MA 575 Linear Models:  
Cedric E. Ginestet, Boston University

*Gauss-Markov Theorem, Weighted Least Squares*

Week 6, Lecture 2

1 Gauss-Markov Theorem

1.1 Assumptions

We make three crucial assumptions on the joint moments of the error terms. These assumptions are required for the Gauss-Markov theorem to hold. Note that this theorem also assumes that the fitted model is linear in the parameters.

i. Firstly, we assume that the expectations of all the error terms are centered at zero, such that

\[ E[E_i | x_i] = 0, \quad i = 1, \ldots, n. \]

ii. Secondly, we also assume that the variances of the error terms are constant for every \( i = 1, \ldots, n \). This assumption is referred to as homoscedasticity.

\[ \text{Var}[E_i|x_i] = \sigma^2, \quad i = 1, \ldots, n. \]

iii. Thirdly, we assume that the error terms are uncorrelated,

\[ \text{Cov}[E_i, E_j|x_i, x_j] = 0, \quad \forall \ i \neq j. \]

1.2 BLUEs

**Definition 1.** Given a random sample, \( Y_1, \ldots, Y_n \overset{\text{ind}}{\sim} f(X, \beta) \); an estimator \( \hat{\beta}(Y_1, \ldots, Y_n) \) of the parameter \( \beta \) is said to be unbiased if

\[ E[\hat{\beta}|X] = \beta, \]

for every \( \beta \in \mathbb{R}^p \).

**Definition 2.** An estimator \( \hat{\beta} \) of a parameter \( \beta \) is said to be Best Linear Unbiased Estimator (BLUE), if it is a linear function of the observed values \( y \), an unbiased estimator of \( \beta \); and if for any other linear unbiased estimator \( \tilde{\beta} \), we have

\[ \text{Var}[\hat{\beta}|X] \leq \text{Var}[\tilde{\beta}|X]. \]

1.3 Proof of Theorem

**Theorem 1.** Under the G-M assumptions, a multiple regression model with mean and variance functions respectively defined as

\[ E[y|X] = X\beta \quad \text{and} \quad \text{Var}[y|X] = \sigma^2 I, \]

the OLS estimator \( \hat{\beta} := (X^T X)^{-1} X^T y \) is BLUE for \( \beta \).
Proof. We need to show that for any arbitrary linear unbiased estimator of $\beta$, denoted $\tilde{\beta}$, the following matrix is negative semidefinite,

$$\text{Var}[\hat{\beta}|X] - \text{Var}[\tilde{\beta}|X] \leq 0.$$ 

(i) Firstly, since both $\hat{\beta}$ and $\tilde{\beta}$ are linear functions of $y$, it follows that there exists two matrices $C$ and $D$ of order $p^* \times n$, with $C := (X^TX)^{-1}X^T$, and such that

$$\hat{\beta} = Cy, \quad \text{and} \quad \tilde{\beta} = (C + D)y.$$ 

(ii) Secondly, as both $\hat{\beta}$ and $\tilde{\beta}$ are also unbiased, we hence have

$$E[\tilde{\beta}|X] = E[(C + D)y|X] = E[Cy|X] + DE[y|X] = \beta + DX\beta,$$

and therefore $DX\beta$ must be zero for $\tilde{\beta}$ to be unbiased. In fact, since unbiasedness holds for every values of $\beta \in \mathbb{R}^p$, it follows that $DX\beta = 0$ for every $\beta$, which implies that

$$DX = 0, \quad \text{and} \quad X^TD^T = 0.$$ 

(iii) Finally, it suffices to compute the variance of $\tilde{\beta}$,

$$\text{Var}[\tilde{\beta}|X] = \text{Var}[(C + D)y|X] = (C + D)\text{Var}[y|X](C + D)^T = \sigma^2(CC^T + CD^T + DC^T + DD^T).$$

