Lecture 34: Properties of the LSE

The following results explain why the LSE is popular.

Gauss-Markov Theorem

Assume a general linear model previously described: $Y = X\beta + \mathscr{E}$ with assumption A2, i.e., $\text{Var}(\mathscr{E}) = \sigma^2 I_n$ and X is of full rank $p < n$. Let $\widehat{\beta}$ be the LSE and $l \in \mathscr{R}^p$ be a fixed vector. Then the $l'\widehat{\beta}$ is the *best linear unbiased estimator* (BLUE) of $l'\beta$ in the sense that it has the minimum variance in the class of unbiased estimators of $l'\beta$ that are linear functions of Y.

Proof.

Since $\widehat{\beta} = (X'X)^{-1}X'Y$, it is a linear function of Y and

$$E(\widehat{\beta}) = (X'X)^{-1}X'E(Y) = (X'X)^{-1}X'X\beta = \beta$$

Thus, $l'\widehat{\beta}$ is unbiased for $l'\beta$.

Let $c'Y$ be any linear unbiased estimator of $l'\beta$, where $c \in \mathscr{R}^p$ is a fixed vector.

Since $c'Y$ is unbiased, $E(c'Y) = c'E(Y) = c'X\beta = l'\beta$ for all β, which implies that $c'X = l'$, i.e., $l = X'c$.

Then

$$\begin{aligned}
\mathrm{Var}(c'Y) &= \mathrm{Var}(c'Y - l'\widehat{\beta} + l'\widehat{\beta}) \\
&= \mathrm{Var}(c'Y - l'\widehat{\beta}) + \mathrm{Var}(l'\widehat{\beta}) \\
&\quad + 2\mathrm{Cov}(c'Y - l'\widehat{\beta}, l'\widehat{\beta}) \\
&= \mathrm{Var}(c'Y - l'\widehat{\beta}) + \mathrm{Var}(l'\widehat{\beta}) \\
&\geq \mathrm{Var}(l'\widehat{\beta})
\end{aligned}$$

where the third equality follows from

$$\begin{aligned}
\mathrm{Cov}(c'Y - l'\widehat{\beta}, l'\widehat{\beta}) &= \mathrm{Cov}(c'Y - l'(X'X)^{-1}X'Y, l'(X'X)^{-1}X'Y) \\
&= \mathrm{Cov}(c'Y, l'(X'X)^{-1}X'Y) - \mathrm{Var}(l'(X'X)^{-1}X'Y) \\
&= c'\mathrm{Var}(Y)X(X'X)^{-1}l - l'(X'X)^{-1}X'\mathrm{Var}(Y)X(X'X)^{-1}l \\
&= \sigma^2 c'X(X'X)^{-1}l - \sigma^2 l'(X'X)^{-1}X'X(X'X)^{-1}l \\
&= \sigma^2 l'(X'X)^{-1}l - \sigma^2 l'(X'X)^{-1}l \\
&= 0.
\end{aligned}$$

An unbiased estimator of σ^2

Because the LSE $\widehat{\beta}$ satisfies $X'X\widehat{\beta} = X'Y$,

$$\begin{aligned} \|Y - X\beta\|^2 &= \|Y - X\widehat{\beta} + X(\widehat{\beta} - \beta)\|^2 + 2(\widehat{\beta} - \beta)'X'(Y - X\widehat{\beta}) \\ &= \|Y - X\widehat{\beta}\|^2 + \|X\widehat{\beta} - X\beta\|^2 \end{aligned}$$

Hence

$$\begin{aligned} E\|Y - X\widehat{\beta}\|^2 &= E\|Y - X\beta\|^2 - E\|X\widehat{\beta} - X\beta\|^2 \\ &= E(Y - X\beta)'(Y - X\beta) - E(\beta - \widehat{\beta})'X'X(\beta - \widehat{\beta}) \\ &= \text{trace}\left(\text{Var}(Y) - \text{Var}(X\widehat{\beta})\right) \\ &= \sigma^2\left[\text{trace}(I_n) - \text{trace}\left(X(X'X)^{-1}X'\text{Var}(Y)X(X'X)^{-1}X'\right)\right] \\ &= \sigma^2\left[n - \text{trace}\left(X(X'X)^{-1}X'X(X'X)^{-1}X'\right)\right] \\ &= \sigma^2\left[n - \text{trace}\left((X'X)^{-1}X'X\right)\right] \\ &= \sigma^2(n-p) \end{aligned}$$

Therefore, an unbiased estimator of σ^2 is $\|Y - X\widehat{\beta}\|^2/(n-p)$.

