Chapter 2
Full Rank Linear Models

2.1 Projection Matrices and the Column Space

Vector spaces, subspaces, and column spaces should be familiar from linear algebra, but are reviewed below.

Definition 2.1. A set $\mathcal{V} \subseteq \mathbb{R}^k$ is a **vector space** if for any vectors $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{V}$, and scalars a and b, the operations of vector addition and scalar multiplication are defined as follows.
1) $(\boldsymbol{x}+\boldsymbol{y})+\boldsymbol{z} = \boldsymbol{x}+(\boldsymbol{y}+\boldsymbol{z})$.
2) $\boldsymbol{x}+\boldsymbol{y} = \boldsymbol{y}+\boldsymbol{x}$.
3) There exists $\boldsymbol{0} \in \mathcal{V}$ such that $\boldsymbol{x}+\boldsymbol{0} = \boldsymbol{x} = \boldsymbol{0}+\boldsymbol{x}$.
4) For any $\boldsymbol{x} \in \mathcal{V}$, there exists $\boldsymbol{y} = -\boldsymbol{x}$ such that $\boldsymbol{x}+\boldsymbol{y} = \boldsymbol{y}+\boldsymbol{x} = \boldsymbol{0}$.
5) $a(\boldsymbol{x}+\boldsymbol{y}) = a\boldsymbol{x}+a\boldsymbol{y}$.
6) $(a+b)\boldsymbol{x} = a\boldsymbol{x}+b\boldsymbol{y}$.
7) (ab) $\boldsymbol{x}$ = a(b $\boldsymbol{x}$).
8) 1 $\boldsymbol{x} = \boldsymbol{x}$.

Hence for a vector space, addition is associative and commutative, there is an additive identity vector $\boldsymbol{0}$, there is an additive inverse $-\boldsymbol{x}$ for each $\boldsymbol{x} \in \mathcal{V}$, scalar multiplication is distributive and associative, and 1 is the scalar identity element.

Two important vector spaces are $\mathbb{R}^k$ and $\mathcal{V} = \{\boldsymbol{0}\}$. Showing that a set $\mathcal{M}$ is a subspace is a common method to show that $\mathcal{M}$ is a vector space.

Definition 2.2. Let $\mathcal{M}$ be a nonempty subset of a vector space $\mathcal{V}$. If i) $a\boldsymbol{x} \in \mathcal{M}\ \forall \boldsymbol{x} \in \mathcal{M}$ and for any scalar a, and ii) $\boldsymbol{x}+\boldsymbol{y} \in \mathcal{M}\ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{M}$, then $\mathcal{M}$ is a vector space known as a **subspace**.

Definition 2.3. The set of all linear combinations of $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ is the vector space known as $span(\boldsymbol{x}_1, ..., \boldsymbol{x}_n) = \{\boldsymbol{y} \in \mathbb{R}^k : \boldsymbol{y} = \sum_{i=1}^n a_i \boldsymbol{x}_i$ for some constants $a_1, ..., a_n\}$.

Definition 2.4. Let $\boldsymbol{x}_1, ..., \boldsymbol{x}_k \in \mathcal{V}$. If $\exists$ scalars $\alpha_1, ..., \alpha_k$ not all zero such that $\sum_{i=1}^k \alpha_i \boldsymbol{x}_i = \boldsymbol{0}$, then $\boldsymbol{x}_1, ..., \boldsymbol{x}_k$ are *linearly dependent.* If $\sum_{i=1}^k \alpha_i \boldsymbol{x}_i = \boldsymbol{0}$ only if $\alpha_i = 0 \ \forall \ i = 1, ..., k$, then $\boldsymbol{x}_1, ..., \boldsymbol{x}_k$ are *linearly independent.* Suppose $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_k\}$ is a linearly independent set and $\mathcal{V} = span(\boldsymbol{x}_1, ..., \boldsymbol{x}_k)$. Then $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_k\}$ is a linearly independent spanning set for $\mathcal{V}$, known as a *basis.*

Definition 2.5. 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 $\boldsymbol{A} = C(\boldsymbol{A})$. Then $C(\boldsymbol{A}) = \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y} = \boldsymbol{A}\boldsymbol{w}$ for some $\boldsymbol{w} \in \mathbb{R}^m\} = \{\boldsymbol{y} : \boldsymbol{y} = w_1\boldsymbol{a}_1 + w_2\boldsymbol{a}_2 + \cdots + w_m\boldsymbol{a}_m$ for some scalars $w_1, ..., w_m\} = span(\boldsymbol{a}_1, ..., \boldsymbol{a}_m)$.

The space spanned by the rows of $\boldsymbol{A}$ is the *row space* of $\boldsymbol{A}$. The row space of $\boldsymbol{A}$ is the column space $C(\boldsymbol{A}^T)$ of $\boldsymbol{A}^T$. Note that

$$\boldsymbol{A}\boldsymbol{w} = [\boldsymbol{a}_1 \ \boldsymbol{a}_2 \ ... \ \boldsymbol{a}_m] \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix} = \sum_{i=1}^m w_i \boldsymbol{a}_i.$$

With the design matrix $\boldsymbol{X}$, different notation is used to denote the columns of $\boldsymbol{X}$ since both the columns and rows $\boldsymbol{X}$ are important. Let

$$\boldsymbol{X} = [\boldsymbol{v}_1 \ \boldsymbol{v}_2 \ ... \ \boldsymbol{v}_p] = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix}$$

be an $n \times p$ matrix. Note that $C(\boldsymbol{X}) = \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y} = \boldsymbol{X}\boldsymbol{b}$ for some $\boldsymbol{b} \in \mathbb{R}^p\}$. Hence $\boldsymbol{X}\boldsymbol{b}$ is a typical element of $C(\boldsymbol{X})$ and $\boldsymbol{A}\boldsymbol{w}$ is a typical element of $C(\boldsymbol{A})$. Note that

$$\boldsymbol{X}\boldsymbol{b} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} \boldsymbol{b} = \begin{bmatrix} \boldsymbol{x}_1^T\boldsymbol{b} \\ \vdots \\ \boldsymbol{x}_n^T\boldsymbol{b} \end{bmatrix} = [\boldsymbol{v}_1 \ \boldsymbol{v}_2 \ ... \ \boldsymbol{v}_p] \begin{bmatrix} b_1 \\ \vdots \\ b_p \end{bmatrix} = \sum_{i=1}^p b_i \boldsymbol{v}_i.$$

If the function $\boldsymbol{X}_f(\boldsymbol{b}) = \boldsymbol{X}\boldsymbol{b}$ where the f indicates that the operation $\boldsymbol{X}_f : \mathbb{R}^p \to \mathbb{R}^n$ is being treated as a function, then $C(\boldsymbol{X})$ is the range of $\boldsymbol{X}_f$. Hence some authors call the column space of $\boldsymbol{A}$ the range of $\boldsymbol{A}$.

Let $\boldsymbol{B}$ be $n \times k$, and let $\boldsymbol{A}$ be $n \times m$. One way to show $C(\boldsymbol{A}) = C(\boldsymbol{B})$ is to show that i) $\forall \boldsymbol{x} \in \mathbb{R}^m$, $\exists \ \boldsymbol{y} \in \mathbb{R}^k$ such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{B}\boldsymbol{y} \in C(\boldsymbol{B})$ so $C(\boldsymbol{A}) \subseteq C(\boldsymbol{B})$, and ii) $\forall \boldsymbol{y} \in \mathbb{R}^k$, $\exists \ \boldsymbol{x} \in \mathbb{R}^m$ such that $\boldsymbol{B}\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \in C(\boldsymbol{A})$ so $C(\boldsymbol{B}) \subseteq C(\boldsymbol{A})$. Another way to show $C(\boldsymbol{A}) = C(\boldsymbol{B})$ is to show that a basis for $C(\boldsymbol{A})$ is also a basis for $C(\boldsymbol{B})$.

Definition 2.6. The *dimension of a vector space* $\mathcal{V} = dim(\mathcal{V})$ = the number of vectors in a basis of $\mathcal{V}$. The *rank of a matrix* $\boldsymbol{A} = rank(\boldsymbol{A}) = dim(C(\boldsymbol{A}))$, the dimension of the column space of $\boldsymbol{A}$. Let $\boldsymbol{A}$ be $n \times m$. Then

$\text{rank}(\boldsymbol{A}) = \text{rank}(\boldsymbol{A}^T) \leq \min(m, n)$. If $\text{rank}(\boldsymbol{A}) = \min(m, n)$, then $\boldsymbol{A}$ has *full rank*, or $\boldsymbol{A}$ is a full rank matrix.

Definition 2.7. The *null space* of $\boldsymbol{A} = N(\boldsymbol{A}) = \{\boldsymbol{x} : \boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}\} =$ *kernel* of $\boldsymbol{A}$. The *nullity* of $\boldsymbol{A} = \dim[N(\boldsymbol{A})]$. The subspace $\mathcal{V}^{\perp} = \{\boldsymbol{y} \in \mathbb{R}^k : \boldsymbol{y} \perp \mathcal{V}\}$ is the *orthogonal complement of* $\mathcal{V}$, where $\boldsymbol{y} \perp \mathcal{V}$ means $\boldsymbol{y}^T\boldsymbol{x} = \boldsymbol{0}\ \forall\ \boldsymbol{x} \in \mathcal{V}$. $N(\boldsymbol{A}^T) = [C(\boldsymbol{A})]^{\perp}$, so $N(\boldsymbol{A}) = [C(\boldsymbol{A}^T)]^{\perp}$.

Theorem 2.1: Rank Nullity Theorem. Let $\boldsymbol{A}$ be $n \times m$. Then $\text{rank}(\boldsymbol{A}) + \dim(N(\boldsymbol{A})) = m$.

Generalized inverses are useful for the non-full rank linear model and for defining projection matrices.

Definition 2.8. A **generalized inverse** of an $n \times m$ matrix $\boldsymbol{A}$ is any $m \times n$ matrix $\boldsymbol{A}^-$ satisfying $\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}$.

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 $\boldsymbol{A}$.

Recall that if $\boldsymbol{A}$ is **idempotent**, then $\boldsymbol{A}^2 = \boldsymbol{A}$. A matrix $\boldsymbol{A}$ is *tripotent* if $\boldsymbol{A}^3 = \boldsymbol{A}$. For both these cases, $\boldsymbol{A} := \boldsymbol{A}^-$ since $\boldsymbol{A}\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}$. It will turn out that symmetric idempotent matrices are projection matrices.

Definition 2.9. Let $\mathcal{V}$ be a subspace of $\mathbb{R}^n$. Then every $\boldsymbol{y} \in \mathbb{R}^n$ can be expressed uniquely as $\boldsymbol{y} = \boldsymbol{w} + \boldsymbol{z}$ where $\boldsymbol{w} \in \mathcal{V}$ and $\boldsymbol{z} \in \mathcal{V}^{\perp}$. Let $\boldsymbol{X} = [\boldsymbol{v}_1\ \boldsymbol{v}_2\ \ldots\ \boldsymbol{v}_p]$ be $n \times p$, and let $\mathcal{V} = C(\boldsymbol{X}) = span(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p)$. Then the $n \times n$ matrix $\boldsymbol{P}_{\mathcal{V}} = \boldsymbol{P}_{\boldsymbol{X}}$ 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 $\boldsymbol{y}$.)

Note: Some authors call a projection matrix an "orthogonal projection matrix," and call an idempotent matrix a "projection matrix."

Theorem 2.2: 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 $C(\boldsymbol{P}_{\boldsymbol{X}}) = C(\boldsymbol{X})$, and $\boldsymbol{P}_{\boldsymbol{X}}$ is symmetric and idempotent. Also, each column $\boldsymbol{p}_i$ of $\boldsymbol{P}_{\boldsymbol{X}}$ satisfies $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{p}_i = \boldsymbol{p}_i \in C(\boldsymbol{X})$.
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 $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{y}$ and ii) $\boldsymbol{y} \perp C(\boldsymbol{X})$ implies $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{0}$.
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) The eigenvalues of $\boldsymbol{P_X}$ are 0 or 1.
j) Let $tr(\boldsymbol{A}) = trace(\boldsymbol{A})$. Then $rank(\boldsymbol{P_X}) = tr(\boldsymbol{P_X}) = rank(\boldsymbol{X})$.
k) $\boldsymbol{P_X}$ is singular unless $\boldsymbol{X}$ is a nonsingular $n \times n$ matrix, and then $\boldsymbol{P_X} = \boldsymbol{I}_n$.
l) Let $\boldsymbol{X} = [\boldsymbol{Z}\ \boldsymbol{X}_r]$ where rank($\boldsymbol{X}$) = rank($\boldsymbol{X}_r$) = r so the columns of $\boldsymbol{X}_r$ form a basis for $C(\boldsymbol{X})$. Then

$$\begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & (\boldsymbol{X}_r^T \boldsymbol{X}_r)^{-1} \end{bmatrix}$$

is a generalized inverse of $\boldsymbol{X}^T\boldsymbol{X}$, and $\boldsymbol{P_X} = \boldsymbol{X}_r(\boldsymbol{X}_r^T\boldsymbol{X}_r)^{-1}\boldsymbol{X}_r^T$.

Two important consequences of the above theorem follow. First, $\boldsymbol{P}$ 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 $\mathcal{C}(\boldsymbol{P}_1) = \mathcal{C}(\boldsymbol{X}_1) \subseteq \mathcal{C}(\boldsymbol{X})$, $\boldsymbol{P}\boldsymbol{P}_1 = \boldsymbol{P}_1$. Hence $\boldsymbol{P}_1\boldsymbol{P} = (\boldsymbol{P}\boldsymbol{P}_1)^T = \boldsymbol{P}_1^T = \boldsymbol{P}_1$.

Some results from linear algebra are needed to prove parts of the above theorem. Unless told otherwise, matrices in this text are real. Then the eigenvalues of a symmetric matrix $\boldsymbol{A}$ are real. If $\boldsymbol{A}$ is symmetric, then rank($\boldsymbol{A}$) = number of nonzero eigenvalues of $\boldsymbol{A}$. Recall that if $\boldsymbol{AB}$ is a square matrix, then $tr(\boldsymbol{AB}) = tr(\boldsymbol{BA})$. Similarly, if $\boldsymbol{A}_1$ is $m_1 \times m_2$, $\boldsymbol{A}_2$ is $m_2 \times m_3$, ..., $\boldsymbol{A}_{k-1}$ is $m_{k-1} \times m_k$, and $\boldsymbol{A}_k$ is $m_k \times m_1$, then $tr(\boldsymbol{A}_1\boldsymbol{A}_2 \cdots \boldsymbol{A}_k) = tr(\boldsymbol{A}_k\boldsymbol{A}_1\boldsymbol{A}_2 \cdots \boldsymbol{A}_{k-1}) = tr(\boldsymbol{A}_{k-1}\boldsymbol{A}_k\boldsymbol{A}_1\boldsymbol{A}_2 \cdots \boldsymbol{A}_{k-2}) = \cdots = tr(\boldsymbol{A}_2\boldsymbol{A}_3 \cdots \boldsymbol{A}_k\boldsymbol{A}_1)$. Also note that a scalar is a 1×1 matrix, so $tr(a) = a$. The next two paragraphs follow Christensen (1987, pp. 335-338) closely.

If $\boldsymbol{P}$ and $\boldsymbol{A}$ are $n \times n$ matrices, then $\boldsymbol{P} = \boldsymbol{A}$ iff $\boldsymbol{Py} = \boldsymbol{Ay}$ for all $\boldsymbol{y} \in \mathbb{R}^n$ iff $\boldsymbol{y}^T\boldsymbol{P} = \boldsymbol{y}^T\boldsymbol{A}$ for all $\boldsymbol{y} \in \mathbb{R}^n$. Let $\mathcal{V}$ be a subspace of $\mathbb{R}^n$. Let $\boldsymbol{y} \in \mathbb{R}^n$ with $\boldsymbol{y} = \boldsymbol{w} + \boldsymbol{z}$ where $\boldsymbol{w} \in \mathcal{V}$ and $\boldsymbol{z} \in \mathcal{V}^\perp$. Let $\boldsymbol{A}$ and $\boldsymbol{P}$ be projection matrices on $\mathcal{V}$. Then $\boldsymbol{Ay} = \boldsymbol{w} = \boldsymbol{Py}$. Since $\boldsymbol{y}$ was arbitrary, $\boldsymbol{A} = \boldsymbol{P}$ and projection matrices are unique. We prove that $\boldsymbol{P_X}$ is symmetric below. Then the projection matrix $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^-\boldsymbol{A}$ is symmetric by replacing $\boldsymbol{X}$ by $\boldsymbol{A}$. Hence $\boldsymbol{Az} = \boldsymbol{A}^T\boldsymbol{z} = \boldsymbol{0}$. Thus $\boldsymbol{A}^2\boldsymbol{y} = \boldsymbol{Aw} = \boldsymbol{w} = \boldsymbol{Ay}$, and $\boldsymbol{A}^2 = \boldsymbol{A}$ since $\boldsymbol{y}$ was arbitrary.

Now suppose $\boldsymbol{A}^2 = \boldsymbol{A} = \boldsymbol{A}^T$, and let $\boldsymbol{w} \in C(\boldsymbol{A})$. Hence $\boldsymbol{w} = \boldsymbol{Aa}$ for some vector $\boldsymbol{a}$. Thus $\boldsymbol{Aw} = \boldsymbol{A}^2\boldsymbol{a} = \boldsymbol{Aa} = \boldsymbol{w}$. Let $\boldsymbol{z} \perp C(\boldsymbol{A}) = C(\boldsymbol{A}^T)$. Then $\boldsymbol{z}^T\boldsymbol{A} = \boldsymbol{z}^T\boldsymbol{A}^T = \boldsymbol{0}$. Thus $\boldsymbol{Ay} = \boldsymbol{Aw} = \boldsymbol{w}$, and $\boldsymbol{A}$ is a projection matrix on $C(\boldsymbol{A})$. Note that $C(P_{\boldsymbol{X}}) \subseteq C(\boldsymbol{X})$ since $\boldsymbol{P_X X} = \boldsymbol{X}$, and $C(\boldsymbol{X}) \subseteq C(P_{\boldsymbol{X}})$ since $\boldsymbol{P_X} = \boldsymbol{XW}$ where $\boldsymbol{W} = (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T$. Thus $C(\boldsymbol{X}) = C(\boldsymbol{P_X})$. To show that $\boldsymbol{P_X X} = \boldsymbol{X}$, let $\boldsymbol{y} = \boldsymbol{w} + \boldsymbol{z}$ with $\boldsymbol{w} = \boldsymbol{Xa}$ and $\boldsymbol{z}^T\boldsymbol{X} = \boldsymbol{0}$. Note that $\boldsymbol{y}^T\boldsymbol{P_X X} = \boldsymbol{w}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{a}^T\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{a}^T\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{w}^T\boldsymbol{X} = \boldsymbol{y}^T\boldsymbol{X}$. Since $\boldsymbol{y}$ was arbitrary, $\boldsymbol{P_X X} = \boldsymbol{X}$. Note that $\boldsymbol{P_X y} = \boldsymbol{P_X}(\boldsymbol{w}+\boldsymbol{z}) = \boldsymbol{P_X w} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{Xa} = \boldsymbol{P_X Xa} = \boldsymbol{Xa} = \boldsymbol{w}$. Thus $\boldsymbol{P_X}$ is a projection matrix on $C(\boldsymbol{X})$.

Note that if $\boldsymbol{G}$ is a generalized linear inverse of a symmetric matrix $\boldsymbol{A}$, then $\boldsymbol{A}^T = \boldsymbol{A}^T\boldsymbol{G}^T\boldsymbol{A}^T = \boldsymbol{A}\boldsymbol{G}^T\boldsymbol{A} = \boldsymbol{A}$. Hence $\boldsymbol{G}^T$ is a generalized linear inverse of $\boldsymbol{A}$. Also, $\boldsymbol{AGAG}^T\boldsymbol{A} = \boldsymbol{AG}^T\boldsymbol{A} = \boldsymbol{A}$. Hence $\boldsymbol{GAG}^T$, a symmetric matrix, is a generalized inverse of $\boldsymbol{A}$. Thus a symmetric matrix $\boldsymbol{A}$ always has a symmetric generalized linear inverse. Hence let $\boldsymbol{B} := (\boldsymbol{X}^T\boldsymbol{X})^-$ be a symmetric matrix. Then $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}^T\boldsymbol{B}\boldsymbol{X} = \boldsymbol{X}^T(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}$ is symmetric since $\boldsymbol{P}_{\boldsymbol{X}}$ is unique, even if $(\boldsymbol{X}^T\boldsymbol{X})^-$ is not symmetric.

