Linear Statistical Models (MAST30025) Summary Notes
Akira Wang, Callum Holmes, Calvin Zhang, Anand Bharadwaj

1 Linear Algebra Review

1.1 Transposition:

The *transpose* of a matrix results when the rows and columns are interchanged.

A matrix X is **symmetric** if and only if (iff) $X^T = X$.

Properties that should be noted:

1. $(X^T)X = X$.
2. $(XY)^T = Y^T X^T$.

1.2 Identity:

The matrix *identity* I is a square matrix of arbitrary size with 1's on the diagonals (diags).

A $k \times k$ identity matrix is denoted as I_k.

1.3 Inverse:

If X is a **square matrix** such that $|A| \neq 0$, then its *inverse* is the matrix X^{-1} of the same size which satisfies the following condition:

$$XX^{-1} = X^{-1}X = I \tag{1}$$

Properties that should be noted:

1. $(X^{-1})^{-1} = X$.
2. $(XY)^{-1} = Y^{-1}X^{-1}$.
3. $(X^T)^{-1} = (X^{-1})^T$

1.4 Orthogonal Vectors and Matrices:

Two $n \times 1$ vectors $\mathbf{x}$ and $\mathbf{y}$ are *orthogonal* iff:

$$\mathbf{x}^T\mathbf{y} = \sum_{i=1}^{n} x_i y_i = 0. \tag{2}$$

A square matrix X is *orthogonal* iff:

$$X^T X = I. \tag{3}$$

If X is orthogonal, then

$$X^{-1} = X^T. \tag{4}$$

1.5 Orthogonality:

$$\text{If } \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \text{ then } \mathbf{x}^T\mathbf{x} = ||\mathbf{x}||^2 = \sum_{i=1}^{n} x_i^2 \tag{5}$$

The square root of $||\mathbf{x}^2||$, denoted by $||\mathbf{x}||$, is called the **norm** or **length** of $\mathbf{x}$.

A matrix X is an orthogonal matrix iff the columns of X form an orthonormal set.

1.6 Eigenvalues and Eigenvectors:

Suppose A is a $k \times k$ matrix and $\mathbf{x}$ is a $k \times 1$ nonzero vector which satisfies the equation

$$A\mathbf{x} = \lambda\mathbf{x}. \tag{6}$$

We say that λ is an *eigenvalue* of A, with associated *eigenvector* $\mathbf{x}$.

To find the eigenvalues, solve the following equation:

$$|A - \lambda I| = 0. \tag{7}$$

To then find the eigenvector(s) of A associated with the eigenvalue(s), solve the linear system of equations:

$$A\mathbf{x} = \lambda\mathbf{x}. \tag{8}$$

The linear system should have an infinite number of solutions, which will always happen for an eigenvector system.

Note:
If A is **symmetric**, then all its eigenvalues are **real**, and its eigenvectors all **orthogonal**.

1.7 Diagonalization:

Let A be a **symmetric** $k \times k$,matrix. Then an **orthogonal** matrix P exists such that

$$P^T AP = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_k \end{bmatrix} \text{,where } \lambda_i, i = 1, 2, \dots, k, \text{are the eigenvalues of } A. \tag{9}$$

We say that P *diagonalizes* A, where A is the *diagonalizable* matrix, and $P^T AP$ is the *diagonalized* matrix.

1.8 Linear Independence:

A set of vectors is *linearly depending* iff there exists some numbers $\alpha_1, \alpha_2, \dots, \alpha_k$, which are not all zero, such that

$$\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \alpha_k\mathbf{x}_k = \mathbf{0}. \tag{10}$$

If all α are zero, then they are are *linearly independent*.

If a set of vectors is linearly dependent, then at least one of the vectors can be written as a **linear combination** of some or all the other vectors.

1.9 Rank:

Some definitions (Yao-Ban):

1. A Tall Matrix has more rows than columns ($m > n$)
2. Short Matrix has more columns than rows ($n > m$)
3. **A matrix X : $r(X) =$ No. columns $\Rightarrow$ X is of *full rank*. Short matrices will never be full rank.**

The *rank* of X, denoted by $r(X)$, is the greatest number of **linearly independent vectors** in the set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k\}$. **Hence** $r(X) \leq k$

Properties that should be noted:

1. For any matrix X, we have $r(X) = r(X^T) = r(X^T X)$
2. The rank of a diagonal matrix is equal to the number of nonzero diagonal entries in the matrix.
3. **The $n \times k$ matrix X is non-singular iff $r(X) = k$.**

1.10 Idempotence:

A square matrix A is *idempotent* iff

$$A^2 = A \tag{11}$$

If A is idempotent, then A is diagonalizable.

Properties that should be noted:

1. **The eigenvalues of any idempotent matrix are always either 0 or 1**
   Proof.

$$\begin{aligned} A^2\mathbf{x} &= A(A\mathbf{x}) \\ &= A(\lambda\mathbf{x}), \text{since } A\mathbf{x} = \lambda\mathbf{x} \\ &= \lambda(A\mathbf{x}) \\ &= \lambda^2\mathbf{x} \end{aligned}$$

   By definition, $\mathbf{x} \neq \mathbf{0}$, so $\lambda = \lambda^2$. Therefore $\lambda = 0, 1$.

2. If A is a **symmetric** and **idempotent** matrix, $r(A) = tr(A)$.

   Proof.
   A is symmetric. This implies it is diagonalizable i.e. $D = P^TAP$.

   As A is idempotent, it has $r(A)$ i.e. k eigenvectors, all of which are either 0 or 1. Hence $tr(D) = r(A)$.

   Therefore

$$\begin{aligned} r(A) &= r(P^TAP) = r(D) \text{ (as } P, P^T \text{ are non-singular matrices)} \\ &= tr(P^TAP) \text{ (by above)} \\ &= tr(PP^TA) \text{ (property of trace)} \\ &= tr(A) \quad (P^TP = I) \end{aligned}$$

1.11 Trace:

The *trace* of a square matrix X, denoted by $tr(X)$, is the sum of its diagonal entries

$$tr(X) = \sum_{i=1}^{k} x_{ii}. \tag{12}$$

Properties that should be noted:

1. If c is a scalar, $tr(cX) = ctr(X)$
2. If XY and YX both exist, $tr(XY) = tr(YX)$

1.12 Quadratic Forms:

Suppose A is a square matrix, and $\mathbf{y}$ is a $k \times 1$ vector containing variables. The quantity

$$q = \mathbf{y}^T A \mathbf{y} \tag{13}$$

is called a *quadratic form* in $\mathbf{y}$, and A is called the matrix of the *quadratic form.*

However, since q is a scalar, it can be re-expressed as

$$q = \sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij} y_i y_j. \tag{14}$$

Example:

$$\text{Let } \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}, \text{and } A = \begin{pmatrix} \boxed{2} & \boxed{3} & \boxed{1} \\ \boxed{1} & \boxed{2} & \boxed{0} \\ \boxed{4} & \boxed{6} & \boxed{3} \end{pmatrix}$$

Then

$$\begin{aligned} \mathbf{y}^T A \mathbf{y} =& \left| \; 2y_1^2 \; \right| + \left| \; 3y_1 y_2 + y_1 y_3 \; \right| + \left| \; y_2 y_1 + 2y_2^2 + 4y_3 y_1 \; \right| + \left| \; 0 + 6y_3 y_2 \; \right| + \left| \; 3y_3^2 \; \right| \\ =& \; 2y_1^2 + 2y_2^2 + 3y_3^2 + 4y_1 y_2 + 5y_1 y_3 + 6y_2 y_3 \end{aligned}$$

This can be found from either the summation formula (14) or by multiplying out the matrices.

Positive Definiteness:

1. If $\mathbf{y}^T A\mathbf{y} > 0 \quad \forall \mathbf{y} \neq \mathbf{0}$, then A and the quadratic form are ***positive definite***.
2. If $\mathbf{y}^T A\mathbf{y} \geq 0 \quad \forall \mathbf{y}$, then A and the quadratic form are ***positive semi-definite***.

1.13 Differentiation of Quadratic Forms:

Suppose we have a vector of variables $\mathbf{y} = (y_1, y_2, \ldots, y_k)^T$, and some scalar function of them

$$z = f(\mathbf{y}). \tag{15}$$

We can then define the derivative of z with respect to $\mathbf{y}$ as follows:

$$\frac{\partial z}{\partial \mathbf{y}} = \begin{bmatrix} \frac{\partial z}{\partial y_1} \\ \frac{\partial z}{\partial y_2} \\ \vdots \\ \frac{\partial z}{\partial y_k} \end{bmatrix} \tag{16}$$

To do so, you can take the **quadratic form** of z and **partial derive** with respect to $\mathbf{y}$.

Example from Figure 14 and 15:

$$\frac{\partial z}{\partial \mathbf{y}} = \begin{bmatrix} 4y_1 + 4y_2 + 5y_3 \\ 4y_2 + 4y_1 + 6y_3 \\ 6y_3 + 5y_1 + 6y_2 \end{bmatrix} \tag{17}$$

Important Property!!!
Added another one here, kinda needed it if you want to prove MLS estimators

1. If $z = \mathbf{a}^T\mathbf{y}$ then $\frac{\partial z}{\partial \mathbf{y}} = \mathbf{a}$
2. If $z = \mathbf{y}^T A\mathbf{y}$, then $\frac{\partial z}{\partial \mathbf{y}} = A\mathbf{y} + A^T\mathbf{y}$
3. In particular, if A is **symmetric**, then $\frac{\partial z}{\partial \mathbf{y}} = 2A\mathbf{y}$.

2 Random Vectors

2.1 Expectation:

Note that we will denote random variables (r.v's) as lower-cases, according to linear algebra notation.

We define the *expectation* of a random vector $\mathbf{y}$ as follows:

$$\text{If } \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix}, \text{then } E[\mathbf{y}] = \begin{bmatrix} E[y_1] \\ E[y_2] \\ \vdots \\ E[y_k] \end{bmatrix} \tag{18}$$

Properties that should be noted:

1. If $\mathbf{a}$ is a vector of constants, then:
   (a) $E[\mathbf{a}] = \mathbf{a}$.
   (b) $E[\mathbf{a}^T\mathbf{y}] = \mathbf{a}^T E[\mathbf{a}]$.
2. If A is a matrix of constants, then $E[A\mathbf{y}] = AE[\mathbf{y}]$.

2.2 Variance:

Recall that the variance of a r.v Y with mean μ is defined to be $E[(Y-\mu)^2]$.

We define the *variance* or *covariance matrix* of the random vector $\mathbf{y}$ to be

$$\text{var}(\mathbf{y}) = E[(\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})^T]. \tag{19}$$

The diagonal elements of the **covariance matrix** are the variances of the elements of $\mathbf{y}$:

$$[\text{var}(\mathbf{y})]_{ii} = y_i, \qquad i = 1, 2, \ldots, k. \tag{20}$$

The off-diagonal elements are the covariances of the elements:

$$[\text{var}(\mathbf{y})]_{ij} = cov(y_i, y_j) = E[(y_i - \mu_i)(y_j - \mu_j)] \tag{21}$$

This means that the covariance matrix is always **symmetric**.

Very Important Property!!!:
Suppose that $\mathbf{y}$ is a random vector with $\text{var}(\mathbf{y}) = V$, then,
if A is a matrix of **constants**, $\text{var}(A\mathbf{y}) = AVA^T$.

Properties that should be noted:

1. If $\mathbf{a}$ is a vector of constants, then $\text{var}(\mathbf{a}^T\mathbf{y}) = \mathbf{a}^T V\mathbf{a}$.
2. V is positive semi-definite, meaning $\mathbf{a}^T V\mathbf{a} \geq 0 \quad \forall \mathbf{a}$

Example. Assume that X is a matrix of full rank ($r(X)$ = number of columns), which implies X^TX is non-singular (invertible). Let

$$\mathbf{z} = (X^TX)^{-1}X^T\mathbf{y} = A\mathbf{y}. \tag{22}$$

Then (using $V = \text{var}(\mathbf{y})$),

$$\begin{aligned}
\text{var}(\mathbf{z}) &= AVA^T \\
&= [(X^TX)^{-1}X^T]\sigma^2 I[(X^TX)^{-1}X^T]^T \\
&= (X^TX)^{-1}X^T(X^T)^T[(X^TX)^{-1}]^T\sigma^2 \qquad (X^T(X^T)^T = X^TX) \\
&= (X^TX)^{-1}[(X^TX)^{-1}]^T(X^TX)\sigma^2 \\
&= (X^TX)^{-1}\sigma^2 \qquad \text{, as } [(X^TX)^{-1}]^T = [(X^TX)^T]^{-1} = [(X^TX)^{-1}].
\end{aligned}$$

Hence

$$\text{var}(\mathbf{z}) = (X^TX)^{-1}\sigma^2 \tag{23}$$

2.3 Matrix Square Root:

The square root of a matrix A is a matrix B such that $B^2 = A$. (Note: this is not a common thing!) If A is **symmetric** and **positive semi-definite**, there is a unique symmetric positive semi-definite square root, called the *principle root*, denoted $A^{\frac{1}{2}} = (P\Lambda^{\frac{1}{2}}P^T)$

Proof.

$$\begin{aligned}
A &= P\Lambda P^T \\
&= (P\Lambda^{\frac{1}{2}}P^T)(P\Lambda^{\frac{1}{2}}P^T)
\end{aligned}$$

So clearly the 'square root' $A^{\frac{1}{2}} = P\Lambda^{\frac{1}{2}}P^T$

2.4 Multivariate Normal Distribution:

We say that

$$\mathbf{x} = A\mathbf{z} + \mathbf{b} \tag{24}$$

follows a multivariate normal distribution, with just $\boldsymbol{\mu} = \mathbf{b}$ and covariance matrix $\sum = AA^T$, and write $\mathbf{x} \sim MVN(\boldsymbol{\mu}, \sum)$.

Any linear combination of multivariate normal's results in another multivariate normal. For example:

$$\text{If } \mathbf{x} \sim MVN(\boldsymbol{\mu}, \sum), \text{then } \mathbf{y} = A\mathbf{x} + \mathbf{b} \sim MVN(A\boldsymbol{\mu} + \mathbf{b}, A\sum A^T). \tag{25}$$

2.5 Random Quadratic Forms:

We have seen that a matrix induces a quadratic form (multivariate function) that looks like $\mathbf{y}^T A\mathbf{y}$. The form becomes a scalar function of r.v's, so it itself is a r.v.

