Regression Advanced Topics

Recap: Probabilistic Model: $y = x^T\beta + \varepsilon \quad \varepsilon \sim N(0, \sigma^2)$

$$\Downarrow$$

$$y = X\beta + \varepsilon \quad \varepsilon \sim N(0, \sigma^2 I_n)$$

Decomposition: $y = \hat{y} + \hat{e}$

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

$$\begin{cases} \hat{y} = X(X^TX)^{-1}X^Ty = H_X y \\ \hat{e} = (I - H_X)y = M_X y \end{cases}$$

since $H_X \cdot X = X \Rightarrow \begin{cases} H_X \cdot \mathbb{1} = \mathbb{1} \\ H_X \cdot X_i = X_i \end{cases} \Rightarrow X^T\hat{e} = 0$

---

1. FWL Theorem

consider the following 2 LR:

① $y = X_1\beta_1 + X_2\beta_2 + \varepsilon \Rightarrow (\hat{\beta}_1, \hat{\beta}_2, \hat{\varepsilon})$

② $M_{X_1} y = M_{X_1} X_2 \beta_2 + \varepsilon \Rightarrow (\tilde{\beta}_2, \tilde{\varepsilon})$

we will have $\begin{cases} \hat{\beta}_2 = \tilde{\beta}_2 \\ \hat{\varepsilon} = \tilde{\varepsilon} \end{cases}$

Pf: Mainly through [Schur Matrix Inverse Decomposition]

分块 inverse equality (矩阵打洞)

assume $\det(D) \neq 0$

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

$$\begin{pmatrix} I & -BD^{-1} \\ & I \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} A - BD^{-1}C & 0 \\ C & D \end{pmatrix}$$

$$\begin{pmatrix} I & \\ -C(A-BD^{-1}C)^{-1} & I \end{pmatrix}\begin{pmatrix} A-BD^{-1}C & \\ C & D \end{pmatrix} = \begin{pmatrix} A-BD^{-1}C & \\ & D \end{pmatrix}$$

$$\begin{pmatrix} (A-BD^{-1}C)^{-1} & \\ & D^{-1} \end{pmatrix}\begin{pmatrix} A-BD^{-1}C & \\ & D \end{pmatrix} = \begin{pmatrix} I & \\ & I \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A-BD^{-1}C)^{-1} & \\ & D^{-1} \end{pmatrix}\begin{pmatrix} I & \\ -C(A-BD^{-1}C)^{-1} & I \end{pmatrix}\begin{pmatrix} I & -BD^{-1} \\ & I \end{pmatrix}$$

$$= \begin{pmatrix} (A-BD^{-1}C)^{-1} & \\ & D^{-1} \end{pmatrix}\begin{pmatrix} I & -BD^{-1} \\ -C(A-BD^{-1}C)^{-1} & I + C(A-BD^{-1}C)^{-1}BD^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} (A-BD^{-1}C)^{-1} & -(A-BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A-BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A-BD^{-1}C)^{-1}BD^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1}+D^{-1}CMBD^{-1} \end{pmatrix}$$

$$\boxed{M := (A-BD^{-1}C)^{-1}}$$

$$(X^TX)^{-1}X^Ty$$

Application:
$$X^TX = \begin{pmatrix} X_1^TX_1 & X_1^TX_2 \\ X_2^TX_1 & X_2^TX_2 \end{pmatrix} := \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} X_1^TX_1 & X_1^TX_2 \\ X_2^TX_1 & X_2^TX_2 \end{pmatrix}^{-1}\begin{pmatrix} X_1^T \\ X_2^T \end{pmatrix} y$$

