回归分析 (01714601)
课程主页：http://staff.ustc.edu.cn/~ynyang/lm2020

第十四讲 R^2的分解

2020.4.3

$$R_{I\cup J}^2 = R_I^2 + (1 - R_I^2) R_{J\bullet I}^2$$

命题3. 对于模型

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{1}\beta_0 + Z\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim (0, \sigma^2)$$

记样本方差-协方差矩阵：$\hat{\Sigma}_{\mathbf{xx}} = Z_{(c)}{}^{\mathsf{T}} Z_{(c)}/(n-1)$，$\hat{\Sigma}_{\mathbf{xy}} = Z_{(c)}{}^{\mathsf{T}} \mathbf{y}/(n-1)$,

以及$\hat{\Sigma}_{yy\bullet\mathbf{x}} = \hat{\Sigma}_{yy} - \hat{\Sigma}_{y\mathbf{x}} \hat{\Sigma}_{\mathbf{xx}}^{-1} \hat{\Sigma}_{\mathbf{x}y}$, 则$R^2 = \dfrac{\hat{\Sigma}_{y\mathbf{x}} \hat{\Sigma}_{\mathbf{xx}}^{-1} \hat{\Sigma}_{\mathbf{x}y}}{\hat{\Sigma}_{yy}}$.

$X = (\mathbf{1}, X_I, X_J)$

命题4. 对于模型 $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{1}\beta_0 + X_I \boldsymbol{\beta}_I + X_J \boldsymbol{\beta}_J + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim (0, \sigma^2)$

其中$I \subset \{1, 2, \ldots, p-1\}, J = \{1, 2, \ldots, p-1\} \setminus I$. 则$R^2 = R^2_{I \cup J}$ 可分解为

$$R^2_{I \cup J} = \frac{\hat{\Sigma}_{y,I} \hat{\Sigma}_{I,I}^{-1} \hat{\Sigma}_{I,y}}{\hat{\Sigma}_{yy}} + \frac{\hat{\Sigma}_{yJ\bullet I} \hat{\Sigma}_{JJ\bullet I}^{-1} \hat{\Sigma}_{Jy\bullet I}}{\hat{\Sigma}_{yy}} \triangleq R_I^2 + (1 - R_I^2) R^2_{J\bullet I}$$

其中$R^2_{J\bullet I} = \dfrac{\hat{\Sigma}_{yJ\bullet I} \hat{\Sigma}_{JJ\bullet I}^{-1} \hat{\Sigma}_{Jy\bullet I}}{\hat{\Sigma}_{yy\bullet I}}$,称为复偏相关系数平方，其中$\hat{\Sigma}_{y,I}$ 为y与$\mathbf{x}_I$的样本协方差阵,$\hat{\Sigma}_{I,I}$为$\mathbf{x}_I$的样本协方差矩阵等等.

定义:

(1) $R^2_{J\bullet I}=\dfrac{\hat{\Sigma}_{yJ\bullet I}\hat{\Sigma}^{-1}_{JJ\bullet I}\hat{\Sigma}_{Jy\bullet I}}{\hat{\Sigma}_{yy\bullet I}}$称为复偏相关系数平方,代表了自变量 X_I 所不能解释的**y**的方差中能被 X_J 所解释的比率(即控制变量X_I后的残差对变量 X_J 回归的复相关系数平方).

(2) $(1-R^2_I)R^2_{J\bullet I}=R^2_{J\cup I}-R^2_I=\dfrac{\hat{\Sigma}_{yJ\bullet I}\hat{\Sigma}^{-1}_{JJ\bullet I}\hat{\Sigma}_{Jy\bullet I}}{\hat{\Sigma}_{yy}}$称为半偏相关系数平方(semipartial correlation),代表了当自变量从X_I增加到$X_{I\cup J}=Z$时，复相关系数平方的增量.

注：当$J=\{j\}$时，$R^2_{j\bullet I}=\dfrac{\hat{\Sigma}_{yj\bullet I}\hat{\Sigma}^{-1}_{jj\bullet I}\hat{\Sigma}_{jy\bullet I}}{\hat{\Sigma}_{yy\bullet I}}=\left(r_{yj\bullet I}\right)^2$, $r_{yj\bullet I}$为y与x_j的偏相关系数.

