UNAP
FINESI

INGENIERIA ESTADISTICA U.N.A. INFORMATICA

IV Semestre

Modelos Discretos

Modelos Lineales Generalizados (GLM)

Mtr. Alcides Ramos Calcina

Modelos Lineales Generalizados (GLM)

Mtr. Alcides Ramos Calcina

Introducción

- El Modelo Lineal Generalizado (MLG), que tratamos en este curso, es la extensión natural del Modelo Lineal clásico.

- Expuesto por Nelder y Wedderburn (1972), convirtiéndose en una solución especialmente adecuada para modelos de dependencia con datos no métricos.

- En el área de ciencias sociales es frecuente trabajar atributos, actitudes o conductas que, se miden de forma no métrica (discreta, nominal u ordinal), por tanto, no se ajusta, al Modelo Lineal clásico, incumpliendo los supuestos de linealidad y normalidad.

- Por ejemplo, la clasificación binaria de apto - no apto, es una situación que requieren de modelos que trabajen con datos dicotómicos, ordinales, categóricos o de elecciones discretas, es decir, de modelos de probabilidad de un evento.

1. Modelo Estadístico

- Un modelo pretende explicar la variación de una respuesta a partir de la relación conjunta de dos fuentes de variabilidad, una de carácter determinista y otra aleatoria, lo que responde a la expresión:

$$\textit{Respuesta} = \textit{componente sistemático} + \textit{componente aleatorio}$$

- Otros autores toman la siguiente expresión:

$$DATOS = MODELO + ERROR$$

Donde:

MODELO: está asociado a la parte sistemática. es la función que se introduce con objeto de explicar los datos.

DATOS: corresponde a las observaciones que se quieren analizar (la variable de respuesta o variable dependiente).

ERROR: es la que contiene la discrepancia o falta de ajuste entre *Datos* y *Modelo*.

1. Modelo Estadístico

1. Criterios para la construcción, formulación y ajuste de modelos

01

Criterio estadístico o principio de bondad de ajuste:

La inclusión de parámetros en el MODELO en beneficio de una mejor representación de los DATOS con la correspondiente disminución del ERROR

02

Criterio lógico o principio de parsimonia:

la selección de los parámetros que formen parte del modelo de tal modo que éste se convierta en una representación simple y sobria de la realidad.

03

Criterio sustantivo o integración teórica:

del modelo en la red conceptual que lo generó.

1. Modelo Estadístico

2. Etapas de la construcción de modelos

iii) Selección del modelo

valorando si el nivel de discrepancia entre los datos observados y los datos ajustados es suficientemente bajo como para optar por el modelo

ii) Estimación de parámetros

calculando el valor de los coeficientes del modelo examinado a partir del conjunto de datos observados

iv) Evaluación del modelo

examinando las observaciones individuales, los datos influyentes y los datos anómalos (outliers), así como comprobando los supuestos de normalidad, linealidad, homoscedasticidad e independencia

i) Especificación del modelo teórico

determinando qué variables son de interés, así como cuáles son las relaciones entre ellas..

v) Interpretación del modelo

Esta fase conlleva una explicación detallada de los parámetros del modelo para comprobar si se cumplen los criterios estadístico, lógico y sustantivo

2. Modelo Lineal (LM)

- La fórmula general del Modelo Lineal es:

$$Y = f(X) + g(\varepsilon)$$

donde toda observación sobre la variable de respuesta es la suma de:

a) los efectos de un grupo de factores o componentes sistemáticos, $f(X)$, que implican un conjunto de parámetros de una población y un conjunto de variables independientes relevantes medidas sobre cada uno de los sujetos con los que se trabaja.

a) función $g(\varepsilon)$, que representa el efecto de los componentes aleatorios y es resultado de una o más distribuciones de probabilidad dependientes de un pequeño número de parámetros.

