# 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

$$\mathbb{E}[E_i|\mathbf{x}_i] = 0, \qquad i = 1, \ldots, n.$$

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**.

$$\mathbb{V}\text{ar}[E_i|\mathbf{x}_i] = \sigma^2, \qquad i = 1, \ldots, n.$$

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

$$\mathbb{C}\text{ov}[E_i, E_j|\mathbf{x}_i, \mathbf{x}_j] = 0, \qquad \forall\ i \neq j.$$

### 1.2 BLUEs

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

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta},$$

*for every* $\boldsymbol{\beta} \in \mathbb{R}^{p^*}$.

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

$$\mathbb{V}\text{ar}[\widehat{\boldsymbol{\beta}}|\mathbf{X}] \leq \mathbb{V}\text{ar}[\widetilde{\boldsymbol{\beta}}|\mathbf{X}].$$

### 1.3 Proof of Theorem

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

$$\mathbb{E}[\mathbf{y}|\mathbf{X}] = \mathbf{X}\boldsymbol{\beta} \qquad \textit{and} \qquad \mathbb{V}\text{ar}[\mathbf{y}|\mathbf{X}] = \sigma^2\mathbf{I},$$

*the OLS estimator* $\widehat{\boldsymbol{\beta}} := (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ *is BLUE for* $\boldsymbol{\beta}$.

*Proof.* We need to show that for any arbitrary linear unbiased estimator of $\boldsymbol{\beta}$, denoted $\widetilde{\boldsymbol{\beta}}$, the following matrix is negative semidefinite,

$$\mathbb{V}\text{ar}[\widehat{\boldsymbol{\beta}}|\mathbf{X}] - \mathbb{V}\text{ar}[\widetilde{\boldsymbol{\beta}}|\mathbf{X}] \leq 0.$$

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

$$\widehat{\boldsymbol{\beta}} = \mathbf{C}\mathbf{y}, \qquad \text{and} \qquad \widetilde{\boldsymbol{\beta}} = (\mathbf{C}+\mathbf{D})\mathbf{y}.$$

**(ii)** Secondly, as both $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ are also *unbiased*, we hence have

$$\begin{aligned}\mathbb{E}[\widetilde{\boldsymbol{\beta}}|\mathbf{X}] &= \mathbb{E}[(\mathbf{C}+\mathbf{D})\mathbf{y}|\mathbf{X}] \\ &= \mathbb{E}[\mathbf{C}\mathbf{y}|\mathbf{X}] + \mathbf{D}\mathbb{E}[\mathbf{y}|\mathbf{X}] \\ &= \boldsymbol{\beta} + \mathbf{D}\mathbf{X}\boldsymbol{\beta},\end{aligned}$$

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

$$\mathbf{D}\mathbf{X} = \mathbf{0}, \qquad \text{and} \qquad \mathbf{X}^T\mathbf{D}^T = \mathbf{0}. \tag{1}$$

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

$$\begin{aligned}\mathbb{V}\text{ar}[\widetilde{\boldsymbol{\beta}}|\mathbf{X}] &= \mathbb{V}\text{ar}[(\mathbf{C}+\mathbf{D})\mathbf{y}|\mathbf{X}] \\ &= (\mathbf{C}+\mathbf{D})\,\mathbb{V}\text{ar}[\mathbf{y}|\mathbf{X}](\mathbf{C}+\mathbf{D})^T \\ &= \sigma^2(\mathbf{C}\mathbf{C}^T + \mathbf{C}\mathbf{D}^T + \mathbf{D}\mathbf{C}^T + \mathbf{D}\mathbf{D}^T).\end{aligned}$$

Observe that by equation (1), we have

$$\mathbf{D}\mathbf{C}^T = \mathbf{D}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = \mathbf{0}, \qquad \text{and} \qquad \mathbf{C}\mathbf{D}^T = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}^T = \mathbf{0},$$

and therefore

$$\begin{aligned}\mathbb{V}\text{ar}[\widetilde{\boldsymbol{\beta}}|\mathbf{X}] &= \sigma^2(\mathbf{C}\mathbf{C}^T + \mathbf{D}\mathbf{D}^T) \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} + \sigma^2\mathbf{D}\mathbf{D}^T \\ &= \mathbb{V}\text{ar}[\widehat{\boldsymbol{\beta}}|\mathbf{X}] + \sigma^2\mathbf{D}\mathbf{D}^T.\end{aligned}$$