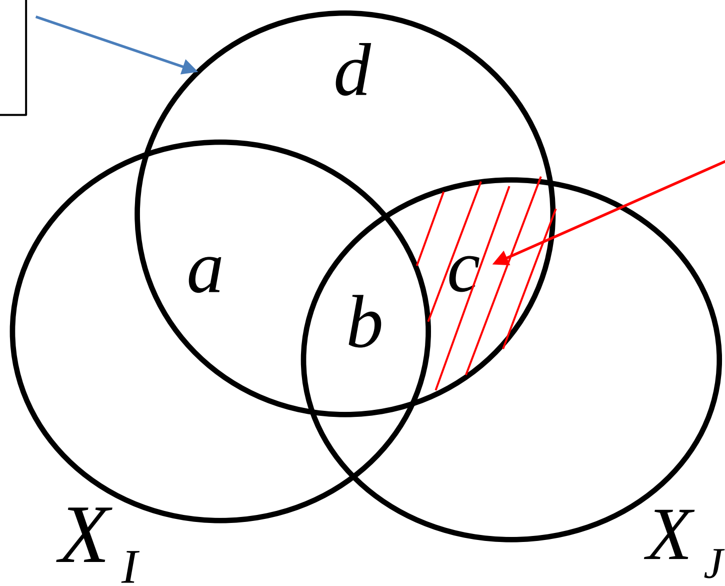
$a+b+c=R^2=R^2_{I\cup J}$： X_I, X_J所能解释的y的方差的比率；

$a+b=R^2_I$： X_I的贡献率(所能解释的y的方差的比率);

$c=R^2_{I\cup J}-R^2_I$： X_J单独的贡献率;

$c/(d+c)=c/(1-a-b)=R^2_{J\bullet I}$:排除X_I的贡献$(a+b)$之后, X_J的贡献率。

y的总方差:
$a+b+c+d=1$



$$R^2_{J\bullet I}=\frac{c}{d+c}$$

$$R^2_{I\cup J}=R^2_I+(1-R^2_I)R^2_{J\bullet I} \Leftrightarrow a+b+c=(a+b)+(1-a-b)\times\frac{c}{d+c}$$

证明：(1)不妨假设所有自变量Z已经中心化，

令$X_J^{\perp}=X_J-P_{X_I}X_J$，我们有正交分解：

$$\hat{\mathbf{y}}-\mathbf{1}\bar{y}=P_{X_I}\mathbf{y}+P_{X_J^{\perp}}\mathbf{y}=P_{X_I}\mathbf{y}+X_J^{\perp}\hat{\boldsymbol{\beta}}_J,$$

所以$SS_{回}=\|\hat{\mathbf{y}}-\mathbf{1}\bar{y}\|^2=\|P_{X_I}\mathbf{y}\|^2+\|P_{X_J^{\perp}}\mathbf{y}\|^2=\|P_{X_I}\mathbf{y}\|^2+\|X_J^{\perp}\hat{\boldsymbol{\beta}}_J\|^2,$

注意

$$\|P_{X_I}\mathbf{y}\|^2=\mathbf{y}^{\mathsf{T}}X_I(X_I^{\mathsf{T}}X_I)^{-1}X_I^{\mathsf{T}}\mathbf{y}=(n-1)\hat{\Sigma}_{y,I}\hat{\Sigma}_{I,I}^{-1}\hat{\Sigma}_{I,y},$$

由推论1，$\hat{\boldsymbol{\beta}}_J=\hat{\Sigma}_{JJ\bullet-J}^{-1}\hat{\Sigma}_{Jy\bullet-J}$，并注意到$X_J^{\perp\mathsf{T}}X_J^{\perp}=(n-1)\hat{\Sigma}_{JJ\bullet-J}$所以

$$\|X_J^{\perp}\hat{\boldsymbol{\beta}}_J\|^2=\hat{\boldsymbol{\beta}}_J^{\mathsf{T}}\left(X_J^{\perp\mathsf{T}}X_J^{\perp}\right)\hat{\boldsymbol{\beta}}_J=(n-1)\hat{\Sigma}_{yy\bullet-J}\hat{\Sigma}_{JJ\bullet-J}^{-1}\hat{\Sigma}_{Jy\bullet-J}$$