2. Modelo Lineal (LM)

- Se habla, entonces, de un Modelo Lineal de primer orden para k variables explicativas y $k + 1$ parámetros si el modelo es lineal en sus parámetros y en sus variables explicativas, respondiendo a la siguiente fórmula general:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

$$Y = \beta_0 + \sum_{j=1}^{k} \beta_j X_j + \varepsilon$$

- Si el modelo es lineal en sus parámetros, pero no en las variables explicativas sería un Modelo Lineal de m-ésimo orden (cuadrático, cúbico, etc.) con km variables independientes y $km + 1$ parámetros. Su formulación es:

$$Y = \beta_0 + \sum_{j=1}^{k} \beta_j X_j + \sum_{j=1}^{k} \beta_{j1} X_{j1}^{2} + \cdots + \sum_{j=1}^{k} \beta_{j1} X_{j1}^{m} + \varepsilon$$

2. Modelo Lineal (LM)

2.1. Limitaciones en el Modelo Lineal

- Sea cual sea el tipo de datos que pretendemos analizar, para una modelización adecuada siempre es básico preguntar y contestar una serie de cuestiones:

Así mismo, se debe corroborar que las respuestas verifiquen:

- Normalidad
- Homogeneidad de varianzas
- Linealidad (aditividad) de los efectos sistemáticos

2. Modelo Lineal (LM)

- Se tiene el Modelo Lineal normal en su forma matricial: $Y = X\beta + \varepsilon$

Donde $\varepsilon = N\left(0, \sigma^2 I\right)$, de forma que,

$$E\left(Y\right) - \mu = X\beta \qquad\qquad Var\left(Y_i\right) = \sigma^2$$

- Sin embargo, el mundo de los datos no es siempre NORMAL.

- Cuando no se verifican las hipótesis del modelo lineal, luego de intentar algunas transformaciones de las variables, surgen algunos inconvenientes:
  - ✓ Algunas transformaciones no están definidas en las fronteras del espacio muestral (como la logit).
  - ✓ No es siempre directa la interpretación de la variable transformada.
  - ✓ Generalmente no existe una técnica única que corrija a la vez todos los defectos.

2. Modelo Lineal (LM)

2.2. Generalización del Modelo Lineal

- Los Modelos Lineales Generalizados (GLM) son una alternativa a transformaciones de la respuesta, justificadas por:
  - ❖ La falta de linealidad
  - ❖ La falta de homogeneidad de la varianza.

- Las hipótesis básicas de un modelo lineal generalizado (GLM) son:
  - ❑ Independencia entre las respuestas
  - ❑ La "respuesta media" cambia con las condiciones, pero no la "forma funcional" de la distribución.
  - ❑ La respuesta media, o alguna transformación de ella, cambia de modo lineal cuando las condiciones cambian.

3. Modelo Lineal Generalizado (GLM)

3.1. Definición de un GLM

- Los modelos lineales generalizados (GLM) amplían el concepto del modelo de regresión lineal. El modelo lineal supone que,

$$E\left(Y/X\right) = X'\beta$$

- Su equivalente, $Y = X'\beta + \varepsilon$

- Desafortunadamente, la restricción a la linealidad no puede tener en cuenta una variedad de situaciones prácticas.

3. Modelo Lineal Generalizado (GLM)

- En los GLM se demuestra que, si la distribución de la variable dependiente Y es un miembro de la familia exponencial, entonces la clase de modelos que conecta la esperanza de Y a una combinación lineal de las variables $X'\beta$ puede tratarse de manera unificada.

- Por consiguiente, denotamos la función que relaciona:

$$\mu = E(Y \mid X) \quad \text{y} \quad \eta = X'\beta$$

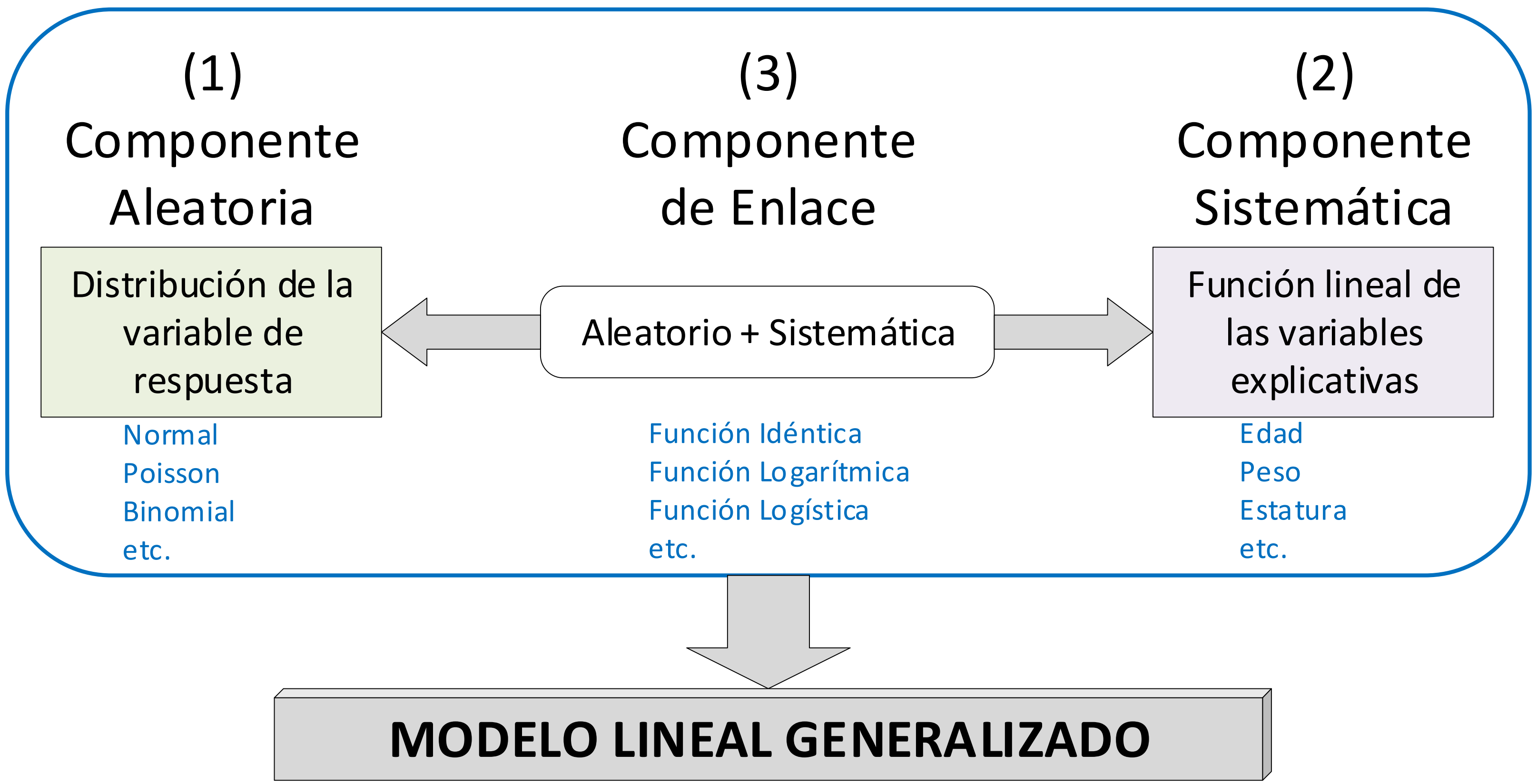
por

$$\eta = G(\mu) \quad \text{o} \quad E(Y \mid X) = G^{-1}(X'\beta)$$

3. Modelo Lineal Generalizado (GLM)

3.2. Componentes de un GLM

- Un modelo lineal generalizado tiene tres componentes básicos:



3. Modelo Lineal Generalizado (GLM)

a) Componente aleatoria

- En muchas aplicaciones, las observaciones de Y son binarias y se identifican como éxito y fracaso, y se modeliza como una distribución binomial.

- En otras ocasiones cada observación es un recuento, con lo que se puede asignar a Y una distribución de Poisson o una distribución binomial negativa.

- Esta componente, identifica la distribución de probabilidad de la variable dependiente. Esto es, cada variable muestral Y_i tiene función de densidad de la forma:

$$f\left(y_i,\theta_i,\phi\right)=\exp\left[\frac{y_i\theta_i-b\left(\theta_i\right)}{\phi}+c\left(y_i,\phi\right)\right]=G\left(y_i,\phi\right)\exp\left[\frac{y_i\theta_i-b\left(\theta_i\right)}{\phi}\right]$$

3. Modelo Lineal Generalizado (GLM)

b) Componente sistemática

- Especifica una función lineal η de los valores fijados $x_{1i}, \cdots, x_{ki}$ de las variables explicativas $X_1, X_2, \cdots, X_k$ dada por:

$$\eta_i := \delta + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \qquad i = 1, 2, 3, \ldots, n$$

donde los β_k son los llamados parámetros del modelo lineal generalizado, incluyendo el llamado intercepto como $\delta = \beta_0$, siendo

3. Modelo Lineal Generalizado (GLM)

- Si se reúnen los valores observados de las variables explicativas en la llamada matriz de diseño:

$$C = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

de tamaño $n\times(1+k)$, los parámetros del modelo en el vector

$$\alpha = \left(\delta, \beta_1, \beta_2, \cdots, \beta_k\right)'$$

entonces, la ecuación lineal puede ser escrita en forma vectorial como $\eta = C.\alpha$

