Oct 23

5. Prediction

1) Regression model $y = \beta^T x + e$, $y, e \in \mathbb{R}$, $\beta, x \in \mathbb{R}^{p+1}$.

New obs. $x_0^T = (1, x_{01}, \cdots, x_{0p})$. $\Rightarrow$ r.v. $y_0 = \beta^T x_0 + e_0$

Point est. $\hat{y}_0 = \widehat{E(y_0)} = \hat{\beta}^T x_0$

$1-\alpha$ C.I. for $E(y_0)$ or $\mu_{y_0|x_0}$: $\hat{\beta}^T x_0 \mp t_{\frac{\alpha}{2}}(n-p')\, \hat{\sigma} \sqrt{x_0^T (X^T X)^{-1} x_0}$

$1-\alpha$ prediction interval: $\hat{\beta}^T x_0 \mp t_{\frac{\alpha}{2}}(n-p')\, \hat{\sigma} \sqrt{1 + x_0^T (X^T X)^{-1} x_0}$

* See P33 for the case of $p=1$.

2) Coef. of determination

Def: $R^2 = \frac{\text{Reg.SS}}{\text{TSS}} = 1 - \frac{\text{Res.SS}}{\text{TSS}}$

$R^2 \in [0, 1]$

For $p=1$, $R^2 = \left(\frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \right)^2$.

6. Lack of fit.

Repeated observations with the same set of independent variables.

H_0: No lack of fit.

m combinations of x, n_i: # observations of the ith combination.

$$\underbrace{\sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2}_{\substack{\text{Res.SS} \\ \sim \sigma^2 \chi^2(n-p', \lambda)}} = \underbrace{\sum_{i=1}^{m} \sum_{j=1}^{n_j} (y_{ij} - \bar{y}_{i\cdot})^2}_{\substack{\text{Pure error S.S.} \\ \sim \sigma^2 \chi^2(n-m)}} + \underbrace{\sum_{i=1}^{m} n_i (\bar{y}_{i\cdot} - \hat{y}_i)^2}_{\substack{\text{Lack of fit SS} \\ \sigma^2 \chi^2(m-p', \lambda)}}$$

Estimates of σ^2:

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n-p'} \quad \text{or} \quad \hat{\sigma}^2_{\text{pure error}} = \frac{\text{Pure error SS}}{n-m}$$

Test stat:

$$F = \frac{\text{Lack of fit SS}/(m-p')}{\text{Pure error SS}/(n-m)} \overset{H_0}{\sim} F(m-p', n-m)$$

Reject H_0 if $F > F_\alpha(m-p', n-m)$.

7. Added variable Plot.

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_i.$$

① $y \sim x_{(-k)}$. $\Rightarrow$ Res. $= \hat{e}_{Y(X_k)}$

② $x_k \sim x_{(-k)}$ $\Rightarrow$ Res. $= \hat{e}_k$

③ $\hat{e}_{Y(X_k)} \sim \hat{e}_k$ $\Rightarrow$ slope $= \hat{\beta}_k$. Explain $\hat{\beta}_k$.

④ $PCC(y, x_k \mid x_{(-k)}) = Corr(\hat{e}_{Y(X_k)}, \hat{e}_k)$. PCC.

PCC as conditional independence test:

Generally, when $(X, Y, Z) \sim MN$, we have

$$PCC(X, Y \mid Z) = 0 \Leftrightarrow X \perp\!\!\!\perp Y \mid Z.$$