Linear statistical models
The less than full rank model

Yao-ban Chan

The less than full rank model

In previous sections we used the linear model

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

in the knowledge (or assumption) that X, of dimension $n \times p$, is of full rank, i.e. $r(X) = p$.

This assumption is important because a full rank X implies that X^TX is invertible, and therefore the normal equations

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

have a unique solution.

$$b = \underline{(X^TX)}^{-1}X^Ty$$

The less than full rank model

Unfortunately, not all linear models fall into this category.

For example, consider the *one-way classification model with fixed effects*.

In this model, samples come from k distinct (sub-)populations, with different characteristics. We wish to determine the differences between these populations.

One-way classification model

ANOVA

For example:

- A medical researcher compares three different types of pain relievers for effectiveness in relieving arthritis;
- A botanist studies the effects of four experimental treatments used to enhance the growth of tomato plants; or
- An engineer investigates the sulfur content in the five major coal seams in a particular geographic region.

One-way classification model

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

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij},$$

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

- k is the number of populations/treatments;
- n_i is the number of samples from the ith population.

One-way classification model

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

$$\mathbf{y} = X \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

The first column of X is the sum of the remaining columns, and therefore X is not of full rank.

One-way classification model

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

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

One-way classification model

The linear model is $r(X)=3$

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

$$\mathbf{y} \quad = \quad X \quad \boldsymbol{\beta} \quad + \quad \boldsymbol{\varepsilon}$$

The less than full rank model

The difficulty with a less than full rank model is that $\underline{X^TX}$ is singular. This means that the normal equations do not have a unique solution.

Ex: $r(X^TX)=3$ X^TX: 4×4 matrix

However, the problem goes deeper than that: not only can we not estimate the parameters, but the parameters themselves are not well defined.

The less than full rank model

(X^TX)

In a one-way classification model, the response variable from population i has a mean of $\mu + \tau_i$. Thus, for our carbon removal example we might have

$$\begin{aligned} \mu + \tau_1 &= 36 \\ \mu + \tau_2 &= 39 \\ \mu + \tau_3 &= 27. \end{aligned}$$

So our parameters might be $\mu = 34, \tau_1 = 2, \tau_2 = 5, \tau_3 = -7$.

However, we can also have $\mu = 30, \tau_1 = 6, \tau_2 = 9, \tau_3 = -3$.

In fact we can choose μ to be any real number, and still describe the system.

Reparametrization

One way we can tackle the less than full rank model is to convert to a full rank model. We can then use all the machinery we have developed.

Example. Consider the one-way classification model with $k = 3$. The less than full rank model for this is

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij},$$

$$\underline{y_{ij} = \mu_i + \varepsilon_{ij}}$$

for $i = 1, 2, 3$, $j = 1, 2, \ldots, n_i$.

However, we can write the mean of each population as

$$\underline{\mu_i = \mu + \tau_i}.$$

Reparametrization

Then we can recast the model as

$$y_{ij} = \mu_i + \varepsilon_{ij},$$

with corresponding matrices

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

Reparametrization

The columns of X are now linearly independent, and so this is a full rank model that we can analyse. Simple matrix calculations give us

$$X^TX = \begin{bmatrix} n_1 & 0 & 0 \\ 0 & n_2 & 0 \\ 0 & 0 & n_3 \end{bmatrix}, \quad (X^TX)^{-1} = \begin{bmatrix} \frac{1}{n_1} & 0 & 0 \\ 0 & \frac{1}{n_2} & 0 \\ 0 & 0 & \frac{1}{n_3} \end{bmatrix}$$

$$X^T\mathbf{y} = \begin{bmatrix} \sum_{i=1}^{n_1} y_{1i} \\ \sum_{i=1}^{n_2} y_{2i} \\ \sum_{i=1}^{n_3} y_{3i} \end{bmatrix}, \quad \mathbf{b} = (X^TX)^{-1}X^T\mathbf{y} = \begin{bmatrix} \sum_{i=1}^{n_1} y_{1i} \Big/ n_1 \\ \sum_{i=1}^{n_2} y_{2i} \Big/ n_2 \\ \sum_{i=1}^{n_3} y_{3i} \Big/ n_3 \end{bmatrix}.$$