For part d), note that if $\boldsymbol{y} = \boldsymbol{w} + \boldsymbol{z}$, then $(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y} = \boldsymbol{z} \in [C(\boldsymbol{X})]^\perp$. Hence the result follows from the definition of a projection matrix by interchanging the roles of $\boldsymbol{w}$ and $\boldsymbol{z}$. Part e) follows from the definition of a projection matrix since if $\boldsymbol{y} \in C(\boldsymbol{X})$ then $\boldsymbol{y} = \boldsymbol{y} + \boldsymbol{0}$ where $\boldsymbol{y} = \boldsymbol{w}$ and $\boldsymbol{0} = \boldsymbol{z}$. If $\boldsymbol{y} \perp C(\boldsymbol{X})$ then $\boldsymbol{y} = \boldsymbol{0} + \boldsymbol{y}$ where $\boldsymbol{0} = \boldsymbol{w}$ and $\boldsymbol{y} = \boldsymbol{z}$. Part g) is a special case of f). In k), $\boldsymbol{P}_{\boldsymbol{X}}$ is singular unless $p = n$ since $\text{rank}(\boldsymbol{X}) = r \leq \min(p, n) < \max(n, p)$ unless $p = n$, and $\boldsymbol{P}_{\boldsymbol{X}}$ is an $n \times n$ matrix. Need $\text{rank}(\boldsymbol{P}_{\boldsymbol{X}}) = n$ for $\boldsymbol{P}_{\boldsymbol{X}}$ to be nonsingular. For h), $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{X}_R} = \boldsymbol{P}_{\boldsymbol{X}_R}$ by f) since each column of $\boldsymbol{P}_{\boldsymbol{X}_r} \in C(\boldsymbol{P}_{\boldsymbol{X}})$. Taking transposes and using symmetry shows $\boldsymbol{P}_{\boldsymbol{X}_R}\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{P}_{\boldsymbol{X}_R}$. For i), if λ is an eigenvalue of $\boldsymbol{P}_{\boldsymbol{X}}$, then for some $\boldsymbol{x} \neq \boldsymbol{0}$, $\lambda\boldsymbol{x} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{x} = \boldsymbol{P}^2_{\boldsymbol{X}}\boldsymbol{x} = \lambda^2\boldsymbol{x}$ since $\boldsymbol{P}_{\boldsymbol{X}}$ is idempotent by c). Hence $\lambda = \lambda^2$ is real since $\boldsymbol{P}_{\boldsymbol{X}}$ is symmetric, so $\lambda = 0$ or $\lambda = 1$. Then j) follows from i) since $\text{rank}(\boldsymbol{P}_{\boldsymbol{X}})$ = number of nonzero eigenvalues of $\boldsymbol{P}_{\boldsymbol{X}} = \text{tr}(\boldsymbol{P}_{\boldsymbol{X}})$.

For l), note that $C(\boldsymbol{X}) = C(\boldsymbol{X}_r)$. Thus $\boldsymbol{X}_r(\boldsymbol{X}_r^T\boldsymbol{X}_r)^{-1}\boldsymbol{X}_r^T = \boldsymbol{P}_{\boldsymbol{X}}$. Then

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} \boldsymbol{Z}^T\boldsymbol{Z} & \boldsymbol{Z}^T\boldsymbol{X}_r \\ \boldsymbol{X}_r^T\boldsymbol{Z} & \boldsymbol{X}_r^T\boldsymbol{X}_r \end{bmatrix} \quad \text{and} \quad \boldsymbol{X}^T\boldsymbol{X} \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & (\boldsymbol{X}_r^T\boldsymbol{X}_r)^{-1} \end{bmatrix} \boldsymbol{X}^T\boldsymbol{X} =$$

$$\begin{bmatrix} \boldsymbol{Z}^T\boldsymbol{X}_r(\boldsymbol{X}_r^T\boldsymbol{X}_r)^{-1}\boldsymbol{X}_r^T\boldsymbol{Z} & \boldsymbol{Z}^T\boldsymbol{X}_r \\ \boldsymbol{X}_r^T\boldsymbol{Z} & \boldsymbol{X}_r^T\boldsymbol{X}_r \end{bmatrix} = \boldsymbol{X}^T\boldsymbol{X}$$

since $\boldsymbol{Z}^T\boldsymbol{P}_X\boldsymbol{Z} = \boldsymbol{Z}^T\boldsymbol{Z}$ because each column of $\boldsymbol{Z} \in C(\boldsymbol{X})$.

Most of the above results apply to full rank and nonfull rank matrices. A corollary of the following theorem is that if $\boldsymbol{X}$ is full rank, then $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{H}$.

Suppose $\boldsymbol{A}$ is $p \times p$. Then the following are equivalent. 1) $\boldsymbol{A}$ is nonsingular, 2) $\boldsymbol{A}$ has a left inverse $\boldsymbol{L}$ with $\boldsymbol{LA} = \boldsymbol{I}_p$, and 3) $\boldsymbol{A}$ has a right inverse $\boldsymbol{R}$ with $\boldsymbol{AR} = \boldsymbol{I}_p$. To see this, note that 1) implies (2) and 3) since $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_p = \boldsymbol{A}\boldsymbol{A}^{-1}$ by the definition of an inverse matrix. Suppose $\boldsymbol{AR} = \boldsymbol{I}_p$. Then the determinant $det(\boldsymbol{I}_p) = 1 = det(\boldsymbol{AR}) = det(\boldsymbol{A})\,\det(\boldsymbol{R})$. Hence $det(\boldsymbol{A}) \neq 0$ and $\boldsymbol{A}$ is nonsingular. Hence $\boldsymbol{R} = \boldsymbol{A}^{-1}\boldsymbol{AR} = \boldsymbol{A}^{-1}$ and 3) implies 1). Similarly 2) implies 1). Also note that $\boldsymbol{L} = \boldsymbol{LI}_p = \boldsymbol{LAR} = \boldsymbol{I}_p\boldsymbol{R} = \boldsymbol{R} = \boldsymbol{A}^{-1}$. Hence in the proof below, we could just show that $\boldsymbol{A}^- = \boldsymbol{L}$ or $\boldsymbol{A}^- = \boldsymbol{R}$.

Theorem 2.3. If $\boldsymbol{A}$ is nonsingular, the unique generalized inverse of $\boldsymbol{A}$ is $\boldsymbol{A}^{-1}$.

Proof. Let $\boldsymbol{A}^-$ be any generalized inverse of $\boldsymbol{A}$. We give two proofs. i) $\boldsymbol{A}^- = \boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}$. ii) $\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$ and $\boldsymbol{A}\boldsymbol{A}^- = \boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}$. Thus $\boldsymbol{A}^- = \boldsymbol{A}^{-1}$. $\square$

2.2 Quadratic Forms

Definition 2.10. Let $\boldsymbol{A}$ be an $n \times n$ matrix and let $\boldsymbol{x} \in \mathbb{R}^n$. Then a **quadratic form** $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^n \sum_{j=1}^n a_{ij}x_ix_j$, and a **linear form** is $\boldsymbol{A}\boldsymbol{x}$. Suppose $\boldsymbol{A}$ is a symmetric matrix. Then $\boldsymbol{A}$ is **positive definite** ($\boldsymbol{A} > 0$) if $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} > 0 \ \forall \ \boldsymbol{x} \neq \boldsymbol{0}$, and $\boldsymbol{A}$ is **positive semidefinite** ($\boldsymbol{A} \geq 0$) if $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \geq 0 \ \forall \ \boldsymbol{x}$.

Notation: The matrix $\boldsymbol{A}$ in a quadratic form $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$ will be **symmetric** unless told otherwise. Suppose $\boldsymbol{B}$ is not symmetric. Since the quadratic form is a scalar, $\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} = (\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x})^T = \boldsymbol{x}^T\boldsymbol{B}^T\boldsymbol{x} = \boldsymbol{x}^T(\boldsymbol{B}+\boldsymbol{B}^T)\boldsymbol{x}/2$, and the matrix $\boldsymbol{A} = (\boldsymbol{B} + \boldsymbol{B}^T)/2$ is symmetric. If $\boldsymbol{A} \geq 0$ then the eigenvalues λ_i 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 $\lambda_n > 0$. Some authors say symmetric $\boldsymbol{A}$ is nonnegative definite if $\boldsymbol{A} \geq 0$, and that $\boldsymbol{A}$ is positive semidefinite if $\boldsymbol{A} \geq 0$ and there exists a nonzero $\boldsymbol{x}$ such that $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = 0$. Then $\boldsymbol{A}$ is singular.

The spectral decomposition theorem is very useful. One application for linear models is defining the square root matrix.

Theorem 2.4: Spectral Decomposition Theorem. Let $\boldsymbol{A}$ be an $n \times n$ symmetric matrix with eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{t}_1), (\lambda_2, \boldsymbol{t}_2), ..., (\lambda_n, \boldsymbol{t}_n)$ where $\boldsymbol{t}_i^T\boldsymbol{t}_i = 1$ and $\boldsymbol{t}_i^T\boldsymbol{t}_j = 0$ if $i \neq j$ for $i = 1, ..., n$. Hence $\boldsymbol{A}\boldsymbol{t}_i = \lambda_i\boldsymbol{t}_i$. Then the *spectral decomposition* of $\boldsymbol{A}$ is

$$\boldsymbol{A} = \sum_{i=1}^n \lambda_i\boldsymbol{t}_i\boldsymbol{t}_i^T = \lambda_1\boldsymbol{t}_1\boldsymbol{t}_1^T + \cdots + \lambda_n\boldsymbol{t}_n\boldsymbol{t}_n^T.$$

Let $\boldsymbol{T} = [\boldsymbol{t}_1 \ \boldsymbol{t}_2 \ \cdots \ \boldsymbol{t}_n]$ be the $n \times n$ orthogonal matrix with ith column $\boldsymbol{t}_i$. Then $\boldsymbol{T}\boldsymbol{T}^T = \boldsymbol{T}^T\boldsymbol{T} = \boldsymbol{I}$. Let $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_n)$ and let $\boldsymbol{\Lambda}^{1/2} = \text{diag}(\sqrt{\lambda_1}, ..., \sqrt{\lambda_n})$. Then $\boldsymbol{A} = \boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}^T$.

Definition 2.11. If $\boldsymbol{A}$ is a positive definite $n \times n$ symmetric matrix with spectral decomposition $\boldsymbol{A} = \sum_{i=1}^n \lambda_i\boldsymbol{t}_i\boldsymbol{t}_i^T$, then $\boldsymbol{A} = \boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}^T$ and

$$\boldsymbol{A}^{-1} = \boldsymbol{T}\boldsymbol{\Lambda}^{-1}\boldsymbol{T}^T = \sum_{i=1}^n \frac{1}{\lambda_i}\boldsymbol{t}_i\boldsymbol{t}_i^T.$$

The *square root matrix* $\boldsymbol{A}^{1/2} = \boldsymbol{T}\boldsymbol{\Lambda}^{1/2}\boldsymbol{T}^T$ is a positive definite symmetric matrix such that $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{A}$.

The following theorem is often useful. Both the expected value and trace are linear operators. Hence $tr(\boldsymbol{A}+\boldsymbol{B}) = tr(\boldsymbol{A}) + tr(\boldsymbol{B})$, and $E[tr(\boldsymbol{X})] = tr(E[\boldsymbol{X}])$ when the expected value of the random matrix $\boldsymbol{X}$ exists.

Theorem 2.5: 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

$$E(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}) = tr(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}.$$

Proof. Two proofs are given. i) Searle (1971, p. 55): Note that $E(\boldsymbol{x}\boldsymbol{x}^T) = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^T$. Since the quadratic form is a scalar and the trace is a linear operator, $E[\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}] = E[tr(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x})] = E[tr(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T)] = tr(E[\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T]) = tr(\boldsymbol{A}\boldsymbol{\Sigma} + \boldsymbol{A}\boldsymbol{\mu}\boldsymbol{\mu}^T) = tr(\boldsymbol{A}\boldsymbol{\Sigma}) + tr(\boldsymbol{A}\boldsymbol{\mu}\boldsymbol{\mu}^T) = tr(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}$.

ii) Graybill (1976, p. 140): Using $E(x_ix_j) = \sigma_{ij} + \mu_i\mu_j$, $E[\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}] = \sum_{i=1}^n\sum_{j=1}^n a_{ij}E(x_ix_j) = \sum_{i=1}^n\sum_{j=1}^n a_{ij}(\sigma_{ij}+\mu_i\mu_j) = tr(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}$. □

Much of the theoretical results for quadratic forms assumes that the e_i are iid $N(0,\sigma^2)$. These exact results are often special cases of large sample theory that holds for a large class of iid zero mean error distributions that have $V(e_i) \equiv \sigma^2$. For linear models, $\boldsymbol{Y}$ is typically an $n \times 1$ random vector. The following theorem from statistical inference will be useful.

Theorem 2.6. Suppose $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$, $g(\boldsymbol{x})$ is a function of $\boldsymbol{x}$ alone, and $h(\boldsymbol{y})$ is a function of $\boldsymbol{y}$ alone. Then $g(\boldsymbol{x}) \perp\!\!\!\perp h(\boldsymbol{y})$.

The following theorem shows that independence of linear forms implies independence of quadratic forms.

Theorem 2.7. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric matrices and $\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{BY}$, then $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{BY}$.

Proof. Let $g(\boldsymbol{AY}) = \boldsymbol{Y}^T\boldsymbol{A}^T\boldsymbol{A}^-\boldsymbol{AY} = \boldsymbol{Y}^T\boldsymbol{AA}^-\boldsymbol{AY} = \boldsymbol{Y}^T\boldsymbol{AY}$, and let $h(\boldsymbol{BY}) = \boldsymbol{Y}^T\boldsymbol{B}^T\boldsymbol{B}^-\boldsymbol{BY} = \boldsymbol{Y}^T\boldsymbol{BB}^-\boldsymbol{BY} = \boldsymbol{Y}^T\boldsymbol{BY}$. Then the result follows by Theorem 2.6. □

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 $\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{BY}$ iff $\boldsymbol{AB}^T = \boldsymbol{0}$ iff $\boldsymbol{BA}^T = \boldsymbol{0}$.

b) If $\boldsymbol{A}$ is a symmetric $n \times n$ matrix, and $\boldsymbol{B}$ is an $m \times n$ matrix, then $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{BY}$ if $\boldsymbol{A\Sigma B}^T = \boldsymbol{0}$ if $\boldsymbol{B\Sigma A}^T = \boldsymbol{B\Sigma A} = \boldsymbol{0}$. Note that if $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}_n$, then $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{BY}$ if $\boldsymbol{AB}^T = \boldsymbol{0}$ if $\boldsymbol{BA} = \boldsymbol{0}$.

Proof. a) Note that

$$\begin{pmatrix}\boldsymbol{u}\\ \boldsymbol{w}\end{pmatrix} = \begin{pmatrix}\boldsymbol{AY}\\ \boldsymbol{BY}\end{pmatrix} = \begin{pmatrix}\boldsymbol{A}\\ \boldsymbol{B}\end{pmatrix}\boldsymbol{Y}$$

has a multivariate normal distribution. Hence $\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{BY}$ iff $\text{Cov}(\boldsymbol{u},\boldsymbol{w}) = \boldsymbol{0}$. Taking transposes shows $\text{Cov}(\boldsymbol{u},\boldsymbol{w}) = \boldsymbol{A\Sigma B}^T = \boldsymbol{0}$ iff $\boldsymbol{B\Sigma A}^T = \boldsymbol{0}$.

b) If $\boldsymbol{A\Sigma B}^T = \boldsymbol{0}$, then $\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{BY}$ by a). Let $g(\boldsymbol{AY}) = \boldsymbol{Y}^T\boldsymbol{A}^T\boldsymbol{A}^-\boldsymbol{AY} = \boldsymbol{Y}^T\boldsymbol{AA}^-\boldsymbol{AY} = \boldsymbol{Y}^T\boldsymbol{AY}$. Then $g(\boldsymbol{AY}) = \boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{BY}$ by Theorem 2.6. □

One of the most useful theorems for proving that $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{BY}$ is Craig's Theorem. Taking transposes shows $\boldsymbol{A\Sigma B} = \boldsymbol{0}$ iff $\boldsymbol{B\Sigma A} = \boldsymbol{0}$. Note that if $\boldsymbol{A\Sigma B} = \boldsymbol{0}$, then $(*)$ holds. Note $\boldsymbol{A\Sigma B} = \boldsymbol{0}$ is a sufficient condition for $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{BY}$ if $\boldsymbol{\Sigma} \geq 0$, but necessary and sufficient if $\boldsymbol{\Sigma} > 0$. If $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{BY}$, then $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{BY}$, but if $\boldsymbol{\Sigma}$ is singular, it is possible that $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{BY}$ even if $\boldsymbol{AY}$ and $\boldsymbol{BY}$ are dependent.

Theorem 2.9: 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 $\boldsymbol{\mu}^T\boldsymbol{A\Sigma B\mu} = 0$.

Proof. For a) and b), $\boldsymbol{A\Sigma B} = \boldsymbol{0}$ implies $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{BY}$ by c) or by Theorems 2.6, 2.7, and 2.8. See Reid and Driscoll (1988) for why $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{BY}$ implies $\boldsymbol{A\Sigma B} = \boldsymbol{0}$ in a).
c) See Driscoll and Krasnicka (1995).

The following theorem is a corollary of Craig's Theorem.

Theorem 2.10. Let $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{I}_n)$, with $\boldsymbol{A}$ and $\boldsymbol{B}$ symmetric. If $\boldsymbol{Y}^T\boldsymbol{AY} \sim \chi^2_r$ and $\boldsymbol{Y}^T\boldsymbol{BY} \sim \chi^2_d$, then $\boldsymbol{Y}^T\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{BY}$ iff $\boldsymbol{AB} = \boldsymbol{0}$.

Theorem 2.11. If $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} > 0$, then the population squared Mahalanobis distance $(\boldsymbol{Y} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) \sim \chi^2_n$.

Proof. Let $\boldsymbol{Z} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{Y} - \boldsymbol{\mu}) \sim N_n(\boldsymbol{0}, \boldsymbol{I})$. Then $\boldsymbol{Z} = (Z_1, ..., Z_n)^T$ where the Z_i are iid $N(0, 1)$. Hence $(\boldsymbol{Y} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) = \boldsymbol{Z}^T\boldsymbol{Z} = \sum_{i=1}^n Z_i^2 \sim \chi^2_n$. □

For large sample theory, the noncentral χ^2 distribution is important. If $Z_1, ..., Z_n$ are independent $N(0, 1)$ random variables, then $\sum_{i=1}^n Z_i^2 \sim \chi^2_n$. The noncentral $\chi^2(n, \gamma)$ distribution is the distribution of $\sum_{i=1}^n Y_i^2$ where $Y_1, ..., Y_n$ are independent $N(\mu_i, 1)$ random variables. Note that if $Y \sim N(\mu, 1)$, then $Y^2 \sim \chi^2(n = 1, \gamma = \mu^2/2)$, and if $Y \sim N(\sqrt{2\gamma}, 1)$, then $Y^2 \sim \chi^2(n = 1, \gamma)$.

