Modelos de Regresión

En muchos problemas existe una relación inherente entre dos o más variables, y resulta necesario explorar la naturaleza de esta relación. El **análisis de regresión** es una técnica estadística para el modelado y la investigación de la relación entre dos o más variables. Por ejemplo, en un proceso químico, supóngase que el rendimiento del producto está relacionado con la temperatura de operación del proceso. El análisis de regresión puede emplearse para construir un modelo que permita predecir el rendimiento para una temperatura dada. Como ilustración, considérense los datos de la siguiente tabla. En ella, Y es la pureza del oxígeno producido en un proceso de destilación químico, y x es el porcentaje de hidrocarburos presentes en el condensador principal de la unidad de destilación. La figura 1 presenta el **diagrama de dispersión** de los datos contenidos en la Tabla 1. El análisis de este diagrama de dispersión indica que, si bien una curva no pasa exactamente por todos los puntos, existe una evidencia fuerte de que los puntos están dispersos de manera aleatoria alrededor de una línea recta. Por consiguiente es razonable suponer que la media de la variable aleatoria Y está relaciona con x por la siguiente relación lineal:

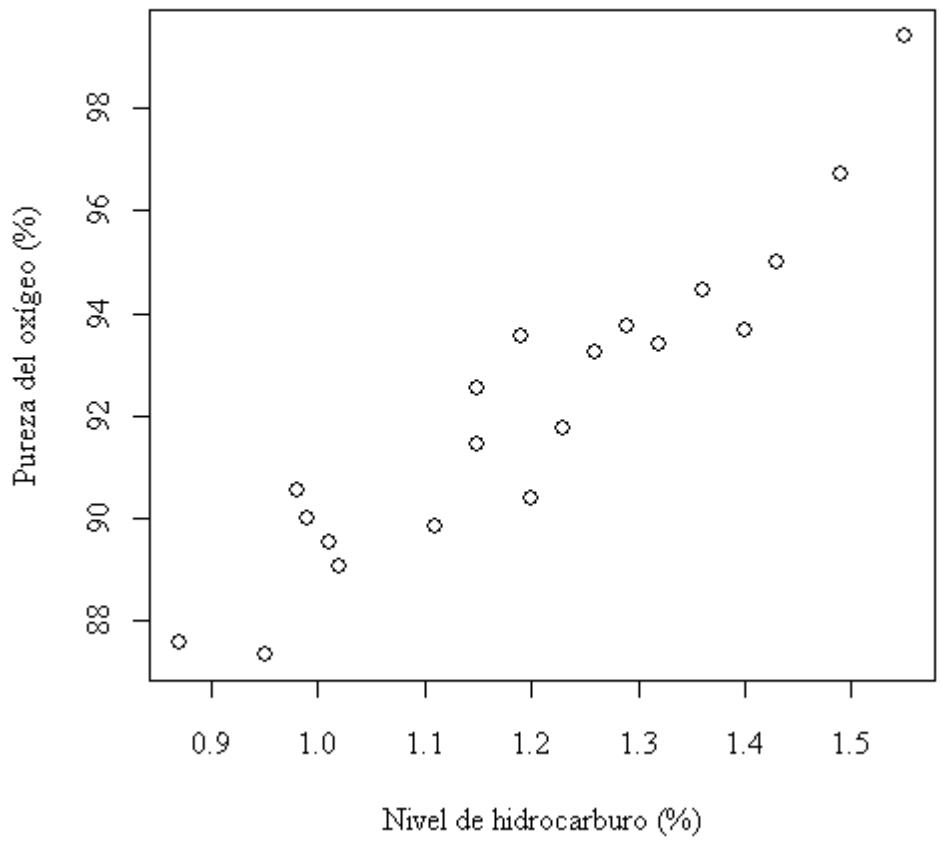
$$E(Y|x) = \beta_0 + \beta_1 x$$

Donde la pendiente y la ordenada al origen de la recta reciben el nombre de **coeficientes de regresión**.