Theorem 3.2:
Let $\mathbf{y}$ be a random vector with $E[\mathbf{y}] = \boldsymbol{\mu}$, $V = \text{var}(\mathbf{y})$, and A be a matrix of constants. Then

$$E[\mathbf{y}^T A\mathbf{y}] = tr(AV) + \mu^T A\boldsymbol{\mu} \tag{26}$$

Example:
Let $\mathbf{y}$ be a 2×1 random vector with

$$E[\mathbf{y}] = \boldsymbol{\mu} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \qquad V = \text{var}(\mathbf{y}) = \begin{bmatrix} 2 & 1 \\ 1 & 5 \end{bmatrix}. \tag{27}$$

$$\text{And let } A = \begin{pmatrix} 4 & 1 \\ 1 & 2 \end{pmatrix}$$

The quadratic form is given as

$$\mathbf{y}^T A\mathbf{y} = 4y_1^2 + 2y_1y_2 + 2y_2^2. \tag{28}$$

The expectation of this form is

$$E[\mathbf{y}^T A\mathbf{y}] = 4E[y_1^2] + 2E[y_1y_2] + 2E[y_2^2]. \tag{29}$$

From the given covariance matrix V,

1. $V_{11} = 2 = \text{var}(y_1) = E[y_1^2] - E[y_1]^2 = E[y_1^2] - 1^2$
2. $V_{22} = 5 = \text{var}(y_2) = E[y_2^2] - E[y_2]^2 = E[y_2^2] - 3^2$

Solving both equations gives $E[y_1^2] = 3$ and $E[y_2^2] = 14$. Finally,

$$V_{12} = V_{21} = 1 = cov(y_1, y_2) = E[y_1 y_2] - E[y_1]E[y_2] = E[y_1 y_2] - (1 \times 3) \tag{30}$$

Solving the equation gives $E[y_1 y_2] = 4$, which gives

$$E[\mathbf{y}^T A \mathbf{y}] = 4 \times 3 + 2 \times 4 + 2 \times 14 = 48. \tag{31}$$

BUT, from Theorem 3.2,

$$\begin{aligned} E[\mathbf{y}^T A \mathbf{y}] &= tr(AV) + \boldsymbol{\mu}^T A \boldsymbol{\mu} \\ &= 9 + 11 + 7 + 21 \\ &= 48. \end{aligned}$$

2.6 Non-central χ^2 Distribution:

Theorem 3.3:
Let $\mathbf{y} \sim MVN(\boldsymbol{\mu}, I)$ be a $k \times 1$ random vector. Then

$$\mathbf{y}^T \mathbf{y} \sim \chi^2_{k,\lambda} \tag{32}$$

where

$$k = \text{number of parameters} \qquad \lambda = \frac{1}{2}\boldsymbol{\mu}^T \boldsymbol{\mu}.$$

WARNING:
R defines λ to be $\boldsymbol{\mu}^T \boldsymbol{\mu}$, so **double** it before putting it in!!!

Properties that should be noted:

1. $\text{var}(x) = 2k + 8\lambda$.
2. The non-centrality parameter λ is 0 iff $\boldsymbol{\mu} = \mathbf{0}$, resulting in a central χ^2_k distribution.

Theorem 3.4:
Let $X^2_{k_1,\lambda_1}, \ldots, X^2_{k_k,\lambda_k}$ be a collection of n independent non-central χ^2 random variables. Then

$$\sum_{i=1}^{n} X^2_{k_i,\lambda_i} \tag{33}$$

has a non-central χ^2 distribution with $k = \sum_{i=1}^n$ d.o.f and non-centrality parameter $\lambda = \sum_{i=1}^n \lambda_i$.
Setting $\lambda_i = 0$ results in a sum of independent χ^2 distributions.

2.7 Distribution of Quadratic Forms:

Theorem 3.5:
Let $\mathbf{y} \sim MVN(\boldsymbol{\mu}, I)$ be a $n \times 1$ random vector and let A be a $n \times n$ **idempotent** matrix. Then

$$\mathbf{y}^T A \mathbf{y} \sim \chi^2_{k,\lambda} \tag{34}$$

where

$$k = r(A) \qquad\qquad \lambda = \frac{1}{2}\boldsymbol{\mu}^T A \boldsymbol{\mu}.$$

Some definitions (Yao-Ban):
Let J_n be an $n \times n$ matrix filled with 1's. Example:

$$J_2 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \tag{35}$$

Example:
Let y_1 and y_2 be independent normal r.v's with means 3 and -2 respectively, and variance 1. Let

$$A = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \frac{1}{2}J_2 \tag{36}$$

Since A is **symmetric** and **idempotent**, it has rank 1. Therefore

$$\mathbf{y}^T A \mathbf{y} = \frac{1}{2}\begin{bmatrix} y_1 & y_2 \end{bmatrix} J_2 \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \frac{1}{2}y_1^2 + y_1 y_2 + \frac{1}{2}y_2^2 \tag{37}$$

has a non-central χ^2 distribution with 1 d.o.f and non-centrality parameter

$$\lambda = \frac{1}{2}\begin{bmatrix}3 & -2\end{bmatrix} J_2 \begin{bmatrix}3\\-2\end{bmatrix} = \frac{1}{4} \tag{38}$$

Theorem 3.6:
Let $\mathbf{y} \sim MVN(\mathbf{0}, I)$ be a $n \times 1$ random vector and let A be a $n \times n$ **idempotent** matrix. Then

$$\mathbf{y}^T A \mathbf{y} \sim \chi^2_k \tag{39}$$

where $k = r(A)$.

Theorem 3.7:
Let $\mathbf{y} \sim MVN(\boldsymbol{\mu}, \sigma^2 I)$ be a $n \times 1$ random vector and let A be a $n \times n$ **idempotent** matrix. Then

$$\frac{1}{\sigma^2}\mathbf{y}^T A \mathbf{y} \sim \chi^2_{k,\lambda} \tag{40}$$

where

$$k = r(A) \qquad\qquad \lambda = \frac{1}{2\sigma^2}\boldsymbol{\mu}^T A \boldsymbol{\mu}.$$

Theorem 3.8:
Let $\mathbf{y} \sim MVN(\boldsymbol{\mu}, V)$ be a $n \times 1$ random vector and let A be a $n \times n$ **idempotent** matrix. Then

$$\frac{1}{\sigma^2}\mathbf{y}^T A \mathbf{y} \sim \chi^2_{k,\lambda} \tag{41}$$

where You had an extra sigma2 below, I've removed it. Also, not sure if you can say $r(AV) = r(A) = k$ as V may not necessarily be invertible so consider removing that.

$$k = r(A) \qquad\qquad \lambda = \frac{1}{2}\boldsymbol{\mu}^T A \boldsymbol{\mu}$$

iff AV is **idempotent** and $r(AV) = r(A) = k$.

Theorem 3.9:
Let $\mathbf{y} \sim MVN(\mathbf{0}, V)$ be a $n \times 1$ random vector and let A be a $n \times n$ **idempotent** matrix. Then

$$\frac{1}{\sigma^2}\mathbf{y}^T A \mathbf{y} \sim \chi^2_k \tag{42}$$

where $k = r(A)$, and AV is **idempotent** and $r(AV) = r(A) = k$.

Theorem 3.10:
Had an extra sigma in this formula Let $\mathbf{y} \sim MVN(\boldsymbol{\mu}, V)$ be a $n \times 1$ random vector. Then

$$\mathbf{y}^T V^{-1} \mathbf{y} \sim \chi^2_{k,\lambda} \tag{43}$$

where $k = n$, see below

$$k = n \qquad\qquad \lambda = \frac{1}{2}\boldsymbol{\mu}^T V^{-1} \boldsymbol{\mu}.$$

2.8 Independence of Quadratic Forms:

Theorem 3.11:
Let $\mathbf{y} \sim MVN(\boldsymbol{\mu}, V)$ be a $n \times 1$ random vector with non-singular variance V, and let A and B be square matrices. Then $\mathbf{y}^T A \mathbf{y}$ and $\mathbf{y}^T B \mathbf{y}$ are independent iff

$$AVB = 0 \tag{44}$$

Example:
Let y_1 and y_2 follow a MVN distribution with covariance matrix

$$V = \begin{bmatrix} a & b \\ c & d \end{bmatrix}. \tag{45}$$

Consider the symmetric matrices

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}. \tag{46}$$

Then

$$\mathbf{y}^T A \mathbf{y} = y_1^2, \qquad \mathbf{y}^T B \mathbf{y} = y_2^2. \tag{47}$$

To find if they are independent, we solve $AVB = 0$ (the zero matrix)

$$\begin{aligned} AVB &= \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix} \end{aligned}$$

This means that $b = 0$ and implies that $cov(y_1, y_2) = 0$. Therefore, the quadratic forms are independent.

Theorem 3.12:
Let $\mathbf{y} \sim MVN(\boldsymbol{\mu}, V)$ be a $n \times 1$ random vector with non-singular variance V, and let A and B be square matrices. Then $\mathbf{y}^T A\mathbf{y}$ and $B\mathbf{y}$ are independent iff

$$BVA = 0 \tag{48}$$

Theorem 3.15: Cochran-Fisher Theorem:
Let $\mathbf{y} \sim MVN(\boldsymbol{\mu}, \sigma^2 I)$ be a $n \times 1$ random vector. Decompose the sum of squares of $\mathbf{y}/\sigma$ into the equation

$$\frac{1}{\sigma^2}\mathbf{y}^T\mathbf{y} = \sum_{i=1}^{m} \frac{1}{\sigma^2}\mathbf{y}^T A_i \mathbf{y} \tag{49}$$

Where $\sum A_i = I$, each A_i idempotent, and $A_i A_j = 0,\ i \neq j$. Then EACH quadratic form is independent with a non-central χ^2 distribution iff $\sum r(A_i) = n$.
This theorem is quite useful for SS_{Total} decomposition later.

3 The Full Rank Model

3.1 Linear Model

The linear model is given as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i, \text{ for all } i = 1, 2, \ldots, n \tag{50}$$

and can be expressed in matrix form (DON'T FORGET COLUMNS OF 1's)

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{22} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}. \tag{51}$$

A common assumption is that $\epsilon \sim MVN(\mathbf{0}, \sigma^2 I)$. However, X and β are **NOT** random vectors and we therefore assume that there is **no uncertainty/error** in these measurements.

For this section, we assume that X is of full rank. Therefore, it means that X^TX is invertable and $(X^TX)^{-1}$ exists.

Important Property!!!
Since X has dimension $n \times (k+1)$ (+1 for columns of 1's), we say the model has *full rank* iff X is full rank - $r(x) = k+1$.

Example:
We wish to analyse the selling price of a house (y), and think that it depends on two variables: age (x_1) and area (x_2). Our linear model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \tag{52}$$

We sample 5 random houses and obtain data:

Price	Age	Area
50	1	1
40	5	1
52	5	2
47	10	2
65	20	3

The model generates 5 linear equations in total

$$\begin{aligned}
50 &= \beta_0 + 1\beta_1 + 1\beta_2 + \epsilon_1 \\
40 &= \beta_0 + 5\beta_1 + 1\beta_2 + \epsilon_2 \\
52 &= \beta_0 + 5\beta_1 + 2\beta_2 + \epsilon_3 \\
47 &= \beta_0 + 10\beta_1 + 2\beta_+\epsilon_4 \\
65 &= \beta_0 + 20\beta_1 + 3\beta_2 + \epsilon_5
\end{aligned}$$

The matrix form of the model is $\mathbf{y} = X\beta + \epsilon$, where

$$\mathbf{y} = \begin{bmatrix} 50 \\ 40 \\ 52 \\ 47 \\ 65 \end{bmatrix}, X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 5 & 1 \\ 1 & 5 & 2 \\ 1 & 10 & 2 \\ 1 & 20 & 3 \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \end{bmatrix}. \tag{53}$$

A direct calculation will show that X is of full rank (more rows than columns). This is an example of *multiple regression*.

3.2 Parameter Estimation Using Least Squares

We can estimate the parameters $\beta_0, \beta_1, \ldots, \beta_k$ by using the ***method of least squares***. To do so, we assume that the error vector ϵ has mean 0 and variance $\sigma^2 I$.

In other words, we assume the model is **unbiased** and that the errors are **independent of the responses** and **uncorrelated with each other**. We do **NOT** necessarily assume that the errors are *independent* of each other.

$$\mathbf{y} = X\beta + \epsilon \tag{54}$$

The error term ϵ is the **only** random term in the model, so

$$\begin{aligned} E[\mathbf{y}] &= X\beta + E[0] \\ &= X\beta \end{aligned}$$

and

$$\text{var}(\mathbf{y}) = \sigma^2 I. \tag{55}$$

Suppose that $b_0, b_1, \ldots, b_k$ are the estimates of the parameters $\beta_0, \beta_1, \ldots, \beta_k$. Then we can **estimate** the expected value of y_i by

$$\hat{E[y_i]} = b_0 + b_1 x_{i1} + \ldots, b_k x_{ik}. \tag{56}$$

The resulting ith residual is defined to be the **difference** between the **observed value (estimate)** and the **estimated value**:

$$e_i = y_i - \hat{E[y_i]} \tag{57}$$

If the estimates are good, then the residuals should be very close to the errors (distance between line and data):

$$\begin{aligned} \epsilon_i &= y_i - \hat{E[y_i]} = y_i - \hat{E[y_i]} + \hat{E[y_i]} - E[y_i] \\ &= e_i + \hat{E[y_i]} - E[y_i]. \end{aligned}$$

We choose our estimates to *minimise* the residuals - specifically, we *minimise* the sum of squares of the residuals. This is the **least squares estimation** of the parameters.

We define the vectors of estimated parameters and residuals as

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix}, \mathbf{e} = \begin{bmatrix} e_0 \\ e_1 \\ \vdots \\ e_k \end{bmatrix}. \tag{58}$$

where we wish choose a $\mathbf{b}$ to minimise $\mathbf{e}^T\mathbf{e}$.

