Departamento de Ingeniería Biomédica
Universidad de los Andes

Regresión Lineal

Modelo de regresión lineal múltiple

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon,$$

Donde y es la variable independiente, $x_1, x_2, \dots, x_n$ las variables independientes y ε el error asociado.

Modelo de regresión lineal simple

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

Sólo se tiene una variable independiente y dos parámetros de regresión $\beta_0\ y\ \beta_1$

Estimación de parámetros: Método de mínimos cuadrados

$$\begin{array}{rcccccccl}
n\hat{\beta}_0 & + & \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} & + & \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} & + \cdots + & \hat{\beta}_k \sum_{i=1}^{n} x_{ik} & = & \sum_{i=1}^{n} y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} & + & \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 & + & \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} & + \cdots + & \hat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} & = & \sum_{i=1}^{n} x_{i1}y_i \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} & + & \hat{\beta}_i \sum_{i=1}^{n} x_{ik}x_{i1} & + & \hat{\beta}_2 \sum_{i=1}^{n} x_{ik}x_{i2} & + \cdots + & \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 & = & \sum_{i=1}^{n} x_{ik}y_i
\end{array}$$

Se resuelve el sistema de ecuaciones para hallar los coeficientes $\beta_0, \dots, \beta_n$

Notación matricial

El sistema expuesto arriba en notación matricial es $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, donde la primera columna de la matriz $\boldsymbol{X}$ es una columna de 1's.

El modelo de regresión es $\boldsymbol{y}^{\wedge} = \boldsymbol{X\beta}^{\wedge}$**,** donde los coeficientes $\boldsymbol{\beta_i}$ están dados por:

$$\boldsymbol{\beta}^{\wedge} = (\boldsymbol{X}'\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}'\boldsymbol{y}$$

Ejemplo:

Table 10-2 Viscosity Data for Example 10-1 (viscosity in centistokes @ 100°C)

Observation	Temperature (x_1, °C)	Catalyst Feed Rate (x_2, lb/h)	Viscosity
1	80	8	2256
2	93	9	2340
3	100	10	2426
4	82	12	2293
5	90	11	2330
6	99	8	2368
7	81	8	2250
8	96	10	2409
9	94	12	2364
10	93	11	2379
11	97	13	2440
12	95	11	2364
13	100	8	2404
14	85	12	2317
15	86	9	2309
16	87	12	2328

La estimación de mínimos cuadrados de $\boldsymbol{\beta}$ es

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$\mathbf{X} = \begin{bmatrix} 1 & 80 & 8 \\ 1 & 93 & 9 \\ 1 & 100 & 10 \\ 1 & 82 & 12 \\ 1 & 90 & 11 \\ 1 & 99 & 8 \\ 1 & 81 & 8 \\ 1 & 96 & 10 \\ 1 & 94 & 12 \\ 1 & 93 & 11 \\ 1 & 97 & 13 \\ 1 & 95 & 11 \\ 1 & 100 & 8 \\ 1 & 85 & 12 \\ 1 & 86 & 9 \\ 1 & 87 & 12 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} 2256 \\ 2340 \\ 2426 \\ 2293 \\ 2330 \\ 2368 \\ 2250 \\ 2409 \\ 2364 \\ 2379 \\ 2440 \\ 2364 \\ 2404 \\ 2317 \\ 2309 \\ 2328 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 80 & 93 & \cdots & 87 \\ 8 & 9 & \cdots & 12 \end{bmatrix} \begin{bmatrix} 1 & 80 & 8 \\ 1 & 93 & 9 \\ \vdots & \vdots & \vdots \\ 1 & 87 & 12 \end{bmatrix} = \begin{bmatrix} 16 & 1458 & 164 \\ 1458 & 133{,}560 & 14{,}946 \\ 164 & 14{,}946 & 1{,}726 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 80 & 93 & \cdots & 87 \\ 8 & 9 & \cdots & 12 \end{bmatrix} \begin{bmatrix} 2256 \\ 2340 \\ \vdots \\ 2328 \end{bmatrix} = \begin{bmatrix} 37{,}577 \\ 3{,}429{,}550 \\ 385{,}562 \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1566.07777 \\ 7.62129 \\ 8.58485 \end{bmatrix}$$

$$\hat{y} = 1566.08 + 7.62x_1 + 8.58x_2$$

<u>Análisis de varianza del modelo de regresión</u>

La variación de la variable de respuesta y está dada por:

$$SS_T = SS_R + SS_E$$

Donde,

SS_T es la suma de cuadrados total definida por: $\sum_{i=1}^{n}(y_i - media(y))^2$

SS_R es la suma de cuadrados de la regresión definida por: $\sum_{i=1}^{n}(y^{\wedge}{}_i - media(y))^2$

SS_E es la suma de cuadrados del error definida por: $\sum_{i=1}^{n}(y_i - y^{\wedge}{}_i)^2$

Así, el coeficiente de determinación (bondad de ajuste del modelo) es,

$$R^2 = \frac{SS_R}{SS_T} \qquad 0 \le R^2 \le 1$$

R^2 indica qué porcentaje de la variabilidad de y es explicada por su relación lineal con X.

Referencia

Desing and Analysis of Experiments. Douglas C. Montgomery. 8th Edition. John Wiley & Sons, Inc. 2013