2.3 Properties of the Least-Square Estimators

Recall that in MLR the LS estimator $\hat{\beta}$ is given by

$$\hat{\beta} \ = \ \left(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p\right)^T == (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$$

Based on the model assumptions, *conditionally on* $\boldsymbol{X}$, the vector $\hat{\beta}$ is a random vector, since it is a function of $\mathbf{y}$ (which is random). To design hypothesis tests for the model parameters, we need to understand distribution of $\hat{\beta}$.

2.3.1 Mean & Covariance of $\hat{\boldsymbol{\beta}}$

Recall that our model assumptions are

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon,$$

with $\mathrm{E}(\varepsilon) = \mathbf{0}$, and $Cov(\varepsilon) = \sigma^2\mathbf{I}_n$.

These assumptions imply that the response $\mathbf{y}$ has mean and variance equal to:

$$\mathrm{E}(\mathbf{y}) = \mathbf{X}\beta, \quad Cov(\mathbf{y}) = \sigma^2\mathbf{I}_n$$

Proposition

The LS estimators $\hat{\beta}$ are ***unbiased***.

Proof

Indeed,

$$\begin{aligned}
\mathrm{E}(\hat{\beta}) &= \mathrm{E}\left((\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y}\right) = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\,\mathrm{E}(\mathbf{y}) \\
&= \underbrace{(\mathbf{X}}_{=\mathbf{I}}{}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{X}\,\beta = \beta
\end{aligned}$$

■

Proposition

The Variance-Covariance matrix of $\hat{\beta}$ is equal to

$$Cov(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

Proof

We can directly compute the covariance matrix of $\hat{\beta}$ as follows:

$$\begin{aligned}
Cov(\hat{\beta}) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Cov(\mathbf{y})\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)^T \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,\sigma^2\,\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2\underbrace{(\mathbf{X}}_{=\mathbf{I}}{}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}
\end{aligned}$$

■

2.3.2 Properties of $\hat{y}$ and r

Using the previous results, we can also show the following properties for the fitted values $\hat{y}$ and the residuals r:

1. $\mathrm{E}(\hat{\mathbf{y}}) = \mathbf{X}\beta$
2. $Cov(\hat{\mathbf{y}}) = \sigma^2\mathbf{H}$

3. $\mathrm{E}(\mathbf{r}) = \mathbf{0}$

4. $Cov(\mathbf{r}) = \sigma^2(\mathbf{I_n} - \mathbf{H})$

5. $\mathrm{E}(\hat{\sigma}^2) = \frac{1}{n-p}\,\mathrm{E}(\mathbf{r^T r}) = \frac{1}{n-p}\sigma^2(n-p) = \sigma^2$

Proposition If we assume that the error terms are normally distributed, then we have that

$$\frac{\mathbf{r^T r}}{\sigma^2} = \frac{RSS}{\sigma^2} \sim \chi^2_{n-p}$$

The (i, j) element of covariance matrix we computed for the vector of β coefficients corresponds to the covariance term $Cov(x_i, x_j)$. So, if we want to extract the variance of β_i, then this will be the term (i, i) element of the matrix, i.e. $((\mathbf{X^T X})^{-1})_{ii}$.

Note that $\hat{\beta}$ and $\hat{\sigma}^2$ are *unbiased* estimators of β and σ^2 respectively, so we can plug-in the variance estimator $\hat{\sigma}^2$ to get an estimator for the covariance of $\hat{\beta}$.

Standard Error of $\hat{\boldsymbol{\beta}}_1$ The standard errors of the $\hat{\beta}_i$ are the square roots of the elements of the diagonal of the covariance matrix $Cov(\hat{\beta}) = \hat{\sigma}^2(\mathbf{X^T X})^{-1}$, namely

$$se(\hat{\beta}_i) = \hat{\sigma}\sqrt{((\mathbf{X^T X})^{-1})_{ii}}$$

2.3.3 The Gauss Markov Theorem

The main reason why we use LS estimation is because of the Gauss-Markov theorem. If the errors are uncorrelated, have equal variance and mean equal to zero, the LS estimators have **the lowest variance within the class of linear estimators**.

Let's consider a more general case. Suppose we are interested in estimating a linear combination of β of the form:

$$\theta = \mathbf{c}^T\beta = \sum_{j=1}^{p} c_j\beta_j,$$

where c_j are real numbers.For example, estimating any element of β or estimating the mean response at a new value x^* are all *special cases of this setup*.

