1 Linear Algebra Review

1.1 Partitioning of Matrices

$$X = \left[\begin{array}{cc|c} 1 & 2 & 3 \\ \hline 4 & 5 & 6 \end{array}\right] = \left[\begin{array}{c|c} X_{11} & X_{12} \\ \hline X_{21} & X_{22} \end{array}\right]$$

$$Y = \left[\begin{array}{cc} 1 & 2 \\ 3 & 4 \\ \hline 5 & 6 \end{array}\right] = \left[\begin{array}{c} Y_{11} \\ \hline Y_{12} \end{array}\right]$$

where

$$X_{11} = \left[\begin{array}{cc} 1 & 2 \end{array}\right]$$
$$X_{12} = \left[\begin{array}{c} 3 \end{array}\right]$$
$$X_{21} = \left[\begin{array}{cc} 4 & 5 \end{array}\right]$$
$$X_{22} = \left[\begin{array}{c} 6 \end{array}\right]$$
$$Y_{11} = \left[\begin{array}{cc} 1 & 2 \\ 3 & 4 \end{array}\right]$$
$$Y_{12} = \left[\begin{array}{cc} 5 & 6 \end{array}\right]$$

Then we can perform matrix multiplication for the submatrices as follows:

$$XY = \left[\begin{array}{c|c} X_{11} & X_{12} \\ \hline X_{21} & X_{22} \end{array}\right] \left[\begin{array}{c} Y_{11} \\ \hline Y_{12} \end{array}\right] = \left[\begin{array}{c} X_{11}Y_{11} + X_{12}Y_{21} \\ \hline X_{21}Y_{11} + X_{22}Y_{21} \end{array}\right]$$

1.2 Rules

Transpose

- $(X^T)^T = X$
- $(XY)^T = Y^T X^T \neq X^T Y^T$
- X is symmetric $\iff X^T = X$

Inverse

If X has an inverse (X^{-1} exists), X is nonsingular (invertible). Let X be nonsingular:

- for any square matrix X, $XX^{-1} = X^{-1}X = I$
- $(X^{-1})^{-1} = X$
- if X and Y are same size and nonsingular, XY is non singular
- $(XY)^{-1} = Y^{-1}X^{-1} \neq X^{-1}Y^{-1}$ (similar property with transpose)

** X is singular $\iff |X| = 0$ **

1.3 Orthogonal Vectors

Definition: $\boldsymbol{x}^T\boldsymbol{y} = \sum_{i=1}^{n} x_i y_i = 0 \iff x, y$ are orthogonal.

- square matrix X is orthogonal $\iff X^T X = I$
- X is orthogonal $\implies X^{-1} = X^T$
- $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ n \end{bmatrix} \implies \quad \boldsymbol{x}^T\boldsymbol{x} = ||\boldsymbol{x}||^2 = \sum_{i=1}^{n} x_i^2, \quad$ where $||\boldsymbol{x}||$ is the norm or length of $\boldsymbol{x}$
- a set of vectors $\{ \boldsymbol{x_1} \cdots \boldsymbol{x_k}\}$ is orthonormal
  $\iff$ each pair of vectors $(\boldsymbol{x_i}, \boldsymbol{x_j})$ is orthogonal, and $||\boldsymbol{x_i}|| = 1$ for all i.

1.4 Eigen

Let A be $n \times n$ matrix, $\boldsymbol{x}$ be $n \times 1$ nonzero vector, with λ being scalar.

$$A\boldsymbol{x} = \lambda\boldsymbol{x} \longleftrightarrow (A - \lambda I)\boldsymbol{x} = 0$$

The eigenvalues, λ is solved by putting

$$|A - \lambda I| = 0$$

The eigenvectors, $\boldsymbol{x}$ is solved by first finding λ and solving the system of equation $A\boldsymbol{x} = \lambda\boldsymbol{x}$.

- P is orthogonal matrix of the same size as $A \implies$ eigenvalues of $P^T AP =$ eigenvalues of A
- $A \in R$ and symmetric $\implies$ its eigenvalues $\lambda \in R$ and eigenvectors orthogonal

1.5 Diagonalisation

A is symmetric $\implies$ orthogonal matrix P exists such that

$$P^T AP = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 \cdots & 0 & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix} = D$$

where λ_i's are the eigenvalues of A.

- columns in P are eigenvectors of A with the respective eigenvalues, and are an orthonormal set
- P diagonalizes A

1.6 Linear Independence

With $\boldsymbol{x_i}$ as vectors,

$$\alpha_1\boldsymbol{x_1} + \alpha_2\boldsymbol{x_2} + \cdots + \alpha_k\boldsymbol{x_k} = 0$$

- The set of vectors are linearly independent if the only way to satisfy the equation is for all $\alpha_i = 0$.
- Else, the vectors are linearly dependent and at least one of the vectors can be written as a linear combination of another.

1.7 Rank

Definition: rank of $X, r(X)$ is the dimension of column space of X (the largest number of linearly independent vectors)

- $r(X) = r(X^T) = r(X^T X)$
- if X is $k \times k$, then X is nonsingular $\iff r(X) = k$
- if X is $n \times k$, P is $n \times n$, Q is $k \times k$,
  then $r(X) = r(PX) = r(XQ)$ (assuming P and Q are nonsingular)
- $r(XY) \leq r(X), r(Y)$
- rank of diagonal matrix = # nonzero diagonals

1.8 Idempotence

Definition: Square matrix A is idempotent $\iff A^2 = A$.

1.9 Trace

Definition: $tr(X) = \sum_{i=1}^{k} x_{ii}$ for a $k \times k$ matrix X.
(Add up the diagonal elements)

- c is a scalar $\implies tr(cX) = ctr(X)$
- $(X \pm Y) = tr(X) \pm tr(Y)$
- XY, YX exist $\implies tr(XY) = tr(YX)$

1.10 Important Theorems

1. Eigenvalues λ of idempotent matrix are always either 0 or 1.
   hint: use $A\boldsymbol{x} = \lambda\boldsymbol{x}$

2. If A is symmetric & idempotent, $r(A) = tr(A)$
   hint: diagonalize A, $r(P^TAP) = r(P^TA) = r(A) = tr(P^TAP) = tr(P^TPA) = tr(A)$

3. Let $A_1, A_2, \cdots A_m$ be a collection of symmetric matrices. Then these are equivalent:

   - $\exists$ orthogonal P, such that P^TA_iP is diagonal
   - $A_iA_j = A_jA_i$ for every pair of matrices

4. Let $A_1, A_2, \cdots A_m$ be a collection of symmetric matrices. Any 2 implies the third:

   - $\forall i, A_i$ are idempotent
   - $\sum_{i=1}^m A_i$ is idempotent
   - $A_iA_j = 0$ for $i \neq j$

5. if (4) are all true, then $r(\sum_{i=1}^m A_i) = \sum_{i=1}^m r(A_i)$

1.11 Quadratic Forms

A is $k \times k$, y is $k \times 1$ vector with variables. *Definition*: $q = y^TAy$ is a quadratic form in y.
A is the matrix of quadratic form.

