Types of Problems–Review for Some of the QUAL problems

Notation: Let $\boldsymbol{A}^T = \boldsymbol{A}'$ be the transpose of $\boldsymbol{A}$.

0) Covariance and Expected Value = Mean, and the Multivariate Normal (MVN) Distribution:

Notation: Unless told otherwise, assume expectations exist and that conformable matrices and vectors are used.

The *population mean* of a random $n \times 1$ vector $\boldsymbol{x} = (x_1, ..., x_n)^T$ is $E(\boldsymbol{x}) = \boldsymbol{\mu} = (E(x_1), ..., E(x_n))^T$ and the $n \times n$ *population covariance matrix*
$\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}} = E(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T = (\sigma_{i,j})$ where $\text{Cov}(x_i, x_j) = \sigma_{i,j}$. The *population covariance matrix* of $\boldsymbol{x}$ with $\boldsymbol{y}$ is

$$\text{Cov}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\Sigma}_{\boldsymbol{x},\boldsymbol{y}} = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{y} - E(\boldsymbol{y}))^T].$$

If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $n \times 1$ random vectors, $\boldsymbol{a}$ a conformable constant vector, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are conformable constant matrices, then

$$E(\boldsymbol{X} + \boldsymbol{Y}) = E(\boldsymbol{X}) + E(\boldsymbol{Y}), \ E(\boldsymbol{a} + \boldsymbol{Y}) = \boldsymbol{a} + E(\boldsymbol{Y}), \ \& \ E(\boldsymbol{AXB}) = \boldsymbol{A}E(\boldsymbol{X})\boldsymbol{B}.$$

Also

$$\text{Cov}(\boldsymbol{a} + \boldsymbol{AX}) = \text{Cov}(\boldsymbol{AX}) = \boldsymbol{A}\text{Cov}(\boldsymbol{X})\boldsymbol{A}^T.$$

Note that $E(\boldsymbol{AY}) = \boldsymbol{A}E(\boldsymbol{Y})$ and $\text{Cov}(\boldsymbol{AY}) = \boldsymbol{A}\text{Cov}(\boldsymbol{Y})\boldsymbol{A}^T$.

If $\boldsymbol{X}$ $(m \times 1)$ and $\boldsymbol{Y}$ $(n \times 1)$ are random vectors, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are conformable constant matrices, then

$$\text{Cov}(\boldsymbol{AX}, \boldsymbol{BY}) = \boldsymbol{A}\text{Cov}(\boldsymbol{X}, \boldsymbol{Y})\boldsymbol{B}^T.$$

If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $E(\boldsymbol{X}) = \boldsymbol{\mu}$, $\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$, and $m_{\boldsymbol{X}}(\boldsymbol{t}) = \exp(\boldsymbol{t}^T\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t})$.

If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and if $\boldsymbol{A}$ is a $q \times p$ matrix, then $\boldsymbol{AX} \sim N_q(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^T)$. If $\boldsymbol{a}$ $(p \times 1)$ and $\boldsymbol{b}$ $(q \times 1)$ are constant vectors, then $\boldsymbol{X} + \boldsymbol{a} \sim N_p(\boldsymbol{\mu} + \boldsymbol{a}, \boldsymbol{\Sigma})$ and $\boldsymbol{AX} + \boldsymbol{b} \sim N_q(\boldsymbol{A\mu} + \boldsymbol{b}, \boldsymbol{A\Sigma A}^T)$.

$$\text{Let } \ \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \ \text{and } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

The conditional distribution of a MVN is MVN. If $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the conditional distribution of $\boldsymbol{X}_1$ given that $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal with mean $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2)$ and covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. That is,

$$\boldsymbol{X}_1 | \boldsymbol{X}_2 = \boldsymbol{x}_2 \sim N_q(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

Notation:

$$\boldsymbol{X}_1 | \boldsymbol{X}_2 \sim N_q(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{X}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}).$$

1) Projection Matrices, Generalized Inverses, and the Column Space $C(\boldsymbol{X})$**:**

Let $\boldsymbol{A} = [\boldsymbol{a}_1 \ \boldsymbol{a}_2 \ ... \ \boldsymbol{a}_m]$ be an $n \times m$ matrix. The space spanned by the columns of $\boldsymbol{A}$ = **column space** of $\boldsymbol{A} = C(\boldsymbol{A})$.

Let $\boldsymbol{X} = [\boldsymbol{v}_1 \ \boldsymbol{v}_2 \ ... \ \boldsymbol{v}_p]$ be an $n \times p$ matrix. Then $C(\boldsymbol{X}) = \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y} = \boldsymbol{X\beta} \text{ for some } \boldsymbol{\beta} \in \mathbb{R}^p\}$.

One way to show $C(\boldsymbol{A}) = C(\boldsymbol{B})$ is to show that i) $\boldsymbol{Ax} = \boldsymbol{By} \in C(\boldsymbol{B})$ and ii) $\boldsymbol{By} = \boldsymbol{Ax} \in C(\boldsymbol{A})$.

The *null space* of $\boldsymbol{A} = N(\boldsymbol{A}) = \{\boldsymbol{x} : \boldsymbol{Ax} = \boldsymbol{0}\} =$ kernel of $\boldsymbol{A}$. The subspace $V^\perp = \{\boldsymbol{y} \in \mathbb{R}^k : \boldsymbol{y} \perp V\}$ is the *orthogonal complement of* V.

$N(\boldsymbol{A}^T) = [C(\boldsymbol{A})]^\perp$, so $N(\boldsymbol{A}) = [C(\boldsymbol{A}^T)]^\perp$.

A **generalized inverse** of an $m \times n$ matrix $\boldsymbol{A}$ is any $n \times m$ matrix $\boldsymbol{A}^-$ satisfying $\boldsymbol{AA}^-\boldsymbol{A} = \boldsymbol{A}$. Other names are conditional inverse, pseudo inverse, g-inverse, and p-inverse. Usually a generalized inverse is not unique, but if $\boldsymbol{A}^{-1}$ exists, then $\boldsymbol{A}^- = \boldsymbol{A}^{-1}$ is unique. Notation: $\boldsymbol{G} := \boldsymbol{A}^-$ means $\boldsymbol{G}$ is a generalized inverse of $\boldsymbol{A}$.

Let V be a subspace of $\mathbb{R}^k$. Then every $\boldsymbol{y} \in \mathbb{R}^k$ can be expressed uniquely as $\boldsymbol{y} = \boldsymbol{w}+\boldsymbol{z}$ where $\boldsymbol{w} \in V$ and $\boldsymbol{z} \in V^\perp$.

