

Lecture 4: Multivariate Linear Regression

Linear regression analysis is one of the most widely used statistical methods. We shall start with the multiple linear regression before studying the multivariate model. Our discussion of the multiple linear regression uses matrix algebra.

1 The classical linear regression model

The linear regression model with a single response and several explanatory variables takes the form

$$Y_i = \beta_0 + \beta_1 Z_{i1} + \beta_2 Z_{i2} + \cdots + \beta_r Z_{ir} + \epsilon_i, \quad i = 1, \dots, n, \quad (1)$$

where the error terms $\{\epsilon_i\}$ satisfy

1. $E(\epsilon_i) = 0$ for all i ;
2. $\text{Var}(\epsilon_i) = \sigma^2$, a constant; and
3. $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ if $i \neq j$.

In matrix notation, Eq. (1) becomes

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & Z_{11} & Z_{12} & \cdots & Z_{1r} \\ 1 & Z_{21} & Z_{22} & \cdots & Z_{2r} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & Z_{n1} & Z_{n2} & \cdots & Z_{nr} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

or

$$\mathbf{Y}_{n \times 1} = \mathbf{Z}_{n \times (r+1)} \boldsymbol{\beta}_{(r+1) \times 1} + \boldsymbol{\epsilon}_{n \times 1},$$

and the assumptions become

1. $E(\boldsymbol{\epsilon}) = \mathbf{0}$; and
2. $\text{Cov}(\boldsymbol{\epsilon}) = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2 \mathbf{I}_{n \times n}$.

The matrix \mathbf{Z} is referred to as the *design matrix*.

2 Least squares estimation

The least squares estimate (LSE) of $\boldsymbol{\beta}$ is obtained by minimizing the sum of squares of errors,

$$S(\boldsymbol{\beta}) = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 Z_{i1} - \cdots - \beta_r Z_{ir})^2 = (\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}).$$

Before deriving the LSE, we first define $\tilde{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$ and $\mathbf{H} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$, assuming that the inverse exists. Then, it is easy to see that \mathbf{H} is symmetric and $\mathbf{H}^2 = \mathbf{H}$. The latter property says that \mathbf{H} is an idempotent matrix.

Define $\tilde{\boldsymbol{\epsilon}} = \mathbf{Y} - \mathbf{Z}\tilde{\boldsymbol{\beta}} = \mathbf{Y} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y} = [\mathbf{I} - \mathbf{H}]\mathbf{Y}$. Then,

$$\mathbf{Z}'\tilde{\boldsymbol{\epsilon}} = \mathbf{Z}'[\mathbf{I} - \mathbf{H}]\mathbf{Y} = [\mathbf{Z}' - \mathbf{Z}'\mathbf{H}]\mathbf{Y} = [\mathbf{Z}' - \mathbf{Z}']\mathbf{Y} = \mathbf{0}.$$

Result 7.1 Let \mathbf{Z} have full rank $(r + 1) \leq n$. The LSE of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}.$$

Proof. Since

$$\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta} = \mathbf{Y} - \mathbf{Z}\tilde{\boldsymbol{\beta}} + \mathbf{Z}\tilde{\boldsymbol{\beta}} - \mathbf{Z}\boldsymbol{\beta} = \mathbf{Y} - \mathbf{Z}\tilde{\boldsymbol{\beta}} + \mathbf{Z}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}),$$

we have

$$\begin{aligned} S(\boldsymbol{\beta}) &= (\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}) \\ &= (\mathbf{Y} - \mathbf{Z}\tilde{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{Z}\tilde{\boldsymbol{\beta}}) + (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{Z}'\mathbf{Z}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &\quad + 2(\mathbf{Y} - \mathbf{Z}\tilde{\boldsymbol{\beta}})'\mathbf{Z}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &= (\mathbf{Y} - \mathbf{Z}\tilde{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{Z}\tilde{\boldsymbol{\beta}}) + (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{Z}'\mathbf{Z}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \end{aligned}$$

because $(\mathbf{Y} - \mathbf{Z}\tilde{\boldsymbol{\beta}})'\mathbf{Z} = \tilde{\boldsymbol{\epsilon}}'\mathbf{Z} = \mathbf{0}$. The first term of $S(\boldsymbol{\beta})$ does not depend on $\boldsymbol{\beta}$ and the second term is the squared length of $\mathbf{Z}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Because \mathbf{Z} has full rank, $\mathbf{Z}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \neq \mathbf{0}$ if $\tilde{\boldsymbol{\beta}} \neq \boldsymbol{\beta}$, so the minimum of $S(\boldsymbol{\beta})$ is unique and occurs at $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$. Consequently, the LSE of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$. Q.E.D.