Naturally, we obtain an estimate of θ by plugging in the LS estimate β in the equation for θ, i.e.

$$\hat{\theta}_{LS} = \mathbf{c}^T\hat{\beta} = \mathbf{c^T(X^TX)^{-1}X^Ty}$$

This is still a linear[9] and unbiased estimator of θ with a mean square error that computes as

$$MSE(\hat{\theta}_{LS}) = \mathrm{E}(\hat{\theta}_{LS} - \theta)^2 = Var(\hat{\theta}_{LS})$$

Now, assume that there is another estimate of θ, which is also linear and unbiased. The following *Theorem* states that $\hat{\theta}_{LS}$ is **always** better in the sense that its MSE is always smaller (or at least, not bigger).

The Gauss-Markov Theorem

Let $\hat{\theta}$ be the least-squares estimate of $\theta = \mathbf{X}\beta$, where $\theta \in \Omega = \mathrm{C}(\mathbf{X})$ and $\mathbf{X}$ may not have full rank. Then among the class of unbiased estimates of $\mathbf{c}^T\theta$, $\mathbf{c}^T\hat{\theta}$ is the unique estimate with minimum variance. We say that $\mathbf{c}^T\hat{\theta}$ is the **best linear unbiased estimate (BLUE)** of $\mathbf{c}^T\theta$.

Proof

We know that $\hat{\theta} = \mathbf{X}\hat{\beta} = \mathbf{H}Y$, where $\mathbf{H}\theta = \mathbf{H}\mathbf{X}\beta = \mathbf{X}\beta = \theta$. Hence, $\mathrm{E}(\mathbf{c}^T\hat{\theta}) = \mathbf{c}^T\mathbf{H}\theta = \mathbf{c}^T\theta$, for all $\theta \in \Omega$, which means that $\mathbf{c}^T\hat{\theta}$ is an unbiased estimator of $\mathbf{c}^T\theta$.

Then, $\mathbf{c}^T\theta = \mathrm{E}(\mathbf{d}^T\mathbf{Y}) = \mathbf{d}^T\theta$ or $(\mathbf{c}-\mathbf{d})^T\theta = 0$, so that $(\mathbf{c}-\mathbf{d})$ is orthogonal to Ω. Therefore, $\mathbf{H}(\mathbf{c}-\mathbf{d}) = 0$ and $\mathbf{H}c = \mathbf{H}d$.

Now,

$$
\begin{aligned}
Var(\mathbf{c}^T\hat{\theta}) &= Var\Big((\mathbf{Hc})^T\mathbf{Y}\Big) \\
&= Var\Big((\mathbf{Hd})^T\mathbf{Y}\Big) \\
&= \sigma^2\mathbf{d}^T\mathbf{H}^T\mathbf{Hd} \\
&= \sigma^2\mathbf{d}^T\mathbf{H}^2\mathbf{d} \\
&= \sigma^2\mathbf{d}^T\mathbf{Hd}
\end{aligned}
$$

so that

$$
\begin{aligned}
Var(\mathbf{d}^T\mathbf{Y}) - Var(\mathbf{c}^T\hat{\theta}) &= Var(\mathbf{d}^T\mathbf{Y}) - Var((\mathbf{Hd})^T\mathbf{Y}) \\
&= \sigma^2(\mathbf{d}^T\mathbf{d} - \mathbf{d}^T\mathbf{Hd}) \\
&= \sigma^2\mathbf{d}^T(\mathbf{I}_n - \mathbf{H})\mathbf{d} \\
&= \sigma^2\mathbf{d}^T(\mathbf{I}_n - \mathbf{H})^T\underbrace{(\mathbf{I}_n - \mathbf{H})\mathbf{d}}_{:=\mathbf{d}_1} \\
&= \sigma^2\mathbf{d}_1^T\mathbf{d}_1 \geq 0
\end{aligned}
$$

with equality only if $(\mathbf{I}_n - \mathbf{H})\mathbf{d} = 0$ or $\mathbf{d} = \mathbf{Hd} = \mathbf{Hc}$. Hence, $\mathbf{c}^T\hat{\theta}$ has minimum variance and is unique.

$\blacksquare$

Corollary

If $\mathbf{X}$ has full rank, then $\mathbf{a}^T\hat{\beta}$ is the BLUE of $\mathbf{a}^T\beta$ for every vector $\mathbf{a}$.