3. Modelo Lineal Generalizado (GLM)

c) Función link o de enlace

- Sea μ_i la esperanza condicional de Y_i dada la condición $x_{i1},\cdots,x_{ik}$, es decir, $\mu_i := E\left(Y \mid x_{i1},x_{i2},\cdots,x_{ik}\right)$, $i = 1, \cdots, n$. Entonces, este enlace está dado por una llamada función de enlace:

$$g\left(\mu_i\right) = \theta_i$$

en cuyo caso resultan $\theta_i = \eta_i$, y el enlace está descrito por la expresión

$$\theta_i = \delta + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

4. Tipos de GLMs

4.1. Lineal

- Supongamos que la variable Y_i, $i = 1, \cdots$, n está normalmente distribuida con esperanza μ_i y varianza σ^2. La función de densidad en los valores y_i viene dada por:

$$f(y_i,\theta_i,\phi) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}(y_i-\mu_i)^2\right]$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{y_i^2}{2\sigma^2}\right].\exp\left[\frac{y_i\mu_i-\frac{\mu_i^2}{2}}{\sigma^2}\right]$$

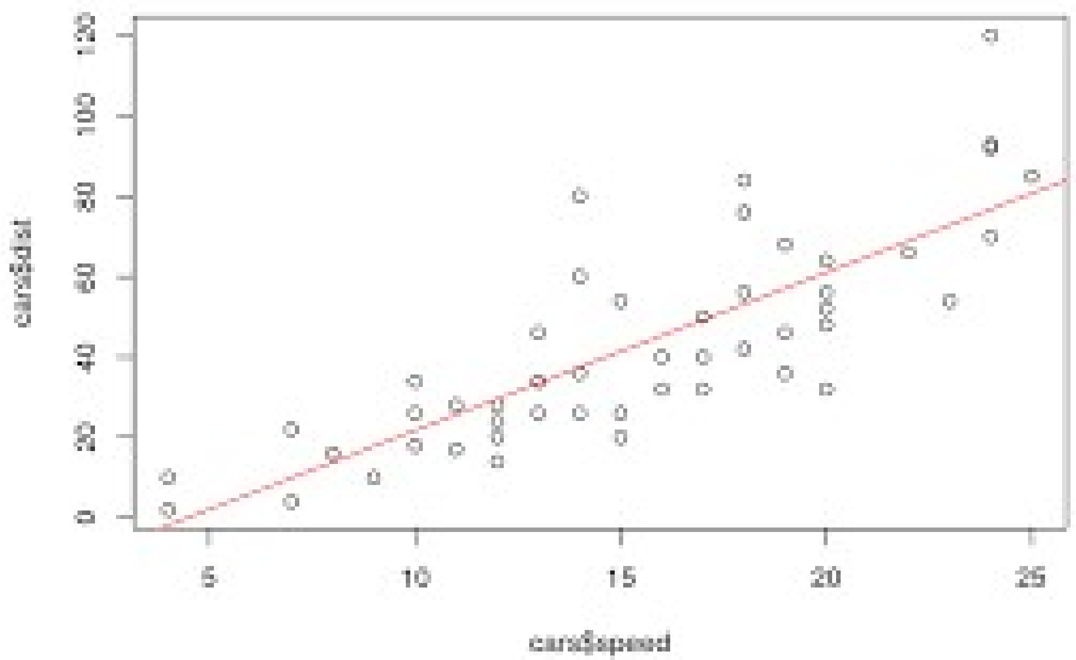
- Aquí, $\theta_i = \mu_i$ son los parámetros naturales y se tiene $\phi = \sigma^2$ como parámetro de dispersión. Además,

$$b(\theta_i) = \frac{\mu_i^2}{2} \qquad G(y_i,\phi) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{y_i^2}{2\sigma^2}\right]$$

4. Tipos de GLMs

- El enlace canónico está dado por la función identidad $g(\mu_i) = \mu_i$. Los GLMs que usan el enlace identidad son llamados modelos lineales.

- En la siguiente figura se muestra una representación gráfica de un modelo de regresión lineal.



