Chapter 4

The Gauss-Markov Theorem

In Chap. 3 we showed that the least squares estimator, $\widehat{\boldsymbol{\beta}}_{LSE}$, in a Gaussian linear model has is *unbiased*, meaning that $E[\widehat{\boldsymbol{\beta}}_{LSE}] = \boldsymbol{\beta}$, and that its variance-covariance matrix is

$$\operatorname{Var}\widehat{\boldsymbol{\beta}}_{LSE} = \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} = \sigma^2\boldsymbol{R}^{-1}(\boldsymbol{R}^{-1})'.$$

The Gauss-Markov theorem says that this variance-covariance (or *dispersion*) is the best that we can do when we restrict ourselved to *linear unbiased estimators*, which means estimators that are linear functions of $\mathcal{Y}$ and are unbiased.

To make these definitions more formal:

Definition 5 (Minimum Dispersion)**.** *Let $\mathcal{T} = (\mathcal{T}_1, \ldots, \mathcal{T}_p)'$ be an estimator of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$. The dispersion of $\mathcal{T}$ is $\boldsymbol{D}(\mathcal{T}) = E[(\mathcal{T} - \boldsymbol{\theta})(\mathcal{T} - \boldsymbol{\theta})']$. If $\mathcal{T}$ is unbiased then its dispersion is simply its variance-covariance matrix, $\boldsymbol{D}(\mathcal{T}) = \operatorname{Var}(\mathcal{T})$. $\mathcal{T}$ is minimum dispersion unbiased estimator of $\boldsymbol{\theta}$ if $\boldsymbol{D}(\tilde{\mathcal{T}}) - \boldsymbol{D}(\mathcal{T})$ is positive semidefinite for any unbiased estimator $\tilde{\mathcal{T}}$. That is*

$$\boldsymbol{a}'[\boldsymbol{D}(\tilde{\mathcal{T}}) - \boldsymbol{D}(\mathcal{T})]\boldsymbol{a} \geq 0 \quad \forall\, \boldsymbol{a} \in \mathbb{R}^p$$

Because the dispersion matrices of unbiased estimators are the variance-covariance matrices, this condition is equivalent to

$$\boldsymbol{a}'\operatorname{Var}(\tilde{\mathcal{T}})\boldsymbol{a} - \boldsymbol{a}'\operatorname{Var}(\mathcal{T})\boldsymbol{a} \geq 0 \Rightarrow \operatorname{Var}(\boldsymbol{a}'\tilde{\mathcal{T}}) - \operatorname{Var}(\boldsymbol{a}'\mathcal{T}) \geq 0$$

Theorem 8 (Gauss-Markov)**.** *In the full-rank case (i.e.* $\operatorname{rank}(\boldsymbol{X}) = p$*) the minimum dispersion linear unbiased estimator of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}}_{LSE}$ with dispersion matrix $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. It is also called the* best linear unbiased estimator *or BLUE of $\boldsymbol{\beta}$.*

Proof. Any linear estimator of $\boldsymbol{\beta}$ can be written as $\boldsymbol{A}\mathcal{Y}$ for some $p \times n$ matrix $\boldsymbol{A}$. (That's what it means to be a linear estimator.) To be an unbiased linear estimator we must have

$$\boldsymbol{\beta} = E[\boldsymbol{A}\mathcal{Y}] = \boldsymbol{A}\,E[\mathcal{Y}] = \boldsymbol{A}\,\boldsymbol{X}\boldsymbol{\beta} \quad \forall\,\boldsymbol{\beta} \in \mathbb{R}^p \;\Rightarrow\; \boldsymbol{A}\boldsymbol{X} = \boldsymbol{I}_p$$

The variance-covariance matrix such a linear unbiased estimator, $\boldsymbol{A}\mathcal{Y}$, is

$$\operatorname{Var}(\boldsymbol{A}\mathcal{Y}) = \boldsymbol{A}\operatorname{Var}(\mathcal{Y})\boldsymbol{A}' = \boldsymbol{A}\sigma^2\boldsymbol{I}_n\boldsymbol{A}' = \sigma^2\boldsymbol{A}\boldsymbol{A}'.$$

