametrizaciones, la utilizada por McCullagh y Nelder es:

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

Cada densidad perteneciente a esta familia, tiene ciertas funciones $a()$, $b()$ y $c()$. Al parámetro θ se lo suele conocer como parámetro *canónico* y ϕ es un parámetro de dispersión. Veamos si la distribución normal pertenece a esta familia.

$$\begin{aligned} f_Y(y|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-(y-\mu)^2/2\sigma^2\right\} \\ &= exp\left\{\frac{(y\mu - \mu^2/2)}{\sigma^2} - \frac{1}{2}\left[y^2/\sigma^2 + \log(2\pi\sigma^2)\right]\right\} \end{aligned}$$

Entonces: $\mu = \theta$, $\phi = \sigma^2$, $a(\phi) = \phi$, $b(\theta) = \theta^2/2$ y $c(y,\phi) = -\frac{1}{2}\left[y^2/\phi + \log(2\pi\phi)\right]$

Familia exponencial (Bernoulli)

Reconozcamos algunos casos más. Comencemos por la distribución Bernoulli.

$$
\begin{aligned}
f_Y(y|p) &= p^y(1-p)^{1-y} \\
&= exp\{y\log(p)+(1-y)\log(1-p)\} \\
&= exp\left\{y\log\left(\frac{p}{1-p}\right)+\log(1-p)\right\} \\
&= exp\left\{\frac{y\log\left(\frac{p}{1-p}\right)+\log(1-p)}{1}+0\right\}
\end{aligned}
$$

Familia exponencial (Bernoulli)

Reconozcamos algunos casos más. Comencemos por la distribución Bernoulli.

$$\begin{aligned} f_Y(y|p) &= p^y(1-p)^{1-y} \\ &= exp\{y\log(p)+(1-y)\log(1-p)\} \\ &= exp\left\{y\log\left(\frac{p}{1-p}\right)+\log(1-p)\right\} \\ &= exp\left\{\frac{y\log\left(\frac{p}{1-p}\right)+\log(1-p)}{1}+0\right\} \end{aligned}$$

Entonces: $\theta = \log\left(\frac{p}{1-p}\right)$, $\phi = 1$, $a(.)$ es la identidad, $b(\theta) = \log(1+e^\theta)$ y $c(.) = 0$.

Familia exponencial (Bernoulli)

Reconozcamos algunos casos más. Comencemos por la distribución Bernoulli.

$$\begin{aligned} f_Y(y|p) &= p^y(1-p)^{1-y} \\ &= exp\{y\log(p)+(1-y)\log(1-p)\} \\ &= exp\left\{y\log\left(\frac{p}{1-p}\right)+\log(1-p)\right\} \\ &= exp\left\{\frac{y\log\left(\frac{p}{1-p}\right)+\log(1-p)}{1}+0\right\} \end{aligned}$$

Entonces: $\theta = \log\left(\frac{p}{1-p}\right)$, $\phi = 1$, $a(.)$ es la identidad, $b(\theta) = \log(1+e^\theta)$ y $c(.) = 0$.

A diferencia del caso normal, el parámetro canónico de la distribución no coincide con el parámetro al que estamos acostumbrados (que representa la media). La función que los vincula es la **función de enlace canónica**, en este caso, la función logística.

Familia exponencial (Bernoulli)

Habiendo determinado que la Bernoulli pertenece a la familia exponencial y que el enlace canónico para este caso es la función logística, estamos en condiciones de construir un *GLM* adecuado a casos donde Y es una varirable cualitativa con 2 opciones.

Familia exponencial (Bernoulli)

Habiendo determinado que la Bernoulli pertenece a la familia exponencial y que el enlace canónico para este caso es la función logística, estamos en condiciones de construir un *GLM* adecuado a casos donde Y es una varirable cualitativa con 2 opciones.

GLM Bernoulli

- Las observaciones $Y_1, Y_2, \ldots, Y_n$ tienen distribución $Ber(\mu_i)$.
- $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$.
- $g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \eta_i$

Familia exponencial (Bernoulli)

Habiendo determinado que la Bernoulli pertenece a la familia exponencial y que el enlace canónico para este caso es la función logística, estamos en condiciones de construir un *GLM* adecuado a casos donde Y es una varirable cualitativa con 2 opciones.

GLM Bernoulli

- Las observaciones $Y_1, Y_2, \ldots, Y_n$ tienen distribución $Ber(\mu_i)$.
- $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$.
- $g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \eta_i$

Recordemos que en este caso $\mathbb{E}(Y) = \mu = p$.

