Fitting linear models by Least squares
Part 3

Given: TS $(\underline{x}_1, y_1) \dots (\underline{x}_n\ y_n)$

Goal: Generate prediction rule $f(\underline{x})$

Simple approach:

- Try linear rule $\ell(\underline{x}) = b_0 + b_1 x_1 .. + b_p x_p$
- Find coefficient vector $\hat{\underline{b}}$ that minimizes resubstitution error:

  $\longrightarrow$ $\hat{\underline{b}} = \underset{\underline{b}}{\operatorname{argmin}} \|\underline{y} - X\underline{b}\|^2$ where

  design matrix $X = \begin{array}{cc} \begin{pmatrix} 1 & \underline{x}_1 \\ 1 & \underline{x}_2 \\ & \vdots \\ 1 & \underline{x}_n \end{pmatrix} & n \\ p & \end{array}$

We always denote # of columns by p and assume $X^1 = (1 \overset{n}{\dots} 1)^T$

Facts

- $\hat{\underline{y}}$ = projetion of $\underline{y}$ on $[X^1 .. X^p]$
- If $\operatorname{rank}(X) = p$ then

  $\hat{\underline{b}} = (X^T X)^{-1} X^T \underline{y}$

  $\hat{\underline{y}} = X\hat{\underline{b}} = X(X^T X)^{-1} X^T \underline{y}$ $\leftarrow$ not operational if $\operatorname{rank}(X) < p$

- If $\operatorname{rank}(X) < p$ then

  $\hat{\underline{b}}$ not unique; there are ∞ many ways to write $\hat{\underline{y}}$ as a lincom of $X^1 \dots X^p$

There is a way to compute $\hat{y}$ even for rank deficient case. May discuss later

For the moment let's assume rank$(X) = p$

Then $\hat{y} = \underbrace{X(X^T X)^{-1} X^T}_{n \times n} y$

H is called "hat matrix"

H is a projection matrix

$H = H^T$ symmetrix

$H^2 = H$ idempotent

Assume $y_i = f(\underline{x}_i) + \varepsilon_i$

ε_i iid $E(\varepsilon_i) = 0$

$V(\varepsilon_i) = \sigma^2$

Then the expected squared estimation error (expectation is taken over the ε's; $\underline{x}_1 \ldots \underline{x}_n$ are held fixed.

$$ESE = \frac{1}{n} E\left(\|\underline{f} - \hat{y}\|^2\right.$$

$$= \frac{1}{n} E\left(\|\underline{f} - Hy\|^2\right)$$

$$= \frac{1}{n} E\left(\|\underline{f} - H\underline{f} - H\underline{\varepsilon})\|^2\right)$$

$$= \frac{1}{n}\left[E\left(\|\underline{f} - H\underline{f}\|^2\right) + E\left(\|H\underline{\varepsilon}\|^2\right)\right]$$

$+$ cross-term $= 0$

$$= \frac{1}{n}\left\{\|(I-H)f\|^2 + E\left(\varepsilon^T H^T H \varepsilon\right)\right\}$$

$$= \frac{1}{n}\left[\|(I-H)f\|^2 + E\sum_{i,j} H_{ij}\,\varepsilon_i \varepsilon_j\right]$$

$$= \frac{1}{n}\Big[\underbrace{\|(I-H)f\|^2}_{\text{squared bias}} + \underbrace{\sigma^2 \sum H_{ii}}_{\substack{\| \\ \text{trace } H = p}} \Big] \longrightarrow \text{variance}$$

Suppose $f \in [X^1 .. X^p] \Rightarrow$ squared bias $= 0$

Estimate is unbiased

Suppose X^p unnecessary because $f \in [X^1 ... X^{p-1}]$. Then performance is better because variance term will decrease and the bias^2 won't increase.