Schur Complement

Oliver_Shu

March 31, 2019

For a definite positive matrix

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{12}^\intercal & A_{22} \end{bmatrix}$$

Its inverse

$$A^{-1} = \begin{bmatrix} B_1 & X \\ Y & B_2 \end{bmatrix}$$

where B_1, B_2, X, Y all have different expressions, listed as follow:

$$B_1 = (A_{11} - A_{12}A_{22}^{-1}A_{12}^\intercal)^{-1} = A_{11}^{-1} + A_{11}^{-1}A_{12}B_2A_{12}^\intercal A_{11}^{-1} \tag{1}$$

$$B_2 = (A_{22} - A_{12}^\intercal A_{11}^{-1}A_{12})^{-1} = A_{22}^{-1} + A_{22}^{-1}A_{12}^\intercal B_1 A_{12}A_{22}^{-1} \tag{2}$$

$$X = -B_1A_{12}A_{22}^{-1} = -A_{11}^{-1}A_{12}B_2 \tag{3}$$

$$Y = -A_{22}^{-1}A_{12}^\intercal B_1 = -B_2A_{12}^\intercal A_{11}^{-1} \tag{4}$$

As long as one of B_1 and B_2 has a concise expression, we can use it to express the other three terms.

We can prove it by doing Gaussian elimination,

$$\begin{bmatrix} I & 0 \\ -A_{12}^\intercal A_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{12}^\intercal & A_{22} \end{bmatrix} \begin{bmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & I \end{bmatrix} = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} - A_{12}^\intercal A_{11}^{-1}A_{12} \end{bmatrix}$$

Denote it as $PAP^\intercal = B$, then $P^{-\intercal}A^{-1}P^{-1} = B^{-1}$, $A^{-1} = P^\intercal B^{-1} P$.

$$\begin{aligned} A^{-1} &= \begin{bmatrix} I & -A_{11}^{-1}A_{12} \\ 0 & I \end{bmatrix} \begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & B_2 \end{bmatrix} \begin{bmatrix} I & 0 \\ -A_{12}^\intercal A_{11}^{-1} & I \end{bmatrix} \\ &= \begin{bmatrix} A_{11}^{-1} & -A_{11}^{-1}A_{12}B_2 \\ 0 & B_2 \end{bmatrix} \begin{bmatrix} I & 0 \\ -A_{12}^\intercal A_{11}^{-1} & I \end{bmatrix} \\ &= \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}B_2A_{12}^\intercal A_{11}^{-1} & -A_{11}^{-1}A_{12}B_2 \\ -B_2A_{12}^\intercal A_{11}^{-1} & B_2 \end{bmatrix} \end{aligned}$$

where $B_2 = (A_{22} - A_{12}^\intercal A_{11}^{-1}A_{12})^{-1}$ is called the Schur complement of element A_{22} of matrix A.

By the same manner,

$$A^{-1} = \begin{bmatrix} B_1 & -B_1A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{12}^\intercal B_1 & A_{22}^{-1} + A_{22}^{-1}A_{12}^\intercal B_1A_{12}A_{22}^{-1} \end{bmatrix}$$

We provide examples from Mutiple Linear Regression to illustrate the power of the expressions above.

1 Adding a regressor

As we all know, adding a regressor can reduce RSS. However, the mean square error of the prediction at a given point increases.

Suppose now we have p regressors (whether the intercept term is included does not matter here) and n smaple points. The initial design matrix reads

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

and the design matrix after adding a regressor reads

$$\tilde{\mathbf{X}} = \begin{bmatrix} x_{11} & \cdots & x_{1p} & x_{1p+1} \\ x_{21} & \cdots & x_{2p} & x_{2p+1} \\ \vdots & \vdots & \vdots & \\ x_{n1} & \cdots & x_{np} & x_{np+1} \end{bmatrix} = [\mathbf{X}, \mathbf{r}_{p+1}]$$

where $\mathbf{r}_{p+1} = [x_{1p+1}, x_{2p+1}, \cdots, x_{np+1}]^\mathsf{T}$.

