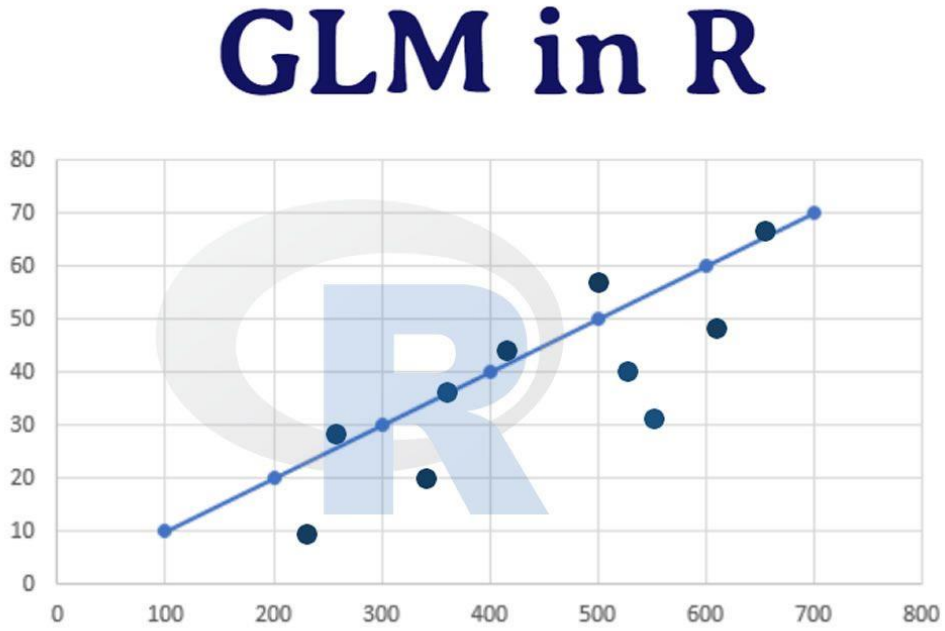
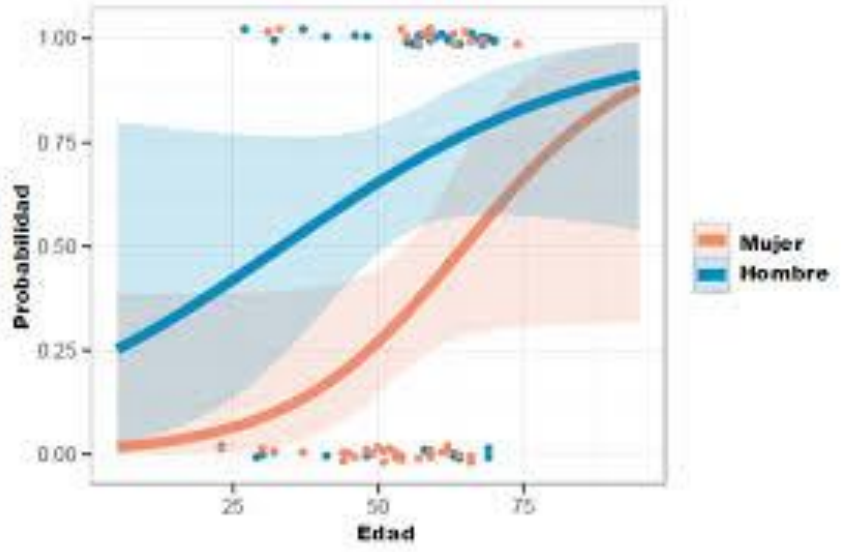




Modelos Lineales Generalizados

Óscar Centeno Mora

Preámbulo

- Hasta ahora hemos tratado, diagnosticado, remediado o “machado” los modelos de la RLM para obtener un modelo que regresión en la predicción de un acontecimiento.

- Sin embargo, hemos hecho un supuesto que, dentro de un plano práctico, muy pocas veces se cumple. Para la RLM hicimos la siguiente suposición:

$$\varepsilon \sim N(0, \sigma^2 I)$$

- Nos hacemos las preguntas: ¿si trabajará con conteos, con variables que poseen una asimetría, con variables categóricas ?

- En los casos anteriores, el supuesto anterior no es sostenible, y debemos optar por otros métodos de estimación más generales, este es, los modelos lineales generalizados (GLM).