The deviations

$$\hat{\epsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 Z_{i1} - \cdots - \hat{\beta}_r Z_{ir}, \quad i = 1, \dots, n,$$

are called the *residuals*. Let $\widehat{\mathbf{Y}} = \mathbf{Z}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{Y}$ denote the *fitted values* of \mathbf{Y} . The \mathbf{H} matrix is called the *hat* matrix. The LS residuals then become

$$\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \widehat{\mathbf{Y}} = [\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\mathbf{Y} = [\mathbf{I} - \mathbf{H}]\mathbf{Y}.$$

Note that $\mathbf{I} - \mathbf{H}$ is also asymmetric and idempotent.

Our prior discussion shows that the residuals satisfy (a) $\mathbf{Z}'\hat{\boldsymbol{\epsilon}} = \mathbf{0}$ and (b) $\widehat{\mathbf{Y}}'\hat{\boldsymbol{\epsilon}} = 0$. Also, the residual sum of squares (RSS) is

$$RSS = \sum_{i=1}^n \hat{\epsilon}_i^2 = \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} = \mathbf{Y}'[\mathbf{I} - \mathbf{H}]\mathbf{Y}.$$

Sum of squares decomposition: Since $\widehat{\mathbf{Y}}'\widehat{\boldsymbol{\epsilon}} = 0$, we have

$$\mathbf{Y}'\mathbf{Y} = (\widehat{\mathbf{Y}} + \widehat{\boldsymbol{\epsilon}})'(\widehat{\mathbf{Y}} + \widehat{\boldsymbol{\epsilon}}) = \widehat{\mathbf{Y}}'\widehat{\mathbf{Y}} + \widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}.$$

Since the first column of \mathbf{Z} is $\mathbf{1}$, the result $\mathbf{Z}'\widehat{\boldsymbol{\epsilon}} = \mathbf{0}$ includes $0 = \mathbf{1}'\widehat{\boldsymbol{\epsilon}}$. Consequently, we have $\bar{Y} = \widehat{\bar{Y}}$. Subtracting $n\bar{Y}^2 = n(\widehat{\bar{Y}})^2$ from the prior decomposition, we have

$$\mathbf{Y}'\mathbf{Y} - n\bar{Y}^2 = \widehat{\mathbf{Y}}'\widehat{\mathbf{Y}} - n(\widehat{\bar{Y}})^2 + \widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}},$$

or

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n \hat{\epsilon}_i^2.$$

The quantity

$$R^2 = 1 - \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

is the coefficient of determination. It is the proportion of the total variation in the Y_i 's *explained* by the predictors Z_1, \dots, Z_p .

Sampling properties:

1. The LSE $\widehat{\boldsymbol{\beta}}$ satisfies $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $\text{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{Z}'\mathbf{Z})^{-1}$.
2. The residuals satisfy $E(\widehat{\boldsymbol{\epsilon}}) = \mathbf{0}$ and $\text{Cov}(\widehat{\boldsymbol{\epsilon}}) = \sigma^2[\mathbf{I} - \mathbf{H}]$.
3. $E(\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}) = (n - r - 1)\sigma^2$ so that, defining

$$s^2 = \frac{\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}}{n - r - 1} = \frac{\mathbf{Y}'[\mathbf{I} - \mathbf{H}]\mathbf{Y}}{n - r - 1},$$

we have $E(s^2) = \sigma^2$.

4. $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\epsilon}}$ are uncorrelated.

