Regression Analysis
Chapter 7: Variance

Shaobo Jin

Department of Mathematics

Different Variance

- We have always assumed $\mathrm{Var}\left(y \mid \boldsymbol{x}\right) = \sigma^2$ so far. In practice, the variance does not have to be the same for all values of $\boldsymbol{x}$.
- Consider $\mathrm{E}\left(Y \mid \boldsymbol{x}\right) = \boldsymbol{x}^T \boldsymbol{\beta}$ and $\mathrm{Var}\left(y \mid \boldsymbol{x}\right) = \sigma^2 / w_i$, where w_1, ..., w_n are known positive numbers.
- The OLS estimator is no longer optimal, since it weights all observations equally.

Weighted Least Squares

The weighted least squares (WLS) estimator minimizes

$$\begin{aligned}\mathrm{RSS}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} w_i \left(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}\right)^2 \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}),\end{aligned}$$

where $\boldsymbol{W}$ is a diagonal matrix with diagonal entries $\{w_i\}$.

The WLS estimator is

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{y}.$$

Property of WLS

1. Under the assumption that $\mathrm{E}\left(\boldsymbol{y} \mid \boldsymbol{X}\right) = \boldsymbol{X}\boldsymbol{\beta}$ is correctly specified,, the WLS estimator is unbiased

$$\mathrm{E}\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \boldsymbol{\beta}.$$

2. Under an extra assumption $\mathrm{Var}\left(\boldsymbol{y} \mid \boldsymbol{X}\right) = \sigma^2 \boldsymbol{W}^{-1}$, the covariance matrix is

$$\begin{aligned} \mathrm{Var}\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) &= \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W} \mathrm{Var}\left(\boldsymbol{y} \mid \boldsymbol{X}\right) \boldsymbol{W} \boldsymbol{X} \\ &= \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1}. \end{aligned}$$

Compared to OLS

1. The WLS estimator
$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y}$$
is an unbiased linear estimator.
2. Consider another unbiased linear estimator
$$\tilde{\boldsymbol{\beta}} = \left[\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W} + \boldsymbol{K}\right]\boldsymbol{y}$$
for some matrix $\boldsymbol{K}$. Since $\tilde{\boldsymbol{\beta}}$ is unbiased, we must have $\boldsymbol{K}\boldsymbol{X}\boldsymbol{\beta} = \mathbf{0}$ for any $\boldsymbol{\beta}$.We can show that
$$\mathrm{Var}\left(\tilde{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) - \mathrm{Var}\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \sigma^2\boldsymbol{K}\boldsymbol{W}^{-1}\boldsymbol{K}^T \geq 0.$$
Hence, the WLS estimator is more efficient than the OLS estimator.
3. The Gauss-Markov Theorem applies to the WLS estimator.

Residuals

- The residual is

$$\hat{e}_i = y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$$

or

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \left[\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\right]\boldsymbol{y}.$$

- The RSS $(\boldsymbol{\beta})$ evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ is

$$\begin{aligned} \text{RSS}\left(\hat{\boldsymbol{\beta}}\right) &= \sum_{i=1}^{n} w_i \left(y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}\right)^2 \\ &= \boldsymbol{y}^T\left[\boldsymbol{W} - \boldsymbol{W}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\right]\boldsymbol{y}. \end{aligned}$$

Hat Matrix

- The residual is

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \left[\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\right]\boldsymbol{y}.$$

  Hence, the hat matrix can be

$$\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}.$$

- The RSS $(\boldsymbol{\beta})$ evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ is

$$\text{RSS}\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{y}^T\left[\boldsymbol{W} - \boldsymbol{W}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}\right]\boldsymbol{y}.$$

  Hence, the hat matrix can be

$$\boldsymbol{H} = \boldsymbol{W}^{1/2}\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}^{1/2}.$$

They have the same diagonal values.