Let $\boldsymbol{X} = [\boldsymbol{v}_1 \ \boldsymbol{v}_2 \ ... \ \boldsymbol{v}_p]$ be $n \times p$, and let $V = C(\boldsymbol{X}) = \text{span}(\boldsymbol{v}_1, ..., \boldsymbol{v}_p)$. Then the $n \times n$ matrix $\boldsymbol{P}_V = \boldsymbol{P}_{\boldsymbol{X}}$ is a **projection matrix** on $C(\boldsymbol{X})$ if $\boldsymbol{P}_{\boldsymbol{X}} \ \boldsymbol{y} = \boldsymbol{w} \ \forall \ \boldsymbol{y} \in \mathbb{R}^n$. (Here $\boldsymbol{y} = \boldsymbol{w} + \boldsymbol{z} = \boldsymbol{w}_{\boldsymbol{y}} + \boldsymbol{z}_{\boldsymbol{y}}$, so $\boldsymbol{w}$ depends on $\boldsymbol{y}$.)

Projection Matrix Theorem: a) $\boldsymbol{P}_{\boldsymbol{X}}$ is unique.
b) $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$ where $(\boldsymbol{X}^T\boldsymbol{X})^-$ is any generalized inverse of $\boldsymbol{X}^T\boldsymbol{X}$.
c) $\boldsymbol{A}$ is a projection matrix on $C(\boldsymbol{A})$ iff $\boldsymbol{A}$ is symmetric and idempotent. Hence $\boldsymbol{P}_{\boldsymbol{X}}$ is a projection matrix on $C(\boldsymbol{P}_{\boldsymbol{X}}) = C(\boldsymbol{X})$.
d) $\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}}$ is the projection matrix on $[C(\boldsymbol{X})]^\perp$.
e) $\boldsymbol{A} = \boldsymbol{P}_{\boldsymbol{X}}$ iff i) $\boldsymbol{y} \in C(\boldsymbol{X})$ implies $A\boldsymbol{y} = \boldsymbol{y}$ and ii) $\boldsymbol{y} \perp C(\boldsymbol{X})$ implies $\boldsymbol{Ay} = \boldsymbol{0}$.
f) $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{X} = \boldsymbol{X}$, and $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{W} = \boldsymbol{W}$ if each column of $\boldsymbol{W} \in C(\boldsymbol{X})$.
g) $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{v}_i = \boldsymbol{v}_i$.
h) If $C(\boldsymbol{X}_R)$ is a subspace of $C(\boldsymbol{X})$, then $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{X}_R} = \boldsymbol{P}_{\boldsymbol{X}_R}\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{P}_{\boldsymbol{X}_R}$.
i) $rank(\boldsymbol{P}_{\boldsymbol{X}}) = tr(\boldsymbol{P}_{\boldsymbol{X}}) = rank(\boldsymbol{X})$.

Note that $\boldsymbol{P}$ is a projection matrix iff $\boldsymbol{P}$ is symmetric and idempotent. Partition $\boldsymbol{X}$ as $\boldsymbol{X} = [\boldsymbol{X}_1 \ \ \boldsymbol{X}_2]$, let $\boldsymbol{P}$ be the projection matrix for $\mathcal{C}(\boldsymbol{X})$ and let $\boldsymbol{P}_1$ be the projection matrix for $\mathcal{C}(\boldsymbol{X}_1)$. Since $\mathcal{C}(\boldsymbol{P}_1) = \mathcal{C}(\boldsymbol{X}_1) \subseteq \mathcal{C}(\boldsymbol{X})$, $\boldsymbol{PP}_1 = \boldsymbol{P}_1$. Hence $\boldsymbol{P}_1\boldsymbol{P} = (\boldsymbol{PP}_1)' = \boldsymbol{P}_1' = \boldsymbol{P}_1$.

1a): Given small $\boldsymbol{X}$, be able to find the projection matrix $\boldsymbol{P}$ for $C(\boldsymbol{X})$.

1b): Given small $\boldsymbol{X}$, be able to find rank($\boldsymbol{X}$), a basis for $C(\boldsymbol{X})$, and $[\mathcal{C}(\boldsymbol{X})]^\perp =$ nullspace of $\boldsymbol{X}^T$.

1c): Be able to show that $\boldsymbol{G} := \boldsymbol{A}^-$.

2) Quadratic Forms $\boldsymbol{Y}'\boldsymbol{AY}$ and terms like $\boldsymbol{AY}$:

The matrix $\boldsymbol{A}$ in a quadratic form $\boldsymbol{x}^T\boldsymbol{Ax}$ is **symmetric.** $\boldsymbol{A}$ is **positive definite** ($\boldsymbol{A} > 0$) if $\boldsymbol{x}^T\boldsymbol{Ax} > 0 \ \forall \ \boldsymbol{x} \neq \boldsymbol{0}$. $\boldsymbol{A}$ is **positive semidefinite** ($\boldsymbol{A} \geq 0$) if $\boldsymbol{x}^T\boldsymbol{Ax} \geq 0 \ \forall \ \boldsymbol{x}$.

Let $\boldsymbol{A}$ be symmetric. If $\boldsymbol{A} \geq 0$ then the eigenvalues of $\boldsymbol{A}$ are real and nonnegative. If $\boldsymbol{A} \geq 0$, let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$. If $\boldsymbol{A} > 0$, then $\lambda_n > 0$.

Theorem 2.5 (Seber and Lee Th. 1.5) expected value of a quadratic form: Let $\boldsymbol{X}$ be a random vector with $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$. Then

$$E(\boldsymbol{X}^T\boldsymbol{AX}) = tr(\boldsymbol{A\Sigma}) + [E(\boldsymbol{X})]^T\boldsymbol{A}E(\boldsymbol{X}) = tr(\boldsymbol{A\Sigma}) + \boldsymbol{\mu}^T\boldsymbol{A\mu}.$$

Theorems 2.6 and 2.7: If $\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{BY}$, then $f(\boldsymbol{AY}) \perp\!\!\!\perp g(\boldsymbol{BY})$ where f and g are functions (such that $f(\boldsymbol{AY})$ only depends on $\boldsymbol{A}$ and $\boldsymbol{AY}$ and $g(\boldsymbol{BY})$ only depends on $\boldsymbol{B}$ and $\boldsymbol{BY}$). Note that $\boldsymbol{Y}'\boldsymbol{AY} = \boldsymbol{Y}'\boldsymbol{A}'\boldsymbol{A}^-\boldsymbol{AY} = f(\boldsymbol{AY})$ (for a quadratic form $\boldsymbol{A}$ is symmetric), $\boldsymbol{Y}'(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y} = \|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}\|^2$, and $\boldsymbol{Y}'\boldsymbol{PY} = \|\boldsymbol{PY}\|^2$ where the squared Euclidean norm $\|\boldsymbol{Z}\|^2 = \boldsymbol{Z}'\boldsymbol{Z}$.

