Modelos dicotómicos: el logit y el probit

- **Logistic Regression**:
  - Logit [P(x)] = $\log\left(\frac{P(x)}{1-P(x)}\right) = \alpha + \beta x$, $P(x) = \frac{\exp(\alpha + \beta x)}{1+\exp(\alpha + \beta x)} = \frac{e^{\alpha + \beta x}}{1+e^{\alpha + \beta x}} = \frac{1}{1+e^{-(\alpha + \beta x)}}$
- **Probit Analysis**:
  - Probit [P(x)] = $\alpha + \beta x$, $P(x) = \phi(\alpha + \beta x)$

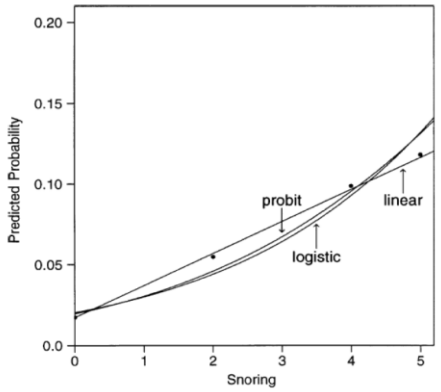


Figure 3.1. Fit of models for snoring and heart disease data.

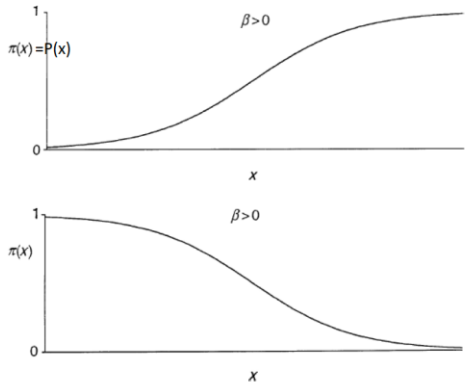


Figure 3.2. Logistic regression functions.

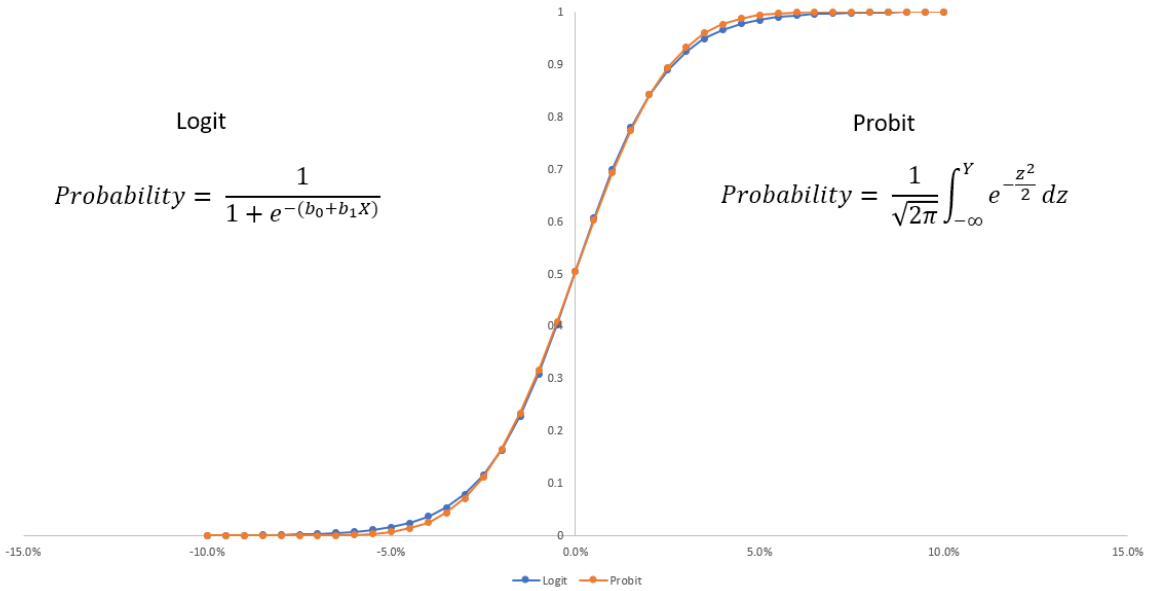
Logit vs Profit Model (S&P500 Impact on KLCI)

Logit

$$Probability = \frac{1}{1 + e^{-(b_0+b_1X)}}$$

Probit

$$Probability = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{Y} e^{-\frac{z^2}{2}}\,dz$$



Óscar Centeno Mora

Preámbulo

- Hasta ahora hemos analizados casos en donde la variable dependiente Y es una variable continua.

