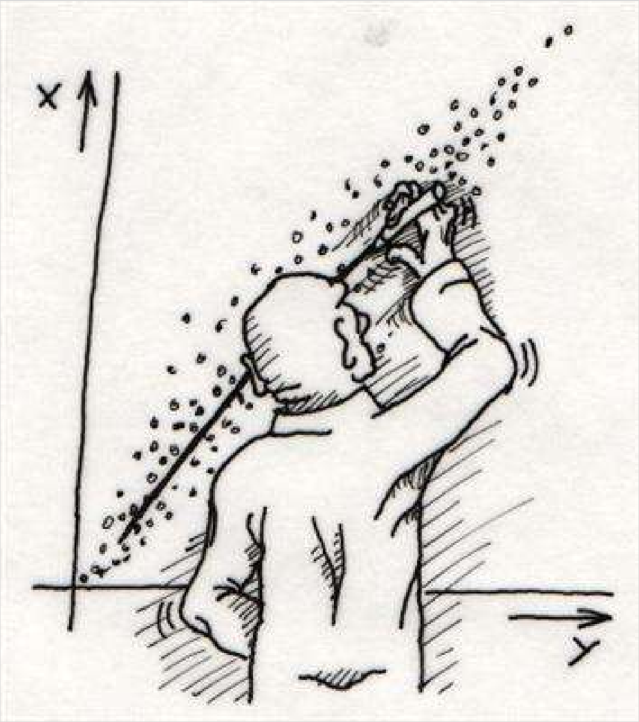


DigitalHouse >
Coding School

DATA SCIENCE

UNIDAD 2
MÓDULO 4

Introducción a la regresión logística

Octubre 2017

Regresión Logística

Objetivos

- Definir el concepto de modelo de regresión logística
- Entender el fundamento matemático de la regresión
- Diseñar un modelo de regresión logística utilizando las librerías statsmodels y Scikit-Learn de Python

Objetivos

- La regresión logística es un abordaje lineal para resolver problemas de clasificación
- Estima la relación existente entre una variable dependiente (target) y diversas variables predictoras (independientes)
- A grandes rasgos mediante una regresión logística se busca estimar la probabilidad que Y sea 1 dados ciertos valores de X : $P(Y=1 \mid X) = ?$
- Si X e Y mantienen una relación de tipo lineal positiva el valor de Y se acercará a 1, conforme se incremente el valor de X

