

## REGRESIÓN NO LINEAL

Como ya indicamos, el propósito de los estudios de regresión es establecer la relación funcional de cierto tipo que “mejor expresa” el valor de una variable en función del de la otra en todos los individuos de la muestra. En las dos secciones previas, el tipo de relación funcional analizado ha sido el lineal.

Cuando el problema de regresión involucra una relación no lineal (de tipo general, es decir, con pendiente y ordenada en el origen cualesquiera en  $\mathbb{R}$ ) de  $Y$  respecto a  $X$ , pueden presentarse distintas situaciones, algunas de las cuales admiten un tratamiento similar al de (o que puede aproximarse por el de) la regresión lineal mínimo-cuadrática.

Entre estos casos pueden mencionarse los siguientes:

- Si la relación funcional de aproximación es del tipo

$$\hat{Y} = a \cdot g(X) + b \quad \text{con } a, b \in \mathbb{R},$$

en el supuesto de que  $g : \mathbb{R} \rightarrow \mathbb{R}$  es una función totalmente especificada (es decir, no depende de ningún término desconocido), como, por ejemplo, relaciones del tipo:

$$\hat{Y} = a \cdot \frac{1}{X} + b \quad \text{con } a, b \in \mathbb{R},$$

$$\hat{Y} = a \cdot \log(X) + b \quad \text{con } a, b \in \mathbb{R},$$

entonces la aplicación del principio de mínimos cuadrados es inmediata y análoga a la del caso lineal general.

De hecho, la *solución óptima del problema de regresión* es la del caso lineal reemplazando  $X$  por  $g(X)$ :

$$\hat{Y} - \bar{Y} = \frac{S_{g(X)Y}}{S_{g(X)}^2} \left( g(X) - \overline{g(X)} \right),$$

Las conclusiones de correlación son también análogas, sin más que recurrir al reemplazamiento anterior, de manera que el grado de dependencia del tipo funcional considerado puede medirse a través del coeficiente:

$$r_{g(X)Y} = \frac{S_{g(X)Y}}{S_{g(X)} \cdot S_Y},$$

y satisfaciéndose que:

$$\min_{a, b \in \mathbb{R}} \text{ECM}[\hat{Y} = a \cdot g(X) + b] = (1 - r_{g(X)Y}^2) S_Y^2.$$

En consecuencia, para elegir el tipo de relación más adecuada entre dos variables entre dos tipos de relaciones diferentes de las incluidas en este apartado, es indistinto optar por la que proporciona menor error cuadrático medio mínimo o por la

que proporciona mayor coeficiente de determinación, puesto que ambos criterios de elección coinciden.

Si en el ejemplo sobre el que hemos estudiado la regresión lineal tratamos de analizar una relación del tipo

$$\hat{Y} = a\sqrt{X} + b,$$

entonces la relación óptima vendría dada por:

$$\hat{Y} - \bar{Y} = \frac{S_{\sqrt{X}Y}}{S_{\sqrt{X}}^2} \left( \sqrt{X} - \overline{\sqrt{X}} \right),$$

el coeficiente de correlación asociado a esa relación óptima sería:

$$r_{\sqrt{X}Y} = \frac{S_{\sqrt{X}Y}}{S_{\sqrt{X}} \cdot S_Y}$$

y el error cuadrático medio mínimo en el que puede incurrirse aproximando el valor de  $Y$  por una relación del tipo considerado vendría dado por:

$$\min_{a,b \in \mathbb{R}} \text{ECM}[\hat{Y} = a\sqrt{X} + b] = (1 - r_{\sqrt{X}Y}^2) S_Y^2.$$