Familia exponencial (Bernoulli)

Habiendo determinado que la Bernoulli pertenece a la familia exponencial y que el enlace canónico para este caso es la función logística, estamos en condiciones de construir un *GLM* adecuado a casos donde Y es una varirable cualitativa con 2 opciones.

GLM Bernoulli

- Las observaciones $Y_1, Y_2, \ldots, Y_n$ tienen distribución $Ber(\mu_i)$.
- $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$.
- $g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \eta_i$

Recordemos que en este caso $\mathbb{E}(Y) = \mu = p$.

Mediante un análisis similar es posible determinar que la cuantía $Bin(m, p)$ también pertenecea a la familia exponencial y que el modelo *GLM* planteado anteriormente también es adecuado para este caso.

Familia exponencial (Poisson)

Veamos un caso más, el de la distribución Poisson.

$$
\begin{aligned}
f_Y(y|\lambda) &= \frac{\lambda^y e^{-\lambda}}{y!} \\
&= e^{y\log(\lambda)-\lambda-\log(y!)} \\
&= exp\{y\log(\lambda)-\lambda-\log(y!)\} \\
&= exp\left\{\frac{y\log(\lambda)-\lambda}{1}+\log(y!)\right\}
\end{aligned}
$$

Familia exponencial (Poisson)

Veamos un caso más, el de la distribución Poisson.

$$\begin{aligned} f_Y(y|\lambda) &= \frac{\lambda^y e^{-\lambda}}{y!} \\ &= e^{y\log(\lambda)-\lambda-\log(y!)} \\ &= exp\{y\log(\lambda)-\lambda-\log(y!)\} \\ &= exp\left\{\frac{y\log(\lambda)-\lambda}{1}+\log(y!)\right\} \end{aligned}$$

Entonces: $\theta = \log(\lambda)$, $\phi = 1$, $a(.)$ es la identidad, $b(\theta) = e^\theta$ y $c(y,\phi) = \log(y!)$.

Familia exponencial (Poisson)

Veamos un caso más, el de la distribución Poisson.

$$\begin{aligned} f_Y(y|\lambda) &= \frac{\lambda^y e^{-\lambda}}{y!} \\ &= e^{y\log(\lambda)-\lambda-\log(y!)} \\ &= exp\{y\log(\lambda)-\lambda-\log(y!)\} \\ &= exp\left\{\frac{y\log(\lambda)-\lambda}{1}+\log(y!)\right\} \end{aligned}$$

Entonces: $\theta = \log(\lambda)$, $\phi = 1$, $a(.)$ es la identidad, $b(\theta) = e^\theta$ y $c(y,\phi) = \log(y!)$.

En este caso el parámetro canónico tampoco coincide con el parámetro al que estamos acostumbrados (que representa la media). La función que los vincula (función de enlace canónica), es la función logarítmica.

Familia exponencial (Poisson)

Como la Poisson pertenece a la familia exponencial y su enlace canónico es el logaritmo, podemos plantear un modelo *GLM* adecuado a casos donde Y es una variable cuantitativa discreta (un conteo).

Familia exponencial (Poisson)

Como la Poisson pertenece a la familia exponencial y su enlace canónico es el logaritmo, podemos plantear un modelo *GLM* adecuado a casos donde Y es una variable cuantitativa discreta (un conteo).

GLM Poisson

- Las observaciones $Y_1, Y_2, \ldots, Y_n$ tienen distribución $Poisson(\mu_i)$.
- $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$.
- $g(\mu_i) = \log(\mu_i) = \eta_i$

Familia exponencial (Poisson)

Como la Poisson pertenece a la familia exponencial y su enlace canónico es el logaritmo, podemos plantear un modelo *GLM* adecuado a casos donde Y es una variable cuantitativa discreta (un conteo).

GLM Poisson

- Las observaciones $Y_1, Y_2, \ldots, Y_n$ tienen distribución $Poisson(\mu_i)$.
- $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$.
- $g(\mu_i) = \log(\mu_i) = \eta_i$

Recordemos que en este caso $\mathbb{E}(Y) = \mu = \lambda$.

Familia exponencial (Poisson)

Como la Poisson pertenece a la familia exponencial y su enlace canónico es el logaritmo, podemos plantear un modelo *GLM* adecuado a casos donde Y es una variable cuantitativa discreta (un conteo).