Tabla 1

Número de observación	Nivel de Hidrocarburo x(%)	Pureza Y(%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.49	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

Figura 1


Pureza del oxígeo (%)
88
90
92
94
96
98
0.9
1.0
1.1
1.2
1.3
1.4
1.5
Nivel de hidrocarburo (%)

Si bien la media de Y es una función lineal de *x*, el valor real observado de Y no cae de manera exacta sobre la recta. La manera apropiada para generalizar este hecho con un **modelo probabilístico lineal** es suponer que el valor esperado de Y es una función lineal de *x*, pero que para un valor fijo de *x* el valor real de Y está determinado por el valor medio de la función (el modelo lineal) más un término que representa un error aleatorio, por ejemplo,

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

donde ε es el error aleatorio. Este modelo recibe el nombre de **modelo de regresión lineal** simple, ya que sólo tiene una variable independiente o **regresor**.

Regresión Lineal Simple

El caso de la **regresión lineal simple** considera sólo un *regresor* o *predictor x*, y una variable dependiente o *respuesta* Y. Supóngase que la verdadera relación entre Y y *x* es una línea rect, y que la observación Y en cada nivel *x* es una variable aleatoria. Tal como ya se indicó, el valor esperado de Y para cada valor de *x* es

$$E(Y|x) = \beta_0 + \beta_1 x$$

donde la ordenada al origen β_0 y la pendiente β_1 son los coeficientes desconocidos de la regresión. Se supone que cada observación, Y, puede describirse por el modelo

$$Y = \beta_0 + \beta_1 x + \varepsilon \qquad \textbf{(1)}$$

donde ε es un error aleatorio con media cero y varianza σ^2. También se supone que los errores aleatorios que corresponden a observaciones diferentes son variables aleatorias no correlacionadas.

Las estimaciones de β_0 y β_1 deben dar como resultado una línea que (en algún sentido) se "ajuste mejor" a los datos. El científico alemán Karl Gauss (1777-1855) propuso estimar los parámetros β_0 y β_1 de la ecuación (**1**) de modo que se minimice la suma de los cuadrados del error. Este criterio para estimar los coeficientes de regresión se conoce como **método de mínimos cuadrados**. Al utilizar la ecuación **(1)**, es posible expresar las *n* observaciones de la muestra como:

$$Y_i = \beta_0 + \beta_1 x + \varepsilon_i, \qquad i = 1,2,\ldots,n$$

y la suma de los cuadrados de las desviaciones de las observaciones con respecto a la recta de regresión es

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (\mathrm{Y}_i - \beta_0 - \beta_1 x_i)^2 .$$

Los estimadores de mínimos cuadrados de β_0 y β_1, $\hat{\beta}_0$ y $\hat{\beta}_1$, deben satisfacer las ecuaciones siguientes:

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_o, \hat{\beta}_1} = -2\sum_{i=1}^{n} (\mathrm{Y}_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_o, \hat{\beta}_1} = -2\sum_{i=1}^{n} (\mathrm{Y}_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

Después de simplificar las expresiones anteriores, se tiene que

$$n\hat{\beta}_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{\mathrm{i}=1}^{n} \mathrm{Y}_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{\mathrm{i}=1}^{n} x_i \mathrm{Y}_i$$

Las últimas ecuaciones reciben el nombre de **ecuaciones normales de mínimos cuadrados.** La solución de estas ecuaciones dan como resultado los estimadores de mínimos cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$.

Definición

Las **estimaciones de mínimos cuadrados** de la ordenada al origen y la pendiente del modelo de regresión lineal simple son

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{y} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

donde $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ y $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Por tanto, la **línea de regresión estimada** o **ajustada** es

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, \ldots, n$$

Los **residuos** se determinan como $\hat{\varepsilon}_{\mathrm{i}} = y_i - \hat{y}_i$. El residuo describe el error en el ajuste del modelo en la *i*-ésima observación y_i. Los residuos proporcionan información sobre la adecuación del modelo ajustado.