Theorem 4.1:
Let $\mathbf{y} = X\beta + \epsilon$ where X is a $n \times (k+1)$ matrix of full rank (non-singular), β is a $(k+1) \times 1$ vector of *parameters*, and ϵ is a $n \times 1$ random vector with mean 0. Then the least squares *estimator* for β is

$$\mathbf{b} = (X^TX)^{-1}X^T\mathbf{y} \tag{59}$$

Example:
Recall that the simple linear regression model can be written as a linear model with two parameters

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i \tag{60}$$

which gives

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}. \tag{61}$$

Then

$$\begin{aligned} X^TX &= \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \vdots & x_3 \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \\ &= \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} \end{aligned}$$

and

$$X^T\mathbf{y} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \vdots & x_3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$
$$= \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}.$$

We have (using the det formula for a 2×2 matrix)

$$(X^TX)^{-1} = \frac{1}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix}. \tag{62}$$

Therefore the **least squares estimator** for β is

$$\begin{aligned} \mathbf{b} &= (X^TX)^{-1}X^T\mathbf{y} \\ &= \frac{1}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix} \\ &= \frac{1}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i \\ n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i \end{bmatrix} \end{aligned}$$

The estimator for the *slope* of the regression line is:

$$\begin{aligned} b_1 &= \frac{n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}, \text{ divide by } \frac{1}{n^2} \text{ to get means} \\ &= \frac{\frac{1}{n}\sum_{i=1}^n x_i y_i - \frac{1}{n^2}\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\frac{1}{n}\sum_{i=1}^n x_i^2 - \frac{1}{n^2}(\sum_{i=1}^n x_i)^2} \\ &= \frac{\bar{xy} - \bar{x}\bar{y}}{\bar{x^2} - \bar{x}^2} \\ &= \frac{cov(x,y)}{\text{var}(x)} \end{aligned}$$

The estimator for the *intercept* of the regression line is:

$$b_0 = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \tag{63}$$

1. An estimator is "good" when two desirable properties are satisfied:
   (a) The estimator is unbiased
   (b) The variance is small
2. **Theorem 4.2**:
   In the general linear model, the least squares estimator $\mathbf{b} = (X^TX)^{-1}X^T\mathbf{y}$ is an ***unbiased*** estimator for β. In other words,

$$E[\mathbf{b}] = \beta, \qquad \text{var}(\mathbf{b}) = (X^TX)^{-1}\sigma^2 \tag{64}$$

Proof.

1. The expectation is unbiased

$$\begin{aligned} E[\mathbf{b}] &= E[(X^TX)^{-1}X^T\mathbf{y}] \\ &= (X^TX)^{-1}X^TE[\mathbf{y}] \\ &= (X^TX)^{-1}X^T(X\beta) \\ &= (X^TX)^{-1}(X^TX)\beta \\ &= \beta \end{aligned}$$

2. The variance is unbiased

$$\begin{aligned} \text{var}(\mathbf{b}) &= \text{var}((X^TX)^{-1}X^T\mathbf{y}), \text{ Using var}(A\mathbf{y}) = AVA^T \\ &= [(X^TX)^{-1}X^T]\sigma^2I[(X^TX)^{-1}X^T]^T \\ &= (X^TX)^{-1}(X^TX)[(X^TX)^{-1}X^T]^{-1}\sigma^2 \\ &= (X^TX)^{-1}\sigma^2 \end{aligned}$$

3.3 BLUE and Gauss-Markov Theorem

Linear (*as to Least Squares*) estimators take the form $L\mathbf{y}$, where L is a matrix of constants. The least squares estimator is a *linear estimator* with $L = (X^TX)^{-1}X^T$.

Suppose we have a model with some parameters β and linear estimators $\mathbf{b}$ for these parameters.

1. Some definitions:
   If $E[\mathbf{b}] = \beta$ and the variances of $b_0, b_1, \ldots, b_k$ are minimised over all linear estimators, then $\mathbf{b}$ is called a ***Best Linear Unbiased Estimator*** of β (BLUE).

2. **Theorem 4.3 (Gauss-Markov Theorem)**:
   In the full rank general linear model $\mathbf{y} = X\beta + \epsilon$, the least squares estimator $\mathbf{b} = (X^T X)^{-1} X^T \mathbf{y}$ is the unique BLUE for β.

Example:
Consider the house price example from above (51 - 52). The variance of the least squares estimators is given by

$$(X^T X)^{-1}\sigma^2 = \begin{bmatrix} 2.31 & 0.16 & -1.88 \\ 0.16 & 0.03 & -0.2 \\ -1.88 & -0.2 & 1.98 \end{bmatrix} \sigma^2. \tag{65}$$

This means that there is **NO unbiased** linear estimator of $(\beta_0, \beta_1, \beta_2)$ which has smaller variance than $(2.31\sigma^2, 0.03\sigma^2, 1.98\sigma^2)$ respectively. This is true even if we do not know what σ^2 is.

Theorem 4.5:
Take the full rank general linear model $\mathbf{y} = X\beta + \epsilon$ and let $\mathbf{t}$ be a $\mathbf{t}$ is a $(k+1) \times 1$ vector of constants. Then the BLUE for $\mathbf{t}^T\beta$ is $\mathbf{t}^T\mathbf{b}$, where $\mathbf{b}$ is the least squares estimator for β.

Example PREDICTION:
Consider the house price example from above (51 - 52, 64). The model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon, \tag{66}$$

where y is the house price, x_1 is its age, and x_2 is its area.

Suppose we are then given a specific house with age x_1^* and area x_2^* that we wish to estimate (predict) what price it will fetch.

1. We want to estimate the linear function of the parameters

$$E[y] = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* = \mathbf{t}^T\beta, \text{ where } \mathbf{t} = \begin{bmatrix} 1 & x_1^* & x_2^* \end{bmatrix}^T. \tag{67}$$

Therefore an unbiased estimator for the specific house price is

$$\mathbf{t}^T\mathbf{b} = \begin{bmatrix} 1 & x_1^* & x_2^* \end{bmatrix} \mathbf{b} = b_0 + b_1 x_1^* + b_2 x_2^* \tag{68}$$

where $\mathbf{b}$ is the least squares estimator for β.

3.4 Variance Estimation

If you recall, we previously assumed that the errors ϵ (and thus $\mathbf{y}$) have a covariance matrix equal to $\sigma^2 I$.

Now, we want to be able to estimate the **common variance** σ^2. The reason we do this is to *create confidence intervals for the true value of the parameters.*

The common variance σ^2 can be written as

$$\sigma^2 = E\Big[\frac{(\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta)}{n}\Big], \tag{69}$$

and so a reasonable estimator for the variance might be

$$\hat{\sigma^2} = \frac{(\mathbf{y} - X\mathbf{b})^T(\mathbf{y} - X\mathbf{b})}{n}. \tag{70}$$

This however turns out to be **slightly biased** and requires some adjustment. We introduce,

Theorem 4.6:
The sample variance s^2, given as

$$s^2 = \frac{(\mathbf{y} - X\mathbf{b})^T(\mathbf{y} - X\mathbf{b})}{n - (k+1)} \tag{71}$$

is an **unbiased estimator** for σ^2.

Proof.

$$
\begin{aligned}
E[s^2] &= E\Big[\frac{(\mathbf{y}-X\mathbf{b})^T(\mathbf{y}-X\mathbf{b})}{n-(k+1)}\Big] \\
&= \frac{1}{n-(k+1)}E[(\mathbf{y}-X\mathbf{b})^T(\mathbf{y}-X\mathbf{b})], \text{ sub in } \mathbf{b}=(X^TX)^{-1}X^T\mathbf{y} \\
&= \frac{1}{n-(k+1)}E[(\mathbf{y}-X(X^TX)^{-1}X^T\mathbf{y})^T(\mathbf{y}-X(X^TX)^{-1}X^T\mathbf{y})] \\
&= \frac{1}{n-(k+1)}E[\mathbf{y}^T(I-X(X^TX)^{-1}X^T)(I-X(X^TX)^{-1}X^T)\mathbf{y}] \\
&= \frac{1}{n-(k+1)}E[\mathbf{y}^T(I-X(X^TX)^{-1}X^T)\mathbf{y}]
\end{aligned}
$$

This is of form $E[\mathbf{y}^TA\mathbf{y}]$ (quadratic form), which we can apply Theorem 3.2:

$$
E[\mathbf{y}^TA\mathbf{y}] = tr(AV) + \boldsymbol{\mu}^TA\boldsymbol{\mu}. \tag{72}
$$

1. This cancels out

$$
\begin{aligned}
\boldsymbol{\mu}^TA\boldsymbol{\mu} &= (X\beta)^T(I-X(X^TX)^{-1}X^T)X\beta \\
&= \beta^TX^TX\beta - \beta^TX^TX(X^TX)^{-1}X^TX\beta \\
&= I - I^2 \\
&= 0.
\end{aligned}
$$

2. So we are left with the trace

$$
\begin{aligned}
tr(AV) &= tr((I_n - X(X^TX)^{-1}X^T)\sigma^2I_n) \\
&= \sigma^2(tr(I_n) - tr(X(X^TX)^{-1}X^T)) \\
&= \sigma^2(n - tr((X^TX)^{-1}X^TX)) \\
&= \sigma^2(n - tr(I_{k+1})) \\
&= \sigma^2(n-(k+1)) + 0.
\end{aligned}
$$

3. This gives us

$$
\begin{aligned}
E[\mathbf{y}^TA\mathbf{y}] &= \sigma^2(n-(k+1)) + 0 \\
&= \sigma^2(n-(k+1)).
\end{aligned}
$$

4. Therefore we have proved that $E[s^2]$ is unbiased,

$$\begin{aligned} E[s^2] &= \frac{1}{n-(k+1)} E[\mathbf{y}^T(I - X(X^TX)^{-1}X^T)\mathbf{y}] \\ &= \frac{1}{n-(k+1)} E[\mathbf{y}^T A\mathbf{y}] \\ &= \frac{\sigma^2(n-(k+1))}{n-(k+1)} \\ &= \sigma^2. \end{aligned}$$

3.5 Diagnostic Plots

3.5.1 Residuals vs Fitted

1. Do the points have large residuals (unequal variances)? [We want small residuals around 0]
2. Is there a trend (bias) in the residuals? [We want no trend]
3. Is there a pattern (correlation) in the residuals? [We want no correlation]

3.5.2 QQ-Plot

We look for the points to follow the line (normally distributed). If not, then we look at how they deviate.

1. Are there a small number of outliers? [We want no outliers]
2. Is there any over / under estimation in the tails? [We don't want any over / under estimation]
3. Are the points skewed? [We want an even spread and no skew]

3.5.3 Scale-Location

1. Are there points with high residuals (unequal variances)? [We want small residuals around 0]
2. Is there a trend in the size of the residuals (heteroskedasticity)? [We want no trend in the size of residuals]

3.5.4 Leverage vs Standardized Residuals

1. Are there points with high residuals (unequal variances)? [We want small residuals around 0]
2. Are there points with high leverage? If so, this can be bad!
3. Are there points with high Cook's Distance (implies bad fit)? [We want all points under 0.5]
4. Is there a pattern (correlation) in the residuals? [We want no correlation and even spread]

3.6 Regression Through The Origin

Previously we had always considered the linear model to include a parameter β_0 (the intercept), which is associated with the 1's columns in the **design matrix** X.

It is perfectly reasonable to assume (from prior knowledge of the data) that **no** intercept is needed. The linear model becomes

$$y_i = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon. \tag{73}$$

The **least squares estimator** is still

$$\mathbf{b} = (X^T X)^{-1} X^T \mathbf{y}. \tag{74}$$

However, the variance estimator becomes

$$s^2 = \frac{SS_{Res}}{n-p}, \tag{75}$$

where $p = k$ instead of $k+1$ since we don't estimate the parameter β_0.

3.7 Standardised Residuals

To assess the *fit* and *model assumptions* of our linear models, we look at our **residuals**.

If there is an extremely large residual, or a pattern in the residuals, we might question our assumptions.

The variance of the errors is *assumed to be* $\sigma^2 I$, but the **variance of the residuals is not** $\sigma^2 I$. In general, the variance of a particular residual depends on how *far away from the centre* the design variables are.

The further away, the smaller variance of the residual.

We can calculate the variance of the residuals

$$\mathbf{e} = \mathbf{y} - X\mathbf{b} = (I - X(X^T X)^{-1} X^T)\mathbf{y}. \tag{76}$$

We define $H = X(X^T X)^{-1} X^T$ as the *hat matrix*,

$$\begin{aligned} \text{var}(\mathbf{e}) &= \text{var}(I - H)\mathbf{y} \\ &= (I - H)\sigma^2 (I - H)^T, \text{ using } AVA^T \\ &= \sigma^2 (I - H). \end{aligned}$$

Then, we can can find the ***standardised residuals*** with

$$z_i = \frac{e_i}{\sqrt{s^2(1 - H_{ii})}}. \tag{77}$$

The standardised residuals have (approximately) equal variance and can therefore be used to compare.

3.8 Leverage and Cook's Distance

Consider the scenario when we calculate the fitted values by $\hat{\mathbf{y}} = X\mathbf{b} = H\mathbf{y}$.. H tends to have its largest values on the diagonal (the best estimate for the mean y_i is generally close to y_i). The size of H_{ii} reflects how much $\hat{y}_i$ is based on y_i, as opposed to the other y_j.

Important!!!

1. If H_{ii} is particularly large, then y_i has a **large effect on the fit**. We thus define the *leverage* of point i as H_{ii}.
2. The leverage of a point must always fall between [0,1].
   $\Rightarrow$ Idempotent matrices (H) are limited between [0,1].

Points with **large leverage** have an unusually large effect on the estimated parameters. If this is combined with a **large residual**, then the corresponding point may distort the fit.

To check this, we calculate the ***Cook's distance*** of each point. This measures the change in the estimated parameters **b** if we **remove the point.**

The definition of **Cook's distance** is

$$D_i = \frac{(\mathbf{b_{(-1)}} - \mathbf{b})^T X^T X (\mathbf{b_{(-1)}} - \mathbf{b})}{(k+1)s^2} = \frac{z_i^2}{k+1}\left(\frac{H_{ii}}{1-H_{ii}}\right). \tag{78}$$

where $\mathbf{b_{(-1)}}$ is the estimated parameters if point i is **removed**, and the **idempotent matrix** $H \in [0, 1)$.

It is generally considered large if it is greater than 1, and small if it is less than 0.5.

3.9 Maximum Likelihood Estimation

MLE's are popular because they have good *asymptotic* properties:

1. As the same size goes to ∞ they are unbiased, normally distributed, and have minimum variance under certain conditions.

We assume that the errors are $MVN(\mathbf{0}, \sigma^2 I)$. In particular, this means that the errors are both **independent and uncorrelated**.

Given observed values **y** of the response variable, the errors ϵ are $\mathbf{y} - X\beta$. Since they are

independent, their joint density is given by

$$\begin{aligned} L(\beta, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\epsilon_i^2/2\sigma^2} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^{n} \epsilon_i^2/(2\sigma^2)} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(\mathbf{y}-X\beta)^T(\mathbf{y}-X\beta)/(2\sigma^2)}. \end{aligned}$$

This is also called the **likelihood** function of the parameters β and σ^2. We maximise the likelihood with respect to β to generate the MLE for β. To do this, we differentiate with respect to β and set the result to be $\mathbf{0}$.

