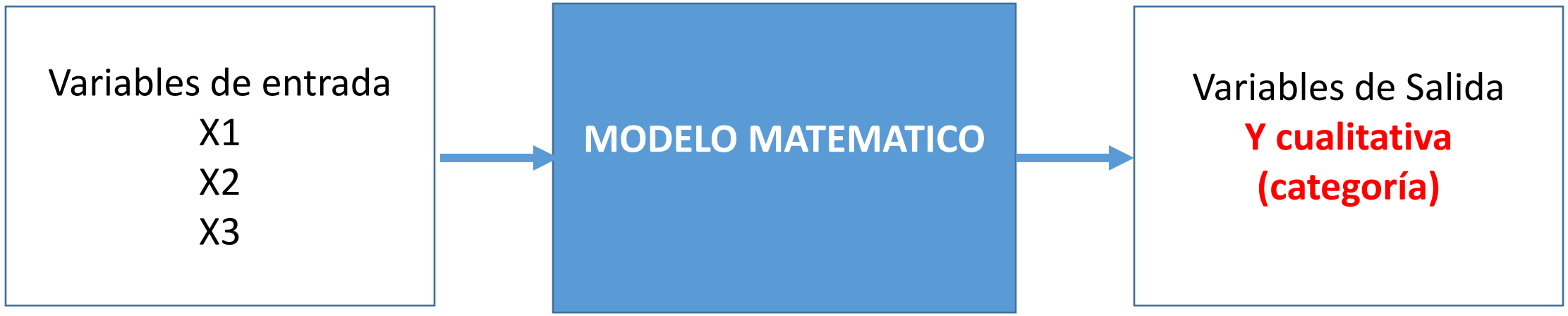
UNIVERSIDAD
NACIONAL
DE COLOMBIA

Universidad Nacional de Colombia
Sede Medellín

Regresión logística

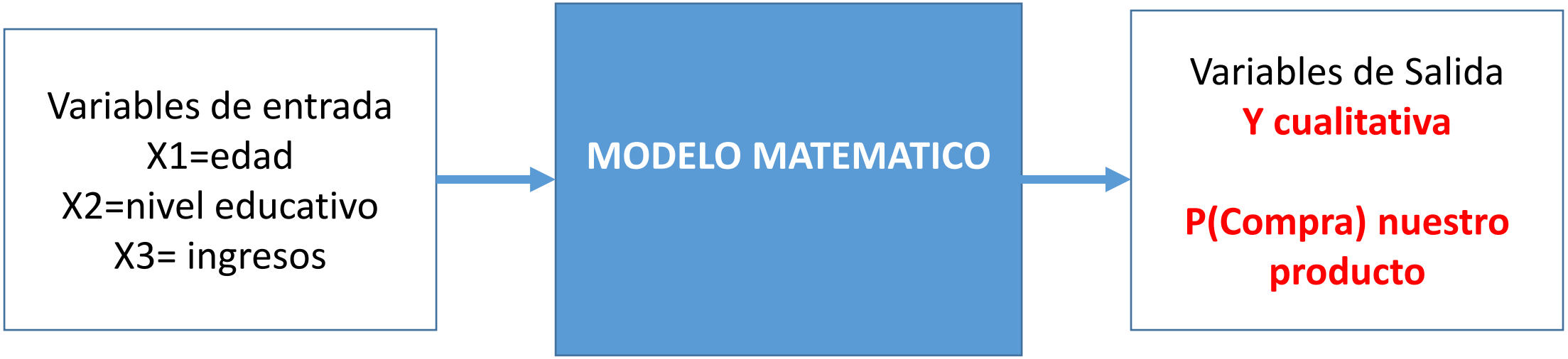
Profesora: Patricia Jaramillo A. Ph.D

Regresión logística



Cuando la variable de salida Y es cualitativa, permite calcular la probabilidad de que, a partir de los valores X se de Y (variable cualitativa)

Ejemplo:

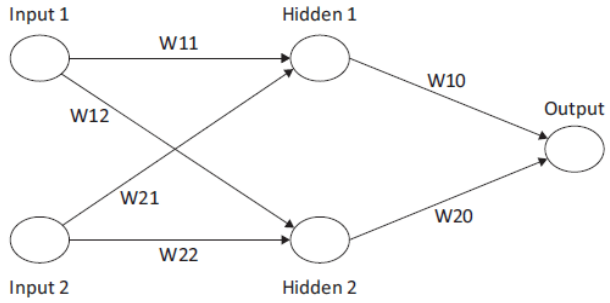


Regresión Logística: Estima la probabilidad P_y de que un individuo pertenezca a una categoría específica (por ejemplo, Y=1, quiere decir que el cliente compra)

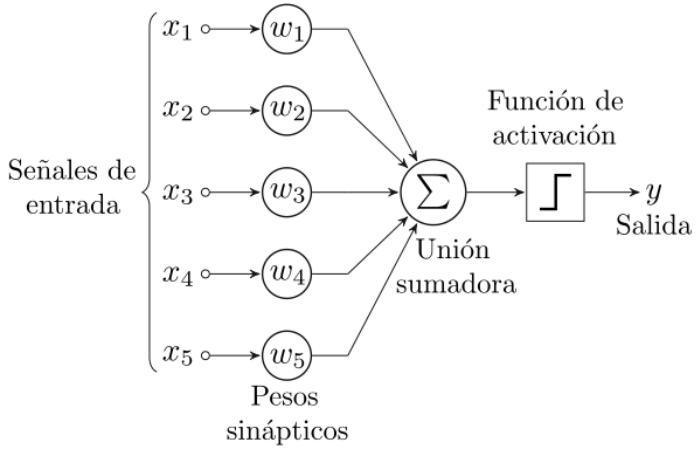
Generalmente se usa para 2 categorías (1 o 0) aunque también *puede usarse con un conjunto mas amplio de mas categorías*

En una Red neuronal de clasificación, la secuencia de operaciones es la siguiente:

- En cada enlace: el valor de entrada se multiplica por un parámetro (peso) w: wiXi
- En cada neurona: entrada: suma de los valores de los enlaces. Salida aplicar una función de activación – que puede ser logística, entre otras.



Perceptrón multicapa



Cada Perceptrón

Es un integración de regresión lineales y logísticas (u otra función de activación)

Veamos por ahora la Regresión logística

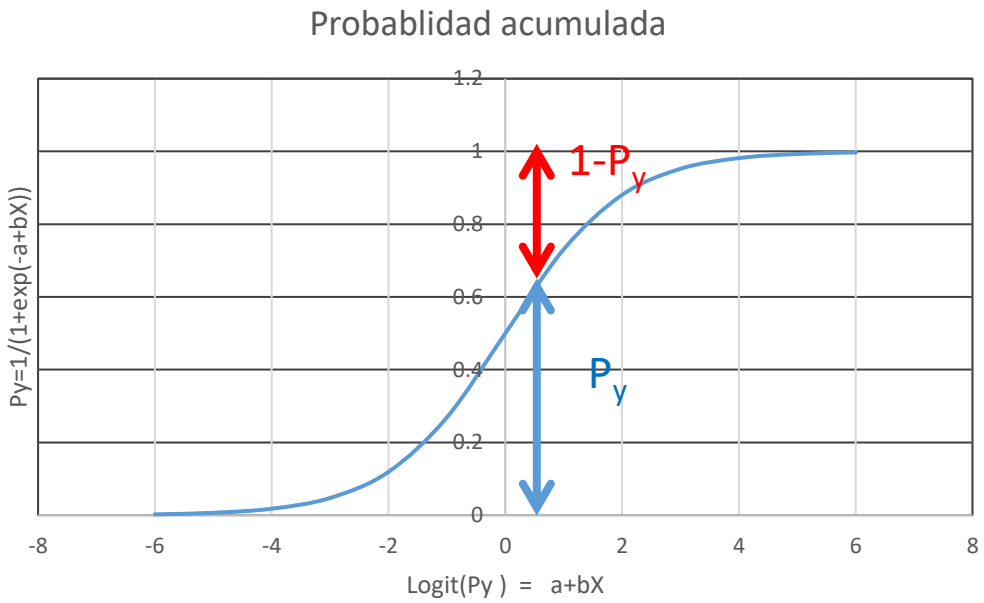
El operador logístico: Transforma la variable cualitativa en un valor de probabilidad.