Como:

$$\overline{\sqrt{X}} = \frac{\sqrt{36} + \sqrt{31} + \sqrt{26} + \sqrt{23} + \sqrt{20} + \sqrt{17} + \sqrt{14} + \sqrt{11} + \sqrt{8} + \sqrt{5}}{10}$$

$$= 4.2181,$$

$$\left( \overline{\sqrt{X}} \right)^2 = \bar{X} = 19.1,$$

$$\Rightarrow S_{\sqrt{X}}^2 = \bar{X} - \left( \overline{\sqrt{X}} \right)^2 = 19.1 - (4.2181)^2 = 1.3076,$$

$$\overline{\sqrt{X} \cdot Y} = \frac{\sqrt{36} \cdot 14.0 + \sqrt{31} \cdot 12.4 + \sqrt{26} \cdot 11.3 + \sqrt{23} \cdot 10.6 + \sqrt{20} \cdot 9.3}{10}$$

$$\frac{+\sqrt{17} \cdot 8.9 + \sqrt{14} \cdot 7.3 + \sqrt{11} \cdot 7.0 + \sqrt{8} \cdot 5.9 + \sqrt{5} \cdot 5.1}{10} = 41.8404,$$

$$\Rightarrow S_{\sqrt{X}Y} = \overline{\sqrt{X} \cdot Y} - \overline{\sqrt{X}} \cdot \bar{Y} = 41.8404 - 4.2181 \cdot 9.18 = 3.1182,$$

$$\frac{S_{\sqrt{X}Y}}{S_{\sqrt{X}}^2} = \frac{3.1182}{1.3076} = 2.3847,$$

$$\bar{Y} - \frac{S_{\sqrt{X}Y}}{S_{\sqrt{X}}^2} \cdot \overline{\sqrt{X}} = 9.18 - 2.3847 \cdot 4.2181 = -0.8789,$$

de donde la relación óptima del tipo considerado será:

$$\hat{Y} = 2.3847 \sqrt{X} - 0.8789.$$

El coeficiente de correlación asociado a esa relación óptima corresponderá a:

$$r_{\sqrt{X}Y} = \frac{3.1182}{\sqrt{1.3076 \cdot 7.5896}} = 0.9898,$$

que tratándose también de una relación de dependencia muy fuerte, es ligeramente inferior a la de tipo lineal. De hecho, el error cuadrático medio mínimo en el que puede incurrirse aproximando el valor de  $Y$  por una relación del tipo considerado vendría dado por:

$$\min_{a,b \in \mathbb{R}} \text{ECM}[\hat{Y} = a\sqrt{X} + b] = (1 - 0.9898^2) \cdot 7.5896 = 0.1540.$$

- En ocasiones, la relación funcional de  $Y$  con  $X$  es tal que el error cuadrático medio asociado es una función de ciertos términos (en principio, no especificados) que permite una determinación directa sencilla del mínimo absoluto. Así ocurre, por ejemplo, con las relaciones lineales con pendiente u ordenada en el origen conocida (como pueden ser  $\hat{Y} = a \cdot X$  con  $a \in \mathbb{R}$ , o  $\hat{Y} = X + b$  con  $b \in \mathbb{R}$ ) o rangos de posibles valores de  $a$  y  $b$  restringidos a ciertos subconjuntos reales, o con las relaciones polinómicas (como las de segundo grado,  $\hat{Y} = a \cdot X^2 + b \cdot X + c$  con  $a, b, c \in \mathbb{R}$ ).

En estos casos, la forma de resolver el problema de regresión consiste en plantear el error cuadrático medio para la relación considerada, determinando a continuación los valores de los términos para los que el error cuadrático medio alcanza el mínimo absoluto. De esta forma se hallaría la *relación óptima según el criterio de mínimos cuadrados*.

Para poder comparar la idoneidad de este tipo de relaciones con otros tipos distintos, habrá que recurrir a determinar el error cuadrático medio mínimo asociado a los tipos comparados.

Si, por ejemplo para los datos del ejemplo de regresión lineal tratamos de ajustar una relación del tipo

$$\hat{Y} = \log(X) + b$$

(en la que los logaritmos se entienden como neperianos) el término no especificado es  $b$ .

Se trata de una relación lineal de  $Y$  respecto a  $\log(X)$ , pero no es de tipo general (puesto que, por un lado, se supone que  $a = 1$ ).