Proof

Now $\theta = \mathbf{X}\beta$ implies that $\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\theta$ and $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\hat{\theta}$. Hence, setting $\mathbf{c}^T = \mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ we have that $\mathbf{a}\hat{\beta}(= \mathbf{c}^T\hat{\theta})$ is the BLUE of $\mathbf{a}\beta(= \mathbf{c}^T\theta)$ for every vector $\mathbf{a}$.

$\blacksquare$

Theorem (Unbiased Estimator of σ^2)

If $\mathrm{E}(\mathbf{Y}) = \mathbf{X}\beta$, where $\mathbf{X}$ is an $n \times p$ matrix of rank r ($r \leq p$), and $Var(\mathbf{Y}) = \sigma^2 \mathbf{I}_n$, then

$$\hat{\sigma}^2 = \frac{(\mathbf{Y} - \hat{\theta})^T(\mathbf{Y} - \hat{\theta})}{n - r} = \frac{RSS}{n - r}$$

is an unbiased estimate of σ^2.

Proof

Consider the full-rank representation $\theta = \mathbf{X}_1\alpha$, where $\mathbf{X}_1$ is $n \times r$ of rank r. Then,

$$\mathbf{Y} - \hat{\theta} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y},$$

where $\mathbf{H} = \mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T$. Using the properties of the Hat matrix we have the following:

$$\begin{aligned}(n - r)\hat{\sigma}^2 &= \mathbf{Y}^T(\mathbf{I}_n - \mathbf{H})^T(\mathbf{I}_n - \mathbf{H})\mathbf{Y} \\ &= \mathbf{Y}^T(\mathbf{I}_n - \mathbf{H})^2\mathbf{Y} \\ &= \mathbf{Y}^T(\mathbf{I}_n - \mathbf{H})\mathbf{Y}\end{aligned}$$

Since $\mathbf{H}\theta = \theta$, we have

$$\mathrm{E}(\mathbf{Y}^T(\mathbf{I}_n - \mathbf{H})\mathbf{Y}) = \sigma^2 tr(\mathbf{I}_n - \mathbf{H}) + \theta^T(\mathbf{I}_n - \mathbf{H})\theta = \sigma^2(n - r)$$

and hence $\mathrm{E}(\hat{\sigma}^2) = \sigma^2$.

■

2.3.4 Maximum Likelihood Estimation

In this section we derive the Maximum Likelihood estimators for the regression model parameters, namely β and σ. In order to write the likelihood, we need to assume a distribution for the error terms and as a result the responses. So, we assume that

$$\mathbf{y} \sim MVN(\mathbf{0}, \sigma^2\mathbf{I}),$$

where *MVN* stand for Multivariate Normal distribution.

Assuming normality, the likelihood function $L(\beta, \sigma^2)$ for the *full rank regression model* is the probability density of $\mathbf{Y}$, namely

$$L(\beta, \sigma^2) = (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{X}\beta||^2\right\}$$

Taking the logarithm of the likelihood, we have (ignoring constants)

$$\ell(\beta, \sigma^2) = \log L(\beta, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}||\mathbf{y} - \mathbf{X}\beta||^2$$

To compute the Maximum Likelihood estimators, we take derivatives with respect to β and σ^2 as follows:

$$\begin{aligned}\frac{\partial\ell}{\partial\beta} &= -\frac{1}{2\sigma^2}(-2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\beta)\\ \frac{\partial\ell}{\partial\sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}||\mathbf{y} - \mathbf{X}\beta||^2\end{aligned}$$

Setting

$$\frac{\partial\ell}{\partial\beta} = 0,$$

we get the estimator of β

$$\hat{\beta}_{ML} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

which is the same as the Least Squares estimator $\hat{\beta}_{LS}$.

$\hat{\beta}_{ML}$ clearly maximizes $\ell(\beta, \sigma^2)$ for any $\sigma^2 > 0$. Hence,

$$L(\beta, \sigma^2) \leq L(\hat{\beta}_{ML}, \sigma^2)$$

for all $\sigma^2 > 0$ with equality **if and only if** $\beta = \hat{\beta}$.

We now wish to maximize $L(\hat{\beta}, \sigma^2)$, or equivalently $\ell(\hat{\beta}, \sigma^2)$ with respect to σ^2.

