SMIII lectures

Week 5

GLS

Generalised least squares

Have considered until now the linear model

$$\boldsymbol{Y} = X\beta + \boldsymbol{\mathcal{E}}$$

where $\boldsymbol{\mathcal{E}}$ is a random vector with

$$E(\boldsymbol{\mathcal{E}}) = \mathbf{0} \text{ and } \mathrm{Var}(\boldsymbol{\mathcal{E}}) = \sigma^2 I.$$

Now assume instead that

$$\mathrm{Var}(\boldsymbol{\mathcal{E}}) = \sigma^2 V$$

where V is a known $n \times n$ positive definite, symmetric matrix.

Symmetry

Observe that

$$\sigma^2 v_{ij} = \mathrm{cov}(\mathcal{E}_i, \mathcal{E}_j) = \mathrm{cov}(\mathcal{E}_j, \mathcal{E}_i) = \sigma^2 v_{ji}.$$

Hence V must be a symmetric matrix.

Positive definite

Definition 5.1

The symmetric $n \times n$ matrix V is said to be

- **positive definite** if

$$\boldsymbol{a}^T V \boldsymbol{a} > 0 \text{ for all } \boldsymbol{a} \neq \mathbf{0}$$

- **non-negative definite** if

$$\boldsymbol{a}^T V \boldsymbol{a} \geq 0 \text{ for all } \boldsymbol{a}.$$

Variance matrices are non-negative definite

If $\boldsymbol{a}$ is a fixed vector then

$$0 \leq \text{var}(\boldsymbol{a}^T \boldsymbol{\mathcal{E}}) = \sigma^2 \boldsymbol{a}^T V \boldsymbol{a}.$$

To be a variance matrix, V must be non-negative definite.

Postive definite

If V is non-negative definite but not positive definite, then there must exist $\boldsymbol{a} \neq \mathbf{0}$ for which

$$\mathrm{var}(\boldsymbol{a}^T \boldsymbol{Y}) = 0.$$

In this case, one of the observations is a linear function of the others.

To eliminate any such redundancy, we assume that V is positive definite.

Ordinary least squares

The ordinary least squares (OLS) estimate of β is

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

It is easy to check that

$$E(\hat{\beta}_{OLS}) = \boldsymbol{\beta}$$

and

$$\mathrm{Var}(\hat{\beta}_{OLS}) = \sigma^2 (X^T X)^{-1} X^T V X (X^T X)^{-1}.$$

OLS not optimal

- $\hat{\beta}_{OLS}$ is an unbiased linear estimator that we can easily compute.
- But, the assumptions of the Gauss-Markov theorem do not hold.
- It is therefore not the **best** linear unbiased estimator.

Square-root matrices

Any square, symmetric matrix V is expressible in the form

$$V = E\Lambda E^T$$

where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ and $E^T E = EE^T = I$.

The matrix $V = E\Lambda E^T$ is non-negative definite if and only if $\lambda_i \geq 0$ and is positive definite if and only if $\lambda_i > 0$ for $i = 1, 2, \ldots, n$.

Square-root matrices

The symmetric square-root of a non-negative definite matrix is

$$V^{\frac{1}{2}} = E\Lambda^{\frac{1}{2}}E^T$$

where

$$\Lambda^{\frac{1}{2}} = \mathrm{diag}(\lambda_1^{\frac{1}{2}}, \lambda_2^{\frac{1}{2}}, \ldots, \lambda_n^{\frac{1}{2}}).$$

For V positive definite the symmetric square root is invertible and

$$V^{-\frac{1}{2}} = E\Lambda^{-\frac{1}{2}}E^T$$

Can check $V^{\frac{1}{2}}$ and $V^{-\frac{1}{2}}$ both symmetric and

$$V^{\frac{1}{2}}V^{\frac{1}{2}} = V,\; V^{-\frac{1}{2}}V^{-\frac{1}{2}} = V^{-1},\; \text{and } V^{-\frac{1}{2}}VV^{-\frac{1}{2}} = I.$$

Generalised least squares

Consider now the regression model

$$\boldsymbol{Y} = X\beta + \boldsymbol{\mathcal{E}}$$

where

$$E(\boldsymbol{\mathcal{E}}) = \mathbf{0} \text{ and } \mathrm{Var}(\boldsymbol{\mathcal{E}}) = \sigma^2 V.$$

Pre-multiplying by $V^{-\frac{1}{2}}$ gives

$$\boldsymbol{Y}_* = X_*\beta + \boldsymbol{\mathcal{E}_*}$$

where

$$\boldsymbol{Y}_* = V^{-\frac{1}{2}}\boldsymbol{Y},\ X_* = V^{-\frac{1}{2}}X,\ \text{and } \boldsymbol{\mathcal{E}}_* = V^{-\frac{1}{2}}\boldsymbol{\mathcal{E}}.$$