$$\tilde{\mathbf{X}}^\mathsf{T}\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X}^\mathsf{T}\mathbf{X} & \mathbf{X}^\mathsf{T}\mathbf{r}_{p+1} \\ \mathbf{r}_{p+1}^\mathsf{T}\mathbf{X} & \mathbf{r}_{p+1}^\mathsf{T}\mathbf{r}_{p+1} \end{bmatrix}$$

Then

$$(\tilde{\mathbf{X}}^\mathsf{T}\tilde{\mathbf{X}})^{-1} = \begin{bmatrix} (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} + b\mathbf{c}\mathbf{c}^\mathsf{T} & -b\mathbf{c} \\ -b\mathbf{c}^\mathsf{T} & b \end{bmatrix}$$

where $b = 1/(\mathbf{r}_{p+1}^\mathsf{T}\mathbf{r}_{p+1} - \mathbf{r}_{p+1}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{r}_{p+1})$ is a positive real number and $\mathbf{c} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{r}_{p+1}$ is a n–dimension column vector. These two complex expressions are of no avail afterwards.

Now suppose the prediction is at point $\tilde{\mathbf{x}} = [x_1, \cdots, x_p, x_{p+1}]$. Let $\mathbf{x} = [x_1, \cdots, x_p]$, $\tilde{\mathbf{x}} = [\mathbf{x}, x_{p+1}]$. The MSE of prediction at $\mathbf{x}$ in the linear model of p regressors is

$$\text{sepred}^2(y|\mathbf{x}) = \sigma^2 + \sigma^2\mathbf{x}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}^\mathsf{T}$$

The MSE of prediction at $\bar{\mathbf{x}}$ in the linear model of $p+1$ regressors is

$$\text{sepred}^2(y|\bar{\mathbf{x}}) = \sigma^2 + \sigma^2\tilde{\mathbf{x}}(\tilde{\mathbf{X}}^\mathsf{T}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{x}}^\mathsf{T}$$

$$\begin{aligned} \tilde{\mathbf{x}}(\tilde{\mathbf{X}}^\mathsf{T}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{x}}^\mathsf{T} &= [\mathbf{x}, x_{p+1}] \begin{bmatrix} (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} + b\mathbf{c}\mathbf{c}^\mathsf{T} & -b\mathbf{c} \\ -b\mathbf{c}^\mathsf{T} & b \end{bmatrix} \begin{bmatrix} \mathbf{x}^\mathsf{T} \\ x_{p+1} \end{bmatrix} \\ &= \mathbf{x}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}^\mathsf{T} + b(\mathbf{x}\mathbf{c} - x_{p+1})^2 \\ &\geq \mathbf{x}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}^\mathsf{T} \end{aligned}$$

Then

$$\text{sepred}(y|\mathbf{x}) \leq \text{sepred}(y|\tilde{\mathbf{x}})$$

2 Added-Variable Plots

Consider the model with intercept. Now we have $p+1$ regressors $\mathbf{X}_1, \cdots, \mathbf{X}_{p+1}$ and we want to see the effect of adding $\mathbf{X}_{p+1}$ to the model that includes $\mathbf{X}_1, \cdots, \mathbf{X}_p$.

Let the two design matrix:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \qquad \mathbf{X}_2 = [\mathbf{X}, \mathbf{r}_{p+1}]$$

where $\mathbf{r}_{p+1} = [x_{1p+1}, x_{2p+1}, \cdots, x_{np+1}]^\top$.

The two projection matrix:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \qquad \mathbf{H}_2 = \mathbf{X}_2(\mathbf{X}_2^\top\mathbf{X}_2)^{-1}\mathbf{X}_2^\top$$

Let $\hat{\mathbf{e}_1}$ be the residuals from the regression of $\mathbf{Y}$ on $\mathbf{X}_1, \cdots, \mathbf{X}_p$,

$$\hat{\mathbf{e}_1} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

which is the part of the response $\mathbf{Y}$ not explained by the regression on $\mathbf{X}_1, \cdots, \mathbf{X}_p$.

Let $\hat{\mathbf{e}_2}$ be the residuals from the regression of $\mathbf{X}_{p+1}$ on $\mathbf{X}_1, \cdots, \mathbf{X}_p$,

$$\hat{\mathbf{e}_2} = (\mathbf{I} - \mathbf{H})\mathbf{r}_{p+1}$$

which is the part of $\mathbf{X}_{p+1}$ not explained by the regression on $\mathbf{X}_1, \cdots, \mathbf{X}_p$.