It practice, it is usually easier to maximise the log-likelihood (denoted as $l(\beta, \sigma^2)$ since log is a monotonic function, with the maximum being at the same point.

$$\begin{aligned} \frac{\partial}{\partial\beta} l(\beta, \sigma^2) &= -\frac{1}{2\sigma^2} \frac{\partial}{\partial\beta} (\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T X\beta + \beta^T (X^T X)\beta) \\ &= -\frac{1}{2\sigma^2} (-2(X^T\mathbf{y}) + 2(X^T X)\beta) = 0 \\ (X^T X)\beta &= X^T\mathbf{y}, \ \textbf{These are just the normal equations!} \end{aligned}$$

Theorem 4.7:
In the full rank general linear model $\mathbf{y} = X\beta + \epsilon$, assume $\epsilon \sim MVN(0, \sigma^2 I)$. Then the MLE for β is also the least squares estimator:

$$\mathbf{b} = (X^T X) {-1} X^T\mathbf{y} \tag{79}$$

Theorem 4.8:
In the full rank general linear model $\mathbf{y} = X\beta + \epsilon$, assume $\epsilon \sim MVN(0, \sigma^2 I)$. Then the MLE for σ^2 is also the least squares estimator:

$$\tilde{\sigma^2} = \frac{SS_{Res}}{n}. \tag{80}$$

However, this is an unbiased estimator since the MLE is only asymptotically unbiased. Therefore, we prefer to use the sample variance which has the same asymptotic properties as $\tilde{\sigma^2}$, but is unbiased for all n:

$$s^2 = \frac{SS_{Res}}{n-p} = \frac{n}{n-p}\tilde{\sigma^2}. \tag{81}$$

3.10 Sufficiency

Previously, we showed that the least squares estimator is the BLUE for β, and that if the errors are **normally distributed**, it is also the MLE. Now we wish to see if the least squares estimators are ***sufficient***. That is, they contain all the information required about the estimates.

Theorem 4.9 (Fisher-Neyman Factorization Theorem):
Let $\mathbf{x}$ be a random variable with parameters $\boldsymbol{\theta}$, and let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a random sample drawn from this distribution, with join density $f(\mathbf{x}_1, \ldots, \mathbf{x}_n; \boldsymbol{\theta})$. Then the statistic $\mathbf{y} = u(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is sufficient for $\boldsymbol{\theta}$ iff f can be expressed as

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_n; \boldsymbol{\theta}) = g(\mathbf{y}; \boldsymbol{\theta}) h(\mathbf{x}_1, \ldots, \mathbf{x}_n). \tag{82}$$

where g is only the data and parameters, and h is only the x_i's.

Theorem 4.10:
In the full rank general linear model $\mathbf{y} = X\beta + \epsilon$, assume $\epsilon \sim MVN(0, \sigma^2 I)$. Then the estimators

$$\mathbf{b} = (X^T X) {-1} X^T \mathbf{y} \qquad\qquad s^2 = \frac{SS_{Res}}{n - p}$$

are **jointly sufficient** for β and σ^2.

Maximum likelihood theory also tells us that asymptotically $\mathbf{b}$ and s^2 have minimum variance. **This is also true for finite samples**:

Theorem 4.11:
In the full rank general linear model $\mathbf{y} = X\beta + \epsilon$, assume $\epsilon \sim MVN(0, \sigma^2 I)$. Then the estimators

$$\mathbf{b} = (X^T X) {-1} X^T \mathbf{y} \qquad\qquad s^2 = \frac{SS_{Res}}{n - p}$$

have the lowest variance among all unbiased estimators of β and σ^2

Theorem 4.11 is a **stronger** condition than BLUE because it includes non-linear estimators. We call this **UMVUE** (Uniformly Minimum Variance Unbiased Estimator). Since $\mathbf{y}$ and $\mathbf{b}$ are linear combinations of ϵ, they both are MVN's (Refer to Theorem 4.2).

3.11 Interval Estimation on Parameters

The least squares estimators gives us excellent **point estimates** for the parameters. To get an idea of how accurate these estimates are, we want to find **interval estimates**.

REMEMBER FROM NOW ON
We will always assume $\epsilon \sim MVN(0, \sigma^2 I)$ for MLEs.

Theorem 4.12:
In the full rank general linear model $\mathbf{y} = X\beta + \epsilon$, assume $\epsilon \sim MVN(0, \sigma^2 I)$. Then

$$\mathbf{b} = (X^T X)^{-1} X^T \mathbf{y} \sim MVN(\beta, (X^T X)^{-1}\sigma^2). \tag{83}$$

Theorem 4.13:
In the full rank general linear model $\mathbf{y} = X\beta + \epsilon$, assume $\epsilon \sim MVN(0, \sigma^2 I)$. Then

$$\frac{(n-p)s^2}{\sigma^2} = \frac{SS_{Res}}{\sigma^2} \sim \chi^2_{n-p}. \tag{84}$$

Proof.
Previously, we showed that SS_{Res} could be expressed as the **quadratic form**

$$SS_{Res} = (\mathbf{y} - X\mathbf{b})^T(\mathbf{y} - X\mathbf{b}) = \mathbf{y}^T A \mathbf{y} \tag{85}$$

where $A = I - X(X^T X)^{-1} X^T$ is **symmetric, idempotent, and has a rank** $r(A) = n - p$.
By assumption, we have $\mathbf{y} \sim MVN(X\beta, \sigma^2 I)$.
By Corollary 3.7, $\mathbf{y}^T A \mathbf{y}$ has a non-central χ^2 with $n - p$ d.o.f and non-centrality parameter

$$\lambda = \frac{1}{2\sigma^2} \boldsymbol{\mu}^T A \boldsymbol{\mu}. \tag{86}$$

But, $\boldsymbol{\mu} = E[\mathbf{y}] = X\beta$, so

$$\begin{aligned}
\lambda &= \frac{1}{2\sigma^2}(X\beta)^T [I - X(X^T X)^{-1} X^T] X\beta \\
&= \frac{1}{2\sigma^2}[\beta^T X^T X\beta - \beta^T X^T X (X^T X)^{-1} X^T X\beta] \\
&= 0
\end{aligned}$$

Hence

$$\frac{SS_{Res}}{\sigma^2} \sim \chi^2_{n-p}. \tag{87}$$

It can also be shown that $r((X^TX)^{-1}X^T) = r(X)$:

$$r(X) \geq r((X^TX)^{-1}X^T) \geq r((X^TX)^{-1}X^T)r(X), \text{ since } r(X), r(Y) \geq r(XY)$$
$$r(X) \geq r((X^TX)^{-1}X^T) \geq r((X^TX)^{-1}X^TX)$$
$$r(X) \geq r((X^TX)^{-1}X^T) \geq r(X)$$
$$r((X^TX)^{-1}X^T) = r(X)$$

Theorem 4.14:
In the full rank general linear model $\mathbf{y} = X\beta + \epsilon$, assume $\epsilon \sim MVN(0, \sigma^2 I)$. Then $\mathbf{b}$ and s^2 are independent.
Proof. Using Theorem 3.13 we have

$$\mathbf{b} = (X^TX)^{-1}X^T\mathbf{y} \qquad \frac{SS_{Res}}{\sigma^2} = \mathbf{y}^T\frac{[I - X(X^TX)^{-1}X^T]}{\sigma^2}\mathbf{y}$$

which are both quadratic forms. We can then let

$$B\mathbf{y} = (X^TX)^{-1}X^T\mathbf{y} \qquad \mathbf{y}^TA\mathbf{y} = \mathbf{y}^T\frac{[I - X(X^TX)^{-1}X^T]}{\sigma^2}\mathbf{y}.$$

Then we solve for independence using (Theorem 3.13 BVA)

$$\begin{aligned} BVA &= (X^TX)^{-1}X^T\sigma^2 I\frac{I - X(X^TX)^{-1}X^T}{\sigma^2} \\ &= (X^TX)^{-1}X^T - (X^TX)^{-1}X^TX(X^TX)^{-1}X^T \\ &= 0. \end{aligned}$$

We can now find a **confidence interval** for a single parameter, β_i (a point estimate).

Formula below didn't have -1 Consider the covariance matrix of $\mathbf{b} \sim MVN(\beta, (X^TX)^{-1}\sigma^2)$:

$$(X^TX)^{-1}\sigma^2 = \begin{bmatrix} c_{00} & c_{01} & \dots & c_{0k} \\ c_{10} & c_{11} & \dots & c_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k0} & c_{k1} & \dots & c_{kk} \end{bmatrix}\sigma^2. \tag{88}$$

The least squares estimator of β_i is b_i. The variance of b_i is the ith diagonal element of the covariance matrix, denoted $c_{ii}\sigma^2$.

The ith diagonal element is what we use to get the confidence interval for ith parameter.

Since b_i is normal,

$$\frac{b_i - \beta_i}{\sigma\sqrt{c_{ii}}} \sim Z, \text{ standard normal} \tag{89}$$

BUT, we do not know what σ^2 is. So we use the t distribution:

$$\left(\frac{b_i - \beta_i}{\sigma\sqrt{c_{ii}}}\right) / \left(\sqrt{\frac{SS_{Res}/\sigma^2}{n-p}}\right) = \left(\frac{b_i - \beta_i}{\sigma\sqrt{c_{ii}}}\right) / \left(\sqrt{\frac{s^2}{\sigma^2}}\right) = \frac{b_i - \beta_i}{s^2\sqrt{c_{ii}}} \sim t_{n-p} \tag{90}$$

Now we can derive a $100(1-\alpha)\%$ confidence interval (using a t distribution with $n-p$ d.o.f):

$$b_i \pm t_{\alpha/2} s\sqrt{c_{ii}}, \tag{91}$$

where c_{ii} is the ith diagonal element of $(X^TX)^{-1}$.

Critical Regions:

1. $Z > \Phi^{-1}(1-\alpha)$
2. $Z < \Phi^{-1}(\alpha)$
3. $Z \geq \Phi^{-1}(1-\frac{\alpha}{2})$

3.12 Interval Estimation on Response Variables

We proved (Lab-04 Q9) that if we want to estimate the function $\mathbf{t}^T\beta$, the BLUE is $\mathbf{t}^T\mathbf{b}$, where $\mathbf{b}$ is the least squares estimator of the parameters.

Since $\mathbf{b}$ is MVN, any linear combination of b's are normally distributed. We have

$$E[\mathbf{t}^T\mathbf{b}] = \mathbf{t}^T\beta. \tag{92}$$

Then, the Variance result give us

$$\text{var}(\mathbf{t}^T\mathbf{b}) = \mathbf{t}^T(X^TX)^{-1}\sigma^2\mathbf{t} = \mathbf{t}^T(X^TX)^{-1}\mathbf{t}\sigma^2. \tag{93}$$

Therefore

$$\frac{\mathbf{t}^T\mathbf{b} - \mathbf{t}^T\beta}{\sqrt{\mathbf{t}^T(X^TX)^{-1}\mathbf{t}\sigma^2}} \sim Z \tag{94}$$

BUT again, we do not know what σ is.

Since $\frac{SS_{Res}}{\sigma^2}$ is independent of $\mathbf{b}$, it is independent of $\mathbf{t}^T\mathbf{b}$.

Hence

$$\frac{\mathbf{t}^T\mathbf{b} - \mathbf{t}^T\beta/\sqrt{\mathbf{t}^T(X^TX)^{-1}\mathbf{t}\sigma^2}}{\sqrt{SS_{Res}/\sigma^2(n-p)}} = \frac{\mathbf{t}^T\mathbf{b} - \mathbf{t}^T\beta}{s\sqrt{\mathbf{t}^T(X^TX)^{-1}\mathbf{t}}} \sim t_{n-p} \tag{95}$$

Using the steps from (83), this gives the $100(1-\alpha)\%$ confidence interval

$$\mathbf{t}^T\mathbf{b} \pm t_{\alpha/2}s\sqrt{\mathbf{t}^T(X^TX)^{-1}\mathbf{t}}. \tag{96}$$

In particular, if we want to find a confidence interval for the expected response to particular set of x variables $x_1^*, x_2^*, \ldots, x_k^*$, we wish to predict

$$E[y] = \beta_0 + \beta_1 x_1^* + \cdots + \beta_k x_k^* = (\mathbf{x}^*)^T\beta \tag{97}$$

where $\mathbf{x}^* = \begin{bmatrix} 1 & x_1^* & x_2^* & \ldots & x_k^* \end{bmatrix}^T$.

This is a linear function of β, and therefore the $100(1-\alpha)\%$ confidence interval for it is

$$(\mathbf{x}^*)^T\beta \pm t_{\alpha/2}s\sqrt{(\mathbf{x}^*)^T(X^TX)^{-1}\mathbf{x}^*}. \tag{98}$$

3.13 Prediction Intervals

1. Given a set of inputs, a 95% confidence interval for the response gives an interval that contains the *expected* response 95% of the time.
2. In contrast, given a set of inputs, a 95% *prediction interval* produces an interval in which we are 95% sure that *any given response with those inputs* lie in.

Because a single observation is more variable than the expected response, a prediction interval is wider than the corresponding confidence interval.

Suppose we have inputs $\mathbf{x}^* = \begin{bmatrix} 1 & x_1^* & x_2^* & \ldots & x_k^* \end{bmatrix}^T$, with corresponding response

$$y^* = (\mathbf{x}^*)^T\beta + \epsilon^* \tag{99}$$

where $\text{var}(\epsilon^*) = \sigma^2$ by assumption.

This will be point estimated by $(\mathbf{x}^*)^T\mathbf{b}$ with an error of

$$y^* - (\mathbf{x}^*)^T\mathbf{b} = (\mathbf{x}^*)^T\beta + \epsilon^* - (\mathbf{x}^*)^T\mathbf{b}, \tag{100}$$

where the only random terms are ϵ^* and $(\mathbf{x}^*)^T\mathbf{b}$.

1. ϵ^* is an error associated with the future observation y^*.
2. $\mathbf{b}$ depends only on the current observations of $\mathbf{y}$.

We can say the are independent, giving us You said the variance was 0 below here, fixed that up. I think you intended to say "variance is this" and since the estimator is unbiased, expectation is 0.

$$\begin{aligned}
\text{var}(y^* - (\mathbf{x}^*)^T\mathbf{b}) &= \text{var}(\epsilon^*) + \text{var}((\mathbf{x}^*)^T\mathbf{b}) \\
&= \sigma^2 + (\mathbf{x}^*)^T(X^TX)^T{-}1\sigma^2\mathbf{x}^* \\
&= [1 + (\mathbf{x}^*)^T(X^TX)^{-1}\mathbf{x}^*]\sigma^2
\end{aligned}$$

This can be used to derive that

$$\frac{y^* - (\mathbf{x}^*)^T\mathbf{b} - 0}{s\sqrt{[1 + (\mathbf{x}^*)^T(X^TX)^{-1}\mathbf{x}^*]}} \sim t_{n-p} \tag{101}$$

Thus a prediction interval for y^* is

$$(\mathbf{x}^*)^T\mathbf{b} \pm t_{\alpha/2}s\sqrt{[1 + (\mathbf{x}^*)^T(X^TX)^{-1}\mathbf{x}^*]}. \tag{102}$$

The only difference with a corresponding confidence interval is the presence of the '1', which makes the interval wider as expected.