Sigue una estructura equivalente a la regresión lineal, así:

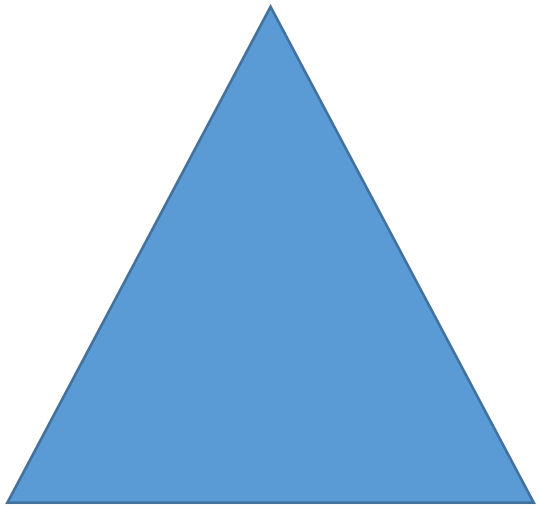
$$P_y = 1/ (1+e^{-(a+bX)})$$

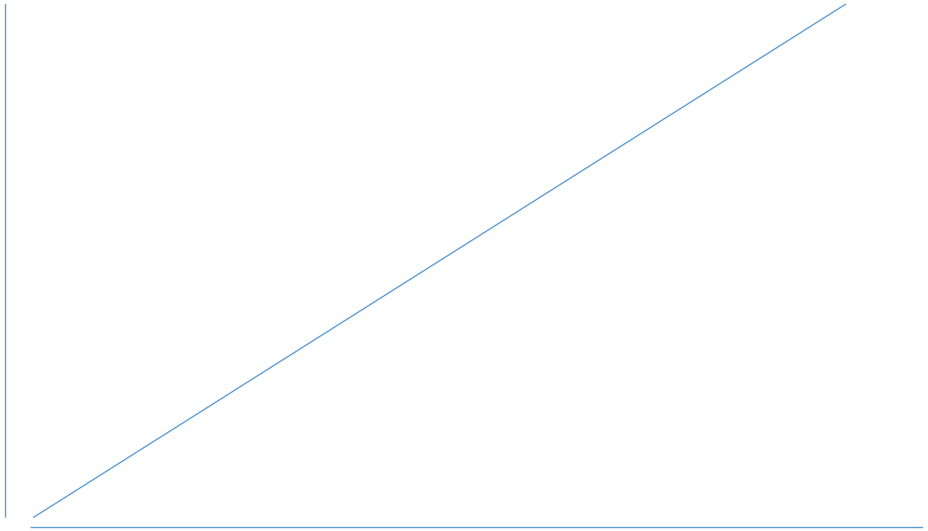
$$Logit(P_y)=ln(P_y /(1- P_y)) = a + bX$$

P_y= probabilidad acumulada de valor 1 (referencia).



No se ajusta una línea sino una función logística S-forma, que siempre da valores entre 0 y 1 .