However, since $\mathbf{D}\mathbf{D}^T$ is a *Gram matrix* of order $p^* \times p^*$, it follows that it is at least **positive semidefinite**, such that $\mathbf{D}\mathbf{D}^T \geq 0$. Therefore, we indeed obtain $\mathbb{V}\text{ar}[\widetilde{\boldsymbol{\beta}}|\mathbf{X}] \geq \mathbb{V}\text{ar}[\widehat{\boldsymbol{\beta}}|\mathbf{X}]$, 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.

$$\mathbb{E}[Y|X=\mathbf{x}_i] = \mathbf{x}_i^T\boldsymbol{\beta}, \qquad \text{and} \qquad \mathbb{V}\text{ar}[Y|X=\mathbf{x}_i] = \frac{\sigma^2}{w_i},$$

where the $w_i$'s are *known positive numbers*, such that

$$w_i > 0, \ \forall \ i = 1, \ldots, n.$$

These weights may naturally come from the number of 'samples', associated with each data points. 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,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \qquad \text{and} \qquad \mathbb{V}\text{ar}[\mathbf{e}|\mathbf{X}] = \sigma^2 \mathbf{W}^{-1},$$

where $\mathbf{W}$ is assumed to be a **diagonal** matrix. It then suffices to specify a statistical criterion, such that

$$\begin{aligned} \text{RSS}(\boldsymbol{\beta}; \mathbf{W}) &:= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{W} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} w_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 \end{aligned} \tag{2}$$

Alternatively, this may be re-expressed in terms of the error vector, $\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$, such that

$$\text{RSS}(\boldsymbol{\beta}; \mathbf{W}) = \mathbf{e}^T \mathbf{W} \mathbf{e} = \sum_{i=1}^{n} \frac{e_i^2}{w_i}.$$

## 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,

$$\mathbf{W} = \mathbf{W}^{1/2} \mathbf{W}^{1/2}, \qquad \text{and} \qquad \mathbf{W}^{1/2} \mathbf{W}^{-1/2} = \mathbf{W}^{-1/2} \mathbf{W}^{1/2} = \mathbf{I};$$

where the diagonal entries in $\mathbf{W}^{1/2}$ and $\mathbf{W}^{-1/2}$ are respectively defined for every $i = 1, \ldots, n$ as

$$(\mathbf{W}^{1/2})_{ii} := \sqrt{w_i}, \qquad \text{and} \qquad (\mathbf{W}^{-1/2})_{ii} := \frac{1}{\sqrt{w_i}}.$$

Once we have performed this decomposition, we can transform our original WLS model, such that we **pre-multiply** both sides in this fashion,

$$\mathbf{W}^{1/2} \mathbf{y} = \mathbf{W}^{1/2} \mathbf{X} \boldsymbol{\beta} + \mathbf{W}^{1/2} \mathbf{e},$$

and define the following terms,

$$\mathbf{z} := \mathbf{W}^{1/2} \mathbf{y}, \qquad \mathbf{M} := \mathbf{W}^{1/2} \mathbf{X}, \qquad \text{and} \qquad \mathbf{d} := \mathbf{W}^{1/2} \mathbf{e}.$$

Observe that the vector of parameters of interest, $\boldsymbol{\beta}$, has not been affected by this change of notation. Using these definitions, we can now re-define our target OLS model as follows,

$$\mathbf{z} = \mathbf{M}\boldsymbol{\beta} + \mathbf{d}.$$

This yields a new RSS, which can be shown to be equivalent to the one described in equation (2),

$$\begin{aligned} \text{RSS}(\boldsymbol{\beta}; \mathbf{W}) &= (\mathbf{z} - \mathbf{M}\boldsymbol{\beta})^T (\mathbf{z} - \mathbf{M}\boldsymbol{\beta}) \\ &= [\mathbf{W}^{1/2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})]^T [\mathbf{W}^{1/2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})] \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{W} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \end{aligned}$$

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

$$\begin{aligned}\mathbb{V}\text{ar}[\mathbf{d}|\mathbf{X}] &= \mathbb{V}\text{ar}[\mathbf{W}^{1/2}\mathbf{e}|\mathbf{X}] \\ &= \mathbf{W}^{1/2}\,\mathbb{V}\text{ar}[\mathbf{e}|\mathbf{X}](\mathbf{W}^{1/2})^T \\ &= \mathbf{W}^{1/2}\sigma^2\mathbf{W}^{-1}(\mathbf{W}^{1/2})^T \\ &= \sigma^2\mathbf{W}^{1/2}\mathbf{W}^{-1/2}\mathbf{W}^{-1/2}(\mathbf{W}^{1/2}) \\ &= \sigma^2\mathbf{I}.\end{aligned}$$