Generalised least squares

Applying the rules for linear transformation of random vectors, we find

$$E(\boldsymbol{\mathcal{E}}_*) = \mathbf{0}$$

and

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{\mathcal{E}}_*) &= \mathrm{Var}(V^{-\frac{1}{2}}\boldsymbol{\mathcal{E}}) \\
&= V^{-\frac{1}{2}}\,\mathrm{Var}(\boldsymbol{\mathcal{E}})\left\{V^{-\frac{1}{2}}\right\}^T \\
&= \sigma^2 V^{-\frac{1}{2}} V V^{-\frac{1}{2}} \\
&= \sigma^2 I.
\end{aligned}$$

Generalised least squares

Hence the Gauss-Markov theorem applies

$$\boldsymbol{Y}_* = X_* \beta + \boldsymbol{\mathcal{E}}_*$$

and the BLUE for β is

$$\hat{\boldsymbol{\beta}} = (X_*^T X_*)^{-1} X_*^T \boldsymbol{Y}_*$$

and

$$\mathrm{Var}(\hat{\beta}) = \sigma^2 (X_*^T X_*)^{-1}.$$

Generalised least squares

Substituting for X_* and $\boldsymbol{Y}_*$ produces the generalised least squares estimates

$$\hat{\beta}_{GLS} = (X^T V^{-1} X)^{-1} X^T V^{-1} \boldsymbol{Y}$$

and

$$\text{Var}(\hat{\beta}_{GLS}) = \sigma^2 (X^T V^{-1} X)^{-1}.$$

Weighted least squares

When V is a diagonal matrix, $V = \text{diag}(v_1, v_2, \ldots, v_n)$ the generalised least squares estimator is called the weighted least squares estimator, with weights

$$w_i = 1/v_i.$$

Many computer packages allow for the direct specification of the weights w_i.

Example

Suppose $Y_1, Y_2, \ldots, Y_n$ are independent with

$$E(Y_i) = \mu$$

and

$$\mathrm{var}(Y_i) = \sigma_i^2.$$

The ordinary least squares estimate is just $\bar{Y}$ with

$$E(\bar{Y}) = \mu$$

and

$$\mathrm{var}(\bar{Y}) = \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2.$$

Example (continued)

The weighted least squares estimate is obtained by taking,

$$X = \mathbf{1},\ \beta = (\mu) \text{ and } V = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2).$$

The weighted least squares estimate is

$$\hat{\mu} = \sum_{i=1}^{n} a_i Y_i$$

where

$$a_i = \frac{1/\sigma_i^2}{\sum_{j=1}^{n} 1/\sigma_j^2}.$$

Example (continued)

It follows that

$$E(\hat{\mu}) = \mu \text{ and } \operatorname{var}(\hat{\mu}) = \left(\sum_{j=1}^{n} \frac{1}{\sigma_j^2} \right)^{-1}.$$

It is easy to check that

$$\operatorname{var}(\hat{\mu}) \leq \operatorname{var}(\bar{Y})$$

with equality if and only if

$$\sigma_1^2 = \sigma_2^2 = \ldots \sigma_n^2.$$

Example (continued)

Intuitively, $\hat{\mu}$ gives greater weight to observations with lower variance, compared to the OLS estimator $\bar{Y}$ which weights all observations equally.

It can also be proved from first principles to be the best linear unbiased estimator for μ. (See Statistical Modelling and Inference II).

Box-Cox

Transformations

When the assumptions of linearity, homoscedasticity and normality are found to be violated, it is sometimes possible to find a simple transformation of the data for which the regression assumptions are more reasonable.

For example, in previous courses, you may have tried transformations such as $\log y$, $\sqrt{y}$, $1/y$ when dealing with a positive variable Y.

A simple approach is to consider each of these transformations and chose the one for which the model diagnostics (residuals) appear most reasonable.

The Box-Cox transformations

A more systematic approach for positive Y, is to consider the family of power transformations, y^λ. A convenient formulation is to consider the family of Box-Cox transformations defined by

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log y & \text{if } \lambda = 0. \end{cases}$$

It can be shown that

$$\lim_{\lambda \to 0} \frac{y^\lambda - 1}{\lambda} = \log y$$

so that $y^{(\lambda)}$ is a continuous function of λ.

Overview

The Box-Cox approach to transformation consists of the following steps.