Then we do the regression of $\hat{\mathbf{e}_1}$ on $\hat{\mathbf{e}_2}$. We will prove the following two statements:

(1) The estimated slope is exactly the estimate $\hat{\beta_{p+1}}$ in the regression of $\mathbf{Y}$ on the whole $p+1$ regressors.

(2) The residuals in the added-variable plot are identical to the residuals from regression of $\mathbf{Y}$ on the whole $p+1$ regressors.

For simpler notation, we denote $\mathbf{r}_{p+1}$ as $\mathbf{r}$ in the proof below.

2.1 Slope

Since we are working with the model with intercept, the means of $\hat{\mathbf{e}_1}$ and $\hat{\mathbf{e}_2}$ are both 0. Then the estimated slope of the regression of $\hat{\mathbf{e}_1}$ on $\hat{\mathbf{e}_2}$ is just

$$\begin{aligned} \text{estimated slope} &= \frac{\hat{\mathbf{e}_1}^\top\hat{\mathbf{e}_2}}{\hat{\mathbf{e}_2}^\top\hat{\mathbf{e}_2}} \\ &= \frac{\mathbf{Y}^\top(\mathbf{I}-\mathbf{H})\mathbf{r}}{\mathbf{r}^\top(\mathbf{I}-\mathbf{H})\mathbf{r}} \end{aligned}$$

Next we find the expression for $\hat{\beta_{p+1}}$, which is the last element of $(\mathbf{X}_2^\top\mathbf{X}_2)^{-1}\mathbf{X}_2^\top\mathbf{Y}$.

$$\mathbf{X}_2^\top\mathbf{X}_2 = \begin{bmatrix} \mathbf{X}^\top\mathbf{X} & \mathbf{X}^\top\mathbf{r} \\ \mathbf{r}^\top\mathbf{X} & \mathbf{r}^\top\mathbf{r} \end{bmatrix}$$

The Schur complement of element $\mathbf{r}^\top\mathbf{r}$ equals

$$(\mathbf{r}^\top\mathbf{r} - \mathbf{r}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{r})^{-1} = (\mathbf{r}^\top(\mathbf{I}-\mathbf{H})\mathbf{r})^{-1}$$

Then

$$(\mathbf{X}_2^\top\mathbf{X}_2)^{-1} = \begin{bmatrix} (\mathbf{X}^\top\mathbf{X})^{-1} + \frac{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{r}\mathbf{r}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}}{\mathbf{r}^\top(\mathbf{I}-\mathbf{H})\mathbf{r}} & -\frac{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{r}}{\mathbf{r}^\top(\mathbf{I}-\mathbf{H})\mathbf{r}} \\ -\frac{\mathbf{r}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}}{\mathbf{r}^\top(\mathbf{I}-\mathbf{H})\mathbf{r}} & \frac{1}{\mathbf{r}^\top(\mathbf{I}-\mathbf{H})\mathbf{r}} \end{bmatrix} \tag{5}$$

$$\begin{bmatrix} * \\ \hat{\beta_{p+1}} \end{bmatrix} = \begin{bmatrix} * & * \\ -\frac{\mathbf{r}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}}{\mathbf{r}^\top (\mathbf{I}-\mathbf{H})\mathbf{r}} & \frac{1}{\mathbf{r}^\top (\mathbf{I}-\mathbf{H})\mathbf{r}} \end{bmatrix} \begin{bmatrix} \mathbf{X}^\top \mathbf{Y} \\ \mathbf{r}^\top \mathbf{Y} \end{bmatrix}$$

$$\begin{aligned} \hat{\beta_{p+1}} &= \frac{-\mathbf{r}^\top \mathbf{H}\mathbf{Y} + \mathbf{r}^\top \mathbf{Y}}{\mathbf{r}^\top (\mathbf{I}-\mathbf{H})\mathbf{r}} \\ &= \frac{\mathbf{r}^\top (\mathbf{I}-\mathbf{H})\mathbf{Y}}{\mathbf{r}^\top (\mathbf{I}-\mathbf{H})\mathbf{r}} \\ &= \text{estimated slope} \end{aligned}$$