Proof. First, $E(\widehat{\boldsymbol{\beta}}) = E[(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}] = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'E(\mathbf{Y}) = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'[\mathbf{Z}\boldsymbol{\beta} + E(\boldsymbol{\epsilon})] = \boldsymbol{\beta}$. Also, $\text{Cov}(\widehat{\boldsymbol{\beta}}) = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\text{Cov}(\mathbf{Y})\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\sigma^2\mathbf{I})\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1} = \sigma^2(\mathbf{Z}'\mathbf{Z})^{-1}$. Next, $E(\widehat{\boldsymbol{\epsilon}}) = [\mathbf{I} - \mathbf{H}]E(\mathbf{Y}) = [\mathbf{I} - \mathbf{H}]\mathbf{Z}\boldsymbol{\beta} = [\mathbf{Z} - \mathbf{Z}]\boldsymbol{\beta} = \mathbf{0}$. Also, $\text{Cov}(\widehat{\boldsymbol{\epsilon}}) = [\mathbf{I} - \mathbf{H}]\text{Cov}(\mathbf{Y})[\mathbf{I} - \mathbf{H}] = \sigma^2[\mathbf{I} - \mathbf{H}]$, because $[\mathbf{I} - \mathbf{H}]$ is idempotent.

Next, using $\text{tr}(\mathbf{H}) = \text{tr}(\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}') = \text{tr}[(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Z}] = \text{tr}(\mathbf{I}_{r+1}) = r + 1$, we have

$$\begin{aligned} E(\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}) &= E[\text{tr}(\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}})] = E[\text{tr}(\widehat{\boldsymbol{\epsilon}}\widehat{\boldsymbol{\epsilon}}')] \\ &= \text{tr}[\text{Cov}(\widehat{\boldsymbol{\epsilon}})] = \text{tr}[\sigma^2(\mathbf{I} - \mathbf{H})] \\ &= \sigma^2[\text{tr}(\mathbf{I}) - \text{tr}(\mathbf{H})] = \sigma^2(n - r - 1). \end{aligned}$$

Finally, $\text{Cov}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\epsilon}}) = \text{Cov}[(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}, (\mathbf{I} - \mathbf{H})\mathbf{Y}] = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\text{Cov}(\mathbf{Y}, \mathbf{Y})(\mathbf{I} - \mathbf{H}) = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\sigma^2\mathbf{I})(\mathbf{I} - \mathbf{H}) = \sigma^2(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{I} - \mathbf{H}) = \sigma^2(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}' - \mathbf{Z}') = \mathbf{0}$ so that $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\epsilon}}$ are uncorrelated.

Gauss's least squares theorem. Let $\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $E(\boldsymbol{\epsilon}) = \mathbf{0}$, $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$, and $\text{Rank}(\mathbf{Z}) = r + 1$. For any \mathbf{c} , the estimator $\mathbf{c}'\hat{\boldsymbol{\beta}}$ of $\mathbf{c}'\boldsymbol{\beta}$ has the smallest possible variance among all linear estimators of the form $\mathbf{a}'\mathbf{Y}$ that are unbiased for $\mathbf{c}'\boldsymbol{\beta}$.

Proof. For any fixed \mathbf{c} , let $\mathbf{a}'\mathbf{Y}$ be an unbiased estimator of $\mathbf{c}'\boldsymbol{\beta}$. Then, $E(\mathbf{a}'\mathbf{Y}) = \mathbf{c}'\boldsymbol{\beta}$, whatever the value of $\boldsymbol{\beta}$. But $E(\mathbf{a}'\mathbf{Y}) = E[\mathbf{a}'(\mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon})] = \mathbf{a}'\mathbf{Z}\boldsymbol{\beta}$. Consequently, $\mathbf{c}'\boldsymbol{\beta} = \mathbf{a}'\mathbf{Z}\boldsymbol{\beta}$ or equivalently, $(\mathbf{c}' - \mathbf{a}'\mathbf{Z})\boldsymbol{\beta} = 0$ for all $\boldsymbol{\beta}$. In particular, choosing $\boldsymbol{\beta} = (\mathbf{c}' - \mathbf{a}'\mathbf{Z})'$, we obtain $\mathbf{c}' = \mathbf{a}'\mathbf{Z}$ for any unbiased estimator.

