Name- Suresh Kumar Prajapati

Class- - M.Sc. (Master in data Science and Applied statistics).

Enroll No— CUSB2302222008

Assignment work

20/03/2024 08:57

Define the coefficient of determination of R^2 and adjusted R^2 for multiple linear regression :-

Let R be the multiple correlation coefficient between y and $X_1, X_2, \ldots X_k$. Then square of multiple correlation coefficient (R^2) is called a coefficient of determination.

The value of R^2 commonly describes how well the sample regression line fits the observed data.

This is also treated as a measure of goodness of fit of the model.

Assuming that the intercept is present in the model as

$$y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \cdots + \beta_k X_{ik} + u_i$$

$i = 1, 2, \ldots n$.

then,

$$R^2 = 1 - \frac{e'e}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{SS_{res}}{SS_T} = \frac{SS_{reg}}{SS_T}$$

where, SS_{res} = sum of squares due to residuals,

SS_T = Total sum of squares.

SS_{reg} = Sum of squares due to regression.

$\because e'e = y'\left[I - X(X'X)^{-1}X'\right]y = y'\bar{H}y$

$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n} l'y$ with $l = (1, 1, \ldots 1)'$

$y = (y_1, y_2 \ldots, y_n)'$

So, $R^2 = 1 - \dfrac{y'\bar{H}y}{y'Ay}$

The limits of R^2 are 0 and 1, i.e,

$0 \le R^2 \le 1$

$R^2 = 0$ indicates the poorest fit of the model.

$R^2 = 1$ indicated the best fit of the model.

$R^2 = 0.95$ indicates that 95% of the variation in Y is explained by R^2.

In simple word, the model is 95% good.

Remark:- Any other value of R^2 between 0 and 1 indicates the adequacy of the fitted model.

Adjusted R^2:-

If more explanatory variables are added to the model, then R^2 increases. In case the variables are irrelevant, then R^2 will still increase and gives an overly optimistic picture.

With a purpose of correction in the overly optimistic picture, adjusted R^2, denoted as $\bar{R}^2$ or adj R^2 is used which is defined as

$$\bar{R}^2 = 1 - \frac{SS_{res}/(n-k)}{SS_T/(n-1)}$$

$$\boxed{\bar{R}^2 = 1 - \left(\frac{n-1}{n-k}\right)(1-R^2)}$$

$(n-k)$ & $(n-1)$ are degree of freedom associated with the distribution of SS_{res} and SS_T respectively

Limitation:-

(1) If the constant term is absent in the model, then R^2 can not be defined. In such cases R^2 can be negative. Some ad-hoc measures based on R^2 for regression line through origin have been proposed in the literature.

Reason that why R^2 is valid only in linear models with intercept term:-

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{y})^2$$

$$= SS_{reg} + SS_{res} - 2\bar{y}\,l'e \quad (\text{because } X'e = 0)$$

Fisher Cochran theorem requires

$TSS = SS_{reg} + SS_{res}$ to hold true in the context of analysis of variance and further to define by R^2. true only. $l'e$ should be zero. i.e. $l'e = l'(y - \hat{y}) = 0$ which is possible only when there is an intercept term in the model.

Some key points:-

① A simple linear regression model with intercept term $y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (i = 1, 2, \dots n)$

where the parameter β_0 and β_1 are estimated as $b_0 = \bar{y} - \beta_1 \bar{x}$, $\quad b_1 = \frac{S_{xy}}{S_{xx}}$

