Multiple Regression

2019年6月28日 16:38

多元回归分析

有多个自变量.

模型: $y=\beta_0+\beta_1 x+\beta_2 x_2+\cdots+\beta_p x_p+\varepsilon$ (允许形成曲面甚至)

其中: $\beta_0,\beta_1\cdots\beta_p$ 为参数, ε 为误差.

方程: $E(y)=\beta_0+\beta_1 x_1+\cdots+\beta_p x_p$

估计方程

$y=b_0+b_1 x+\cdots+b_p x_p$.

估计流程: 与 simple linear regression 相似.

估计方法: min the sum of squares

即 $\min\Sigma(y_i-\hat{y}_i)^2$

例:

X_1: 运输距离 (km)

X_2: 运输次数.

Y: 运输总时间.

	X_1	X_2	Y
①	100	4	9.2
②	50	3	4.8
③	100	4	8.9
④	100	2	6.5

③	100	4	8.9
④	100	2	6.5
⑤	50	2	4.2
⑥	80	2	6.2
⑦	75	3	7.4
⑧	65	4	6.0
⑨	90	3	7.6
⑩	90	2	6.1

$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$

最小二乘法

要使 $Q = \Sigma (y_i - \hat{y}_i)^2$ min

即使 $Q = \Sigma (y_i - \hat{y}_i) = \Sigma (y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 \cdots - \beta_p x_p)$ min.

Q分别对 $\beta_0, \beta_1 \cdots \beta_p$ 求偏导数，令其为0 即得

$$\begin{cases} \frac{\partial Q}{\partial \beta_0} = \Sigma (y_i - \beta_0 - \beta_1 x_1 - \cdots - \beta_p x_p)(-1) = 0 \\ \frac{\partial Q}{\partial \beta_1} = \Sigma (y_i - \beta_0 - \beta_1 x_1 - \cdots - \beta_p x_p)(-x_1) = 0 \\ \vdots \\ \frac{\partial Q}{\partial \beta_p} = \Sigma (y_i - \beta_0 - \beta_1 x_1 - \cdots - \beta_p x_p)(-x_p) = 0 \end{cases}$$

解上述方程组即得各参数的估计值 $\beta_0, \beta_1 \cdots \beta_p$.

多元得：$\hat{y} = -0.869 + 0.0611 x_1 + 0.923 x_2$.

b_1：平均多送 1km，时间延长 0.0611h

b_2：平均多送 1次，时间延长 0.923h.

距离 $x_1 = 102$
次数 $x_2 = 6$ ↓ 预测.

$\hat{y} =$ [illegible]

$$\hat{y} = -0.887 + 0.0611 \times 102 + a_{23} \times 6$$
$$= 10.98.$$

注：自变量中不仅可以有连续型变量，也可以有离散型变量或类别型变量。

如在上例中加入 x_3（车的类别） 0, 1, 2.

若在数据中有类别型变量则将其转化为 0,1… 变量即可。

如：车型 1，车型 2，车型 3.

则 车 1：(1,0,0)

车 2：(0,1,0)

车 3：(0,0,1)