所以

$$R^2 = R^2_{I\cup J} = \frac{\| \hat{\mathbf{y}} - \mathbf{1}\bar{y} \|^2}{\| \mathbf{y} - \mathbf{1}\bar{y} \|^2} = \frac{\| P_{X_I}\mathbf{y} \|^2}{\| \mathbf{y} - \mathbf{1}\bar{y} \|^2} + \frac{\| X_J^{\perp}\hat{\boldsymbol{\beta}}_J \|^2}{\| \mathbf{y} - \mathbf{1}\bar{y} \|^2}$$

$$= \frac{\hat{\Sigma}_{y,I}\hat{\Sigma}^{-1}_{I,I}\hat{\Sigma}_{I,y}}{\hat{\Sigma}_{yy}} + \frac{\hat{\Sigma}_{yJ\bullet I}\hat{\Sigma}^{-1}_{JJ\bullet I}\hat{\Sigma}_{Jy\bullet I}}{\hat{\Sigma}_{yy}}$$

$$\hat{=} R^2_I + \Delta R^2$$

(2) $$\Delta R^2 = \frac{\hat{\Sigma}_{yJ\bullet I}\hat{\Sigma}^{-1}_{JJ\bullet I}\hat{\Sigma}_{Jy\bullet I}}{\hat{\Sigma}_{yy}} = \frac{\hat{\Sigma}_{yJ\bullet I}\hat{\Sigma}^{-1}_{JJ\bullet I}\hat{\Sigma}_{Jy\bullet I}}{\hat{\Sigma}_{yy\bullet I}} \times \left(\frac{\hat{\Sigma}_{yy\bullet I}}{\hat{\Sigma}_{yy}} \right)$$

$$= R^2_{I\bullet J} \times \left(\frac{\hat{\Sigma}_{yy} - \hat{\Sigma}_{y,I}\hat{\Sigma}^{-1}_{I,I}\hat{\Sigma}_{I,y}}{\hat{\Sigma}_{yy}} \right)$$

$$= R^2_{I\bullet J} \times \left(1 - R^2_I\right)$$

方差膨胀因子

自变量之间的相关性称为复共线性，
复共线性对于LS估计的方差有何影响？

模型$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{1}\beta_0 + Z\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim (0, \sigma^2)$, 其中

$$\boldsymbol{\beta} = \left(\beta_0, \beta_1, \ldots, \beta_{p-1}\right)^{\mathsf{T}},$$
$$X = \left(\mathbf{x}_{(0)} = \mathbf{1}, \mathbf{x}_{(1)} \ldots, \mathbf{x}_{(p-1)}\right),$$

由命题2, $k \geq 1, \quad \mathbf{x}_{(k)}^{\perp} = \mathbf{x}_{(k)} - P_{X_{(-k)}} \mathbf{x}_{(k)}$,

$$\operatorname{var}(\hat{\beta}_k \mid X) = \sigma^2 / \| \mathbf{x}_{(k)}^{\perp} \|^2 = \sigma^2 / [(n-1)\hat{\Sigma}_{\mathbf{x}_k \mathbf{x}_k \bullet \mathbf{x}_{-k}}].$$

$\| \mathbf{x}_{(k)}^{\perp} \|$的大小依赖于$\mathbf{x}_{(k)}$与其它列的线性相关程度。

如果$\mathbf{x}_k$与Z的其它列正交，则 $\operatorname{var}_0(\hat{\beta}_k \mid X) = \dfrac{\sigma^2}{\|\mathbf{x}_{(k)} - \mathbf{1}\bar{x}_{(k)}\|^2}$

如果$\mathbf{x}_k$与其它列不正交, 则

$$\operatorname{var}(\hat{\beta}_k \mid X) = \frac{\sigma^2}{\|\mathbf{x}_{(k)}^{\perp}\|^2} = \frac{\|\mathbf{x}_{(k)} - \mathbf{1}\bar{x}_{(k)}\|^2}{\|\mathbf{x}_{(k)}^{\perp}\|^2} \times \frac{\sigma^2}{\|\mathbf{x}_{(k)} - \mathbf{1}\bar{x}_k\|^2}$$

该方差比正交情形下的方差扩大的倍数为$\|\mathbf{x}_{(k)} - \mathbf{1}\bar{x}_k\|^2 / \|\mathbf{x}_{(k)}^{\perp}\|^2$

定义：对于带截距项的回归，设计阵$X =（\mathbf{x}_{(0)} = \mathbf{1}, \mathbf{x}_{(1)} = \ldots, \mathbf{x}_{(p-1)})$,

对$1 \le k \le p-1$,定义方差膨胀因子(variance inflation factor)

$$\mathrm{VIF}_k = \|\mathbf{x}_{(k)} - \mathbf{1}\bar{x}_k\|^2 / \|\mathbf{x}_{(k)}^{\perp}\|^2 = \hat{\Sigma}_{\mathbf{x}_k\mathbf{x}_k} / \hat{\Sigma}_{\mathbf{x}_k\mathbf{x}_k \bullet \mathbf{x}_{-k}}$$