Ejemplo

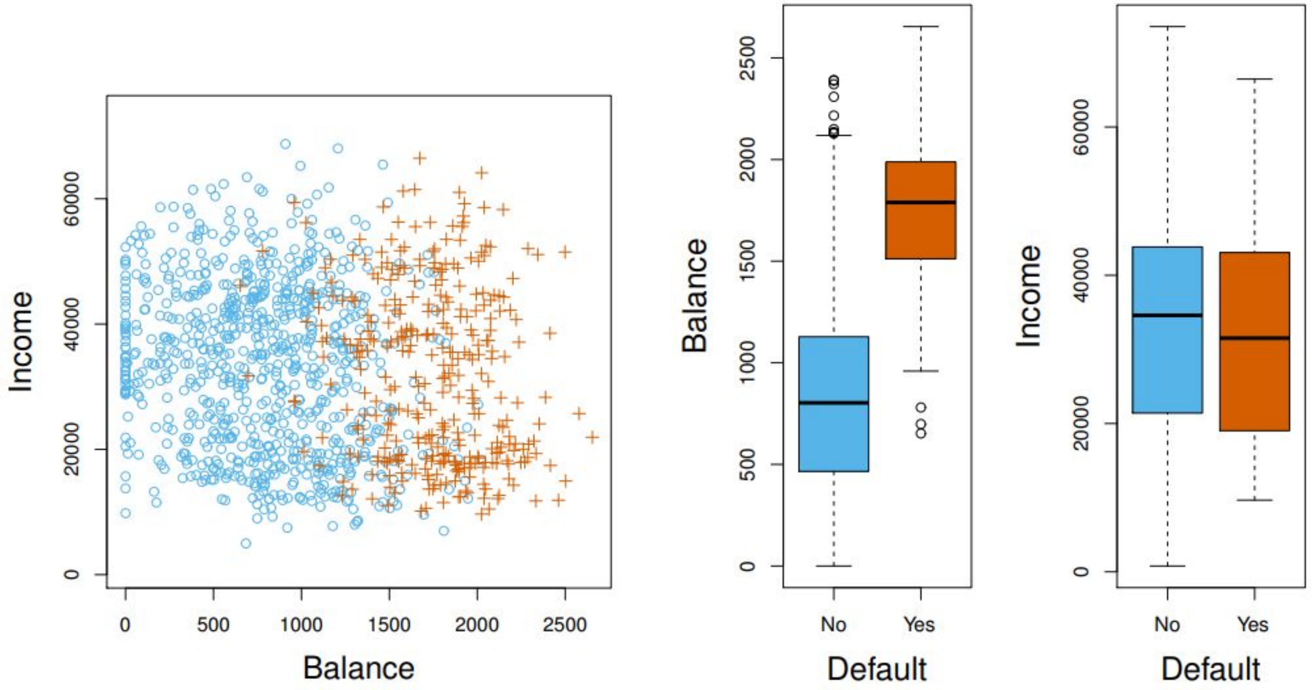
DigitalHouse>
Coding School

- Queremos predecir la probabilidad de que un cliente no pague su préstamos (entre en Default)

$$Y = \begin{cases} 0 & \text{if } \texttt{No} \\ 1 & \text{if } \texttt{Yes}. \end{cases}$$

- Los features son
  - ingresos
  - balance
- Queremos predecir P(Y = 1 | balance)
- Si la p(Y=1|balance) > 0.5 => default = Yes (podríamos elegir otro umbral)

Ejemplo


Income
0
20000
40000
60000
0
500
1000
1500
2000
2500
Balance
Balance
0
500
1000
1500
2000
2500
No
Yes
Default
Income
0
20000
40000
60000
No
Yes
Default

Un abordaje intuitivo sobre la regresión logística

- La regresión logística es similar a la Lineal , pero con una diferencia crucial:
  - las variables predictoras pueden ser tanto categóricas como continuas igual que en Regresión Lineal
  - la variable target que se busca modelar es **categórica**, usualmente de tipo dicotómica

- Basándonos en esta idea, mediante una regresión logística, podemos conocer, por ejemplo, cuál de dos tipos de usuarios es más probable que adquiera un producto X.
- En función de lo anterior podemos establecer lo que en negocios de denomina segmentación de clientes
- En investigación clínica por ejemplo podemos usar este modelo para generar predicciones sobre un tipo de tumor

Regresión Lineal vs. Logística

- ¿Por qué no estimar p(Y=1 | X) con una regresión lineal? En ese caso, nuestro modelo asumiría la siguiente forma:

$$p(X) = \beta_0 + \beta_1 X_1$$

- donde para abreviar, definimos $p(X) = p(Y = 1|X)$

- Arrojaría valores fuera del rango válido para una probabilidad (0,1)

