Regresión estadística

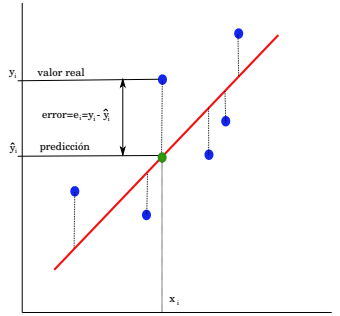
Josemari Sarasola

sigmalitika.hirusta.io

Regresión estadística

Objetivo

Ajustar una recta a la nube de puntos formada por las dos variables (x_i, y_i). de forma que la recta se acerque lo más posible a los puntos. La variable y es la variable dependiente. x es la variable independiente o causal.



Regresión estadística

Cómo se construye la recta?

La recta de regresión será la que minimice la suma de los cuadrados de los errores, la que llamaremos recta de regresión minimocuadrática. Por lo tanto, siendo la recta $\hat{y} = a + bx$, el problema se plantea de esta forma:

$$\min_{a,b} \sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - a - bx_i)^2$$

Regresión estadística

Cómo se construye la recta?

Para resolver el problema de minimización, derivamos la expresión a minimizar respecto a los parametros a y b y los igualamos a 0. De esta forma obtenemos las ecuaciones normales minimocuadráticas:

$$\frac{\partial \sum_i (y_i - a - bx_i)^2}{\partial a} = 0 \rightarrow \sum_i y_i = na + b \sum_i x_i$$
$$\frac{\partial \sum_i (y_i - a - bx_i)^2}{\partial b} = 0 \rightarrow \sum_i x_i y_i = a \sum_i x_i + b \sum_i x_i^2$$

Cómo se construye la recta?

Resolviendo las dos ecuaciones anteriores, por ejemplo con la regla de Cramer, así se estiman los parámetros a y b:

$$b = \frac{s_{xy}}{s_x^2}$$

$$a = \overline{y} - b\overline{x}$$

Bondad del ajuste

En la regresión de mínimos cuadrados se cumple la siguiente igualdad:

$$s_y^2 = s_{\hat{y}}^2 + s_e^2$$

s_y^2 es la varianza total, la varianza que pertenece a la variable dependiente y y que deseamos explicar a través de la variable x. $s_{\hat{y}}^2$ es la varianza explicada, la parte de varianza total que explicamos a través de la regresión con x, y s_e^2 la varianza residual, la parte de varianza total que no explicamos a través de la regresión con x. Así pues:

varianza total=varianza explicada+varianza residual

Bondad de ajuste: en inglés, *goodness of fit*

Bondad de ajuste

Por lo tanto, en relación a la varianza total, cuanto mayor sea la varianza explicada, mejor se ajustará la línea de regresión a los datos. De esta forma, podemos utilizar el denominado coeficiente de determinación como medida de bondad de ajuste:

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

R^2: en inglés, *coefficient of determination*.

Bondad de ajuste

El coeficiente de determinación R^2 toma valores en el intervalo $[0, 1]$ y puede utilizarse la siguiente regla general para su interpretación, bajo reserva de estudios similares y sin tener en cuenta el error muestral:

- $0 < R^2 < 0{,}3 \rightarrow$ ajuste escaso;
- $0{,}3 < R^2 < 0{,}6 \rightarrow$ bondad de ajuste media;
- $R^2 > 0{,}6 \rightarrow$ ajuste bueno;
- $R^2 = 1 \rightarrow$, ajuste total: todos los puntos están alineados.

Bondad de ajuste

Estas dos propiedades son útiles para el calculo del coeficiente de determinación:

- $\sum_i e_i = 0 \rightarrow \overline{e} = 0 \rightarrow s_e^2 = \frac{\sum_i e_i^2}{n}$;
- $\sum_i y_i = \sum_i \hat{y}_i \rightarrow \overline{y} = \overline{\hat{y}}$.

Diagnosis del modelo: diagrama de error

El diagrama de error no es más que un diagrama cartesiano, con valores x_i en el eje de abcisas, y errores e_i en el eje de ordenadas.

Para qué se utiliza? Al estimar la recta de regresión, se suponen implicitamente una serie de hipótesis o supuestos. El diagrama de errores se utiliza para ver si se cumplen o no dichas hipótesis, es decir, para realizar lo que se denomina la diagnosis del modelo.

Cómo se interpreta el diagrama de error?

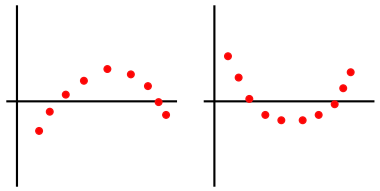


Figura: **Error de especificación**: se ha escogido una línea inadecuada, por ejemplo en lugar de una curva, una recta.