GLM Poisson

- Las observaciones $Y_1, Y_2, \ldots, Y_n$ tienen distribución $Poisson(\mu_i)$.
- $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$.
- $g(\mu_i) = \log(\mu_i) = \eta_i$

Recordemos que en este caso $\mathbb{E}(Y) = \mu = \lambda$.

Debe tenerse en cuenta que una limitación importante de este modelo es que una de las propiedades de esta distribución es la *equidispersión*: $\mathbb{E}(Y) = \mathbb{V}ar(Y)$.

Regresión logística

El objetivo de esta clase de modelos es el mismo que el que venimos persiguiendo desde el inicio del curso:

Regresión logística

El objetivo de esta clase de modelos es el mismo que el que venimos persiguiendo desde el inicio del curso:

Obtener el modelo que presente el mejor ajuste (de la forma más parsimoniosa posible) para describir la relación entre la variable de respuesta y el conjunto de variables explicativas.

Regresión logística

El objetivo de esta clase de modelos es el mismo que el que venimos persiguiendo desde el inicio del curso:

Obtener el modelo que presente el mejor ajuste (de la forma más parsimoniosa posible) para describir la relación entre la variable de respuesta y el conjunto de variables explicativas.

Como se dijo anteriormente, las dos grandes diferencias entre el modelo de *RLM* y el modelo de regresión logística (*RL*) son la naturaleza de la variable dependiente Y y la función de enlace.

- **Regresión lineal**: $\mathbb{E}(Y|\mathbf{X}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$
- **Regresión logística**: $logit\left(\mathbb{E}(Y|\mathbf{X})\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$

A partir de la formulación anterior:

$$
\begin{aligned}
Y \sim Ber(\pi(\mathbf{x})) \qquad & \text{Siendo } \pi(\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{x}) \\
\log\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) &= \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \\
\pi(\mathbf{x}) &= \frac{exp\{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k\}}{1+exp\{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k\}}
\end{aligned}
$$

A partir de la formulación anterior:

$$
\begin{aligned}
Y \sim Ber(\pi(\mathbf{x})) \qquad & \text{Siendo } \pi(\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{x}) \\
\log\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) &= \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \\
\pi(\mathbf{x}) &= \frac{exp\{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k\}}{1+exp\{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k\}}
\end{aligned}
$$

El cociente $\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}$ es conocido como *odds*.

A partir de la formulación anterior:

$$\begin{aligned} Y \sim Ber(\pi(\mathbf{x})) \qquad & \text{Siendo } \pi(\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{x}) \\ \log\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) &= \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \\ \pi(\mathbf{x}) &= \frac{exp\{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k\}}{1+exp\{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k\}} \end{aligned}$$

El cociente $\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}$ es conocido como *odds*.

Estas ecuaciones nos permiten observar que:

- El efecto de cada variable explicativa es lineal en el logaritmo de los *odds*.

A partir de la formulación anterior:

$$
\begin{aligned}
Y \sim Ber(\pi(\mathbf{x})) \qquad & \text{Siendo } \pi(\mathbf{x}) = \mathbb{P}(Y = 1|\mathbf{x}) \\
\log\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) &= \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \\
\pi(\mathbf{x}) &= \frac{exp\{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k\}}{1+exp\{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k\}}
\end{aligned}
$$

El cociente $\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}$ es conocido como *odds*.

Estas ecuaciones nos permiten observar que:

- El efecto de cada variable explicativa es lineal en el logaritmo de los *odds*.
- El efecto de cada variable explicativa es NO lineal sobre la media concicional de Y.

Regresión Logística (interpretación)

La interpretación de los coeficientes no es tan inmediata como en el GLM normal y requiere de una adecuada comprensión de los *odds*.

Regresión Logística (interpretación)

La interpretación de los coeficientes no es tan inmediata como en el GLM normal y requiere de una adecuada comprensión de los *odds*.

- Cuando $\pi(\mathbf{x}) \to 0$ entonces $\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} \to 0$
- Cuando $\pi(\mathbf{x}) \to 1$ entonces $\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} \to +\infty$

Regresión Logística (interpretación)

La interpretación de los coeficientes no es tan inmediata como en el GLM normal y requiere de una adecuada comprensión de los *odds*.