- $y^TAy > 0$and $\forall y \neq 0 \Longrightarrow$ it is positive definite
- $y^TAy \geq 0$ instead, it is positive semi-definite
- symmetric matrix A is positive definite $\Longleftrightarrow \lambda$ (eigenvalues) > 0
- symmetric matrix A is positive semi-definite $\Longleftrightarrow \lambda \geq 0$

1.12 Vector Differentiation

- $z = \alpha^Ty$, $\frac{\partial z}{\partial y} = \alpha$ (α is a vector of constant)
- $z = y^Ty$, $\frac{\partial z}{\partial y} = 2y$
- $z = y^TAy$, $\frac{\partial z}{\partial y} = Ay + A^Ty$

1.13 Extra Problems

1. **Prove if symmetric A has $\lambda = 0$ or 1, it is idempotent.**
   hint: diagonalize A such that $A = PDP^T$, giving
   $A^2 = PD^2P^T = PDP^T = A$
   since D has 0 or 1 on diagonals, so $D^2 = D$

2. **If square matrix A is positive semidefinite, then λ (eigenvalues) ≥ 0**
   hint: $0 \leq x^T Ax = \lambda x^T x$
   since $x^T Ax \geq 0$, we have $x^T x > 0$ giving $\lambda \geq 0$

2 Random Vectors

2.1 Expectation

Let a be a vector of constants, A a matrix of constants.

- $E[a] = a$
- $E[a^T y] = a^T E[y]$
- $E[Ay] = AE[y]$

2.2 Variance

Definition: var $y = E[(y - \mu)(y - \mu)^T] = V$, with $\mu = E[y]$

- $[\text{var } y]_{ii} = \text{var } y_i$
- $[\text{var } y]_{ij} = \text{cov}(y_i, y_j) = E[y_i y_j] - \mu_i \mu_j$
- var $a^T y = a^T V a$
- var $Ay = AVA^T$

*** V is always positive semidefinite ***

2.3 Matrix Square Root

Definition: A is symmetric and positive semidefinite
$\Longrightarrow \exists$ unique symmetric positive semidefinite square root, $A^{1/2}$.

$$\begin{aligned} A &= PDP^T = PD^{1/2}D^{1/2}P^T \\ &= PD^{1/2}P^T PD^{1/2}P^T \\ \text{giving } A^{1/2} &= PD^{1/2}P^T \end{aligned}$$

2.4 Multivariate Normal Distribution (MVN)

Definition: Let z be $k \times 1$, with $z_i \sim N(0,1)$. A is $n \times k$, b is $n \times 1$. Then

$$x = Az + b \sim MVN(\mu, \Sigma)$$

with $\mu = b$, $\Sigma = AA^T$

- any linear combination of MVN results in another MVN
- if $z = (z_1, z_2)^T \sim$ MVN, then z_1, z_2 are independent $\iff$ they are uncorrelated
- let $x = \mu + \Sigma^{1/2} z \sim$ MVN (μ, Σ)
  $z = \Sigma^{-1/2}(x - \mu) \sim N(0,1)$ if $\Sigma^{-1/2}$ exists

The distribution of $y \sim MVN(X\beta, \sigma^2 I_n)$ can be expressed as below:

$$f(y) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp -(y - X\beta)^T (y - X\beta)/2\sigma^2$$

2.5 Random Quadratic Form

Expectation of Random Quadratic Form

$\boldsymbol{y}$, a random vector with $E[\boldsymbol{y}] = \boldsymbol{\mu}$, var($\boldsymbol{y}$)= V; A, a matrix of constants.

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

hint for proof:

$$\begin{aligned} E[\boldsymbol{y}^T A \boldsymbol{y}] &= E[\sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij} y_i y_j] \\ &= \sum_{i=1}^{k} \sum_{j=1}^{k} a_{ij} (\sigma_{ij} + \mu_i \mu_j) \end{aligned}$$

Non-central χ^2 Distribution

Let $\boldsymbol{y} \sim MVN(\mu, I)$. Then

$$x = \boldsymbol{y}^T \boldsymbol{y} = \sum_{i=1}^{k} y_i^2 \sim \chi^2_{k,\lambda}$$

where degrees of freedom $= k$, and noncentrality parameter $\lambda = \frac{1}{2} \boldsymbol{\mu}^T \boldsymbol{\mu}$.

- $\lambda = 0 \iff \mu = 0$. In this case, x is the sum of k independent standard normals, following χ^2_k.
- for a general case, var$(x) = 2k + 8\lambda$

Sum of χ^2 s

Let $X^2_{k_1,\lambda_1}, \cdots, X^2_{k_n,\lambda_n}$ be n independent noncentral χ^2 random variables, with $X^2_{k_i,\lambda_i} \sim \chi^2_{k_i,\lambda_i}$.
Then

$$\sum_{i=1}^{n} X^2_{k_i,\lambda_i} \sim \chi^2_{\boldsymbol{k},\boldsymbol{\lambda}}$$

with $\boldsymbol{k} = \sum_{i=1}^{n} k_i, \boldsymbol{\lambda} = \sum_{i=1}^{k} \lambda_i$.

- if $\lambda_i = 0, \forall i$, the distribution above results in $\chi^2_{\boldsymbol{k}}$.

Distribution of Quadratic Forms

1. Let $y \sim MVN(\mu, I)$ be a $n \times 1$ random vector, A be $n \times n$ symmetric matrix.

$$y^T Ay \sim \chi^2_{k,}, \lambda = \frac{1}{2}\mu^T A\mu \iff A \text{ is idempotent}, r(A) = k$$

2. Let $y \sim MVN(\mu, \sigma^2 I_n)$ be a $n \times 1$ random vector, A be $n \times n$ symmetric matrix.

$$\frac{1}{\sigma^2} y^T Ay \sim \chi^2_{k,}, \lambda = \frac{1}{2\sigma^2}\mu^T A\mu \iff A \text{ is idempotent}, r(A) = k$$

   *** Note: $\frac{1}{\sigma} y \sim MVN(\frac{1}{\sigma}\mu, I)$ ***

3. Let $y \sim MVN(\mu, V)$ be a $n \times 1$ random vector, A be $n \times n$ symmetric matrix.

$$y^T Ay \sim \chi^2_{k,}, \lambda = \frac{1}{2}\mu^T A\mu \iff AV \text{ is idempotent}, r(AV) = k$$

Independence of Quadratic Forms

1. Let $y \sim MVN(\mu, V)$, V be a nonsingular matrix, A, B be $n \times n$ symmetric matrix.
   Then $y^T Ay$ and $y^T By$ are independent $\iff AVB = 0$.

2. Let $y \sim MVN(\mu, \sigma^2 I)$, A, B be $n \times n$ symmetric matrix.
   Then $y^T Ay$ and $y^T By$ are independent $\iff AB = 0$.

3. Let $y \sim MVN(\mu, V)$, A be $n \times n$ symmetric matrix, and B a $m \times n$ matrix.
   Then $y^T Ay$ and By are independent $\iff BVA = 0$.

Combining the Theorems above, we have

Let $y \sim MVN(\mu, I)$, $A_1, \cdots, A_m$ be a set of $n \times n$ symmetric matrix.

Any two below implies the third:

- All A_i are idempotent
- $\sum_{i=1}^m A_i$ is idempotent
- $A_i A_j = 0 \forall i \neq j$

AND the following are true:

- $\forall i, y^T Ay \sim \chi^2_{r(A_i), \lambda_i}; \lambda_i = \frac{1}{2}\mu^T A\mu$
- $y^T A_i y, y^T A_j y$ are independent for $i \neq j$
- $\sum_{i=1}^m r(A_i) = r(\sum_{i=1}^m A_i)$

Cochran-Fisher Theorem

When $\sum_i A_i = I$, and let $y \sim MVN(\mu, \sigma^2 I)$ be a $n \times 1$ random vector. Decomposing the sum of squares into quadratic forms:

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

Then the quadratic forms

$$\frac{1}{\sigma^2} y^T A_i y \sim \chi^2_{r(A_i)}, \lambda$$
$$\lambda = \frac{1}{2\sigma^2}\mu^T A_i \mu \text{ are independent } \iff \sum_{i=1}^m r(A_i) = n$$

3 Full Rank Model

3.1 Definition of Full Rank

$$\boldsymbol{y} = X\boldsymbol{\beta} + \epsilon$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The linear model above as **full rank** when the design matrix X has full rank.