Planteamos la función de  $b$  correspondiente al error cuadrático medio y estudiamos sus posibles mínimos. La función viene dada por:

$$\begin{aligned}
 G(b) &= \overline{[Y - (\log(X) + b)]^2} \\
 &= \overline{Y^2 + (\log(X))^2 + b^2 + 2 \log(X) b - 2 \log(X) Y - 2 Y b} \\
 &= \overline{Y^2} + \overline{(\log(X))^2} + b^2 + 2 \overline{\log(X)} b - 2 \overline{\log(X) Y} - 2 \overline{Y} b \\
 &= b^2 + 2 [\overline{\log(X)} - \overline{Y}] b + \overline{Y^2} + \overline{(\log(X))^2} - 2 \overline{\log(X) Y},
 \end{aligned}$$

que es una función polinómica de segundo grado en  $b$ . Para estudiar sus puntos críticos, puede hallarse la derivada primera:

$$\frac{dG(b)}{db} = 2b + 2 [\overline{\log(X)} - \overline{Y}] = 0 \Leftrightarrow b = \overline{Y} - \overline{\log(X)}.$$

Como

$$\begin{aligned}
 \overline{\log(X)} &= \frac{\log(36) + \log(31) + \log(26) + \log(23) + \log(20)}{10} \\
 &+ \frac{\log(17) + \log(14) + \log(11) + \log(8) + \log(5)}{10} = 2.7966, \\
 \overline{Y} = 9.18 &\Rightarrow \overline{Y} - \overline{\log(X)} = 6.3834 > 0,
 \end{aligned}$$

al ser

$$\frac{d^2 G(b)}{d^2 b} = 2 > 0$$

se concluye que la relación óptima es la que corresponde a la aproximación:

$$\widehat{Y} = \log(X) + 6.3834.$$

Si se quisiera comparar la relación

$$\widehat{Y} = \log(X) + 6.3834$$

con la óptima de tipo lineal, bastaría con comparar sus errores cuadrático medios.

En el caso de la relación lineal (general) óptima, éste sería el valor:

$$\min_{a,b \in \mathbb{R}} \text{ECM}[\widehat{Y} = a \cdot X + b] = (1 - r_{XY}^2) S_Y^2 = (1 - 0.9967^2) \cdot 7.5896 = 0.0500.$$

En el caso de la relación  $\widehat{Y} = \log(X) + 6.3834$ , el error cuadrático medio sería:

$$\begin{aligned}
 \min_{b \in (0, +\infty)} \text{ECM}[\widehat{Y} = \log(X) + b] &= \min_{b \in (0, +\infty)} G(b) = G(6.3834) \\
 &= 6.3834^2 + 2 [\overline{\log(X)} - \overline{Y}] \cdot 6.3834 + \overline{Y^2} + \overline{(\log(X))^2} - 2 \overline{\log(X) Y}
 \end{aligned}$$

que, al ser:

$$\begin{aligned} \overline{(\log(X))^2} &= \frac{(\log(36))^2 + (\log(31))^2 + (\log(26))^2 + (\log(23))^2 + (\log(20))^2}{10} \\ &\quad + \frac{(\log(17))^2 + (\log(14))^2 + (\log(11))^2 + (\log(8))^2 + (\log(5))^2}{10} = 8.1710, \\ \overline{\log(X)Y} &= \frac{\log(36) \cdot 14.0 + \log(31) \cdot 12.4 + \log(26) \cdot 11.3 + \log(23) \cdot 10.6 + \log(20) \cdot 9.3}{10} \\ &\quad + \frac{\log(17) \cdot 8.9 + \log(14) \cdot 7.3 + \log(11) \cdot 7.0 + \log(8) \cdot 5.9 + \log(5) \cdot 5.1}{10} = 27.2406, \\ \overline{Y^2} &= 91.862, \end{aligned}$$

viene dado por:

$$G(6.3834) = 6.3834^2 - 2 \cdot 6.3834 \cdot 6.3834 + 91.862 + 8.1710 - 2 \cdot 27.2406 = 4.8040,$$

que es muy superior al error asociado a la relación lineal general, por lo que ésta describiría de forma más adecuada la verdadera relación de  $Y$  respecto a  $X$  que la de tipo logarítmico que estábamos analizando en este apartado.

