回归习题

中国人民大学统计学院 黄实磊

1、证明帽子矩阵的秩

帽子矩阵$H=X(X^TX)^{-1}X^T$为对称幂等矩阵

则$rank(H)=tr(H)=tr(X(X^TX)^{-1}X^T)=tr((X^TX)^{-1}X^TX)=p+1$

2、证明SSR和SSE独立

$Y-\bar{Y}=(I-H_1)Y$，$H_1=LL^T/n$

$Y-\hat{Y}=(I-H)Y$

$SSR=(\hat{Y}-\bar{Y})^T(\hat{Y}-\bar{Y})=[(H-H_1)Y]^T(H-H_1)Y$

$SSE=[(I-H)Y]^T(I-H)Y$

由$(I-H)X=0$和$X^T(I-H)=0$可知$(I-H)H_1=0$，$H_1(I-H)=0$

(利用这个性质可以证明$1/n\le(H)_{ii}\le 1$)

得到$cov((H-H_1)Y,(I-H)Y)=(H-H_1)cov(Y,Y)(I-H)=0$

可知正态假设下SSR和SSE独立

3、证明遗漏变量有偏

真实模型为$Y=X\beta+\varepsilon=(X_1,X_2)(\beta_1^T,\beta_2^T)^T+\varepsilon$

如果只用X_1估计模型，β_1最小二乘估计为$(X_1^TX_1)^{-1}X_1^TY$

期望为$E((X_1^TX_1)^{-1}X_1^TY)=E[(X_1^TX_1)^{-1}X_1^T(X_1\beta_1+X_2\beta_2+\varepsilon)]=\beta_1+\beta_2(X_1^TX_1)^{-1}X_1^TX2$

4、中心化回归

未中心化模型为

$$Y=X\beta+\varepsilon,X=(X_0,X_1),\beta=(\beta_0,\beta_1^T)^T,X_0=(1,1,1,\ldots,1,1,1)^T$$

记$H_1=L*L^T/n,L=(1,1,1,\ldots,1,1,1)^T$ 中心化模型为

$$\begin{aligned}(I-H_1)Y&=(I-H_1)(X_0,X_1)(\beta_0,\beta_1^T)^T+(I-H_1)\varepsilon\\&=(I-H_1)X_1\beta_1+\varepsilon-H_1(Y-X\beta)\\&=(\beta_0+\bar{X}_1\beta_1-\bar{Y})+(I-H_1)X_1\beta_1+\varepsilon\end{aligned}$$

记$\beta_0^{new}=\beta_0+\bar{X}_1\beta_1-\bar{Y},\tilde{Y}=(I-H_1)Y,\tilde{X}_1=(I-H_1)X_1,X_{new}=(X_0,\tilde{X}_1)$

中心化模型为

$$\tilde{Y}=\beta_0^{new}+\tilde{X}_1\beta_1+\varepsilon$$

也就是说，即使中心化，也无法忽略β_0^{new}项。回归模型的统一框架是带截距项的回归，不带截距项的回归只是$\beta_0=0$的特例，当$\beta_0=0$时，忽略β_0仍可以带来β_1的无偏估计，但中心化并不能从理论上带来$\beta_0^{new}=0$。我们将会发现, 中心化只能带来$\hat{\beta}_0^{new}=0$，中心化处理后自由度仍然是$n-p-1$, 而不是$n-p$。

中心化模型的最小二乘估计为

$$\begin{pmatrix} \hat{\beta}_0^{new} \\ \hat{\beta}_{1c} \end{pmatrix} = (X_{new}^T X_{new})^{-1} X_{new}^T \tilde{Y}$$

其中

$$X_{new}^T X_{new} = (X_0, \tilde{X}_1)^T (X_0, \tilde{X}_1) = \begin{pmatrix} X_0^T X_0 & 0 \\ 0 & \tilde{X}_1^T \tilde{X}_1 \end{pmatrix}$$

易知$\hat{\beta}_0^{new} = 0$

未中心化模型的参数估计为

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (X^T X)^{-1} X^T Y$$

利用分块矩阵的逆可以证得 $\hat{\beta}_1 = \hat{\beta}_{1c}$

中心化处理后计算残差

$$\begin{aligned} \tilde{Y} - \hat{\tilde{Y}} &= \tilde{Y} - \hat{\beta}_0^{new} - \tilde{X}_1 \hat{\beta}_{1c} \\ &= \tilde{Y} - \hat{\beta}_1 \tilde{X}_1 \\ &= (I - H_1)Y - (I - H_1)X_1 \hat{\beta}_1 \end{aligned}$$

利用分块矩阵的逆可以证明$Y - \hat{Y}$会等于上式，也就是说残差的估计和未中心化模型是一样的。

即σ^2无偏估计为$\hat{\sigma}^2 = \frac{1}{n-p-1}(Y - \hat{Y})^T (Y - \hat{Y})$

中心化模型认识误区的补充说明：我们重新审视估计的$\hat{\beta}_0^{new}$和$\hat{\beta}_{1c}$。它估计的模型是$\tilde{Y} = \beta_0^{new} + \tilde{X}_1 \beta_1 + \varepsilon$。从**估计**的角度，最小二乘估计的时候$\beta_0^{new}$被视为一个常数。然而，我们同时也有$\beta_0^{new} = \beta_0 + \bar{X}_1 \beta_1 - \bar{Y}$表示它不是常数，这说明中心化模型背后的假定已经发生了变化：样本均值$\bar{Y}$被视作对Y做平移的**常数**，而不是随机变量。最小二乘估计算法下，这样的处理能够保证$\hat{\beta}_0^{new}$等于0，$\hat{\beta}_{1c}$的估计也会等于原模型的$\hat{\beta}_1$。从估计的角度，它节约了$\hat{\beta}_0^{new}$，但从模型的角度，β_0^{new}本身并不能被忽略，$\hat{\tilde{Y}}$可以写为$\tilde{X}_1 \hat{\beta}_1$，但$\tilde{Y}$并不能写为$\tilde{X}_1 \beta_1 + \varepsilon$。

5、证明偏F检验和t统计量等价

偏$F = \frac{SSE_j - SSE}{SSE/(n-p-1)}$，$t = \hat{\beta}_i^2 / c_{ii} \hat{\sigma}^2$

其中$c_{ii} = (X^T X)_{ii}^{-1}$

只需证明$SSE_j - SSE = \hat{\beta}_i^2 / c_{ii}$

不妨假设删除的是最后一个变量

记$X = (X_{p-1}, X_p)$，$H = X[X^T X]^{-1} X^T, H_{p-1} = X_{p-1}[X_{p-1}^T X_{p-1}]^{-1} X_{p-1}^T$

则$SSE_j - SSE = Y^T (H - H_{p-1}) Y$

由

$$X^TX = \begin{pmatrix} X_{p-1}^T X_{p-1} & X_{p-1}^T X_p \\ X_p^T X_{p-1} & X_p^T X_p \end{pmatrix}$$

利用分块矩阵的求逆公式以及$c_{pp} = (X^TX)_{pp}^{-1}$以及$\hat{\beta}_p = [(X^TX)^{-1}X^TY]_p$ 即得证。