命题5. $\mathrm{VIF}_k = \frac{1}{1-R_k^2} \geq 1, R_k^2 = \frac{\| \hat{\mathbf{x}}_{(k)} - \mathbf{1}\bar{x}_k \|^2}{\| \mathbf{x}_{(k)} - \mathbf{1}\bar{x}_k \|^2} = \frac{\hat{\Sigma}_{k,-k}\hat{\Sigma}_{-k,-k}^{-1}\hat{\Sigma}_{-k,k}}{\hat{\Sigma}_{k,k}}$,

其中$\hat{\mathbf{x}}_{(k)} = \mathrm{P}_{X_{(-k)}}\mathbf{x}_{(k)}$。

证：因为 $\mathbf{x}_{(k)} = \hat{\mathbf{x}}_{(k)} \oplus \mathbf{x}_{(k)}^{\perp}$, 且$\mathbf{x}_{(k)}^{\perp} \perp \hat{\mathbf{x}}_{(k)}$，$\mathbf{x}_{(k)}^{\perp} \perp \mathbf{1}$，所以

$$(\mathbf{x}_{(k)} - \mathbf{1}\bar{x}_k) = (\hat{\mathbf{x}}_{(k)} - \mathbf{1}\bar{x}_k) \oplus \mathbf{x}_k^{\perp}$$

所以$\frac{1}{\mathrm{VIF}_k} = \frac{\| \mathbf{x}_{(k)}^{\perp} \|^2}{\| \mathbf{x}_{(k)} - \mathbf{1}\bar{x}_k \|^2} = \frac{\| \mathbf{x}_{(k)} - \mathbf{1}\bar{x}_k \|^2 - \| \hat{\mathbf{x}}_{(k)} - \mathbf{1}\bar{x}_k \|^2}{\| \mathbf{x}_{(k)} - \mathbf{1}\bar{x}_k \|^2} = 1 - R_k^2$

$0 \leq R_k^2 \leq 1 \Rightarrow \mathrm{VIF}_k \geq 1$。

注：R_k^2为$\mathbf{x}_{(k)}$对其它自变量回归的复相关系数平方,度量了$\mathbf{x}_{(k)}$与其它自变量的相关性，故VIF与复共线性有关。

例子：两变量回归

我们以两个自变量的模型为例，考察LS估计以及R^2.

例1. 假设数据$(y_i, x_{i1}, x_{i2}), i=1,2,\ldots,n$,满足模型：

$$y_i = a + bx_{i1} + cx_{i2} + \varepsilon_i,\quad \varepsilon_i \ \text{iid} \sim (0, \sigma^2),\ \varepsilon_i \text{ 与 } x_{i1}, x_{i2} \text{ 独立。}$$

记号：

$$s_{12} = \sum (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2),$$

$$s_{11} = \sum (x_{i1} - \bar{x}_1)^2$$

$$s_{y1} = \sum (y_i - \bar{y})(x_{i1} - \bar{x}_1) \text{等等}$$

对于两变量回归，我们有如下结果：

(1) $$\hat{b}=\frac{s_{y1}-s_{y2}s_{12}/s_{22}}{s_{11}-s_{12}s_{21}/s_{22}}=\sqrt{\frac{s_{yy}}{s_{11}}}\left(\frac{r_{y1}-r_{y2}r_{12}}{1-r_{12}^2}\right),$$

$$\operatorname{var}(\hat{b})=\frac{\sigma^2}{s_{11}-s_{12}^2/s_{22}}=\frac{\sigma^2}{s_{11}}\times\frac{1}{1-r_{12}^2},VIF=\frac{1}{1-r_{12}^2}$$

(2) $$\Delta R^2=R_{12}^2-R_1^2=\frac{\left(r_{y2}-r_{y1}r_{12}\right)^2}{1-r_{12}^2},\quad R_{2\bullet 1}^2=\frac{\left(r_{y2}-r_{y1}r_{12}\right)^2}{(1-r_{y1}^2)(1-r_{12}^2)}=(r_{y2\bullet 1})^2,$$

其中$r_{y2\bullet 1}=\dfrac{r_{y2}-r_{y1}r_{12}}{\sqrt{(1-r_{y1}^2)(1-r_{12}^2)}}$为$y$和$x_2$的偏相关系数.