Definition 2.12. Suppose $Y_1, ..., Y_n$ are independent $N(\mu_i, 1)$ random variables so that $\boldsymbol{Y} = (Y_1, ..., Y_n)^T \sim N_n(\boldsymbol{\mu}, \boldsymbol{I}_n)$. Then $\boldsymbol{Y}^T\boldsymbol{Y} = \sum_{i=1}^n Y_i^2 \sim \chi^2(n, \gamma = \boldsymbol{\mu}^T\boldsymbol{\mu}/2)$, a *noncentral* $\chi^2(n, \gamma)$ *distribution*, with n degrees of freedom and *noncentrality parameter* $\gamma = \boldsymbol{\mu}^T\boldsymbol{\mu}/2 = \frac{1}{2}\sum_{i=1}^n \mu_i^2 \geq 0$. The noncentrality parameter $\delta = \boldsymbol{\mu}^T\boldsymbol{\mu} = 2\gamma$ is also used. If $W \sim \chi^2_n$, then $W \sim \chi^2(n, 0)$ so $\gamma = 0$. The χ^2_n distribution is also called the *central* χ^2 *distribution.*

Some of the proof ideas for the following theorem came from Marden (2012, pp. 48, 96-97). Recall that if $Y_1, ..., Y_k$ are independent with moment generating functions (mgfs) $m_{Y_i}(t)$, then the mgf of $\sum_{i=1}^k Y_i$ is $m_{\sum_{i=1}^k Y_i}(t) = \prod_{i=1}^k m_{Y_i}(t)$. If $Y \sim \chi^2(n,\gamma)$, then the probability density function (pdf) of Y is rather hard to use, but is given by

$$f(y) = \sum_{j=0}^{\infty} \frac{e^{-\gamma}\gamma^j}{j!} \; \frac{y^{\frac{n}{2}+j-1}e^{-y/2}}{2^{\frac{n}{2}+j}\Gamma(\frac{n}{2}+j)} = \sum_{j=0}^{\infty} p_\gamma(j) f_{n+2j}(y)$$

where $p_\gamma(j) = P(W = j)$ is the probability mass function of a Poisson(γ) random variable W, and $f_{n+2j}(y)$ is the pdf of a χ^2_{n+2j} random variable. If $\gamma = 0$, define $\gamma^0 = 1$ in the first sum, and $p_0(0) = 1$ with $p_0(j) = 0$ for $j > 0$ in the second sum. For computing moments and the moment generating function, the integration and summation operations can be interchanged. Hence $\int_0^\infty f(y)dy = \sum_{j=0}^\infty p_\gamma(j) \int_0^\infty f_{n+2j}(y)dy = \sum_{j=0}^\infty p_\gamma(j) = 1$. Similarly, if $m_{n+2j}(t) = (1-2t)^{-(n+2j)/2}$ is the mgf of a χ^2_{n+2j} random variable, then the mgf of Y is $m_Y(t) = E(e^{tY}) = \int_0^\infty e^{ty} f(y)dy = \sum_{j=0}^\infty p_\gamma(j) \int_0^\infty e^{ty} f_{n+2j}(y)dy = \sum_{j=0}^\infty p_\gamma(j) m_{n+2j}(t)$.

Theorem 2.12. a) If $Y \sim \chi^2(n,\gamma)$, then the moment generating function of Y is $m_Y(t) = (1-2t)^{-n/2}\exp(-\gamma[1-(1-2t)^{-1}]) =$
$(1-2t)^{-n/2}\exp[2\gamma t/(1-2t)]$ for $t < 0.5$.

b) If $Y_i \sim \chi^2(n_i,\gamma_i)$ are independent for $i = 1,...,k$, then
$\sum_{i=1}^k Y_i \sim \chi^2\left(\sum_{i=1}^k n_i, \sum_{i=1}^k \gamma_i\right)$.

c) If $Y \sim \chi^2(n,\gamma)$, then $E(Y) = n + 2\gamma$ and $V(Y) = 2n + 8\gamma$.

Proof. Two proofs are given. a) i) From the above remarks, and using $e^x = \sum_{j=0}^\infty \frac{x^j}{j!}$, $m_Y(t) = \sum_{j=0}^\infty \frac{e^{-\gamma}\gamma^j}{j!}(1-2t)^{-(n+2j)/2} = (1-2t)^{-n/2}\sum_{j=0}^\infty \frac{e^{-\gamma}\left(\frac{\gamma}{1-2t}\right)^j}{j!} =$

$$(1-2t)^{-n/2}\exp\left(-\gamma + \frac{\gamma}{1-2t}\right) = (1-2t)^{-n/2}\exp\left(\frac{2\gamma t}{1-2t}\right).$$

ii) Let $W \sim N(\sqrt{\delta},1)$ where $\delta = 2\gamma$. Then $W^2 \sim \chi^2(1,\delta/2) = \chi^2(1,\gamma)$. Let $W \perp\!\!\!\perp X$ where $X \sim \chi^2_{n-1} \sim \chi^2(n-1,0)$, and let $Y = W^2 + X \sim \chi^2(n,\gamma)$ by b). Then $m_{W^2}(t) =$

$$E(e^{tW^2}) = \int_{-\infty}^{\infty} e^{tw^2}\frac{1}{\sqrt{2\pi}}\exp\left[\frac{-1}{2}(w-\sqrt{\delta})^2\right]dw =$$

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left[\frac{2}{2}tw^2 - \frac{1}{2}(w^2 - 2\sqrt{\delta}\;w + \delta)\right]dw =$$

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left[\frac{-1}{2}(w^2 - 2tw^2 - 2\sqrt{\delta}\;w + \delta)\right]dw =$$

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-1}{2}(w^2(1-2t) - 2\sqrt{\delta}w + \delta)\right] dw = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-1}{2}A\right] dw$$

where $A = [\sqrt{1-2t}\ \ (w-b)]^2 + c$ with

$$b = \frac{\sqrt{\delta}}{1-2t} \quad \text{and} \quad \text{c} = \frac{-2\text{t}\delta}{1-2\text{t}}$$

after algebra. Hence $m^2_W(t) =$

$$e^{-c/2}\sqrt{\frac{1}{1-2t}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\frac{1}{1-2t}}} \exp\left[\frac{-1}{2} \frac{1}{\frac{1}{1-2t}}(w-b)^2\right] dw = e^{-c/2}\sqrt{\frac{1}{1-2t}}$$

since the integral $= 1 = \int_{-\infty}^{\infty} f(w)dw$ where $f(w)$ is the $N(b, 1/(1-2t))$ pdf. Thus

$$m_{W^2}(t) = \frac{1}{\sqrt{1-2t}} \exp\left(\frac{t\delta}{1-2t}\right).$$

So $m_Y(t) = m_{W^2+X}(t) = m_{W^2}(t) m_X(t) =$

$$\frac{1}{\sqrt{1-2t}} \exp\left(\frac{t\delta}{1-2t}\right) \left(\frac{1}{1-2t}\right)^{(n-1)/2} = \frac{1}{(1-2t)^{n/2}} \exp\left(\frac{t\delta}{1-2t}\right) =$$

$$(1-2t)^{-n/2} \exp\left(\frac{2\gamma t}{1-2t}\right).$$

b) i) By a), $m_{\sum_{i=1}^k Y_i}(t) =$

$$\prod_{i=1}^k m_{Y_i}(t) = \prod_{i=1}^k (1-2t)^{-n_i/2} \exp(-\gamma_i[1-(1-2t)^{-1}]) =$$

$$(1-2t)^{-\sum_{i=1}^k n_i/2} \quad \exp\left(-\sum_{i=1}^k \gamma_i[1-(1-2t)^{-1}]\right),$$

the $\chi^2\left(\sum_{i=1}^k n_i, \sum_{i=1}^k \gamma_i\right)$ mgf.

ii) Let $Y_i = \boldsymbol{Z}_i^T \boldsymbol{Z}_i$ where the $\boldsymbol{Z}_i \sim N_{n_i}(\boldsymbol{\mu}_i, \boldsymbol{I}_{n_i})$ are independent. Let

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}_1 \\ \boldsymbol{Z}_2 \\ \vdots \\ \boldsymbol{Z}_k \end{pmatrix} \sim N_{\sum_{i=1}^k n_i}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \vdots \\ \boldsymbol{\mu}_k \end{pmatrix}, \boldsymbol{I}_{\sum_{i=1}^k n_i}\right] \sim N_{\sum_{i=1}^k n_i}(\boldsymbol{\mu}_{\boldsymbol{Z}}, \boldsymbol{I}_{\sum_{i=1}^k n_i}).$$

Then $\boldsymbol{Z}^T\boldsymbol{Z} = \sum_{i=1}^k \boldsymbol{Z}_i^T\boldsymbol{Z}_i = \sum_{i=1}^k Y_i \sim \chi^2\left(\sum_{i=1}^k n_i, \gamma_{\boldsymbol{Z}}\right)$ where

$$\gamma_{\boldsymbol{Z}} = \frac{\boldsymbol{\mu}_{\boldsymbol{Z}}^T\boldsymbol{\mu}_{\boldsymbol{Z}}}{2} = \sum_{i=1}^k \frac{\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i}{2} = \sum_{i=1}^k \gamma_i.$$

c) i) Let $W \sim \chi^2(1,\gamma) \perp\!\!\!\perp X \sim \chi^2_{n-1} \sim \chi^2(n-1,0)$. Then by b) $Y = W+X \sim \chi^2(n,\gamma)$. Let $Z \sim N(0,1)$ and $\delta = 2\gamma$. Then $\sqrt{\delta}+Z \sim N(\sqrt{\delta},1)$, and $W = (\sqrt{\delta}+Z)^2$. Thus $E(W) = E[(\sqrt{\delta}+Z)^2] = \delta+2\sqrt{\delta}E(Z)+E(Z^2) = \delta+1$. Using the binomial theorem

$$(x+y)^n = \sum_{i=0}^n \binom{n}{i} x^i y^{n-i}$$

with $x = \sqrt{\delta}$, $y = Z$, and $n = 4$, $E(W^2) = E[(\sqrt{\delta}+Z)^4] =$

$$E[\delta^2 + 4\delta^{3/2}Z + 6\delta Z^2 + 4\sqrt{\delta}Z^3 + Z^4] = \delta^2 + 6\delta + 3$$

since $E(Z) = E(Z^3) = 0$, and $E(Z^4) = 3$ by Problem 2.8. Hence $V(W) = E(W^2) - [E(W)]^2 = \delta^2+6\delta+3-(\delta+1)^2 = \delta^2+6\delta+3-\delta^2-2\delta-1 = 4\delta+2$. Thus $E(Y) = E(W) + E(X) = \delta + 1 + n - 1 = n + \delta = n + 2\gamma$, and $V(Y) = V(W) + V(X) = 4\delta + 2 + 2(n-1) = 8\delta + 2n$.

ii) Let $Z_i \sim N(\mu_i, 1)$ so $E(Z_i^2) = \sigma^2 + \mu_i^2 = 1 + \mu_i^2$. By Problem 2.8, $E(Z_i^3) = \mu_i^3 + 3\mu_i$, and $E(Z_i^4) = \mu_i^4 + 6\mu_i^2 + 3$. Hence $Y \sim \chi^2(n,\gamma)$ where $Y = \boldsymbol{Z}^T\boldsymbol{Z} = \sum_{i=1}^n Z_i^2$ where $\boldsymbol{Z} \sim N_n(\boldsymbol{\mu}, \boldsymbol{I})$. So $E(Y) = \sum_{i=1}^n E(Z_i^2) = \sum_{i=1}^n (1+\mu_i^2) = n + \boldsymbol{\mu}^T\boldsymbol{\mu} = n + 2\gamma$, and $V(Y) = \sum_{i=1}^n V(Z_i^2) =$

$$\sum_{i=1}^n [E(Z_i^4) - (E[Z_i^2])^2] = \sum_{i=1}^n [\mu_i^4 + 6\mu_i^2 + 3 - \mu_i^4 - 2\mu_i^2 - 1] = \sum_{i=1}^n [4\mu_i^2 + 2]$$

$= 2n + 4\boldsymbol{\mu}^T\boldsymbol{\mu} = 2n + 8\gamma$. □

For the following theorem, see Searle (1971, p. 57). Most of the results in Theorem 2.14 are corollaries of Theorem 2.13. Recall that the matrix in a quadratic form is symmetric, unless told otherwise.

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

For the following theorem, note that if $\boldsymbol{A} = \boldsymbol{A}^T = \boldsymbol{A}^2$, then $\boldsymbol{A}$ is a projection matrix since $\boldsymbol{A}$ is symmetric and idempotent. An $n \times n$ projection matrix $\boldsymbol{A}$ is not a full rank matrix unless $\boldsymbol{A} = \boldsymbol{I}_n$. See Theorem 2.2 j) and k). Often results are given for $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{I})$, and then the $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ case is handled as in c) and g) below, since $\boldsymbol{Y}/\sigma \sim N_n(\boldsymbol{0}, \boldsymbol{I})$.

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 $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \sim \chi^2_r$ 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

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

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

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

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

Note that $\boldsymbol{A}$ is a projection matrix iff $\boldsymbol{A}$ is idempotent in b) since $\boldsymbol{A}$ is symmetric. Thus b) is a special case d). To see that c) holds, note $\boldsymbol{Z} = \boldsymbol{Y}/\sigma \sim N_n(\boldsymbol{0}, \boldsymbol{I})$. Hence by b)

$$\frac{\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y}}{\sigma^2} = \boldsymbol{Z}^T\boldsymbol{A}\boldsymbol{Z} \sim \chi^2_r$$

iff $\boldsymbol{A}$ is idempotent of rank r. Much of Theorem 2.14 follows from Theorem 2.13. For f), we give another proof from Christensen (1987, p. 8). Since $\boldsymbol{A}$ is a projection matrix with $\text{rank}(\boldsymbol{A}) = r$, let $\{\boldsymbol{b}_1, ..., \boldsymbol{b}_r\}$ be an orthonormal basis for $C(\boldsymbol{A})$ and let $\boldsymbol{B} = [\boldsymbol{b}_1\ \boldsymbol{b}_2\ ...\ \boldsymbol{b}_r]$. Then $\boldsymbol{B}^T\boldsymbol{B} = \boldsymbol{I}_r$ and the projection matrix $\boldsymbol{A} = \boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T = \boldsymbol{B}\boldsymbol{B}^T$. Thus $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{Y} = \boldsymbol{Z}^T\boldsymbol{Z}$ where $\boldsymbol{Z} = \boldsymbol{B}^T\boldsymbol{Y} \sim N_r(\boldsymbol{B}^T\boldsymbol{\mu}, \boldsymbol{B}^T\boldsymbol{I}\boldsymbol{B}) \sim N_r(\boldsymbol{B}^T\boldsymbol{\mu}, \boldsymbol{I}_r)$. Thus $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} = \boldsymbol{Z}^T\boldsymbol{Z} \sim \chi^2(r, \boldsymbol{\mu}^T\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\mu}/2) \sim \chi^2(r, \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}/2)$ by Definition 2.12.

The following theorem is useful for constructing ANOVA tables. See Searle (1971, pp. 60-61).

Theorem 2.15: Generalized Cochran's Theorem. Let $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}_i = \boldsymbol{A}_i^T$ have rank r_i for $i = 1, ..., k$, and let $\boldsymbol{A} = \sum_{i=1}^k \boldsymbol{A}_i = \boldsymbol{A}^T$ have

rank r. Then $\boldsymbol{Y}^T\boldsymbol{A}_i\boldsymbol{Y} \sim \chi^2(r_i, \boldsymbol{\mu}^T\boldsymbol{A}_i\boldsymbol{\mu}/2)$, and the $\boldsymbol{Y}^T\boldsymbol{A}_i\boldsymbol{Y}$ are independent, and $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \sim \chi^2(r, \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}/2)$, iff
I) any 2 of a) $\boldsymbol{A}_i\boldsymbol{\Sigma}$ are idempotent $\forall i$,
b) $\boldsymbol{A}_i\boldsymbol{\Sigma}\boldsymbol{A}_j = \boldsymbol{0} \;\; \forall i < j$,
c) $\boldsymbol{A}\boldsymbol{\Sigma}$ is idempotent
are true; or II) c) is true and d) $r = \sum_{i=1}^k r_i$;
or III) c) is true and e) $\boldsymbol{A}_1\boldsymbol{\Sigma}, .., \boldsymbol{A}_{k-1}\boldsymbol{\Sigma}$ are idempotent and $\boldsymbol{A}_k\boldsymbol{\Sigma} \geq 0$ is singular.

2.3 Least Squares Theory

Definition 2.13. Estimating equations are used to find estimators of unknown parameters. The least squares criterion and log likelihood for maximum likelihood estimators are important examples.

Estimating equations are often used with a model, like $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, and often have a variable $\boldsymbol{\beta}$ that is used in the equations to find the estimator $\hat{\boldsymbol{\beta}}$ of the vector of parameters in the model. For example, the log likelihood $\log(L(\boldsymbol{\beta}, \sigma^2))$ has $\boldsymbol{\beta}$ and σ^2 as variables for a parametric statistical model where $\boldsymbol{\beta}$ and σ^2 are fixed unknown parameters, and maximizing the log likelihood with respect to these variables gives the maximum likelihood estimators of the parameters $\boldsymbol{\beta}$ and σ^2. So the term $\boldsymbol{\beta}$ is both a variable in the estimating equations, which could be replaced by another variable such as $\boldsymbol{\eta}$, and a vector of parameters in the model. In the theorem below, we could replace $\boldsymbol{\eta}$ by $\boldsymbol{\beta}$ where $\boldsymbol{\beta}$ is a vector of parameters in the linear model and a variable in the least squares criterion which is an estimating equation.

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**
$\sum_{i=1}^n r_i^2(\boldsymbol{\eta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}\|^2$.

Proof. Following Seber and Lee (2003, pp. 36-38), let $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{\theta}} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y} \in C(\boldsymbol{X})$, $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Y} \in [C(\boldsymbol{X})]^{\perp}$, and $\boldsymbol{\theta} \in C(\boldsymbol{X})$. Then $(\boldsymbol{Y} - \hat{\boldsymbol{\theta}})^T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = (\boldsymbol{Y} - \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y})^T(\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y} - \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{\theta}) = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{Y} - \boldsymbol{\theta}) = 0$ since $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{\theta} = \boldsymbol{\theta}$. Thus $\|\boldsymbol{Y} - \boldsymbol{\theta}\|^2 = (\boldsymbol{Y} - \hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T(\boldsymbol{Y} - \hat{\boldsymbol{\theta}} + \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) =$

$$\|\boldsymbol{Y} - \hat{\boldsymbol{\theta}}\|^2 + \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 + 2(\boldsymbol{Y} - \hat{\boldsymbol{\theta}})^T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \geq \|\boldsymbol{Y} - \hat{\boldsymbol{\theta}}\|^2$$

with equality iff $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 = 0$ iff $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\eta}$. Since $\hat{\boldsymbol{\theta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ the result follows. □

Definition 2.14. The **normal equations** are

$$\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}.$$

To see that the normal equations hold, note that $\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} \perp C(\boldsymbol{X})$ by Theorem 1.2 c) (and Theorem 2.20 i)). Thus $\boldsymbol{r} \in [C(\boldsymbol{X})]^\perp = N(\boldsymbol{X}^T)$, and $\boldsymbol{X}^T(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \boldsymbol{0}$. Hence $\boldsymbol{X}^T\hat{\boldsymbol{Y}} = \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$.

The maximum likelihood estimator uses the log likelihood as an estimating equation. Note that it is crucial to observe that the likelihood function is a function of $\boldsymbol{\theta}$ (and that $y_1, ..., y_n$ act as fixed constants). Also, if the MLE $\hat{\boldsymbol{\theta}}$ exists, then $\hat{\boldsymbol{\theta}} \in \Theta$, the parameter space.

Definition 2.15. Let $f(\boldsymbol{y}|\boldsymbol{\theta})$ be the joint pdf of $Y_1, ..., Y_n$. If $\boldsymbol{Y} = \boldsymbol{y}$ is observed, then **the likelihood function** $L(\boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{\theta})$. For each sample point $\boldsymbol{y} = (y_1, ..., y_n)$, let $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ be a parameter value at which $L(\boldsymbol{\theta}|\boldsymbol{y})$ attains its maximum as a function of $\boldsymbol{\theta}$ with $\boldsymbol{y}$ held fixed. Then a maximum likelihood estimator (**MLE**) of the parameter $\boldsymbol{\theta}$ based on the sample $\boldsymbol{Y}$ is $\hat{\boldsymbol{\theta}}(\boldsymbol{Y})$.