- condition: $r(X) = k + 1$
- this means that $X^T X$ is nonsingular
- NOTE: X, β are not random vectors

Let $\boldsymbol{y} = X\boldsymbol{\beta} + \epsilon$ be a full rank model, where X is a $n \times (k+1)$ matrix, $\boldsymbol{\beta}$ a vector of $(k \times 1)$ parameters, and ϵ a $(n \times 1)$ random vector with mean 0. We have

- $\mathrm{E}[\boldsymbol{y}] = E[X\boldsymbol{\beta} + \epsilon] = E[X\boldsymbol{\beta}] + E[\epsilon] = X\boldsymbol{\beta} + 0 = X\boldsymbol{\beta}$
- $\mathrm{Var}(\boldsymbol{y}) = \mathrm{Var}(X\boldsymbol{\beta} + \epsilon) = \mathrm{Var}(\epsilon)$

3.2 Least Squares Estimation (LSE)

Assumptions:

- error vector ϵ has mean 0, variance $\sigma^2 I$
- errors are independent of responses y
- errors are uncorrelated with each other (not necessarily independent)

The LSE

Let $\boldsymbol{y} = X\boldsymbol{\beta} + \epsilon$ be a full rank model. Let ϵ have mean 0 and variance $\sigma^2 I$.

The LSE for $\boldsymbol{\beta}$ is:

$$b = (X^T X)^{-1} X^T y$$

with

$$E[b] = \boldsymbol{\beta},\ \mathrm{Var}\ (b) = (X^T X)^{-1}\sigma^2$$

Derivation

- $\hat{E[y_i]} = b_0 + b_1 x_{i1} +, \cdots, + b_k x_{ik}$ which estimates the expected value of y
- residual, $e_i = y_i - \hat{E[y_i]}$
- errors, $\epsilon_i = y_i - E[y_i] = e_i + \hat{E[y_i]} - E[y_i]$

If our estimates are good, the residuals should be small (Note residuals is the difference between observed value, y_i and estimated value, $\hat{E[y_i]}$). Hence, the LSE method aims to **minimise** the sum of squares of residuals.

$$\begin{aligned}\sum_{i=1}^{n} e_i^2 = e^T e &= (y - Xb)^T (y - Xb) \\ &= y^T y - 2(X^T y)^T b + b^T (X^T X) b\end{aligned}$$

hint: use $(b^T X^T y)^T = y^T X b$.

To **minimise**, we put

$$\frac{\partial e^T e}{\partial b} = 0$$

giving

$$\begin{aligned}-2X^T y + 2(X^T X)b &= 0 \\ X^T X b &= X^T y\end{aligned}$$

which is the normal equations.

With the assumption that X is of full rank and $X^T X$ is nonsingular, we can solve for b to obtain the LSE:

$$b = (X^T X)^{-1} X^T y$$

The Best Linear Unbiased Estimator (BLUE) - Gauss-Markov Theorem

In a full rank linear model $\boldsymbol{y} = X\boldsymbol{\beta} + \epsilon$, the LSE $b = (X^T X)^{-1} X^T y$ is the unique BLUE for $\boldsymbol{\beta}$.
hint: prove by contraposition. (Suppose we have another unbiased linear estimator, $b^* = [(X^T X)^{-1} X^T B] + y$, with B being a $(k+1) \times n$ matrix.)

Estimation of Linear Functions

In a full rank linear model $\boldsymbol{y} = X\boldsymbol{\beta} + \epsilon$, $\boldsymbol{t}$ a vector of constants, the BLUE for $\boldsymbol{t}^T \boldsymbol{\beta}$ is $\boldsymbol{t}^T b$, where b is the LSE for $\boldsymbol{\beta}$.

3.3 Variance Estimation

$$s^2 = \frac{(y - Xb)^T(y - Xb)}{n - (k + 1)} \text{ is the unbiased estimator for } \sigma^2$$

Define the sum of squares of residuals,

$$SS_{Res} = (y - Xb)^T(y - Xb); \text{ with the of parameters, } p = k + 1$$

Then

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

and

$$\begin{aligned} E[s^2] &= \frac{1}{n-p} E[y^T(I - X(X^TX)^{-1}X^T)y] \\ &= \frac{1}{n-p} tr((I - X(X^TX)^{-1}X^T)\sigma^2 I) \\ &= \sigma^2(n - p) \end{aligned}$$

using the expectation of quadratic forms.

3.4 The Hat Matrix

Define the Hat matrix, H to be

$$H = X(X^TX)^{-1}X^T$$

Properties of Hat Matrix

- H is symmetric and idempotent
- $r(H) = p = k + 1$
- $I - H$ is symmetric and idempotent
- $r(I - H) = n - p$
- $HX = X(X^TX)^{-1}X^TX = X$
- $X^TH = X^T$

3.5 Diagnostics

- var $(\epsilon) = \sigma^2 I$
- but the variance of residuals, var $(e) \neq \sigma^2 I$
- we know in general, the further away X is from the center, the lower the variance of residuals

Using the Hat Matrix,

$$\begin{aligned} \text{Var } e &= \text{Var } (I - X(X^T X)^{-1} X^T) y \\ &= \text{Var } (I - H) y \\ &= \sigma^2 (I - H) \end{aligned}$$

Standardised Residuals

$$z_i = \frac{e_i}{\sqrt{s^2 (I - H_{ii})}}$$

*** var(z_i) ≈ 1***

Leverage

Leverage of a point i is H_{ii}.

- the larger H_{ii}, the larger the influence of each point y_i on the fit
- this results in large effect on estimated parameters and may distort the fit

Cook's Distance

$$\begin{aligned} D_i &= \frac{(b_{(-i)} - b)^T X^T X (b_{(-i)} - b)}{(k+1) s^2} \\ &= \frac{1}{k+1} z_i^2 (\frac{H_{ii}}{I - H_{ii}}) \end{aligned}$$

where $b_{(-i)}$ is the estimated parameter if the point i is removed from the model. Generally,

- $D_i > 1$ is considered large
- $D_i < 0.5$ is considered small

3.6 Diagnostics Plots & Interpretation

Things to look out for:

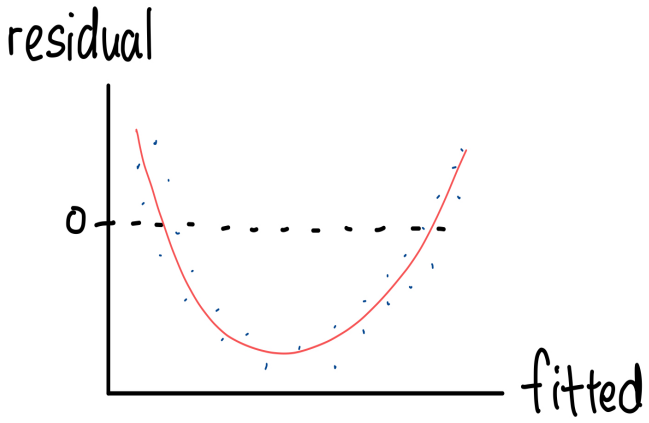

Figure 1: Residuals vs fitted plot

- large residual -> unequal variance
- trend -> may be bias
- pattern -> sign of high correlation between parameters