(3) $$R^2=R_1^2+\Delta R^2=\frac{r_{y1}^2+r_{y2}^2-2r_{y1}r_{y2}r_{12}}{1-r_{12}^2}.$$

(1) 由推论1,

$$\hat{b}=\frac{s_{y1\bullet 2}}{s_{11\bullet 2}}=\frac{s_{y1}-s_{y2}s_{12}/s_{22}}{s_{11}-s_{12}s_{21}/s_{22}}=\sqrt{\frac{s_{yy}}{s_{11}}}\left(\frac{r_{y1}-r_{y2}r_{12}}{1-r_{12}^2}\right)$$

$$\mathrm{var}(\hat{b})=\frac{\sigma^2}{s_{11\bullet 2}}=\frac{\sigma^2}{s_{11}-s_{12}^2/s_{22}}=\frac{\sigma^2}{s_{11}}\times\frac{1}{1-r_{12}^2},\qquad VIF=\frac{1}{1-r_{12}^2}$$

$$R_{I\cup J}^2=\frac{\hat{\Sigma}_{y,I}\hat{\Sigma}_{I,I}^{-1}\hat{\Sigma}_{I,y}}{\hat{\Sigma}_{yy}}+\frac{\hat{\Sigma}_{yJ\bullet I}\hat{\Sigma}_{JJ\bullet I}^{-1}\hat{\Sigma}_{Jy\bullet I}}{\hat{\Sigma}_{yy}}\hat{=}R_I^2+\Delta R^2=R_I^2+(1-R_I^2)R_{J\bullet I}^2,$$

(2) 由命题4, $R_1^2=\dfrac{\hat{\Sigma}_{y1}\hat{\Sigma}_{11}^{-1}\hat{\Sigma}_{1y}}{\hat{\Sigma}_{yy}}=r_{y1}^2$

$$\Delta R^2 = R_{12}^2 - R_1^2 = \frac{\hat{\Sigma}_{y2\bullet 1}\hat{\Sigma}_{22\bullet 1}^{-1}\hat{\Sigma}_{2y\bullet 1}}{\hat{\Sigma}_{yy}} = \frac{(\hat{\Sigma}_{y2} - \hat{\Sigma}_{y1}\hat{\Sigma}_{11}^{-1}\hat{\Sigma}_{12})^2(\hat{\Sigma}_{22} - \hat{\Sigma}_{21}\hat{\Sigma}_{11}^{-1}\hat{\Sigma}_{12})^{-1}}{\hat{\Sigma}_{yy}}$$

$$= \frac{\left(r_{y2} - r_{y1}r_{12}\right)^2}{1 - r_{12}^2} = \frac{\left(r_{y2} - r_{y1}r_{12}\right)^2}{(1 - r_{y1}^2)(1 - r_{12}^2)} \times (1 - r_{y1}^2)$$

偏相关系数(3个变量情形)

$$r_{12\bullet 3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

$$= r_{y2\bullet 1}^2 (1 - r_{y1}^2)$$

(3)所以$R^2 = R_1^2 + \Delta R^2 = r_{y1}^2 + \dfrac{\left(r_{y2} - r_{y1}r_{12}\right)^2}{1 - r_{12}^2} = \dfrac{r_{y1}^2 + r_{y2}^2 - 2r_{y1}r_{y2}r_{12}}{1 - r_{12}^2}$

总结如下：$R_1^2 = r_{y1}^2$

偏相关系数:$R_{2\bullet 1} = r_{y2\bullet 1} = \dfrac{r_{y2} - r_{y1}r_{12}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{y1}^2}}$,

半偏相关系数: $\Delta R^2 = \dfrac{r_{y2} - r_{y1}r_{12}}{\sqrt{1-r_{12}^2}} = R_{2\bullet 1}\sqrt{1-r_{y1}^2}$,

$$R^2 = R_1^2 + \Delta R^2 = R_1^2 + R_{2\bullet 1}(1-R_1^2) = \frac{r_{y1}^2 + r_{y2}^2 - 2r_{y1}r_{y2}r_{12}}{1-r_{12}^2}.$$

对称地我们也有$R_{12}^2 - R_2^2 = \dfrac{(r_{y1} - r_{y2}r_{12})^2}{1-r_{12}^2} = r_{y1\bullet 2}^2(1-r_{y2}^2)$，$R_{1\bullet 2}^2 = (r_{y1\bullet 2})^2$

$$R^2 = R_{12}^2 = r_{y2}^2 + \frac{(r_{y1} - r_{y2}r_{12})^2}{1-r_{12}^2} = \frac{r_{y1}^2 + r_{y2}^2 - 2r_{y1}r_{y2}r_{12}}{1-r_{12}^2}$$