Reparametrization

Therefore, the least squares estimates for each of the population means are the means of the samples drawn from that population:

$$\hat{\mu_i} = \bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}.$$

Linear functions of the parameters, of the form $\mathbf{t}^T\boldsymbol{\beta}$, are estimated using $\mathbf{t}^T\mathbf{b}$. For example, the function $\mu_1 - \mu_2$ is estimated by

$$\bar{y}_1 - \bar{y}_2 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_{1i} - \frac{1}{n_2}\sum_{i=2}^{n_2} y_{2i}.$$

Reparametrization

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 variance σ^2 (but different means). The estimator for this variance is

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

Reparametrization

For the example,

$$
\begin{aligned}
s^2 & \\
&= \frac{1}{n-3}\left[\sum_{i=1}^{3}\sum_{j=1}^{n_i} y_{ij}^2 - \left[\begin{array}{ccc} \sum_{i=1}^{n_1} y_{1i} & \sum_{i=1}^{n_2} y_{2i} & \sum_{i=1}^{n_3} y_{3i} \end{array}\right]\left[\begin{array}{c} \sum_{i=1}^{n_1} y_{1i}\Big/ n_1 \\ \sum_{i=1}^{n_2} y_{2i}\Big/ n_2 \\ \sum_{i=1}^{n_3} y_{3i}\Big/ n_3 \end{array}\right]\right] \\
&= \frac{1}{n-3}\left[\sum_{i=1}^{3}\sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{3}\frac{1}{n_i}\left(\sum_{j=1}^{n_i} y_{ij}\right)^2\right] \\
&= \frac{1}{n-3}\sum_{i=1}^{3}\left[\sum_{j=1}^{n_i} y_{ij}^2 - \frac{1}{n_i}\left(\sum_{j=1}^{n_i} y_{ij}\right)^2\right].
\end{aligned}
$$

Reparametrization

This can be written as a 'pooled' variance

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

where s_i^2 are the individual population variance estimators

$$s_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2 .$$

More generally, for a one-way classification model with k levels,

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

Reparametrization

In general, it is always possible to re-parameterise a less than full rank model into a full rank model.

However, this is not always desirable.

For the one-way classification model, we have a nice interpretation of the (re-)parameters as the population means. But this is not always possible.

Reparametrization

Example. Consider the *two-way* classification model (without interaction), with two levels of each factor:

$$y_{ij} = \underline{\mu} + \underline{\tau_i} + \underline{\beta_j} + \varepsilon_{ij}, i, j = 1, 2.$$

The design matrix for this model is

$$X = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}.$$

5 columns

$\text{rank}(X) = 3$

Reparametrization

It is obvious that the first column is the sum of the next two columns, and also the sum of the 4th and 5th columns. Thus $r(X) = 3$.

This means that we have to remove 2 parameters — which ones? This makes interpretability much harder!

Fortunately, we do not have to re-parameterise our models: we can develop theory for the less than full rank model.

Conditional inverses

X^TX inverse does not exist

We start with more linear algebra. Here we introduce the concept of conditional inverses.

Definition 6.1

Let A be a $\underbrace{n \times p}$ matrix. The $p \times n$ matrix A^c is called a *conditional inverse* for A if and only if

$$AA^cA = A.$$

Conditional inverses

$$AA^cA = A$$

If A is nonsingular, $A^{-1}AA^cA = A^{-1}A$

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

$\Rightarrow A^cA = I$

Example. Consider the (singular) matrices

$$A = \begin{bmatrix} 2 & 4 & 2 \\ 1 & 0 & -1 \\ 3 & 1 & -2 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{4} & -\frac{1}{2} & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Conditional inverses

We have

$$\begin{aligned}
\underline{AA_1A} &= \begin{bmatrix} 2 & 4 & 2 \\ 1 & 0 & -1 \\ 3 & 1 & -2 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{4} & -\frac{1}{2} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 & 4 & 2 \\ 1 & 0 & -1 \\ 3 & 1 & -2 \end{bmatrix} \\
&= \begin{bmatrix} 2 & 4 & 2 \\ 1 & 0 & -1 \\ 3 & 1 & -2 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \\
&= \begin{bmatrix} 2 & 4 & 2 \\ 1 & 0 & -1 \\ 3 & 1 & -2 \end{bmatrix} = \underline{A}.
\end{aligned}$$

Therefore A_1 is a conditional inverse for A.