Definition 2.16. Let the log likelihood of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ be $\log[L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)]$. If $\hat{\boldsymbol{\theta}}_2$ is the MLE of $\boldsymbol{\theta}_2$, then the *log profile likelihood* is $\log[L_p(\boldsymbol{\theta}_1)] = \log[L(\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_2)]$.

We can often fix σ and then show $\hat{\boldsymbol{\beta}}$ is the MLE by direct maximization. Then the MLE $\hat{\sigma}$ or $\hat{\sigma}^2$ can be found by maximizing the log profile likelihood function $\log[L_p(\sigma)]$ or $\log[L_p(\sigma^2)]$ where $L_p(\sigma) = L(\sigma, \boldsymbol{\beta} = \hat{\boldsymbol{\beta}})$.

Remark 2.1. a) Know how to find the max and min of a function h that is continuous on an interval [a,b] and differentiable on (a, b). Solve $h'(x) \equiv 0$ and find the places where $h'(x)$ does not exist. These values are the **critical points**. Evaluate h at a, b, and the critical points. One of these values will be the min and one the max.

b) Assume h is continuous. Then a critical point θ_o is a local max of $h(\theta)$ if h is increasing for $\theta < \theta_o$ in a neighborhood of θ_o and if h is decreasing for $\theta > \theta_o$ in a neighborhood of θ_o. The first derivative test is often used.

c) If h is strictly concave $\left(\frac{d^2}{d\theta^2}h(\theta) < 0 \text{ for all } \theta\right)$, then any local max of h is a global max.

d) Suppose $h'(\theta_o) = 0$. The 2nd derivative test states that if $\frac{d^2}{d\theta^2}h(\theta_o) < 0$, then θ_o is a local max.

e) If $h(\theta)$ is a continuous function on an interval with endpoints $a < b$ (not necessarily finite), and differentiable on (a, b) and if the **critical point is unique**, then the critical point is a **global maximum** if it is a local maximum (because otherwise there would be a local minimum and the critical point would not be unique). To show that $\hat{\theta}$ is the MLE (the global maximizer of $h(\theta) = \log L(\theta)$), show that $\log L(\theta)$ is differentiable on (a, b). Then show that $\hat{\theta}$ is the unique solution to the equation $\frac{d}{d\theta}\log L(\theta) = 0$ and that the

2nd derivative evaluated at $\hat{\theta}$ is negative: $\frac{d^2}{d\theta^2} \log L(\theta)|_{\hat{\theta}} < 0$. Similar remarks hold for finding $\hat{\sigma}^2$ using the profile likelihood.

Theorem 2.17. Let $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e} = \hat{\boldsymbol{Y}} + \boldsymbol{r}$ where $\boldsymbol{X}$ is full rank, and $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I})$. Then the MLE of $\boldsymbol{\beta}$ is the least squares estimator $\hat{\boldsymbol{\beta}}$ and the MLE of σ^2 is $RSS/n = (n-p)MSE/n$.

Proof. The $Y_i = Y_i|\boldsymbol{x}_i$ are independent $N(\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma^2)$ random variables with probability density functions (pdfs) $f_{Y_i}(y_i)$. Let y_i be the observed values of Y_i. Thus the likelihood function

$$L(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^n f_{Y_i}(y_i) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2\right) =$$

$$(2\pi\sigma^2)^{-n/2} \exp\left(\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2\right) = (2\pi\sigma^2)^{-n/2} \exp\left(\frac{-1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2\right).$$

The least squares criterion $Q(\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 = \sum_{i=1}^n r_i^2(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 = (\boldsymbol{y} - \boldsymbol{X\beta})^T(\boldsymbol{y} - \boldsymbol{X\beta})$. For fixed σ^2, maximizing the likelihood is equivalent to maximizing

$$\exp\left(\frac{-1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2\right),$$

which is equivalent to minimizing $\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2$. But the least squares estimator minimizes $\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2$ by Theorem 2.16. Hence $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$.

Let $Q = \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2$. Then the MLE of σ^2 can be found by maximizing the log profile likelihood $\log(L_P(\sigma^2))$ where

$$L_P(\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(\frac{-1}{2\sigma^2}Q\right).$$

Let $\tau = \sigma^2$. Then

$$\log(L_p(\sigma^2)) = c - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}Q,$$

and

$$\log(L_p(\tau)) = c - \frac{n}{2}\log(\tau) - \frac{1}{2\tau}Q.$$

Hence

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

or $-n\tau + Q = 0$ or $n\tau = Q$ or

$$\hat{\tau} = \frac{Q}{n} = \hat{\sigma}^2 = \frac{\sum_{i=1}^n r_i^2}{n} = \frac{n-p}{n}MSE,$$

which is a unique solution.

Now

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

Thus by Remark 2.1, $\hat{\sigma}^2$ is the MLE of σ^2. □

Now assume the $n \times p$ matrix $\boldsymbol{X}$ has full rank p. There are two ways to compute $\hat{\boldsymbol{\beta}}$. Use $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, and use sample covariance matrices. The population OLS coefficients are defined below. Let $\boldsymbol{x}_i^T = (1, \boldsymbol{u}_i^T)$ where $\boldsymbol{u}_i$ is the vector of nontrivial predictors. Let $\frac{1}{n}\sum_{j=1}^n X_{jk} = \overline{X}_{ok} = \overline{u}_{ok}$ for $k = 2, ..., p$. The subscript "ok" means sum over the first subscript j. Let $\overline{\boldsymbol{u}} = (\overline{u}_{o,2}, ..., \overline{u}_{o,p})^T$ be the sample mean of the $\boldsymbol{u}_i$. Note that regressing on $\boldsymbol{u}$ is equivalent to regressing on $\boldsymbol{x}$ if there is an intercept β_1 in the model.

Definition 2.17. Using the above notation, let $\boldsymbol{x}_i^T = (1, \boldsymbol{u}_i^T)$, and let $\boldsymbol{\beta}^T = (\beta_1, \boldsymbol{\beta}_2^T)$ where β_1 is the intercept and the slopes vector $\boldsymbol{\beta}_2 = (\beta_2, ..., \beta_p)^T$. Let the population covariance matrices

$$\text{Cov}(\boldsymbol{u}) = E[(\boldsymbol{u} - E(\boldsymbol{u}))(\boldsymbol{u} - E(\boldsymbol{u}))^T] = \boldsymbol{\Sigma}_{\boldsymbol{u}}, \quad \text{and}$$

$$\text{Cov}(\boldsymbol{u}, Y) = E[(\boldsymbol{u} - E(\boldsymbol{u}))(Y - E(Y))] = \boldsymbol{\Sigma}_{\boldsymbol{u}Y}.$$

Then the population coefficients from an OLS regression of Y on $\boldsymbol{x}$ (even if a linear model does not hold) are

$$\beta_1 = E(Y) - \boldsymbol{\beta}_2^T E(\boldsymbol{u}) \quad \text{and} \quad \boldsymbol{\beta}_2 = \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{u}\mathrm{Y}}.$$

Definition 2.18. Let the sample covariance matrices be

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}} = \frac{1}{n-1}\sum_{i=1}^n (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T \text{ and } \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \frac{1}{n-1}\sum_{i=1}^n (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(Y_i - \overline{Y}).$$

Let the method of moments or maximum likelihood estimators be $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}} = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(\boldsymbol{u}_i - \overline{\boldsymbol{u}})^T$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{u}_i - \overline{\boldsymbol{u}})(Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{u}_i Y_i - \overline{\boldsymbol{u}}\,\overline{Y}$.

Refer to Definitions 1.27, 1.28, and 1.33 for the notation "$\hat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}$ as $n \to \infty$," which means that $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$, or that $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}$. Note that $\boldsymbol{D} = \boldsymbol{X}_1^T\boldsymbol{X}_1 - n\overline{\boldsymbol{u}}\,\overline{\boldsymbol{u}}^T = (n-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}$.

Theorem 2.18: Seber and Lee (2003, p. 106). Let $\boldsymbol{X} = (\boldsymbol{1} \quad \boldsymbol{X}_1)$. Then $\boldsymbol{X}^T\boldsymbol{Y} = \begin{pmatrix} n\overline{Y} \\ \boldsymbol{X}_1^T\boldsymbol{Y} \end{pmatrix} = \begin{pmatrix} n\overline{Y} \\ \sum_{i=1}^n \boldsymbol{u}_i Y_i \end{pmatrix}$, $\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} n & n\overline{\boldsymbol{u}}^T \\ n\overline{\boldsymbol{u}} & \boldsymbol{X}_1^T\boldsymbol{X}_1 \end{pmatrix}$,

$$\text{and} \quad (\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} \frac{1}{n} + \overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{u}} & -\overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{u}} & \boldsymbol{D}^{-1} \end{pmatrix}$$

where the $(p-1) \times (p-1)$ matrix $\boldsymbol{D}^{-1} = [(n-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}]^{-1} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}/(n-1)$.

Theorem 2.19: Second way to compute $\hat{\boldsymbol{\beta}}$:
a) If $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}$ exists, then $\hat{\beta}_1 = \overline{Y} - \hat{\boldsymbol{\beta}}_2^T\overline{\boldsymbol{u}}$ and

$$\hat{\boldsymbol{\beta}}_2 = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y}.$$

b) Suppose that $(Y_i, \boldsymbol{u}_i^T)^T$ are iid random vectors such that σ_Y^2, $\boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}$, and $\boldsymbol{\Sigma}_{\boldsymbol{u}Y}$ exist. Then $\hat{\beta}_1 \xrightarrow{P} \beta_1$ and

$$\hat{\boldsymbol{\beta}}_2 \xrightarrow{P} \boldsymbol{\beta}_2 \quad \text{as} \quad \text{n} \to \infty.$$

Proof. Note that

$$\boldsymbol{Y}^T\boldsymbol{X}_1 = (Y_1 \cdots Y_n)\begin{bmatrix} \boldsymbol{u}_1^T \\ \vdots \\ \boldsymbol{u}_n^T \end{bmatrix} = \sum_{i=1}^n Y_i\boldsymbol{u}_i^T$$

and

$$\boldsymbol{X}_1^T\boldsymbol{Y} = [\boldsymbol{u}_1 \cdots \boldsymbol{u}_n]\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \sum_{i=1}^n \boldsymbol{u}_i Y_i.$$

So

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{n} + \overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{u}} & -\overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{u}} & \boldsymbol{D}^{-1} \end{bmatrix}\begin{bmatrix} \boldsymbol{1}^T \\ \boldsymbol{X}_1^T \end{bmatrix}\boldsymbol{Y} =$$

$$\begin{bmatrix} \frac{1}{n} + \overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{u}} & -\overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{u}} & \boldsymbol{D}^{-1} \end{bmatrix}\begin{bmatrix} n\overline{Y} \\ \boldsymbol{X}_1^T\boldsymbol{Y} \end{bmatrix}.$$

Thus $\hat{\boldsymbol{\beta}}_2 = -n\boldsymbol{D}^{-1}\overline{\boldsymbol{u}}\,\overline{Y} + \boldsymbol{D}^{-1}\boldsymbol{X}_1^T\boldsymbol{Y} = \boldsymbol{D}^{-1}(\boldsymbol{X}_1^T\boldsymbol{Y} - n\overline{\boldsymbol{u}}\,\overline{Y}) =$

$$\boldsymbol{D}^{-1}\left[\sum_{i=1}^n \boldsymbol{u}_i Y_i - n\overline{\boldsymbol{u}}\,\overline{Y}\right] = \frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}}{n-1}n\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y}. \quad \text{Then}$$

$\hat{\beta}_1 = \overline{Y} + n\overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{u}}\,\overline{Y} - \overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1}\boldsymbol{X}_1^T\boldsymbol{Y} = \overline{Y} + [n\overline{Y}\overline{\boldsymbol{u}}^T\boldsymbol{D}^{-1} - \boldsymbol{Y}^T\boldsymbol{X}_1\boldsymbol{D}^{-1}]\overline{\boldsymbol{u}}$ $= \overline{Y} - \hat{\boldsymbol{\beta}}_2^T\boldsymbol{u}$. The convergence in probability results hold since sample means and sample covariance matrices are consistent estimators of the population means and population covariance matrices. □

It is important to note that the convergence in probability results are for iid $(Y_i, \boldsymbol{u}_i^T)^T$ with second moments and nonsingular $\boldsymbol{\Sigma}_{\boldsymbol{u}}$: a linear model

$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ does not need to hold. Also, $\boldsymbol{X}$ is a random matrix, and the least squares regression is conditional on $\boldsymbol{X}$. When the linear model does hold, the second method for computing $\hat{\boldsymbol{\beta}}$ is still valid even if $\boldsymbol{X}$ is a constant matrix, and $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}$ by the LS CLT. Some properties of the least squares estimators and related quantities are given below, where $\boldsymbol{X}$ is a constant matrix. The population results of Definition 2.17 were also shown when

$$\begin{bmatrix} Y \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \sim N_p \left[\begin{pmatrix} E(Y) \\ E(\boldsymbol{u}) \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \boldsymbol{\Sigma}_{Y\boldsymbol{u}} \\ \boldsymbol{\Sigma}_{\boldsymbol{u}Y} & \boldsymbol{\Sigma}_{\boldsymbol{uu}} \end{pmatrix} \right]$$

in Remark 1.5. Also see Theorem 1.40. The following theorem is similar to Theorem 1.2.

Theorem 2.20. Let $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e} = \hat{\boldsymbol{Y}} + \boldsymbol{r}$ where $\boldsymbol{X}$ has full rank p, $E(\boldsymbol{e}) = \boldsymbol{0}$, and $\text{Cov}(\boldsymbol{e}) = \sigma^2 \boldsymbol{I}$. Let $\boldsymbol{P} = \boldsymbol{P}_{\boldsymbol{X}}$ be the projection matrix on $C(\boldsymbol{X})$ so $\hat{\boldsymbol{Y}} = \boldsymbol{PX}$, $\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$, and $\boldsymbol{PX} = \boldsymbol{X}$ so $\boldsymbol{X}^T\boldsymbol{P} = \boldsymbol{X}^T$.
i) The predictor variables and residuals are orthogonal. Hence the columns of $\boldsymbol{X}$ and the residual vector are orthogonal: $\boldsymbol{X}^T\boldsymbol{r} = \boldsymbol{0}$.
ii) $E(\boldsymbol{Y}) = \boldsymbol{X\beta}$.
iii) $\text{Cov}(\boldsymbol{Y}) = \text{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$.
iv) The fitted values and residuals are uncorrelated: $\text{Cov}(\boldsymbol{r}, \hat{\boldsymbol{Y}}) = \boldsymbol{0}$.
v) The least squares estimator $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$: $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.
vi) $\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

Proof. i) $\boldsymbol{X}^T\boldsymbol{r} = \boldsymbol{X}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y} = \boldsymbol{0Y} = \boldsymbol{0}$, while ii) and iii) are immediate.
iv) $\text{Cov}(\boldsymbol{r}, \hat{\boldsymbol{Y}}) = E([\boldsymbol{r} - E(\boldsymbol{r})][\hat{\boldsymbol{Y}} - E(\hat{\boldsymbol{Y}})]^T) =$

$$E([(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y} - (\boldsymbol{I} - \boldsymbol{P})E(\boldsymbol{Y})][\boldsymbol{PY} - \boldsymbol{P}E(\boldsymbol{Y})]^T) =$$

$$E[(\boldsymbol{I} - \boldsymbol{P})[\boldsymbol{Y} - E(\boldsymbol{Y})][\boldsymbol{Y} - E(\boldsymbol{Y})]^T\boldsymbol{P}] = (\boldsymbol{I} - \boldsymbol{P})\sigma^2\boldsymbol{IP} = \sigma^2(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{P} = \boldsymbol{0}.$$

v) $E(\hat{\boldsymbol{\beta}}) = E[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E[\boldsymbol{Y}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\beta$ $= \boldsymbol{\beta}$.
vi) $\text{Cov}(\hat{\boldsymbol{\beta}}) = \text{Cov}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}] = \text{Cov}(\boldsymbol{AY}) = \boldsymbol{A}\text{Cov}(\boldsymbol{Y})\boldsymbol{A}^T =$

$$\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{IX}(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}. \ \square$$

Definition 2.19. Let $\boldsymbol{a}, \boldsymbol{b}$, and $\boldsymbol{c}$ be $n \times 1$ constant vectors. A 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 for any other unbiased linear estimator $\boldsymbol{b}^T\boldsymbol{Y}$ of $\boldsymbol{c}^T\boldsymbol{\theta}$, $Var(\boldsymbol{a}^T\boldsymbol{Y}) \leq Var(\boldsymbol{b}^T\boldsymbol{Y})$.

The following theorem is useful for finding the BLUE when $\boldsymbol{X}$ has full rank. Note that if W is a random variable, then the covariance matrix of

W is $\text{Cov}(W) = \text{Cov}(W, W) = V(W)$. Note that the theorem shows that $\boldsymbol{b}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{b}^T\boldsymbol{PY} = \boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is the BLUE of $\boldsymbol{b}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{a}^T\boldsymbol{\beta}$ where $\boldsymbol{a}^T = \boldsymbol{b}^T\boldsymbol{X}$ and $\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta}$. Also, if $\boldsymbol{b}^T\boldsymbol{Y}$ is an unbiased estimator of $\boldsymbol{a}^T\boldsymbol{\beta} = \boldsymbol{b}^T\boldsymbol{X}\boldsymbol{\beta}$, then $\boldsymbol{b}^T\boldsymbol{PY} = \boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is a better unbiased estimator in that $V(\boldsymbol{b}^T\boldsymbol{PY}) \leq V(\boldsymbol{b}^T\boldsymbol{Y})$. Since $\boldsymbol{X}$ is full rank, $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable with BLUE $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ for every $p \times 1$ constant vector $\boldsymbol{A}$. Note that the e_i are uncorrelated with zero mean, but not necessarily independent or identically distributed in the following theorem. Note that if $\boldsymbol{b} = \boldsymbol{d} = \boldsymbol{Pb}$, then $\boldsymbol{Pb} = \boldsymbol{PPb} = \boldsymbol{Pb} = \boldsymbol{d}$. The proof of the more general Theorem 3.2 c) also proves Theorem 2.21.

Theorem 2.21: 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 $\boldsymbol{a}$.

Proof. Let $\boldsymbol{b}^T\boldsymbol{Y}$ be any linear unbiased estimator of $\boldsymbol{a}^T\boldsymbol{\beta}$. Then $E(\boldsymbol{b}^T\boldsymbol{Y}) = \boldsymbol{a}^T\boldsymbol{\beta} = \boldsymbol{b}^T E(\boldsymbol{Y}) = \boldsymbol{b}^T\boldsymbol{X}\boldsymbol{\beta}$ for any $\boldsymbol{\beta} \in \mathbb{R}^p$, the parameter space of $\boldsymbol{\beta}$. Hence $\boldsymbol{a}^T = \boldsymbol{b}^T\boldsymbol{X}$. The least squares estimator $\boldsymbol{a}^T\hat{\boldsymbol{\beta}} = \boldsymbol{a}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{d}^T\boldsymbol{Y} = \boldsymbol{b}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{b}^T\boldsymbol{PY}$ is a linear unbiased estimator of $\boldsymbol{a}^T\boldsymbol{\beta}$ since $E(\boldsymbol{a}^T\hat{\boldsymbol{\beta}}) = \boldsymbol{a}^T\boldsymbol{\beta}$. Now $V(\boldsymbol{b}^T\boldsymbol{Y}) - V(\boldsymbol{a}^T\hat{\boldsymbol{\beta}}) = V(\boldsymbol{b}^T\boldsymbol{Y}) - V(\boldsymbol{b}^T\boldsymbol{PY}) = \text{Cov}(\boldsymbol{b}^T\boldsymbol{Y}) - \text{Cov}(\boldsymbol{b}^T\boldsymbol{PY}) = \sigma^2\boldsymbol{b}^T\boldsymbol{b} - \sigma^2\boldsymbol{b}^T\boldsymbol{Pb} = \sigma^2\boldsymbol{b}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{b} = \sigma^2\boldsymbol{z}^T\boldsymbol{z} \geq 0$ with equality iff $\boldsymbol{z} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{b} = \boldsymbol{0}$ iff $\boldsymbol{b} = \boldsymbol{d} = \boldsymbol{Pb}$ iff $\boldsymbol{b}^T\boldsymbol{Y} = \boldsymbol{b}^T\boldsymbol{PY} = \boldsymbol{a}^T\hat{\boldsymbol{\beta}}$. Since $\boldsymbol{b}^T\boldsymbol{Y}$ was an arbitrary unbiased linear estimator, the least squares estimator $\boldsymbol{a}^T\hat{\boldsymbol{\beta}}$ is BLUE. □

Lai et al. (1979) note that if $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \to \boldsymbol{0}$ as $n \to \infty$, then $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$. Also see Zhang (2019). The following theorem gives some properties of the least squares estimators $\hat{\boldsymbol{\beta}}$ and MSE under the normal least squares model. Similar properties will be developed without the normality assumption.