3.14 Deriving CIs and PIs

Let the mean response be $x = x^*$, then we have

$$\begin{aligned} var(\hat{y}) &= var(\beta_0 + \beta_1 x^*) \\ &= var(\beta_0) + var(\beta_1 x^*) + 2cov(\beta_0, \beta_1 x^*) \\ &= var(\beta_0) + (x^*)^2 var(\beta_1) + 2x^* cov(\beta_0, \beta_1) \\ &= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}\right) + \sigma^2\left(\frac{(x^*)^2}{SS_{xx}}\right) - \frac{2x^*\bar{x}}{SS_{xx}} \\ &= \sigma^2\left(\frac{1}{n} + [\frac{\bar{x}^2 + (x^*)^2 - 2x^*\bar{x}}{SS_{xx}}]\right) \\ &= \sigma^2\left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{SS_{xx}}\right) \end{aligned}$$

$$\begin{aligned} var(\hat{y}) &= \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{SS_{xx}}\right) \\ &= \sigma^2\left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SS_{xx}}\right) \end{aligned}$$

3.15 Joint Confidence Intervals (Confidence Region)

Finding confidence intervals individually **for each parameter** is misleading. If we find more than one 95% interval, we **DO NOT** have 95% confidence that **all of them will be satisfied at once**. The more confidence intervals we have, the more likely it is that at least one will be wrong. You will need to find a ***joint confidence region*** for a number of parameters at the same time.

Definition 4.16:
Let $\chi^2_{\gamma_1}, \chi^2_{\gamma_2}$ be independent χ^2 r.v with γ_1, γ_2 d.o.f. Then

$$\frac{\chi^2_{\gamma_1}/\gamma_1}{\chi^2_{\gamma_2}/\gamma_2} \sim F_{\gamma_1,\gamma_2} \tag{103}$$

Once again, the least squares estimator for β is

$$\mathbf{b} = (X^T X)^{-1} X^T \mathbf{y} \sim MVN(\beta, (X^T X^{-1}\sigma^2). \tag{104}$$

From Corollary 3.10, the quadratic form

$$\frac{(\mathbf{b} - \beta)^T X^T X (\mathbf{b} - \beta)}{\sigma^2} \sim \chi^2_p \tag{105}$$

where p is the number of parameters (design variables) in the model.

We also know that

$$\frac{(n-p)s^2}{\sigma^2} \sim \chi^2_{n-p} \tag{106}$$

Since $\mathbf{b}$ and s^2 are independent, the two χ^2 variables above are independent, which means that

$$\left(\frac{(\mathbf{b}-\beta)^T X^T X(\mathbf{b}-\beta)}{p\sigma^2}\right) / \left(\frac{(n-p)s^2}{(n-p)\sigma^2}\right) = \frac{(\mathbf{b}-\beta)^T X^T X(\mathbf{b}-\beta)}{ps^2} \sim F_{p,n-p} \tag{107}$$

IMPORTANT NOTE!!!
Because the test statistic F is based on $\mathbf{b}-\beta$ which we want to minimise, we use a **right-hand tail** of the F distribution to create a **one sided confidence region**.

$$\mathbf{b}-\beta \sim MVN(\beta-\beta, (X^TX)^{-1}\sigma^2) \Rightarrow MVN(\mathbf{0}, (X^TX)^{-1}\sigma^2) \tag{108}$$

Let f_α be the critical value of the F distribution with $p.n-p$ d.o.f, and probability α. Then

$$Pr[(\mathbf{b}-\beta)^T X^T X(\mathbf{b}-\beta)/ps^2 \leq f_\alpha] = 1-\alpha. \tag{109}$$

Since X^TX is always positive definite, we get the confidence region

$$(\mathbf{b}-\beta)^T X^T X(\mathbf{b}-\beta) \leq ps^2 f_\alpha. \tag{110}$$

Note: The region has the form of an ellipse or ellipsoid.

3.16 Generalised Least Squares

Previously, we made the assumptions that **the errors ϵ have mean 0 and variance $\sigma^2 I$** - However, they do not always hold.

If the errors **DO NOT** have **0** mean, we should find another model.

Suppose that ϵ is multivariate normal but with a positive definite variance V. The maximum likelihood estimator now minimises

$$\mathbf{e}^T V^{-1}\mathbf{e} = (\mathbf{y}-X\mathbf{b})^T V^{-1}(\mathbf{y}-X\mathbf{b}) \tag{111}$$

and thus satisfies the **normal equations**

$$X^T V^{-1} X\mathbf{b} = X^T V^{-1}\mathbf{y}. \tag{112}$$

This gives the *generalised least squares estimators*

$$\mathbf{b} = (X^T V^{-1} X)^{-1} X^T V^{-1}\mathbf{y}. \tag{113}$$

If we let $V = \sigma^2 I$, then this reduces to the **ordinary least squares**.

We can prove that the following holds (the generalised least squares estimator is BLUE):

$$\begin{aligned} E[\mathbf{b}] &= \beta \\ \text{var}(\mathbf{b}) &= (X^T V^{-1} X)^{-1}. \end{aligned}$$

In this situation, the errors are **uncorrelated but do not have a common variance**:

$$(\mathbf{y} - X\mathbf{b})^T V^{-1} (\mathbf{y} - X\mathbf{b}) = \sum_{i=1}^{n} \left(\frac{e_i}{\sigma_i}\right)^2. \tag{114}$$

In other words, we *weight* each residual by **the inverse of the corresponding standard deviation**. So a point with high variance influences **b** less than a point with low variance.

3.17 Transformations

Certain kinds of relationships (in particular **multiplicative** relationships) also require the transformation of the *response* variable.

For example, if the true underlying model is

$$y = \alpha_1 e^{\alpha_2 x} \epsilon, \tag{115}$$

then we would transform the response variable to $ln(y)$:

$$ln(y) = ln(\alpha_1) + \alpha_2 x + ln(\epsilon). \tag{116}$$

We can then fit a linear model to $ln(y)$ with design variable x and recover the original coefficients with

$$\alpha_1 = e^{\beta_0} \qquad\qquad \alpha_2 = \beta_1$$

Here are some certain signs which may indicate that **a transformation is required**:

1. All the values are positive
2. The distribution of the data is skewed
3. There is an obvious non-linear relationship with another variable
4. The variances show a relationship with one of the variables

Logarithmic transformations are very common because they convert multiplicative effects into additive ones. Useful transformations are:

1. $ln(y), x$	exponential
2. $ln(y), ln(x)$	power law
3. $\sqrt{y}$	areas, or occurences inside areas
4. $\sqrt[3]{y}$	volumes
5. $\frac{1}{y}$	rates
6. $ln\left(\frac{y}{1-y}\right)$	proportions

4 Inference for the Full Rank Model

To recap, the full rank model is

$$\mathbf{y} = X\beta + \epsilon \tag{117}$$

where X is $n \times p$, $n \geq p$, $r(X) = p$, and the errors $\epsilon \sim N(\mathbf{0}, \sigma^2 I)$ (for some theorems MVN).

We now want to test *model relevance* - does our model contribute anything at all?

If none of the x variables have any relevance for predicting y, then all the parameters β will be $\mathbf{0}$. We test for this using the **null hypothesis**

$$H_0 : \beta = \mathbf{0}. \tag{118}$$

Alternatively, if *at least some* of the x variables are relevant to predicting y, then the corresponding parameters will be nonzero. So our alternative hypothesis is

$$H_1 : \beta \neq \mathbf{0}. \tag{119}$$

To test these hypothesis, we once again assume that the errors $\epsilon \sim MVN(\mathbf{0}, \sigma^2 I)$.

4.1 ANOVA

1. If $\beta = \mathbf{0}$, then $\mathbf{y} = \epsilon$ consists entirely of errors. In this case, $\mathbf{y}^T\mathbf{y}$, the SS_{Res} measure the variability of the errors.
2. However, if $\beta \neq \mathbf{0}$, then $\mathbf{y} = X\beta + \epsilon$. In this case, some of $\mathbf{y}^T\mathbf{y}$ will come from errors, but some will come from the model prediction.

By separating $\mathbf{y}^T\mathbf{y}$ into these two parts, we can compare them to see how well the model is doing.

The sum of squares of the residual SS_{Res} is

$$\begin{aligned} SS_{Res} &= (\mathbf{y} - X\mathbf{b})^T(\mathbf{y} - X\mathbf{b}) \\ &= (\mathbf{y} - H\mathbf{y})^T(\mathbf{y} - H\mathbf{y}), \text{ Hatton Matrix transform, } H \text{ is idempotent} \\ &= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T H\mathbf{y} + \mathbf{y}^T H^2\mathbf{y} \\ &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T H\mathbf{y} \\ &= \mathbf{y}^T H\mathbf{y} - \mathbf{y}^T X(X^T X)^{-1}X^T\mathbf{y} \end{aligned}$$

which means that

$$\mathbf{y}^T\mathbf{y} = \mathbf{y}^T X(X^T X)^{-1}X^T\mathbf{y} + SS_{Res}. \tag{120}$$

We call $\mathbf{y}^T X(X^T X)^{-1}X^T\mathbf{y} = \hat{\mathbf{y}}^T\hat{\mathbf{y}} = \mathbf{b}^T X^T X\mathbf{b}$ the **regression sum of squares** and will now *denote it as* SS_{Reg}. This reflects the variation in the response variable y that is explained by the model.

We call the **total variation in the response variable** $SS_{Total} = \mathbf{y}^T\mathbf{y}$, which can be divided into

$$SS_{Total} = SS_{Reg} + SS_{Res}. \tag{121}$$

Example:(Two extremes of the spectrum)

Suppose that there is no error, so that $\mathbf{y} = X\beta$. We have

$$\begin{aligned} SS_{Reg} &= \mathbf{y}^T X(X^TX)^{-1}X^T\mathbf{y} \\ &= \beta^T X^T X(X^TX)^{-1}X^TX\beta \\ &= \beta^T X^T X\beta \\ &= \mathbf{y}^T\mathbf{y} = SS_{Total} \end{aligned}$$

and $SS_{Res} = 0$.

On the contrary, suppose that there is **no signal**, so that $\beta = \mathbf{0}$ and $\mathbf{y} = \epsilon$. If we let the least squares estimator $\mathbf{b} = \beta = \mathbf{0}$ then

$$\begin{aligned} SS_{Res} &= (\mathbf{y} - X\mathbf{b})^T(\mathbf{y} - X\mathbf{b}) \\ &= \mathbf{y}^T\mathbf{y} = SS_{Total} \end{aligned}$$

and $SS_{Reg} = 0$.

Theorem 5.1 / 4.13:
In the full rank general linear model $\mathbf{y} = X\beta + \epsilon$,

$$SS_{Res}/\sigma^2 \sim \chi^2_{n-p} \tag{122}$$

Theorem 5.2:
In the full rank general linear model $\mathbf{y} = X\beta + \epsilon$,

$$SS_{Reg}/\sigma^2 \sim \chi^2_{p,\lambda} \tag{123}$$

where

$$\lambda = \frac{1}{2\sigma^2}\beta^T X^T X\beta. \tag{124}$$

Proof:
By definition

$$\frac{SS_{Reg}}{\sigma^2} = \frac{1}{\sigma^2}\mathbf{y}^T H\mathbf{y}. \tag{125}$$

By assumption, $\mathbf{y} \sim MVN(X\beta, \sigma^2 I)$, and $H = X(X^TX)^{-1}X^T$ is an idempotent and symmetric matrix. Therefore, we know that its rank is equal to its trace

$$\begin{aligned} r(H) &= tr(H) \\ &= tr((X^TX)^{-1}X^TX) \\ &= tr(I_p) \\ &= p \end{aligned}$$

By **Theorem 3.5, 3.6, 3.7**, SS_{Reg}/σ^2 has a non-central χ^2 distribution with $k+1$ d.o.f and non-centrality parameter

$$\begin{aligned} \lambda &= \frac{1}{2\sigma^2}(X\beta)^T X(X^TX)^{-1}X^T(X\beta) \\ &= \frac{1}{2\sigma^2}\beta^T X^T X\beta. \end{aligned}$$

Theorem 5.3 (EXAM):
In the full rank general linear model $\mathbf{y} = X\beta + \epsilon$, SS_{Res} and SS_{Reg} are independent.
We can write $SS_{Reg} = \mathbf{b}^T X^T X\mathbf{b}$ and observe that $\mathbf{b}$ and s^2 are independent (which SS_{Res} depends on).

Now, to test $H_0 : \beta = 0$, we observe that the non-centrality parameter λ for SS_{Reg}/σ^2 must be 0. Thus, under H_0,

$$\frac{SS_{Reg}/p\sigma^2}{SS_{Res}/(n-p)\sigma^2} = \frac{SS_{Reg}/p}{SS_{Res}/(n-p)} = \frac{MS_{Reg}}{MS_{Res}} \sim F_{p,n-p} \tag{126}$$

If H_0 was not true, the expected value of MS_{Reg} is

$$E\left[\frac{SS_{Reg}}{p}\right] = E\left[\frac{\mathbf{y}^T H\mathbf{y}}{p}\right] = tr(HV) + \boldsymbol{\mu}^H\boldsymbol{\mu} = \sigma^2 + \frac{1}{2}\beta^T X^T X\beta. \tag{127}$$

The expected value of the denominator MS_{Res} is

$$E\left[\frac{SS_{Res}}{n-p}\right] = E[s^2] = \sigma^2. \tag{128}$$

So if $\beta = 0, E\left[\frac{SS_{Reg}}{p}\right] = \sigma^2$ and the statistic should be close to 1. However, if $\beta \neq 0$, we end up with $E\left[\frac{SS_{Reg}}{p}\right] > \sigma^2$ since X^TX is positive definite.

Therefore, we should use a **one-tailed test and reject H_0 if the statistic is large**.

4.2 The General Linear Hypothesis

The *general linear hypothesis* is the structure of all possible linear hypothesis, given by

$$H_0 : C\beta = \boldsymbol{\delta}^* \qquad\qquad H_1 : C\beta \neq \boldsymbol{\delta}^*$$

where C is an $r \times p$ matrix with rank $r \leq p$, and $\boldsymbol{\delta}^*$ is an $r \times 1$ vector of constants.