Theorem 2.8. Let $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. a) Let $\boldsymbol{u} = \boldsymbol{AY}$ and $\boldsymbol{w} = \boldsymbol{BY}$. Then $\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{BY}$ iff $\text{Cov}(\boldsymbol{u}, \boldsymbol{w}) = \boldsymbol{A\Sigma B}^T = \boldsymbol{0}$ iff $\boldsymbol{B\Sigma A}^T = \boldsymbol{0}$. Note that if $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_n$, then $\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{BY}$ if $\boldsymbol{AB}^T = \boldsymbol{0}$ if $\boldsymbol{BA}^T = \boldsymbol{0}$.

b) If $\boldsymbol{A}$ is a symmetric $n \times n$ matrix, and $\boldsymbol{B}$ is an $m \times n$ matix, then $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{BY}$ iff $\boldsymbol{A\Sigma B}^T = \boldsymbol{0}$ iff $\boldsymbol{B\Sigma A} = \boldsymbol{0}$.

Craig's Theorem: Let $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

a) If $\boldsymbol{\Sigma} > 0$, then $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{BY}$ iff $\boldsymbol{A\Sigma B} = \boldsymbol{0}$ iff $\boldsymbol{B\Sigma A} = \boldsymbol{0}$.

b) If $\boldsymbol{\Sigma} \geq 0$, then $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{BY}$ if $\boldsymbol{A\Sigma B} = \boldsymbol{0}$ (or if $\boldsymbol{B\Sigma A} = \boldsymbol{0}$).

c) If $\boldsymbol{\Sigma} \geq 0$, then $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{BY}$ iff

$(*)$ $\boldsymbol{\Sigma A \Sigma B \Sigma} = \boldsymbol{0}, \boldsymbol{\Sigma A \Sigma B \mu} = \boldsymbol{0}, \boldsymbol{\Sigma B \Sigma A \mu} = \boldsymbol{0}$, and $\boldsymbol{\mu}^T\boldsymbol{A\Sigma B\mu} = 0$.

Note that if $\boldsymbol{A\Sigma B} = \boldsymbol{0}$, then $(*)$ holds.

Theorem 2.13. If $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} > 0$, then $\boldsymbol{Y}^T\boldsymbol{AY} \sim \chi^2(\text{rank}(\boldsymbol{A}), \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{A\mu}/2)$ iff $\boldsymbol{A\Sigma}$ is idempotent.

Remark 1: If the theorem is for $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{I})$ and $\boldsymbol{Z} \sim N_n(E(\boldsymbol{Z}), \sigma^2\boldsymbol{I})$, then use $\boldsymbol{Y} = \boldsymbol{Z}/\sigma \sim N_n(\boldsymbol{\mu} = E(\boldsymbol{Z})/\sigma, \boldsymbol{I})$.

Theorem 2.14. Let $\boldsymbol{A} = \boldsymbol{A}^T$ be symmetric.

a) If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is a projection matrix, then $\boldsymbol{Y}^T\boldsymbol{Y} \sim \chi^2(\text{rank}(\boldsymbol{\Sigma}))$ where $\text{rank}(\boldsymbol{\Sigma}) = tr(\boldsymbol{\Sigma})$.

b) If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{I})$, then $\boldsymbol{Y}^T\boldsymbol{AY} \sim \chi^2_r$ iff $\boldsymbol{A}$ is idempotent with $\text{rank}(\boldsymbol{A}) = tr(\boldsymbol{A}) = r$.

c) Let $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Then

$$\frac{\boldsymbol{Y}^T\boldsymbol{AY}}{\sigma^2} \sim \chi^2_r \ \text{ or } \ \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{AY} \sim \sigma^2\, \chi^2_{\mathrm{r}}$$

iff $\boldsymbol{A}$ is idempotent of rank r.

d) If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} > 0$, then $\boldsymbol{Y}^T\boldsymbol{AY} \sim \chi^2_r$ iff $\boldsymbol{A\Sigma}$ is idempotent with $\text{rank}(\boldsymbol{A}) = r = \text{rank}(\boldsymbol{A\Sigma})$.

e) If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ then $\dfrac{\boldsymbol{Y}^T\boldsymbol{Y}}{\sigma^2} \sim \chi^2\left(n, \dfrac{\boldsymbol{\mu}^T\boldsymbol{\mu}}{2\sigma^2}\right)$.

f) If $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{I})$ then $\boldsymbol{Y}^T\boldsymbol{AY} \sim \chi^2(r, \boldsymbol{\mu}^T\boldsymbol{A\mu}/2)$ iff $\boldsymbol{A}$ is idempotent with $\text{rank}(\boldsymbol{A}) = tr(\boldsymbol{A}) = r$.

g) If $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2\boldsymbol{I})$ then $\dfrac{\boldsymbol{Y}^T\boldsymbol{AY}}{\sigma^2} \sim \chi^2\left(r, \dfrac{\boldsymbol{\mu}^T\boldsymbol{A\mu}}{2\sigma^2}\right)$ iff $\boldsymbol{A}$ is idempotent with $\text{rank}(\boldsymbol{A}) = tr(\boldsymbol{A}) = r$.

3) MLE: The following problem is typical. It is assumed than $\sigma > 0$ and $\boldsymbol{\beta} \in \mathbb{R}^p$.

Suppose $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \epsilon_i$ with $Q(\boldsymbol{\beta}) \geq 0$. Let c_n be a constant that does not depend on $\boldsymbol{\beta}$ or σ^2. Suppose the likelihood function is

$$L(\boldsymbol{\beta}, \sigma^2) = c_n \, \frac{1}{\sigma^n} \exp\left(\frac{-1}{2\sigma^2} Q(\boldsymbol{\beta})\right).$$

a) Suppose that $\hat{\boldsymbol{\beta}}_Q$ minimizes $Q(\boldsymbol{\beta})$. Show that $\hat{\boldsymbol{\beta}}_Q$ is the MLE of $\boldsymbol{\beta}$.

b) Then find the MLE $\hat{\sigma}^2$ of σ^2.