1

$$\Rightarrow \hat{\beta}_1 = MX_1^Ty - MBD^{-1}X_2^Ty$$

$$= M\,(X_1^T - BD^{-1}X_2^T)\,y$$

$$= M\,(X_1^T - X_1^T X_2 (X_2^T X_2)^{-1} X_2^T)\,y$$

$$= M \cdot \underline{X_1^T M_{X_2}} \, y$$

$$= (X_1^T X_1 - X_1^T X_2 (X_2^T X_2)^{-1} X_2^T X_1)^{-1} X_1^T M_{X_2} y$$

$$= (X_1^T M_{X_2} X_1)^{-1} (X_1^T M_{X_2} y)$$

$$= (X_1^T M_{X_2} M_{X_2} X_1)^{-1} (X_1^T M_{X_2} M_{X_2} y)$$

$$= \tilde{\beta}_1$$

can give $\underline{\hat{\beta}_1 \sim N(0, \sigma^2 (X_1^T M_{X_2} X_1)^{-1})}$

② $\hat{\varepsilon} = y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2$

$$\Rightarrow \underline{M_{X_2}\hat{\varepsilon} = \hat{\varepsilon}} = M_{X_2} y - M_{X_2} X_1 \hat{\beta}_1$$

$\boxed{X_2^T \hat{\varepsilon} = 0}$

$$= M_{X_2} y - M_{X_2} X_1 \tilde{\beta}_1$$

$$= \tilde{\varepsilon}$$

#

Conclusion:

① regression totally $\rightarrow (\hat{\beta}_1, \hat{\beta}_2)$

= a) $y \sim 1 + x_1 + \varepsilon \Rightarrow \tilde{y}$

b) $x_2 \sim 1 + x_1 + \varepsilon \Rightarrow \tilde{x}_2$

c) $\tilde{y} \sim 1 + \tilde{x}_2 \Rightarrow \underline{\tilde{\beta}_2 = \hat{\beta}_2}$

2-stage regression

② $y - X\hat{\beta} = \hat{e} = y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2$

$$= \tilde{e}$$

$$= M_{X_2}(y - X_1\hat{\beta}_1) = M_{X_1}(y - X_2\hat{\beta}_2)$$

2. Analysis on some Measure

① R^2

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

→ Further $$R^2 = \mathrm{corr}(y, \hat{y})$$

$$= \frac{[(I - \frac{1}{n}J)y]^T [(I - \frac{1}{n}J)\hat{y}]}{[y^T(I - \frac{1}{n}J)y]^{\frac{1}{2}} [\hat{y}^T(I - \frac{1}{n}J)\hat{y}]^{\frac{1}{2}}}$$

$$= \frac{y^T(I - \frac{1}{n}J)(H_X - \frac{1}{n}J)y}{(SST)^{\frac{1}{2}} \cdot (SSR)^{\frac{1}{2}}}$$

$$= \frac{y^T(H_X - \frac{1}{n}J)y}{(SST)^{\frac{1}{2}} (SSR)^{\frac{1}{2}}}$$

$$= \frac{(SSR)^{\frac{1}{2}}}{(SST)^{\frac{1}{2}}}$$

$Y \in \mathbb{R}^n$ observation $\qquad X \in \mathbb{R}^{n \times k}$ observation

② consider $Y \in \mathbb{R}$ v.s. $X \in \mathbb{R}^k$ (correlation)

$$\mathrm{mcorr}(Y, X)^2 := \max_a \mathrm{corr}^2(Y, Xa)$$

$$= \max_a \frac{\mathrm{cov}(Y, Xa)^2}{\mathrm{var}[Y]\,\mathrm{var}[Xa]}$$

$$= \max_a \frac{(y^T(I - \frac{1}{n}J) \cdot Xa)^2}{(y^T(I - \frac{1}{n}J)y)\ (a^T X^T (I - \frac{1}{n}J) X a)}$$

$$= \frac{1}{y^T(I-\frac{1}{n}J)y} \max_a \frac{(y^T X_d a)^2}{a^T X_d^T X_d a}$$

$$= \frac{1}{y^T(I-\frac{1}{n}J)y} \max_b \frac{[y^T X_d (A^{\frac{1}{2}})^{-1} b]^2}{b^T b}$$