Example:
Consider the null hypothesis of model relevance, $H_0 : \beta = 0$.

We can express this in the form of the general linear hypothesis by letting $C = I_p$ (which has rank p) and $\boldsymbol{\delta}^* = 0$.

Example:
Now consider a regression model with 4 parameters (3 predictors with 1 intercept parameter)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i. \tag{129}$$

Let

$$C = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \qquad \boldsymbol{\delta}^* = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{130}$$

If we test $H_0 : C\beta = \boldsymbol{\delta}^*$, (each row of C is a particular hypothesis), this is equivalent to

$$\begin{gathered} \beta_1 - \beta_2 = 0 \\ \beta_2 - \beta_3 = 0 \\ \therefore \beta_1 = \beta_2 = \beta_3. \end{gathered}$$

To develop a test statistic, we start with $C\mathbf{b} - \boldsymbol{\delta}^*$, the least squares estimator for $C\beta - \boldsymbol{\delta}^*$.

We have

$$E(C\mathbf{b} - \boldsymbol{\delta}^*) = C\beta - \boldsymbol{\delta}^*$$
$$\text{var}(C\mathbf{b} - \boldsymbol{\delta}^*) = C(X^TX)^{-1}C^T\sigma^2.$$

From **Corollary 3.10** (using $V^{-1} = C(X^TX)^{-1}C^T\sigma^2$), the quadratic form

$$\frac{(C\mathbf{b} - \boldsymbol{\delta}^*)^T[C(X^TX)^{-1}C^T]^{-1}(C\mathbf{b} - \boldsymbol{\delta}^*)}{\sigma^2} \sim \chi^2_{r,\lambda} \tag{131}$$

where

$$\lambda = \frac{1}{2\sigma^2}(C\mathbf{b} - \boldsymbol{\delta}^*)^T[C(X^TX)^{-1}C^T]^{-1}(C\mathbf{b} - \boldsymbol{\delta}^*). \tag{132}$$

(We know $C(X^TX)^{-1}C^T$ has an inverse since it is a consequence of full rank models.)

Under the null hypothesis, then $C\beta = \boldsymbol{\delta}^*$, and the statistic will be

$$\frac{(C\mathbf{b} - \boldsymbol{\delta}^*)^T[C(X^TX)^{-1}C^T]^{-1}(C\mathbf{b} - \boldsymbol{\delta}^*)/r}{SS_{Res}/(n-p)} \sim F_{r,n-p} \tag{133}$$

(You can think of r as the *number of hypotheses*.)

Now, we check the expected value of the numerator to justify a one-tailed test.

$$E\left[\frac{(C\mathbf{b} - \boldsymbol{\delta}^*)^T[C(X^TX)^{-1}C^T]^{-1}(C\mathbf{b} - \boldsymbol{\delta}^*)}{r}\right] = \sigma^2 + \frac{1}{r}(C\mathbf{b} - \boldsymbol{\delta}^*)^T[C(X^TX)^{-1}C^T]^{-1}(C\mathbf{b} - \boldsymbol{\delta}^*), \tag{134}$$

where $C(X^TX)^{-1}C^T$ is positive definite. Hence, we can reject H_0 when the statistic is large.

4.3 Splitting β (Testing if part of β is 0)

If we find that the alternative hypothesis $H_1 : \beta \neq 0$ holds true, we cannot say which β_i are nonzero - only that **at least one is not**. If a particular β_i is zero, then it is best to remove it from the model.

This is because it will otherwise only fit noise (**over-fitting**) and reduce the ability of the model to predict.

We split the parameter vector

$$\beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{r-1} \\ \hline \beta_r \\ \vdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} \gamma_1 \\ \hline \gamma_2 \end{bmatrix} \tag{135}$$

and test the hypotheses

$$H_0 : \gamma_1 = 0 \qquad\qquad H_1 : \gamma_1 \neq 0.$$

Note we are comparing the two models: in H_1, the **full model**

$$\mathbf{y} = X\beta + \epsilon, \tag{136}$$

and in H_0, the **reduced model**

$$\mathbf{y} = X_2\gamma_2 + \epsilon_2 \tag{137}$$

where X_2 contains the last $p - r$ columns of $X = \big[X_1 | X_2\big]$.

Now, we can do this in the framework of the general linear hypothesis.

Let $C = \big[I_r | 0\big]$ and $\boldsymbol{\delta}^* = 0$. THen $C\beta = \boldsymbol{\delta}^*$ iff $\gamma_1 = 0$.

We define the **regression sum of squares** (SS_{Reg}) for γ_1 in the presence of γ_2 as

$$\begin{aligned} R(\gamma_1|\gamma_2) &= (C\mathbf{b} - \boldsymbol{\delta}^*)^T [C(X^TX)^{-1}C^T]^{-1}(C\mathbf{b} - \boldsymbol{\delta}^*) \\ &= \hat{\gamma_1}^T A_{11}^{-1} \hat{\gamma_1}, \end{aligned}$$

where $\hat{\gamma_1}$ is the least squares estimator for γ_1, and A_{11} is the $r \times r$ principal minor of $(X^TX)^{-1}$.

Our test statistic is

$$\frac{R(\gamma_1|\gamma_2)/r}{SS_{Res}/(n-p)} \sim F_{r,n-p} \tag{138}$$

and under the null hypothesis $H_0 : \gamma_1 = 0$, so we reject the null when the statistic is too large.

Theorem 5.4:

$$\begin{aligned} R(\gamma_1|\gamma_2) &= R(\beta) - R(\gamma_2) \\ &= \mathbf{b}^T X^T X \mathbf{b} - \mathbf{b_2}^T X_2^T X_2 \mathbf{b_2}, \end{aligned}$$

where $R(\beta)$ is the **regression sum of squares** for the **full model**

$$\mathbf{y} = X\beta + \epsilon = \left[X_1 | X_2\right] \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix} + \epsilon, \tag{139}$$

and $R(\gamma_2)$ is the **regression sum of squares** for the **reduced model**

$$\mathbf{y} = X_2\gamma_2 + \epsilon. \tag{140}$$

We will content ourselves with showing that

$$\begin{aligned} E[\hat{\gamma_1}^T A_{11}^{-1} \hat{\gamma_1}] &= E[R(\beta) - R(\gamma_2)] \\ &= E[\mathbf{y}^T [X(X^T X)^{-1} X^T - X_2 (X_2^T X_2)^{-1} X_2^T] \mathbf{y}]. \end{aligned}$$

Lemma 5.5:
Suppose that

$$A = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right], A^{-1} = B = \left[\begin{array}{c|c} B_{11} & B_{12} \\ \hline B_{21} & B_{22} \end{array}\right], \tag{141}$$

and B_{22}^{-1} exists. Then

$$A_{11}^{-1} = B_{11} - B_{12} B_{22}^{-1} B_{21}. \tag{142}$$

Refer to slides 56 - 60 for an example.

4.4 Corrected Sum of Squares

In general, our ANOVA table tests $H_0 : \beta_1 = \cdots = \beta_k = 0$ versus the alternative that some $\beta_i = 0, i \in \{1, \ldots, k\}$.

NOTE: The below is only for the test $\gamma_1 = 0$ in the presence of $\gamma_2 = \beta_0$ only. Do not use the below formulations if more than just the intercept is being 'ignored'.

Source of variation	Sum of squares	degrees of freedom	Mean square	F ratio
Regression				
Full model	$R(\boldsymbol{\beta}) = \mathbf{y}^T H \mathbf{y}$	$k+1$		
Reduced model	$\left(\sum_{i=1}^n y_i\right)^2 / n$	1		
γ_1 in presence of γ_2	$R(\gamma_1 \vert \gamma_2)$	k	$\frac{R(\gamma_1 \vert \gamma_2)}{k}$	$\frac{R(\gamma_1 \vert \gamma_2)/k}{MS_{Res}}$
Residual	$\mathbf{y}^T\mathbf{y} - R(\boldsymbol{\beta})$	$n-k-1$	$\frac{SS_{Res}}{n-p}$	
Total	$\mathbf{y}^T\mathbf{y}$	n		

The SS_{Reg} for the **reduced model** comes from

$$\mathbf{y}^T\mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}\mathbf{y} = \left(\sum_{i=1}^n y_i\right)\frac{1}{n}\left(\sum_{i=1}^n y_i\right) = \left(\sum_{i=1}^n y_i\right)^2 / n. \tag{143}$$

(Reduced model is $\mathbf{y} = \beta_0 + \epsilon$ so we only have columns of $\mathbf{1}$'s.)

This ANOVA table is sometimes presented differently. Observe that

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n} = \mathbf{y}^T\mathbf{y} - R(\gamma_2). \tag{144}$$

This is called the **corrected sum of squares**, and $R(\gamma_2)$ the **correction factor**.

We break down the corrected sum of squares into $R(\gamma_1|\gamma_2)$ and SS_{Res}, and test using an F statistic ratio. The end result is the same as before, but the table looks slightly different.

Source of variation	Sum of squares	degrees of freedom	Mean square	F ratio
Regression	$SS_{Reg} - \left(\sum_{i=1}^n y_i\right)^2 / n$	k	$\frac{R(\gamma_1 \vert \gamma_2)}{k}$	$\frac{R(\gamma_1 \vert \gamma_2)/k}{MS_{Res}}$
Residual	SS_{Res}	$n-k-1$	$\frac{SS_{Res}}{n-k-1}$	
Total	$\mathbf{y}^T\mathbf{y} - \left(\sum_{i=1}^n y_i\right)^2 / n$	$n-1$		

4.5 Sequential Testing

Consider the scenario when we have a number of explanatory variables in a model, but it is not obvious if all of them are relevant. We could fit a model using all of them, but we would run into the risk of **overfitting** (using irrelevant variables to explain noise by coincidence).

Ideally, we prefer to fit a **parsimonious model**, a model with minimal number of variables. This is because a parsimonious model is *less likely* to suffer from overfitting.

Consider the series of models

$$\begin{aligned} y &= \beta_0 + \epsilon^{(0)} \\ y &= \beta_0 + \beta_1 x_1 + \epsilon^{(1)} \\ &\vdots \\ y &= \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon^{(k)}. \end{aligned}$$

We denote the corresponding X matrix by $X^{(j)}$, which are the first $j+1$ columns of X.

The regression sum of squares for each of these models is calculated in the usual way:

$$R(\beta_0, \beta_1, \ldots, \beta_j) = \mathbf{y}^T X^{(j)} ((X^{(j)})^T X^{(j)})^{-1} (X^{(j)})^T \mathbf{y}. \tag{145}$$

Note that these are full regression sum of squares (we are looking at the total variation explained by the model in the presence of no other parameters).

By taking the difference between the sum of squares, we get the extra variation explained as we add variables to the model one at a time:

$$\begin{aligned} R(\beta_1|\beta_0) &= R(\beta_0, \beta_1) - R(\beta_0) \\ R(\beta_2|\beta 1, \beta_0) &= R(\beta_0, \beta_1, \beta_2) - R(\beta_0, \beta_1) \\ &\vdots \\ R(\beta_k|\beta_0, \beta_1, \ldots, \beta_{k-1}) &= R(\beta) - R(\beta_0, \beta_1, \ldots, \beta_{k-1}). \end{aligned}$$

Theorem 5.6:
In the full rank general linear model $\mathbf{y} = X\beta + \epsilon$, assume $\epsilon \sim N(0, \sigma^2 I)$. Then

$$\frac{1}{\sigma^2}\mathbf{y}^T\mathbf{y} = \frac{1}{\sigma^2} SS_{Res} + \frac{1}{\sigma^2} R(\beta_0) + \frac{1}{\sigma^2} R(\beta_1|\beta_0) + \frac{1}{\sigma^2} R(\beta_2|\beta_0, \beta_1) + \cdots + \frac{1}{\sigma^2} R(\beta_k|\beta_0, \beta_1, \ldots, \beta_{k-1}) \tag{146}$$

and the quadratic forms on the right are all **independent** non-central χ^2 distributions. SS_{Res} will have $n-p$ d.o.f whilst the rest all have 1 d.o.f each.

Proof:
We require the following **Lemma**.

Lemma 5.7:
Let $A = \begin{bmatrix} A_1 | A_2 \end{bmatrix}$ be a matrix of **full rank**. Then the matrix

$$A(A^T A)^{-1} A^T - A_1 (A_1^T A_1)^{-1} A_1^T \tag{147}$$

is idempotent.

The sum follows from the definition. To prove the rest, we use **Theorem 3.14**.

Let $X^{(j)}$ be the first $j+1$ columns of X, and

$$H_j = X^{(j)}((X^{(j)})^T X^{(j)})^{-1}(X^{(j)})^T,$$
$$R_j = \mathbf{y}^T H \mathbf{y} = R(\beta_0, \ldots, \beta_j).$$

Then

$$R(\beta_j | \beta_0, \ldots, \beta_{j-1}) = R_j - R_{j-1} = \mathbf{y}^T (H_j - H_{j-1}) \mathbf{y}. \tag{148}$$

From **Lemma 5.7**, we know $H_j - H_{j-1}$ is idempotent. Furthermore, both H_0 and $I - H_p$ are idempotent.

This we have a set of idempotent matrices with sum I. Hence **Theorem 3.14** applies and the quadratic forms all have independent non-central χ^2 distributions.

To show the d.o.f, observe that the sum of the ranks is n and $r(I - H_p) = n - p$. Now there are p remaining terms with a total rank of p. Hence each term has rank 1.

Each sequential regression sum of squares has 1 d.o.f. Therefore under the hypothesis $\beta_j = 0$, the test statistic is

$$\frac{R(\beta_j | \beta_0, \ldots, \beta_{j-1})}{SS_{Res}/(n-p)} \sim F_{1,n-p} \tag{149}$$

Note that this is still not entirely satisfactory, because the result will depend heavily on the order of the parameters considered. Different orderings can result in different sets of parameters being included in the final model.

4.6 Selection

4.6.1 Forward Selection

Forward selection starts off with an empty model, and adds the variable which is found to be **most significant**. Significance is measured in relation to the current model, so all tests are conducted in the presence of already included parameters, but not the other parameters. When no variables are significant enough to add, we stop and take the current model as the final model.

1. Start with an empty model.
2. Calculate the F-values for the null hypothesis $H_0 : \beta_i = 0$, for all parameters **not** in the model, in the **presence** of parameters **already** in the model.
3. If none of the tests are significant (we **do not reject** any null hypothesis), then stop.
4. Otherwise, add the most significant parameter (the parameter with the **largest** F-value).
5. Return to step 2.