4. Tipos de GLMs

4.2. Log-lineal

- Supongamos que la Y_i, $i = 1, \cdots, n$, es una variable de Poisson con parámetro $\lambda_i > 0$. Su función de densidad viene dada por

$$f(y_i, \lambda_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i\,!}$$

- para todo $y_i \in \{0, 1, 2, \cdots,\}$. Escribiendo esta función en la forma:

$$f(y_i, \lambda_i) = \frac{\exp\left[y_i \ln \lambda_i - \lambda_i\right]}{y_i\,!} = \exp\left[y_i \ln \lambda_i - \lambda_i - \ln(y_i\,!)\right]$$

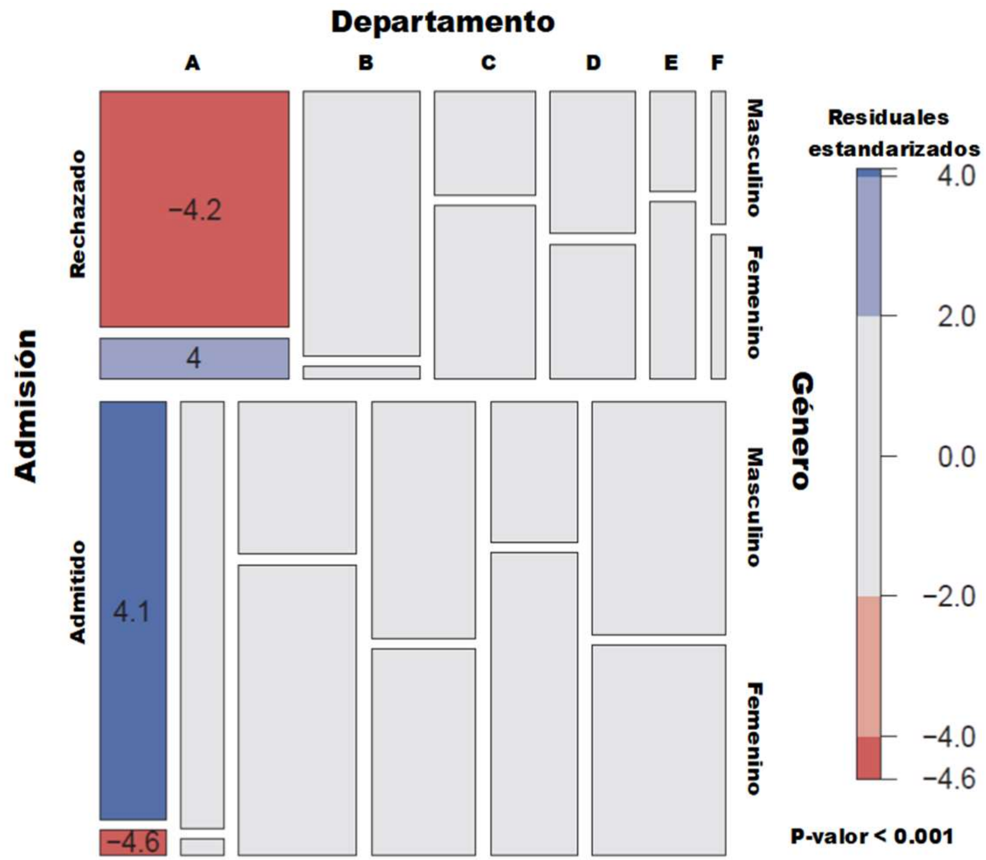
- Aquí, $\theta_i = \ln\lambda_i$ son los parámetros naturales y el parámetro de dispersión es $\phi = 1$. Además,

$$b(\theta_i) = e^{\theta_i} = \lambda_i \qquad\qquad G(y_i, \phi) = \frac{1}{y_i\,!}$$

4. Tipos de GLMs

- El enlace canónico es $g(\lambda_i) = \ln \lambda_i$. Los GLMs que usan el enlace log son llamados modelos loglineales.

- En la figura se muestra una representación gráfica (en mosaico) de un modelo de regresión loglineal, que muestra los residuos estandarizados de las contribuciones de las celdas en el valor del estadístico de prueba correspondiente.



4. Tipos de GLMs

4.3. Logístico (Y es Bernoulli)

- Muchas variables categóricas tienen únicamente dos categorías. La observación para cada caso puede ser clasificada como éxito o fracaso.