Conditional inverses

$AA_2A = A$

But it can also be shown that

$$A_2 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & -3 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

is also a conditional inverse for A! So conditional inverses are *not* unique.

That is why we speak of *a* conditional inverse for A, not *the* conditional inverse for A.

Of course, if A is nonsingular, then the conditional inverse is uniquely the regular inverse. We can use this in the above example to show that A is singular.

Finding a conditional inverse

For a square matrix to have a regular inverse, it must satisfy some nonsingularity conditions. However, this is not the case for a conditional inverse.

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 nonsingular 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.*

Finding a conditional inverse

Proof. Let's assume M is the principal (top left) minor of A. We write

$$A = \left[\begin{array}{c|c} M & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right].$$

The procedure constructs a $p \times n$ matrix B which can be partitioned as

$$B = \left[\begin{array}{c|c} M^{-1} & 0 \\ \hline 0 & 0 \end{array}\right].$$

Finding a conditional inverse

Then we have

$$\begin{aligned} ABA &= \left[\begin{array}{c|c} M & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right] \left[\begin{array}{c|c} M^{-1} & 0 \\ \hline 0 & 0 \end{array}\right] \left[\begin{array}{c|c} M & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right] \\ &= \left[\begin{array}{c|c} I & 0 \\ \hline A_{21}M^{-1} & 0 \end{array}\right] \left[\begin{array}{c|c} M & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right] \\ &= \left[\begin{array}{c|c} M & A_{12} \\ \hline A_{21} & A_{21}M^{-1}A_{12} \end{array}\right]. \end{aligned}$$

We merely have to show that $A_{21}M^{-1}A_{12} = A_{22}$.

Finding a conditional inverse

This follows because $r(A)$ is the size of M; we can write all other columns of A as linear combinations of the first $r(A)$ columns. In other words, there exists a matrix R such that

$$\begin{aligned}
\left[\frac{A_{12}}{A_{22}}\right] &= \left[\frac{M}{A_{21}}\right] R \\
A_{12} &= MR \\
R &= M^{-1}A_{12} \\
A_{22} &= A_{21}R \\
&= A_{21}M^{-1}A_{12}.
\end{aligned}$$

Finding a conditional inverse

$$-\begin{pmatrix}2\\1\\3\end{pmatrix}+\begin{pmatrix}4\\0\\1\end{pmatrix}=\begin{pmatrix}2\\-1\\-2\end{pmatrix}$$

Example. From the previous example,

$$\underline{A} = \begin{bmatrix} 2 & 4 & 2 \\ 1 & 0 & -1 \\ 3 & 1 & -2 \end{bmatrix}.$$

It can be seen that $\underline{r(A) = 2}$, so we take the principal 2×2 minor

$$\underline{M = \begin{bmatrix} 2 & 4 \\ 1 & 0 \end{bmatrix}.}$$

$$(M^{-1})^T$$

Finding a conditional inverse

Then

Step 2
$$(M^{-1})^T = -\frac{1}{4}\begin{bmatrix} 0 & -4 \\ -1 & 2 \end{bmatrix}^T = \boxed{\begin{bmatrix} 0 & \frac{1}{4} \\ 1 & -\frac{1}{2} \end{bmatrix}}$$

and

Step 3
$$A^c = \begin{bmatrix} 0 & \frac{1}{4} & 0 \\ 1 & -\frac{1}{2} & 0 \\ 0 & 0 & 0 \end{bmatrix}^T = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{4} & -\frac{1}{2} & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

This is the conditional inverse A_1 of the earlier example, so we have seen that it works.

On the other hand, if we take the lower left 2×2 minor, following the procedure gives us A_2. So this procedure can produce more than one conditional inverse. □

Conditional inverse properties

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

- $\underline{r(A) = r(AA^c) = r(A^cA)}$;
- $(A^c)^T = (A^T)^c$; ~
- A^cA, AA^c, $I - A^cA$ and $I - AA^c$ are idempotent;
- $A = A(A^TA)^c(A^TA)$ and $A^T = (A^TA)(A^TA)^cA^T$.

$$\text{Pf: } r(A) \geq r(\underline{AA^c}) \geq r(\underline{AA^cA}) \underset{\uparrow \text{ Def}}{=} r(\underline{A}) \Rightarrow r(AA^c) = r(A)$$

Conditional inverse properties

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

- $A(A^TA)^cA^T$ is unique, symmetric ✓, and idempotent ✓;
- $r(A(A^TA)^cA^T) = r$;
- $I - A(A^TA)^cA^T$ is unique, symmetric and idempotent;
- $r(I - A(A^TA)^cA^T) = n - r$.