Next, $\mathbf{c}'\hat{\boldsymbol{\beta}} = \mathbf{c}'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y} \equiv \mathbf{a}^*\mathbf{Y}$, where $\mathbf{a}^* = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{c}$. Since $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, so $\mathbf{c}'\boldsymbol{\beta} = \mathbf{a}^*\mathbf{Y}$ is an unbiased estimator of $\mathbf{c}'\boldsymbol{\beta}$. Consequently, for any \mathbf{a} satisfying the unbiased requirement $\mathbf{c}' = \mathbf{a}'\mathbf{Z}$,

$$\begin{aligned} \text{Var}(\mathbf{a}'\mathbf{Y}) &= \text{Var}(\mathbf{a}'\mathbf{Z}\boldsymbol{\beta} + \mathbf{a}'\boldsymbol{\epsilon}) = \text{Var}(\mathbf{a}'\boldsymbol{\epsilon}) = \sigma^2\mathbf{a}'\mathbf{a} \\ &= \sigma^2(\mathbf{a} - \mathbf{a}^* + \mathbf{a}^*)'(\mathbf{a} - \mathbf{a}^* + \mathbf{a}^*) \\ &= \sigma^2[(\mathbf{a} - \mathbf{a}^*)'(\mathbf{a} - \mathbf{a}^*) + (\mathbf{a}^*)'\mathbf{a}^*], \end{aligned}$$

where $(\mathbf{a} - \mathbf{a}^*)'\mathbf{a}^* = (\mathbf{a} - \mathbf{a}^*)'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{c} = (\mathbf{a}'\mathbf{Z} - (\mathbf{a}^*)'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{c} = (\mathbf{c}' - \mathbf{c}')(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{c} = 0$. Because \mathbf{a}^* is fixed, and $(\mathbf{a} - \mathbf{a}^*)'(\mathbf{a} - \mathbf{a}^*)$ is positive unless $\mathbf{a} = \mathbf{a}^*$, $\text{Var}(\mathbf{a}'\mathbf{Y})$ is minimized by the choice of $(\mathbf{a}^*)'\mathbf{Y} = \mathbf{c}'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y} = \mathbf{c}'\hat{\boldsymbol{\beta}}$. Q.E.D.

3 Inference

For the multiple linear regression model in (1), we further assume that $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$.

Result 7.4 $\hat{\boldsymbol{\beta}}$ is also the maximum likelihood estimate of $\boldsymbol{\beta}$. In addition, $\hat{\boldsymbol{\beta}} \sim N_{r+1}[\boldsymbol{\beta}, \sigma^2(\mathbf{Z}'\mathbf{Z})^{-1}]$ and is independent of the residuals $\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\beta}}$. Let $\tilde{\sigma}^2$ be the maximum likelihood estimate of σ^2 . Then,

$$n\tilde{\sigma}^2 = \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} \sim \sigma^2\chi_{n-r-1}^2.$$

Proof. Follows what we discussed before for the MLE of multivariate model random sample. Q.E.D.

Note that the MLE $\tilde{\sigma}^2$ of σ^2 is $\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}/n$, which is different from the LSE of $\hat{\sigma}^2$.

Result 7.5. For the Guassin MLR model, a $100(1 - \alpha)$ percent confidence region for $\boldsymbol{\beta}$ is given by

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{Z}'\mathbf{Z}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leq (r + 1)\tilde{\sigma}^2 F_{r+1, n-r-1}(\alpha).$$

Also, simultaneous $100(1 - \alpha)$ percent confidence intervals for the β_j are

$$\hat{\beta}_i \pm \sqrt{\text{Var}(\hat{\beta}_i)}\sqrt{(r + 1)F_{r+1, n-r-1}(\alpha)}, \quad i = 0, 1, \dots, r,$$

where $\text{Var}(\hat{\beta}_i)$ is the diagonal element of $\tilde{\sigma}^2(\mathbf{Z}'\mathbf{Z})^{-1}$ corresponding to $\hat{\beta}_i$.

Remark: The R command for MLR is `lm`, which stands for linear model.