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

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

$$l'e = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i)$$

$$= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)$$

$$= \sum_{i=1}^{n} (y_i - \bar{y} + b_1 \bar{x} - b_1 x_i)$$

$$= \sum_{i=1}^{n} [(y_i - \bar{y}) - b_1 (x_i - \bar{x})]$$

$$= \sum_{i=1}^{n} (y_i - \bar{y}) - b_1 \sum_{i=1}^{n} (x_i - \bar{x})$$

$$= 0$$

In case of multiple regression

$y = \beta_0 l + X\beta + \epsilon$ where parameter β_0 and β are estimated as $\hat{\beta}_0 = \bar{y} - b\bar{x}$ & $b = (X'X)^{-1}X'y$ respectively

$$l'e = l'(y - \hat{y})$$

$$= l'(y - \hat{\beta}_0 - Xb)$$

$$= l''(y - \bar{y} + \bar{x}b - Xb)$$

20/03/2024 08:

$$\ell'(y-\bar{y}) + \ell(x-\bar{x})b = 0$$

② R^2 is sensitive to extreme values, so R^2 lacks robustness.

③ R^2 always increase with an increase in the number of explanatory variable in the model. This indicates that the model is getting better, which is not really correct.

Relationship of analysis of variance test and coefficient of determination :-

Assuming the β_1 to be an intercept term, then for $H_0 : \beta_2 : \beta_3 = \dots = \beta_k = 0$, then F-statistic in analysis of variance test is.

$$F = \frac{MS_{reg}}{MS_{res}} = \frac{(n-k)\,SS_{reg}}{(k-1)\,SS_{res}} = \frac{(n-k)\,SS_{reg}}{(k-1)\,(SS_T - SS_{reg})}$$

$$\boxed{F = \frac{(n-k)\,R^2}{(k-1)\,(1-R^2)}}$$

So, R^2 is the coefficient of determination.

So F and R^2 are closely related.

when $R^2 = 0$, then $F = 0$

when $R^2 = 1$, then $F = \infty$

So both F and R^2 vary directly.

If F is highly significant, it implies that we can reject H_0 i.e. y is linearly related to X's.

Prove that OLSE is the best linear unbiased estimator of beta for M.R.:- Gauss-Markov Theorem:-

Proof:- The OLSE of β is

$$b = (X'X)^{-1}X'y$$

which is linear function of y. Consider the arbitrary linear estimator

$$b^* = a'y$$

of linear parametric function $l'\beta$ where the elements of a are arbitrary constants.

Then for b^*

$$E(b^*) = E(a'y) = a'X\beta$$

and so b^* is an unbiased estimator of $l'\beta$ when

$$E(b^*) = a'X\beta = l'\beta$$

$$\Rightarrow a'X = l$$

Since we wish to consider only those estimators that are linear and unbiased, so we restrict ourselves to those estimators for which

$$a'X = l'$$

20/03/2024 08:

Furthers,

$$\text{Var}(a'y) = a' \text{var}(y)\, a = \sigma^2 a'a$$

$$\text{Var}(l'b) = l' \text{var}(b)\, l$$

$$= \sigma^2 a' X (X'X)^{-1} X' a$$

Consider,

$$\text{Var}(a'y) - \text{var}(l'b) = \sigma^2 a' [I - X(X'X)^{-1}X'] a$$

$$= \sigma^2 a' [I - H] a$$

$\because (I - H)$ is a +ve semi-definite matrix, so

$$\text{Var}(a'y) - \text{var}(l'b) \geq 0$$

This reveals that if b^* is any linear unbiased estimator then its variance must be no smaller than that of b.

Consequently b is the best linear unbiased estimator, where 'best' refers to the fact that b is efficient with the class of linear and unbiased estimators.

Likelihood Ratio test :-

The likelihood ratio test statistic is:

$$\lambda = \frac{\max L(\beta, \sigma^2 | y, X)}{\max L(\beta, \sigma^2 | y, X, R\beta = r)} = \frac{\hat{L}(\Omega)}{\hat{L}(\omega)}$$

where Ω is the whole parametric space and ω is the sample space.

20/03/2024 08

If both the likelihoods are maximized, one constrained and the other unconstrained then the value of the unconstrained will not be smaller than the value of constrained. Hence $\lambda \geq 1$.

First, we discuss the likelihood ratio test for a more straightforward case when

$R = I_k$ and $r = \beta_0$ i.e. $\beta = \beta_0$

This will give us a better and detailed understanding of the minor details and then we generalized it for $R\beta = r$, in general.