2.2 Residuals

In a simple linear regression, residual $\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x})$. Here since the means of $\hat{\mathbf{e}_1}$ and $\hat{\mathbf{e}_2}$ are 0, the residuals from the regression of $\hat{\mathbf{e}_1}$ on $\hat{\mathbf{e}_2}$ are just

$$\begin{aligned} &\hat{\mathbf{e}_1} - (\text{estimated slope})\hat{\mathbf{e}_2} \\ =&(\mathbf{I}-\mathbf{H})\mathbf{Y} - \frac{\mathbf{r}^\top (\mathbf{I}-\mathbf{H})\mathbf{Y}}{\mathbf{r}^\top (\mathbf{I}-\mathbf{H})\mathbf{r}}(\mathbf{I}-\mathbf{H})\mathbf{r} \end{aligned}$$

We want to prove it is equal to the residual from the regression of $\mathbf{Y}$ on $\mathbf{X}_1, \cdots, \mathbf{X}_{p+1}$, which is $(\mathbf{I}-\mathbf{H}_2)\mathbf{Y}$, and this is equivalent to prove

$$(\mathbf{H}_2 - \mathbf{H})\mathbf{Y} = \frac{\mathbf{r}^\top (\mathbf{I}-\mathbf{H})\mathbf{Y}}{\mathbf{r}^\top (\mathbf{I}-\mathbf{H})\mathbf{r}}(\mathbf{I}-\mathbf{H})\mathbf{r}$$

From (5),

$$\begin{aligned} \mathbf{H}_2 &= \mathbf{X}_2(\mathbf{X}_2^\top \mathbf{X}_2)^{-1}\mathbf{X}_2^\top \\ &= \begin{bmatrix} \mathbf{X} & \mathbf{r} \end{bmatrix} \begin{bmatrix} (\mathbf{X}^\top \mathbf{X})^{-1} + \frac{(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{r}\mathbf{r}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}}{\mathbf{r}^\top (\mathbf{I}-\mathbf{H})\mathbf{r}} & -\frac{(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{r}}{\mathbf{r}^\top (\mathbf{I}-\mathbf{H})\mathbf{r}} \\ -\frac{\mathbf{r}^\top \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}}{\mathbf{r}^\top (\mathbf{I}-\mathbf{H})\mathbf{r}} & \frac{1}{\mathbf{r}^\top (\mathbf{I}-\mathbf{H})\mathbf{r}} \end{bmatrix} \begin{bmatrix} \mathbf{X}^\top \\ \mathbf{r}^\top \end{bmatrix} \\ &= \mathbf{H} + \frac{\mathbf{H}\mathbf{r}\mathbf{r}^\top \mathbf{H} - \mathbf{r}\mathbf{r}^\top \mathbf{H} - \mathbf{H}\mathbf{r}\mathbf{r}^\top + \mathbf{r}\mathbf{r}^\top}{\mathbf{r}^\top (\mathbf{I}-\mathbf{H})\mathbf{r}} \end{aligned}$$

Then

$$\begin{aligned} (\mathbf{H}_2 - \mathbf{H})\mathbf{Y} &= \frac{1}{\mathbf{r}^\top (\mathbf{I}-\mathbf{H})\mathbf{r}}(\mathbf{H}\mathbf{r}\mathbf{r}^\top \mathbf{H} - \mathbf{r}\mathbf{r}^\top \mathbf{H} - \mathbf{H}\mathbf{r}\mathbf{r}^\top + \mathbf{r}\mathbf{r}^\top)\mathbf{Y} \\ &= \frac{1}{\mathbf{r}^\top (\mathbf{I}-\mathbf{H})\mathbf{r}}(\mathbf{I}-\mathbf{H})\mathbf{r}\mathbf{r}^\top (\mathbf{I}-\mathbf{H})\mathbf{Y} \\ &= \frac{\mathbf{r}^\top (\mathbf{I}-\mathbf{H})\mathbf{Y}}{\mathbf{r}^\top (\mathbf{I}-\mathbf{H})\mathbf{r}}(\mathbf{I}-\mathbf{H})\mathbf{r} \end{aligned}$$

Proof is completed.