Solution: a) For fixed $\sigma > 0$, $L(\boldsymbol{\beta}, \sigma^2)$ is maximized by minimizing $Q(\boldsymbol{\beta}) \geq 0$. So $\hat{\boldsymbol{\beta}}_Q$ maximizes $L(\boldsymbol{\beta}, \sigma^2)$ regardless of the value of $\sigma^2 > 0$. So $\hat{\boldsymbol{\beta}}_Q$ is the MLE.

b) Let $Q = Q(\hat{\boldsymbol{\beta}}_Q)$. Then the MLE $\hat{\sigma}^2$ is found by maximizing the profile likelihood, $L_p(\sigma^2) = L(\hat{\boldsymbol{\beta}}_Q, \sigma^2) = c_n \dfrac{1}{\sigma^n} \exp\left(\dfrac{-1}{2\sigma^2} Q\right)$. Let $\tau = \sigma^2$. The $L_p(\tau) = c_n \dfrac{1}{\tau^{n/2}} \exp\left(\dfrac{-1}{2\tau} Q\right)$, and the log profile likelihood $\log L_p(\tau) = d - \dfrac{n}{2} \log(\tau) - \dfrac{Q}{2\tau}$. Thus

$$\frac{d \ \log L_p(\tau)}{d\tau} = \frac{-n}{2\tau} + \frac{Q}{2\tau^2} \stackrel{set}{=} 0$$

or $-n\tau + Q = 0$ or $\hat{\tau} = \hat{\sigma}^2 = Q/n$, unique. Then

$$\frac{d^2 \ \log L_p(\tau)}{d\tau^2} = \frac{n}{2\tau^2} - \left.\frac{2Q}{2\tau^3}\right|_{\hat{\tau}} = \frac{n}{2\tau^2} - \frac{2n\hat{\tau}}{2\hat{\tau}^3} = \frac{-n}{2\hat{\tau}^2} < 0$$

which proves that $\hat{\sigma}^2$ is the MLE of σ^2.

Note: A negative second derivative shows that $\hat{\sigma}^2$ is a local max. The result that $\hat{\sigma}^2$ was the unique solution to setting the first derivative of the profile likelihood equal to zero makes $\hat{\sigma}^2$ the global max.

Common errors: Students use $Q(\boldsymbol{\beta})$ instead of $Q(\hat{\boldsymbol{\beta}})$ in the profile likelihood. Students forget to write the word "unique."

Variant: $Q(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X\beta}\|^2 = (\boldsymbol{Y} - \boldsymbol{X\beta})^T(\boldsymbol{Y} - \boldsymbol{X\beta})$ is the least squares criterion. Recognize that $Q(\boldsymbol{\beta})$ is minimized by $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS}$, and proceed as in the above problem.

Note: If the e_i are iid $N(0, \sigma^2)$ and least squares is used, then the MLE of $\boldsymbol{\beta}$ is the least squares estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ and the MLE of σ^2 is

$$\hat{\sigma}^2_M = \frac{n-p}{n} MSE = \frac{1}{n} \sum_{i=1}^n r_i^2.$$

4) LS Estimators for $p \leq 2$:

Given a least squares model with $p \leq 2$, derive or find the least squares estimator $\hat{\boldsymbol{\beta}}$.

Tip: If the LS model is $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ for $i = 1, ..., n$, then the LS criterion is $Q(\boldsymbol{\beta}) = \sum_{i=1}^n (Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 = \sum_{i=1}^n r_i^2(\boldsymbol{\beta})$.

To derive the LS estimator, let $Q(\beta_1, \beta_2) = \sum_{i=1}^n (Y_i - \beta_1 - \beta_2 x_i)^2$ be the residual sum of squares where β_i vary on $\mathbb{R}$. Take the partial derivatives, set them to 0, and solve for the least squares estimators. If $p = 2$, we will assume 2nd derivatives do not need to be taken. If $p = 1$, show the solution is unique and show that the second derivative evaluated at $\hat{\beta}$ is positive. The β_i could be replaced by other symbols such as η_i.

Location model: $Y_i = \beta + e_i$ or $\boldsymbol{Y} = \boldsymbol{1}\beta + \boldsymbol{e}$. The parameter β could be replaced with μ or θ. The LS criterion $Q(\beta) = \sum_{i=1}^n (Y_i - \beta)^2$, and $\hat{\beta} = \overline{Y}$, the sample mean.

$$\text{Proof} : \frac{dQ(\beta)}{d\beta} = -2 \sum_{i=1}^n (Y_i - \beta).$$

Setting the derivative equal to 0 and calling the unique solution $\hat{\beta}$ gives $\sum_{i=1}^n Y_i = n\hat{\beta}$ or $\hat{\beta} = \overline{Y}$. The second derivative

$$\frac{d^2Q(\beta)}{d\beta^2} = 2n > 0,$$

hence $\hat{\beta}$ is the global minimizer.

Simple linear regression (SLR): $Y_i = \beta_1 + x_i\beta_2 + e_i$ or $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ where $\boldsymbol{X} = [\boldsymbol{1} \;\; \boldsymbol{x}]$ and $\boldsymbol{\beta} = (\beta_1 \; \beta_2)^T$. The LS criterion $Q(\beta_1, \beta_2) = \sum_{i=1}^n (Y_i - \beta_1 - x_i\beta_2)^2$.

The least squares (OLS) line is $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$ where the slope

$$\hat{\beta}_2 \equiv \hat{\beta} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})Y_i}{\sum_{j=1}^n (X_j - \overline{X})^2} = \sum_{i=1}^n k_i Y_i$$

with

$$k_i = \frac{X_i - \overline{X}}{\sum_{j=1}^n (X_j - \overline{X})^2} = \frac{X_i - \overline{X}}{(n-1)S_X^2},$$

and the intercept $\hat{\beta}_1 \equiv \hat{\alpha} = \overline{Y} - \hat{\beta}_2 \overline{X}$.

By the **chain rule,**

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^n (Y_i - \beta_1 - \beta_2 X_i)$$

and

$$\frac{\partial^2 Q}{\partial \beta_1^2} = 2n.$$

Similarly,

$$\frac{\partial Q}{\partial \beta_2} = -2\sum_{i=1}^n X_i(Y_i - \beta_1 - \beta_2 X_i)$$

and

$$\frac{\partial^2 Q}{\partial \beta_2^2} = 2\sum_{i=1}^n X_i^2.$$

Setting the first partial derivatives to zero and calling the solutions $\hat{\beta}_1$ and $\hat{\beta}_2$ shows that the OLS estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ satisfy the **normal equations**:

$$\sum_{i=1}^n Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^n X_i \;\; \text{and}$$

$$\sum_{i=1}^n X_i Y_i = \hat{\beta}_1 \sum_{i=1}^n X_i + \hat{\beta}_2 \sum_{i=1}^n X_i^2.$$

The first equation gives $\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}$.