- In the first instance, treat λ as an unknown parameter and obtain an estimate $\hat{\lambda}$ from the data;
- Perform the usual regression diagnostics on the transformed data.
  - If they are satisfactory, adopt $\hat{\lambda}$.
  - If they are not satisfactory, conclude that no suitable transformationis available.
- When a transformation is adopted, $\hat{\lambda}$ is treated as a known constant and analysis of the transformed data proceeds in the usual way.

Remark

A more refined approach would be to treat $\hat{\lambda}$ as a parameter estimate throughout the entire analysis.

However, this would make the analyses substantially more complicated and Box and Cox (1964) argued that the benefits would be very small.

The profile likelihood

The method of estimation for λ is a variant on maximum likelihood called profile likelihood and consists of the following steps:

1. Obtain the full log-likelihood function, $\ell(\lambda, \boldsymbol{\beta}, \sigma^2; \boldsymbol{y})$.
2. For fixed λ, let $\hat{\boldsymbol{\beta}}_\lambda$ and $\hat{\sigma}^2_\lambda$ be the MLEs on $\boldsymbol{y}^{(\lambda)}$.
3. The **profile likelihood** is defined by

$$\hat{\ell}(\lambda) = \ell(\lambda, \hat{\boldsymbol{\beta}}_\lambda, \hat{\sigma}^2_\lambda; \boldsymbol{y}).$$

4. Maximizing $\hat{\ell}(\lambda)$ gives the overall maximum likelihood estimate for λ.
5. The profile likelihood can also be used to obtain an approximate 95% confidence interval for λ. This allows us to choose a convenient nearby value for λ. For example, if $\hat{\lambda} = -0.483$ we might choose to use -0.5 etc.
6. In practice, when dealing with non-negative data, a constant *a* is sometimes added to all data values to avoid zeros.

The profile likelihood

The profile likelihood is given by

$$\hat{\ell}(\lambda) = \text{const} - \frac{n}{2} \log \text{RSS}(z^{(\lambda)})$$

where RSS is the residual sum of squares when

$$z^{(\lambda)} = \frac{y^{(\lambda)}}{\dot{y}^{\lambda-1}}$$

is taken as response variable in the multiple regression

$$\boldsymbol{\eta} = X\boldsymbol{\beta}$$

and $\dot{y}^{\lambda-1}$ is the geometric mean of $y_1^{\lambda-1}, \ldots, y_n^{\lambda-1}$.

The profile likelihood

Let $\boldsymbol{x}_i$ be the ith row of X. The transformed regression model postulates

$$Y_i^{(\lambda)} \sim N(\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma^2).$$

Applying the transformation rule for a continuous random variable yields

$$f(y_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i^{(\lambda)} - \boldsymbol{x}_i^T \beta)^2}{2\sigma^2}\right) y_i^{\lambda-1}$$

The profile likelihood

The full log-likelihood is therefore,

$$
\begin{aligned}
\ell(\lambda, \boldsymbol{\beta}, \sigma^2; \boldsymbol{y}) &= \log \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i^{(\lambda)} - \boldsymbol{x}_i^T \boldsymbol{\beta})^2}{2\sigma^2} \right) y_i^{\lambda-1} \right\} \\
&= -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i^{(\lambda)} - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 \\
&\qquad + n \log \dot{y}^{\lambda-1}
\end{aligned}
$$

The profile likelihood

Maximising ℓ with respect to β yields

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

and then maximising with respect to σ^2 yields

$$\hat{\sigma^2_\lambda} = \frac{\text{RSS}(\boldsymbol{y}^{(\lambda)})}{n}.$$

Substituting into ℓ, we obtain

$$\begin{aligned}
\hat{\ell}(\lambda) &= \ell(\lambda, \hat{\boldsymbol{\beta}}_\lambda, \hat{\sigma^2_\lambda}; \boldsymbol{y}) \\
&= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \text{RSS}(\boldsymbol{y}^{(\lambda)}) + \frac{n}{2}\log n - \frac{n}{2} + n \log \dot{y}^{\lambda-1} \\
&= \text{const} - \frac{n}{2}\log \frac{\text{RSS}(\boldsymbol{y}^{(\lambda)})}{(\dot{y}^{\lambda-1})^2}
\end{aligned}$$

The profile likelihood

Finally, observe that $\mathrm{RSS}(c\boldsymbol{w}) = c^2\,\mathrm{RSS}(\boldsymbol{w})$ for any scalar x and vector $\boldsymbol{w}$.

Hence

$$\hat{\ell}(\lambda) = \mathrm{const} - \frac{n}{2}\log \mathrm{RSS}(\boldsymbol{z}^{(\lambda)}).$$