2.3 Variance of estimators

$\hat{\text{Var}}(\text{estimated slope}) = \frac{\tilde{\sigma}^2}{\mathbf{r}^\top (\mathbf{I}-\mathbf{H})\mathbf{r}} = \frac{\text{RSS}}{(n-2)\mathbf{r}^\top (\mathbf{I}-\mathbf{H})\mathbf{r}}$.

$\hat{\text{Var}}(\hat{\beta_{p+1}}) = \frac{\hat{\sigma}^2}{\mathbf{r}^\top (\mathbf{I}-\mathbf{H})\mathbf{r}} = \frac{\text{RSS}}{(n-p-1)\mathbf{r}^\top (\mathbf{I}-\mathbf{H})\mathbf{r}}$. Here RSS corresponds to that of the regression of $\mathbf{Y}$ on the whole $p+1$ regressors. d.f. differs.

3 Corrected sum of squares and cross products

Consider the model with intercept.

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \qquad \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Let

$$\mathcal{X} = \begin{bmatrix} x_{11} - \bar{x_1} & \cdots & x_{1p} - \bar{x_p} \\ x_{21} - \bar{x_1} & \cdots & x_{2p} - \bar{x_p} \\ \vdots & \vdots & \vdots \\ x_{n1} - \bar{x_1} & \cdots & x_{np} - \bar{x_p} \end{bmatrix} \qquad \mathcal{Y} = \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix}$$

where $\bar{x_i}$ is the mean of the i–th regressor,

$$\bar{x_i} = \frac{1}{n}\sum_{k=1}^{n} x_{ki}, \qquad i = 1, \cdots, p$$

Let $\bar{\mathbf{x}} = [\bar{x_1}, \cdots, \bar{x_p}]$, it is obvious that

$$\mathbf{X}^\mathsf{T}\mathbf{X} = \begin{bmatrix} n & n\bar{\mathbf{x}} \\ n\bar{\mathbf{x}}^\mathsf{T} & \mathcal{X}^\mathsf{T}\mathcal{X} + n\bar{\mathbf{x}}^\mathsf{T}\bar{\mathbf{x}} \end{bmatrix}$$
$$\mathbf{X}^\mathsf{T}\mathbf{Y} = \begin{bmatrix} n\bar{y} \\ \mathcal{X}^\mathsf{T}\mathcal{Y} + n\bar{y}\bar{\mathbf{x}}^\mathsf{T} \end{bmatrix}$$

The Schur complement of element A_{22} of $\mathbf{X}^\mathsf{T}\mathbf{X}$ is simple:

$$(A_{22} - A_{12}^\mathsf{T}A_{11}^{-1}A_{12})^{-1} = (\mathcal{X}^\mathsf{T}\mathcal{X} + n\bar{\mathbf{x}}^\mathsf{T}\bar{\mathbf{x}} - n\bar{\mathbf{x}}^\mathsf{T}n^{-1}n\bar{\mathbf{x}})^{-1} = (\mathcal{X}^\mathsf{T}\mathcal{X})^{-1}$$

Also

$$A_{11}^{-1} + A_{11}^{-1}A_{12}B_2A_{12}^\mathsf{T}A_{11} = \frac{1}{n} + \frac{1}{n}n\bar{\mathbf{x}}(\mathcal{X}^\mathsf{T}\mathcal{X})^{-1}n\bar{\mathbf{x}}^\mathsf{T}\frac{1}{n} = \frac{1}{n} + \bar{\mathbf{x}}(\mathcal{X}^\mathsf{T}\mathcal{X})^{-1}\bar{\mathbf{x}}^\mathsf{T}$$
$$-A_{11}^{-1}A_{12}B_2 = -\frac{1}{n}n\bar{\mathbf{x}}(\mathcal{X}^\mathsf{T}\mathcal{X})^{-1} = -\bar{\mathbf{x}}(\mathcal{X}^\mathsf{T}\mathcal{X})^{-1}$$