- Cuando $\pi(\mathbf{x}) \to 0$ entonces $\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} \to 0$
- Cuando $\pi(\mathbf{x}) \to 1$ entonces $\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} \to +\infty$

Así, diremos que eventos poco (muy) probables, tienen una *"chance"* baja (alta) de ocurrir.

Regresión Logística (interpretación)

La interpretación de los coeficientes no es tan inmediata como en el GLM normal y requiere de una adecuada comprensión de los *odds*.

- Cuando $\pi(\mathbf{x}) \to 0$ entonces $\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} \to 0$
- Cuando $\pi(\mathbf{x}) \to 1$ entonces $\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})} \to +\infty$

Así, diremos que eventos poco (muy) probables, tienen una *"chance"* baja (alta) de ocurrir.

El valor en el que la probabilidad de éxito de la Bernoulli vale $1/2$ corresponde al valor 1 en términos de *"chance"*.

Regresión logística (interpretación)

Para obtener la interpretación de los coeficientes, ralizaremos el mismo proceso llevado a cabo en la regresión lineal. Compararemos el valor de $g(\mu(\mathbf{x}))$ en x y $x+1$. Sin pérdida de generalidad, realizaremos esta comparación sobre dos valores de x_1 **dejando las demás variables constantes**.

Regresión logística (interpretación)

Para obtener la interpretación de los coeficientes, ralizaremos el mismo proceso llevado a cabo en la regresión lineal. Compararemos el valor de $g(\mu(\mathbf{x}))$ en x y $x+1$. Sin pérdida de generalidad, realizaremos esta comparación sobre dos valores de x_1 **dejando las demás variables constantes**.

- $\log(odds|x1 = x+1) = \beta_0 + \beta_1(x+1) + \ldots + \beta_k x_k$
- $\log(odds|x1 = x) = \beta_0 + \beta_1 x + \ldots + \beta_k x_k$

Regresión logística (interpretación)

Para obtener la interpretación de los coeficientes, ralizaremos el mismo proceso llevado a cabo en la regresión lineal. Compararemos el valor de $g(\mu(\mathbf{x}))$ en x y $x+1$. Sin pérdida de generalidad, realizaremos esta comparación sobre dos valores de x_1 **dejando las demás variables constantes**.

- $\log(odds|x1 = x+1) = \beta_0 + \beta_1(x+1) + \ldots + \beta_k x_k$
- $\log(odds|x1 = x) = \beta_0 + \beta_1 x + \ldots + \beta_k x_k$

Al restarlas:

$$\begin{aligned} \log(odds|x1 = x+1) - \log(odds|x1 = x) &= \beta_1 \\ \log\left(\frac{odds|x_1=x+1}{odds|x_1=x}\right) &= \beta_1 \\ \frac{odds|x_1=x+1}{odds|x_1=x} &= e^{\beta_1} \end{aligned}$$

Regresión logística (interpretación)

Para obtener la interpretación de los coeficientes, ralizaremos el mismo proceso llevado a cabo en la regresión lineal. Compararemos el valor de $g(\mu(\mathbf{x}))$ en x y $x+1$. Sin pérdida de generalidad, realizaremos esta comparación sobre dos valores de x_1 **dejando las demás variables constantes**.

- $\log(odds|x1 = x+1) = \beta_0 + \beta_1(x+1) + \ldots + \beta_k x_k$
- $\log(odds|x1 = x) = \beta_0 + \beta_1 x + \ldots + \beta_k x_k$

Al restarlas:

$$\begin{aligned} \log(odds|x1 = x+1) - \log(odds|x1 = x) &= \beta_1 \\ \log\left(\frac{odds|x_1=x+1}{odds|x_1=x}\right) &= \beta_1 \\ \frac{odds|x_1=x+1}{odds|x_1=x} &= e^{\beta_1} \end{aligned}$$

De esta forma estimamos el *odds-ratio* (OR) entre x_1 y Y.

Regresión logística (interpretación)

A partir de la última linea de la diapositiva anterior se puede establecer que:

$$odds|x_1 = x + 1 = e^{\beta_1} \times odds|x_1 = x$$

Regresión logística (interpretación)

A partir de la última linea de la diapositiva anterior se puede establecer que:

$$odds|x_1 = x + 1 = e^{\beta_1} \times odds|x_1 = x$$

Así vemos que cada variable tiene un aumento *multiplicativo* en el *odds*. Diremos que: dejando las demás variables constantes, un aumento de una unidad en x_1 aumenta el odds tanto como $\left(e^{\beta_1} - 1\right)$ %.

