R Lab : Estimating σ^2

Valay Shah BS2026

31 August 2022

We have seen that in a linear model σ^2 (which is the variance of the residual error ϵ), can be estimated by the following formula-

$$\hat{\sigma}^2 = \frac{||\vec{Y} - X\hat{\beta}||^2}{n - r(X)}$$

We estimate this in R using the following code -

```
agri<-data.frame(
variety=c(1,2,2,3,3,2,1,1),yield=c(210.3,245.0,248.9,212.3,2
stringsAsFactors=FALSE)
agri
names(agri)
```

```
A data.frame: 8 × 2
variety yield
<dbl><dbl>
1 210.3
2 245.0
2 248.9
3 212.3
3 230.4
```

```
2 250.1
1 213.5
1 212.4
'variety''yield'

agri$variety = factor(agri$variety)
fit=lm(yield~variety-1,agri)
fit
model.matrix(fit)
summary(fit)


  (Intercept) variety2 variety3
1           1        0        0
2           1        1        0
3           1        1        0
4           1        0        1
5           1        0        1
6           1        1        0
7           1        0        0
8           1        0        0


Call:
lm(formula = yield ~ variety, data = agri)

Residuals:
      1        2        3        4        5        6        7
-1.7667  -3.0000   0.9000  -9.0500   9.0500   2.1000   1.4333   0.3
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  212.067      3.496  60.663  2.3e-08 ***
variety2      35.933      4.944   7.268 0.000771 ***
variety3       9.283      5.527   1.680 0.153887
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 6.055 on 5 degrees of freedom
Multiple R-squared:  0.9179,Adjusted R-squared:  0.8851
F-statistic: 27.96 on 2 and 5 DF,  p-value: 0.00193
```

As seen from the output the Residual standard error: 6.055 gives us the estimate of σ^2 and the degrees of freedom $= 5$ denote the value of $n - r(X)$.

Note that the command -

```
    agri$variety = factor(agri$variety)}
```

is very important. Otherwise which would treat the variable Variety as integer type.

Factors are the data objects which are used to categorize the data and store it as levels, which are in-turn used as separate categories to fit the model.