Then

$$(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = \begin{bmatrix} \frac{1}{n} + \bar{\mathbf{x}}(\mathcal{X}^\mathsf{T}\mathcal{X})^{-1}\bar{\mathbf{x}}^\mathsf{T} & -\bar{\mathbf{x}}(\mathcal{X}^\mathsf{T}\mathcal{X})^{-1} \\ -(\mathcal{X}^\mathsf{T}\mathcal{X})^{-1}\bar{\mathbf{x}}^\mathsf{T} & (\mathcal{X}^\mathsf{T}\mathcal{X})^{-1} \end{bmatrix}$$

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y} \\ &= \begin{bmatrix} \frac{1}{n} + \bar{\mathbf{x}}(\mathcal{X}^\mathsf{T}\mathcal{X})^{-1}\bar{\mathbf{x}}^\mathsf{T} & -\bar{\mathbf{x}}(\mathcal{X}^\mathsf{T}\mathcal{X})^{-1} \\ -(\mathcal{X}^\mathsf{T}\mathcal{X})^{-1}\bar{\mathbf{x}}^\mathsf{T} & (\mathcal{X}^\mathsf{T}\mathcal{X})^{-1} \end{bmatrix} \begin{bmatrix} n\bar{y} \\ \mathcal{X}^\mathsf{T}\mathcal{Y} + n\bar{y}\bar{\mathbf{x}}^\mathsf{T} \end{bmatrix} \\ &= \begin{bmatrix} \bar{y} - \bar{\mathbf{x}}(\mathcal{X}^\mathsf{T}\mathcal{X})^{-1}\mathcal{X}^\mathsf{T}\mathcal{Y} \\ (\mathcal{X}^\mathsf{T}\mathcal{X})^{-1}\mathcal{X}^\mathsf{T}\mathcal{Y} \end{bmatrix} \end{aligned}$$

This is to say

$$\hat{\boldsymbol{\beta}}^* = (\mathcal{X}^\mathsf{T}\mathcal{X})^{-1}\mathcal{X}^\mathsf{T}\mathcal{Y} \qquad \hat{\beta}_0 = \bar{y} - \bar{\mathbf{x}}\hat{\boldsymbol{\beta}}^*$$

where $\hat{\boldsymbol{\beta}}^* = [\hat{\beta}_1, \cdots, \hat{\beta}_p]^\mathsf{T}$.

4 Interpretation of R^2

In mutiple linear regression, R^2 is defined as $1 - \frac{\text{RSS}}{\text{SYY}}$.

The means of $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ are both $\bar{y}$. If we subtract mean $\bar{y}$ from the two vectors,

$$\mathbf{Y} - \bar{y}\mathbb{1}_n = \mathcal{Y} \qquad \hat{\mathbf{Y}} - \bar{y}\mathbb{1}_n = \mathcal{X}\hat{\boldsymbol{\beta}}^*$$

Then

$$\begin{aligned}
\text{RSS} &= ||\mathbf{Y} - \hat{\mathbf{Y}}||^2 \\
&= ||\mathcal{Y} - \mathcal{X}\hat{\boldsymbol{\beta}}^*||^2 \\
&= \text{SYY} - \hat{\boldsymbol{\beta}}^{*\mathsf{T}}\mathcal{X}^\mathsf{T}\mathcal{X}\hat{\boldsymbol{\beta}}^* \\
&= \text{SYY} - \mathcal{Y}^\mathsf{T}\mathcal{X}^\mathsf{T}(\mathcal{X}^\mathsf{T}\mathcal{X})^{-1}\mathcal{X}^\mathsf{T}\mathcal{Y} \\
R^2 &= \frac{\hat{\boldsymbol{\beta}}^{*\mathsf{T}}\mathcal{X}^\mathsf{T}\mathcal{X}\hat{\boldsymbol{\beta}}^*}{\text{SYY}} = \frac{\mathcal{Y}^\mathsf{T}\mathcal{X}^\mathsf{T}(\mathcal{X}^\mathsf{T}\mathcal{X})^{-1}\mathcal{X}^\mathsf{T}\mathcal{Y}}{\text{SYY}}
\end{aligned}$$

Note $\mathcal{X}^\mathsf{T}\mathcal{X}\hat{\boldsymbol{\beta}}^* = \mathcal{X}^\mathsf{T}\mathcal{Y}$.

