Model Adequacy checking

Wednesday, 21 August 2019 11:06 AM

$E(Y) = X\beta$ β is model parameters

$Y \sim N(X\beta, I_n\sigma^2)$ σ^2 unknown

1) Errors are uncorrelated
2) Errors are independent.
3) Errors are normally distributed

Coefficient of determination (R^2)

$$R^2 = \frac{\text{Variation of } Y \text{ explained by model}}{\text{Total Variation}}$$

$$= \frac{SS\,Model}{SS\,T}$$

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

⊛ Larger value of R^2 is considered better for the model

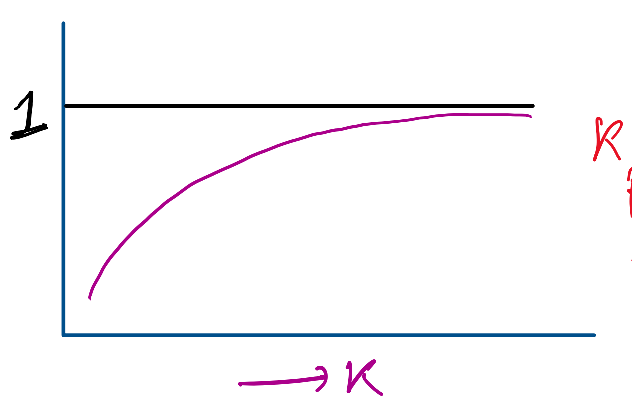
⊛ $R^2 \in [0,1]$

$SST = SS\,Model + SS\,Error$

$$1 = \frac{SS\,Model}{SST} + \frac{SS\,Error}{SST} \Rightarrow R^2 = 1 - \frac{SSE}{SST}$$

As $R^2 = 1 - (SSE/SST)$.

If we increase the number of regression variables such that (X^TX) remains invertible then R^2 will increase with the number of regressor variable



Adjusted R^2

$$R^2_{adj} = 1 - \frac{SSE/df(SSE)}{SST/df(SST)}$$

$$= 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

(1)

When R^2_{adj} attains maxima then we can stop and build the model.

ERROR ANALYSIS

$$\underset{\sim}{e} = \underset{\sim}{y} - \hat{\underset{\sim}{y}} = (I - P_x)\underset{\sim}{Y} \sim N(0, \sigma^2(I-P_x))$$

$$\frac{\underset{\sim}{e}^T\underset{\sim}{e}}{\sigma^2} \sim \chi^2_{n-k-1}$$

$$e_i = y_i - \hat{y}_i \sim N(0, \sigma^2(1-h_{ii}))$$

Assuming $H = P_x$

$$cov(e_i, e_j) = cov(y_i - \hat{y}_i, y_j - \hat{y}_j)$$
$$= \sigma^2(-h_{ij}) = \sigma^2(I-H)_{ij}$$

⊛ estimated errors are correlated in general

Define

$$r_i = \frac{e_i}{\sqrt{\hat\sigma^2(1-h_{ii})}}$$

We can estimate $\hat{\sigma}^2 = \frac{SSE}{n-k-1}$

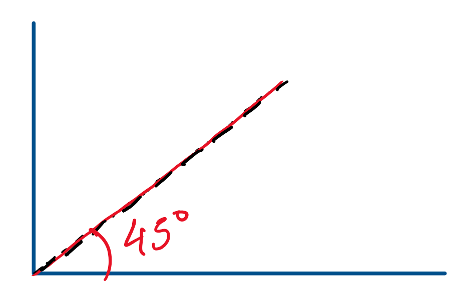
r_i will follow $N(0,1)$ when $n \uparrow \infty$

because $\hat{\sigma}^2 \longrightarrow \sigma^2$ with Prob 1 (Chebyshev)

If we plot the histogram of empirical density then we are supposed to get the pdf of $N(0,1)$.

We also can do q-q plot for r_i's. If the errors are normally distributed then only qq plot will give a straight line passing through origin with slope 1.

(2)

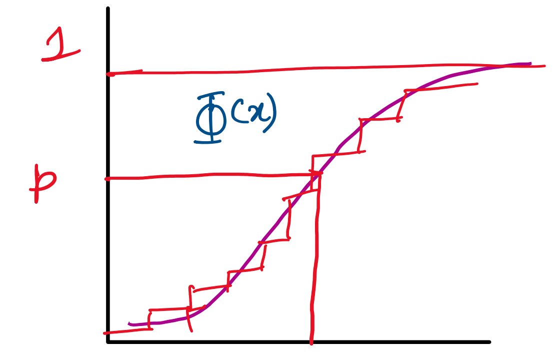


$q = x$ such that

$P(Z \le x) = p$

$$\frac{\#(r_i \le x)}{n} = p$$

$$\frac{\sum 1\{r_i \le x\}}{n} = p$$



p-p plot

$e_i = y_i - \hat{y}_i$ prediction error of y_i but the same y_i has been used to get $\hat{\beta}$ and $\hat{y}_i$.

⊛ If we remove (y_i, x_i) from the dataset and predict y_i based on other $(n-1)$ data then what will be the change corresponding

$$\frac{e_i}{\sqrt{V(e_i)}}$$

where $e_{(i)} = y_i - \hat{y}_i$

$\hat{y}_{(i)}$ = predicted value of y_i based on rest $(n-1)$ observations.

$\sum_{i=1}^{n} e_{(i)}^2$ = Predicted (Residual) error sum of squares

$\sum_{i=1}^{n}(y_i - \hat{y}_{(i)})^2 = PRESS$.

Leave out / Jack-Knife

$$e_{(i)} = \frac{e_i}{1-h_{ii}} \sim N\left(0, \frac{\sigma^2}{1-h_{ii}}\right)$$ This can be shown

Standardize $e_{(i)}$ as

$$\frac{e_{(i)}}{\sqrt{\sigma^2/(1-h_{ii})}} = \frac{e_i/(1-h_{ii})}{\sqrt{\sigma^2/(1-h_{ii})}}$$

$$= \frac{e_i}{\sqrt{\hat\sigma^2(1-h_{ii})}}$$

↳ (n-1) observations

Here we need to estimate σ^2 based on $(n-1)$ observations

(3)