제 7 장 Weighted Least Squares

▷ Underlying regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots \beta_p x_{ip} + \epsilon_i, \quad \epsilon_i \sim \ iid$$

▷ Ch 7 and Ch 8 investigate non-iid case.

Ch 7 deals with the heteroscedasticity,

Ch 8 treats the autocorrelation problem.

▷ Equal variance assumtion is relaxed : ϵ_i : independent with variance σ_i^2

Ch 6 : transformation and OLS cf

Ch 7 Use wieghted least squares(WLS)

Minimize : $\sum_i w_i (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))^2, \quad w_i \propto \dfrac{1}{\sigma_i^2}$

w_i : *weights inversely proportional to the variances of the errors*

Any observation with a small weight will be
severely discounted by WLS
eg. If $w_i = 0$ the effect of WLS is to exclude the i -th observation.

Note : If the weights are unknown : two stage procedure

step 1 : Use OLS to estimate w_i

step 2 : WLS with w_i

◆ Heteroscedastic models

There are three types :

type 1 : *we can expect the variance structure from information in the raw data.*

type 2 : *observations are average of individual sampling units taken over well-defined groups(clusters)*

type 3 : *structure of heteroscedasticity is determined empirically*

▷ Type I : We can expect the variance structure from information in the raw data.

◎ **Example** : Supervisor data (Industrial data)

data : Table 6.9 on p.176,

X : no. of supervised workers

Y : no. of supervisors

model : $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \sigma_i^2 = k^2 x_i^2, \qquad k > 0$

minimize : $\sum w_i (y_i - (\beta_0 + \beta_1 x_i))^2, \qquad w_i = \dfrac{1}{x_i^2}$

equivalently, use OLS with model :

$$\frac{y_i}{x_i} = \frac{\beta_0}{x_i} + \beta_1 + \frac{\epsilon_i}{x_i},$$

This approach may be considered in multiple regression.

eg $\qquad y_i = \beta_0 + \beta_1 x_{i1} + \cdots \beta_p x_{ip} + \epsilon_i, \qquad$ If $\sigma_i^2 = k^2 x_{i2}^2$

WLS : with $w_i = \dfrac{1}{x_{i2}^2}$

or

OLS : $\qquad \dfrac{y_i}{x_{i2}} = \beta_0 \dfrac{1}{x_{i2}} + \beta_1 \dfrac{x_{i1}}{x_{i2}} + \beta_2 + \beta_3 \dfrac{x_{i3}}{x_{i2}} + \cdots + \beta_p \dfrac{x_{ip}}{x_{i2}} + \dfrac{\epsilon_i}{x_{i2}}$

▷ Type 2 : occurs in large scale surveys ;

Observations are average of individual sampling units taken over well-defined groups(clusters)

◎ **Example** : College expense data variable in Table 7.1

A set of schools are selected and interview a prescribed number of randomly selected students at each school.

$$Var(\overline{y_i}) = \frac{\sigma^2}{n_i}, \qquad n_i \text{ : \# of obs at } i\text{-th institution}$$

Minimize $\sum n_i \left(y_i - \left(\beta_0 + \sum_j \beta_j x_{ij} \right) \right)^2$

⇒ WLS with $w_i = n_i$

or

OLS $\sqrt{n_i} y_i = \beta_0 \sqrt{n_i} + \beta_1 x_{i1} \sqrt{n_i} + \cdots \beta_6 x_{i6} \sqrt{n_i} + \epsilon_{isqrt} n_i$

; 7 predictors with no intercept

▷ Type 3 : Structure of heteroscedasticity is determined empirically.
there is no prior indication that the variances are not equal.
residual plots can serve as a first step
to detect heteroscedasticity.

$\Rightarrow$ Two stage estimation

If there are replicated measurements on y.
eg Fig 7.2

$$y_{ij} = \beta_0 + \beta_1 x_j + \epsilon_{ij}, \qquad i = 1, \ldots, n_j, \qquad Var(\epsilon_{ij}) = \sigma_j^2$$

$$\Rightarrow \quad \widehat{y_{ij}} = \widehat{y_j}$$

$$e_{ij} = (y_{ij} - \widehat{y_{ij}}) = (y_{ij} - \overline{y_j}) + (\overline{y_j} - \widehat{y_j})$$

pure error lack of fit error

$$s_j^2 = \frac{1}{n_j - 1} \sum (y_{ij} - \overline{y_j})^2,$$

$$\Rightarrow \quad w_i = \frac{1}{s_j^2}$$

But the presence of replications on the response variable for a given value of X is rather uncommon when data are collected in a non-experimental setting,
$\Rightarrow$ A more plausible way to investigate heteroscedasticity in (multiple) regression is by clustering obs according to prior, natural, and meaningful associations.

◎ **Example** : Education expenditure data in sec 5.7

use only 1975 data (p.198), Variable list : Table 7.2

Objective : to get the best representation of the relationship between expenditure on education and the other variables.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

Grouping geographically : 4 groups $j = 1, 2, 3, 4$

j-th group error variance : $c_j^2 \sigma^2$

Minimize $S_w = S_1 + S_2 + S_3 + S_4$

$$S_j = \sum_{i \in I_j} \frac{1}{c_j^2} (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}))^2, \qquad j = 1, 2, 3, 4$$

I_j contains index of j-th group

c_j : unknown

⇒ ① use OLS :

result Table 7.4 on p. 200

residual plot Fig7.3-7.7 : unequal variance

#49(Alaska) is influential, high leverage, outlier

Alaska is a state with very small population and a boom in revenue from oil...

⇒ better omit the #49 and do separate analysis...

Table 7.5 shows : the values of coefs changed significantly...

residual plot(Fig 7.8-7.9) still shows unequal variance

② use WLS with $\hat{c}_j^2 = \dfrac{\dfrac{1}{n_j - 1} \sum_{i \in I_j} e_i^2}{\dfrac{1}{n} \sum_i^n e_i^2}$, Table 7.6

result Table 7.7 : $R^2_{OLS} > R^2_{WLS}$ due to the definition of OLS

residual plot (original y scale에 대한 그림임)

Note : region에 대하여 group을 나누고 WLS를 시행함으로 region에 대한 unequal variance는 상당부분 해소됨을 볼 수 있다.

$R^2 = 0.477$ ⇒ the search for other factors must continue

▷ Binary reponse data (Logistic model)

eg ① x : different dose of a drug or poison

y : death or survival

② x : discount offered

y : purchase or non purchase

$Var(Y) = p(1-p)$ depends on $p = P(Y=1) = E(Y) = \beta_0 + \beta_1 X$

⇒ Transformations to stabilize variance (*and normalizing*) : See Table 6.5 on p. 161

Ch 12 : more in details