- Sin embargo, las regresiones no se limitan únicamente al caso de predecir un tipo valor de tipo continuo, y pueden ser utilizadas tanto para datos dicotómicos, nominales, y hasta ordinales, esto es, el caso para variables cualitativas en la variable Y.

- El presente tema solo abarcará el caso de datos dicotómicos.

- Veremos que dentro de la familia de distribuciones, la binomial puede optar por funciones de enlace tipo logit o probit (el cloglog no será cubierto en el presente capítulo).

- Estudiaremos la regresión logística y la regresión probit.

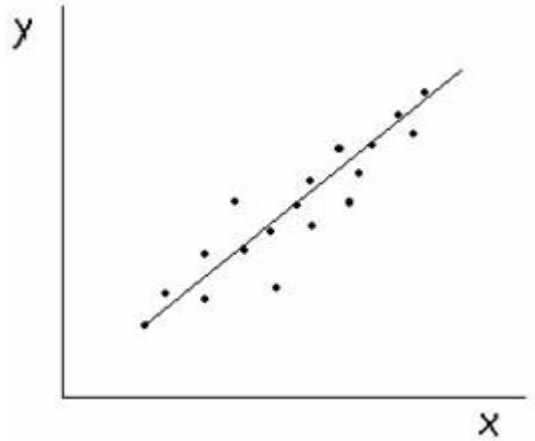
Preámbulo

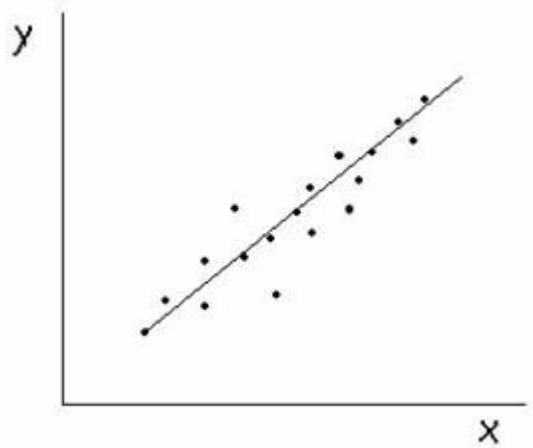
Regresión bivariada

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

$$Y \sim X$$

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





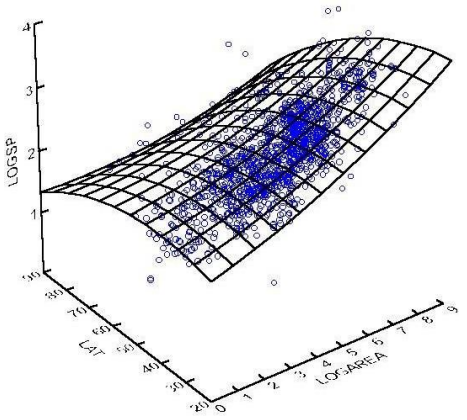
Regresión multivariada

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

Índice

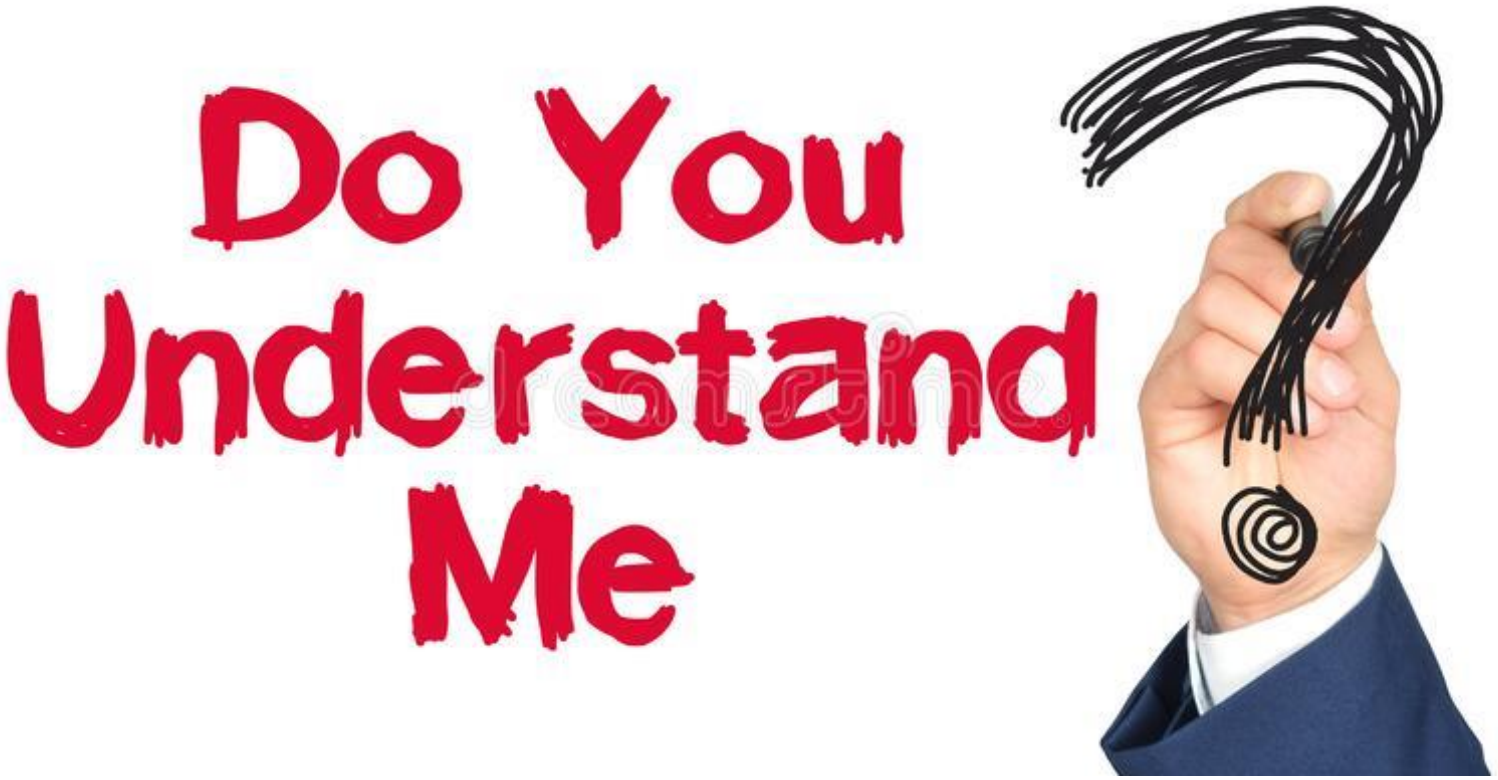
Introducción

- En el análisis de la información, concretamente en el análisis de regresión, una regresión binaria estima una relación entre una o más variables explicativas (X) y una única variable binaria de salida (Y).

- Generalmente, se trata de modelar la probabilidad de las dos alternativas (hombre-Mujer, sano-enfermo, mora-ok, etc…), en lugar de simplemente generar un valor único, como en la regresión lineal.

- Una relación lineal en este caso no sería lo adecuado: tratamos de estimar si el resultado será cualquier dicotómico planteado o especificado por la variable dependiente Y.

- En este caso, al no predecir un conjunto de valores en el dominio de $\mathbb{R}$ (no será un valor continuo), nos interesará predecir, de forma probabilística, para el conjunto de variables explicativas (X), la probabilidad de que la observación i, pertenezca a una de las dos variables dependientes.



¿lo anterior se entendió?

Índice

GLM y la familia binomial

- Recordemos lo visto en la clase anterior :
- **Familia Binomial.** Esta distribución es adecuada únicamente para las variables que representan una respuesta binaria o un número de eventos.

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

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: $Y \sim Bin(1, \pi)$.

