Relevancy and R^2 in Regression

2019年7月1日 10:34

回归中的相关度和决定系数.

Pearson Correlation Coefficient.
(皮尔逊相关系数)

用于衡量两个值的线性相关的强度. $r \in (-1,1)$.

r: 正向相关 >0, 负向相关 <0, 无相关性 =0.

$$\rho = Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

即 $$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \cdot \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

R^2 (决定系数).

若模型自变量不变, 则因变量的变异程度会减少 R^2.

如: $R^2 = 0.8$

即若自变量 $x=x_0$ 则有 80% 的 $y=\hat{y}$ 而 20% 的 $y \neq \hat{y}$.

对于 simple linear regression

$$R^2 = r \cdot r.$$

对于 multiply regression.

对于 multiply regression.

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

注: SSR → regression 回归模型 : $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ 由于模型的对应而引起的差异.

Sum sqrt 平方和

SST → total 总和 : $\sum_{i=1}^{n}(y_i - \bar{y})^2$ 总的差异量.

SSE → error 误差值 : $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ 总的误差值.

$$SSR + SSE = SST.$$

局限性: R^2会随着自变量的增多从而变大.

修正: 样本总数.

$$R^2_{adjusted} = 1 - \frac{(1-R^2)(N-1)}{N-P-1}$$

修正. 预测量总数.