Theorem 2.22. Suppose $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ is full rank, $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, and $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$.
a) $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$.
b) $\dfrac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\boldsymbol{X}^T\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\sigma^2} \sim \chi^2_p$.
c) $\hat{\boldsymbol{\beta}} \perp\!\!\!\perp MSE$.
d) $\dfrac{RSS}{\sigma^2} = \dfrac{(n-p)MSE}{\sigma^2} \sim \chi^2_{n-p}$.

Proof. Let $\boldsymbol{P} = \boldsymbol{P}_{\boldsymbol{X}}$.

a) Since $\boldsymbol{A} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is a constant matrix,

$$\hat{\boldsymbol{\beta}} = \boldsymbol{AY} \sim N_p(\boldsymbol{A}E(\boldsymbol{Y}), \boldsymbol{A}\text{Cov}(\boldsymbol{Y})\boldsymbol{A}^T) \sim$$

$$N_p((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}) \sim$$

$$N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}).$$

b) The population Mahalanobis distance of $\hat{\boldsymbol{\beta}}$ is

$$\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T \boldsymbol{X}^T \boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\sigma^2} = (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T [\text{Cov}(\hat{\boldsymbol{\beta}})]^{-1}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \sim \chi^2_p$$

by Theorem 2.11.

c) Since $\text{Cov}(\hat{\boldsymbol{\beta}}, \boldsymbol{r}) = \text{Cov}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}, (\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}) =$ $\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{I}(\boldsymbol{I}-\boldsymbol{P}) = \boldsymbol{0}$, $\hat{\boldsymbol{\beta}} \perp\!\!\!\perp \boldsymbol{r}$. Thus $\hat{\boldsymbol{\beta}} \perp\!\!\!\perp RSS = \|\boldsymbol{r}\|^2$, and $\hat{\boldsymbol{\beta}} \perp\!\!\!\perp MSE$.

d) Since $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$ and $\boldsymbol{X}^T\boldsymbol{P} = \boldsymbol{X}^T$, it follows that $\boldsymbol{X}^T(\boldsymbol{I}-\boldsymbol{P}) = \boldsymbol{0}$ and $(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{X} = \boldsymbol{0}$. Thus $RSS = \boldsymbol{r}^T\boldsymbol{r} = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y} =$

$$(\boldsymbol{Y}-\boldsymbol{X\beta})^T(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{Y}-\boldsymbol{X\beta}) = \boldsymbol{e}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{e}.$$

Since $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, then by Theorem 2.14 c), $\boldsymbol{e}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{e}/\sigma^2 \sim \chi^2_{n-p}$ where $n-p = rank(\boldsymbol{I}-\boldsymbol{P}) = tr(\boldsymbol{I}-\boldsymbol{P})$. □

2.3.1 Hypothesis Testing

Suppose $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ where rank$(\boldsymbol{X}) = p$, $E(\boldsymbol{e}) = \boldsymbol{0}$ and Cov$(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$. Let $\boldsymbol{L}$ be an $r \times p$ constant matrix with rank$(\boldsymbol{L}) = r$, let $\boldsymbol{c}$ be an $r \times 1$ constant vector, and consider testing $H_0 : \boldsymbol{L\beta} = \boldsymbol{c}$. First theory will be given for when $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. The large sample theory will be given for when the iid zero mean e_i have $V(e_i) = \sigma^2$. Note that the normal model will satisfy the large sample theory conditions.

The partial F test, and its special cases the ANOVA F test and the Wald t test, use $\boldsymbol{c} = \boldsymbol{0}$. Let the **full model** use Y, $x_1 \equiv 1$, $x_2, ..., x_p$, and let the **reduced model** use Y, $x_1 = x_{j_1} \equiv 1$, $x_{j_2}, ..., x_{j_k}$ where $\{j_1, ..., j_k\} \subset \{1, ..., p\}$ and $j_1 = 1$. Here $1 \le k < p$, and if $k = 1$, then the model is $Y_i = \beta_1 + e_i$. Hence the full model is $Y_i = \beta_1 + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + e_i$, while the reduced model is $Y_i = \beta_1 + \beta_{j_2} x_{i,j_2} + \cdots + \beta_{j_k} x_{i,j_k} + e_i$. In matrix form, the full model is $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ and the reduced model is $\boldsymbol{Y} = \boldsymbol{X}_R\boldsymbol{\beta}_R + \boldsymbol{e}_R$ where the columns of $\boldsymbol{X}_R$ are a proper subset of the columns of $\boldsymbol{X}$. i) The **partial F test** has $H_0 : \beta_{j_{k+1}} = \cdots = \beta_{j_p} = 0$, or H_0 : the reduced model is good, or $H_0 : \boldsymbol{L\beta} = \boldsymbol{0}$ where $\boldsymbol{L}$ is a $(p-k) \times p$ matrix where the ith row of $\boldsymbol{L}$ has a 1 in the j_{k+i}th position and zeroes elsewhere. In particular, if $\beta_1, ..., \beta_k$ are the only β_i in the reduced model, then $\boldsymbol{L} = [\boldsymbol{0} \;\; \boldsymbol{I}_{p-k}]$ and $\boldsymbol{0}$ is a $(p-k) \times k$ matrix. Hence $r = p - k =$ number of predictors in the full model but not in the reduced model. ii) The **ANOVA F test** is the special case of the partial F test where the reduced model is $Y_i = \beta_1 + \epsilon_i$. Hence $H_0 : \beta_2 = \cdots = \beta_p = 0$, or H_0 : none of the nontrivial predictors $x_2, ..., x_p$ are needed in the linear model, or $H_0 : \boldsymbol{L\beta} = \boldsymbol{0}$ where $\boldsymbol{L} = [\boldsymbol{0} \;\; \boldsymbol{I}_{p-1}]$ and $\boldsymbol{0}$ is a $(p-1) \times 1$ vector. Hence $r = p - 1$. iii) The **Wald t test** uses the reduced model that deletes the jth predictor from the full model. Hence $H_0 : \beta_j = 0$, or H_0 : the jth predictor x_j is not needed in the linear model given that the other predictors are in the model, or $H_0 : \boldsymbol{L}_j\boldsymbol{\beta} = 0$ where $\boldsymbol{L}_j = [0, ..., 0, 1, 0, ..., 0]$ is a $1 \times p$ row vector with a 1 in the jth position for $j = 1, ..., p$. Hence $r = 1$.

A way to get the test statistic F_R for the partial F test is to fit the full model and the reduced model. Let RSS be the RSS of the full model, and let $RSS(R)$ be the RSS of the reduced model. Similarly, let MSE and $MSE(R)$ be the MSE of the full and reduced models. Let $df_R = n - k$ and $df_F = n - p$ be the degrees of freedom for the reduced and full models. Then $F_R = \dfrac{RSS(R) - RSS}{rMSE}$ where $r = df_R - df_F = p - k =$ number of predictors in the full model but not in the reduced model.

If $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$, then

$$\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c} \sim N_r(\boldsymbol{L}\boldsymbol{\beta} - \boldsymbol{c}, \sigma^2\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T).$$

If H_0 is true then $\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c} \sim N_r(\boldsymbol{0}, \sigma^2\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)$, and by Theorem 2.11

$$rF_1 = \frac{1}{\sigma^2}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c})^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) \sim \chi^2_r.$$

Let $rF_R = \sigma^2 rF_1/MSE$. If H_0 is true, $rF_R \xrightarrow{D} \chi^2_r$ for a large class of zero mean error distributions. See Theorem 2.26 c).

From Definition 1.25, if $\boldsymbol{Z}_n \xrightarrow{D} \boldsymbol{Z}$ as $n \to \infty$, then $\boldsymbol{Z}_n$ converges in distribution to the random vector $\boldsymbol{Z}$, and "$\boldsymbol{Z}$ is the limiting distribution of $\boldsymbol{Z}_n$" means that the distribution of $\boldsymbol{Z}$ is the limiting distribution of $\boldsymbol{Z}_n$. The notation $\boldsymbol{Z}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ means $\boldsymbol{Z} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Remark 2.2. a) $\boldsymbol{Z}$ is the limiting distribution of $\boldsymbol{Z}_n$, and does not depend on the sample size n (since $\boldsymbol{Z}$ is found by taking the limit as $n \to \infty$).

b) When $\boldsymbol{Z}_n \xrightarrow{D} \boldsymbol{Z}$, the distribution of $\boldsymbol{Z}$ can be used to approximate probabilities $P(\boldsymbol{Z}_n \le \boldsymbol{c}) \approx P(\boldsymbol{Z} \le \boldsymbol{c})$ at continuity points $\boldsymbol{c}$ of the cdf $F_{\boldsymbol{Z}}(\boldsymbol{z})$. Often the limiting distribution is a continuous distribution, so all points $\boldsymbol{c}$ are continuity points.

c) Often the two quantities $\boldsymbol{Z}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{Z}_n \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ behave similarly. A big difference is that the distribution on the RHS (right hand side) can depend on n for $\sim$ but not for $\xrightarrow{D}$. In particular, if $\boldsymbol{Z}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{A}\boldsymbol{Z}_n + \boldsymbol{b} \xrightarrow{D} N_m(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$, provided the RHS does not depend on n, where $\boldsymbol{A}$ is an $m \times k$ constant matrix and $\boldsymbol{b}$ is an $m \times 1$ constant vector.

d) We often want a normal approximation where the RHS can depend on n. Write $\boldsymbol{Z}_n \sim AN_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for an approximate multivariate normal distribution where the RHS may depend on n. For normal linear model, if $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2 I)$, then $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$. If the e_i are iid with $E(e_i) = 0$ and $V(e_i) = \sigma^2$, use the multivariate normal approximation $\hat{\boldsymbol{\beta}} \sim AN_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$ or $\hat{\boldsymbol{\beta}} \sim AN_p(\boldsymbol{\beta}, MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1})$. The RHS depends on n since the number of rows of $\boldsymbol{X}$ is n.

Theorem 2.23. Suppose $\hat{\boldsymbol{\Sigma}}_n$ and $\boldsymbol{\Sigma}$ are positive definite and symmetric. If $\boldsymbol{W}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\hat{\boldsymbol{\Sigma}}_n \xrightarrow{P} \boldsymbol{\Sigma}$, then $\boldsymbol{Z}_n = \hat{\boldsymbol{\Sigma}}_n^{-1/2}(\boldsymbol{W}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{I})$, and $\boldsymbol{Z}_n^T \boldsymbol{Z}_n = (\boldsymbol{W}_n - \boldsymbol{\mu})^T \hat{\boldsymbol{\Sigma}}_n^{-1}(\boldsymbol{W}_n - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_k$.

Proof. $\boldsymbol{Z}_n = (\hat{\boldsymbol{\Sigma}}_n^{-1/2} - \boldsymbol{\Sigma}^{-1/2} + \boldsymbol{\Sigma}^{-1/2})(\boldsymbol{W}_n - \boldsymbol{\mu}) =$ $(\hat{\boldsymbol{\Sigma}}_n^{-1/2} - \boldsymbol{\Sigma}^{-1/2})(\boldsymbol{W}_n - \boldsymbol{\mu}) + \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{W}_n - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{0} + N_k(\boldsymbol{0}, \boldsymbol{I}) \sim N_k(\boldsymbol{0}, \boldsymbol{I})$ by Slutsky's Theorem 1.34 b). Hence $\boldsymbol{Z}_n^T \boldsymbol{Z}_n \xrightarrow{D} \chi^2_k$. □

See Remark 2.3 for why Theorem 2.24 is useful.

Theorem 2.24. If $W_n \sim F_{r,d_n}$ where the positive integer $d_n \to \infty$ as $n \to \infty$, then $rW_n \xrightarrow{D} \chi^2_r$.

Proof. If $X_1 \sim \chi^2_{d_1} \perp\!\!\!\perp X_2 \sim \chi^2_{d_2}$, then

$$\frac{X_1/d_1}{X_2/d_2} \sim F_{d_1,d_2}.$$

If $U_i \sim \chi^2_1$ are iid then $\sum_{i=1}^k U_i \sim \chi^2_k$. Let $d_1 = r$ and $k = d_2 = d_n$. Hence if $X_2 \sim \chi^2_{d_n}$, then

$$\frac{X_2}{d_n} = \frac{\sum_{i=1}^{d_n} U_i}{d_n} = \overline{U} \xrightarrow{P} E(U_i) = 1$$

by the law of large numbers. Hence if $W \sim F_{r,d_n}$, then $rW_n \xrightarrow{D} \chi^2_r$. □

The following theorem is analogous to the central limit theorem and the theory for the t–interval for μ based on $\overline{Y}$ and the sample standard deviation (SD) S_Y. If the data $Y_1, ..., Y_n$ are iid with mean 0 and variance σ^2, then $\overline{Y}$ is asymptotically normal and the t–interval will perform well if the sample size is large enough. The result below suggests that the OLS estimators $\hat{Y}_i$ and $\hat{\boldsymbol{\beta}}$ are good if the sample size is large enough. The condition $\max h_i \to 0$ in probability usually holds if the researcher picked the design matrix $\boldsymbol{X}$ or if the $\boldsymbol{x}_i$ are iid random vectors from a well behaved population. Outliers can cause the condition to fail. Convergence in distribution, $\boldsymbol{Z}_n \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, means the multivariate normal approximation can be used for probability calculations involving $\boldsymbol{Z}_n$. When $p = 1$, the univariate normal distribution can be used. See Sen and Singer (1993, p. 280) for the theorem, which implies that $\hat{\boldsymbol{\beta}} \approx N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}))$. Let $h_i = \boldsymbol{H}_{ii}$ where $\boldsymbol{H} = \boldsymbol{P}_{\boldsymbol{X}}$. Note that the following theorem is for the full rank model since $\boldsymbol{X}^T\boldsymbol{X}$ is nonsingular.

Theorem 2.25, 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 $\text{VAR}(e_i) = \sigma^2$. Also assume that $\max_i(h_1, ..., h_n) \to 0$ in probability as $n \to \infty$ and

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

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

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2 \ \boldsymbol{W}). \tag{2.1}$$

Equivalently,

$$(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2 \ \boldsymbol{I}_p). \tag{2.2}$$

If $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{W}$, then $\hat{\boldsymbol{\Sigma}}_n = nMSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Hence

$$\hat{\boldsymbol{\beta}} \sim AN_p(\boldsymbol{\beta}, MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}), \quad \text{and}$$

$$rF_R = \frac{1}{MSE}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c})^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) \xrightarrow{D} \chi^2_r \tag{2.3}$$

as $n \rightarrow \infty$ if $H_0 : \boldsymbol{L\beta} = \boldsymbol{c}$ is true so that $\sqrt{n}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) \xrightarrow{D} N_r(\boldsymbol{0}, \sigma^2 \ \boldsymbol{LWL}^T)$.

Definition 2.20. A test with test statistic T_n is a *large sample right tail δ test* if the test rejects H_0 if $T_n > a_n$ and $P(T_n > a_n) = \delta_n \rightarrow \delta$ as $n \rightarrow \infty$ when H_0 is true.

Typically we want $\delta \leq 0.1$, and the values $\delta = 0.05$ or $\delta = 0.01$ are common. (An analogy is a large sample $100(1-\delta)\%$ confidence interval or prediction interval.)

Remark 2.3. Suppose $P(W \leq \chi^2_q(1-\delta)) = 1-\delta$ and $P(W > \chi^2_q(1-\delta)) = \delta$ where $W \sim \chi^2_q$. Suppose $P(W \leq F_{q,d_n}(1-\delta)) = 1-\delta$ when $W \sim F_{q,d_n}$. Also write $\chi^2_q(1-\delta) = \chi^2_{q,1-\delta}$ and $F_{q,d_n}(1-\delta) = F_{q,d_n,1-\delta}$. Suppose $P(W > z_{1-\delta}) = \delta$ when $W \sim N(0,1)$, and $P(W > t_{d_n,1-\delta}) = \delta$ when $W \sim t_{d_n}$.

i) Theorem 2.24 is important because it can often be shown that a statistic $T_n = rW_n \xrightarrow{D} \chi^2_r$ when H_0 is true. Then tests that reject H_0 when $T_n > \chi^2_r(1-\delta)$ or when $T_n/r = W_n > F_{r,d_n}(1-\delta)$ are both large sample right tail δ tests if the positive integer $d_n \rightarrow \infty$ as $n \rightarrow \infty$. Large sample F tests and intervals are used instead of χ^2 tests and intervals since the F tests and intervals are more accurate for moderate n.

ii) An analogy is that if test statistic $T_n \xrightarrow{D} N(0,1)$ when H_0 is true, then tests that reject H_0 if $T_n > z_{1-\delta}$ or if $T_n > t_{d_n,1-\delta}$ are both large sample right tail δ tests if the positive integer $d_n \rightarrow \infty$ as $n \rightarrow \infty$. Large sample t tests and intervals are used instead of Z tests and intervals since the t tests and intervals are more accurate for moderate n.

iii) Often $n \geq 10p$ starts to give good results for the OLS output for error distributions not too far from $N(0,1)$. Larger values of n tend to be needed if the zero mean iid errors have a distribution that is far from a normal distribution. Also see Theorem 1.5.

Theorem 2.26, Partial F Test Theorem. Suppose $H_0 : \boldsymbol{L\beta} = \boldsymbol{0}$ is true for the partial F test. 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.

Proof sketch. a) Seber and Lee (2003, p. 100) show that

$$RSS(R) - RSS = (\boldsymbol{L}\hat{\boldsymbol{\beta}})^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}}).$$

b) 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, $F_R \sim F_{p-k,n-p}$.

Alternatively, let $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}_n)$ where $\boldsymbol{X}$ is an $n \times p$ matrix of rank p. Let $\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2]$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T\ \boldsymbol{\beta}_2^T)^T$ where $\boldsymbol{X}_1$ is an $n \times k$ matrix and $r = p-k$. Consider testing $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$. (The columns of $\boldsymbol{X}$ can be rearranged so that H_0 corresponds to the partial F test.) Let $\boldsymbol{P}$ be the projection matrix on $C(\boldsymbol{X})$. Then $\boldsymbol{r}^T\boldsymbol{r} = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y} = \boldsymbol{e}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{e} =$
$(\boldsymbol{Y} - \boldsymbol{X\beta})^T(\boldsymbol{I} - \boldsymbol{P})(\boldsymbol{Y} - \boldsymbol{X\beta})$ since $\boldsymbol{PX} = \boldsymbol{X}$ and $\boldsymbol{X}^T\boldsymbol{P} = \boldsymbol{X}^T$ imply that $\boldsymbol{X}^T(\boldsymbol{I} - \boldsymbol{P}) = \boldsymbol{0}$ and $(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{X} = \boldsymbol{0}$.