$= \underline{\text{cov}(y)}$

$\underline{A := X_d^T X_d}$

$\approx$ "covariance matrix"

a) $\leq \| (A^{\frac{1}{2}})^{-1} X_d^T y \|_2^2$

$= y^T X_d (A^{\frac{1}{2}})^{-1} (A^{\frac{1}{2}})^{-1} X_d^T y$

when $b = c \cdot (A^{\frac{1}{2}})^{-1} X_d^T y$

b) since $b = A^{\frac{1}{2}} a$

$\Rightarrow a = c \cdot A^{-1} X_d^T y$

$= c \cdot (X_d^T X_d)^{-1} X_d^T y := a_M$

To conclude, $\text{mcorr}^2(y, X) = \frac{\text{cov}[y, X]\,\text{cov}[X, X]^{-1}\,\text{cov}[X, y]}{\text{var}[y]}$

when $\boxed{a_M = c \cdot (X_d^T X_d)^{-1} X_d^T y}$ $c \in \mathbb{R}$.

Also, we claim $a_M \propto (\hat{\beta}_1, \ldots\ldots \hat{\beta}_{k-1})$ $\rightarrow$ apply FWL

Pf sketch: $a_M \propto (X_d^T X_d)^{-1} X_d^T y = (X_d^T X_d)^{-1} X_d^T y_d$

$\Rightarrow \boxed{y_d = X_d \cdot a_M + \varepsilon}$

where $\begin{cases} y_d = M_1 y \\ X_d = M_1 y \end{cases}$ $M_1 = I - (\mathbb{1}^T \mathbb{1})^{-1} \mathbb{1}^T y$

$\boxed{y = \mathbb{1}^T \beta_0 + X\beta + \varepsilon}$

FWL implies that, ① $\underline{a_M \propto (\hat\beta_1, \dots, \hat\beta_{k-1})}$

② 2 residuals are the same !

$$\Rightarrow m\,corr^2(y, X) = corr^2(y, X a_M)$$

$$= corr^2(y, \mathbb{1}^T \hat\beta_0 + X a_M)$$

$$= corr^2(y, \hat y)$$

$$= R^2$$

#

③ consider model $\underline{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon}$

our interest is: $\boxed{\text{how } y \text{ correlated with } x_1}$

<u>idea</u>: <u>we should kick out the influence of $\boxed{1, x_2}$.</u>

$$\begin{cases} \tilde y = M_2 y \\ \tilde x_1 = M_2 x_1 \end{cases} \longrightarrow r^2_{y1:2} := corr^2(\tilde y, \tilde x_1)$$

$$= R^2_{\tilde y(\tilde x_1)}$$

补常数

$$= R^2_{\tilde y : 1 + \tilde x_1}$$

Note that $\underline{X_2^T M_2 = 0}$

$\Rightarrow X_2^T \tilde x_1 = 0$

$\Rightarrow \mathbb{1}^T \tilde x_1 = 0 \Rightarrow \underline{\mathbb{1}^T H_{\tilde x_1} = 0}$

$$= \frac{(H_{\tilde x_1} \tilde y)^T (I - \frac{1}{n} J)(H_{\tilde x_1} \tilde y)}{\tilde y^T (I - \frac{1}{n} J)\, \tilde y}$$

$$H_{\tilde x_1} = \tilde x_1 (\tilde x_1^T \tilde x_1)^{-1} \tilde x_1^T$$

$\mathbb{1}^T H_{\tilde{x}} = \mathbb{1}^T$ since $\tilde{X}_1^T H_{\tilde{x}_1} = X_1^T$

$\Rightarrow \mathbb{1}^T H_{\tilde{x}_1} \cdot \tilde{y} = \mathbb{1}^T \tilde{y} = 0$

$$\boxed{H_{\tilde{X}_1} = \frac{1}{n} J + H_{\tilde{x}_1}}$$

$\tilde{X}_1 = [\mathbb{1} ; \tilde{x}_1]$

Consider $\tilde{y}^T (I - H_{\tilde{x}_1}) \tilde{y}$

$$= \left[(I - H_{\tilde{x}_1}) \tilde{y} \right]^T \left[(I - H_{\tilde{x}_1}) \tilde{y} \right]$$