Regresión Lineal vs. Logística

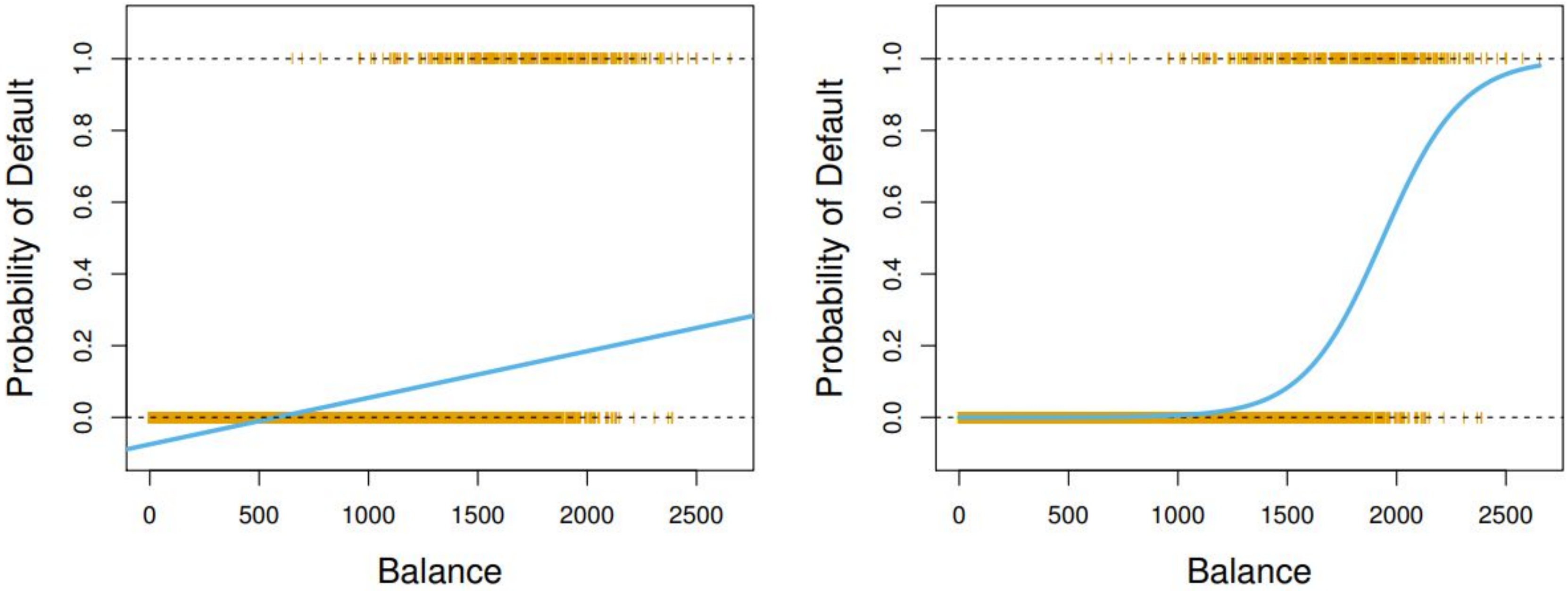
DigitalHouse >
Coding School

- Tenemos que buscar una función que nos garantice que las estimaciones que hagamos estarán dentro del rango válido de una probabilidad.
    - Podemos usar la función logística:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

- Vemos ahora que, sin que importe qué valores tome X siempre vamos a predecir valores dentro del rango 0-1.
- La función logística devuelve una curva en forma de “S” siempre entre 0 y 1.
- No es la única función para conseguir este resultado

Regresión Lineal vs. Logística



Regresión Lineal vs. Logística

- Si manipulamos un poco la función logística que vimos hace algunas slides podemos llegar a la siguiente expresión:

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X}$$

- La cantidad p(X)/1-p(X) se denomina "odds-ratio" y lo que expresa es la relación entre la probabilidad de que y=1 (**p(X)**) y la probabilidad de que y=0 (**1-p(X)**).
- El odds-ratio toma valores entre 0 e infinito.

Regresión Lineal vs. Logística

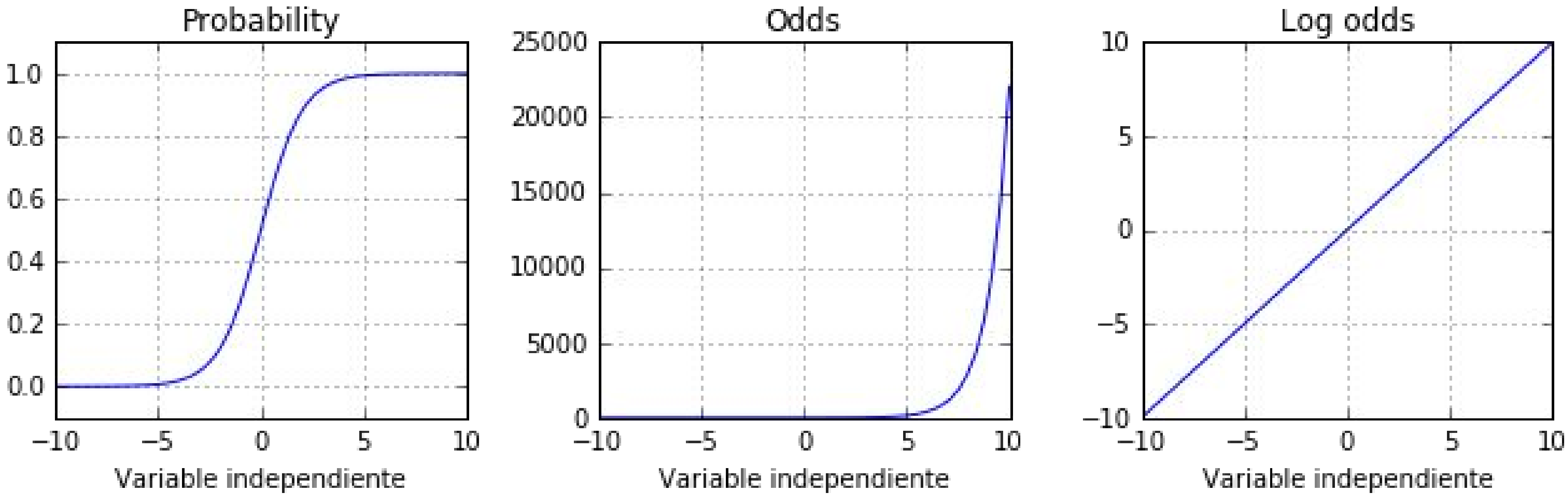
DigitalHouse >
Coding School

- Si tomamos logaritmos de la expresión anterior (odds-ratio)