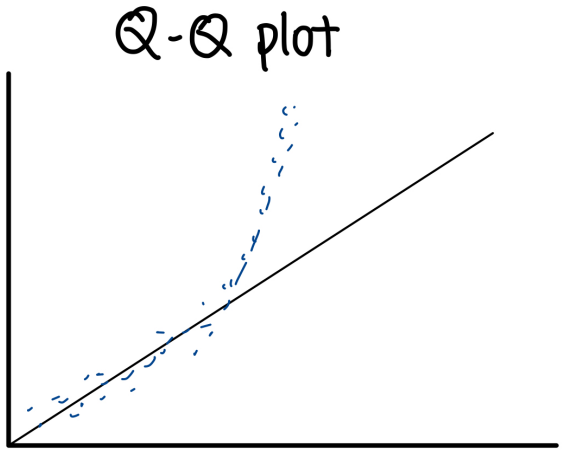

Figure 2: QQ-plot

- look for outliers (in the example, the outliers are at the right-tail)
- look for over/under-estimated tails
- look for skewness (indicating not normal errors)

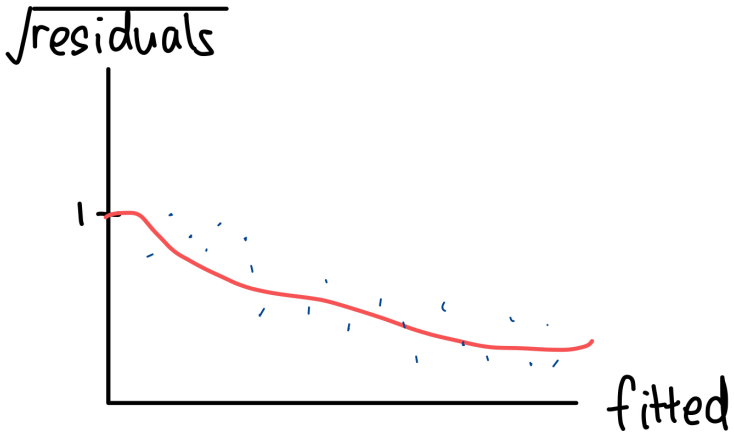

Figure 3: Square root of residuals vs fitted plot

- large residual
- trend in size (indicating heteroskedasticity)

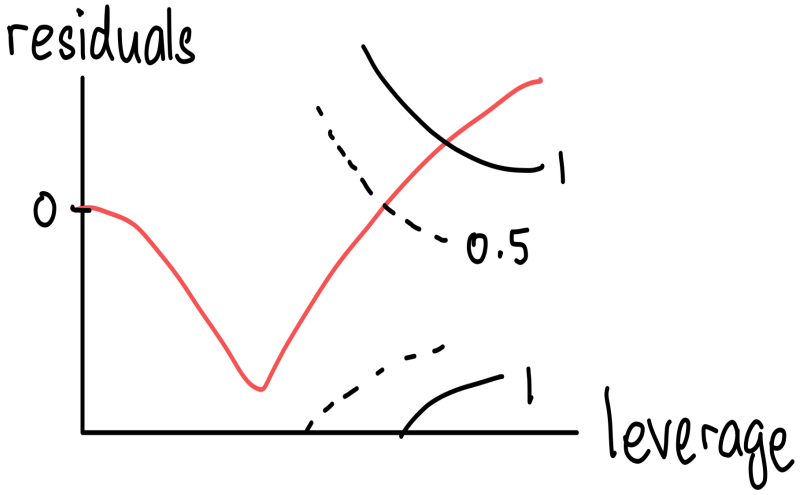

Figure 4: Leverage plot

- points with large residual indicating unequal variance
- high leverage & Cook's distance indicating large influence on fit
- significant pattern indicating correlation between parameters

3.7 Maximum Likelihood Estimation (MLE)

MLE refers to the maximum probability of observing the estimated parameter given data. It assumes:

- $\epsilon \sim MVN(0, \sigma^2 I)$
- errors are independent

*** Note that there are more assumptions for MLE than LSE***

The sketch of derivation of MLE is as follows:

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

where we obtain the same normal equations as that of the LSE.

However, we note that the variance of the MLE of β is unbiased:

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

Comparison of LSE and MLE:

	LSE	MLE
Assumptions	error term has mean 0, variance 1	$\epsilon \sim MVN(0, \sigma^2 I)$
Independence	independent errors	independent errors
Bias	unbiased	biased

Interval Estimation for MLE

We know that $\epsilon \sim MVN(0, \sigma^2 I)$.

Distributions of estimators from LSE:

$$\begin{aligned} b &= (X^T X)^{-1} X^T y \sim MVN(\beta, (X^T X)^{-1}\sigma^2) \\ \frac{(n-p)s^2}{\sigma^2} &= \frac{SS_{Res}}{\sigma^2} \sim \chi^2_{n-p} \end{aligned}$$

Noting that s^2 and b are independent.

3.8 Other Distributions

t - distribution

Let $Z \sim N(0,1), X_\gamma^2 \sim \chi_\gamma^2$; Z and X_γ^2 are independent. The t-distribution is defined as follows:

$$\frac{Z}{\sqrt{X_\gamma^2/\gamma}} \sim t_\gamma$$

with degrees of freedom γ.

F - distribution

Let $X_{\gamma_1}^2 \sim \chi_{\gamma_1}^2, X_{\gamma_2}^2 \sim \chi_{\gamma_2}^2$, with $X_{\gamma_1}, X_{\gamma_2}$ being independent. The F-distribution is defined as follows:

$$\frac{X_{\gamma_1}^2/\gamma_1}{X_{\gamma_2}^2/\gamma_2} \sim F_{\gamma_1,\gamma_2}$$

3.9 Confidence Interval

Deriving Interval Estimation

For a linear model with p parameters, n rows. Let var $(b_i\ ) = c_{ii}\sigma^2$, where c_{ii} is the ith diagonal of $(X^TX)^{-1}$.

1. Confidence interval for β_i

$$\frac{b_i - \beta_i}{\sigma\sqrt{c_{ii}}} \sim N(0,1)$$

$$\frac{b_i - \beta_i}{\sigma\sqrt{c_{ii}}} \Big/ \sqrt{\frac{SS_{Res}/\sigma^2}{n-p}} = \frac{b_i - \beta_i}{s\sqrt{c_{ii}}} \sim t_{n-p}$$

   giving

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

2. Confidence interval for $t^T\beta$

$$\frac{t^Tb - E[t^Tb]}{\sqrt{\text{ var } t^Tb}} = \frac{t^Tb - t^T\beta}{s\sqrt{t^T(X^TX)^{-1}t}} \sim t_{n-p}$$

   giving

$$t^Tb \pm t_{\alpha/2} s\sqrt{t^T(X^TX)^{-1}t}$$

3. Confidence interval for σ^2

$$\frac{(n-p)s^2}{\sigma^2} \sim \chi_{n-p}^2$$

   giving the region

$$P[\chi_{n-p}^2(\frac{\alpha}{2}) \leq \frac{(n-p)s^2}{\sigma^2} \leq \chi_{n-p}^2(1-\frac{\alpha}{2})] = 1-\alpha$$

Deriving Prediction Intervals

Let y^* be a new observation,

$$y^* = (x^*)^T\beta + \epsilon^*$$

with

$$\text{Var } (\epsilon^*) = \sigma^2$$

$y^* - x^*)^T b$ is the estimator. To find its variance:

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

Then we have:

$$\frac{y^* - (x^*)^T b}{s\sqrt{I + (x^*)^T (X^T X)^{-1} x^*}} \sim t_{n-p}$$

Joint Confidence Interval

Using the quadratic form:

$$(b - \beta)^T X^T X (b - \beta)/ps^2 \sim F_{p,n-p}$$

gives the $(1 - \alpha)\%$ confidence region:

$$(b - \beta)^T X^T X (b - \beta) \leq ps^2 f_\alpha;$$

The $(1 - \alpha)$ % region above is greater than the rectangle region that assumes the parameters are independent.