- En este caso, la variable Y_i tiene distribución de Bernoulli con parámetro p_i. Su función de densidad es

$$f(y_i, p_i) = \exp\left[y_i \ln\left(\frac{p_i}{1-p_i}\right) - \ln\left(\frac{1}{1-p_i}\right) \right]$$

para todo $0 < p_i < 1$ y $y_i \in \{0,1\}$. Los parámetros naturales son $\theta_i = \ln\left(\frac{p_i}{1-p_i}\right)$

y el parámetro de dispersión es $\phi = 1$.

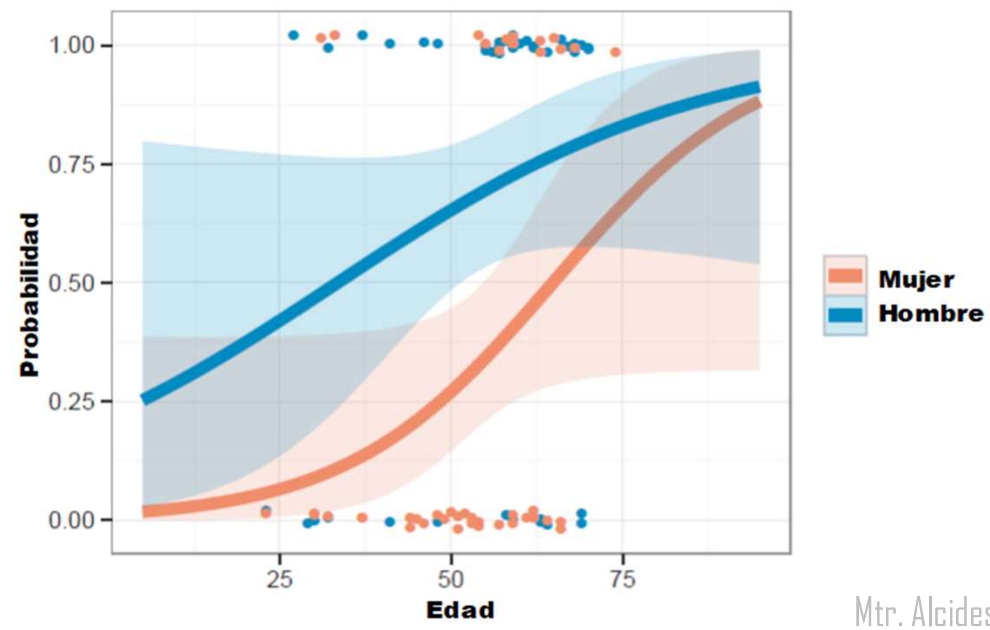
4. Tipos de GLMs

- Además, $b(\theta_i) = \ln\left(1 + e^{\theta_i}\right) = \ln\left(\frac{1}{1-p_i}\right)$, $G(y_i, \phi) = 1$

- El enlace canónico $g(p_i) = \ln\left(\frac{p_i}{1-p_i}\right)$

  es llamado el logit de p_i. Los GLMs que usan el enlace logit son llamados modelos logit o logísticos.

- En la figura se muestra un ejemplo de un gráfico logit condicional. En él se muestran los puntos separados de las edades y las curvas ajustadas, ambas estratificadas por género.



4. Tipos de GLMs

4.4. Logístico (Y es Binomial)

- La variable Y_i tiene distribución Binomial con parámetros n y p_i. Su función de densidad es

$$f(y_i, n, p_i) = \binom{n}{y_i} p^{y_i} (1-p)^{n-y_i}$$
$$= \binom{n}{y_i} \exp\left[y_i \ln\left(\frac{p_i}{1-p_i}\right) - \ln\left(\frac{1}{1-p_i}\right)^n \right]$$

para todo $0 < p_i < 1$ y $y_i \in \{0, 1, \ldots, n\}$. Los parámetros naturales son $\theta_i = \ln\left(\frac{p_i}{1-p_i}\right)$ y el parámetro de dispersión es $\phi = 1$. Además,

$$b(\theta_i) = \ln\left(1+e^{\theta_i}\right)^n = \ln\left(\frac{1}{1-p_i}\right)^n, \qquad G(y_i, \phi) = \binom{n}{y_i}$$

4. Tipos de GLMs

- El enlace canónico

$$g(p_i) = \ln\left(\frac{p_i}{1-p_i}\right)$$

es llamado el logit de p_i.

Los GLMs que usan el enlace logit son llamados modelos logit o logísticos.

FINESI

Modelos Discretos

IV Semestre

GRACIAS

https://aulavirtual2.unap.edu.pe/