- Hay otras situaciones, en las que la relación funcional de aproximación de  $Y$  respecto a  $X$ ,  $Y = G(X)$  con  $G$  no especificada totalmente sino dependiente de ciertos términos, no se ajusta a ninguno de los dos casos precedentes (en otras palabras, no permite una aplicación directa sencilla del principio de mínimos cuadrados). En este caso se encontrarían, entre otras las relaciones de tipo exponencial  $Y = b \cdot e^{aX}$ , potencial  $Y = b \cdot X^a$ , hiperbólico  $Y = \frac{1}{a \cdot X + b}$ .

Si, a pesar de ello, fuera posible transformar la relación inicial para expresarla a través de una relación lineal entre variables  $Z = h(Y)$  y  $T = H(X)$  entonces podría determinarse una **aproximación de la solución óptima**, sin más que aplicar el método de mínimos cuadrados entre  $Z$  y  $T$ , y despejando posteriormente  $Y$  en la relación óptima.

Así las relaciones de tipo exponencial, potencial e hiperbólico, citadas anteriormente pueden transformarse de la forma siguiente.

$$Y = b \cdot e^{aX} \Leftrightarrow Z = a \cdot T + d \text{ con } a, d \in \mathbb{R}, \quad Z = \log Y, T = X, d = \log b,$$

$$Y = b \cdot X^a \Leftrightarrow Z = a \cdot T + d \text{ con } a, d \in \mathbb{R}, \quad Z = \log Y, T = \log X, d = \log b,$$

$$Y = \frac{1}{a \cdot X + b} \Leftrightarrow Z = a \cdot T + b \text{ con } a, b \in \mathbb{R}, \quad Z = \frac{1}{Y}, T = X.$$

El procedimiento que se sigue en estos casos consiste simplemente en:

- hallar la relación óptima mínimo-cuadrática entre  $Z$  y  $T$ ;
- despejar  $Y$  en la relación óptima anterior. Esta última relación será una buena aproximación de la óptima.

**Observación:** En general, esta aproximación de la relación óptima no coincidirá exactamente con la que habríamos obtenido si hubiéramos podido aplicar directamente mínimos cuadrados a la relación original  $Y = G(X)$ . Esta afirmación se debe a que al transformar la variable  $Y$  el criterio de mínimos cuadrados busca el error cuadrático medio mínimo entre los transformados, y no entre los originales.

En general tampoco será cierto que el error cuadrático medio mínimo de la relación transformada, coincida con el de la relación inicial (salvo que uno de ellos se anule, en cuyo caso también lo hará el otro).

Para poder comparar la idoneidad de este tipo de relaciones con otros tipos distintos, habrá que recurrir a determinar el error cuadrático medio mínimo asociado a los tipos comparados. Para los casos que incluimos en este apartado, el error cuadrático medio debería hallarse aplicando la expresión original del error.

Alternativamente, y siguiendo una idea similar a la de la correlación lineal, podría considerarse el coeficiente adimensional

$$z_{XY}^2 = 1 - \frac{\text{ECM} [\hat{Y} = G^*(X)]}{S_Y^2},$$

denominado *razón de correlación* (que coincidiría con  $r_{XY}^2$  en el caso de que  $G$  fueran todas las funciones lineales generales de  $Y$  respecto a  $X$ ) y que tomaría valores inferiores o iguales a 1 con interpretaciones similares a las del coeficiente de determinación. Las comparaciones a través de la razón de correlación (eligiendo como más idóneo el tipo de relación con mayor razón de correlación) coincidiría con las comparaciones a través del error cuadrático medio.