Setting

$$\frac{\partial \ell}{\partial \sigma^2} = 0,$$

we get a *stationary* value of

$$\hat{\sigma}^2_{ML} = \frac{||\mathbf{y} - \mathbf{X}\beta||^2}{n}.$$

Then,

$$\ell(\hat{\beta}_{ML}, \hat{\sigma}^2_{ML}) - \ell(\hat{\beta}_{ML}, \sigma^2) = -\frac{n}{2}\left(\log\left(\frac{\hat{\sigma}^2_{ML}}{\sigma^2}\right) + 1 - \frac{\hat{\sigma}^2_{ML}}{\sigma^2}\right) \geq 0$$

since $x \leq e^{x-1}$ and therefore $\log x \leq x - 1$ for $x \geq 0$ (with equality when $x = 1$).

This implies that

$$L(\beta, \sigma^2) \leq L(\hat{\beta}_{ML}, \hat{\sigma}^2_{ML}), \text{ for all } \sigma^2 > 0$$

with equality **if and only if** $\beta = \hat{\beta}_{ML}$ and $\sigma^2 = \hat{\sigma}^2_{ML}$.

Thus, $\hat{\beta}_{ML}$ and $\hat{\sigma^2}_{ML}$ are the **maximum likelihood estimators** of β and σ^2 and the maximum value of the likelihood is computed as

$$L(\hat{\beta}_{ML}, \hat{\sigma}^2_{ML}) = (2\pi\hat{\sigma}^2_{ML})^{-n/2}e^{-n/2}.$$

2.3.5 Distribution of the Least-Squares estimates

Recall the assumption for the Normal Linear regression model:

$$\mathbf{y} \sim \mathbf{N_n}(\mathbf{X}\beta, \sigma^2\mathbf{I})$$

Any affine transformation of $\mathbf{y}$ will also have a Normal distribution. In fact, we can show that the elements of $\mathbf{y}$ are **jointly** Normal. Therefore, always conditional on $\mathbf{X}$, we can show that

$$\hat{\beta} = (\mathbf{X^TX})^{-1}\mathbf{X^TY} \sim \mathbf{N_p}(\beta, \sigma^2(\mathbf{X^TX})^{-1})$$

$$\hat{\mathbf{y}} = \mathbf{Hy} \sim \mathbf{N_n}(\mathbf{X}\beta, \sigma^2\mathbf{H})$$

$$\hat{\mathbf{r}} = (\mathbf{I_n} - \mathbf{H})\mathbf{y} \sim \mathbf{N_n}(\mathbf{0}, \sigma^2(\mathbf{I_n} - \mathbf{H}))$$

Indeed, for the **fitted values** $\hat{\mathbf{y}}$ and the estimated **residuals** $\mathbf{r}$ we can calculate the mean and covariance matrices as follows:

$$\begin{aligned} \mathrm{E}(\hat{\mathbf{y}}) &= \mathbf{H}\ \mathrm{E}(\mathbf{y}) = \mathbf{HX}\beta = \mathbf{X}\beta \\ Cov(\hat{\mathbf{y}}) &= \mathbf{H}\sigma^2\mathbf{H}^T = \sigma^2\mathbf{H} \\ \mathrm{E}(\mathbf{r}) &= (\mathbf{I_n} - \mathbf{H})\mathbf{X}\beta = \mathbf{0} \\ Cov(\mathbf{r}) &= (\mathbf{I_n} - \mathbf{H})\sigma^2(\mathbf{I_n} - \mathbf{H})^T = \sigma^2(\mathbf{I_n} - \mathbf{H}) \end{aligned}$$

Residuals' Properties

Although $\mathbf{r}$ is a vector of dimension n, it always lies in a subspace of dimension $(n-p)$ (the error space). In fact, $\mathbf{r}$ behaves like a random vector with a distribution

$$\mathbf{r} \sim \mathbf{N}_{n-p}(\mathbf{0}, \sigma^2\mathbf{I}_{n-p})$$

Therefore, it can be shown that

Proposition

$$\hat{\sigma}^2 = \frac{||\mathbf{r}||^2}{n-p} \sim \sigma^2\frac{\chi^2_{n-p}}{n-p}$$

In addition, $\hat{\mathbf{y}}$ and $\mathbf{r}$ are uncorrelated since they are in orthogonal spaces. Since they also have a joint normal distribution, they are independent[10].