Likelihood ratio test for $H_0: \beta = \beta_0$

Let the null hypothesis related to $K \times 1$ vector β is

$$H_0: \beta = \beta_0$$

where β_0 is specified by the investigator. The element of β_0 can take on any value including zero. The concerned alternative hypothesis is.

$$H_1: \beta \neq \beta_0$$

$\because \epsilon \sim N(0, \sigma^2 I)$ in $y = X\beta + \epsilon$ so.

$y \sim N(X\beta, \sigma^2 I)$. Thus the whole parametric space and sample space are Ω and ω respectively given by

20/03/2024 08:

$$\Omega = \{(\beta, \sigma^2): -\infty < \beta_i < \infty, \sigma^2 > 0 \quad i = 1, 2, \ldots k\}$$

$$\omega : \{(\beta, \sigma^2) : \beta = \beta_0, \sigma^2 > 0\}$$

The unconstrained likelyhood under Ω

$$L(\beta, \sigma^2 | y, X) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right]$$

This is maximized over Ω when

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

$$\hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})(y - X\hat{\beta})$$

where $\hat{\beta}$ and $\hat{\sigma}^2$ are maximum likelihood estimates of β and σ^2 which are values maximizing the likelyhood function

$$\hat{L}(\Omega) = \max L(\beta, \sigma^2 | y, X)$$

$$= \frac{1}{\left[\frac{2\pi}{n}(y - X\hat{\beta})'(y - X\hat{\beta})\right]^{n/2}} \exp\left[-\frac{(y - X\tilde{\beta})'(y - X\tilde{\beta})}{\left(\frac{2(y - X\tilde{\beta})'(y - X\tilde{\beta})}{n}\right)}\right]$$

$$= \frac{n^{n/2} \exp\left(-\frac{n}{2}\right)}{\left(\frac{2\pi}{n}\right)^{n/2}\left[(y - X\tilde{\beta})'(y - X\tilde{\beta})\right]^{n/2}}$$

The constrained likelyhood under ω is

$$\hat{L}(\omega) = \max L(\beta, \sigma^2 | y, X, \beta = \beta_0)$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta_0)'(y - X\beta_0)\right]$$

20/03/2024 0

$$\hat{L}(\omega) = \frac{n^{n/2} \exp\left(-\frac{n}{2}\right)}{(2\pi)^{n/2}\left[(y - x\beta_0)'(y - x\beta_0)\right]^{n/2}}$$

The likelyhood ratio is.

$$\frac{\hat{L}(\Omega)}{\hat{L}(\omega)} = \frac{\left(\frac{n^{n/2} \exp(-n/2)}{(2\pi)^{n/2}\left[(y - x\tilde{\beta})'(y - x\tilde{\beta})\right]^{n/2}}\right)}{\left(\frac{n^{n/2} \exp(-n/2)}{(2\pi)^{n/2}\left[(y - x\tilde{\beta}_0)'(y - x\tilde{\beta}_0)\right]^{n/2}}\right)}$$

$$= \left[\frac{(y - x\beta_0)'(y - x\beta_0)}{(y - x\hat{\beta})'(y - x\hat{\beta})}\right]^{n/2}$$

$$= \left(\frac{\hat{\sigma}_\omega^2}{\hat{\sigma}^2}\right)^{n/2} = (\lambda)^{n/2}$$

where $$\lambda = \frac{(y - x\beta_0)'(y - x\beta_0)}{(y - x\hat{\beta})(y - x\hat{\beta})}$$

is the ratio of the quadratic forms.