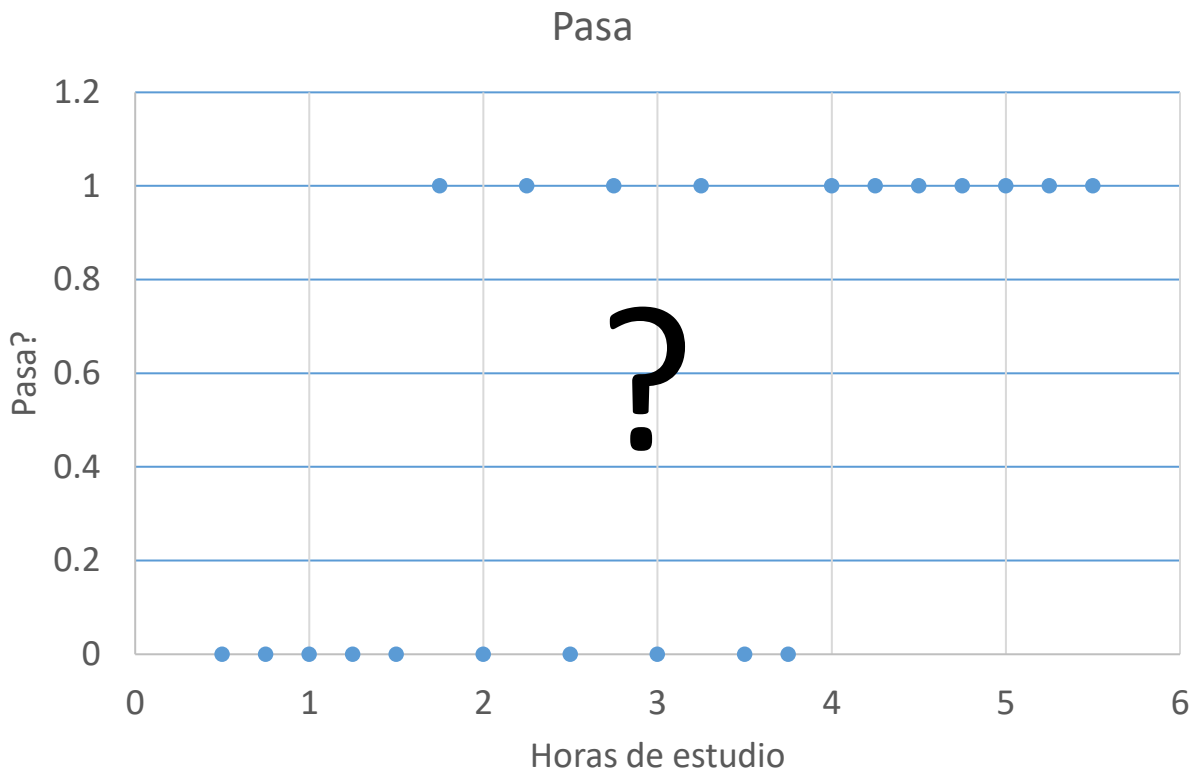
Ejemplo

Se cuenta con datos de estudiantes que pasaron una prueba académica y se les consultó cuántas horas habían dedicado al estudio.

Se desea saber que relación entre las horas de estudio y si paso o no

- Variable de entrada X: horas de estudio
- Variable de salida Y:
  - 1: pasa
  - 0: No pasa

	Horas	Pasa
1	0.5	0
2	0.75	0
3	1	0
4	1.25	0
5	1.5	0
6	1.75	1
7	2	0
8	2.25	1
9	2.5	0
10	2.75	1
11	3	0
12	3.25	1
13	3.5	0
14	3.75	0