Cómo se interpreta el diagrama de error?

Figura: **Autocorrelación**: los errores consecutivos están relacionados. Es mala señal.

Cómo se interpreta el diagrama de error?

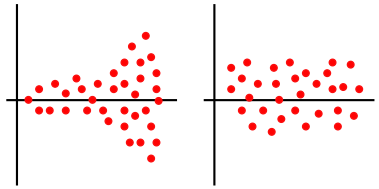


Figura:

- Dispersión de errores irregular: **heterocedasticidad** (mala señal).
- Dispersión de errores uniforme: **homocedasticidad** (buena señal).

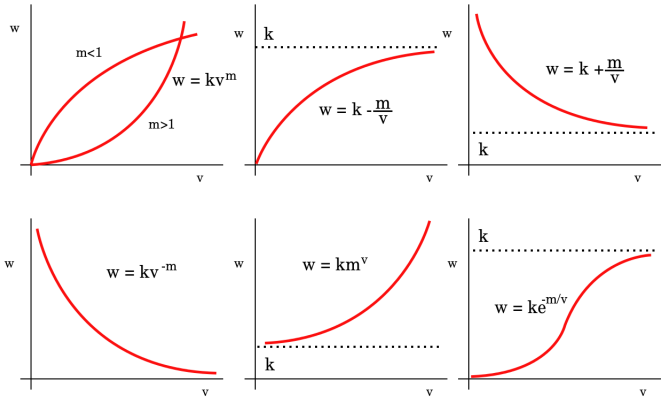
Cómo se interpreta el diagrama de error?

Para dar por bueno el modelo de regresión obtenido, los errores deben distribuirse al azar o caoticamente, sin tendencia clara.

Modelos no lineales

Hemos aprendido a ajustar una recta. Cómo se ajusta una curva? En general, se linealizará la curva, es decir, la convertiremos en recta, y tras estimar a y b , volveremos a la curva.

Modelos no lineales: catálogo de curvas



Modelos no lineales: linealización

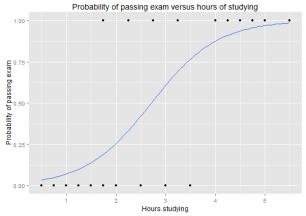
L	Linealización ($y = a + bx$)	a	b
$w = kv^m$	$\ln w = \ln k + m \ln v$	$a = \ln k$	$b = m$
$w = k - \frac{m}{v}$	$w = k + (-m)\frac{1}{v}$	$a = k$	$b = -m$
$w = k + \frac{m}{v}$	$w = k + m\frac{1}{v}$	$a = k$	$b = m$
$w = kv^{-m}$	$\ln w = \ln k + (-m) \ln v$	$a = \ln k$	$b = -m$
$w = km^v$	$\ln w = \ln k + (\ln m)v$	$a = \ln k$	$b = \ln m$
$w = ke^{-\frac{m}{v}}$	$\ln w = \ln k + (-m)\frac{1}{v}$	$a = \ln k$	$b = -m$

Modelos no lineales: cálculo de R^2

El cálculo de R^2 en curvas se debe hacer siempre en la forma linealizada, es decir, los valores de la variable dependiente y, sus predicciones $\hat{y}$ y los errores se deben calcular respecto de la recta. Si no se hace así, el coeficiente de determinación puede salir del intervalo $[0, 1]$.

Modelo logit

El modelo logit o regresión logistica es una curva de regresión particular, en la que la variable independiente se denomina *dosis* y la variable dependiente es la probabilidad de un suceso. Por tanto, los valores de la variable dependiente se deben limitar al intervalo [0,1].
En su forma normal, a mayor dosis, mayor probabilidad.



Modelo logit

En el anterior ejemplo, aparece una curva logística típica. Los valores en el eje de ordenadas aparecen acotados en el intervalo [0,1], ya que se trata de una probabilidad. Por otra parte, los incrementos de probabilidad se dan en un intervalo reducido de la variable dosis. Para dosis muy pequeñas y grandes, los incrementos de probabilidad son muy pequeños (por ejemplo, a partir de un cierto número de horas estudio la probabilidad de aprobar no se incrementará mucho, ya que de por sí será alta. La curva logística presenta pues dos asíntotas horizontales para $y = 0$ e $y = 1$ respectivamente.

Modelo logit

Para una dosis x y una probabilidad p, esta es la forma linealizada de la curva logística:

$$\ln\left(\frac{p}{1-p}\right) = a + bx$$

Como en cualquier otra curva, el cálculo del coeficiente de determinación se hará a partir de la forma linealizada.