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1$$

- vemos que el logaritmo del odds-ratio tiene una relación lineal con X.
- En el modelo lineal, los $\boldsymbol{\beta}_p$ eran el cambio promedio en Y ante un cambio unitario en X.
- En la regresión logística, incrementar una unidad en X, cambia el logaritmo del odds-ratio en $\boldsymbol{\beta}_p$. O, lo que es lo mismo, multiplica el odds por $e^{\beta p}$.
- La relación entre p(X) y X no es una línea recta => **cuánto cambia p(X) ante un cambio unitario en X** depende de los valores de X. Aún así, el signo de $\boldsymbol{\beta}_p$ expresa la dirección de cambio en p(X) (independientemente del valor de X).

Regresión Lineal vs. Logística



Regresión Logística - Estimando los coeficientes

- Los coeficientes se estiman mediante el método de máxima verosimilitud (un método de estimación más general que el de mínimos cuadrados). Lo que se busca es estimar los β_0 y β_1 que tengan una mayor probabilidad relativa de haber generado los datos observados.

- Los resultados de la estimación de la probabilidad de default en base al balance pueden verse en la tabla siguiente

Regresión Logística - Estimando los coeficientes

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.6513	0.3612	−29.5	<0.0001
balance	0.0055	0.0002	24.9	<0.0001

- Se observa que un incremento de $1 en el balance incrementa 0.0055 unidades el "log odds ratio".
- Hay un estadístico z que es análogo al estadístico t en regresión lineal.
  - Ho β_1 = 0 (o en otras palabras que la probabilidad de default no depende del balance
  - Ha β_1 <> 0

Regresión Logística - Estimando los coeficientes

- También podrían usarse variables cualitativas como features.
- Veamos, por ejemplo, cómo influye la condición de ser estudiante sobre la probabilidad de default.
- Realizamos un procedimiento análogo al que usábamos para introducir variables cualitativas en una regresión lineal: incluimos variables dummy. En este caso, X=1 si es estudiante y X=0 si no lo es.

Regresión Logística - Estimando los coeficientes

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−3.5041	0.0707	−49.55	<0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

- En este caso, el coeficiente es positivo, lo cual indica que ser estudiante tiene una relación positiva con ser potencial moroso.
- ¿Qué pueden decir de la significación del coeficiente?

Regresión Logística - Haciendo predicciones

- Una vez que hemos estimados los coeficientes del modelo podemos hacer predicciones y computar la probabilidad de default para algún valor dado de balance.
- Por ejemplo, para un balance $1000 tenemos que está por debajo del 1%

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1,000}}{1+e^{-10.6513+0.0055\times 1,000}} = 0.00576,$$

- En cambio, para un balance de $2000 es mucho más grande y está cerca del 58%

Regresión Logística - Estimando los coeficientes

Podemos estimar las probabilidades de entrar en default siendo estudiante y no siéndolo (la variable dummy cuyos coeficientes habíamos estimado previamente).

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431,$$

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292.$$

Regresión Logística Múltiple

DigitalHouse> Coding School

- Ahora, de forma análoga al caso de regresión lineal, pensemos en el problema de predecir una variable cualitativa binaria con una serie de p features.
- Nuestro modelo de regresión logística múltiple quedaría definido

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_p X_p}}$$

- Pudiendo ser reescrito en términos de logs odds

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

Regresión Logística Múltiple

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

- Veamos los resultados de aplicar este modelo para predecir la probabilidad de default según el ingreso, el balance y la condición de estudiante.
- ¿Qué pueden decir de los resultados?
- ¿Ven algo raro?

Práctica guiada: ejecutando un modelo de regresión logística

LAB: implementando un modelo de regresión logística en datos de cáncer de mama

Conclusiones

- Modelo para abordar problemas de clasificación con una variable target cualitativa o categórica, generalmente, binaria

- La relación entre la variable dependiente y los predictores es lineal al realizar la transformación logística de los datos

- Pueden interpretarse los valores predichos por el modelo como “probabilidades” de cada uno de las categorías de la variable.

- Podemos realizar la interpretación de la influencia de las variables predictoras en términos de odd-ratio