Now we must show that

$$\mathrm{Var}(\boldsymbol{a}'\boldsymbol{A}\mathcal{Y}) - \mathrm{Var}(\boldsymbol{a}'\widehat{\boldsymbol{\beta}}_{LSE}) = \sigma^2\boldsymbol{a}'\left(\boldsymbol{A}\boldsymbol{A}' - (\boldsymbol{X}'\boldsymbol{X})^{-1}\right)\boldsymbol{a} \geq 0,\ \forall\ \boldsymbol{a} \in \mathbb{R}^p.$$

In other words, the symmetric matrix, $\left(\boldsymbol{A}\boldsymbol{A}' - (\boldsymbol{X}'\boldsymbol{X})^{-1}\right)$, must be positive semi-definite. Consider

$$\begin{aligned}\boldsymbol{A}\boldsymbol{A}' =&[\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'][\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']' \\ =&[\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'][\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']' + [\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'][(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}]'+ \\ &(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}[\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']' + [(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'][\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}] \\ =&[\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'][\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']' + (\boldsymbol{X}'\boldsymbol{X})^{-1},\end{aligned}$$

showing that $\boldsymbol{A}\boldsymbol{A}' - (\boldsymbol{X}'\boldsymbol{X})^{-1}$ is the positive semi-definite matrix $[\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'][\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']'$. Therefore $\widehat{\boldsymbol{\beta}}_{LSE}$ is the BLUE for $\boldsymbol{\beta}$. □

Corollary 7. *If* $\mathrm{rank}(\boldsymbol{X}) = p < n$, *the best linear unbiased estimator of* $\boldsymbol{a}'\boldsymbol{\beta}$ *is* $\boldsymbol{a}'\widehat{\boldsymbol{\beta}}_{LSE}$.

To extend the Gauss-Markov theorem to the rank-deficient case we must define

Definition 6 (Estimable linear function)**.** *An estimable linear function of the parameters* $\boldsymbol{\beta}$ *in the linear model,* $\mathcal{Y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$, *is any function of the form* $\boldsymbol{l}'\boldsymbol{\beta}$ *where* $\boldsymbol{l}$ *is in the row span of* $\boldsymbol{X}$. *That is,* $\boldsymbol{l}'\boldsymbol{\beta}$ *is estimable if and only if there exists* $\boldsymbol{c} \in \mathbb{R}^n$ *such that* $\boldsymbol{l} = \boldsymbol{X}'\boldsymbol{c}$.

The coefficients of the estimable functions form a $\mathrm{rank}(\boldsymbol{X}) = k$-dimensional linear subspace of $\mathbb{R}^p$. In the full-rank this subspace is all of $\mathbb{R}^p$ so any linear combination $\boldsymbol{l}'\boldsymbol{\beta}$ is estimable.

In the rank-deficient case (i.e. $\mathrm{rank}(\boldsymbol{X}) = k < p$), consider the singular value decomposition $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'$ with $\boldsymbol{D}$ a diagonal matrix having non-negative, non-increasing diagonal elements, the first k of which are positive and the last $p-k$ are zero. Let $\boldsymbol{U}_k$ be the first k columns of $\boldsymbol{U}$, $\boldsymbol{D}_k$ be the first k rows and k columns of $\boldsymbol{D}$, and $\boldsymbol{V}_k$ be the first k columns of $\boldsymbol{V}$. The coefficients $\boldsymbol{l}$ for an estimable linear function must lie in the column span of $\boldsymbol{V}_k$ because

$$\boldsymbol{l} = \boldsymbol{X}'\boldsymbol{c} = \boldsymbol{V}_k\underbrace{\boldsymbol{D}_k\boldsymbol{U}_k'\boldsymbol{c}}_{\boldsymbol{a}} = \boldsymbol{V}_k\boldsymbol{a}$$