There are several equivalent formulas for the slope $\hat{\beta}_2$.

$$\hat{\beta}_2 \equiv \hat{\beta} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{\sum_{i=1}^n X_i Y_i - \frac{1}{n}(\sum_{i=1}^n X_i)(\sum_{i=1}^n Y_i)}{\sum_{i=1}^n X_i^2 - \frac{1}{n}(\sum_{i=1}^n X_i)^2}$$

$$= \frac{\sum_{i=1}^n (X_i - \overline{X}) Y_i}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{\sum_{i=1}^n X_i Y_i - n \overline{X} \ \overline{Y}}{\sum_{i=1}^n X_i^2 - n (\overline{X})^2} = \hat{\rho} s_Y / s_X.$$

Here the sample correlation $\hat{\rho} \equiv \hat{\rho}(X, Y) = \text{corr}(X, Y) =$

$$\frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{(n-1) s_X s_Y} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}}$$

where the sample standard deviation

$$s_W = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (W_i - \overline{W})^2}$$

for $W = X, Y$. Notice that the term $n-1$ that occurs in the denominator of $\hat{\rho}, s_Y^2$, and s_X^2 can be replaced by n as long as n is used in all 3 quantities.

SLR through the origin: $Y_i = x_i \beta + e_i$ or $Y = \boldsymbol{x} \beta + \boldsymbol{e}$. The LS criterion $Q(\beta) = \sum_{i=1}^n (Y_i - x_i \beta)^2$, and $\hat{\beta} = \sum_{i=1}^n x_i Y_i / \sum_{i=1}^n x_i^2$.

Known intercept: $Y_i = a + x_i \beta + e_i$ where the intercept a is known. $Q(\beta) = \sum_{i=1}^n (Y_i - a - x_i \beta)^2$.

Known slope: $Y_i = \beta + x_i b + e_i$ where the slope b is known. $Q(\beta) = \sum_{i=1}^n (Y_i - \beta - x_i b)^2$. Here, β may be replaced by α.

5) WLS:

For the WLS model $Y|\boldsymbol{x} = \boldsymbol{x}^T \boldsymbol{\beta} + e$ where the e_i are independent wtih $E(e_i) = 0$ and $V(e_i) = \sigma_i^2$. Hence $\boldsymbol{Y} = \boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $E(\boldsymbol{e}) = \boldsymbol{0}$ and $Cov(\boldsymbol{e}) = diag(\sigma_i^2)$.

An alternative model is $Y|\boldsymbol{x}^T \boldsymbol{\beta} = \boldsymbol{x}^T \boldsymbol{\beta} + \boldsymbol{u}$ where the u_i are independent with $E(u_i) = 0$ and $V(u_i) = \tau_i^2$. Hence $\boldsymbol{Y} = \boldsymbol{Y}|\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$ where $E(\boldsymbol{u}) = \boldsymbol{0}$ and $Cov(\boldsymbol{u}) = diag(\tau_i^2)$.

6) Non-full rank linear models:

The **nonfull rank linear model** is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ has rank $r < p \leq n$, and $\boldsymbol{X}$ is an $n \times p$ matrix.

Theorem 3.1. i) $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^- \boldsymbol{X}^T$ is the unique projection matrix on $C(\boldsymbol{X})$ and does not depend on the generalized inverse $(\boldsymbol{X}^T\boldsymbol{X})^-$.

ii) $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^- \boldsymbol{X}^T \boldsymbol{Y}$ does depend on $(\boldsymbol{X}^T\boldsymbol{X})^-$ and is not unique.

iii) $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{P}\boldsymbol{Y}$, $\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$ and $RSS = \boldsymbol{r}^T \boldsymbol{r}$ are unique and so do not depend on $(\boldsymbol{X}^T\boldsymbol{X})^-$.

iv) $\hat{\boldsymbol{\beta}}$ is a solution to the *normal equations*: $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$.

v) Rank$(\boldsymbol{P}) = r$ and rank$(\boldsymbol{I} - \boldsymbol{P}) = n - r$.

vi) If Cov$(\boldsymbol{Y})$ = Cov$(\boldsymbol{e}) = \sigma^2 \boldsymbol{I}$, then $MSE = \dfrac{RSS}{n-r} = \dfrac{\boldsymbol{r}^T\boldsymbol{r}}{n-r}$ is an unbiased estimator of σ^2.

vii) Let the columns of $\boldsymbol{X}_1$ form a basis for $C(\boldsymbol{X})$. For example, take r linearly independent columns of $\boldsymbol{X}$ to form $\boldsymbol{X}_1$. Then $\boldsymbol{P} = \boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T$.

7) Estimability and the Gauss Markov Theorem:

Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be constant vectors. Then $\boldsymbol{a}^T\boldsymbol{\beta}$ is **estimable** if there exists a linear unbiased estimator $\boldsymbol{b}^T\boldsymbol{Y}$ so $E(\boldsymbol{b}^T\boldsymbol{Y}) = \boldsymbol{a}^T\boldsymbol{\beta}$. Also, $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable iff $\boldsymbol{a}^T = \boldsymbol{b}^T\boldsymbol{X}$ iff $\boldsymbol{a} = \boldsymbol{X}^T\boldsymbol{b}$ iff $\boldsymbol{a} \in C(\boldsymbol{X}^T)$.

The linear estimator $\boldsymbol{a}^T\boldsymbol{Y}$ of $\boldsymbol{c}^T\boldsymbol{\theta}$ is the best linear unbiased estimator (BLUE) of $\boldsymbol{c}^T\boldsymbol{\theta}$ if $E(\boldsymbol{a}^T\boldsymbol{Y}) = \boldsymbol{c}^T\boldsymbol{\theta}$, and if for any other unbiased linear estimator $\boldsymbol{b}^T\boldsymbol{Y}$ of $\boldsymbol{c}^T\boldsymbol{\theta}$, $V(\boldsymbol{a}^T\boldsymbol{Y}) \leq V(\boldsymbol{b}^T\boldsymbol{Y})$. Note that $E(\boldsymbol{b}^T\boldsymbol{Y}) = \boldsymbol{c}^T\boldsymbol{\theta}$.