Estimating σ^2

The estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}\left(\hat{\boldsymbol{\beta}}\right)}{n-p},$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector. We can show that

$$\mathrm{E}\left(\hat{\sigma}^2\right) = \sigma^2,$$

under the assumptions that

C1 $\mathrm{E}(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$ is correctly specified,

C2 $\mathrm{E}(\boldsymbol{e} \mid \boldsymbol{X} = \boldsymbol{x}) = \mathbf{0}$,

C2 $\mathrm{Var}(\boldsymbol{e} \mid \boldsymbol{X} = \boldsymbol{x}) = \sigma^2 \boldsymbol{W}^{-1}$.

Misspecified Variance

Suppose that the true model is

$$\begin{aligned}\mathrm{E}(\boldsymbol{y} \mid \boldsymbol{X}) &= \boldsymbol{X}\boldsymbol{\beta}, \\ \mathrm{Var}(\boldsymbol{y} \mid \boldsymbol{X}) &= \sigma^2 \boldsymbol{W}^{-1},\end{aligned}$$

where $\boldsymbol{W} > 0$ is a diagonal matrix.

But we assume that $\mathrm{Var}(\boldsymbol{y} \mid \boldsymbol{X}) = \sigma^2 \boldsymbol{I}$ and estimate $\boldsymbol{\beta}$ by OLS:

$$\hat{\boldsymbol{\beta}}_{OLS} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y}.$$

1. The OLS estimator is still unbiased.
2. The variance of the OLS estimator is

$$\mathrm{Var}\left(\hat{\boldsymbol{\beta}}_{OLS} \mid \boldsymbol{X}\right) = \sigma^2 \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W}^{-1} \boldsymbol{X} \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}.$$

Sandwich Estimator

Since the variance is misspecified, we cannot estimate $\operatorname{Var}\left(\hat{\boldsymbol{\beta}}_{OLS} \mid \boldsymbol{X}\right)$ by $\hat{\sigma}^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$.

Let $\hat{\boldsymbol{M}}$ be an estimator of $\sigma^2 \boldsymbol{W}^{-1}$. We can estimate $\operatorname{Var}\left(\hat{\boldsymbol{\beta}}_{OLS} \mid \boldsymbol{X}\right)$ by the sandwich estimator

$$\left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \hat{\boldsymbol{M}} \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$$

- One popular estimator $\hat{\boldsymbol{M}}$ is

$$\operatorname{diag}\left\{\frac{\hat{e}_i^2}{\left(1 - h_{ii}\right)^2}\right\},$$

where $h_{ii} = \boldsymbol{x}_i^T \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{x}_i$ is called the leverage.

Generalized Least Squares

The generalized least squares (GLS) assumes that

$$\begin{aligned} \mathrm{E}(\boldsymbol{y} \mid \boldsymbol{X}) &= \boldsymbol{X}\boldsymbol{\beta}, \\ \mathrm{Var}(\boldsymbol{y} \mid \boldsymbol{X}) &= \boldsymbol{\Sigma}, \end{aligned}$$

where $\boldsymbol{\Sigma} > 0$.

If $\boldsymbol{\Sigma}$ is known, the GLS estimator minimizes

$$\mathrm{RSS}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

The GLS estimator is given by

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{y}.$$

Feasible Generalized Least Squares

GLS is infeasible when $\boldsymbol{\Sigma}$ is unknown. The feasible GLS estimator uses an estimator of $\boldsymbol{\Sigma}$ instead, i.e.,

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{y}.$$

For example, suppose that in WLS we do not know $\boldsymbol{W}^{-1}$.

1. We obtain the OLS estimator and its residuals.
2. We estimate $\sigma^2\boldsymbol{W}^{-1}$ by matrix $\boldsymbol{M}$.
3. The feasible GLS estimator is obtained by using $\boldsymbol{M}^{-1}$ as the weights.

Motivation

Suppose that we have fitted the following model

$$\hat{\mathrm{E}}\left(Y \mid \boldsymbol{x}\right) = 324.0 - 2.52x_1 - 4.40x_2 + 0.024x_2^2.$$

1. We want to find out the value of x_2 that minimize $\hat{\mathrm{E}}\left(Y \mid \boldsymbol{x}\right)$. The minimizer is given by $x_2 = -0.5\beta_2/\beta_3$.
2. We also want a confidence interval for x_2 that attains the minimum of $\hat{\mathrm{E}}\left(Y \mid \boldsymbol{x}\right)$.