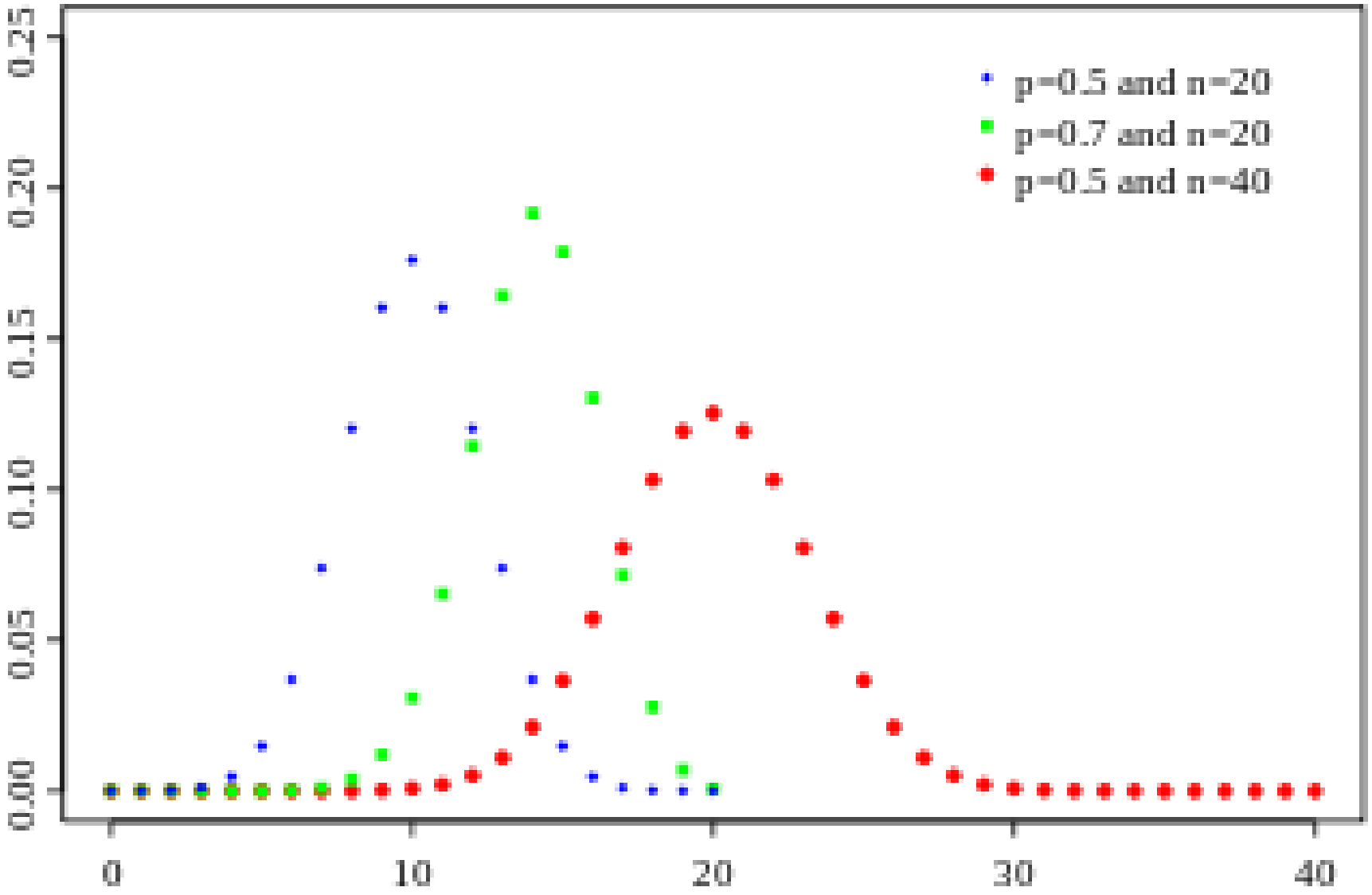
- Una distribución binomial se presenta matemáticamente:

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

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

La regresión logística

- Tanto la regresión lineal como la regresión logística son tipos de modelos lineales generalizados (GLM). Matemáticamente, los GLM se pueden expresar como:

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

En donde la función $g(*)$ se conoce como la una función invertible g, denominada función enlace (o link), y es la que determina, según la familia de funciones de probabilidad, el enlace que le corresponderá a la parte determinística del modelo.