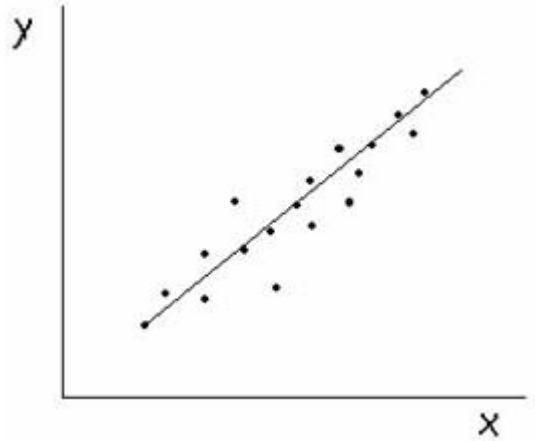
Preámbulo

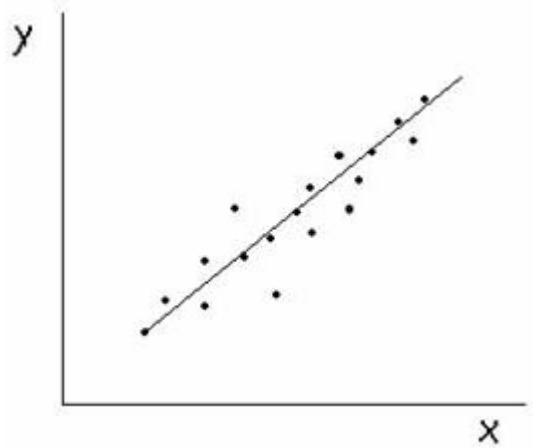
Regresión bivariada

Una variable dependiente (Y)
Una variable independiente (X)

$$Y \sim X$$

$$\begin{bmatrix} Y1 \\ Y2 \\ \cdots \\ Y_n \end{bmatrix} \sim \begin{bmatrix} X1 \\ X2 \\ \cdots \\ X_n \end{bmatrix}$$





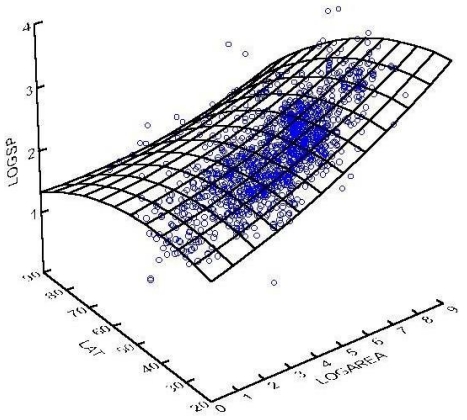
Regresión multivariada

Una variable dependiente
Dos o más variables independientes

$$Y \sim X$$

$$\begin{bmatrix} Y1 \\ Y2 \\ \cdots \\ Y_n \end{bmatrix} \sim \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

¿podemos aplicar otro tipo de métodos de predicción para la explicación de Y?



Índice

1 Introducción

2 Componente de los GLM

3 Familia de los GLM

4 GLM para datos binarios

5 Otras aproximaciones

Índice

1 Introducción

Introducción

- Los modelos lineales generalizados son una extensión de los modelos lineales para el caso de que la distribución condicional de la variable respuesta no sea normal (por ejemplo discreta: Bernouilli, Binomial, Poisson, …).

- En los modelo lineales se supone que:

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = X\beta = \hat{Y}$$

En los modelos lineales generalizados se introduce una función invertible *g*, denominada función enlace (o link):

$$g\big(E(Y|X)\big) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = g^{-1}(X\beta)$$

Introducción

- Un modelo lineal generalizado (GLM) es una generalización flexible de la regresión lineal ordinaria que permite variables de respuesta que tienen modelos de distribución de errores distintos de una distribución normal.

- El GLM generaliza la regresión lineal al permitir que el modelo lineal esté relacionado con la variable de respuesta a través de una función de enlace y al permitir que la magnitud de la varianza de cada medición sea una función de su valor predicho (¿? los momentos).

- ¿Más específicamente, en dónde radica la diferencia de los MLR y los GLM?