Conditional inverse properties

Proof.

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

$$
\begin{aligned}
A(A^TA)^cA^TA(A^TA)^cA^T &= [A(A^TA)^cA^TA]\,(A^TA)^cA^T \\
&\overset{\#4\text{ on slide }31}{=} A(A^TA)^cA^T. \quad ✓
\end{aligned}
$$

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

R Example

```
> library(MASS)
> A <- matrix(c(2,-6,3,1,6,4,-2,-1,0),3,3)
> det(A)
[1] 89
> A # non-singular
     [,1] [,2] [,3]
[1,]    2    1   -2
[2,]   -6    6   -1
[3,]    3    4    0
```

R Example

conditional inverse of A

```
> Ac <- ginv(A)
> Ac
```

```
            [,1]        [,2]      [,3]
[1,]  0.04494382 -0.08988764 0.1235955
[2,] -0.03370787  0.06741573 0.1573034
[3,] -0.47191011 -0.05617978 0.2022472
```

```
> solve(A)
```

inverse of A

```
            [,1]        [,2]      [,3]
[1,]  0.04494382 -0.08988764 0.1235955
[2,] -0.03370787  0.06741573 0.1573034
[3,] -0.47191011 -0.05617978 0.2022472
```

R Example

```
> A <- matrix(c(2,-6,3,1,6,4,3,0,7),3,3)
> det(A)
[1] 0
> A # singular
     [,1] [,2] [,3]
[1,]    2    1    3
[2,]   -6    6    0
[3,]    3    4    7
```

R Example

solve(A) ⟶ Error

```
> Ac <- ginv(A)
> Ac
            [,1]         [,2]       [,3]
[1,] 0.025713835 -0.084240416 0.03659883
[2,] 0.009149708  0.080454330 0.04369774
[3,] 0.034863543 -0.003786086 0.08029658
> round(A %*% Ac %*% A, 5)
     [,1] [,2] [,3]
[1,]    2    1    3
[2,]   -6    6    0
[3,]    3    4    7
```

M=

$AA^cA = A$

R Example

```
> Ac2 <- matrix(0,3,3)
> Ac2[1:2,1:2] <- t(solve(A[1:2,1:2]))   Step 2
> Ac2 <- t(Ac2)                          Step 3
> Ac2

          [,1]        [,2] [,3]
[1,] 0.3333333 -0.05555556    0
[2,] 0.3333333  0.11111111    0
[3,] 0.0000000  0.00000000    0

> A %*% Ac2 %*% A

     [,1] [,2] [,3]
[1,]    2    1    3
[2,]   -6    6    0
[3,]    3    4    7
```

R Example

$$A = \begin{pmatrix} 2 & 1 & 3 \\ -6 & 6 & 0 \\ 3 & 4 & 7 \end{pmatrix}$$

$$\underbrace{\begin{pmatrix} -6 & 6 \\ 3 & 4 \end{pmatrix}}_{M}$$

```
     > Ac3 <- matrix(0,3,3)
Step2> Ac3[2:3,1:2] <- t(solve(A[2:3,1:2]))
Step3> Ac3 <- t(Ac3)
     > Ac3

          [,1]        [,2]      [,3]
     [1,]    0 -0.09523810 0.1428571
     [2,]    0  0.07142857 0.1428571
     [3,]    0  0.00000000 0.0000000

     > A %*% Ac3 %*% A

          [,1] [,2] [,3]
     [1,]    2    1    3
     [2,]   -6    6    0
     [3,]    3    4    7
```

$$\begin{pmatrix} 0 & 0 & 0 \\ X & X & 0 \\ X & X & 0 \end{pmatrix}$$

R Example

```
> library(Matrix)
> rankMatrix(A)[1]
```

$r(A) = 2$

```
[1] 2
```

```
> rankMatrix(Ac2 %*% A)[1]
```

$r(A^cA) = 2$

```
[1] 2
```

```
> round(A %*% ginv(t(A) %*% A) %*% t(A) %*% A, 5)
```

```
     [,1] [,2] [,3]
[1,]    2    1    3
[2,]   -6    6    0
[3,]    3    4    7
```