Delta Method

The delta method can be used if we are interested in the standard errors for a nonlinear function of regression coefficients.

- Let $\boldsymbol{\theta}$ be a $k \times 1$ parameter vector with estimator $\hat{\boldsymbol{\theta}}$ such that
$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \boldsymbol{\Omega})$$
for a positive definite matrix $\boldsymbol{\Omega}$.
- Consider a scalar-valued function $g(\boldsymbol{\theta})$ such that
$$g\left(\hat{\boldsymbol{\theta}}\right) \approx g(\boldsymbol{\theta}) + \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right).$$
- The delta method says that
$$g\left(\hat{\boldsymbol{\theta}}\right) \sim N\left(g(\boldsymbol{\theta}), \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T \boldsymbol{\Omega} \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right).$$

Delta Method: Confidence Interval

An approximate $1-\alpha$ confidence interval for $g(\boldsymbol{\theta})$ can be

$$g\left(\hat{\boldsymbol{\theta}}\right) \pm \lambda_{1-\alpha/2}\sqrt{\left(\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T \boldsymbol{\Omega} \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}}$$

Example

Consider $\boldsymbol{\beta} = \begin{bmatrix}\beta_0 & \beta_1 & \beta_2 & \beta_3\end{bmatrix}^T$. We are interested in $g(\boldsymbol{\beta}) = -0.5\beta_2/\beta_3$. Hence,

$$\frac{\partial g(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \begin{bmatrix} 0 \\ 0 \\ -0.5/\beta_3 \\ 0.5\beta_2/\beta_3^2 \end{bmatrix}.$$

Bootstrap

The bootstrap is a computationally intensive approach for computing standard errors, confidence intervals, hypothesis testing, etc. It can be used if

1. some assumptions are not satisfied,
2. or the closed form expression of a quantity of interest is hard to obtain.

Suppose that we have a sample $(y_i, \boldsymbol{x}_i)$, $i = 1, ..., n$.

Residual Bootstrap

Algorithm 1: Residual Bootstrap

Fit the model $\mathrm{E}\left(Y \mid \boldsymbol{X} = \boldsymbol{x}\right) = \boldsymbol{x}^T \boldsymbol{\beta}$ using data ;
Obtain the residuals $\hat{e}_i$;
Specify an integer B ;
for *each integer b from 1 to B* **do**
    Sample with replacement of size n from the residuals $\hat{e}_i$. Denote the sample by $\hat{e}_i^*$;
    Generate a new bootstrap sample $(y_i^*, \boldsymbol{x}_i)$, where $y_i^* = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} + \hat{e}_i^*$ for all i ;
    Fit the model using the bootstrap sample $(y_i^*, \boldsymbol{x}_i^*)$, $i = 1, ..., n$;
    Obtain the bootstrap estimate $\hat{\boldsymbol{\beta}}_{(b)}^*$;
end
The distribution of $\hat{\boldsymbol{\beta}}$ is approximated by the distribution of $\hat{\boldsymbol{\beta}}_{(b)}^*$, $b = 1, ..., B$

Case Bootstrap

Algorithm 2: Case Bootstrap

```
Specify an integer B ;
for each integer b from 1 to B do
    Sample with replacement of size n from (y_i, x_i) ;
    Fit the model using the bootstrap sample (y_i^*, x_i^*), i = 1, ..., n ;
    Obtain the bootstrap estimate β̂*_(b) ;
end
The distribution of β̂ is approximated by the distribution of β̂*_(b), b = 1, ..., B
```

Specify an integer B ;
for *each integer* b *from 1 to* B **do**
  Sample with replacement of size n from $(y_i, \boldsymbol{x}_i)$;
  Fit the model using the bootstrap sample $(y_i^*, \boldsymbol{x}_i^*)$, $i = 1, ..., n$;
  Obtain the bootstrap estimate $\hat{\boldsymbol{\beta}}^*_{(b)}$;
end
The distribution of $\hat{\boldsymbol{\beta}}$ is approximated by the distribution of $\hat{\boldsymbol{\beta}}^*_{(b)}$, $b = 1, ..., B$