- Los modelos lineales hacen siempre referencia a modelos de regresión lineal normal con una variable de respuesta continua, asumen además que los residuos / errores siguen una distribución normal. En el modelo lineal generalizado, por otro lado, permite que los residuos tengan otras distribuciones de la familia exponencial de distribuciones. Esto permite una mayor flexibilidad de optar por algo aparte de la distribución normal en los residuos.

Índice

1 Introducción

2 Componente de los GLM

Componentes de los GLM

Un modelo lineal generalizado tiene tres componentes básicos:

- **Componente aleatoria**: identifica la variable respuesta y su distribución de probabilidad.
- **Componente sistemático**: especifica las variables explicativas (independientes o predictoras) utilizadas en la función predictora lineal.
- **Función link o de enlace**: es una función del valor esperado de Y , E(Y), como una combinación lineal de las variables predictoras.
- Veamos que es cada componentes.

Componentes de los GLM

- **Componente aleatorio**

El componente aleatoria de un GLM consiste en una variable aleatoria Y con observaciones independientes $(y_1, \dots, y_N)$.

En muchas aplicaciones, las observaciones de Y son binarias y se identifican como éxito y fracaso. Aunque de modo más general, cada Y_i indica el número de éxitos de entre un número fijo de ensayos, 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. Finalmente, si las observaciones son continuas se puede asumir para Y una distribución normal.

Todos estos modelos se pueden incluir dentro de la llamada familia exponencial de distribuciones

$$f(y_i|\theta_i) = a(\theta_i) * b(y_i) * \exp[y_i Q(\theta_i)]$$

de modo que $Q(\theta)$ recibe el nombre de parámetro natural.

Componentes de los GLM

Componente sistemático

La componente sistemático de un GLM especifica las variables explicativas, que entran en forma de efectos fijos en un modelo lineal, es decir, las variables x_j se relacionan mediante:

$$\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Esta combinación lineal de variables explicativas se denomina predictor lineal. Alternativamente, se puede expresar como un vector $(\eta_1, \ldots, \eta_N)$ tal que:

$$\eta_i = \sum_j \beta_j x_{ij}$$

Componentes de los GLM

Función de enlace

Se denota el valor esperado de Y como μ = E(Y), entonces la función link especifica una función $g(\cdot)$ que relaciona μ con el predictor lineal como

$$g(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Así, la función link $g(\cdot)$ relaciona las componentes aleatoria y sistemáticos. De este modo, para $i = 1, \ldots, N$,

$$\mu_i = E(Y_i)$$

$$\eta_i = g(\mu_i) = \sum_j \beta_j x_{ij}$$

Los modelos de regresión lineal típicos para respuestas continuas son un caso particular de los GLM.

Estos modelos generalizan la regresión ordinaria de dos modos: permitiendo que Y tenga distribuciones diferentes a la normal y, por otro lado, incluyendo distintas funciones link de la media.

Índice

1 Introducción

2 Componente de los GLM

3 Familia de los GLM

Familia de los GLM

- ¿Podemos entonces especificar diversas distribuciones para la variable Y y sus residuos?
- La posibilidad de especificar una distribución que no sea la normal y una función de enlace que no sea la identidad es la principal mejora que aporta el modelo lineal generalizado respecto al modelo lineal general. Hay muchas combinaciones posibles de distribución y función de enlace, varias de las cuales pueden ser adecuadas para un determinado conjunto de datos, por lo que su elección puede estar guiada por consideraciones teóricas a priori y por las combinaciones que parezcan funcionar mejor.
- **Binomial.** Esta distribución es adecuada únicamente para las variables que representan una respuesta binaria o un número de eventos.
- **Gamma.** Esta distribución es adecuada para las variables con valores de escala positivos que se desvían hacia valores positivos más grandes. Si un valor de datos es menor o igual que 0 o es un valor perdido, el correspondiente caso no se utilizará en el análisis.
- **De Gauss inversa.** Esta distribución es adecuada para las variables con valores de escala positivos que se desvían hacia valores positivos más grandes. Si un valor de datos es menor o igual que 0 o es un valor perdido, el correspondiente caso no se utilizará en el análisis.
- **Binomial negativa.** Esta distribución considera el número de intentos necesarios para lograr *k* éxitos y es adecuada para variables que tengan valores enteros que no sean negativos. Si un valor de datos no es entero, es menor que 0 o es un valor perdido, el correspondiente caso no se utilizará en el análisis. El valor del parámetro auxiliar de la distribución binomial negativa puede ser cualquier número mayor o igual que 0; se puede establecer en un valor fijo o dejar que lo estime el procedimiento. Cuando el parámetro auxiliar se establece en 0, utilizar esta distribución equivale a utilizar la distribución de Poisson.