Pasa
1.2
1
0.8
0.6
0.4
0.2
0
Pasa?
?
0
1
2
3
4
5
6
Horas de estudio

Implementación práctica aproximada

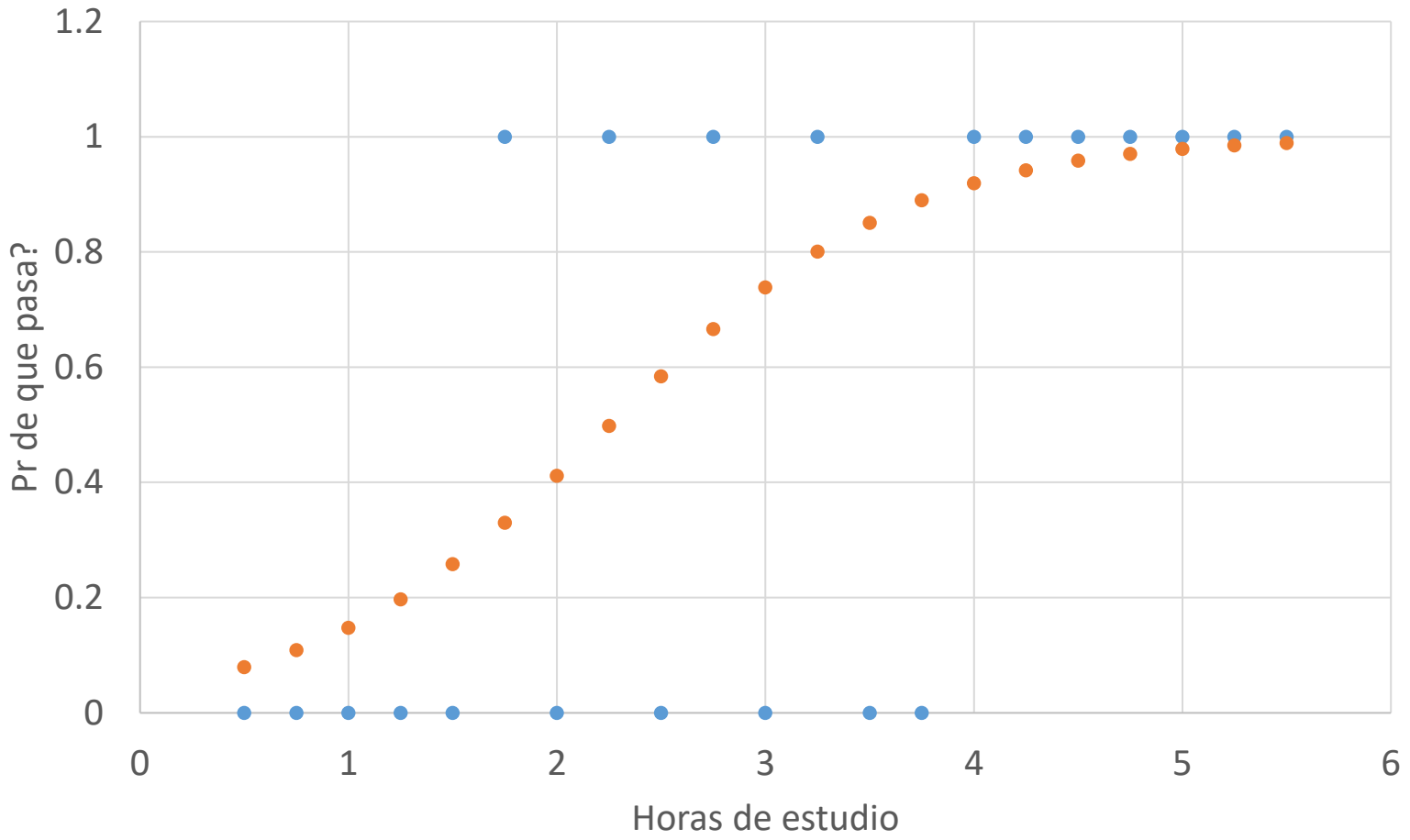
X horas ordenados	Categoría paso o no?	Y (1 si pasa)	Py= suma acumulada de los que pasan (Y=1)	(1-Py)=suma acumulada de los que no pasan (Y=0)	Logit Ln(Py/(1-py))

Se aplica regresión lineal a la columna X y a columna logit

Ver actividad…

a	-2.66809
b	0.519397

Ln(py/(1-py))=-2.66809 + 0.51939 Horas



La relación lineal es:

logit (Probabilidad pasa) = **-3.15339** + **1.397634***Horas

a	-3.15339
b	1.397634

Ejemplo: Calcular la probabilidad de que ase si estudio 4 horas

logit(Probabilidad Pase) = **-3.15339** + **1.397634***4 = 2.4371

P(pase)= $1/(1+e^{-2.4371})$ = 0.91 equivalente a 91% de probabilidad de que pase

Regresión logística Incluyendo varias variables de entrada

- Sigue una estructura equivalente a la regresión lineal:

$$\text{Logit}(P_y)=\ln(P_y/(1-P_y)) = a + b_1*X_1 + b_2*X_2 +\ldots+ b_k*X_k$$

$$P_Y= \frac{1}{1+e^{-(a+b_1X_1+\ldots+b_kX_k)}}$$

Equivalente a función perceptrón de una capa siempre en una red neuronal:

$$P_Y= \frac{1}{(1+e^{-f(x)})}$$

Entremos en mas detalles:

El “radio de odds” (posibilidades): $r = P_y /(1- P_y))$

r significa que, por ejemplo, de cada 100 observaciones, P_y*100 contra (100 - P_y*100) fueron categoría 1.

De: $\ln(P_y /(1- P_y)) = \ln (r) = a + b_1*X_1 + b_2*X_2 +\ldots+ b_k*X_k$

$$r= e^{(a + b1*X1 + b2*X2 +\ldots+ bk*Xk)}$$

ln (r) sigue una función lineal : Si X_i cambia en 1 unidad y las demás X quedan igual, r cambia en e^{bi}, y P_y cambia en 1/(1+r).