3.10 Other Cases of Parameter Estimations

Generalised Least Squares

The condition where $\epsilon \sim MVN(0, V)$ instead. Through similar derivations,

$$\begin{aligned} e^T V^{-1} e &= (y - Xb)^T V^{-1} (y - Xb) \\ &\vdots \\ X^T V^{-1} X b &= X^T V^{-1} y \\ b &= (X^T V^{-1} X)^{-1} X^T V^{-1} y \end{aligned}$$

Gauss-Markov still holds, and we have

$$E[b] = \beta$$

$$\text{Var } b = (X^T V^{-1} X)^{-1}$$

Weighted Least Squares

The condition where errors are uncorrelated but do not share a common variance. We can estimate through Maximum Likelihood methods. and minimise the following:

$$(y - Xb)^T V^{-1} (y - Xb) = \sum_{i=1}^{n} (\frac{e_i}{\sigma_i})^2$$

3.11 Transformations

Signs that transformations may be required:

- distribution of data is skewed
- obvious non-linear relationship with another variable
- variances show a relationship with other variables
- all values of observations are very large (consider log transform)

3.12 Extra

Interpretation of Confidence Interval for β_i

Let β_i measure the percentage of population, and let y be the log of birth rate per 100 females.

Let $(-a, -b)$ be the confidence interval of β_i, with a, b being positive numbers. This means that for every 1% increase of population, the log birth rate *reduces* by an amount between a and b.

4 Inference for Full Rank Model

4.1 Model Relevance

Assume $\epsilon \sim MVN(0, \sigma^2 I)$.

1. To test whether all parameters β contribute to y:

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0$$

*** Note that this is a two-tailed test ***

2. To test whether the residuals have a pattern (ANOVA) under H_0,

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

*** Note that this is a one-tailed test ***

and

$$\begin{aligned} SS_{Total} &= y^T y \\ SS_{Res} &= (y - Xb)^T (y - Xb) \\ SS_{Reg} &= y^T X (X^T X)^{-1} X^T y \\ &= b^T X^T X b = y^T X b = y^T H y \end{aligned}$$

4.2 General Linear Hypothesis

Let C be $r \times p$ matrix of rank $r \leq p$, δ^* a $r \times 1$ vector of constants.

$$H_0 : C\beta = \delta^*$$
$$H_1 : C\beta \neq \delta^*$$

and we have

$$E[Cb - \delta^*] = C\beta - \delta^*$$
$$\text{Var } (Cb - \delta^*) = C(X^T X)^{-1} C^T \sigma^2$$

Under H_0, the critical region can be found:

$$\frac{(Cb - \delta^*)^T [C(X^T X)^{-1} C^T]^{-1} (Cb - \delta^*)/r}{SS_{Res}/(n-p)} \sim F_{r,n-p}$$

Noting that the numerator was found as:

$$(Cb - \delta^*)^T [C(X^T X)^{-1} C^T]^{-1} (Cb - \delta^*)/\sigma^2 \sim \chi^2_{r,\lambda}$$

with

$$\lambda = \frac{1}{2} \mu^T V^{-1} \mu$$

When $C = I$, we have the denominator

$$(b - \beta^*)^T X^T X (b - \beta^*) = (y - X\beta^*)^T (y - X\beta^*)$$

so

$$\frac{SS_{Res}}{n-p} \sim \chi^2_{n-p}$$

4.3 Testing the Subset of a Model

To test part of β. Split

$$\beta = \left[\begin{array}{c} \beta_0 \\ \vdots \\ \beta_{r-1} \\ \hline \beta_r \\ \vdots \\ \beta_k \end{array}\right] = \left[\begin{array}{c} \gamma_1 \\ \hline \gamma_2 \end{array}\right]$$

Then we have the following hypothesis test,

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

$$\text{Full model } : y = X\beta + \epsilon$$
$$\text{Reduced model } : y = X_2\gamma_2 + \epsilon_2$$

Let $C = [I_r|0], \delta^* = 0$. Then $C\beta = \delta^* \iff \gamma_1 = 0$.
Define the regression sum of squares $(\gamma_1|\gamma_2)$:

$$\begin{aligned} R(\gamma_1|\gamma_2) &= (Cb - \delta^*)^T (C(X^TX)^{-1}C^T)^{-1}(Cb - \delta^*) \\ &= \hat{\gamma_1}^T A_{11}^{-1} \hat{\gamma_1} \end{aligned}$$

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

The test statistic acquired is:

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

THEOREM

$$R(\gamma_1|\gamma_2) = R(\beta) - R(\gamma_2)$$

where $R(\beta)$ is the regression sum of squares for the *full model*,
and $R(\gamma_2)$ is the regression sum of squares for the *reduced model*.

ANOVA Table

Regression	SSE	d.f.
Full model	$R(\beta) = y^T H y$	p
Reduced	$R(\gamma_2)$	$p - r$
$\gamma_1\|\gamma_2$	$R(\gamma_1\|\gamma_2)$	r
Residual	$y^T y - R(\beta)$	$n - p$
Total	$y^T y$	n

4.4 Corrected Sum of Squares

For this, we do not consider β_0 (the intercept) in the model.

$$H_0 : \beta_1 = \cdots = \beta_k = 0$$
$$H_1 : \text{ some } \beta_i \neq 0$$

with $i \in \{1, \cdots, k\}$.

So the sum of squares of reduced model, $R(\gamma_2) = (\sum_{i=1}^{n} y_i)^2/n$.

This gives the corrected sum of squares as

$$\begin{aligned} y^T y - R(\gamma_2) &= \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2/n \\ &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \end{aligned}$$

The equivalent ANOVA table is as follows:

		d.f.
Regression	$SS_{Reg} - (\sum y_i)^2/n$	k
Residual	SS_{Res}	$n-k-1$
Total	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	n

R: summary(model)

- the F-statistic provided in summary(model) is the model relevance test with corrected Sum of Squares
- the p-values provided for $\beta_0, \beta_1, \cdots \beta_k$ individually are testing for the significance of each variable

4.5 Sequential Testing

The parameters and variables we select incrementally (or decrementally) may be correlated.

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

where SS_{Res} has $n-p$ degrees of freedom and the others have 1 degree of freedom. This gives:

$$\frac{R(\beta_k|\beta_0,\beta_1,\cdots\beta_{k-1})}{SS_{Res}/(n-p)} \sim F_{1,n-p}$$

- forward selection and backward elimination (F-values used) may result in different model outcomes
- same goes for stepwise selection (AIC is used) as a goodness-of-fit measure

t-test as partial test

When only testing for 1 parameter, the t-test can be used, as it is equivalent to F-test when there is only 1 parameter considered.

$$H_0 : \beta_i = 0$$
$$H_1 : \beta_i \neq 0$$

Find the confidence interval (CI) for β_0 for the hypothesis test. If CI includes 0, do not reject H_0.

$$\frac{b_i}{s\sqrt{c_{ii}}} \sim t_{n-p}$$
$$\equiv \frac{R(\beta_k|\beta_0,\beta_1,\cdots\beta_{k-1})}{SS_{Res}/(n-p)} = \frac{b_i^2}{c_{ii}s^2} \sim F_{1,n-p}$$

4.6 Goodness-of-fit measures

Sample Variance, s^2

- when a parameter is added, p increases so $n - p$ decreases
- the smaller s^2, the better the parameter (i.e decreases SS_{Res})

R^2

$$R^2 = 1 - \frac{SS_{Res}}{SS_{total} - (\sum y_i)^2/n}$$

- proportion of corrected SS_{total} explained by the model
- the larger R^2, the more variation is explained by the model
- downside: irrelevant variables may be included

Adjusted R^2

$$I - \frac{n-1}{n-1-k}(I - R^2)$$

AIC

$$\begin{aligned} AIC &= -2\log(\text{likelihood}) + 2p \\ &= n\log(\frac{SS_{Res}}{n}) + 2p + \text{ const.} \end{aligned}$$

- likelihood vs penalty term for # of parameters
- smaller AIC is preferred

BIC

$$\begin{aligned} BIC &= -2\log(\text{likelihood}) + p\log n \\ &= n\log(\frac{SS_{Res}}{n}) + p\log n + \text{ const }. \end{aligned}$$

- $\log n > 2$, so BIC penalises heavier
- promotes model with fewer variables

Mallow's C_p

$$C_p = \frac{SS_{Res}(\text{intermediate model})}{s^2(\text{full model})} + 2p - n$$

where p is the # of parameters in the intermediate model.