4.6.2 Backward Elimination

A conceptually similar method is *backward elimination*.

1. Start with the full model.
2. Calculate the F-values for the null hypothesis $H_0 : \beta_i = 0$, for all parameters in the model, in the **presence** of the other parameters in the model.
3. If all of the tests are significant (we **reject** the null hypothesis), then stop.
4. Otherwise, we remove the least significant parameter (the parameter with the **smallest** F-value).
5. Return to stop 2.

4.6.3 Stepwise Selection

Stepwise selection functions similarly to forward or backward selection, but with the possibility of either adding or eliminating a variable at each step. In order to assess the appropriateness of a model, we use a **goodness-of-fit** measure.

We give a procedure using a goodness-of-fit measure called ***Akaike's Information Criterion*** (AIC), but it is trivual to adjust for any other goodness-of-fit statistic.

1. Start with any model.
2. Compute the AIC of all models which either have one extra variable or one less variable than the current model.
3. If the AIC of all such models is more than the AIC of the current model, stop.
4. Otherwise, change to the model with the lowest AIC.
5. Return to step 2.

4.6.4 Akaike's Information Criterion

The AIC is given as

$$\begin{aligned} AIC &= -2\ln(L(\beta, \sigma^2)) + 2p \\ &= n\ln\left(\frac{SS_{Res}}{n}\right) + 2p + c. \end{aligned}$$

A variant of the Bayesian Information Gain is given as

$$\begin{aligned} BIC &= -2\ln(L(\beta, \sigma^2)) + p\ln(n) \\ &= n\ln\left(\frac{SS_{Res}}{n}\right) + p\ln(n) + c. \end{aligned}$$

4.6.5 t-test

We can also use a t-test for a partial test of one parameter. That is, to test $H_0 : \beta_i = 0$ against $H_1 : \beta_i \neq 0$ in the presence of all other parameters.

Recall our CI for β_i is

$$b_i \pm t_{\alpha/2} s\sqrt{c_{ii}}, \tag{150}$$

where c_{ii} is the (i,i)th entry of the covariance matrix $c = (X^T X)^{-1}$.

If this CI **includes** 0, we fail to reject H_0. Otherwise, we can reject it.

It should be noted that t^2_{n-p} is the equivalent of $F_{1,n-p}$.

4.7 Shrinkage

For *ridge regression*, the SS_{Res} includes a parameter which **penalizes** the size of the parameters. We choose $\mathbf{b}$ to minimise

$$(\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta) + \lambda \mathbf{b}^T\mathbf{b}. \tag{151}$$

The λ parameter controls the amount of *shrinkage* of the parameters. The penalized least squares estimators can be calculated to be

$$\mathbf{b} = (X^T X + \lambda I)^{-1} X^T \mathbf{y}. \tag{152}$$

5 The Less Than Full Rank Model

For Full Rank Models, we assumed the design matrix X to be of full rank. This is *important* since a full rank X implies that $X^T X$ is invertible, and therefore the normal equations

$$X^T X\mathbf{b} = X^T\mathbf{y} \tag{153}$$

have a **unique** solution.

For Less Than Full Rank Models, samples come from k distinct populations, and we wish to determine the differences between these populations (*Think nominal / ordinal attributes*).

5.1 One-way Classification Model

Let y_{ij} be the jth sample taken from the ith population. Then the mode we use is given as

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \tag{154}$$

for $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n_i$, k is the number of populations / treatments, and n_i is the number of samples from the ith population.

The matrix representation is

$$\begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{21} \\ y_{22} \\ \vdots \\ y_{k,n_k} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \vdots \\ \tau_k \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{21} \\ \epsilon_{22} \\ \vdots \\ \epsilon_{k,n_k} \end{bmatrix}, \tag{155}$$

where μ is the overall mean, and the first column of X is the sum of the remaining columns (therefore not full rank).

Example:
Three different treatment methods for removing organic carbon from tar sand wastewater are compared: airflotation, foam separation, and ferric-chloride coagulation. A study is conducted and the amounts of carbon removed are:

AF	FS	FCC
34.6	38.8	26.7
35.1	39.0	26.7
35.3	40.1	27.0

The matrix representation of the linear model is

$$\begin{bmatrix} 34.6 \\ 35.1 \\ 35.3 \\ 38.8 \\ 39.0 \\ 40.1 \\ 26.7 \\ 26.7 \\ 27.0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{31} \\ \epsilon_{32} \\ \epsilon_{33} \end{bmatrix}. \tag{156}$$

The difficulty with a less than full rank model is that X^TX is singular. This means that the normal equations have an infinite number of solutions (we can choose μ to be any real number).

5.2 Reparametrization

One way to address the issue stated above is to convert the less than full rank model to a full rank model.

Example:
Consider the one-way classification model with $k = 3$, then the less than full rank model is

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \tag{157}$$

for $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, n_i$.

However, we can write the **mean** of each population as

$$\mu_i = \mu + \tau_i. \tag{158}$$

This allows us to recast the model as

$$y_{ij} = \mu_i + \epsilon_{ij}, \tag{159}$$

with corresponding matrices

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{bmatrix}, \beta = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}. \tag{160}$$

Since the columns of X are now linearly independent, X becomes a full rank model we can analyse. Therefore, the least squares estimates for each of the population means are the means of the samples drawn from that population:

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{i=1}^{n} y_{ij}. \tag{161}$$

The standard assumption that *the errors are normally distributed with mean* $\mathbf{0}$ *and variance* $\sigma^2 I$ is interpreted in this context to mean that ***all populations have a common pooled variance*** σ^2 ***but different means***. The standard estimator for this variance is

$$s^2 = \frac{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T X (X^T X)^{-1} X^T \mathbf{y}}{n - p} = \frac{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T X \mathbf{b}}{n - 3}. \tag{162}$$

This can be written as the *pooled variance*

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2}{n_1 + n_2 + n_3 - 3}. \tag{163}$$

5.2.1 Difference Between 2 Populations

Linear functions of the parameters, of the form $\mathbf{t}^T\beta$, are estimated using $\mathbf{t}^T\mathbf{b}$. **To find a difference between two populations** $\mu_1 - \mu_2$, we can estimate it by

$$\frac{1}{n_1}\sum_{i=1}^{n_1} y_{1j} - \frac{1}{n_2}\sum_{i=2}^{n_2} y_{2j}. \tag{164}$$

5.2.2 General Pooled Variance

The pooled variance for k sub-populations is given as

$$s^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{\sum_{i=1}^{k}(n_i - 1)}. \tag{165}$$

5.3 Conditional Inverses

Definition 6.1:
Let A be an $n \times p$ matrix. The $p \times n$ matrix A^c is called a **conditional inverse** for A iff

$$AA^cA = A. \tag{166}$$

Proof:

$$\begin{aligned} A^{-1}AA^cA &= A^{-1}A \\ A^cA &= I \\ A^c &= A^{-1}. \end{aligned}$$

However, the opposite does not hold:

$$A^cAA^c \neq A^c \tag{167}$$

If A is non-singular and square, then $A^{-1} = A^c$, so conditional inverses are an *extension* of regular inverses to non-square and singular matrices.

It should be noted that there are an *infinite* number of conditional inverses:

$$\begin{aligned} ABA &= ACA \\ A(\alpha\beta + (1-\alpha)C)A &= \alpha A + (1-\alpha)A \\ &= A. \end{aligned}$$

Theorem 6.2:
Let A be a $n \times p$ matrix. Then A has a conditional inverse. Moreover, conditional inverses can be constructed as follows:

1. Find a minor M of A which is **non-singular** and of dimension $r(A) \times r(A)$.
2. Replace M in A with $(M^{-1})^T$ and the other entries with zeros.
3. Transpose the resulting matrix.

5.3.1 Properties of Conditional Inverse

Let A be a $n \times p$ matrix of rank r, where $n \geq p \geq r$. Then

1. $r(A) = r(AA^c) = r(A^cA)$
2. $(A^c)^T = (A^T)^c$
3. $A^cA, AA^c, I - A^cA, I - AA^c$ are all idempotent
4. $A = A(A^TA)^c(A^TA)$ and $A^T = (A^TA)(A^TA)^cA^T$

We say that an expression involving a conditional inverse is *unique* if it is the same **no matter** what conditional inverse we use.

1. $A(A^TA)^cA^T$ is unique, symmetric, and idempotent
2. $r(A(A^TA)^cA^T) = r$
3. $I - A(A^TA)^cA^T$ is unique, symmetric, and idempotent
4. $r(I - A(A^TA)^cA^T) = n - r$

Proof:

$$\begin{aligned}[A(A^TA)^cA^T]^T &= A[(A^TA)^c]^TA^T \\ &= A[(A^TA)^T]^cA^T \\ &= A(A^TA)^cA^T.\end{aligned}$$

$$\begin{aligned}A(A^TA)^cA^TA(A^TA)^cA^T &= [A(A^TA)^cA^TA](A^TA)^cA^T \\ &= A(A^TA)^cA^T.\end{aligned}$$

To show that $A = A(A^TA)^cA^TA$, we can use the property that if $A^TA = 0 \Rightarrow A = 0$.

Let $M = A - A(A^TA)^cA^TA$,

$$\begin{aligned}M^TM &= [A - A(A^TA)^cA^TA]^T[A - A(A^TA)^cA^TA] \\ &= A^TA - A^TA(A^TA)^cA^TA - A^TA[(A^TA)^c]^TA^TA + A^TA[(A^TA)^c]^T(A^TA)^c(A^TA)^TA^TA \\ &= A^TA - A^TA - A^TA + (A^TA)[(A^TA)^c]^T(A^TA)^c(A^TA)^T(A^TA) \\ &= A^TA - A^TA - A^TA + (A^TA)(A^TA)^c(A^TA)^c(A^TA)(A^TA) \\ &= A^TA - A^TA - A^TA + A^TA \\ &= 0.\end{aligned}$$

Therefore, $M = A - A(A^TA)^cA^TA = 0$, as required.

5.4 Solving The Normal Equations

Theorem 6.3:
The system $A\mathbf{x} = \mathbf{g}$ is consistent iff the rank of $\begin{bmatrix} A & | & \mathbf{g} \end{bmatrix}$ is equal to the rank of A.

Theorem 6.4:
The general linear model $y = X\beta + \epsilon$, the normal equations

$$X^TX\mathbf{b} = X^T\mathbf{y} \tag{168}$$

are consistent.

Theorem 6.5:
Let $A\mathbf{x} = \mathbf{g}$ be a consistent system. Then $A^c\mathbf{g}$ is a solution to the system, where A^c is any conditional inverse for A.

From this theorem, we see that

$$\mathbf{b} = (X^T X)^c X^T \mathbf{y} \tag{169}$$

solves the normal equations, for any conditional inverse.

However, in the less than full rank model, different conditional inverses may result in different solutions.

Theorem 6.6:
Let $A\mathbf{x} = \mathbf{g}$ be a consistent system. Then

$$\mathbf{x} = A^c \mathbf{g} + (I - A^c A)\mathbf{z} \tag{170}$$

solves the system, where $\mathbf{z}$ is an arbitrary $p \times 1$ vector.

Thus, for the normal equations, any vector of the form

$$\mathbf{b} = (X^T X)^c X^T \mathbf{y} + [I - (X^T X)^c X^T X]\mathbf{z} \tag{171}$$

satisfies the equations.

Theorem 6.7:
Let $A\mathbf{x} = \mathbf{g}$ be a consistent system and let $\mathbf{x}_0$ be any solution to the system. Then for any A^c,

$$\mathbf{x}_0 = A^c \mathbf{g} + (I - A^c A)\mathbf{z} \tag{172}$$

where z=x_0.

5.5 Estimability

Now that we can find all solutions to the normal equations, which solution should we use?

Definition 6.8:
In the general linear model $\mathbf{y} = X\beta + \epsilon$, a function $\mathbf{t}^T\beta$ is said to be ***estimable*** if there exists a vector $\mathbf{c}$ such that $E[\mathbf{c}^T\mathbf{y}] = \mathbf{t}^T\beta$

In other words, a quantity is estimable if there is a linear unbiased estimator for it.

Theorem 6.9:
In the general linear model $\mathbf{y} = X\beta + \epsilon$, $\mathbf{t}^T\beta$ is estimable iff there is a solution to the linear system $X^T X\mathbf{z} = \mathbf{t}$.

Theorem 6.10:
In the general linear model $\mathbf{y} = X\beta + \epsilon$, $\mathbf{t}^T\beta$ is estimable iff

$$\mathbf{t}^T(X^TX)^cX^TX = \mathbf{t}^T, \tag{173}$$

for all conditional inverses of (X^TX).

In other words, using Theorem 6.9 and 6.10, suppose $\mathbf{t}^T\beta$ is estimable. By Theorem 6.9, there exists a solution to the system $X^TX\mathbf{z} = \mathbf{t}$.

This means that

$$X^TX(X^TX)^c\mathbf{t} = \mathbf{t}. \tag{174}$$

By taking the transposes, we see that this gives the required condition.

Example:
Consider the matrix

$$(X^TX)^c = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{175}$$

and consider the quantity $\beta_1 - \beta_2$ which corresponds to

$$\mathbf{t} = \begin{bmatrix} 0 & 1 & -1 \end{bmatrix}^T. \tag{176}$$

Then,

$$\begin{aligned} \mathbf{t}^T(X^TX)^c(X^TX) &= \begin{bmatrix} 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 & -1 \end{bmatrix} \\ &= \mathbf{t}^T. \end{aligned}$$

Hence we see that $\beta_1 - \beta_2$ is **estimable**.

Theorem 6.11:
In the general linear model $\mathbf{y} = X\beta + \epsilon$, suppose $\mathbf{t}^T\beta$ is estimable. Then the BLUE for $\mathbf{t}^T\beta$ is $\mathbf{z}^TX^T\mathbf{y}$, where $\mathbf{z}$ is a solution to the system $X^TX\mathbf{z} = \mathbf{t}$. Furthermore, this estimate is the *same* for any solution of the system, and can be **written as** $\mathbf{t}^T\mathbf{b}$, where $\mathbf{b}$ is *any* solution to the normal equations.
This is because $E[\mathbf{b}] = \beta$ **is not valid** for the less than full rank model!

5.5.1 One-Way Classification

For the one-way classification model with any number of levels,

$$\mu + \tau_i \tag{177}$$

is **ALWAYS** estimable.

Theorem 6.13:
If $\mathbf{z}$ is a linear combination of estimable functions, then $\mathbf{z}$ is estimable.

Of particular interest among **treatment contrasts** is the contrast of the form $\tau_i - \tau_j, ij$. This is because

$$\tau_i - \tau_j = (\mu + \tau_i) - (\mu + \tau_j) \tag{178}$$

is the difference between the *mean responses* in populations i and j.