Regresión logística (interpretación)

A partir de la última linea de la diapositiva anterior se puede establecer que:

$$odds|x_1 = x + 1 = e^{\beta_1} \times odds|x_1 = x$$

Así vemos que cada variable tiene un aumento *multiplicativo* en el *odds*. Diremos que: dejando las demás variables constantes, un aumento de una unidad en x_1 aumenta el odds tanto como $\left(e^{\beta_1} - 1\right)$ %.

Un error muy común entre los usuarios de esta clase de modelos es decir que:

Según el modelo de regresión logística, la probabilidad de éxito aumenta e^{β} por cada unidad que aumente x

Regresión logística (interpretación)

A partir de la última linea de la diapositiva anterior se puede establecer que:

$$odds|x_1 = x + 1 = e^{\beta_1} \times odds|x_1 = x$$

Así vemos que cada variable tiene un aumento *multiplicativo* en el *odds*. Diremos que: dejando las demás variables constantes, un aumento de una unidad en x_1 aumenta el odds tanto como $\left(e^{\beta_1} - 1\right)$ %.

Un error muy común entre los usuarios de esta clase de modelos es decir que:

Según el modelo de regresión logística, la probabilidad de éxito aumenta e^{β} por cada unidad que aumente x

No confundamos probabilidad con *"chance"*.

Regresión logística (supuestos)

De manera indirecta, otra diferencia clave de la regresión logística respecto de la resgresión lineal es que en la regresión logística **NO** se asume homoscedasticidad.

Regresión logística (supuestos)

De manera indirecta, otra diferencia clave de la regresión logística respecto de la resgresión lineal es que en la regresión logística **NO** se asume homoscedasticidad.

Recordando como la caracterizamos previamente:

GLM Bernoulli (regresión logística)

- Las observaciones $Y_1, Y_2, \ldots, Y_n$ tienen distribución $Ber(\pi(\mathbf{x}))$.
- $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$.
- $\log\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) = \eta_i$

La distribución de Y es por naturaleza heteroscedástica ya que:

$$\begin{aligned} \mathbb{E}(Y) &= \pi(\mathbf{x}) \\ \mathbb{V}ar(Y) &= \pi(\mathbf{x})(1-\pi(\mathbf{x})) \end{aligned}$$

Regresión logística (supuestos)

De manera indirecta, otra diferencia clave de la regresión logística respecto de la resgresión lineal es que en la regresión logística **NO** se asume homoscedasticidad.

Recordando como la caracterizamos previamente:

GLM Bernoulli (regresión logística)

- Las observaciones $Y_1, Y_2, \ldots, Y_n$ tienen distribución $Ber(\pi(\mathbf{x}))$.
- $\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$.
- $\log\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) = \eta_i$

La distribución de Y es por naturaleza heteroscedástica ya que:

$$\begin{aligned} \mathbb{E}(Y) &= \pi(\mathbf{x}) \\ \mathbb{V}ar(Y) &= \pi(\mathbf{x})(1-\pi(\mathbf{x})) \\ \mathbb{V}ar(Y) &= \mathbb{E}(Y)(1-\mathbb{E}(Y)) \end{aligned}$$

Ejemplo

Se cuenta con una muestra de 100 personas, de las cuales se sabe su edad y su diagnóstico de enfermedad coronaria (enfermo o sano).

Ejemplo

Se cuenta con una muestra de 100 personas, de las cuales se sabe su edad y su diagnóstico de enfermedad coronaria (enfermo o sano).

Vayamos al  y veamos como ajustar nuestro primer *GLM* y que habría pasado si hubiésemos ajustado un modelo de regresión lineal.

En la próxima clase

La próxima clase:

- Definiremos la *devianza* como medida de bondad de ajuste.
- Veremos como se estiman los parámetros de un GLM no normal.
- Evaluaremos las predicciones de estos modelos, haciendo hincapié en la regresión logística.

Bibliografía

- Carmona, Francesc (2003). *Modelos Lineales (notas de curso)*. Departament d'Estadística.
- Farraway, Julian (2014). *Linear Models with R, second edition*. Chapman Hall/CRC.
- McCullagh, P. y J.A. Nelder (1983). *Generalized Linear Models*. Chapman Hall/CRC.
- Rencher, Alvin y Bruce Schaalje (2008). *Linear Models in Statistics, second edition*. John Wiley Sons, Inc.

¿Preguntas?

Muchas Gracias