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

$$\mathbf{z} = \mathbf{M}\boldsymbol{\beta} + \mathbf{d}, \qquad \text{and} \qquad \mathbb{V}\text{ar}[\mathbf{d}|\mathbf{X}] = \sigma^2\mathbf{I}.$$

It simply remains to compute the actual form of $\boldsymbol{\beta}$ with respect to $\mathbf{W}$, such that

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{WLS}} &= (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{z} \\ &= ((\mathbf{W}^{1/2}\mathbf{X})^T\mathbf{W}^{1/2}\mathbf{X})^{-1}(\mathbf{W}^{1/2}\mathbf{X})^T\mathbf{z} \\ &= (\mathbf{X}^T\mathbf{W}^{1/2}\mathbf{W}^{1/2}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{1/2}\mathbf{W}^{1/2}\mathbf{y} \\ &= (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}.\end{aligned}$$

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

$$\mathbf{M} = \begin{vmatrix} \sqrt{w_1} & \sqrt{w_1}x_{11} & \dots & \sqrt{w_1}x_{1p} \\ \sqrt{w_2} & \sqrt{w_2}x_{21} & \dots & \sqrt{w_2}x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sqrt{w_n} & \sqrt{w_n}x_{n1} & \dots & \sqrt{w_n}x_{np} \end{vmatrix}, \qquad \text{and} \qquad \mathbf{z} = \begin{vmatrix} \sqrt{w_1}y_1 \\ \sqrt{w_2}y_2 \\ \vdots \\ \sqrt{w_n}y_n \end{vmatrix}$$

The regression problem simply involves finding the WLS fitted values $\widehat{\mathbf{z}} := \mathbf{M}\widehat{\boldsymbol{\beta}}$.

## 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

$$\mathbb{V}\text{ar}[\mathbf{e}|\mathbf{X}] := \boldsymbol{\Sigma}^{-1}, \tag{3}$$

where the generalized residual sum of squares becomes

$$\text{RSS}(\boldsymbol{\beta}; \boldsymbol{\Sigma}) := (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

which can also be re-written as

$$\text{RSS}(\boldsymbol{\beta}; \boldsymbol{\Sigma}) = \sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{ij}^2(y_i - \mathbf{x}_i^T\boldsymbol{\beta})(y_j - \mathbf{x}_j^T\boldsymbol{\beta}),$$

where $\sigma_{ij}^2 := \boldsymbol{\Sigma}_{ij}$. Noting that the inverse of a positive definite matrix is also positive definite, we require that

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^T \qquad \text{and} \qquad \boldsymbol{\Sigma} > 0,$$

which implies that for every non-zero $\mathbf{v} \in \mathbb{R}^n$, we have $\mathbf{v}^T \mathbf{\Sigma} \mathbf{v} > 0$. Every **symmetric positive definite** matrices can be Cholesky decomposed, such that

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^T,$$

where $\mathbf{L}$ 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

$$\mathbf{L}^T \mathbf{y} = \mathbf{L}^T \mathbf{X} \boldsymbol{\beta} + \mathbf{L}^T \mathbf{e},$$

and define the following terms,

$$\mathbf{z} := \mathbf{L}^T \mathbf{y}, \qquad \mathbf{M} := \mathbf{L}^T \mathbf{X}, \qquad \text{and} \qquad \mathbf{d} := \mathbf{L}^T \mathbf{e}.$$

Then, we can again apply the standard OLS minimization machinery, after having verified that the variance of $\mathbf{d}$ is simply $\mathbf{I}$. Moreover, it is straightforward to see that the Gauss-Markov theorem also holds under these more general assumptions, such that the GLS estimator

$$\widehat{\boldsymbol{\beta}}_{\text{GLS}} := (\mathbf{X}^T \mathbf{\Sigma} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{\Sigma} \mathbf{y},$$

is also BLUE, amongst the class of unbiased linear estimators in a model, whose variance function is $\mathbb{V}\text{ar}[\mathbf{e}|\mathbf{X}] := \mathbf{\Sigma}^{-1}$.

## References

Casella, G. and Berger, R. (2002). *Statistical Inference (2nd edition).* Duxbury, New York.

Christensen, R. (2011). *Plane Answers to Complex Questions: The Theory of Linear Models (4th edition).* Springer, New York.