Observe that by equation (1), we have

$$DC^T = DX(X^TX)^{-1} = 0, \quad \text{and} \quad CD^T = (X^TX)^{-1}X^TD^T = 0,$$

and therefore

$$\text{Var}[\tilde{\beta}|X] = \sigma^2(CC^T + DD^T) = \sigma^2(X^TX)^{-1} + \sigma^2DD^T = \text{Var}[\hat{\beta}|X] + \sigma^2DD^T.$$ 

However, since $DD^T$ is a Gram matrix of order $p^* \times p^*$, it follows that it is at least positive semidefinite, such that $DD^T \geq 0$. Therefore, we indeed obtain $\text{Var}[\tilde{\beta}|X] \geq \text{Var}[\hat{\beta}|X]$, as required. 

This theorem can be generalized to weighted least squares (WLS) estimators. A more geometric proof of the Gauss-Markov theorem can be found in Christensen (2011), using the properties of the hat matrix. However, this latter proof technique is less natural as it relies on comparing the variances of the fitted values corresponding to two different estimators, as a proxy for the actual variances of these estimators. Finally, yet another proof can be found in Casella and Berger (2002), on p. 544.

2 Weighted Least Squares (WLS)

The classical OLS setup can be extended by including a set of weights associated with each data point.

$$E[Y|X = x_i] = x_i^T\beta, \quad \text{and} \quad \text{Var}[Y|X = x_i] = \frac{\sigma^2}{w_i}.$$
where the \( w_i \)'s are known positive numbers, such that
\[
w_i > 0, \quad \forall \ i = 1, \ldots, n.
\]
These weights may naturally come from the number of 'samples', associated with each data point. This is especially the case, when every data point is a sample average of some quantity, such as the number of cancer cases in a particular geographical location. This extension can be formulated using matrix notation as follows,
\[
y = X\beta + e \quad \text{and} \quad \text{Var}[e|X] = \sigma^2 W^{-1},
\]
where \( W \) is assumed to be a diagonal matrix. It then suffices to specify a statistical criterion, such that
\[
\text{RSS}(\beta; W) := \frac{1}{2} \sum_{i=1}^{n} w_i (y_i - x_i^T \beta)^2
\]
Alternatively, this may be re-expressed in terms of the error vector, \( e = y - X\beta \), such that
\[
\text{RSS}(\beta; W) = e^T We = \sum_{i=1}^{n} \frac{e_i^2}{w_i}.
\]

2.1 Fitting WLS using the OLS Framework

It is useful to try to re-formulate this WLS optimization into the standard OLS framework that we have already encountered. Hence, consider the following matrix decompositions,
\[
W = W^{1/2}W^{1/2}, \quad \text{and} \quad W^{1/2}W^{-1/2} = W^{-1/2}W^{1/2} = I;
\]
where the diagonal entries in \( W^{1/2} \) and \( W^{-1/2} \) are respectively defined for every \( i = 1, \ldots, n \) as
\[
(W^{1/2})_{ii} := \sqrt{w_i}, \quad \text{and} \quad (W^{-1/2})_{ii} := \frac{1}{\sqrt{w_i}}.
\]
Once we have performed this decomposition, we can transform our original WLS model, such that we pre-multiply both sides in this fashion,
\[
W^{1/2}y = W^{1/2}X\beta + W^{1/2}e,
\]
and define the following terms,
\[
z := W^{1/2}y, \quad M := W^{1/2}X, \quad \text{and} \quad d := W^{1/2}e.
\]
Observe that the vector of parameters of interest, \( \beta \), has not been affected by this change of notation. Using these definitions, we can now re-define our target OLS model as follows,
\[
z = M\beta + d.
\]
This yields a new RSS, which can be shown to be equivalent to the one described in equation (2),
\[
\text{RSS}(\beta; W) = (z - M\beta)^T (z - M\beta)
= [W^{1/2}(y - X\beta)]^T [W^{1/2}(y - X\beta)]
= (y - X\beta)^T W(y - X\beta).
\]
This can be directly minimized using the standard machinery that we have developed for minimizing unweighted residual sum of squares. In addition, note that the variance function is given by

\[ \text{Var}[d|X] = W^{1/2} \text{Var}[e|X](W^{1/2})^T = W^{1/2} \sigma^2 W^{-1} (W^{1/2})^T = \sigma^2 W^{-1/2} W^{-1/2} (W^{1/2}) = \sigma^2 I. \]