Familia de los GLM

- **Normal.** Es adecuada para variables de escala cuyos valores adoptan una distribución simétrica con forma de campana en torno a un valor central (la media). La variable dependiente debe ser numérica.
- **Poisson.** Esta distribución considera el número de ocurrencias de un evento de interés en un período fijo de tiempo y es apropiada para variables que tengan valores enteros que no sean negativos. Si un valor de datos no es entero, es menor que 0 o es un valor perdido, el correspondiente caso no se utilizará en el análisis.
- **Tweedie.** Esta distribución es adecuada para variables que puedan representarse mediante mezclas de Poisson de distribuciones gamma; la distribución es una "mezcla" en el sentido de que combina las propiedades de distribuciones continuas (toma valores reales no negativos) y discretas (masa de probabilidad positiva en un único valor, 0). La variable dependiente debe ser numérica y los valores de los datos deben ser iguales o mayores que cero. Si un valor de datos es menor que 0 o es un valor perdido, el correspondiente caso no se utilizará en el análisis. El valor fijo del parámetro de la distribución de Tweedie puede ser cualquier número mayor que uno y menor que dos.
- **Multinomial.** Esta distribución es adecuada para variables que representan una respuesta ordinal. La variable dependiente puede ser numérica o de cadena, y debe tener como mínimo dos valores válidos distintos de los datos.

Familia de los GLM

A partir de las familias, surgen entonces también distintas formas de enlazas la regresión a los datos:

- **Identidad.** $f(x)=x$. No se transforma la variable dependiente. Este enlace se puede utilizar con cualquier distribución.
- **Log-log complementario.** $f(x)=\log(-\log(1-x))$. Es apropiada únicamente para la distribución binomial.
- **Cauchit acumulada.** $f(x) = \tan(\pi\ (x - 0.5))$, aplicada a la probabilidad acumulada de cada categoría de la respuesta. Es apropiada únicamente para la distribución multinomial.
- **Log-log complementario acumulado.** $f(x)=\ln(-\ln(1-x))$, aplicada a la probabilidad acumulada de cada categoría de la respuesta. Es apropiada únicamente para la distribución multinomial.
- **Logit acumulado.** $f(x)=\ln(x\ /\ (1-x))$, aplicada a la probabilidad acumulada de cada categoría de la respuesta. Es apropiada únicamente para la distribución multinomial.
- **Log-log negativo acumulado.** $f(x)=-\ln(-\ln(x))$, aplicada a la probabilidad acumulada de cada categoría de la respuesta. Es apropiada únicamente para la distribución multinomial.
- **Probit acumulada.** $f(x)=\Phi^{-1}(x)$, aplicada a la probabilidad acumulada de cada categoría de la respuesta, donde Φ^{-1} es la función de distribución acumulada normal estándar inversa. Es apropiada únicamente para la distribución multinomial.

Familia de los GLM

- **Logaritmo.** $f(x)=\log(x)$. Este enlace se puede utilizar con cualquier distribución.
- **Complemento log.** $f(x)=\log(1-x)$. Es apropiada únicamente para la distribución binomial.
- **Logit.** $f(x)=\log(x / (1-x))$. Es apropiada únicamente para la distribución binomial.
- **Binomial negativa.** $f(x)=\log(x / (x+k^{-1}))$, donde k es el parámetro auxiliar de la distribución binomial negativa. Es apropiada únicamente para la distribución binomial negativa.
- **Log-log negativo.** $f(x)=-\log(-\log(x))$. Es apropiada únicamente para la distribución binomial.
- **Poder de probabilidad.** $f(x)=[(x/(1-x))^{\alpha}-1]/\alpha$, si $\alpha \neq 0$. $f(x)=\log(x)$, si $\alpha=0$. α es la especificación de número necesaria y debe ser un número real. Es apropiada únicamente para la distribución binomial.
- **Probit.** $f(x)=\Phi^{-1}(x)$, donde Φ^{-1} es la función de distribución acumulada normal estándar inversa. Es apropiada únicamente para la distribución binomial.
- **Potencia.** $f(x)=x^{\alpha}$, si $\alpha \neq 0$. $f(x)=\log(x)$, if $\alpha=0$. α es la especificación de número necesaria y debe ser un número real. Este enlace se puede utilizar con cualquier distribución.