#4 of slide 31

$A(A^TA)^cA^TA = A$

R Example

```
> A %*% ginv(t(A) %*% A) %*% t(A)

            [,1]        [,2]       [,3]
[1,]  0.16516801 -0.09938476 0.35778514
[2,] -0.09938476  0.98816848 0.04259347
[3,]  0.35778514  0.04259347 0.84666351

> AtAc2 <- matrix(0,3,3)
> AtAc2[1:2,1:2] <- solve((t(A) %*% A)[1:2,1:2])
> A %*% AtAc2 %*% t(A)

            [,1]        [,2]       [,3]
[1,]  0.16516801 -0.09938476 0.35778514
[2,] -0.09938476  0.98816848 0.04259347
[3,]  0.35778514  0.04259347 0.84666351
```

Solving the normal equations

Let us now solve the normal equations

$$X^T X\mathbf{b} = X^T\mathbf{y}.$$

First, we must make sure that they *have* a solution! (existence)

→ **Theorem 6.3**

have solution(s)

The system $A\mathbf{x} = \mathbf{g}$ is consistent if and only if the rank of $\left[\, A \,\middle|\, \mathbf{g} \,\right]$ is equal to the rank of A.

$$r([A, g]) \geq r(A)$$

Solving the normal equations

Proof. ($\Leftarrow$) Since $r([\ A \mid \mathbf{g}\ ]) = r(A)$, $\mathbf{g}$ must be a linear combination of the columns of A.

Therefore there exist constants $x_1, x_2, \ldots, x_p$ so that

$$x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \ldots + x_p\mathbf{a}_p = \mathbf{g}$$

where $\mathbf{a}_i$ is the ith column of A.

Solving the normal equations

But if we put this into matrix notation and set

$$\mathbf{x} = \boxed{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}},$$

then this is exactly the system $\underline{A\mathbf{x} = \mathbf{g}}$.

Therefore the system is consistent.

Solving the normal equations

Apply Thm 6.3 to

Theorem 6.4

In the general linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, the normal equations

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

are consistent.

$$\underline{r\left([X^TX, X^Ty]\right)} = r\left([X^TX]\right)$$

$$\geq$$

Solving the normal equations

✓

Proof. It is obvious that $r(X^TX) \le r([\ X^TX \mid X^T\mathbf{y}\ ])$, as adding a column cannot decrease the number of linearly independent columns.

we want to prove

$$r([X^TX, X^Ty]) \le r(X^TX)$$

However,

$$\begin{aligned}
r([\ X^TX \mid X^T\mathbf{y}\ ]) &\overset{✓}{=} r(\overset{=A}{X^T}[\overset{=B}{X \mid \mathbf{y}}]) \\
&\le r(X^T) \\
&= r(X^TX).
\end{aligned}$$

$$r(AB) \le r(A)$$

Theorem 6.3 now shows that the normal equations are consistent.

Solving the normal equations

$Ax = g$

A is nonsingular ⇒ $x = A^{-1}g$

A is singular ⇒ $x = A^c g$

Now that we know the normal equations always have a solution, how can we find one?

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.

We want to verify $\boxed{A(A^c g) = g}$

Proof. Since $A\mathbf{x}^* = \mathbf{g}$ for some $\mathbf{x}^*$,

$$AA^c\mathbf{g} \overset{g=Ax^*}{=} AA^c(A\mathbf{x}^*) \overset{def}{=} A\mathbf{x}^* = \mathbf{g}.$$

Therefore, $A^c\mathbf{g}$ solves the system.

Solving the normal equations

From this theorem, we see that

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

solves the normal equations, for any conditional inverse.

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

Solving the normal equations

Example. Suppose we have a one-way classification model with two classes and one sample from each class. The design matrix is

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

Supposing that $\mathbf{y}^T = [6, 8]$ we get

$$X^T X = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \quad X^T \mathbf{y} = \begin{bmatrix} 14 \\ 6 \\ 8 \end{bmatrix}.$$

Solving the normal equations

The normal equations are

$$\begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 14 \\ 6 \\ 8 \end{bmatrix}.$$

Since the first column of $X^T X$ is the sum of the next two, $X^T X$ is not of full rank. It is easy to see that $r(X^T X) = 2$.