The next theorem shows that the least squares estimator of an estimable function $\boldsymbol{a}^T\boldsymbol{\beta}$ is $\boldsymbol{a}^T\hat{\boldsymbol{\beta}} = \boldsymbol{b}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{b}^T\boldsymbol{P}\boldsymbol{Y}$. Note that $\boldsymbol{b}^T\boldsymbol{Y}$ is also an unbiased estimator of $\boldsymbol{a}^T\boldsymbol{\beta}$ since $E(\boldsymbol{b}^T\boldsymbol{Y}) = \boldsymbol{b}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{a}^T\boldsymbol{\beta}$.

Theorem 3.2 (see Seber and Lee Th 3.2) Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where where $\boldsymbol{X}$ has rank $r \leq p \leq n$, $E(\boldsymbol{e}) = \boldsymbol{0}$, and $\text{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$.

a) The quantity $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable iff $\boldsymbol{a}^T = \boldsymbol{b}^T\boldsymbol{X}$ iff $\boldsymbol{a} = \boldsymbol{X}^T\boldsymbol{b}$ (for some constant vector $\boldsymbol{b}$) iff $\boldsymbol{a} \in C(\boldsymbol{X}^T)$.

b) Let $\hat{\boldsymbol{\theta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta}$. Suppose there exists a constant vector $\boldsymbol{c}$ such that $E(\boldsymbol{c}^T\hat{\boldsymbol{\theta}}) = \boldsymbol{c}^T\boldsymbol{\theta}$. Then among the class of linear unbiased estimators of $\boldsymbol{c}^T\boldsymbol{\theta}$, the least squares estimator $\boldsymbol{c}^T\hat{\boldsymbol{\theta}}$ is the unique BLUE.

c) **Gauss Markov Theorem**: If $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable and a least squares estimator $\hat{\boldsymbol{\beta}}$ is any solution to the normal equations $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$, then $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is the unique BLUE of $\boldsymbol{a}^T\boldsymbol{\beta}$.

Proof: a) If $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable, then $\boldsymbol{a}^T\boldsymbol{\beta} = E(\boldsymbol{b}^T\boldsymbol{Y}) = \boldsymbol{b}^T\boldsymbol{X}\boldsymbol{\beta}$ for all $\boldsymbol{\beta} \in \mathbb{R}^p$. Thus $\boldsymbol{a}^T = \boldsymbol{b}^T\boldsymbol{X}$ or $\boldsymbol{a} = \boldsymbol{X}^T\boldsymbol{b}$. Hence $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable iff $\boldsymbol{a}^T = \boldsymbol{b}^T\boldsymbol{X}$ iff $\boldsymbol{a} = \boldsymbol{X}^T\boldsymbol{b}$ iff $\boldsymbol{a} \in C(\boldsymbol{X}^T)$.

b) Since $\hat{\boldsymbol{\theta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{P}\boldsymbol{Y}$, it follows that $E(\boldsymbol{c}^T\hat{\boldsymbol{\theta}}) = E(\boldsymbol{c}^T\boldsymbol{P}\boldsymbol{Y}) = \boldsymbol{c}^T\boldsymbol{P}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{c}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{c}^T\boldsymbol{\theta}$. Thus $\boldsymbol{c}^T\hat{\boldsymbol{\theta}} = \boldsymbol{c}^T\boldsymbol{P}\boldsymbol{Y} = (\boldsymbol{P}\boldsymbol{c})^T\boldsymbol{Y}$ is a linear unbiased estimator of $\boldsymbol{c}^T\boldsymbol{\theta}$. Let $\boldsymbol{d}^T\boldsymbol{Y}$ be any other linear unbiased estimator of $\boldsymbol{c}^T\boldsymbol{\theta}$. Hence $E(\boldsymbol{d}^T\boldsymbol{Y}) = \boldsymbol{d}^T\boldsymbol{\theta} = \boldsymbol{c}^T\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in C(\boldsymbol{X})$. So $(\boldsymbol{c}-\boldsymbol{d})^T\boldsymbol{\theta} = 0$ for all $\boldsymbol{\theta} \in C(\boldsymbol{X})$. Hence $(\boldsymbol{c}-\boldsymbol{d}) \in [C(\boldsymbol{X})]^{\perp}$ and $\boldsymbol{P}(\boldsymbol{c}-\boldsymbol{d}) = \boldsymbol{0}$, or $\boldsymbol{P}\boldsymbol{c} = \boldsymbol{P}\boldsymbol{d}$. Thus $V(\boldsymbol{c}^T\hat{\boldsymbol{\theta}}) = V(\boldsymbol{c}^T\boldsymbol{P}\boldsymbol{Y}) = V(\boldsymbol{d}^T\boldsymbol{P}\boldsymbol{Y}) = \sigma^2\boldsymbol{d}^T\boldsymbol{P}^T\boldsymbol{P}\boldsymbol{d} = \sigma^2\boldsymbol{d}^T\boldsymbol{P}\boldsymbol{d}$. Then $V(\boldsymbol{d}^T\boldsymbol{Y}) - V(\boldsymbol{c}^T\hat{\boldsymbol{\theta}}) = V(\boldsymbol{d}^T\boldsymbol{Y}) - V(\boldsymbol{d}^T\boldsymbol{P}\boldsymbol{Y}) = \sigma^2[\boldsymbol{d}^T\boldsymbol{d} - \boldsymbol{d}^T\boldsymbol{P}\boldsymbol{d}] = \sigma^2\boldsymbol{d}^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{d} = \sigma^2\boldsymbol{d}^T(\boldsymbol{I}_n - \boldsymbol{P})^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{d} = \boldsymbol{g}^T\boldsymbol{g} \geq 0$ with equality iff $\boldsymbol{g} = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{d} = \boldsymbol{0}$, or $\boldsymbol{d} = \boldsymbol{P}\boldsymbol{d} = \boldsymbol{P}\boldsymbol{c}$. Thus $\boldsymbol{c}^T\hat{\boldsymbol{\theta}}$ has minimum variance and is unique.

c) Since $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable, $\boldsymbol{a}^T\hat{\boldsymbol{\beta}} = \boldsymbol{b}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}$. Then $\boldsymbol{a}^T\hat{\boldsymbol{\beta}} = \boldsymbol{b}^T\hat{\boldsymbol{\theta}}$ is the unique BLUE of $\boldsymbol{a}^T\boldsymbol{\beta} = \boldsymbol{b}^T\boldsymbol{\theta}$ by b).