- smaller C_p is preferred

4.7 Shrinkage

A more holistic approach to model selection.

Ridge regression

$$min_b \sum_{i=1}^{n} e_i^2 + \lambda \sum_{j=1}^{k} b_j^2$$

- λ : penalty term
- parameters $\neq 0$
- instead they are closer to 0, so are less penalised

LASSO

$$min_b \sum_{i=1}^{n} e_i^2 + \lambda \sum_{j=1}^{k} |b_j|$$

- parameters may be 0
- for variable/parameter selection

4.8 Extra

R output for anova(null, model):

Model 1: $y = x_1$

Model 2: $y = x_1 + x_2$

Model	Res. df	RSS	Df	Sum of Sq	F	Pr(>F)
1	137	9.3				
2	136	4.7	1	4.6	134	<2.2e-16

Identifying the sum of squares:

- $SS_{Res} = 4.7$
- $SS_{Reg}(\gamma_2) = 9.3$
- $SS_{Reg}(\gamma_1|\gamma_2) = 4.6$
- noting that $SS_{Reg}(\gamma_2) - SS_{Res} = SS_{Reg}(\gamma_1|\gamma_2)$

5 Less than Full Rank Model

5.1 Definition of Less than Full Rank

X is not of full rank. For example, the 1-way classification mode:

$$\begin{aligned} y_{ij} &= \mu + \tau_i + \epsilon_j \\ & i = 1, 2, \cdots, k; j = 1, 2, \cdots, n_i \end{aligned}$$

- k is the # of treatments
- n_i is the # pf samples from the ith population

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

We see that the first column of X is the sum, of the remaining columns.

5.2 Reparametrization (1-way classification)

One way to let X be of full rank is through reparametrization.

The 1-way classification model:

$$\begin{aligned} y_{ij} &= \mu + \tau_i + \epsilon_j \\ & i = 1, 2, \cdots, k; j = 1, 2, \cdots, n_i \end{aligned}$$

Rewrite $\mu_i = \mu + \tau_i$, we have

$$y_{ij} = \mu_i + \epsilon_{ij}$$

Giving design matrix X to have full rank, and

$$\beta = \begin{bmatrix} \tau_1 \\ \vdots \\ \tau_k \end{bmatrix}$$

5.3 Conditional Inverse

A is $n \times p$ matrix, A^c is $p \times n$ matrix, $\iff AA^cA = A$.

Algorithm:

1. Find minor M of A, which is nonsingular and of dimension $r(A) \times r(A)$.
2. Replace M in A with $(M^{-1})^T$, other entries with 0.
3. Transpose the resulting matrix.

Conditional Inverse Properties

A is $n \times p$ matrix, $r(A) = r$.

A^c is not always unique

- $r(A) = r(AA^c) = r(A^cA)$
- $(A^c)^T = (A^T)^c$
- $A^cA, AA^c, I - A^cA, I - AA^c$ are idempotent
- $A = A(A^TA)^c(A^TA)$; $A^T = (A^TA)(A^TA)^cA^T$
- $A(A^TA)^cA^T$ is unique, symmetric and idempotent (similarly for $A(A^TA)^cA^T$)
- $r(A(A^TA)^cA^T) = r(A) = r$
- $r(I - A(A^TA)^cA^T) = n - r$

5.4 Solving Normal Equations

Consistent System

- System $Ax = g$ is consistent (at least 1 solution) $\iff r(A|g) = r(A)$.
- Let $Ax = g$ be consistent. Then $A^c g$ is a solution, where A^c is any conditional inverse for A.

  Proof: $Ax^* = g$ for some x^*,

  $AA^c g = AA^c Ax^* = Ax^* = g$

  hence, $A^c g$ solves the equation.
- in general linear models, $X^T Xb = X^T y$ is consistent, and $b = (X^T X)^c X^T y$ solves the equation.

The third holds due to:

$$r(X^T X|X^T y) \geq r(X^T X);$$

but

$$\begin{aligned} r(X^T X|X^T y) &= r(X^T(X|y)) \\ &\leq r(X^T) \text{ (holds true in general)} \\ &= r(X^T X) \end{aligned}$$

*** plugging $A = X^T X, g = X^T y$***

Solving the system

Let $Ax = g$ be a consistent system. Then

$$x = A^c g + (I - A^c A)z$$

solves the system, where z is an arbitrary $p \times 1$ matrix, and A^c is any conditional inverse of A.

- a less than full rank model has an infinite # of solutions for its normal equations
- vector of the form $b = (X^T X)^c X^T y + (I - (X^T X)^c X^T X)z$ satisfies the equations
- similarly, given x_0 as any solution to the system $Ax = g$, then for any A^c,

$$x_0 = A^c g + (I - A^c A)z \text{ where } z = x_0$$

5.5 Estimability

In general linear models, functin $t^T\beta$ is estimable if $\exists$ a vector c, such that

$$E[c^T y] = t^T\beta$$

- $t^T\beta$ is estimable $\iff \exists$ solution to $X^TXz = t$ linear system
- $t^T\beta$ is estimable $\iff t^T(X^TX)^cX^TX = t^T$

  $\forall$ conditional inverse of X^TX

Gauss-Markov Theorem

The *BLUE* for $t^T\beta$ is z^TX^Ty, where z is a solution to $X^TXz = t$.

- this estimate is the same for any solution of the system
- estimate can be written as t^Tb, where b is any solution of the normal equation
- $z_1^TXy = t^Tb = z_0^TX^Ty = z_0^TX^TXb = (X^TXz_0)^Tb$ as the BLUE is unique

Estimable Functions

In linear models, $y = X\beta + \epsilon$, elements of $X\beta$ are estimable

- $\mu + \tau_1 = \begin{bmatrix}1 & 1 & 0 & 0\end{bmatrix}\beta$ is estimable
- let $t_1^T\beta, \cdots, t_k^T\beta$ be estimable functions, and let $z = \alpha_1 t^T\beta, \cdots, \alpha_k t_k^T\beta$. Then z is estimable, with its BLUE being $\alpha_1 t^Tb, \cdots, \alpha_k t_k^Tb$

5.6 Treatment Contrast

$$\alpha_1\tau_1 + \alpha_2\tau_2 + \cdots + \alpha_k\tau_k$$

with condition

$$\sum_{i=1}^{k}\alpha_i = 0$$

- wipe out effect of overall mean response to describe the differences between populations
- estimate treatment effects relative to each other
- in 1-way classification, any treatment contrast is estimable

R

	contr. treatment	contr. sum
Intercept	$\mu + \tau_1$	$\mu + \frac{1}{k}\sum \tau_i$
f_1	-	$\tau_1 - \frac{1}{k}\sum \tau_i$
f_2	$\tau_2 - \tau_1$	$\tau_2 - \frac{1}{k}\sum \tau_i$
$\vdots$	$\vdots$	$\vdots$
f_k	$\tau_k - \tau_1$	-