Notación

En ocasiones es conveniente dar símbolos especiales al numerador y denominador en las fórmulas de los estimadores de mínimos cuadrados.
Dados los datos $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ sean

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i \right)^2}{n}$$

y

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

A partir de esta notación, los estimadores de mínimos cuadrados serán:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Ejemplo

Es momento de ajustar un modelo de regresión lineal simple a los datos de pureza del oxígeno de la Tabla 1. Con esos datos pueden calcularse las cantidades siguientes:

$$n = 20, \quad \sum_{i=1}^{20} x_i = 23.92, \quad \sum_{i=1}^{20} y_i = 1843.21, \quad \bar{x} = 1.20, \quad \bar{y} = 92.16$$

$$\sum_{i=1}^{20} y_i^2 = 170044.53, \quad \sum_{i=1}^{20} x_i^2 = 29.29, \quad \sum_{i=1}^{20} x_i y_i = 2214.66$$

$$S_{xx} = \sum_{i=1}^{20} x_i^2 - \frac{\left(\sum_{i=1}^{20} x_i\right)^2}{20} = 29.29 - \frac{(23.92)^2}{20} = 0.68$$

y

$$S_{xy} = \sum_{i=1}^{20} x_i y_i - \frac{\left(\sum_{i=1}^{20} x_i\right)\left(\sum_{i=1}^{20} y_i\right)}{20} = 2214.66 - \frac{(23.92)(1843.21)}{20} = 1018$$

Por consiguiente, las estimaciones de mínimos cuadrados de la pendiente y la ordenada al origen son

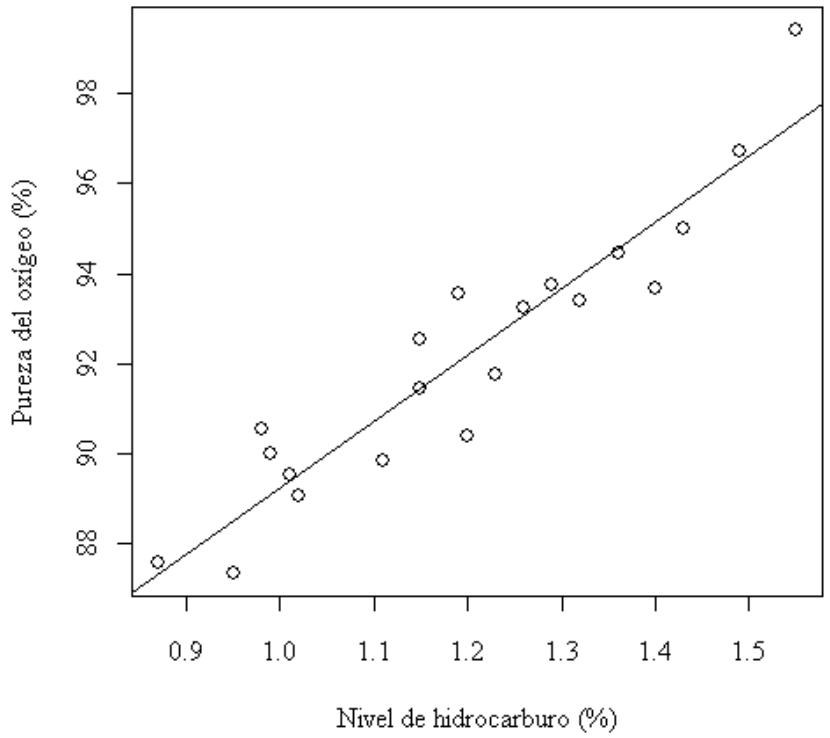
$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{18.18}{0.68} = 14.97 \quad \text{y} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 92.16 - (14.97)1.20 = 74.20$$

El modelo de regresión lineal simple ajustado es

$$\hat{y}_i = 74.20 + 14.97 x_i, \quad i = 1,2,\ldots,20$$

La gráfica de este modelo aparece en la figura 2, junto con los datos de la muestra.