- El marco de referencia de la regresión logística dentro de los GLM es fácil de ver cuando se observan las ecuaciones generalizadas para la regresión lineal y logística. Recuerde que los modelos de regresión lineal se definen mediante la ecuación:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \ , \epsilon \sim N(0, I\sigma^2)$$

La regresión logística

- La regresión logística se define de manera similar:

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

- Sin embargo, a diferencia de la regresión lineal, tenga en cuenta que el cálculo de la respuesta no es directo. ¿Por qué decimos esto?

- El lado de la ecuación con la variable de respuesta se conoce como logaritmo de probabilidades.

La regresión logística

- En un contexto binario, las probabilidades son la probabilidad de un evento "positivo“ ($Y = 1$), dividido por la probabilidad de un evento "negativo“ ($Y = 0$).

$$\frac{p(x)}{1 - p(x)} = \frac{P[Y = 1|X = x]}{P[Y = 0|X = x]}$$

La ecuación de la regresión logística garantiza que se calcule un valor entre **0** y **1**. Esto es evidente cuando se aplica la transformación *logit inversa*, lo que da como resultado una predicción de probabilidad “directa” :

$$p(x_i) = P[Y_i = 1|X_i = x_i] = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}$$

Tenga en cuenta que esta es una predicción de probabilidad, no un valor numérico. Este valor de probabilidad debe traducirse en una predicción categórica.

La regresión logística

- ¿Entonces, qué significa todo esto?

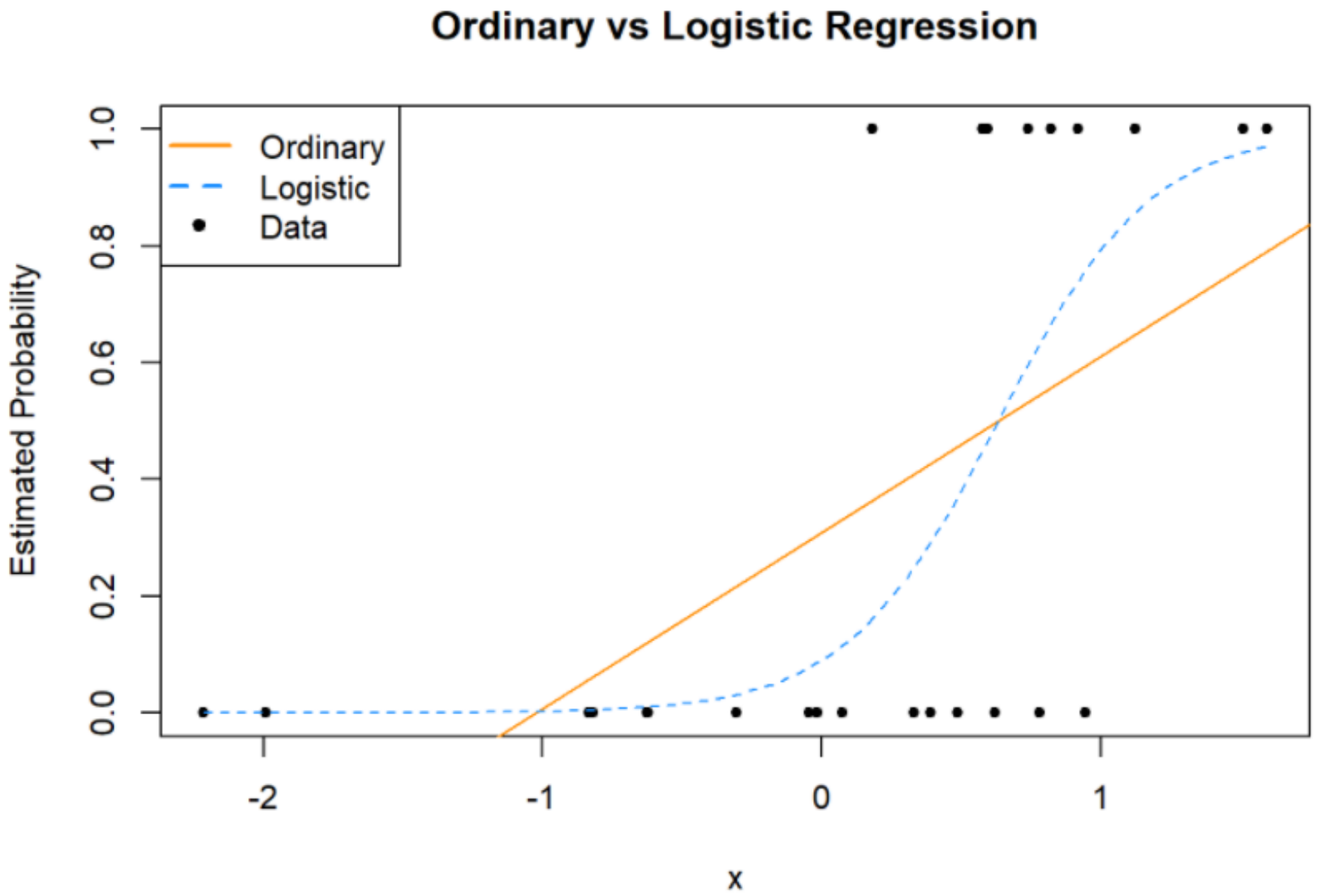
- En pocas palabras, la regresión lineal debe usarse para predecir una variable de respuesta cuantitativa (es decir, numérica), mientras que la regresión logística debe usarse para predecir una variable de respuesta cualitativa (es decir, categórica).