5.7 Estimating σ^2

Let r be rank of X. We know

$$SS_{Res} = (y - Xb)^T(y - Xb) = y^T[I - X(X^TX)^cX^T]y$$

Then

$$\begin{aligned} E[SS_{Res}] &= E[y^T(I - H)y] \\ &= tr(I - H)\sigma^2 \\ &= r((I - H)\sigma^2) \\ &= (n - r)\sigma^2 \end{aligned}$$

Therefore,

$$s^2 = SS_{Res}/(n - r)$$

The distribution for estimating σ^2

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

if $t^T\beta$ is estimable, t^Tb is independent of s^2

5.8 Interval Estimation

If t is estimable,

$$\begin{aligned}\text{Var } (t^T b) &= \text{ Var } (t^T (X^T X)^c X^T y) \\ &= t^T (X^T X)^c X^T \sigma^2 I X (X^T X)^c t \\ &= \sigma^2 t^T (X^T X)^c t\end{aligned}$$

Writing a $(1-\alpha)\%$ confidence interval,

$$\frac{t^T b - t^T \beta}{s\sqrt{t^T (X^T X)^c t}} \sim t_{n-r}$$

giving

$$t^T b \pm t_{\alpha/2} s \sqrt{t^T (X^T X)^c t}$$

Difference between class 3 and average of other 2 classes

$$\begin{aligned}\tau_3 - \frac{1}{2}\tau_2 - \frac{1}{2}\tau_1 &= (\tau_3 - \tau_1) - \frac{1}{2}(\tau_2 - \tau_1) \\ &= \text{ class f3} - \frac{1}{2} \text{ class f2}\end{aligned}$$

Then we have

$$C = \begin{bmatrix} 0 & -0.5 & 1 \end{bmatrix}$$

an estimable quantity.

Interval Estimation with R

- the estimates in model's summary are treatment contrasts
- if we want to test $\gamma_3 - \gamma_1$, then the standard error of the quantity is the standard error of the 3rd parameter given in the model summary

5.9 Prediction Interval for $t^T \beta$

Set up:

$$\begin{aligned} t^T b &\sim N(t^T \beta, \sigma^2 t^T (X^T X)^c t), \\ y^* &= t^T \beta + \epsilon \text{ so} \\ t^T b - y^* &\sim N(0, \sigma^2 (I + t^T (X^T X)^c t), \\ &\text{where } r = rank(X) \end{aligned}$$

we can write the $(1-\alpha)\%$ prediction interval as:

$$\frac{t^T b - y^*}{s\sqrt{I + t^T (X^T X)^c t}} \sim t_{n-r}$$

$$t^T b \pm t_{\alpha/2} s \sqrt{I + t^T (X^T X)^c t}$$

6 Inference for Less than Full Rank Model

6.1 Testability

H_0 is *testable* if $\exists$ a set of estimable functions $c_1^T\beta, \cdots, c_m^T\beta$ such that H_0 is true $\iff$

$$c_1^T\beta = c_2^T\beta = \cdots = 0$$

and $c_1, \cdots, c_m$ are linearly independent.

The distribution of testable elements

Suppose $C\beta = 0$ is testable, C is a $m \times p$ matrix, $r(C) = m$.

$$\frac{(Cb)^T[C(X^TX)^cC^T]^{-1}Cb}{\sigma^2} \sim \chi^2_{m,\lambda}$$

with

$$\lambda = \frac{(C\beta)^T[C(X^TX)^cC^T]^{-1}C\beta}{2\sigma^2}$$

Proof (hint)

$$Cb = C(X^TX)^cX^Ty \sim MVN(C\beta, C(X^TX)^cC^T\sigma^2)$$

And Cb, s^2 are independent. We have the test statistic

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

6.2 1-factor model

Important formulas for 1-way classification:

$$X^TX = \begin{bmatrix} n & n_1 & \cdots & \cdots & n_k \\ n_1 & n_1 & 0 & \cdots & 0 \\ n_2 & 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 \\ n_k & 0 & \vdots & \vdots & n_k \end{bmatrix} \quad (X^TX)^c = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 1/n_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \ldots & 1/n_k \end{bmatrix}$$

$$X^Ty = \begin{bmatrix} \sum y_{ij} \\ \sum y_{1j} \\ \vdots \\ \sum y_{kj} \end{bmatrix} \quad b = (X^TX)^cX^Ty = \begin{bmatrix} 0 \\ \bar{y_1} \\ \vdots \\ \bar{y_k} \end{bmatrix}$$

6.3 2-factor model

Important notes for 2-way classification:

Assume level of each factor affects μ.

$$y_{ijk} = \mu + \tau_i + \beta_j + \epsilon_{ijk}$$

In an additive 2-factor model, every contrast in τ and β is estimable.

R's Additive 2-factor model

In R, μ_{ij} is expressed as:

$$\mu_{ij} = \mu + \tau_i + \beta_j$$

Example of the parameters of model in a table:

Intercept	$\mu_{11} = \mu + \tau_1 + \beta_1$
f2	$\mu_{21} - \mu_{11} = \tau_2 - \tau_1$
f3	$\mu_{31} - \mu_{11} = \tau_3 - \tau_1$
g2	$\mu_{12} - \mu_{11} = \beta_2 - \beta_1$
g3	$\mu_{13} - \mu_{11} = \beta_3 - \beta_1$

6.4 Interaction model

- the effect of 1 predictor depends on the level of another predictor

The model is of the form:

$$y_{ijk} = \mu + \tau_i + \beta_j + \zeta_{ij} + \epsilon_{ijk}$$

There is no interaction $\iff$

$$(\zeta_{ij} - \zeta_{ij'}) - (\zeta_{i'j} - \zeta_{i'j'}) = 0 \tag{1}$$

$\forall i \neq i^{'}, j \neq j^{'}$. and all quantities are estimable.

- require $(a-1)(b-1)$ equations, where there are $a \times b$ is the size of ζ matrix
- $\mu_{ij} - \mu_{ij'} = \mu_{i'j} - \mu_{i'j'}$ reduces to the equation 1 above.
- μ_{ij} are elements of $X\beta$, so linear combinations of them are also estimable

R's 2-factor Interaction model

In R, μ_{ij} is expressed as:

$$\mu_{ij} = \mu + \tau_i + \beta_j + \zeta_{ij}$$

Example of the parameters of model in a table:

Intercept	$\mu_{11} = \mu + \tau_1 + \beta_1 + \zeta_{11}$
f2	$\mu_{21} - \mu_{11} = \tau_2 - \tau_1 + \zeta_{21} - \zeta_{11}$
f3	$\mu_{31} - \mu_{11} = \tau_3 - \tau_1 + \zeta_{31} - \zeta_{11}$
g2	$\mu_{12} - \mu_{11} = \beta_2 - \beta_1 + \zeta_{12} - \zeta_{11}$
g3	$\mu_{13} - \mu_{11} = \beta_3 - \beta_1 + \zeta_{13} - \zeta_{11}$
f2: g2	$\mu_{22} - \mu_{21} - \mu_{12} + \mu_{11} = \zeta_{22} - \zeta_{21} - \zeta_{12} + \zeta_{11}$
$\vdots$	$\vdots$
fi: gj	$\mu_{ij} - \mu_{i1} - \mu_{1j} + \mu_{11} = \zeta_{ij} - \zeta_{i1} - \zeta_{1j} + \zeta_{11}$

6.5 ANCOVA model

This model applies when there is at least 1 categorical and 1 continuous predictor.