Proof:
For the one-way classification for $k = 3$,

$$\begin{aligned}
\mathbf{t}^T &= \mathbf{t}^T (X^T X)^c X^T X \\
&= \mathbf{t}^T \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{n_1} & 0 & 0 \\ 0 & 0 & \frac{1}{n_2} & 0 \\ 0 & 0 & 0 & \frac{1}{n_3} \end{bmatrix} \begin{bmatrix} n & n_1 & n_2 & n_3 \\ n_1 & n_1 & 0 & 0 \\ n_2 & 0 & n_2 & 0 \\ n_3 & 0 & 0 & n_3 \end{bmatrix} \\
&= \mathbf{t}^T \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \\
\mathbf{t}^T \begin{bmatrix} -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} &= \mathbf{0} \\
\mathbf{t}^T &= \begin{bmatrix} 0 & 1 & 1 & 1 \end{bmatrix}.
\end{aligned}$$

Hence, only the treatment contrasts $(\mu + \tau_i)$ are estimable.

5.6 Estimating σ^2 in the Less Than Full Rank Model

In the full rank model, we estimated σ^2 using the sample variance

$$s^2 = \frac{SS_{Res}}{n - p}, \tag{179}$$

where n is the sample size, p is the number of parameters, and SS_{Res} is the Sum of Squares of the Residuals

$$SS_{Res} = (\mathbf{y} - X\mathbf{b})^T(\mathbf{y} - X\mathbf{b}) = \mathbf{y}^T[I - X(X^TX)^{-1}X^T]\mathbf{y} = \mathbf{y}^T H\mathbf{y}. \tag{180}$$

Theorem 6.14:
For the Less Than Full Rank Model,

$$SS_{Res} = \mathbf{y}^T[I - X(X^TX)^cX^T]\mathbf{y}. \tag{181}$$

Theorem 6.15:
In the general linear model $\mathbf{y} = X\beta + \epsilon$, suppose X has rank r, then an unbiased estimator for σ^2 is

$$\frac{SS_{Res}}{n-r} \tag{182}$$

5.7 Interval Estimation in the Less Than Full Rank Model

Theorem 6.16:
In the general linear model $\mathbf{y} = X\beta + \epsilon$, then

$$\frac{(n-r)s^2}{\sigma^2} = \frac{SS_{Res}}{\sigma^2} \sim \chi^2_{n-r} \tag{183}$$

Theorem 6.17:
In the general linear model $\mathbf{y} = X\beta + \epsilon$, if $\mathbf{t}^T\beta$ is estimable, then $\mathbf{t}^T\mathbf{b}$ is independent of s^2.

I think the $vart^T\beta$ below is meant to be $t^T\mathbf{b}$, which is a typo on YaoBan's part. $t^T\mathbf{b}$ is the rv whose variance we are interested in.

We have $\text{var}(\mathbf{t}^T\beta) = \sigma^2\mathbf{t}^T(X^TX)^c\mathbf{t}$. Thus a CI for the estimable quantity $\mathbf{t}^T\beta$ is

$$\mathbf{t}^T\mathbf{b} \pm t_{\alpha/2}s\sqrt{\mathbf{t}^T(X^TX)^c\mathbf{t}}. \tag{184}$$

needed bars for below Between two populations y_1, y_2, we can derive a CI from first principles

$$\bar{y_1} - \bar{y_2} \pm t_{\alpha/2}s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \tag{185}$$

6 Inference for the Less Than Full Rank Model

6.1 Testability

A testable hypothesis is a form of the General Linear Hypothesis $H_0 : C\beta = \mathbf{0}$ and is testable iff

$$C = C(X^TX)^cX^TX, \tag{186}$$

where $m = r(C), r = r(X)$, and C is $m \times p$.

Example:
Consider the null hypothesis that the means of all three populations are equal. This is the equivalent as $H_0 : \tau_1 = \tau_2 = \tau_3$.

This can be re-expressed in the General Linear Hypothesis $H_0 : C\beta = \mathbf{0}$, where

$$C = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \beta = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix}. \tag{187}$$

Since $(\mu + \tau_1) - (\mu + \tau_2) = \tau_1 - \tau_2$ is **treatment contrast**, it is estimable. Likewise, $\tau_2 - \tau_3$ is also estimable.

Since our matrix X^TX does not have an inverse, our estimator SS_{Res} relies on the conditional inverse of X^TX. Our new sample variance estimator is $\frac{SS_{Res}}{n-r}, r = r(X)$.

Formula below was missing an m. Hence, our proposed test statistic is

$$\frac{(C\mathbf{b})^T[C(X^TX)^cC^T]^{-1}C\mathbf{b}/m}{s^2} \sim F_{m,n-r} \tag{188}$$

Recall that $m = r(C)$.

Theorem 7.3:
In the general linear model $\mathbf{y} = X\beta + \epsilon$, assume $\epsilon \sim MVN(\mathbf{0}, \sigma^2 I)$. Suppose that $C\beta = \mathbf{0}$ is testable. Then $C\mathbf{b}$ and s^2 are independent. **Proof:** Using Theorem 3.13 ($BVA = 0$),

$$C\mathbf{b} = C(X^TX)^cX^T\mathbf{y}, SS_{Res} = \mathbf{y}^T[I - H]\mathbf{y}, \tag{189}$$

where $H_{hat} = X(X^TX)^cX^T$. Then $B = C(X^TX)^cX^T$, and $A = I - H$.

$$\begin{aligned} BVA &= C(X^TX)^cX^T\text{var}(\mathbf{y})[I - H] \\ &= [C(X^TX)^cX^T]\sigma^2 I[I - H] \\ &= [C(X^TX)^cX^T - C(X^TX)^cX^T]\sigma^2 \\ &= 0. \end{aligned}$$

6.2 Two-Factor Models

Theorem 7.4:
In an *additive* two-factor mode, every contrast in the τ's and β's is estimable.

The most common hypothesis that we wish to test are if all levels are equal to each other:

$$\tau_1 = \tau_2 = \cdots = \tau_a \tag{190}$$
$$\beta_1 = \beta_2 = \cdots = \beta_b \tag{191}$$

Because they are **all** composed of treatment contrasts **for one factor**, the are testable.

6.3 Interaction

Theorem 7.5:
For the linear model (with interaction)

$$y_{ijk} = \mu + \tau_i + \beta_j + \xi_{ij} + \epsilon_{ijk}, \tag{192}$$

there is **no interaction** iff

$$(\xi_{ij} - \xi_{ij'}) - (\xi_{i'j} - \xi_{i'j'}) = 0, \tag{193}$$

for all $i \neq i', j \neq j'$.
Moreover, **these quantities are all estimable**. ($\mu + \tau_i$, $\beta + \xi_i$)

This is because if there is no interaction between these levels, the difference in means that results from switching a factor from j to j' is the same regardless if the other factor is at level i or i'.

6.3.1 Considerations for Interaction

1. If we have one sample per combination of factors, it is *impossible* to account for, or test for interaction. This is because $n = r(X)$, and the resulting $n - r$ end up becoming 0.

   This means that we treat each combination of factors as a separate population. If we only have one sample from each population, then we have **no way** to estimate them,

2. If we test for interaction and find that there is none, we should theoretically **still use the residual sum of squares** from the **full model WITH interaction**, unless there is a convincing reason to think that there is no interaction.

   Think of it as "we cannot be sure that there is no interaction, but we just haven't found any".

   However, for practical purposes it is okay to use the SS_{Res} from an additive model (since it may take away too many d.o.f).

3. It is possible to have interaction between three or more factors. In practice, most people only look at two-factor interactions.

6.3.2 Examples of Interaction

When looking at plots, we want to see if:

- Are the lines parallel to each other (if it is additive, it should just be a constant factor separating the two lines)?
- If they are not parallel, change the ordering of the design variables.
- If the lines are still not parallel, then it is most likely interaction (non-constant distances between each line which means something is affecting it)

If we have interaction, we use the model

$$y_{ijk} = \mu + \tau_i + \beta_j + \xi_{ij} + \epsilon_{ijk}, \tag{194}$$

I disagree; I think it is estimable. Theorem 6.12...? **WHERE** $\mu + \tau_i + \beta_j + \xi_{ij}$ **IS NOT ESTIMABLE**.

6.4 ANCOVA

If we have one or more categorical predictors **AND** one or more continuous predictors in our model, then

$$y_{ij} = \mu + \tau_i + \beta x_{ij} + \xi_i x_{ij} + \epsilon_{ij}. \tag{195}$$

We can think of this linear model as fitting **several** regression lines - one for each population (**assuming equal variances across populations**).

Interaction in this case means that the slopes of the regression lines (effect of continuous predictor) are *different* for each population.

However, a model *without* interaction assumes that the slopes are the same (may have different intercepts), and follow this model:

$$y_{ij} = \mu + \tau_i + \beta x_{ij} + \epsilon_{ij}. \tag{196}$$

6.5 Testing for Interaction

Our null hypothesis will be

$$H_0 : \xi_1 = \xi_2 = \cdots = \xi_j \tag{197}$$

(i.e. equal slopes that may have different intercept values).

Our model without interaction will be the design matrix that includes μ, τ, β which means β becomes estimable.

To test the significance of a factor, we use the same null hypothesis as usual

$$H_0 : \tau_1 = \tau_2 = \cdots = \tau_i \tag{198}$$

or

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_i \tag{199}$$

7 Experimental Design

An **observational study** can not determine causality, only correlation. Instead, you need a **designed experiment/trial**. Even then, we can *never* be sure.

7.1 Hill's Criteria

- Strength of association
- Consistency of association (from one study to another)
- Consistent with existing knowledge
- * Monotonic response (increase A makes B more likely)
- * Temporal relationship (A must come before B)
- * Plausibility of alternatives (is there another explanation?)
- Predictive value of link (can you see it?)

7.2 Completely Randomised Design (CRD)

We use a completely randomised design for a single factor (treatment) with k levels. To analyse, we use a one-way classification model

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} = \mu_i + \epsilon_{ij}. \tag{200}$$

Theorem 8.1:
In a completely randomised design with n test units, the allocation of test units to factor levels which minimises

$$\sum_{i=1}^{k} \text{var}(\hat{\mu}_i) = \sigma^2 \sum_{i=1}^{k} \frac{1}{n_i} \tag{201}$$

is

$$n_i = \frac{n}{k}, \tag{202}$$

where n is a multiple of k (the number of factors).

7.3 Randomised Complete Block Design (CBD)

We can use a CBD when we have one factor of interest (the treatment) and one nuisance/confounding factor (the *blocking* factor).

Remember!!! k is the number of treatments.

We partition the experimental units int blocks of size k, which are homogeneous in the blocking factor. In each block, we apply one treatment to each experimental unit chosen

randomly. Since blocks which have a size which is an integer multiple of k also work, an equal number of units in each block will recieve each treatment.

TO analyse, we use a two-way classification model:

$$y_{ijk} = \mu + \beta_i + \tau_j + \epsilon_{ijk}, \tag{203}$$

where the β_i are the block effects and the τ_j are the treatment effects.

Theorem 8.2:
In the general linear model, write

$$\mathbf{y} = \left[X_1 | X_2\right] \begin{bmatrix}\beta_1 \\ \hline \beta_2\end{bmatrix} + \epsilon. \tag{204}$$

Then $\begin{bmatrix}\mathbf{b}_1 \\ \hline \mathbf{b}_2\end{bmatrix}$ is a solution to the normal equations iff $\mathbf{b}_2$ is a solution to the **reduced normal equations**

$$X_2^T[I - H_1]X_2\mathbf{b}_2 = X_2^T[I - H_1]\mathbf{y}, \tag{205}$$

where $H_1 = X_1(X_1^T X_1)^c X_1^T$.

7.3.1 Comparison of CBD vs CRD

If there is really no blocking effect, then the CRD is better than the CBD. However, if there is an effect due to the blocks then ignoring it means your estimators are less accurate than they could be.

DESIGN PRINCIPLE
Segment the study population into homogeneous groups to reduce variation.

Suppose we want to compare two drugs and a control ($k = 3$) using mice.

We could take 30 mice and assign 10 randomly to each treatment (CRD).

Alternatively, we could take 10 groups (blocks), each of three mice from the same family, then assign all 3 treatments to the mice in each group. Within each group, the treatmenbts are assignmed randomly (CBD). The genetic similarity of the blocks will reduce the error variance.

7.4 Latin Square Designs

Latin square designs are used when there are two confounding factors or more. This is because a CBD will grow exponentially in size (and therefore cost), whilst a Latin square uses

fewer experimental units. **REMEMBER!!! Latin squares must have the condition that the treatment and blocking factors to have the same number of levels t.**

Our solution will have: each treatment appearing exactly once in *each* row and *each* column.

Note that choosing a Latin Square at random helps avoid the effects of lurking variables - given a Latin square, a random permutation of the rows and columns gives another Latin square.

The model for a Latin Square design is an **additive three-way classification model**:

$$y_{ijk} = \mu + \beta_i + \gamma_j + \tau_k + \epsilon_{ijk}, \tag{206}$$

where $1 \leq i, j \leq t, k = k(i, j)$ where k is determined by i, j.

In this model, μ is the overall mean, β_i is the effect of the level i of the first confounding factor, γ_j is the effect of level j of the second confounding factor, and τ_k is the k-th treatment effect.

β, γ are the **nuisance terms** that correspond to X_1. τ are the treatment effects and correspond to X_2 (permutation matrix that are zero except for a single 1 for each row and column).
The sum of all permutation matrices P_i is J_t, the matrix of all ones.

7.5 Balanced Incomplete Block Designs (BIBD)

Suppose we have t treatment levels, and b blocks of size $k < t$.

This happens when there is a natural block size. Examples

- twins are pairs.
- car tyres are either 2 or 4.

Therefore, a design is called a BIBD iff

- Each treatment occurs at most **once** in a block.
- Each treatment occurs exactly $r = bk/t$ times (first order balance)
- Each pair of treatments occurs in the same number of blocks λ (second order balance).

In this case, there are $t(t-1)/2$ difference pairs of treatments and $bk(k-1)/2$ available slots, so we must have

$$\lambda = \frac{bk(k-1)}{t(t-1} = r\frac{k-1}{t-1}. \tag{207}$$

Given both t, k, we can always find a BIBD with $b = \binom{t}{k}$ blocks, by taking all possible subsets of size k.

In this case,

$$r = \binom{t}{k}\frac{k}{t} = \binom{t-1}{k-1}, \lambda = \binom{t-1}{k-1}\frac{k-1}{t-1} = \binom{t-2}{k-2}. \tag{208}$$