We will write the $p \times (p-k)$ matrix formed by the last $p-k$ columns of $\boldsymbol{V}$ as $\boldsymbol{V}_{p-k}$ so that

$$\boldsymbol{\beta} = \boldsymbol{V}\boldsymbol{V}'\boldsymbol{\beta} = \begin{bmatrix}\boldsymbol{V}_k & \boldsymbol{V}_{(p-k)}\end{bmatrix}\begin{bmatrix}\boldsymbol{V}_k' \\ \boldsymbol{V}_{p-k}'\end{bmatrix}\boldsymbol{\beta} = \boldsymbol{V}_k\boldsymbol{\gamma} + \boldsymbol{V}_{p-k}\boldsymbol{\delta}$$

where $\boldsymbol{\gamma} = \boldsymbol{V}_k'\boldsymbol{\beta}$ and $\boldsymbol{\delta} = \boldsymbol{V}_{p-k}'\boldsymbol{\beta}$ are the estimable and inestimable parts of the parameter vector in the $\boldsymbol{V}$ basis.

Now any estimable function is of the form

$$\boldsymbol{l}'\boldsymbol{\beta} = \boldsymbol{a}'\boldsymbol{V}_k'\boldsymbol{\beta} = \boldsymbol{a}'\boldsymbol{\gamma} + \boldsymbol{0} = \boldsymbol{a}'\boldsymbol{\gamma},$$

where $\boldsymbol{\gamma}$ is the parameter in the full-rank model $\mathcal{Y} \sim \mathcal{N}(\boldsymbol{D}_k\boldsymbol{U}_k\boldsymbol{\gamma}, \sigma^2\boldsymbol{I}_n)$.

So anything we say about estimable functions of $\boldsymbol{\beta}$ can be transformed into a statement about $\boldsymbol{\gamma}$ in the full rank model and anything we say about the fitted values, $\boldsymbol{X}\boldsymbol{\beta}$, or the residuals can be expressed in terms of the full-rank $\boldsymbol{D}_k\boldsymbol{U}_k\boldsymbol{\gamma}$. In particular, the hat matrix, $\boldsymbol{H} = \boldsymbol{U}_k\boldsymbol{U}_k'$, and has $\mathrm{rank}(\boldsymbol{H}) = k$ and the projection into the orthogonal (residual) space is $\boldsymbol{I}_n - \boldsymbol{H}$.

Corollary 8 (Gauss-Markov extension to rank-deficient cases)**.** $\boldsymbol{l}'\widehat{\boldsymbol{\beta}}_{LSE} = \boldsymbol{a}'\widehat{\boldsymbol{\gamma}}_{LSE}$ *is the BLUE for any estimable linear function,* $\boldsymbol{l}'\boldsymbol{\beta}$*, of* $\boldsymbol{\beta}$*.*

Proof. By the Gauss-Markov theorem $\widehat{\gamma}_{LSE}$ is the BLUE for $\boldsymbol{\gamma}$ and $\boldsymbol{l}'\boldsymbol{\beta} = \boldsymbol{a}'\boldsymbol{\gamma}$ is a linear function of $\boldsymbol{\gamma}$. □

Theorem 9. *Suppose that* $k = rank(\boldsymbol{X}) \leq p$*. Then an unbiased estimator of* σ^2 *is*

$$S^2 = \frac{\|\mathcal{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2}{n-k} = \frac{\|\hat{\boldsymbol{\epsilon}}\|^2}{n-k} = \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{n-k}.$$

Proof. The simple proof is to observe that this estimator is the unbiased estimator of σ^2 for the full-rank version of the model, $\mathcal{Y} \sim \mathcal{N}(\boldsymbol{D}_k\boldsymbol{U}_k\boldsymbol{\gamma}, \sigma^2\boldsymbol{I}_n)$. □