Residual and SSR

The ith component of the n-dimensional vector $Y - X\widehat{\beta}$ is $Y_i - x_i'\widehat{\beta}$, which is called the ith residual.

The vector $Y - X\widehat{\beta}$ is then called the residual vector and $\|Y - X\widehat{\beta}\|^2$ is the sum of squared residuals and denoted by SSR.

The unbiased estimator of σ^2 we derived is then equal to $\mathrm{SSR}/(n-p)$.

Examples.

- Since simple linear regression is a special case of the general linear model, the SSR defined here is the same as the SSR defined in simple linear regression.
- In the case of one-way ANOVA, the LSE $(\bar{Y}_1, ..., \bar{Y}_k)'$ and

$$\mathrm{SSR} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$$

- In the case of two-way balanced ANOVA, if $c > 1$, then

$$\mathrm{SSR} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} (Y_{ijk} - \bar{Y}_{ij\cdot})^2$$

Residual and SSR

The ith component of the n-dimensional vector $Y - X\widehat{\beta}$ is $Y_i - x_i'\widehat{\beta}$, which is called the ith residual.

The vector $Y - X\widehat{\beta}$ is then called the residual vector and $\|Y - X\widehat{\beta}\|^2$ is the sum of squared residuals and denoted by SSR.

The unbiased estimator of σ^2 we derived is then equal to $\mathrm{SSR}/(n-p)$.

Examples.

- Since simple linear regression is a special case of the general linear model, the SSR defined here is the same as the SSR defined in simple linear regression.
- In the case of one-way ANOVA, the LSE $(\bar{Y}_1, ..., \bar{Y}_k)'$ and

$$\mathrm{SSR} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$$

- In the case of two-way balanced ANOVA, if $c > 1$, then

$$\mathrm{SSR} = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} (Y_{ijk} - \bar{Y}_{ij\cdot})^2$$

The result for the two-way ANOVA can be seen as follows.

$$E(Y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

Hence, the LSE of $E(Y_{ijk})$, which is a linear function of regression parameters, is

$$\widehat{\mu} + \widehat{\alpha}_i + \widehat{\beta}_j + \widehat{\gamma}_{ij} = \bar{Y}_{ij\cdot}$$

Thus, the (i,j,k)th residual is $Y_{ijk} - \bar{Y}_{ij\cdot}$.

The residual vector, however, is 0 when $c = 1$.

Correlation between the LSE and the residual vector

The LSE and the residual vector is always uncorrelated, because

$$\begin{aligned}
\mathrm{Cov}(\widehat{\beta}, Y - X\widehat{\beta}) &= \mathrm{Cov}\left((X'X)^{-1}X'Y, [I_n - X(X'X)^{-1}X']Y\right) \\
&= (X'X)^{-1}X'\mathrm{Var}(Y)[I_n - X(X'X)^{-1}X'] \\
&= \sigma^2(X'X)^{-1}X'[I_n - X(X'X)^{-1}X'] \\
&= \sigma^2[(X'X)^{-1}X' - (X'X)^{-1}X'X(X'X)^{-1}X'] \\
&= 0
\end{aligned}$$

We now consider distributions of the LSE and SSR under normality.

The result for the two-way ANOVA can be seen as follows.

$$E(Y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

Hence, the LSE of $E(Y_{ijk})$, which is a linear function of regression parameters, is

$$\widehat{\mu} + \widehat{\alpha}_i + \widehat{\beta}_j + \widehat{\gamma}_{ij} = \bar{Y}_{ij\cdot}$$

Thus, the (i,j,k)th residual is $Y_{ijk} - \bar{Y}_{ij\cdot}$.

The residual vector, however, is 0 when $c = 1$.

Correlation between the LSE and the residual vector

The LSE and the residual vector is always uncorrelated, because

$$\begin{aligned}
\mathrm{Cov}(\widehat{\beta}, Y - X\widehat{\beta}) &= \mathrm{Cov}\left((X'X)^{-1}X'Y, [I_n - X(X'X)^{-1}X']Y\right) \\
&= (X'X)^{-1}X'\mathrm{Var}(Y)[I_n - X(X'X)^{-1}X'] \\
&= \sigma^2 (X'X)^{-1}X'[I_n - X(X'X)^{-1}X'] \\
&= \sigma^2 [(X'X)^{-1}X' - (X'X)^{-1}X'X(X'X)^{-1}X'] \\
&= 0
\end{aligned}$$

We now consider distributions of the LSE and SSR under normality.

The result for the two-way ANOVA can be seen as follows.

$$E(Y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

Hence, the LSE of $E(Y_{ijk})$, which is a linear function of regression parameters, is

$$\widehat{\mu} + \widehat{\alpha}_i + \widehat{\beta}_j + \widehat{\gamma}_{ij} = \bar{Y}_{ij\cdot}$$

Thus, the (i,j,k)th residual is $Y_{ijk} - \bar{Y}_{ij\cdot}$.

