UNIVERSITY of DELAWARE

Modern Machine Learning
Gauss-Markov Theorem

Kenneth E. Barner
Department of Electrical and Computer Engineering
University of Delaware

Gauss-Markov Theorem

Objective: Prove that the least-squares determine linear model parameters, $\widehat{\boldsymbol{\beta}}_{LS}$, have the smallest variance among all linear unbiased estimates

Gauss-Markov Theorem

Theorem Assumptions:

- $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$,
  - where $\mathbf{X}$ is fixed (non-random) and $\boldsymbol{\epsilon}$ components are iid $N(0, \sigma^2)$
- $\widehat{\boldsymbol{\beta}} = \mathbf{C}\mathbf{y}$ be a linear unbiased estimator of $\boldsymbol{\beta}$

Then if for all $\boldsymbol{a} \in \mathbb{R}^{(p+1)}$

$$MSE\left(\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}_{LS}\right) \leq MSE\left(\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}\right)$$ [MSE = Mean Squared Error]

We say that $\widehat{\boldsymbol{\beta}}_{LS}$ is the Best Linear Unbiased Estimator (BLUE) of $\boldsymbol{\beta}$

Gauss-Markov Theorem

To prove the result consider first the: **bias-variance decomposition of the MSE**

Let $\boldsymbol{Z} = \boldsymbol{a}^T\boldsymbol{\beta}$ and $\widehat{\boldsymbol{Z}} = \boldsymbol{a}^T\widehat{\boldsymbol{\beta}}$ then

$$\begin{aligned} MSE\left(\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}\right) &= E\left[\left(\boldsymbol{a}^T(\boldsymbol{\beta}-\widehat{\boldsymbol{\beta}})\right)^2\right] = E\left[(\boldsymbol{Z}-\widehat{\boldsymbol{Z}})^{\mathbf{2}}\right] \\ &= E\left[\boldsymbol{Z}^2 - 2\boldsymbol{Z}\widehat{\boldsymbol{Z}} + \widehat{\boldsymbol{Z}}^2\right] \\ &= E[\boldsymbol{Z}^2] - 2E\left[\boldsymbol{Z}\widehat{\boldsymbol{Z}}\right] + E\left[\widehat{\boldsymbol{Z}}^2\right] \\ &= \boldsymbol{Z}^2 - 2\boldsymbol{Z}E\left[\widehat{\boldsymbol{Z}}\right] + Var\left[\widehat{\boldsymbol{Z}}\right] + E\left[\widehat{\boldsymbol{Z}}\right]^2 \qquad (*) \\ &= \left(\boldsymbol{Z} - E\left[\widehat{\boldsymbol{Z}}\right]\right)^2 + \mathrm{Var}\left[\widehat{\boldsymbol{Z}}\right] \end{aligned}$$

$\mathrm{bias}^2 := \left(\boldsymbol{Z} - E\left[\widehat{\boldsymbol{Z}}\right]\right)^2$, $\mathrm{variance} := \mathrm{Var}\left[\widehat{\boldsymbol{Z}}\right]$

Therefore, if $\widehat{\boldsymbol{\beta}}$ is unbiased then $MSE\left(\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}\right) = \mathrm{Var}\left[\widehat{\boldsymbol{Z}}\right]$

NOTE: derivation (*) above utilizes known result $Var\left[\widehat{\boldsymbol{Z}}\right] = E\left[\widehat{\boldsymbol{Z}}^2\right] - E\left[\widehat{\boldsymbol{Z}}\right]^2$

Gauss-Markov Theorem

To prove the Gauss-Markov theorem it suffices to show that

$$Var\left(\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}_{LS}\right) \leq Var\left(\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}\right), \forall \boldsymbol{a} \in \mathbb{R}^{(p+1)}$$

Proof: Let $\widehat{\boldsymbol{\beta}} = \mathbf{Cy}$, where $\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-\mathbf{1}}\mathbf{X}^T + \mathbf{D}$, for some $\mathbf{D} \in \mathbb{R}^{(p+1)\times n}$ then

$$\begin{aligned} E\left[\widehat{\boldsymbol{\beta}}\right] &= E\left[\left((\mathbf{X}^T\mathbf{X})^{-\mathbf{1}}\mathbf{X}^T + \mathbf{D}\right)\mathbf{y}\right] \\ &= E\left[\left((\mathbf{X}^T\mathbf{X})^{-\mathbf{1}}\mathbf{X}^T + \mathbf{D}\right)(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})\right] \\ &= (\mathbf{I} + \mathbf{DX})\boldsymbol{\beta} \end{aligned}$$

Observation: It *must* hold that $\mathbf{DX} = \mathbf{0}$ for $\widehat{\boldsymbol{\beta}}$ to be unbiased

Now calculate $\mathrm{Var}\left(\boldsymbol{a}^T\widehat{\boldsymbol{\beta}}\right)$

Gauss-Markov Theorem

Then we have

$$\begin{aligned}
\operatorname{Var}(\hat{\boldsymbol{\beta}}) &= \operatorname{Var}(\mathbf{C}\mathbf{y}) \\
&= \mathbf{C}\operatorname{Var}(\mathbf{y})\mathbf{C}^T = \sigma^2\mathbf{C}\mathbf{C}^T \quad [\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \Longrightarrow \operatorname{Var}(\mathbf{y}) = \sigma^2\boldsymbol{I}] \\
&= \sigma^2\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{D}\right)\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T + \mathbf{D}\right)^T \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} + \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}^T \\
&\quad + \sigma^2\mathbf{D}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} + \sigma^2\mathbf{D}\mathbf{D}^T \\
&= \sigma^2\left[(\mathbf{X}^T\mathbf{X})^{-1} + \mathbf{D}\mathbf{D}^T\right] \qquad [\text{since } \mathbf{D}\mathbf{X} = \mathbf{0}]
\end{aligned}$$

Note: $\mathbf{X}^T\mathbf{X}$ and $\mathbf{D}\mathbf{D}^T$ are positive semidefinite

$$\Longrightarrow \boldsymbol{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{a} \geq 0, \qquad \boldsymbol{a}^T(\mathbf{D}^T\mathbf{D})\boldsymbol{a} \geq 0$$

Therefore

$$\begin{aligned}\mathrm{Var}\left(\boldsymbol{a}^T\hat{\beta}\right) &= \underline{\boldsymbol{a}^T\sigma^2\left[(\mathbf{X}^T\mathbf{X})^{-\mathbf{1}} + \mathbf{D}\mathbf{D}^T\right]\boldsymbol{a}} \geq \underline{\boldsymbol{a}^T\sigma^2(\mathbf{X}^T\mathbf{X})^{-\mathbf{1}}\boldsymbol{a}} \\ &\geq \mathrm{Var}\left(\boldsymbol{a}^T\hat{\beta}_{LS}\right)\end{aligned}$$

Since $\mathrm{Var}\left(\boldsymbol{a}^T\hat{\beta}_{LS}\right) = \mathrm{Var}\left(\boldsymbol{a}^T(\mathbf{X}^T\mathbf{X})^{-\mathbf{1}}\boldsymbol{y}\right) = \boldsymbol{a}^T\sigma^2(\mathbf{X}^T\mathbf{X})^{-\mathbf{1}}\boldsymbol{a}$

This concludes the proof

Result: The Least Squares estimate $\hat{\beta}_{LS}$ is unbiased and has the smallest weight variance $\Longrightarrow$ it is the Best Linear Unbiased Estimate (BLUE)