Suppose that $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$ is true so that $\boldsymbol{Y} \sim N_n(\boldsymbol{X}_1\boldsymbol{\beta}_1, \sigma^2\boldsymbol{I}_n)$. Let $\boldsymbol{P}_1$ be the projection matrix on $C(\boldsymbol{X}_1)$. By the above argument, $\boldsymbol{r}_R^T\boldsymbol{r}_R = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{Y} = (\boldsymbol{Y} - \boldsymbol{X}_1\boldsymbol{\beta}_1)^T(\boldsymbol{I} - \boldsymbol{P}_1)(\boldsymbol{Y} - \boldsymbol{X}_1\boldsymbol{\beta}_1) = \boldsymbol{e}_R^T(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{e}_R$ where $\boldsymbol{e}_R \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}_n)$ when H_0 is true. Or use RHS $= \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{Y}$

$$-\boldsymbol{\beta}_1^T\boldsymbol{X}_1^T(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{Y} + \boldsymbol{\beta}_1^T\boldsymbol{X}_1^T(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{X}_1\boldsymbol{\beta}_1 - \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{X}_1\boldsymbol{\beta}_1,$$

and the last three terms equal 0 since $\boldsymbol{X}_1^T(\boldsymbol{I}-\boldsymbol{P}_1) = \boldsymbol{0}$ and $(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{X}_1 = \boldsymbol{0}$.

Hence
$$\frac{\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}}{\sigma^2} \sim \chi^2_{n-p} \perp\!\!\!\perp \frac{\boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}}{\sigma^2} \sim \chi^2_r$$
by Theorem 2.14 c) using $\boldsymbol{e}$ and $\boldsymbol{e}_R$ instead of $\boldsymbol{Y}$, and Craig's Theorem 2.9 b) since $n-p = rank(\boldsymbol{I}-\boldsymbol{P}) = tr(\boldsymbol{I}-\boldsymbol{P})$, $r = rank(\boldsymbol{P}-\boldsymbol{P}_1) = tr(\boldsymbol{P}-\boldsymbol{P}_1) = p-k$, and $(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{P}-\boldsymbol{P}_1) = \boldsymbol{0}$.

If $X_1 \sim \chi^2_{d_1} \perp\!\!\!\perp X_2 \sim \chi^2_{d_2}$, then
$$\frac{X_1/d_1}{X_2/d_2} \sim F_{d_1,d_2}.$$

Hence
$$\frac{\boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}/r}{\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}/(n-p)} = \frac{\boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}}{rMSE} \sim F_{r,n-p}$$
when H_0 is true. Since $RSS = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}$ and $RSS(R) = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{Y}$, $RSS(R) - RSS = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P}_1-[\boldsymbol{I}-\boldsymbol{P}])\boldsymbol{Y} = \boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}$, and thus
$$F_R = \frac{\boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}}{rMSE} \sim F_{r,n-p}.$$

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

d) By Theorem 2.24, if $W_n \sim F_{r,d_n}$ then $rW_n \xrightarrow{D} \chi^2_r$ as $n \to \infty$ and $d_n \to \infty$. Hence the result follows by c). □

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	$\frac{\boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_R)\boldsymbol{Y}/r}{\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}/(n-p)}$

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

Source	df	SS	MS	F	p-value
Regression	p-1	$SSR = \boldsymbol{Y}^T(\boldsymbol{P} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\boldsymbol{Y}$	MSR	$F_0 = \frac{MSR}{MSE}$	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 $\mathbf{0}$. Hence if $\boldsymbol{e} \sim N_n(\mathbf{0}, \sigma^2\boldsymbol{I})$ 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 $\boldsymbol{Y} \sim N_n(\mathbf{1}\beta_1, \sigma^2\boldsymbol{I})$, 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 $rank(\boldsymbol{P} - \frac{1}{n}\mathbf{1}\mathbf{1}^T) = tr(\boldsymbol{P} - \frac{1}{n}\mathbf{1}\mathbf{1}^T) = p - 1$. Hence under normality, $F_0 \sim F_{p-1,n-p}$.

Let $X \sim t_{n-p}$. Then $X^2 \sim F_{1,n-p}$. The two tail Wald t test for $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$ is equivalent to the corresponding right tailed F test since rejecting H_0 if $|X| > t_{n-p}(1-\delta)$ is equivalent to rejecting H_0 if $X^2 > F_{1,n-p}(1-\delta)$.

Definition 2.21. The **pvalue** of a test is the probability, assuming H_0 is true, of observing a test statistic as extreme as the test statistic T_n actually observed. For a right tail test, pvalue $= P_{H_0}$(of observing a test statistic $\geq T_n$).

Under the OLS model where $F_R \sim F_{q,n-p}$ when H_0 is true (so the e_i are iid $N(0,\sigma^2)$), the pvalue $= P(W > F_R)$ where $W \sim F_{q,n-p}$. In general, we can only estimate the pvalue. Let pval be the estimated pvalue. Then pval $= P(W > F_R)$ where $W \sim F_{q,n-p}$, and pval $\overset{P}{\to}$ pvalue an $n \to \infty$ for the large sample partial F test. The pvalues in output are usually actually pvals (estimated pvalues).

Definition 2.22. Let $Y \sim F(d_1, d_2) \sim F(d_1, d_2, 0)$. Let $X_1 \sim \chi^2(d_1, \gamma) \perp\!\!\!\perp X_2 \sim \chi^2(d_2, 0)$. Then $W = \dfrac{X_1/d_1}{X_2/d_2} \sim F(d_1, d_2, \gamma)$, a *noncentral F distribution* with d_1 and d_2 numerator and denominator degrees of freedom, and noncentrality parameter γ.

Theorem 2.27, distribution of F_R under normality when H_0 may not hold. Assume $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ where $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. Let $\boldsymbol{X} = [\boldsymbol{X}_1 \ \ \boldsymbol{X}_2]$ be full rank, and let the reduced model $\boldsymbol{Y} = \boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{e}_R$. Then

$$F_R = \frac{\boldsymbol{Y}^T(\boldsymbol{P} - \boldsymbol{P}_1)\boldsymbol{Y}/r}{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}/(n-p)} \sim F\left(r, n-p, \frac{\boldsymbol{\beta}^T\boldsymbol{X}^T(\boldsymbol{P} - \boldsymbol{P}_1)\boldsymbol{X\beta}}{2\sigma^2}\right).$$

If $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$ is true, then $\gamma = 0$.

Proof. Note that the denominator is the MSE, and $(n-p)MSE/\sigma^2 \sim \chi^2_{n-p}$ by the proof of Theorem 2.26. By Theorem 2.14 f),

$$\boldsymbol{Y}^T(\boldsymbol{P} - \boldsymbol{P}_1)\boldsymbol{Y}/\sigma^2 \sim \chi^2\left(r, \frac{\boldsymbol{\beta}^T\boldsymbol{X}^T(\boldsymbol{P} - \boldsymbol{P}_1)\boldsymbol{X\beta}}{2\sigma^2}\right)$$

where $r = rank(\boldsymbol{P} - \boldsymbol{P}_1) = tr(\boldsymbol{P} - \boldsymbol{P}_1) = p - k$ since $\boldsymbol{P} - \boldsymbol{P}_1$ is a projection matrix (symmetric and idempotent). □

Consider the test $H_0 : \boldsymbol{L\beta} = \boldsymbol{c}$ versus $H_1 : \boldsymbol{L\beta} \neq \boldsymbol{c}$, and suppose H_0 is true. Then $\sqrt{n}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) \xrightarrow{D} N_r(\boldsymbol{0}, \sigma^2 \boldsymbol{LWL}^T)$. Hence

$$rF_0 = \frac{1}{MSE}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c})^T(\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T)^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) \xrightarrow{D} \chi^2_p,$$

and rejecting H_0 if $F_0 > F_{r,n-p,1-\delta}$ is a large sample right tail δ test for a large class of zero mean error distributions. Seber and Lee (2003, pp. 100-101) show that $F_0 \sim F_{r,n-p}$ if H_0 is true and $\boldsymbol{e} \sim N_p(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, but the above result is far stronger: if the iid e_i has to satisfy $e_i \sim N(0, \sigma^2)$, OLS inference would rarely be useful.

Remark 2.4. Suppose tests and confidence intervals are derived under the assumption $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Then by the LS CLT and Remark 2.3, the inference tends to give large sample tests and confidence intervals for a large class of zero mean error distributions. For linear models, often the error distribution has heavier tails than the normal distribution. See Huber and Ronchetti (2009, p. 3). If some points stick out a bit in residual and/or response plots, then the error distribution likely has heavier tails than the normal distribution. See Figure 1.1.

2.4 WLS and Generalized Least Squares

Definition 2.23. Suppose that the response variable and at least one of the predictor variables is quantitative. Then the *generalized least squares* (GLS) model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{2.4}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. Also $E(\boldsymbol{e}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}) = \sigma^2 \boldsymbol{V}$ where $\boldsymbol{V}$ is a known $n \times n$ positive definite matrix.

Definition 2.24. The *GLS estimator*

$$\hat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{Y}. \tag{2.5}$$

The fitted values are $\hat{\boldsymbol{Y}}_{GLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{GLS}$.

Definition 2.25. Suppose that the response variable and at least one of the predictor variables is quantitative. Then the *weighted least squares* (WLS) model with weights $w_1, ..., w_n$ is the special case of the GLS model where $\boldsymbol{V}$ is diagonal: $\boldsymbol{V} = \text{diag}(\text{v}_1, ..., \text{v}_\text{n})$ and $w_i = 1/v_i$. Hence

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{2.6}$$

$E(\boldsymbol{e}) = \boldsymbol{0}$, and $\text{Cov}(\boldsymbol{e}) = \sigma^2 \text{diag}(\text{v}_1, ..., \text{v}_\text{n}) = \sigma^2 \text{diag}(1/\text{w}_1, ..., 1/\text{w}_\text{n})$.

Definition 2.26. The *WLS estimator*

$$\hat{\boldsymbol{\beta}}_{WLS} = (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{Y}. \tag{2.7}$$

The fitted values are $\hat{\boldsymbol{Y}}_{WLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{WLS}$.

Definition 2.27. The *feasible generalized least squares* (FGLS) model is the same as the GLS estimator except that $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$ is a function of an unknown $q \times 1$ vector of parameters $\boldsymbol{\theta}$. Let the estimator of $\boldsymbol{V}$ be $\hat{\boldsymbol{V}} = \boldsymbol{V}(\hat{\boldsymbol{\theta}})$. Then the FGLS estimator

$$\hat{\boldsymbol{\beta}}_{FGLS} = (\boldsymbol{X}^T \hat{\boldsymbol{V}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \hat{\boldsymbol{V}}^{-1} \boldsymbol{Y}. \tag{2.8}$$

The fitted values are $\hat{\boldsymbol{Y}}_{FGLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{FGLS}$. The *feasible weighted least squares* (FWLS) estimator is the special case of the FGLS estimator where $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$ is diagonal. Hence the estimated weights $\hat{w}_i = 1/\hat{v}_i = 1/v_i(\hat{\boldsymbol{\theta}})$. The FWLS estimator and fitted values will be denoted by $\hat{\boldsymbol{\beta}}_{FWLS}$ and $\hat{\boldsymbol{Y}}_{FWLS}$, respectively.

Notice that the ordinary least squares (OLS) model is a special case of GLS with $\boldsymbol{V} = \boldsymbol{I}_n$, the $n \times n$ identity matrix. It can be shown that the GLS estimator minimizes the GLS criterion

$$Q_{GLS}(\boldsymbol{\eta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta})^T \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}).$$

Notice that the FGLS and FWLS estimators have $p+q+1$ unknown parameters. These estimators can perform very poorly if $n < 10(p+q+1)$.

The GLS and WLS estimators can be found from the OLS regression (without an intercept) of a transformed model. Typically there will be a constant in the model: the first column of $\boldsymbol{X}$ is a vector of ones. Let the symmetric, nonsingular $n \times n$ square root matrix $\boldsymbol{R} = \boldsymbol{V}^{1/2}$ with $\boldsymbol{V} = \boldsymbol{R}\boldsymbol{R}$. Let $\boldsymbol{Z} = \boldsymbol{R}^{-1}\boldsymbol{Y}$, $\boldsymbol{U} = \boldsymbol{R}^{-1}\boldsymbol{X}$ and $\boldsymbol{\epsilon} = \boldsymbol{R}^{-1}\boldsymbol{e}$.

Theorem 2.28. a)

$$\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.9}$$

follows the OLS model since $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$.

b) The GLS estimator $\hat{\boldsymbol{\beta}}_{GLS}$ can be obtained from the OLS regression (without an intercept) of $\boldsymbol{Z}$ on $\boldsymbol{U}$.

c) For WLS, $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$. The corresponding OLS model $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is equivalent to $Z_i = \boldsymbol{u}_i^T\boldsymbol{\beta} + \epsilon_i$ for $i = 1, ..., n$ where $\boldsymbol{u}_i^T$ is the ith row of $\boldsymbol{U}$. Then $Z_i = \sqrt{w_i}\, Y_i$ and $\boldsymbol{u}_i = \sqrt{w_i}\, \boldsymbol{x}_i$. Hence $\hat{\boldsymbol{\beta}}_{WLS}$ can be obtained from the OLS regression (without an intercept) of $Z_i = \sqrt{w_i}\, Y_i$ on $\boldsymbol{u}_i = \sqrt{w_i}\, \boldsymbol{x}_i$.

Proof. a) $E(\boldsymbol{\epsilon}) = \boldsymbol{R}^{-1}E(\boldsymbol{e}) = \boldsymbol{0}$ and

$$\text{Cov}(\boldsymbol{\epsilon}) = \boldsymbol{R}^{-1}\text{Cov}(\boldsymbol{e})(\boldsymbol{R}^{-1})^T = \sigma^2\boldsymbol{R}^{-1}\boldsymbol{V}(\boldsymbol{R}^{-1})^T$$

$$= \sigma^2\boldsymbol{R}^{-1}\boldsymbol{R}\boldsymbol{R}(\boldsymbol{R}^{-1}) = \sigma^2\boldsymbol{I}_n.$$

Notice that OLS without an intercept needs to be used since $\boldsymbol{U}$ does not contain a vector of ones. The first column of $\boldsymbol{U}$ is $\boldsymbol{R}^{-1}\boldsymbol{1} \neq \boldsymbol{1}$.

b) Let $\hat{\boldsymbol{\beta}}_{ZU}$ denote the OLS estimator obtained by regressing $\boldsymbol{Z}$ on $\boldsymbol{U}$. Then

$$\hat{\boldsymbol{\beta}}_{ZU} = (\boldsymbol{U}^T\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{Z} = (\boldsymbol{X}^T(\boldsymbol{R}^{-1})^T\boldsymbol{R}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{R}^{-1})^T\boldsymbol{R}^{-1}\boldsymbol{Y}$$

and the result follows since $\boldsymbol{V}^{-1} = (\boldsymbol{R}\boldsymbol{R})^{-1} = \boldsymbol{R}^{-1}\boldsymbol{R}^{-1} = (\boldsymbol{R}^{-1})^T\boldsymbol{R}^{-1}$.

c) The result follows from b) if $Z_i = \sqrt{w_i}\, Y_i$ and $\boldsymbol{u}_i = \sqrt{w_i}\, \boldsymbol{x}_i$. But for WLS, $\boldsymbol{V} = \text{diag}(\text{v}_1, ..., \text{v}_\text{n})$ and hence $\boldsymbol{R} = \text{diag}(\sqrt{\text{v}_1}, ..., \sqrt{\text{v}_\text{n}})$. Hence

$$\boldsymbol{R}^{-1} = \text{diag}(1/\sqrt{\text{v}_1}, ..., 1/\sqrt{\text{v}_\text{n}}) = \text{diag}(\sqrt{\text{w}_1}, ..., \sqrt{\text{w}_\text{n}})$$

and $\boldsymbol{Z} = \boldsymbol{R}^{-1}\boldsymbol{Y}$ has ith element $Z_i = \sqrt{w_i}\, Y_i$. Similarly, $\boldsymbol{U} = \boldsymbol{R}^{-1}\boldsymbol{X}$ has ith row $\boldsymbol{u}_i^T = \sqrt{w_i}\, \boldsymbol{x}_i^T$. □

Remark 2.5. Standard software produces WLS output and the ANOVA F test and Wald t tests are performed using this output.

Remark 2.6. The FGLS estimator can also be found from the OLS regression (without an intercept) of $\boldsymbol{Z}$ on $\boldsymbol{U}$ where $\boldsymbol{V}(\hat{\boldsymbol{\theta}}) = \boldsymbol{R}\boldsymbol{R}$. Similarly the FWLS estimator can be found from the OLS regression (without an intercept) of $Z_i = \sqrt{\hat{w}_i} Y_i$ on $\boldsymbol{u}_i = \sqrt{\hat{w}_i} \boldsymbol{x}_i$. But now $\boldsymbol{U}$ is a random matrix instead of a constant matrix. Hence these estimators are highly nonlinear. OLS output can be used for exploratory purposes, but the p–values are generally not correct. The Olive (2018) bootstrap tests may be useful for FGLS and FWLS. See Chapter 4.

Under regularity conditions, the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$ is a consistent estimator of $\boldsymbol{\beta}$ when the GLS model holds, but $\hat{\boldsymbol{\beta}}_{GLS}$ should be used because it generally has higher efficiency.

Definition 2.28. Let $\hat{\boldsymbol{\beta}}_{ZU}$ be the OLS estimator from regressing $\boldsymbol{Z}$ on $\boldsymbol{U}$. The vector of fitted values is $\hat{\boldsymbol{Z}} = \boldsymbol{U}\hat{\boldsymbol{\beta}}_{ZU}$ and the vector of residuals is $\boldsymbol{r}_{ZU} = \boldsymbol{Z} - \hat{\boldsymbol{Z}}$. Then $\hat{\boldsymbol{\beta}}_{ZU} = \hat{\boldsymbol{\beta}}_{GLS}$ for GLS, $\hat{\boldsymbol{\beta}}_{ZU} = \hat{\boldsymbol{\beta}}_{FGLS}$ for FGLS, $\hat{\boldsymbol{\beta}}_{ZU} = \hat{\boldsymbol{\beta}}_{WLS}$ for WLS, and $\hat{\boldsymbol{\beta}}_{ZU} = \hat{\boldsymbol{\beta}}_{FWLS}$ for FWLS. For GLS, FGLS, WLS, and FWLS, a *residual plot* is a plot of $\hat{Z}_i$ versus $r_{ZU,i}$ and a *response plot* is a plot of $\hat{Z}_i$ versus Z_i.

Inference for the GLS model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ can be performed by using the partial F test for the equivalent no intercept OLS model $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Following Section 1.3.7, create $\boldsymbol{Z}$ and $\boldsymbol{U}$, fit the full and reduced model using the "no intercept" or "intercept = F" option. Let pval be the estimated pvalue.

The 4 step partial F test of hypotheses: i) State the hypotheses H_0: the reduced model is good H_A: use the full model
ii) Find the test statistic $F_R =$

$$\left[\frac{SSE(R) - SSE(F)}{df_R - df_F}\right] / MSE(F)$$

iii) Find the pval = $\mathrm{P}(F_{df_R - df_F, df_F} > F_R)$. (On exams often an F table is used. Here $df_R - df_F = p - q =$ number of parameters set to 0, and $df_F = n - p$.)
iv) State whether you reject H_0 or fail to reject H_0. Reject H_0 if pval $\leq \delta$ and conclude that the full model should be used. Otherwise, fail to reject H_0 and conclude that the reduced model is good.

Assume that the GLS model contains a constant β_1. The GLS ANOVA F test of $H_0 : \beta_2 = \cdots = \beta_p$ versus H_A: not H_0 uses the reduced model that contains the first column of $\boldsymbol{U}$. The GLS ANOVA F test of $H_0 : \beta_i = 0$ versus $H_A : \beta_i \neq 0$ uses the reduced model with the ith column of $\boldsymbol{U}$ deleted. For the special case of WLS, the software will often have a `weights` option that will also give correct output for inference.

Freedman (1981) shows that the nonparametric bootstrap can be useful for the WLS model with the e_i independent. For this case, the sandwich estimator is $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}_{OLS}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ with $\hat{\boldsymbol{W}} = n\ diag(r_1^2, ..., r_n^2)/(n-p)$ where the r_i are the OLS residuals and $\boldsymbol{W} = \sigma^2\boldsymbol{V}$. See Hinkley (1977), MacKinnon and White (1985), and White (1980).