The residual vector, however, is 0 when $c = 1$.

Correlation between the LSE and the residual vector

The LSE and the residual vector is always uncorrelated, because

$$\begin{aligned}
\mathrm{Cov}(\widehat{\beta}, Y - X\widehat{\beta}) &= \mathrm{Cov}\left((X'X)^{-1}X'Y, [I_n - X(X'X)^{-1}X']Y\right) \\
&= (X'X)^{-1}X'\mathrm{Var}(Y)[I_n - X(X'X)^{-1}X'] \\
&= \sigma^2(X'X)^{-1}X'[I_n - X(X'X)^{-1}X'] \\
&= \sigma^2[(X'X)^{-1}X' - (X'X)^{-1}X'X(X'X)^{-1}X'] \\
&= 0
\end{aligned}$$

We now consider distributions of the LSE and SSR under normality.

Theorem.

Consider the general linear model $Y = X\beta + \mathscr{E}$ with assumption A1, i.e., $\mathscr{E} \sim N(0, \sigma^2 I_n)$, where $\sigma^2 > 0$ is unknown.

(i) The LSE $\widehat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$ and $l'\widehat{\beta}$ is the UMVUE of $l'\beta$ for any $l \in \mathscr{R}^p$.

(ii) SSR/σ^2 has the central chi-square distribution with degrees of freedom $n-p$ and the UMVUE of σ^2 is $\mathrm{SSR}/(n-p)$.

(iii) SSR and $\widehat{\beta}$ are independent.

(iv) The MLE of β and σ^2 are respectively $\widehat{\beta}$ and $\widehat{\sigma}^2 = \mathrm{SSR}/n$.

Proof.

The result $\widehat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$ follows from the fact that $\widehat{\beta} = (X'X)^{-1}X'Y$ is a linear function of $Y \sim N(X\beta, \sigma^2 I_n)$ and $E(\widehat{\beta}) = \beta$, $\mathrm{Var}(\widehat{\beta}) = \sigma^2(X'X)^{-1}$.

The joint pdf of Y is

$$\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{\frac{-(Y-X\beta)'(Y-X\beta)}{2\sigma^2}\right\} = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{\frac{-\|Y-X\beta\|^2}{2\sigma^2}\right\}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{\|Y - X\widehat{\beta}\|^2 + \|X\widehat{\beta} - X\beta\|^2}{2\sigma^2} \right\}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{\|Y - X\widehat{\beta}\|^2 + \|X\widehat{\beta}\|^2 + \|X\beta\|^2 - 2\beta' X' X\widehat{\beta}}{2\sigma^2} \right\}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ \frac{\beta' X' Y}{\sigma^2} - \frac{\|Y - X\widehat{\beta}\|^2 + \|X\widehat{\beta}\|^2}{2\sigma^2} - \frac{\|X\beta\|^2}{2\sigma^2} \right\}$$

This pdf is from an exponential family with $(X'Y, \|Y - X\widehat{\beta}\|^2 + \|X\widehat{\beta}\|^2)$ as a complete and sufficient statistic for (β, σ^2).

Since $X\widehat{\beta} = X(X'X)^{-1}X'Y$ is a function of $X'Y$ with X considered as fixed, $(X'Y, \|Y - X\widehat{\beta}\|^2)$ is complete and sufficient.

Since $l'\widehat{\beta}$ is unbiased for $l'\beta$ and $\widehat{\beta}$ is a function of a complete and sufficient statistic, $l'\widehat{\beta}$ is the UMVUE of $l'\beta$.

This completes the proof of (i).

We have already shown that $\mathrm{SSR}/(n-p)$ is unbiased for σ^2.

Since $\mathrm{SSR} = \|Y - X\widehat{\beta}\|^2$ is a function of a complete and sufficient statistic, it is the UMVUE of σ^2.

From

$$Y'Y = Y'[X(X'X)^{-1}X']Y + Y'[I_n - X(X'X)^{-1}X']Y$$

and the fact that the rank of $X(X'X)^{-1}X'$ is p and the rank of $I_n - X(X'X)^{-1}X'$ is $n-p$, by Cochran's theorem, SSR/σ^2 has the chi-square distribution with degrees of freedom $n-p$ and noncentrality parameter

$$\sigma^{-2}\beta'X'[I_n - X(X'X)^{-1}X']X\beta = \sigma^2\beta'(X'X - X'X) = 0$$

This proves (ii).