Familia de los GLM

Table 1: Function $\Psi(\mathbf{x}_i'\boldsymbol{\beta})$ of Generalized Linear Model

Family	Link	Mean Function	$\Psi(\mathbf{x}_i'\boldsymbol{\beta})$
gaussian	identity	$\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$	$1/\sigma^2$
binomial	logit	$\mu_i = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})}$	$\mu_i(1-\mu_i)$
binomial	probit	$\mu_i = \Phi(\mathbf{x}_i'\boldsymbol{\beta})$	$\frac{\phi(\mathbf{x}_i'\boldsymbol{\beta})^2}{\Phi(\mathbf{x}_i'\boldsymbol{\beta})(1-\Phi(\mathbf{x}_i'\boldsymbol{\beta}))}$
binomial	cloglog	$\mu_i = 1-\exp(-\exp(\mathbf{x}_i'\boldsymbol{\beta}))$	$\frac{1-\mu_i}{\mu_i}[\log(1-\mu_i)]^2$
poisson	log	$\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$	μ_i
poisson	identity	$\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$	$1/\mu_i$
poisson	sqrt	$\mu_i = (\mathbf{x}_i'\boldsymbol{\beta})^2$	4
gamma	inverse	$\mu_i = (\mathbf{x}_i'\boldsymbol{\beta})^{-1}$	$a\mu_i^2$
gamma	identity	$\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$	a/μ_i^2
gamma	log	$\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$	a
inverse gaussian	inverse squared	$\mu_i = (\mathbf{x}_i'\boldsymbol{\beta})^{-1/2}$	$\lambda\mu_i^3/4$

Índice

1 Introducción

2 Componente de los GLM

3 Familia de los GLM

4 GLM para datos binarios

GLM para datos binarios

- En muchos casos las respuestas tienen solo dos categorías del tipo SI/NO, FALSO/VERDADERO, HOMBRE/MUJER, PAGO/DEUDA, etc…

- Se puede definir una variable Y que tome dos posibles valores 1 (éxito) y 0 (fracaso), es decir, una variable que se aproxima a una distrinución binomiral del tipo: $Y \sim Bin(1, \pi)$.