2.5 Summary

1) The set of all linear combinations of $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ is the vector space known as *span*$(\boldsymbol{x}_1, ..., \boldsymbol{x}_n) = \{\boldsymbol{y} \in \mathbb{R}^k : \boldsymbol{y} = \sum_{i=1}^n a_i \boldsymbol{x}_i$ for some constants $a_1, ..., a_n\}$.

2) 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 $\boldsymbol{A} = C(\boldsymbol{A})$. Then $C(\boldsymbol{A}) = \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y} = \boldsymbol{A}\boldsymbol{w}$ for some $\boldsymbol{w} \in \mathbb{R}^m\} = \{\boldsymbol{y} : \boldsymbol{y} = w_1\boldsymbol{a}_1 + w_2\boldsymbol{a}_2 + \cdots + w_m\boldsymbol{a}_m$ for some scalars $w_1, ..., w_m\} = span(\boldsymbol{a}_1, ..., \boldsymbol{a}_m)$.

3) A **generalized inverse** of an $n \times m$ matrix $\boldsymbol{A}$ is any $m \times n$ matrix $\boldsymbol{A}^-$ satisfying $\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}$.

4) The **projection matrix** $\boldsymbol{P} = \boldsymbol{P}_{\boldsymbol{X}}$ onto the column space of $\boldsymbol{X}$ is unique, symmetric, and idempotent. $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$, and $\boldsymbol{P}\boldsymbol{W} = \boldsymbol{W}$ if each column of $\boldsymbol{W} \in C(\boldsymbol{X})$. The eigenvalues of $\boldsymbol{P}_{\boldsymbol{X}}$ are 0 or 1. Rank$(\boldsymbol{P}) = tr(\boldsymbol{P})$. Hence $\boldsymbol{P}$ is singular unless $\boldsymbol{X}$ is a nonsingular $n \times n$ matrix, and then $\boldsymbol{P} = \boldsymbol{I}_n$. If $C(\boldsymbol{X}_R)$ is a subspace of $C(\boldsymbol{X})$, then $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{X}_R} = \boldsymbol{P}_{\boldsymbol{X}_R}\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{P}_{\boldsymbol{X}_R}$.

5) $\boldsymbol{I}_n - \boldsymbol{P}$ is the projection matrix on $[C(\boldsymbol{X})]^\perp$.

6) Let $\boldsymbol{A}$ be a positive definite symmetric matrix. The *square root matrix* $\boldsymbol{A}^{1/2}$ is a positive definite symmetric matrix such that $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{A}$.

7) The matrix $\boldsymbol{A}$ in a quadratic form $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$ will be **symmetric** unless told otherwise.

8) **Theorem 2.5.** Let $\boldsymbol{x}$ be a random vector with $E(\boldsymbol{x}) = \boldsymbol{\mu}$ and Cov$(\boldsymbol{x}) = \boldsymbol{\Sigma}$. Then $E(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}) = tr(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu}$.

9) **Theorem 2.7.** If $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric matrices and $\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{B}\boldsymbol{Y}$, then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{Y}$.

10) The important part of **Craig's Theorem** is that if $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Y}^T\boldsymbol{B}\boldsymbol{Y}$ if $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B} = \boldsymbol{0}$.

11) **Theorem 2.14.** Let $\boldsymbol{A} = \boldsymbol{A}^T$ be symmetric. b) If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{I})$, then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \sim \chi^2_r$ iff $\boldsymbol{A}$ is idempotent of rank r. c) If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, then $\boldsymbol{Y}^T\boldsymbol{A}\boldsymbol{Y} \sim \sigma^2 \ \chi^2_r$ iff $\boldsymbol{A}$ is idempotent of rank r.

12) Often theorems are given for when $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \boldsymbol{I})$. If $\boldsymbol{Y} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, then apply the theorem using $\boldsymbol{Z} = \boldsymbol{Y}/\sigma \sim N_n(\boldsymbol{0}, \boldsymbol{I})$.

13) Suppose $Y_1, ..., Y_n$ are independent $N(\mu_i, 1)$ random variables so that $\boldsymbol{Y} = (Y_1, ..., Y_n)^T \sim N_n(\boldsymbol{\mu}, \boldsymbol{I}_n)$. Then $\boldsymbol{Y}^T\boldsymbol{Y} = \sum_{i=1}^n Y_i^2 \sim \chi^2(n, \gamma = \boldsymbol{\mu}^T\boldsymbol{\mu}/2)$, a *noncentral* $\chi^2(n,\gamma)$ *distribution*, with n degrees of freedom and *noncentrality parameter* $\gamma = \boldsymbol{\mu}^T\boldsymbol{\mu}/2 = \frac{1}{2}\sum_{i=1}^n \mu_i^2 \geq 0$. The noncentrality parameter $\delta = \boldsymbol{\mu}^T\boldsymbol{\mu} = 2\gamma$ is also used.

14) **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**
$\sum_{i=1}^n r_i^2(\boldsymbol{\eta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\eta}\|^2$.

15) Let $\boldsymbol{x}_i^T = (1, \boldsymbol{u}_i^T)$, and let $\boldsymbol{\beta}^T = (\beta_1, \boldsymbol{\beta}_2^T)$ where β_1 is the intercept and the slopes vector $\boldsymbol{\beta}_2 = (\beta_2, ..., \beta_p)^T$. Let the population covariance matrices Cov$(\boldsymbol{u}) = \boldsymbol{\Sigma}_{\boldsymbol{u}}$, and Cov$(\boldsymbol{u}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{u}Y}$. If the $(Y_i, \boldsymbol{u}_i^T)^T$ are iid, then the population coefficients from an OLS regression of Y on $\boldsymbol{x}$ are

$$\beta_1 = E(Y) - \boldsymbol{\beta}_2^T E(\boldsymbol{u}) \text{ and } \boldsymbol{\beta}_2 = \boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{u}\mathrm{Y}}.$$

16) **Theorem 2.19: Second way to compute $\hat{\boldsymbol{\beta}}$:** a) If $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1}$ exists, then $\hat{\beta}_1 = \overline{Y} - \hat{\boldsymbol{\beta}}_2^T \overline{\boldsymbol{u}}$ and

$$\hat{\boldsymbol{\beta}}_2 = \frac{n}{n-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y}.$$

b) Suppose that $(Y_i, \boldsymbol{u}_i^T)^T$ are iid random vectors such that σ_Y^2, $\boldsymbol{\Sigma}_{\boldsymbol{u}}^{-1}$, and $\boldsymbol{\Sigma}_{\boldsymbol{u}Y}$ exist. Then $\hat{\beta}_1 \xrightarrow{P} \beta_1$ and $\hat{\boldsymbol{\beta}}_2 \xrightarrow{P} \boldsymbol{\beta}_2$ as $\mathrm{n} \to \infty$ even if the OLS model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ does not hold.

17) **Theorem 2.20.** Let $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e} = \hat{\boldsymbol{Y}} + \boldsymbol{r}$ where $\boldsymbol{X}$ is full rank, $E(\boldsymbol{e}) = \boldsymbol{0}$, and $\mathrm{Cov}(\boldsymbol{e}) = \sigma^2 \boldsymbol{I}$. Let $\boldsymbol{P} = \boldsymbol{P}_{\boldsymbol{X}}$ be the projection matrix on $C(\boldsymbol{X})$ so $\hat{\boldsymbol{Y}} = \boldsymbol{PX}$, $\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$, and $\boldsymbol{PX} = \boldsymbol{X}$ so $\boldsymbol{X}^T\boldsymbol{P} = \boldsymbol{X}^T$.
i) The predictor variables and residuals are orthogonal. Hence the columns of $\boldsymbol{X}$ and the residual vector are orthogonal: $\boldsymbol{X}^T\boldsymbol{r} = \boldsymbol{0}$.
ii) $E(\boldsymbol{Y}) = \boldsymbol{X\beta}$.
iii) $\mathrm{Cov}(\boldsymbol{Y}) = \mathrm{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$.
iv) The fitted values and residuals are uncorrelated: $\mathrm{Cov}(\boldsymbol{r}, \hat{\boldsymbol{Y}}) = \boldsymbol{0}$.
v) The least squares estimator $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$: $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.
vi) $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

18) **LS CLT.** Suppose that the e_i are iid and

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

Then the least squares (OLS) estimator $\hat{\boldsymbol{\beta}}$ satisfies

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

Also,

$$(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{I}_p).$$

19) **Theorem 2.26, Partial F Test Theorem.** Suppose $H_0 : \boldsymbol{L\beta} = \boldsymbol{0}$ is true for the partial F test. 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 $rF_R \xrightarrow{D} \chi_r^2$.
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.

2.6 Complements

A good reference for quadratic forms and the noncentral χ^2, t, and F distributions is Johnson and Kotz (1970, ch. 28-31).

The theory for GLS and WLS is similar to the theory for the OLS MLR model, but the theory for FGLS and FWLS is often lacking or huge sample sizes are needed. However, FGLS and FWLS are often used in practice because usually $\boldsymbol{V}$ is not known and $\hat{\boldsymbol{V}}$ must be used instead. See Eicker (1963, 1967).

Least squares theory can be extended in at least two ways. For the first extension, see Chang and Olive (2010) and Chapter 10. The second extension of least squares theory is to an autoregressive $AR(p)$ time series model: $Y_t = \phi_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + e_t$. In matrix form, this model is $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e} =$

$$\begin{bmatrix} Y_{p+1} \\ Y_{p+2} \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & Y_p & Y_{p-1} & \dots & Y_1 \\ 1 & Y_{p+1} & Y_p & \dots & Y_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & Y_{n-1} & Y_{n-2} & \dots & Y_{n-p} \end{bmatrix} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_p \end{bmatrix} + \begin{bmatrix} e_{p+1} \\ e_{p+2} \\ \vdots \\ e_n \end{bmatrix}.$$

If the $AR(p)$ model is stationary, then under regularity conditions, OLS partial F tests are large sample tests for this model. See Anderson (1971, pp. 210–217).

2.7 Problems

Problems from old qualifying exams are marked with a Q since these problems take longer than quiz and exam problems.

2.1^Q. Suppose $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ where the errors are independent $N(0, \sigma^2)$. Then the likelihood function is

$$L(\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left(\frac{-1}{2\sigma^2}\|\boldsymbol{Y} - \boldsymbol{X\beta}\|^2\right).$$

a) Since the least squares estimator $\hat{\boldsymbol{\beta}}$ minimizes $\|\boldsymbol{Y} - \boldsymbol{X\beta}\|^2$, show that $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$.

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

2.2^Q. Suppose $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ where the errors are iid double exponential $(0, \sigma)$ where $\sigma > 0$. Then the likelihood function is

$$L(\boldsymbol{\beta}, \sigma) = \frac{1}{2^n}\frac{1}{\sigma^n} \exp\left(\frac{-1}{\sigma}\sum_{i=1}^n |Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}|\right).$$

Suppose that $\tilde{\boldsymbol{\beta}}$ is a minimizer of $\sum_{i=1}^n |Y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}|$.

a) By direct maximization, show that $\tilde{\boldsymbol{\beta}}$ is an MLE of $\boldsymbol{\beta}$ regardless of the value of σ.

b) Find an MLE of σ by maximizing

$$L(\sigma) \equiv L(\tilde{\boldsymbol{\beta}}, \sigma) = \frac{1}{2^n} \frac{1}{\sigma^n} \exp\left(\frac{-1}{\sigma} \sum_{i=1}^n |Y_i - \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}}| \right).$$

2.3^Q. Suppose $Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i$ where the errors are independent $N(0, \sigma^2/w_i)$ where $w_i > 0$ are known constants. Then the likelihood function is

$$L(\boldsymbol{\beta}, \sigma^2) = \left(\prod_{i=1}^n \sqrt{w_i} \right) \left(\frac{1}{\sqrt{2\pi}} \right)^n \frac{1}{\sigma^n} \exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^n w_i (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 \right).$$

a) Suppose that $\hat{\boldsymbol{\beta}}_W$ minimizes $\sum_{i=1}^n w_i (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2$. Show that $\hat{\boldsymbol{\beta}}_W$ is the MLE of $\boldsymbol{\beta}$.

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

2.4^Q. Suppose $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{V})$ for known positive definite $n \times n$ matrix $\boldsymbol{V}$. Then the likelihood function is

$$L(\boldsymbol{\beta}, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}} \right)^n \frac{1}{|\boldsymbol{V}|^{1/2}} \frac{1}{\sigma^n} \exp\left(\frac{-1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X\beta})^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X\beta}) \right).$$

a) Suppose that $\hat{\boldsymbol{\beta}}_G$ minimizes $(\boldsymbol{y} - \boldsymbol{X\beta})^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X\beta})$. Show that $\hat{\boldsymbol{\beta}}_G$ is the MLE of $\boldsymbol{\beta}$.

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

2.5. Find the vector $\boldsymbol{a}$ such that $\boldsymbol{a}^T \boldsymbol{Y}$ is an unbiased estimator for $E(Y_i)$ if the usual linear model holds.

2.6. Write the following quantities as $\boldsymbol{b}^T \boldsymbol{Y}$ or $\boldsymbol{Y}^T \boldsymbol{AY}$ or $\boldsymbol{AY}$.

a) $\overline{Y}$, b) $\sum_i (Y_i - \hat{Y}_i)^2$, c) $\sum_i (\hat{Y}_i)^2$, d) $\hat{\boldsymbol{\beta}}$, e) $\hat{\boldsymbol{Y}}$

2.7. Show that $\boldsymbol{I} - \boldsymbol{H} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T$ is idempotent, that is, show that $(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}) = (\boldsymbol{I} - \boldsymbol{H})^2 = \boldsymbol{I} - \boldsymbol{H}$.

2.8. Let $Y \sim N(\mu, \sigma^2)$ so that $E(Y) = \mu$ and $\text{Var}(Y) = \sigma^2 = E(Y^2) - [E(Y)]^2$. If $k \geq 2$ is an integer, then

$$E(Y^k) = (k-1)\sigma^2 E(Y^{k-2}) + \mu E(Y^{k-1}).$$

Let $Z = (Y - \mu)/\sigma \sim N(0, 1)$. Hence $\mu_k = E(Y - \mu)^k = \sigma^k E(Z^k)$. Use this fact and the above recursion relationship $E(Z^k) = (k - 1)E(Z^{k-2})$ to find a) μ_3 and b) μ_4.

2.9. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be matrices with the same number of rows. If $\boldsymbol{C}$ is another matrix such that $\boldsymbol{A} = \boldsymbol{BC}$, is it true that $rank(\boldsymbol{A}) = rank(\boldsymbol{B})$? Prove or give a counterexample.

2.10. Let $\boldsymbol{x}$ be an $n \times 1$ vector and let $\boldsymbol{B}$ be an $n \times n$ matrix. Show that $\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{B}^T\boldsymbol{x}$.

(The point of this problem is that if $\boldsymbol{B}$ is not a symmetric $n \times n$ matrix, then $\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$ where $\boldsymbol{A} = \dfrac{\boldsymbol{B} + \boldsymbol{B}^T}{2}$ is a symmetric $n \times n$ matrix.)

2.11. Consider the model $Y_i = \beta_1 + \beta_2 X_{i,2} + \cdots + \beta_p X_{i,p} + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$. The least squares estimator $\hat{\boldsymbol{\beta}}$ minimizes

$$Q_{OLS}(\boldsymbol{\eta}) = \sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^T\boldsymbol{\eta})^2$$

and the weighted least squares estimator minimizes

$$Q_{WLS}(\boldsymbol{\eta}) = \sum_{i=1}^{n} w_i(Y_i - \boldsymbol{x}_i^T\boldsymbol{\eta})^2$$

where the w_i, Y_i and $\boldsymbol{x}_i$ are known quantities. Show that

$$\sum_{i=1}^{n} w_i(Y_i - \boldsymbol{x}_i^T\boldsymbol{\eta})^2 = \sum_{i=1}^{n}(\tilde{Y}_i - \tilde{\boldsymbol{x}}_i^T\boldsymbol{\eta})^2$$

by identifying $\tilde{Y}_i$, and $\tilde{\boldsymbol{x}}_i$. (Hence the WLS estimator is obtained from the least squares regression of $\tilde{Y}_i$ on $\tilde{\boldsymbol{x}}_i$ without an intercept.)

2.12. Suppose that $\boldsymbol{X}$ is an $n \times p$ matrix but the rank of $\boldsymbol{X} < p < n$. Then the normal equations $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{Y}$ have infinitely many solutions. Let $\hat{\boldsymbol{\beta}}$ be a solution to the normal equations. So $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y}$. Let $\boldsymbol{G} = (\boldsymbol{X}^T\boldsymbol{X})^-$ be a generalized inverse of $(\boldsymbol{X}^T\boldsymbol{X})$. Assume that $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$ and $\text{Cov}(\boldsymbol{Y}) = \sigma^2\boldsymbol{I}$. It can be shown that all solutions to the normal equations have the form $\boldsymbol{b_z}$ given below.

a) Show that $\boldsymbol{b_z} = \boldsymbol{G}\boldsymbol{X}^T\boldsymbol{Y} + (\boldsymbol{G}\boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{I})\boldsymbol{z}$ is a solution to the normal equations where the $p \times 1$ vector $\boldsymbol{z}$ is arbitrary.

b) Show that $E(\boldsymbol{b_z}) \neq \boldsymbol{\beta}$.

(Hence some authors suggest that $\boldsymbol{b_z}$ should be called a solution to the normal equations but not an estimator of $\boldsymbol{\beta}$.)

c) Show that $\text{Cov}(\boldsymbol{b_z}) = \sigma^2 \boldsymbol{G}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{G}^T$.

d) Although $\boldsymbol{G}$ is not unique, the projection matrix $\boldsymbol{P} = \boldsymbol{X}\boldsymbol{G}\boldsymbol{X}^T$ onto $C(\boldsymbol{X})$ is unique. Use this fact to show that $\hat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{b_z}$ does not depend on $\boldsymbol{G}$ or $\boldsymbol{z}$.

e) There are two ways to show that $\boldsymbol{a}^T\boldsymbol{\beta}$ is an estimable function. Either show that there exists a vector $\boldsymbol{c}$ such that $E(\boldsymbol{c}^T\boldsymbol{Y}) = \boldsymbol{a}^T\boldsymbol{\beta}$, or show that $\boldsymbol{a} \in C(\boldsymbol{X}^T)$. Suppose that $\boldsymbol{a} = \boldsymbol{X}^T\boldsymbol{w}$ for some fixed vector $\boldsymbol{w}$. Show that $E(\boldsymbol{a}^T\boldsymbol{b_z}) = \boldsymbol{a}^T\boldsymbol{\beta}$.

(Hence $\boldsymbol{a}^T\boldsymbol{\beta}$ is estimable by $\boldsymbol{a}^T\boldsymbol{b_z}$ where $\boldsymbol{b_z}$ is any solution of the normal equations.)

f) Suppose that $\boldsymbol{a} = \boldsymbol{X}^T\boldsymbol{w}$ for some fixed vector $\boldsymbol{w}$. Show that $Var(\boldsymbol{a}^T\boldsymbol{b_z}) = \sigma^2\boldsymbol{w}^T\boldsymbol{P}\boldsymbol{w}$.

2.13. Let $\boldsymbol{P}$ be a projection matrix.
a) Show that $\boldsymbol{P}$ is a generalized inverse of $\boldsymbol{P}$.
b) Show that $\boldsymbol{P} = \boldsymbol{P}(\boldsymbol{P}^T\boldsymbol{P})^-\boldsymbol{P}^T$.

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

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

a) Suppose that $\hat{\boldsymbol{\beta}}_Q$ minimizes $Q(\boldsymbol{\beta})$. Show that $\hat{\boldsymbol{\beta}}_Q$ is an MLE of $\boldsymbol{\beta}$.
b) Then find an MLE $\hat{\sigma}$ of σ.

2.15^Q. 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.