Ej: $b_j = 0.693$. r cambia en: $\exp(b_j) = 2$:

- Si X_j es una variable numérica (por ejemplo edad), cada incremento de un año dobla r.

- Si X_i es binaria (ejemplo género, donde 1 es hombre, si la observación es hombre, la respuesta es 2 veces mas probable a ser Y=1 , que si es mujer.

$$r = e^{(a + b1*X1 + b2*X2 + \ldots + bk*Xk)}$$

$$= \prod_{j=0}^{K} \exp(b_j x_j)$$

- La multiplicatoria, significa que los modelos logísticos son multiplicativos mas que aditivos.
- La solución a una regresión logística es encontrar los valores de los parámetros b_i que maximicen la verosimilitud de los datos, equivalente al producto de las probabilidad (predichas) de las N observaciones.

Recuerde: exp(z+t) = exp(z)*exp(t)

Calidad del ajuste

- Los modelos de regresión lineal se ajustan minimizando la suma de los residuos al cuadrado.

- Los modelos de regresión logística se ajustan maximizando la log-verosimilitud de los datos con resultado Y=1 .

Idea de Máxima Verosimilitud

Encuentra los coeficientes b_k que generan más a menudo la muestra observada: aquellos que maximizan la probabilidad de que un suceso observado haya ocurrido

En general

Valores absolutos de los coeficientes: indican menor o mayor impacto en las predicciones, pero:

- No significan importancia
- **Significa contribución de una variable a la probabilidad: esta no depende de ella sola sino de la interacción con los otros**
- Si las unidades de X´s son diferentes los coeficientes no son comparables a no ser que se normalicen o se autoescalen (su desviación estándar sea igual a 1)

- Esto solo es válido si realmente son independientes (poco frecuente).
- Grandes magnitudes de los coeficientes y grandes errores estándar de los coeficientes pueden ser señales de correlaciones entre variables dependientes.
- Multicolinealidad: cuando 2 campos son altamente correlacionados: hay redundancia

Ejemplo

- En un hospital se desea diseñar un plan para emergencias neonatal y se requiere diseñar salas y equipos.
- Los bebes recién nacidos tienen 5 minutos para enviarlos a emergencias si en el test Apgar (en una escala de 0 a 10) tienen un valor por debajo de 7.
- Aunque esos valores son escazos, los hospitales no pueden correr riesgos y no tener suficientes recursos para atenderlos rapidamente.

Ejemplo extraído de: Albright and Winston (2017).

- Datos: 2010 natality public-use data file (http://mng.bz/pnGy) de 50 estados de Estados Unidos 26,000 nacimientos en el data frame *data*.

- La variable de salida es 1: en alto riesgo, a la que se le asigna la probabilidad P.

- Las variables independientes incluyen datos de la madre, del padre y del proceso del embarazo. Solo se usan las variables a priori del nacimiento.

Variable	Type	Description
tRisk	Logical	TRUE if 5-minute Apgar score < 7; FALSE otherwise
WGT	Numeric	Mother's prepregnancy weight
PREVIS	Numeric (integer)	Number of prenatal medical visits
IG_REC	Logical	TRUE if smoker; FALSE otherwise
ESTREC3	Categorical	Two categories: <37 weeks (premature) and >=37

Variable	Type	Description
DPLURAL	Categorical	Birth plurality, three categories: single/twin/triplet+
ULD_MECO	Logical	TRUE if moderate/heavy fecal staining of amniotic fluid
ULD_PRECIP	Logical	TRUE for unusually short labor (< three hours)
ULD_BREECH	Logical	TRUE for breech (pelvis first) birth position
URF_DIAB	Logical	TRUE if mother is diabetic
URF_CHYPER	Logical	TRUE if mother has chronic hypertension
URF_PHYPER	Logical	TRUE if mother has pregnancy-related hypertension
URF_ECLAM	Logical	TRUE if mother experienced eclampsia: pregnancy-related seizures

Para el ejemplo se descubrió que:

- Nacimientos prematuros y trillizos son predictores fuertes de riesgo

- Otras variables importantes son el peso previo de las madres (si es alto hay mas riesgo), el número de visitas prenatales (a mas visitas menor el riesgo), entre otras.

- Hay una correlación positiva entre: la madre fuma y riesgo.