FWL thm

$$\boxed{\begin{aligned} &= \text{residual}(\tilde{y} : \tilde{x}_1) \\ &\quad \begin{cases} x_1 : 1 + x_2 + \tilde{x}_1 \\ y : 1 + x_2 + \tilde{y} \end{cases} \\ &= \text{residual}(y : 1 + x_1 + x_2) \end{aligned}}$$

$$= SSE(x_1, x_2)$$

$$= \frac{(H_{\tilde{x}_1} \tilde{y})^T (H_{\tilde{x}_1} \tilde{y})}{\tilde{y}^T \tilde{y}}$$

$$= \frac{\tilde{y}^T H_{\tilde{X}_1} \tilde{y}}{SSE(x_2)}$$

$$= \frac{\tilde{y}^T (\frac{1}{n} J + H_{\tilde{x}_1}) \tilde{y}}{SSE(x_2)}$$

$$= \frac{\tilde{y}^T H_{\tilde{x}_1} \tilde{y}}{SSE(x_2)}$$

$$= \frac{\tilde{y}^T (I - (I - H_{\tilde{x}_1})) \tilde{y}}{SSE(x_2)}$$

$$= \frac{SSE(x_2) - \tilde{y}^T (I - H_{\tilde{x}_1}) \tilde{y}}{SSE(x_2)}$$

$$= \frac{SSE(x_2) - SSE(x_1, x_2)}{SSE(x_2)}$$

Note that $r^2_{y1:2} = \dfrac{SSE(x_2) - SSE(x_1, x_2)}{SSE(x_2)} \sim F(1, n-2-1+1)$

笔记整理 (9) → 子弱猹

Some Statistics

① $\hat{e} = y - \hat{y} = \underline{\underline{(I - H_x)\, y}} \Rightarrow \hat{e} \sim N(0, \sigma^2 (I - H))$

② $\mathbb{E}[\hat{e}^T \hat{e}] = \mathbb{E}[y^T (I - H_x) y] = \sigma^2 (n - p - 1)$

$= \mathbb{E}[\varepsilon^T (I - H_x)\, \varepsilon]$

$\Rightarrow$ estimator of σ^2 : $\dfrac{\hat{e}^T \hat{e}}{n-p-1} = \dfrac{SSE}{n-p-1}$

③ $\mathrm{cov}(\hat{\beta}, \hat{e}) = \mathrm{cov}((X^T X)^{-1} X^T y,\ (I - H) y) = 0$

$\Rightarrow \mathrm{cov}(\hat{\beta}, y) = \mathrm{cov}(\hat{\beta}, \hat{y})$

④ $\dfrac{\frac{SSR}{p}}{\frac{SSE}{n-p-1}} \sim F(p, n-p-1)$ $\quad \begin{cases} \dfrac{SSR}{\sigma^2} \sim \chi^2(p) \\ \dfrac{SSE}{\sigma^2} \sim \chi^2(n-p-1) \end{cases}$

⑤ $\hat{\beta} = (X^T X)^{-1} X^T y \sim N(\beta, \sigma^2 (X^T X)^{-1})$

$\Rightarrow \hat{\beta}_j \sim N(\beta_j, \sigma^2 h_{jj})$

$\Rightarrow \dfrac{\hat{\beta}_j - \beta_j}{\sqrt{h_{jj}}\, \hat{\sigma}} \sim t(n-p-1)$

⑥ $\hat{Y}_n \sim N(x_n^T \beta,\ \sigma^2 (x_n^T (X^T X)^{-1} x_n))$ $\quad \underline{\hat{Y}_n = x_n^T \hat{\beta}}$

$Y_n \sim N(x_n^T \beta,\ \sigma^2)$ $\quad \underline{Y_n = x_n^T \beta + \varepsilon_n}$

$$\Rightarrow \begin{cases} \dfrac{\hat{Y}_n - Y_n}{\sqrt{1 + x_n^T (X^T X)^{-1} x_n} \cdot \hat{\sigma}} \sim t(n-p-1) \\[2ex] \dfrac{\hat{Y}_n - E[Y_n]}{\sqrt{x_n^T (X^T X)^{-1} x_n} \quad \hat{\sigma}} \sim t(n-p-1) \end{cases}$$