Here comes the two interpretation of R^2:

(1) square of the correlation between the observed values $\mathbf{Y}$ and the fitted values $\hat{\mathbf{Y}}$.

(2) sqaure of the maximum of the correlation between $\mathbf{Y}$ and any linear combination of the regressors.

4.1 Correlation between the observed values and the fitted values

$$\begin{aligned}
\rho^2_{\mathbf{Y}\hat{\mathbf{Y}}} &= \frac{((\mathcal{X}\hat{\boldsymbol{\beta}}^*)^\mathsf{T}\mathcal{Y})^2}{(\mathcal{Y}^\mathsf{T}\mathcal{Y})(\hat{\boldsymbol{\beta}}^{*\mathsf{T}}\mathcal{X}^\mathsf{T}\mathcal{X}\hat{\boldsymbol{\beta}}^*)} \\
&= \frac{(\hat{\boldsymbol{\beta}}^{*\mathsf{T}}\mathcal{X}^\mathsf{T}\mathcal{Y})^2}{\text{SYY}(\hat{\boldsymbol{\beta}}^{*\mathsf{T}}\mathcal{X}^\mathsf{T}\mathcal{X}\hat{\boldsymbol{\beta}}^*)} \\
&= \frac{\hat{\boldsymbol{\beta}}^{*\mathsf{T}}\mathcal{X}^\mathsf{T}\mathcal{X}\hat{\boldsymbol{\beta}}^*}{\text{SYY}} \\
&= R^2
\end{aligned}$$

4.2 Correlation between the observed values and linear combinations of the regressors

Let the design matrix $\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \triangleq \begin{bmatrix} \mathbb{1}_n & \mathbf{r}_1 & \cdots & \mathbf{r}_p \end{bmatrix}$. Then a linear combination of regressors $\mathbf{r}$ can be written in the form $\mathbf{r} = [\mathbf{r}_1, \cdots, \mathbf{r}_p]\mathbf{a}$, where $\mathbf{a}$ is a p–dimension column vector.

The mean of $\mathbf{r}$ is $\bar{\mathbf{x}}\mathbf{a}$,

$$\mathbf{r} - \bar{\mathbf{x}}\mathbf{a}\mathbb{1}_n = \mathcal{X}\mathbf{a}$$

Then

$$
\begin{aligned}
\rho^2_{\mathbf{Y},\mathbf{r}} &= \frac{((\mathcal{X}\mathbf{a})^\mathsf{T}\mathbf{Y})^2}{(\mathcal{Y}^\mathsf{T}\mathcal{Y})(\mathbf{a}^\mathsf{T}\mathcal{X}^\mathsf{T}\mathcal{X}\mathbf{a})} \\
&= \frac{1}{\mathrm{SYY}}\frac{((\mathcal{X}^\mathsf{T}\mathcal{Y})^\mathsf{T}\mathbf{a})^2}{\mathbf{a}^\mathsf{T}\mathcal{X}^\mathsf{T}\mathcal{X}\mathbf{a}}
\end{aligned}
\tag{6}
$$

For a positive definite matrix $M \in \mathbb{R}^{p\times p}$ and column vectors $a, b \in \mathbb{R}^p$, we can use spectral decomposition to show

$$
(a^\mathsf{T}Ma)(b^\mathsf{T}M^{-1}b) \geq (b^\mathsf{T}a)^2
$$

which means

$$
\frac{(b^\mathsf{T}a)^2}{a^\mathsf{T}Ma} \leq b^\mathsf{T}M^{-1}b
$$

and the equality is achieved if and only if $a = k(M^{-1}b)$, $k \in \mathbb{R}$.

Then from (6),

$$
\begin{aligned}
\max_{\mathbf{r}=\mathcal{X}\mathbf{a},\ \mathbf{a}\in\mathbb{R}^p} \rho^2_{\mathbf{Y},\mathbf{r}} &= \frac{\mathcal{Y}^\mathsf{T}\mathcal{X}(\mathcal{X}^\mathsf{T}\mathcal{X})^{-1}\mathcal{X}^\mathsf{T}\mathcal{Y}}{\mathrm{SYY}} \\
&= R^2
\end{aligned}
$$