2.16. Suppose that $\boldsymbol{G}$ is a generalized inverse of a symmetric matrix $\boldsymbol{A}$.

a) Show that $\boldsymbol{G}^T$ is a generalized inverse of $\boldsymbol{A}$.

b) Show that $\boldsymbol{G}\boldsymbol{A}\boldsymbol{G}^T$ is a generalized inverse of $\boldsymbol{A}$. (Hence, since a generalized inverse always exists, a symmetric generalized inverse of a symmetric matrix $\boldsymbol{A}$ always exists.)

2.17. (Searle (1971, p. 217)): Let $\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 4 & 3 \\ 3 & -1 & 2 & -2 \\ 5 & -4 & 0 & -7 \end{bmatrix}$ and show that $\boldsymbol{A}^- = \frac{1}{7}\begin{bmatrix} 1 & 2 & 0 \\ 3 & -1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ is a generalized inverse of $\boldsymbol{A}$.

2.18. Find the projection matrix $\boldsymbol{P}$ for $C(\boldsymbol{X})$ where $\boldsymbol{X}$ is the 2×1 vector $\boldsymbol{X} = (1, 2)^T$.

2.19. Let $\boldsymbol{y} \sim N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is positive definite. Let $\boldsymbol{A}$ be a symmetric $p \times p$ matrix.

a) Let $\boldsymbol{x} = \boldsymbol{y} - \boldsymbol{\theta}$. What is the distribution of $\boldsymbol{x}$?

b) Show that

$$E[(\boldsymbol{y} - \boldsymbol{\theta})^T \boldsymbol{A}(\boldsymbol{y} - \boldsymbol{\theta})] = E[\boldsymbol{x}^T A \boldsymbol{x}]$$

is a function of $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ but not of $\boldsymbol{\theta}$.

2.20. (Hocking (2003, p. 61): Let $\boldsymbol{y} \sim N_3(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ where $\boldsymbol{y} = (Y_1, Y_2, Y_3)^T$ and $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)^T$.

Let $\boldsymbol{A} = \frac{1}{2}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ and $\boldsymbol{B} = \frac{1}{6}\begin{bmatrix} 1 & 1 & -2 \\ 1 & 1 & -2 \\ -2 & -2 & 4 \end{bmatrix}$.

Are $\boldsymbol{y}^T \boldsymbol{A} \boldsymbol{y}$ and $\boldsymbol{y}^T \boldsymbol{B} \boldsymbol{y}$ independent? Explain.

2.21^Q. Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$. Assume $\boldsymbol{X}$ has full rank. Let $\boldsymbol{r}$ be the vector of residuals. Then the residual sum of squares RSS $= \boldsymbol{r}^T \boldsymbol{r}$. The sum of squared fitted values is $\hat{\boldsymbol{Y}}^T \hat{\boldsymbol{Y}}$. Prove that $\boldsymbol{r}^T \boldsymbol{r}$ and $\hat{\boldsymbol{Y}}^T \hat{\boldsymbol{Y}}$ independent (or dependent).
(Hint: write each term as a quadratic form.)

2.22. Let $\boldsymbol{B} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$.

a) Find rank($\boldsymbol{B}$).

b) Find a basis for $\mathcal{C}(\boldsymbol{B})$.

c) Find $[C(\boldsymbol{B})]^\perp =$ nullspace of $\boldsymbol{B}^T$.

d) Show that $\boldsymbol{B}^- = \begin{bmatrix} 1 & -1 \\ 1 & 0 \end{bmatrix}$ is a generalized inverse of $\boldsymbol{B}$.

2.23. Suppose that $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\text{Cov}(\boldsymbol{e}) = \sigma^2 \boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2} \boldsymbol{\Sigma}^{1/2}$ where $\boldsymbol{\Sigma}^{1/2}$ is nonsingular and symmetric. Hence $\boldsymbol{\Sigma}^{-1/2} \boldsymbol{Y} = \boldsymbol{\Sigma}^{-1/2} \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\Sigma}^{-1/2} \boldsymbol{e}$. Find $\text{Cov}(\boldsymbol{\Sigma}^{-1/2} \boldsymbol{e})$. Simplify.

2.24. Let $\boldsymbol{y} \sim N_2(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ where $\boldsymbol{y} = (Y_1, Y_2)^T$ and $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$. Let $\boldsymbol{A} = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix}$.

Are $\boldsymbol{y}^T \boldsymbol{A} \boldsymbol{y}$ and $\boldsymbol{y}^T \boldsymbol{B} \boldsymbol{y}$ independent? Explain.

2.25. Assuming the assumptions of the least squares central limit theorem hold, what is the limiting distribution of $\sqrt{n}\ \ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ if $(\boldsymbol{X}'\boldsymbol{X})/n \to \boldsymbol{W}^{-1}$ as $n \to \infty$?

$$\sqrt{n}\ \ (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D}$$

2.26. Let the model be $Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + ... + \beta_{10} x_{i10} + e_i$. The model in matrix form is $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ where $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. Let $\boldsymbol{P}$ be the projection matrix on $C(\boldsymbol{X})$ where the $n \times p$ matrix $\boldsymbol{X}$ has full rank p. What is the distribution of $\boldsymbol{Y}^T \boldsymbol{PY}$?

Hint: If $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{I})$, then $\boldsymbol{Y}^T \boldsymbol{AY} \sim \chi^2(\text{rank}(\boldsymbol{A}), \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{A\mu}/2)$ iff $\boldsymbol{A} = \boldsymbol{A}^T$ is idempotent. $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I})$, so $\dfrac{\boldsymbol{Y}}{\sigma} \sim N_n\left(\dfrac{\boldsymbol{X\beta}}{\sigma}, \boldsymbol{I}\right)$. **Simplify.**

2.27. Let $\boldsymbol{Y}' = \boldsymbol{Y}^T$. Let $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I})$. Recall that $E(\boldsymbol{Y}'\boldsymbol{AY}) = tr(\boldsymbol{A}Cov(\boldsymbol{Y})) + E(\boldsymbol{Y}')\boldsymbol{A}E(\boldsymbol{Y})$.
Find $E(\boldsymbol{Y}'\boldsymbol{Y}) = E(\boldsymbol{Y}'\boldsymbol{IY})$.

2.28. Let $\boldsymbol{y} \sim N_2(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ where $\boldsymbol{y} = (Y_1, Y_2)^T$ and $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$. Let $\boldsymbol{A} = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix} 1/4 & \sqrt{3}/4 \\ \sqrt{3}/4 & 3/4 \end{bmatrix}$.

Are $\boldsymbol{Ay}$ and $\boldsymbol{By}$ independent? Explain.

2.29. Let $\boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$.

a) Find rank($\boldsymbol{X}$).
b) Find a basis for $C(\boldsymbol{X})$.
c) Find $[C(\boldsymbol{X})]^{\perp}$ = nullspace of $\boldsymbol{X}^T$.

2.30^Q. Let $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ where $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$. Assume $\boldsymbol{X}$ has full rank and that the first column of $\boldsymbol{X} = \boldsymbol{1}$ so that a constant is in the model. Let $\boldsymbol{r}$ be the vector of residuals. Then the residual sum of squares RSS = $\boldsymbol{r}^T \boldsymbol{r} = \|(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}\|^2$. The sample mean $\overline{\boldsymbol{Y}} = \frac{1}{n}\boldsymbol{1}^T \boldsymbol{Y}$. Prove that $\boldsymbol{r}^T \boldsymbol{r}$ and $\overline{\boldsymbol{Y}}$ independent (or dependent).
(Hint: If $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{AY} \perp\!\!\!\perp \boldsymbol{BY}$ iff $\boldsymbol{A\Sigma B}^T = \boldsymbol{0}$.
So prove whether $(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y} \perp\!\!\!\perp \dfrac{1}{n}\boldsymbol{1}^T \boldsymbol{Y}$.)

2.31. Let the full model be $Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + e_i$ and let the reduced model be $Y_i = \beta_1 + \beta_3 x_{i3} + e_i$ for $i = 1, ..., n$. Write the full model as $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e} = \boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{e}$, and consider testing $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$ where $\boldsymbol{\beta}_1$ corresponds to the reduced model. Let $\boldsymbol{P}_1$ be the projection matrix on $C(\boldsymbol{X}_1)$ and let $\boldsymbol{P}$ be the projection matrix on $C(\boldsymbol{X})$.

Then $F_R = \dfrac{n-p}{q} \dfrac{\boldsymbol{Y}^T(\boldsymbol{P} - \boldsymbol{P}_1)\boldsymbol{Y}}{\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}}$.

Assume $\boldsymbol{\epsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. Assume H_0 is true.

a) What is q?

b) What is the distribution of $\boldsymbol{Y}^T(\boldsymbol{P}-\boldsymbol{P}_1)\boldsymbol{Y}$?

c) What is the distribution of $\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{Y}$?

d) What is the distribution of F_R?

2.32^Q. If $\boldsymbol{P}$ is a projection matrix, prove a) the eigenvalues of $\boldsymbol{P}$ are 0 or 1, b) $rank(\boldsymbol{P}) = tr(\boldsymbol{P})$.

2.33^Q. Suppose that $\boldsymbol{AY}$ and $\boldsymbol{BY}$ are independent where $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric matrices. Are $\boldsymbol{Y'AY}$ and $\boldsymbol{Y'BY}$ independent? (Hint: show that the quadratic form $\boldsymbol{Y'AY}$ is a function of $\boldsymbol{AY}$ by using the definition of the generalized inverse $\boldsymbol{A}^-$.)

2.34. Craig's theorem states that if $\boldsymbol{x} \sim N_n(\boldsymbol{\mu}, \boldsymbol{V})$ and if $\boldsymbol{A}$ and $\boldsymbol{B}$ are symmetric matrices, then the quadratic forms $\boldsymbol{x'Ax}$ and $\boldsymbol{x'Bx}$ are independent iff i) $\boldsymbol{VAVBV} = \boldsymbol{0}$, ii) $\boldsymbol{VAVB\mu} = \boldsymbol{0}$, iii) $\boldsymbol{VBVA\mu} = \boldsymbol{0}$, and iv) $\boldsymbol{\mu' AVB\mu} = \boldsymbol{0}$. Here $\boldsymbol{V}$ is positive semidefinite. Hence $\boldsymbol{V}$ could be singular. Notice that $\boldsymbol{V}$ is symmetric since it is a covariance matrix.

Suppose that $\boldsymbol{AVB} = \boldsymbol{0}$. Are $\boldsymbol{x'Ax}$ and $\boldsymbol{x'Bx}$ are independent? Explain briefly.

2.35^Q. 2.35. Let $\boldsymbol{Y}$ be an $n \times 1$ random vector and $\boldsymbol{A}$ an $n \times n$ symmetric matrix. Let $E(\boldsymbol{Y}) = \boldsymbol{\theta}$ and $\text{Cov}(\boldsymbol{Y}) = \boldsymbol{\Sigma} = (\sigma_{ij})$.

a) Prove that $E(\boldsymbol{Y}^T\boldsymbol{AY}) = tr(\boldsymbol{A\Sigma}) + \boldsymbol{\theta}^T\boldsymbol{A\theta}$.

b) Let $E(Y_i) = \theta$ for all i, $\sigma_{ii} = \sigma^2$ for all i, and $\sigma_{ij} = \rho\sigma^2$ for $i \neq j$ where $-1 < \rho < 1$. Show that $\sum_i (Y_i - \overline{Y})^2$ is an unbiased estimator of $\sigma^2(1-\rho)(n-1)$. Hint: write $\sum_i (Y_i - \overline{Y})^2 = \boldsymbol{Y}^T\boldsymbol{AY}$ and use a).

c) Show when $\sum_i (Y_i - \overline{Y})^2$ and $\overline{Y}$ are independent if $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$. State the theorems clearly wherever used in your proof.

2.36^Q (NIU, summer 1991). Consider the regression model $Y_i = \beta x_i + e_i$ for $i = 1, ..., n$ where the e_i are iid $N(0, \sigma^2)$.

a) Show that the least squares estimator of β is

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}.$$

b) Express $\hat{\beta}$ as a linear combination of the responses and derive its mean and variance.

c) Show that $\hat{Y}_i = \hat{\beta} x_i$ is an unbiased estimator of $E(Y_i)$and derive its variance.

d) Derive the maximum likelihood estimators of β and σ^2.

2.37^Q. a) For an $n \times 1$ vector $\boldsymbol{Y}$ with $E(\boldsymbol{Y}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{Y}) = \boldsymbol{\Sigma}$, show $E(\boldsymbol{Y}^T\boldsymbol{AY}) = trace(\boldsymbol{A\Sigma}) + \boldsymbol{\mu}^T\boldsymbol{A\mu}$. is normality necessary here?

b) Consider the usual full rank linear model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ is $n \times p$, the first column of $\boldsymbol{X}$ is $\boldsymbol{1}$, $\boldsymbol{\beta}$ is $p \times 1$ and $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$.

i) Write down an ANOVA table to test $(\beta_2, ..., \beta_p)^T = \boldsymbol{0}$, giving expressions for the regression sum of squares (SSR) and the error sum of squares (SSE).

ii) Find $E(SSR)$ and $E(SSE)$ when H_0 is true.

iii) Derive the distribution of SSE/σ^2 if H_0 is true. State any theorems used.

2.38^Q**.** a) Define a generalized inverse of a matrix $\boldsymbol{A}$.

b) i) Suppose $\boldsymbol{X}$ is $n \times p$ with rank $r < p$. Give the formula for the projection matrix $\boldsymbol{P}$ onto the column space of $\boldsymbol{X}$.

ii) For

$$\boldsymbol{X} = \begin{bmatrix} 1 & -2 \\ 1 & -2 \\ 1 & -2 \end{bmatrix},$$

calculate $\boldsymbol{P}$.

iii) With $\boldsymbol{X}$ as above and $\boldsymbol{Y} = (1, 2, 3)^T$, calculate the error sum of squares SSE.

2.39^Q**.** Consider the usual full rank model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ is $n \times p$ and $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T \ \boldsymbol{\beta}_2^T)^T$ where $\boldsymbol{\beta}_i$ is $p_i \times 1$.

a) Write down the complete ANOVA table for the test $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$, including the expected mean squares.

b) Prove that $SSE(R) - SSE$ and MSE are independent.

c) If H_0 is true, show $F_R \sim F_{p_2, n-p}$.

2.40^Q**.** Let $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} > 0$, and let $\boldsymbol{A}$ be a symmetric matrix.

a) State the necessary and sufficient condition(s) for $\boldsymbol{Y}^T \boldsymbol{AY}$ to be a chi-square random variable.

b) Suppose $rank(\boldsymbol{\Sigma}) = n$ and $\boldsymbol{B\Sigma A} = \boldsymbol{0}$ where $\boldsymbol{B}$ is a $q \times n$ matrix. Prove that $\boldsymbol{Y}^T \boldsymbol{AY}$ and $\boldsymbol{BY}$ are independent.

c) If $\boldsymbol{\mu} = \mu \boldsymbol{1}$ and $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$ where $\sigma^2 > 0$, prove that
$\overline{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ and $\frac{1}{n-1} \sum_{i=1}^n (Y_i - \overline{Y})^2$ are independent.

2.41^Q**.** Let $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ where $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, $\boldsymbol{X}$ is an $n \times p$ matrix of rank p, and $\boldsymbol{\beta}$ is a $p \times 1$ vector.

a) Write down (do not derive) the MLEs of $\boldsymbol{\beta}$ and σ^2.

b) If $\hat{\sigma}^2$ is the MLE of σ^2, derive the distribution of $(n-p)\hat{\sigma}^2/\sigma^2$.

c) Prove that $\hat{\boldsymbol{\beta}}$ (MLE of $\boldsymbol{\beta}$) and $\hat{\sigma}^2$ are independent.

d) Now suppose $\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{V})$ where $\boldsymbol{V}$ is a known positive definite matrix. Write down the MLE of $\boldsymbol{\beta}$.

2.42^Q**.** a) Suppose $\boldsymbol{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}$ be an $n \times n$ symmetric matrix.

i) Show $E[(\boldsymbol{Y} - \boldsymbol{\mu})^T \boldsymbol{A} (\boldsymbol{Y} - \boldsymbol{\mu})] = tr(\boldsymbol{A\Sigma})$. Is normality of $\boldsymbol{Y}$ necessary here?

ii) State a necessary and sufficient condition for $(\boldsymbol{Y} - \boldsymbol{\mu})^T \boldsymbol{A} (\boldsymbol{Y} - \boldsymbol{\mu})$ to be a chi-square random variable.

iii) State a necessary and sufficient condition for $(\boldsymbol{Y} - \boldsymbol{\mu})^T \boldsymbol{A} (\boldsymbol{Y} - \boldsymbol{\mu})$ and $\boldsymbol{BY}$ to be independent where $\boldsymbol{B}$ is an $q \times n$ matrix.

b) Suppose $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I})$ where $\boldsymbol{X}$ is an $n \times p$ matrix of rank p and $\boldsymbol{\beta}$ is $p \times 1$.

i) Derive the distribution of $\frac{1}{\sigma}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}$ where $\boldsymbol{H}$ is the projection matrix onto the column space $C(\boldsymbol{X})$.

ii) Derive the distribution of $u = \dfrac{\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}}{\sigma^2}$.

iii) Show that u and $v = \boldsymbol{H}\boldsymbol{Y}$ are independent.

2.43^Q. Consider the regression model $y_i = \beta x_i + e_i$ for $i = 1, ..., n$ where the e_i are iid $N(0, \sigma^2)$.

a) Derive the least squares estimator of β.

b) Write down an unbiased estimator of σ^2.

c) Derive the maximum likelihood estimators of β and σ^2.

2.44^Q. Let Y_1 and Y_2 be iindependent random variables with mean θ and 2θ respectively. Find the least squares estimate of θ and the residual sum of squares.

2.45^Q. a) By the least squares central limit theorem, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\, \boldsymbol{W})$. Hence the limiting distribution of of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is the $N_p(\boldsymbol{0}, \sigma^2\, \boldsymbol{W})$ distribution. Let $\boldsymbol{A}$ be a constant $r \times p$ matrix. Find the limiting distribution of $\boldsymbol{A}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.

b) Suppose $\boldsymbol{Z}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{I})$. Let $\boldsymbol{A}$ be a constant $r \times k$ matrix. Find the limiting distribution of $\boldsymbol{A}(\boldsymbol{Z}_n - \boldsymbol{\mu})$.

2.46.

2.47.

2.48.

2.49.

2.50.

2.51.

2.52.

2.53.

R Problems

Use the command *source("G:/linmodpack.txt")* **to download the functions** and the command *source("G:/linmoddata.txt")* **to download the data. See Preface or Section 11.1.** Typing the name of the `linmodpack` function, e.g. *regbootsim2*, will display the code for the function. Use the `args` command, e.g. *args(regbootsim2)*, to display the needed arguments for the function. For the following problem, the *R* commands can be copied and pasted from (http://parker.ad.siu.edu/Olive/linmodrhw.txt) into *R*.

2.54. Generalized and weighted least squares are each equivalent to a least squares regression without intercept. Let $\boldsymbol{w}' = \boldsymbol{w}^T$. Let $\boldsymbol{V} = \text{diag}(1, 1/2, 1/3, ..., 1/9) = \text{diag}(w_i)$ where $n = 9$ and the weights $w_i = i$ for $i = 1, ..., 9$. Let $\boldsymbol{x}' = (1, x_1, x_2, x_3)$. Then the weighted least squares with weight vector $\boldsymbol{w}' = (1, 2, ..., 9)$ is equivalent to the OLS regression of $\sqrt{w_i}\, Y_i = Z_i$ on $\boldsymbol{u}$ where $\boldsymbol{u} = \sqrt{w_i}\boldsymbol{x} = (\sqrt{w_i}, \sqrt{w_i}x_1, \sqrt{w_i}x_2, \sqrt{w_i}x_3)'$. There is no intercept because the vector of ones has been replaced by a vector of the $\sqrt{w_i}$'s. Copy and paste the commands for this problem into *R*. The commands fit weightd least squares and the equivalent OLS regression without an intercept. Include one page of output in *Word*.