- (De manera más general, predecir una variable de respuesta categórica se conoce como clasificación). ¿Por qué?

- Visualmente, el modelo lineal genera una línea recta y el modelo logístico genera una curva "S".

- Veamos esto en la siguiente diapositiva.

La regresión logística



La regresión logística

En el proceso de estimación de una regresión logística, debemos:

- Analizar las variables predictoras (X) y la variable dependiente (Y).
- Estimar el modelo de la regresión logística.
- Verificar la bondad y ajuste del modelo.
- Interpretar los coeficientes.
- Trazar relaciones de estimación.
- Se puede hacer un análisis de los residuos (diagnóstico).
- Predecir una observación para conocer su probabilidad de clasificación.

La regresión logística

Se debe de tener en cuenta que los modelos de regresión logística no hacen el mismo supuesto que hace la regresión lineal, que incluyen:

- Relación lineal (entre la respuesta y los predictores)
- Normalidad multivariante (de los predictores)
- Poca o nula multicolinealidad (de los predictores)
- Sin autocorrelación
- Homocedasticidad

La regresión logística

- A modo de referencia, la siguiente tabla resume las diferencias entre la regresión lineal y logística con respecto al marco GLM:

	Linear Regression	Logistic Regression
Distribution of $Y \mid \mathbf{X} = \mathbf{x}$	$N(\mu(\mathbf{x}), \sigma^2)$	$\text{Bern}(p(\mathbf{x}))$
Distribution Name	Normal	Bernoulli (Binomial)
$\text{E}[Y \mid \mathbf{X} = \mathbf{x}]$	$\mu(\mathbf{x})$	$p(\mathbf{x})$
Support	Real: $(-\infty, \infty)$	Integer: $0, 1$
Usage	Numeric Data	Binary (Class: Yes/No) Data
Link Name	Identity	Logit
Link Function	$\eta(\mathbf{x}) = \mu(\mathbf{x})$	$\eta(\mathbf{x}) = \log\left(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right)$
Mean Function	$\mu(\mathbf{x}) = \eta(\mathbf{x})$	$p(\mathbf{x}) = \frac{e^{\eta(\mathbf{x})}}{1+e^{\eta(\mathbf{x})}} = \frac{1}{1+e^{-\eta(\mathbf{x})}}$

Índice

La regresión probit.

- Un modelo probit es un tipo de regresión dicotómica en donde la variable dependiente puede tomar solo dos valores.

- La palabra es un acrónimo, viene de ***prob***abilidad + un***it*** (unidad).[1]

- El propósito del modelo es estimar la probabilidad de que una observación con características particulares caerá en una categoría específica.

- Además, clasificando las observaciones basadas en sus probabilidades predichas es un tipo de modelo de clasificación binario.

- ¿Por qué hablamos de clasificar acá también?

La regresión probit.

- Un modelo probit es una especificación popular para un modelo de respuesta ordinal o binario (cualitativa).

- Como tal, trata el mismo conjunto de problemas que la regresión logística utilizando técnicas similares.