In summary, we therefore have recovered, after some transformations, a standard OLS model taking the form,

\[ z = M\beta + d, \quad \text{and} \quad \text{Var}[d|X] = \sigma^2 I. \]

It simply remains to compute the actual form of \( \beta \) with respect to \( W \), such that

\[ \hat{\beta}_{WLS} = (M^T M)^{-1} M^T z = ((W^{1/2} X)^T W^{1/2} X)^{-1} (W^{1/2} X)^T z \]

\[ = (X^T W^{1/2} X)^{-1} X^T W^{1/2} y = (X^T W X)^{-1} X^T W y. \]

The equivalence between the WLS and the OLS framework is best observed by considering the entries of \( M \) and \( z \). The new **weighted design matrix** and vector of observations are now,

\[
M = \begin{bmatrix}
\sqrt{w_1} & \sqrt{w_1} x_{11} & \cdots & \sqrt{w_1} x_{1p} \\
\sqrt{w_2} & \sqrt{w_2} x_{21} & \cdots & \sqrt{w_2} x_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
\sqrt{w_n} & \sqrt{w_n} x_{n1} & \cdots & \sqrt{w_n} x_{np}
\end{bmatrix}, \quad \text{and} \quad z = \begin{bmatrix}
\sqrt{w_1} y_1 \\
\sqrt{w_2} y_2 \\
\vdots \\
\sqrt{w_n} y_n
\end{bmatrix}
\]

The regression problem simply involves finding the WLS fitted values \( \hat{z} := M \hat{\beta} \).

### 2.2 Generalized Least Squares (GLS)

The WLS extension of OLS can be further generalized by considering any symmetric and positive definite matrix, such that

\[ \text{Var}[e|X] := \Sigma^{-1}, \quad (3) \]

where the generalized residual sum of squares becomes

\[ \text{RSS}(\beta; \Sigma) := (y - X\beta)^T \Sigma (y - X\beta), \]

which can also be re-written as

\[ \text{RSS}(\beta; \Sigma) = \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma_{ij}^2 (y_i - x_i^T \beta)(y_j - x_j^T \beta), \]

where \( \sigma_{ij}^2 := \Sigma_{ij} \). Noting that the inverse of a positive definite matrix is also positive definite, we require that

\[ \Sigma = \Sigma^T \quad \text{and} \quad \Sigma > 0, \]

Department of Mathematics and Statistics, Boston University
which implies that for every non-zero \( v \in \mathbb{R}^n \), we have \( v^T \Sigma v > 0 \). Every symmetric positive definite matrices can be Cholesky decomposed, such that
\[
\Sigma = LL^T,
\]
where \( L \) is a lower triangular matrix of dimension \( n \times n \). As a result, we can perform the same manipulations that we have conducted for WLS, such that if we pre-multiply both sides of our GLS equation we obtain
\[
L^T y = L^T X \beta + L^T e,
\]
and define the following terms,
\[
z := L^T y, \quad M := L^T X, \quad \text{and} \quad d := L^T e.
\]
Then, we can again apply the standard OLS minimization machinery, after having verified that the variance of \( d \) is simply \( I \). Moreover, it is straightforward to see that the Gauss-Markov theorem also holds under these more general assumptions, such that the GLS estimator
\[
\hat{\beta}_{GLS} := (X^T \Sigma X)^{-1} X^T \Sigma y,
\]
is also BLUE, amongst the class of unbiased linear estimators in a model, whose variance function is \( \text{Var}[e|X] := \Sigma^{-1} \).

References