Gauss Markov Theorem-Full Rank Case: Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ is full rank, $E(\boldsymbol{e}) = \boldsymbol{0}$, and $\text{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$. Then $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is the unique BLUE of $\boldsymbol{a}^T\boldsymbol{\beta}$ for every constant $p \times 1$ vector $\boldsymbol{a}$.

Notation: $\boldsymbol{\beta}$ is "estimable" by $\hat{\boldsymbol{\beta}}$ for the full rank model, but not for the non-full rank model.

8) Hypothesis Testing:

Theorem 2.16. Let $\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\eta} \in C(\boldsymbol{X})$ where $Y_i = \boldsymbol{x}_i^T\boldsymbol{\eta} + r_i(\boldsymbol{\eta})$ and the residual $r_i(\boldsymbol{\eta})$ depends on $\boldsymbol{\eta}$. The **least squares estimator** $\hat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\eta} \in \mathbb{R}^p$ that minimizes the **least squares criterion**
$\sum_{i=1}^n r_i^2(\boldsymbol{\eta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}\|^2$.

LS CLT (Least Squares Central Limit Theorem): Consider the MLR model $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ and assume that the zero mean errors are iid with $E(e_i) = 0$ and $\text{VAR}(e_i) =$

σ^2. Also assume that $\max_i(h_1, ..., h_n) \to 0$ in probability as $n \to \infty$ and

$$\frac{\boldsymbol{X}^T \boldsymbol{X}}{n} \to \boldsymbol{W}^{-1}$$

as $n \to \infty$. Then the least squares (OLS) estimator $\hat{\boldsymbol{\beta}}$ satisfies

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2 \ \boldsymbol{W}). \tag{1}$$

Partial F Test Theorem: Suppose $H_0 : \boldsymbol{L\beta} = \boldsymbol{0}$ is true for the partial F test where $\boldsymbol{L}$ is a full rank $r \times p$ matrix. Under the OLS full rank model, a)

$$F_R = \frac{1}{rMSE}(\boldsymbol{L}\hat{\boldsymbol{\beta}})^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}}).$$

b) If $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, then $F_R \sim F_{r,n-p}$.
c) For a large class of zero mean error distributions $rF_R \xrightarrow{D} \chi^2_r$.
d) The partial F test that rejects $H_0 : \boldsymbol{L\beta} = \boldsymbol{0}$ if $F_R > F_{r,n-p}(1-\delta)$ is a large sample right tail δ test for the OLS model for a large class of zero mean error distributions.

Assume H_0 is true. By the OLS CLT, $\sqrt{n}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{L\beta}) = \sqrt{n}\boldsymbol{L}\hat{\boldsymbol{\beta}} \xrightarrow{D} N_r(\boldsymbol{0}, \sigma^2 \ \boldsymbol{LWL}^T)$. Thus $\sqrt{n}(\boldsymbol{L}\hat{\boldsymbol{\beta}})^T(\sigma^2 \boldsymbol{LWL}^T)^{-1}\sqrt{n}\boldsymbol{L}\hat{\boldsymbol{\beta}} \xrightarrow{D} \chi^2_r$. Let $\hat{\sigma}^2 = MSE$ and $\hat{\boldsymbol{W}} = n(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Then

$$n(\boldsymbol{L}\hat{\boldsymbol{\beta}})^T[MSE \ \boldsymbol{L}n(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}\boldsymbol{L}\hat{\boldsymbol{\beta}} = rF_R \xrightarrow{D} \chi^2_r.$$

Partial F test: Let the full model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ with a constant β_1 in the model: $\boldsymbol{1}$ is the 1st column of $\boldsymbol{X}$. Let the reduced model $\boldsymbol{Y} = \boldsymbol{X}_R\boldsymbol{\beta}_R + \boldsymbol{e}$ also have a constant in the model where the columns of $\boldsymbol{X}_R$ are a subset of k of the columns of $\boldsymbol{X}$. Let $\boldsymbol{P}_R$ be the projection matrix on $C(\boldsymbol{X}_R)$ so $\boldsymbol{PP}_R = \boldsymbol{P}_R$. Then $F_R = \dfrac{SSE(R) - SSE(F)}{rMSE(F)}$ where $r = df_R - df_F = p - k$ = number of predictors in the full model but not in the reduced model. $MSE = MSE(F) = SSE(F)/(n-p)$ where $SSE = SSE(F) = \boldsymbol{Y}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$. $SSE(R) - SSE(F) = \boldsymbol{Y}^T(\boldsymbol{P} - \boldsymbol{P}_R)\boldsymbol{Y}$ where $SSE(R) = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P}_R)\boldsymbol{Y}$.

Now assume $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$, and when H_0 is true, $\boldsymbol{Y} \sim N_n(\boldsymbol{X}_R\boldsymbol{\beta}_R, \sigma^2\boldsymbol{I})$. Since $(\boldsymbol{I} - \boldsymbol{P})(\boldsymbol{P} - \boldsymbol{P}_R) = \boldsymbol{0}$, $[SSE(R) - SSE(F)] \perp\!\!\!\perp MSE(F)$ by Craig's Theorem. When H_0 is true, $\boldsymbol{\mu} = \boldsymbol{X}_R\boldsymbol{\beta}_R$ and $\boldsymbol{\mu}^T\boldsymbol{A\mu} = 0$ where $\boldsymbol{A} = (\boldsymbol{I} - \boldsymbol{P})$ or $\boldsymbol{A} = (\boldsymbol{P} - \boldsymbol{P}_R)$. Hence the noncentrality parameter is 0, and by Theorem 2.14 g), $SSE \sim \sigma^2\chi^2_{n-p}$ and $SSE(R) - SSE(F) \sim \sigma^2\chi^2_{p-k}$ since $rank(\boldsymbol{P} - \boldsymbol{P}_R) = tr(\boldsymbol{P} - \boldsymbol{P}_R) = p - k$. Hence under H_0, $F_R \sim F_{p-k,n-p}$.

An ANOVA table for the partial F test is shown below, where $k = p_R$ is the number of predictors used by the reduced model, and $r = p - p_R = p - k$ is the number of predictors in the full model that are not in the reduced model.

Source	df	SS	MS	F
Reduced	$n - p_R$	$SSE(R) = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P}_R)\boldsymbol{Y}$	MSE(R)	$F_R = \frac{SSE(R)-SSE}{rMSE} =$
Full	$n - p$	$SSE = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$	MSE	$\dfrac{\boldsymbol{Y}^T(\boldsymbol{P} - \boldsymbol{P}_R)\boldsymbol{Y}/r}{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}/(n-p)}$

The ANOVA F test is the special case where $k = 1$, $\boldsymbol{X}_R = \boldsymbol{1}$, $\boldsymbol{P}_R = \boldsymbol{P}_1$, and $SSE(R) - SSE(F) = SSTO - SSE = SSR$.