Figura 2



Con el empleo del modelo de regresión ajustado, es posible predecir una pureza de oxígeno de $\hat{y} = 89.17\%$ cuando el nivel de hidrocarburo es $x = 1.00\%$. La pureza de 89.17% puede interpretarse como una estimación de la pureza promedio verdadera de la población cuando $x = 1.00\%$, o como una estimación de la nueva observación cuando $x = 1.00\%$. Claro está que estas estimaciones se encuentran sujetas a un error; esto es, es poco probable que una observación futura de la pureza sea exactamente 89.17% cuando el nivel de hidrocarburo sea de 1.00%. En secciones subsecuentes se verá cómo utilizar los intervalos de confianza y los de predicción para describir el error al hacer estimaciones a partir de un modelo de regresión.

Propiedades de los estimadores de mínimos cuadrados y estimación de σ^2

Resulta sencillo describir las propiedades estadísticas de los estimadores de mínimos cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$. Recuérdese que se ha supuesto que el término de error ε en el modelo $Y = \beta_0 + \beta_1 x + \varepsilon$ es una variable aleatoria con media cero y varianza σ^2. Puesto que los valores de *x* son fijos, Y es una variable aleatoria con media $\mu_{Y|x} = \beta_0 + \beta_1 x$ y varianza σ^2. Por consiguiente, los valores de $\hat{\beta}_0$ y $\hat{\beta}_1$ dependen de los valores de *y* observados; por tanto, los estimadores de mínimos cuadrados de los coeficientes de

regresión pueden verse como variables aleatorias. A continuación se investiga el sesgo y las propiedades de la varianza de los estimadores de mínimos cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$.

No es difícil demostrar que $\hat{\beta}_0$ y $\hat{\beta}_1$ son estimadores insesgados de β_0 y β_1, respectivamente, es decir $\mathrm{E}\hat{\beta}_0 = \beta_0$ y $\mathrm{E}\hat{\beta}_1 = \beta_1$.

Por otro lado, se tiene:

$$\mathrm{Var}\hat{\beta}_0 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right], \quad \mathrm{Var}\hat{\beta}_1 = \frac{\sigma^2}{S_{xx}} \quad \text{y} \quad \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2\frac{\bar{x}}{S_{xx}}$$

Para obtener inferencias con respecto a los coeficientes de regresión β_0 y β_1, es necesario estimar la varianza σ^2 que aparece en las expresiones para $\mathrm{Var}\hat{\beta}_0$ y $\mathrm{Var}\hat{\beta}_1$. El parámetro σ^2, que es la varianza del término de error ε en el modelo de regresión, refleja la variación aleatoria alrededor de la verdadera recta de regresión.

La estimación para σ^2, $\hat{\sigma}^2$, esta dada por

$$\hat{\sigma}^2 = \frac{SS_E}{n-2} \qquad \textbf{(2)}$$

donde

$$\begin{aligned} SS_E &= \sum_{i=1}^{n} \hat{\varepsilon}_i \\ &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \hat{\beta}_1 S_{xy} \\ &= S_{yy} - \hat{\beta}_1 S_{xy}, \quad \text{con} \quad S_{yy} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 \end{aligned}$$

Ejemplo (contin.)

A continuación se encuentra la estimación de la varianza σ^2 utilizando para ello los datos de la Tabla 1, obteniendose $\hat{\sigma}^2 = 1.17$.

Definición

En una regresión lineal simple, el **error estándar estimado de la pendiente** es

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

y el **error estándar de la ordenada al origen** es

$$se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]}$$

donde $\hat{\sigma}^2$ se calcula con la ecuación (**2**).