Previously we showed that the LSE $\widehat{\beta}$ and the residual vector $Y - X\widehat{\beta}$ are uncorrelated.

Since both of them are linear functions of Y, they are independent and, thus, their functions $\widehat{\beta}$ and $\mathrm{SSR} = \|Y - X\widehat{\beta}\|^2$ are independent.

The proof of (iii) is completed.

The log likelihood function is

$$\ell(\beta,\sigma^2) = -\frac{\|Y - X\widehat{\beta}\|^2 + \|X\widehat{\beta} - X\beta\|^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2)$$

It is clear that $\widehat{\beta}$ maximizes this function over $\beta \in \mathscr{R}^p$.

To maximize $\ell(\widehat{\beta},\sigma^2)$ over $\sigma^2 > 0$, we obtain that the MLE of σ^2 is $\widehat{\sigma}^2 = \|Y - X\widehat{\beta}\|^2/n = \mathrm{SSR}/n$.

This finishes the proof of (iv).

Fisher information matrix

To derive the Fisher information matrix about (β,σ^2), we differentiate the log likelihood:

$$\ell(\beta,\sigma^2) = -\frac{\|Y - X\beta\|^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2), \qquad \frac{\partial \ell(\beta,\sigma^2)}{\partial \beta} = \frac{X'(Y - X\beta)}{\sigma^2}$$

$$\frac{\partial \ell(\beta,\sigma^2)}{\partial \sigma^2} = \frac{\|Y - X\beta\|^2}{2\sigma^4} - \frac{n}{2\sigma^2}, \qquad \frac{\partial^2 \ell(\beta,\sigma^2)}{\partial \beta \partial \beta'} = -\frac{X'X}{\sigma^2}$$

The log likelihood function is

$$\ell(\beta,\sigma^2) = -\frac{\|Y - X\widehat{\beta}\|^2 + \|X\widehat{\beta} - X\beta\|^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2)$$

It is clear that $\widehat{\beta}$ maximizes this function over $\beta \in \mathscr{R}^p$.

To maximize $\ell(\widehat{\beta},\sigma^2)$ over $\sigma^2 > 0$, we obtain that the MLE of σ^2 is $\widehat{\sigma}^2 = \|Y - X\widehat{\beta}\|^2/n = \mathrm{SSR}/n$.

This finishes the proof of (iv).

Fisher information matrix

To derive the Fisher information matrix about (β,σ^2), we differentiate the log likelihood:

$$\ell(\beta,\sigma^2) = -\frac{\|Y - X\beta\|^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2), \qquad \frac{\partial\ell(\beta,\sigma^2)}{\partial\beta} = \frac{X'(Y - X\beta)}{\sigma^2}$$

$$\frac{\partial\ell(\beta,\sigma^2)}{\partial\sigma^2} = \frac{\|Y - X\beta\|^2}{2\sigma^4} - \frac{n}{2\sigma^2}, \qquad \frac{\partial^2\ell(\beta,\sigma^2)}{\partial\beta\partial\beta'} = -\frac{X'X}{\sigma^2}$$

$$\frac{\partial^2 \ell(\beta,\sigma^2)}{\partial \sigma^4} = -\frac{\|Y - X\beta\|^2}{\sigma^6} + \frac{n}{2\sigma^4}, \qquad \frac{\partial^2 \ell(\beta,\sigma^2)}{\partial \beta \partial \sigma^2} = -\frac{X'(Y - X\beta)}{\sigma^4}$$

$$E\left[\frac{\partial^2 \ell(\beta,\sigma^2)}{\partial \beta \partial \beta'}\right] = -\frac{X'X}{\sigma^2}$$

$$E\left[\frac{\partial^2 \ell(\beta,\sigma^2)}{\partial \sigma^4}\right] = -\frac{E\|Y - X\beta\|^2}{\sigma^6} + \frac{n}{2\sigma^4}$$

$$= -\frac{\text{trace}[\mathrm{Var}(Y)]}{\sigma^6} + \frac{n}{2\sigma^4} = -\frac{n}{2\sigma^4}$$

$$E\left[\frac{\partial^2 \ell(\beta,\sigma^2)}{\partial \beta \partial \sigma^2}\right] = -\frac{X'E(Y - X\beta)}{\sigma^4} = 0$$

Thus, the Fisher information matrix is

$$\frac{1}{\sigma^2}\begin{pmatrix} X'X & 0 \\ 0 & \frac{n}{2\sigma^2} \end{pmatrix}$$

The UMVUE $l'\widehat{\beta}$ attains the information lower bound, whereas the UMVUE of σ^2 does not attain the information lower bound, since the variance of $\mathrm{SSR}/(n-p)$ is $2\sigma^4/(n-p)$.