ANOVA table: $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ with a constant β_1 in the model: $\boldsymbol{1}$ is the 1st column of $\boldsymbol{X}$. $MS = SS/df$.

$SSTO = \boldsymbol{Y}^T(\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T)\boldsymbol{Y} = \sum_{i=1}^n (Y_i - \overline{Y})^2$, $SSE = \sum_{i=1}^n r_i^2$, $SSR = \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2$, $SSTO = SSR + SSE$. SSTO is the SSE (residual sum of squares) for the location model $\boldsymbol{Y} = \boldsymbol{1}\beta_1 + \boldsymbol{e}$ that contains a constant but no nontrivial predictors. The location model has projection matrix $\boldsymbol{P}_1 = \boldsymbol{1}(\boldsymbol{1}^T\boldsymbol{1})^{-1}\boldsymbol{1}^T = \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T$. Hence $\boldsymbol{P}\boldsymbol{P}_1 = \boldsymbol{P}_1$ and $\boldsymbol{P}\boldsymbol{1} = \boldsymbol{P}_1\boldsymbol{1} = \boldsymbol{1}$.

Source	df	SS	MS	F	p-value
Regression	p-1	$SSR = \boldsymbol{Y}^T(\boldsymbol{P} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T)\boldsymbol{Y}$	MSR	$F_0 = MSR/MSE$	for H_0:
Residual	n-p	$SSE = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$	MSE		$\beta_2 = \cdots = \beta_p = 0$

The matrices in the quadratic forms for SSR and SSE are symmetric and idempotent and their product is $\boldsymbol{0}$. Hence if $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ so $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$, then $SSE \perp\!\!\!\perp SSR$ by Craig's Theorem. If H_0 is true under normality, then $\boldsymbol{Y} \sim N_n(\boldsymbol{1}\beta_1, \sigma^2\boldsymbol{I})$, and by Theorem 2.14 g), $SSE \sim \sigma^2\chi^2_{n-p}$ and $SSR \sim \sigma^2\chi^2_{p-1}$ since $rank(\boldsymbol{I} - \boldsymbol{P}) = tr(\boldsymbol{I} - \boldsymbol{P}) = n - p$ and $rank(\boldsymbol{P} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T) = tr(\boldsymbol{P} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^T) = p - 1$. Hence under normality, $F_0 \sim F_{p-1,n-p}$.

9) Expected Value, Covariance Matrix and Large Sample Theory for least squares quantities:

For the full rank model, $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ where $E(\boldsymbol{Y}) = \boldsymbol{X\beta}$, $E(\boldsymbol{e}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}) = \text{Cov}(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}$, $E(\boldsymbol{AY}) = \boldsymbol{AX\beta}$ and $\text{Cov}(\boldsymbol{AY}) = \sigma^2\boldsymbol{A}\boldsymbol{A}^T$.

$\boldsymbol{A} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is used for $\hat{\boldsymbol{\beta}} = \boldsymbol{AY}$. $\boldsymbol{A} = \boldsymbol{I} - \boldsymbol{P} = \boldsymbol{I} - \boldsymbol{H}$ is used for the residual vector $\boldsymbol{Y} - \hat{\boldsymbol{Y}} = \boldsymbol{AY}$. $\boldsymbol{A} = \boldsymbol{P} = \boldsymbol{H}$ is used for the vector of fitted values $\hat{\boldsymbol{Y}}$.

For the full rank Gaussian linear model, $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$, and if $\boldsymbol{A}$ is $k \times n$ with rank k, then $\boldsymbol{AY} \sim N_k(\boldsymbol{AX\beta}, \sigma^2\boldsymbol{A}\boldsymbol{A}^T)$.

If $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{W})$, and $\boldsymbol{A}$ is $k \times p$ with rank k, then $\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{A\beta}) \xrightarrow{D} N_k(\boldsymbol{0}, \sigma^2\ \boldsymbol{AWA}^T)$.

The non-full rank model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ also has $E(\boldsymbol{Y}) = \boldsymbol{X\beta}$, $E(\boldsymbol{e}) = \boldsymbol{0}$, $\text{Cov}(\boldsymbol{e}) = \text{Cov}(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}$, $E(\boldsymbol{AY}) = \boldsymbol{AX\beta}$ and $\text{Cov}(\boldsymbol{AY}) = \sigma^2\boldsymbol{A}\boldsymbol{A}^T$.

For the non-full rank model $\boldsymbol{A} = (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$ is used for $\hat{\boldsymbol{\beta}} = \boldsymbol{AY}$ and $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$.

You should be able to handle the linear model written in different ways. The residual bootstrap model $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{e}^*$ with $E(\boldsymbol{e}^*) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}^*) = \text{Cov}(\boldsymbol{Y}^*) = \hat{\sigma}^2\boldsymbol{I}$. The parametric bootstrap model $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{e}^*$ with $\boldsymbol{Y}^* \sim N_n(\boldsymbol{X}\hat{\boldsymbol{\beta}}, MSE\ \boldsymbol{I})$. In numerical linear algebra, the least squares solution to "$\boldsymbol{Ax} = \boldsymbol{b}$" is of interest where the problem is actually the multiple linear regression model $\boldsymbol{b} = \boldsymbol{Ax} + \boldsymbol{\epsilon}$ where $\boldsymbol{A}$ has full rank p, and we will assume that $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$, and $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}_n$.

References:

Christensen, R. (2020), *Plane Answers to Complex Questions: the Theory of Linear Models,* 5th ed., Springer, New York, NY.

Graybill, F.A. (2000), *Theory and Application of the Linear Model,* Brooks/Cole, Pacific Grove, CA.

Olive, D.J. (2023), *Theory of Linear Models*, online course notes, see (http://parker.ad.siu.edu/Olive/linmodbk.htm).

Rencher, A.C., and Schaalje, G.B. (2008), *Linear Models in Statistics*, 2nd ed., Wiley, Hoboken, NJ.

Searle, S.R., and Gruber, M.H.J. (2017), *Linear Models*, 2nd ed., Wiley, Hoboken, NJ.

Seber, G.A.F., and Lee, A.J. (2003), *Linear Regression Analysis,* 2nd ed., Wiley, New York, NY.

Zimmerman, D.L. (2020), *Linear Model Theory: Exercises and Solutions*, Springer, New York, NY.