Now we simplify the numerator in λ as follows.

$$(y - x\beta_0)'(y - x\beta_0) = \left[(y - x\hat{\beta}) + (\hat{\beta} - \beta_0)x\right]'$$

$$\left[(y - x\hat{\beta}) + x(\hat{\beta} - \beta_0)\right]$$

=

$$(y - X\hat{\beta})'(y - X\hat{\beta}) + 2y'\left[I - X(X'X)^{-1}X'\right]X(\hat{\beta} - \beta_0) + (\hat{\beta} - \beta_0)'X'X(\tilde{\beta} - \beta_0)$$

$$= (y - X\hat{\beta})'(y - X\hat{\beta}) + (\hat{\beta} - \beta_0)'X'X(\hat{\beta} - \beta_0)$$

Thus,

$$\lambda = \frac{(y - X\hat{\beta})'(y - X\tilde{\beta}) + (\hat{\beta} - \beta_0)'X'X(\hat{\beta} - \beta_0)}{(y - X\hat{\beta})'(y - X\hat{\beta})}$$

$$\lambda = 1 + \frac{(\hat{\beta} - \beta_0)X'X(\hat{\beta} - \beta_0)}{(y - X\hat{\beta})'(y - X\hat{\beta})}$$

$$\lambda - 1 = \lambda_0 = \frac{(\hat{\beta} - \beta_0)X'X(\hat{\beta} - \beta_0)}{(y - X\hat{\beta})'(y - X\hat{\beta})}$$

where $0 \leq \lambda_0 < \infty$

Confidence interval Estimation:-

The confidence intervals in a multiple regression model can be constructed for individual regression coefficients as well as jointly.

Confidence interval on the individual regression coefficient:-

Assuming, ϵ_i's are identically and independently distributed following $N(0, \sigma^2)$ in $y = X\beta + \epsilon$, we have,

$$y \sim N(X\beta, \sigma^2 I)$$

$$b \sim N(\beta, \sigma^2 (X'X)^{-1})$$

Thus the marginal distribution of any regression coefficient estimate

$$b_j \sim N(\beta_j, \sigma^2 c_{jj})$$

where c_{jj} is the j^{th} diagonal element of $(X'X)^{-1}$

Thus

$$t_j = \frac{b_j - \beta_j}{\sqrt{\hat{\sigma}^2 c_{jj}}} \sim t_{(n-k)} \text{ under } H_0, \; j = 1, 2 \ldots$$

where $\hat{\sigma}^2 = \frac{SS_{res}}{n-k} = \frac{y'y - b'X'y}{n-k}$

So the $100(1-\alpha)\%$ confidence interval for $\beta_j \; (j = 1, 2, \ldots k)$ is obtained as follow.

$$P\left[-t_{\alpha/2, n-k} \leq \frac{b_j - \beta_j}{\sqrt{\hat{\sigma}^2 c_{jj}}} \leq t_{\alpha/2, n-k}\right] = 1 - \alpha$$

$$P\left[b_j - t_{\alpha/2,n-k}\sqrt{\hat{\sigma}^2 c_{jj}} \le \beta_j \le b_j + t_{\alpha/2,n-k}\sqrt{\hat{\sigma}^2 c_{jj}}\right] = 1-\alpha$$

Thus the confidence interval is

$$\left(b_j - t_{\alpha/2,n-k}\sqrt{\hat{\sigma}^2 c_{jj}},\ b_j + t_{\alpha/2,n-k}\sqrt{\hat{\sigma}^2 c_{jj}}\right)$$

Simultaneous confidence intervals on Regression coefficients :—

A set of confidence intervals that are true simultaneously with probability $(1-\alpha)$ are called simultaneous or joint confidence intervals.

It is relatively easy to define a joint confidence region for β in multiple regression model.

Since,

$$\frac{(b-\beta)' X'X(b-\beta)}{k\, MS_{res}} \sim F_{k,n-k}$$

$$\Rightarrow P\left[\frac{(b-\beta)' X'X (b-\beta)}{k\, MS_{res}} \le F_\alpha(k, n-k)\right] = 1-\alpha$$

So, a $100(1-\alpha)\%$ joint confidence region for all parameters in β is.

$$\frac{(b-\beta)' X'X (b-\beta)}{k\, MS_{res}} \le F_\alpha(k, n-k)$$

which describes an elliptically shaped region.