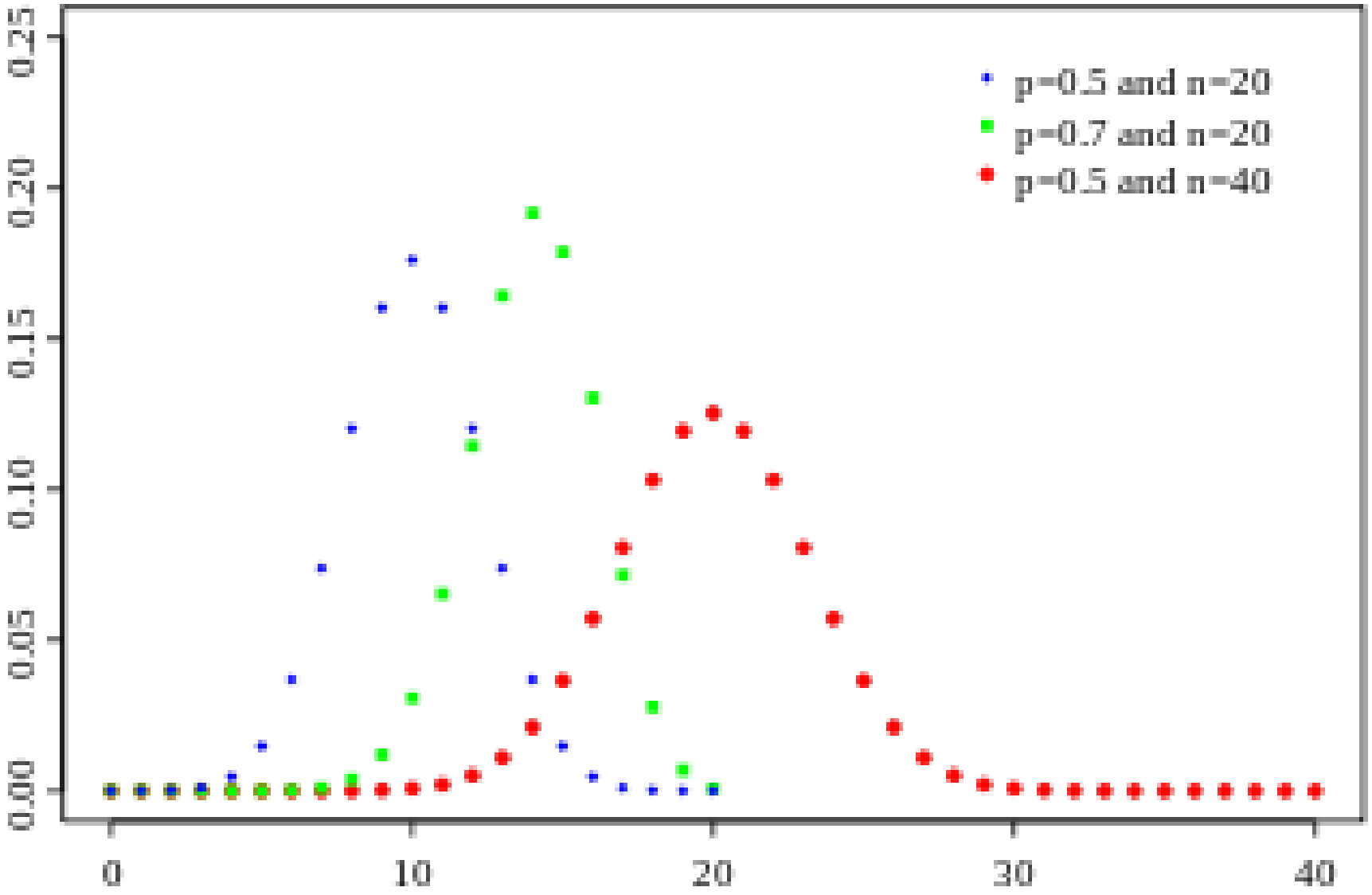
- Una distribución binomial se presenta matemáticamente:

$$P_x = \binom{n}{x} p^x q^{n-x}$$

Donde:

n: $número\ de\ ensayos\ o\ experimentos$
x: $número\ de\ éxitos$

GLM para datos binarios



GLM para datos binarios

- En el caso de un GLM para datos binarios:

$$f(y|\pi) = \pi^{y}\,(1-\pi)^{1-y}$$

$$f(y|\pi)' = (1-\pi)\left(\frac{\pi}{1-\pi}\right)^{y}$$

$$f(y|\pi)' = (1-\pi)\exp\left[y log\left(\frac{\pi}{1-\pi}\right)\right]$$

Con $y = [0,1]$

Dependiente como se vaya a definir el parámetro natural $Q(\pi)$, se podría optar por una modelo logístico, un probit, o un cloglog. Dos dos primeros serán los expuestos más adelante.

Índice

1 Introducción

2 Componente de los GLM

3 Familia de los GLM

4 GLM para datos binarios

5 Otras aproximaciones

Otras aproximaciones

- Más allá de la RLM, y de los GLM, aún podemos hablar de otros aproximaciones en el contexto de las regresiones.

- Existen los modelos no lineals.
- No paramétricos.
- Por capas, o las estimaciones de Deep learning.
- Modelos fraccionados.
- Por árboles.
- Etc, etc., etc.,

- Pero debemos, obligatoriamente, hacernos la siguiente pregunta…

Otras aproximaciones

¿Hay una aproximación mejor que las otras…

Conclusión

El capítulo presentó los fundamentos de los GLM.

Los GLM poseen como componentes una parte aleatoria, sistemática, y una función de enlace.

Dentro de los GLM hay funciones de distribuciones y funciones de enlace, los cuales lo hacen una opción mucho más basta que la RLM.

El próximo capítulo presente los modelos con variables dicotómicos: el probit y el logit.


The End