The model has the form:

$$y_{ij} = \mu + \tau_i + \beta x_{ij} + \zeta_i x_{ij} + \epsilon_{ij}$$

- if there is interaction, the slopes of the regression line are different for each population/treatment

R's ANCOVA

In R, there is a regression line for each sub-population:

$$y = \alpha_i + \beta_i x$$

Example of the parameters of model in a table:

Intercept	$\alpha_1 = \mu + \tau_1$
f2	$\alpha_2 - \alpha_1 = \tau_2 - \tau_1$
f3	$\alpha_3 - \alpha_1 = \tau_3 - \tau_1$
x	$\beta_1 = \beta + \zeta_1$
f2: x	$\beta_2 - \beta_1 = \zeta_2 - \zeta_1$
f3: x	$\beta_3 - \beta_1 = \zeta_3 - \zeta_1$

7 Experimental Design

7.1 Principles

Control

- control group does not receive treatment, as a basis for comparison

Blocking

- given known confounding factor, partition the population into blocks
- this decreases variation between blocks

Randomization

- reduces the effect of confounding factors
- allocating treatments randomly to otherwise homogeneous pool of subjects

Replication

- increases precision - minimises variance of estimators leading to balanced design

7.2 Types of Design

Complete Randomised Design (CRD)

Use Case: 1-way classification, no confounding factors

The model is usually a 1-way classification model:

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

By Reparametrization, we obtain the estimates of β as

$$b = (X^T X)^c X^T y = \begin{bmatrix} 0 \\ \bar{y_{.1}} \\ \vdots \\ \bar{y_{.k}} \end{bmatrix}$$

so $\hat{\mu_i} = \bar{y}$

To find the variance of $\mu_1 = \mu + \tau_1$,

$$\begin{aligned}
\text{Var } (\hat{\mu_i}) &= \text{Var } (t^T b) = t^T (X^T X)^c t \sigma^2 \\
&= \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \end{bmatrix} (X^T X)^c \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \end{bmatrix} \sigma^2 \\
&= \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 1/n_1 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \dots & 1/n_k \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \end{bmatrix} \sigma^2 \\
&= \sigma^2 / n_1
\end{aligned}$$

THEOREM (Calculate number of test units)

In CRD with n test units, the allocation of test units to factor levels that minimises

$$\sum_{i=1}^{k} \text{var} \hat{\mu_i} = \sigma^2 \sum_{i=1}^{k} \frac{1}{n_i}$$

is $n_i = \frac{n}{k}$, assuming $n \bmod k = 0$.

To solve:

- $f(n_1, \cdots, n_k, \lambda) = \sigma^2 \sum_{i=1}^{k} \frac{1}{n_i} + \lambda(\sum_{i=1}^{k} n_i = n)$
- let $\frac{\partial f}{\partial n_i} = 0 = \frac{\sigma^2}{n_i^2} + \lambda$
- so $n_i^2 = \frac{\sigma^2}{\lambda}$

Complete Block Design (CBD)

Use Case: 2-way classification, 1 confounding factors

- blocks of size $k = \#$ of treatments
- 1 treatment for each experimental unit
- do not need to account for interaction as blocking already removed the block effects

The linear model form is as below:

$$y = \left[\begin{array}{c|c} X_1 & X_2 \end{array}\right] \left[\begin{array}{c} \mu \\ \beta \\ \hline \tau \end{array}\right] + \epsilon$$

where μ, β are nuisance parameters.

Solving CBD's Normal Equations

In general linear model

$$y = \left[\begin{array}{c|c} X_1 & X_2 \end{array}\right] \left[\begin{array}{c} \beta_1 \\ \hline \beta_2 \end{array}\right] + \epsilon$$

$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$ is a solution to the normal equations $\iff$ b_2 is a solution to the reduced normal equations:

$$X_2^T[I - H]X_2 b_2 = X_2^T[I - H]y$$

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

Since $I - H_1$ is symmetric and idempotent, it is the same as writing

$$\begin{aligned} y &= ([I - H_1]X_2)\beta_2 + \epsilon \\ &= X_{2|1}\beta_2 + \epsilon \end{aligned}$$

with derivations, we have the estimators

$$\begin{aligned} b_2 &= (X_{2|1}^T X_{2|1})^c X_{2|1}^T y \\ &= \begin{bmatrix} \bar{y_{.1}} - \bar{y_{..}} \\ \vdots \\ \bar{y_{.k}} - \bar{y_{..}} \end{bmatrix} \end{aligned}$$

$t^T\tau$ estimates in CBD

Let J_k be a $k \times k$ matrix of ones, 1_k a vector of ones.

If $t^T\tau$ is estimable, then

$$\begin{aligned} t^T(X_{2|1}^T X_{2|1})^c X_{2|1}^T X_{2|1} &= t^T \\ t^T \frac{1}{b} I_k b[I_k - \frac{1}{k} J_k] &= t^T \\ t^T[I_k - \frac{1}{k} J_k]1_k &= t^T 1_k \\ t^T[1_k - 1_k] = 0 &= t^T 1_k \end{aligned}$$

Then

$$t^T \begin{bmatrix} \bar{y_{.1}} - \bar{y_{..}} \\ \vdots \\ \bar{y_{.k}} - \bar{y_{..}} \end{bmatrix} = t^T \begin{bmatrix} \bar{y_{.1}} \\ \vdots \\ \bar{y_{.k}} \end{bmatrix}$$

This means that the estimates of CBD and CRD are the same.

Differences in CRD & CBD

Model	CRD	CBD
Variance	lower	higher
Degrees of freedom (for s^2)	$b \times k - k$	$b \times k - (b + k - 1)$

Latin Squares

Use Case: 3-way classification, 2 confounding factors

- generally more efficient than CBD
- each treatment appears exactly once in each row and column
- can use ANCOVA to test for significance of parameters

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

one k for each (i, j) pair, $1 \leq (i, j) \leq t$

giving the form

$$y = X_1 \begin{bmatrix} \mu \\ \beta \\ \gamma \end{bmatrix} + X_2\tau + \epsilon$$

with derivations, we will get the estimator of β as

$$b_2 = \begin{bmatrix} \bar{y_{..1}} - \bar{y_{..}} \\ \vdots \\ \bar{y_{..t}} - \bar{y_{..}} \end{bmatrix}$$

which is equivalent to the estimates from CBD and CRD.

BIBD

Use Case: 1 confounding factor but CBD impossible

- t treatment levels, b blocks of size $k < t$ (when there's natural block size)
- each treatment occurs at most once in a block
- each treatment exactly $r = bk/t$ times
- each pair of treatment occur in the same # of blocks,

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

- $t(t-1)/2$ different pairs of treatments
- $bk(k-1)/2$ available slots

We can always find BIBD with $b = \binom{t}{k}$ by taking all possible subsets of size k. Then

$$\begin{aligned} r &= \frac{bk}{t} = \binom{t}{k}\frac{k}{t} \\ \lambda &= r\frac{k-1}{r-1} = \binom{t}{k}\frac{k}{t}\frac{k-1}{t-1} \\ &= \binom{t-1}{k-1}\frac{k-1}{t-1} = \binom{t-2}{k-2} \end{aligned}$$

BIBD Model

$$y_{ij} = \mu + \beta_i + \tau_j + \epsilon_{ij}$$

for $1 \le i \le b$, $j \in S(i)$, where $S(i)$ are the treatments in block i.

The solution to the reduced normal equations is:

$$b_2 = \frac{k}{\lambda t}q;$$

$$q = t - X_2^T X_1 b$$

- unlike previous models, there is bias in the estimator b_2

Steiner System, $S(t, k, n)$

- $n-$element set S
- $k-$element subset of S (blocks)
- each $t-$element subset of S appears in exactly 1 block