- El modelo probit, que emplea una función de enlace probit, se suele estimar utilizando el procedimiento estándar de máxima verosimilitud , que se denomina una regresión probit.

- Los modelos Probit fueron presentados por Chester Bliss en 1934; Ronald Fisher propuso un método rápido para calcular las estimaciones de máxima verosimilitud para ellos como apéndice del trabajo de Bliss en 1935.

La regresión probit.

- Supongamos que una variable de respuesta Y es dicotómica, es decir, que puede tener solo dos resultados posibles que denotaremos como 1 y 0.

- Por ejemplo, Y puede representar la presencia o ausencia de una determinada condición, éxito o falla de algún dispositivo, responder sí o no en una encuesta, etc.

- También tenemos un vector de regresores *que denotaremos por* X, que se supone influyen en el resultado de Y. Específicamente, suponemos que el modelo toma la forma:

$$E(Y|X) = P[Y = 1|X] = \Phi(X'\beta) = \Phi(\beta_0 + \beta_1 X)$$

donde P denota la probabilidad, y Φ es la función de distribución acumulada (FDA) de la distribución normal estándar. Los parámetros β se estiman mediante máxima verosimilitud y denotan .

La regresión probit.

Recordamos que:

$$\Phi(z) = P(Z < z), \qquad Z \sim N(0,1)$$

de tal forma que el coeficiente Probit de β_1 es el cambio en z asociado con un cambio de una unidad en X.

- Aunque el efecto sobre z de un cambio en X es lineal, la relación entre z y la variable dependiente de Y, es no lineal desde que hicimos la elección de que Φ es la función de distribución acumulada, y esta no es lineal respecto al componente de X.

- Dado que la variable dependiente es una función no lineal de los regresores, el coeficiente de X no tiene una interpretación simple (caso contrario del Logit).

- Generalicemos el caso a para 2 o más variables predictoras.

La regresión probit.

- Asumiendo que la variable Y es dicotómica, podemos decir que el modelo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + u$$

mediante

$$P(Y = 1|X_1, X_2, \ldots, X_p = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)$$

es el modelo Probit poblacional con múltiples regresores $X_1, X_2, \ldots, X_p$, y $\Phi()$ es la función de distribución normal estándar acumulada.

La probabilidad predicha de que $Y = 1$ dado las variables $X_1, X_2, \ldots, X_p$ se suele calculas en dos pasos

1. Calcular $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$
2. Buscar el valor de $\Phi(z)$ como probabilidad asociada a una normal estándar (en R mediante el *pnorm()*)