Si, por ejemplo para los datos del ejemplo de regresión lineal tratáramos de ajustar una relación del tipo potencial

$$\hat{Y} = b \cdot X^a \quad \text{con } a, b \in \mathbb{R},$$

se procedería como sigue:

- como

$$Y = b \cdot X^a \Leftrightarrow \log Y = a \cdot \log X + \log b \quad \text{con } a, \log b \in \mathbb{R},$$

se hallaría la relación óptima del tipo lineal entre  $Z = \log Y$  y  $T = \log X$ , que sería:

$$Z - \bar{Z} = \frac{S_{ZT}}{S_T^2}(T - \bar{T}).$$

Como

$$\begin{aligned} \bar{Z} = \overline{\log(Y)} &= \frac{\log(14.0) + \log(12.4) + \log(11.3) + \log(10.6) + \log(9.3)}{10} \\ &+ \frac{\log(8.9) + \log(7.3) + \log(7.0) + \log(5.9) + \log(5.1)}{10} = 2.1696, \end{aligned}$$

$$\begin{aligned} \overline{ZT} = \overline{\log(Y) \cdot \log(Y)} &= \frac{\log(36) \cdot \log(14.0) + \log(31) \cdot \log(12.4)}{10} \\ &+ \frac{\log(26) \cdot \log(11.3) + \log(23) \cdot \log(10.6) + \log(20) \cdot \log(9.3)}{10} \\ &+ \frac{\log(17) \cdot \log(8.9) + \log(14) \cdot \log(7.3) + \log(11) \cdot \log(7.0)}{10} \\ &+ \frac{\log(8) \cdot \log(5.9) + \log(5) \cdot \log(5.1)}{10} = 6.2505, \end{aligned}$$

$$S_{TZ} = S_{\log(X) \log(Y)} = 6.2505 - 2.7966 \cdot 2.1696 = 0.1830,$$

$$S_T^2 = S_{\log(X)}^2 = 8.1710 - 2.7966^2 = 0.3500,$$

la relación óptima del tipo  $Z = a \cdot T + d$  se obtiene para

$$a^* = \frac{0.1830}{0.3500} = 0.5222 \quad d^* = 2.1696 - 0.5229 \cdot 2.7966 = 0.7083.$$

– Despejando  $Y$  en función de  $X$  en la relación precedente, tendremos una buena aproximación de la relación óptima del tipo  $Y = b \cdot X^a$  la dada por :

$$Y = e^{d^*} \cdot X^{a^*} \equiv Y = 2.0321 \cdot X^{0.5222}.$$

Para comparar esta relación con otras anteriores, podríamos hallar el error cuadrático medio asociado, que vendría dado por:

$$\begin{aligned} \text{ECM} \left[ \hat{Y} = 2.0321 \cdot X^{0.5222} \right] &= \frac{1}{10} \sum_{t=1}^{10} [Y(\omega_t) - 2.0321 \cdot X(\omega_t)^{0.5222}]^2 \\ &= \frac{[14.0 - 2.0321 \cdot 36^{0.5222}]^2 + [12.4 - 2.0321 \cdot 31^{0.5222}]^2 + [11.3 - 2.0321 \cdot 26^{0.5222}]^2}{10} \\ &+ \frac{[10.6 - 2.0321 \cdot 23^{0.5222}]^2 + [9.3 - 2.0321 \cdot 20^{0.5222}]^2 + [8.9 - 2.0321 \cdot 176^{0.5222}]^2}{10} \\ &+ \frac{[7.3 - 2.0321 \cdot 14^{0.5222}]^2 + [7.0 - 2.0321 \cdot 11^{0.5222}]^2}{10} \\ &+ \frac{[5.9 - 2.0321 \cdot 8^{0.5222}]^2 + [5.1 - 2.0321 \cdot 5^{0.5222}]^2}{10} = 0.1636, \end{aligned}$$

de manera que las relaciones de tipo potencial son menos adecuadas que las lineales generales y un poco menos que las del tipo  $Y = a\sqrt{X} + b$ .