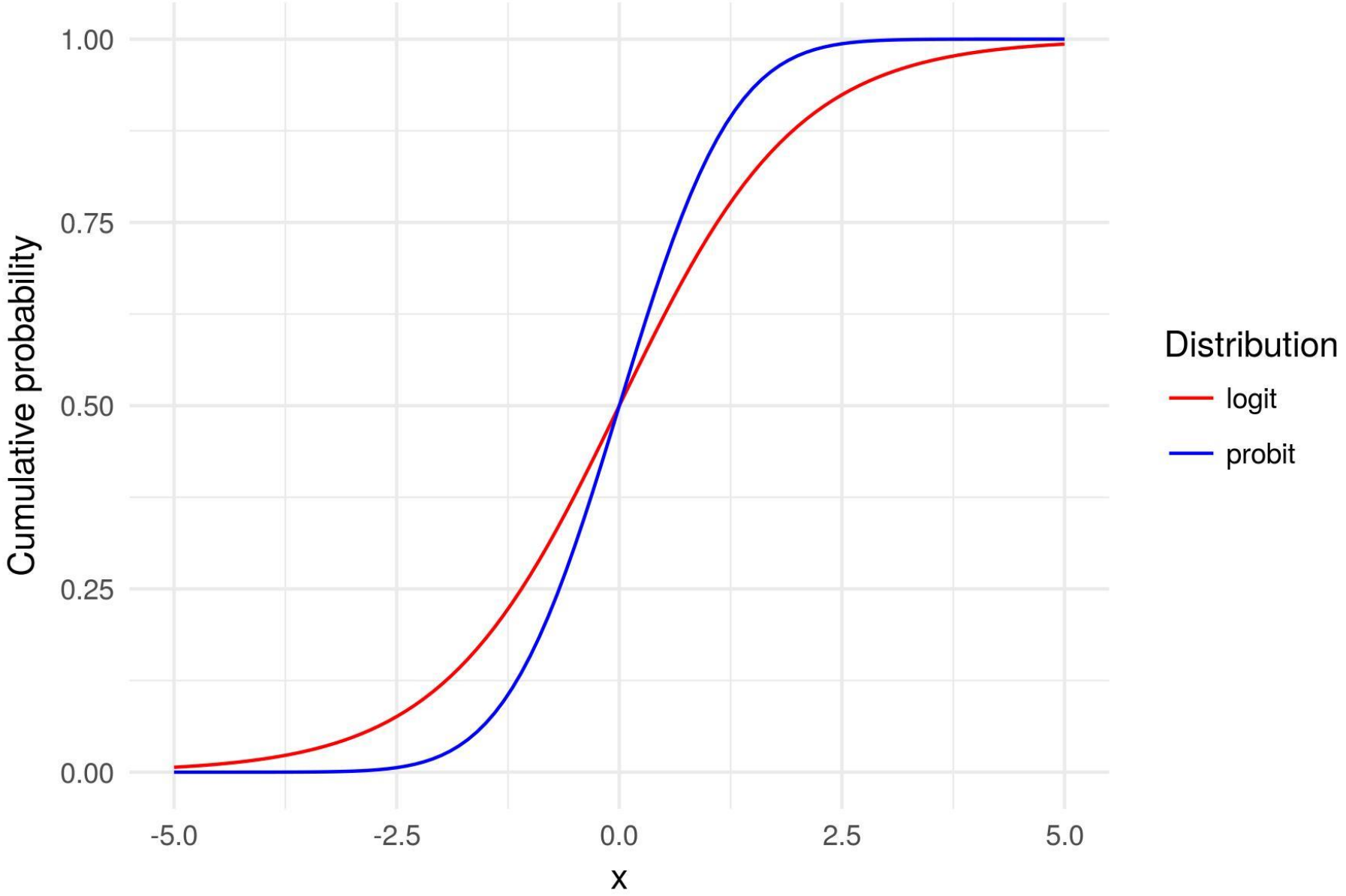
La regresión probit.

- Logit y probit difieren en cómo definen el enlace de la función de familias bonimiales.
- El modelo logit usa algo llamado función de distribución acumulativa de la distribución logística.
- El modelo probit usa algo llamado función de distribución acumulativa de la distribución normal estándar para definir Φ() . Ambas funciones tomarán cualquier número y lo reescalarán para que caiga entre 0 y 1.
- Por lo tanto, sea lo que sea lo que sea igual a $\beta_0 + \beta x$, la función puede transformarlo para producir una probabilidad predicha.
- Cualquier función que devuelva un valor entre cero y uno funcionaría, pero existe un modelo teórico más profundo que sustenta logit y probit que requiere que la función se base en una distribución de probabilidad. Los CDF normales estándar y logísticos resultan ser convenientes matemáticamente y se programan en casi cualquier paquete estadístico de propósito general.

La regresión probit.

- ¿Logit es mejor que probit o viceversa?
- Ambos métodos producirán inferencias similares (aunque no idénticas).
- Logit, también conocido como regresión logística, es más popular en las ciencias de la salud como la epidemiología, en parte porque los coeficientes se pueden interpretar en términos de razones de probabilidades.
- Los modelos probit se pueden generalizar para tener en cuenta las variaciones de error no constantes en entornos econométricos más avanzados (conocidos como modelos probit heterocedásticos) y, por lo tanto, los economistas y politólogos los utilizan en algunos contextos.
- Si estas aplicaciones más avanzadas no son relevantes, entonces no importa el método que elija.

La regresión probit.



Otras aproximaciones

¿Hay una aproximación mejor que las otras…

Conclusión

- El presente capítulo estudio para los GLM, las familia de datos binomiales con función de enlace logit y probit.

- Dentro de los logit, vemos que son aproximaciones en donde se puede utilizar el coeficiente para interpretar los resultados.

- La regresión de probit toma su iniciativa a partir de una curva acumulada, obteniendo una no linealidad con las X, lo cual no permite una interpretación.


The End