Solving the normal equations

To find a conditional inverse of X^TX, we apply Theorem 6.2, using the nonsingular minor $\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$.

This gives

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

and therefore

$$\mathbf{b} = (X^TX)^cX^T\mathbf{y} = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 14 \\ 6 \\ 8 \end{bmatrix} = \begin{bmatrix} 8 \\ -2 \\ 0 \end{bmatrix}.$$

Solving the normal equations

However, using the minor $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ gives the conditional inverse

$$(X^T X)^c = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

which gives the solution

$$\mathbf{b} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 14 \\ 6 \\ 8 \end{bmatrix} = \begin{bmatrix} 0 \\ 6 \\ 8 \end{bmatrix}.$$

Both these solutions solve the normal equations, and are equally valid! This is the problem with the less than full rank model.

Carbon removal example

Example. Consider the earlier carbon removal example. We have

$$X^TX = \begin{bmatrix} 9 & 3 & 3 & 3 \\ 3 & 3 & 0 & 0 \\ 3 & 0 & 3 & 0 \\ 3 & 0 & 0 & 3 \end{bmatrix}$$

so a conditional inverse is

$$(X^TX)^c = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & \frac{1}{3} \end{bmatrix}.$$

Carbon removal example

We can also calculate

$$X^T\mathbf{y} = \begin{bmatrix} 303.3 \\ 105 \\ 117.9 \\ 80.4 \end{bmatrix}.$$

Using the conditional inverse above gives us a solution to the normal equations:

$$\mathbf{b} = (X^T X)^c X^T\mathbf{y} = \begin{bmatrix} 0 \\ 35 \\ 39.3 \\ 26.8 \end{bmatrix}.$$

Solving the normal equations

If the model is less than full rank, the normal equations have an infinite number of solutions.

Theorem 6.6

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

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

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

Solving the normal equations

Proof. We know that $A^c\mathbf{g}$ solves the system, so

$$\begin{aligned} A\mathbf{x} &= A\left[A^c\mathbf{g} + (I - A^cA)\mathbf{z}\right] \\ &= AA^c\mathbf{g} + (A - AA^cA)\mathbf{z} \\ &= \mathbf{g} + (A - A)\mathbf{z} = \mathbf{g}. \end{aligned}$$

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

$$\mathbf{b} = (X^TX)^cX^T\mathbf{y} + [I - (X^TX)^cX^TX]\mathbf{z}$$

satisfies the equations.

Solving the normal equations

Example. In the two-class example above, one solution to the normal equations was $(X^TX)^cX^T\mathbf{y} = \begin{bmatrix} 8 & -2 & 0 \end{bmatrix}^T$.

Using the same conditional inverse we have

$$(X^TX)^cX^TX = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix}.$$

Solving the normal equations

Then another solution to the normal equations is

$$\begin{aligned}
\mathbf{b} &= (X^TX)^cX^T\mathbf{y} + [I - (X^TX)^cX^TX]\mathbf{z} \\
&= \begin{bmatrix} 8 \\ -2 \\ 0 \end{bmatrix} + \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix} \right) \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} \\
&= \begin{bmatrix} 8 - z_3 \\ -2 + z_3 \\ z_3 \end{bmatrix}
\end{aligned}$$

for arbitrary z_3.

For example, $\begin{bmatrix} 7 \\ -1 \\ 1 \end{bmatrix}$ is a solution.

Solving the normal equations

The converse of the above theorem is also true: *all* solutions to the system can be expressed in this form.

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*,*

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

where $\mathbf{z} = \mathbf{x}_0$*.*

Solving the normal equations

Proof. Since $\mathbf{x}_0$ solves the system, we have

$$\begin{aligned} A^c\mathbf{g} + (I - A^cA)\mathbf{z} &= A^c\mathbf{g} + (I - A^cA)\mathbf{x}_0 \\ &= A^c\mathbf{g} + \mathbf{x}_0 - A^cA\mathbf{x}_0 \\ &= A^c\mathbf{g} + \mathbf{x}_0 - A^c\mathbf{g} = \mathbf{x}_0. \end{aligned}$$

For the normal equations, this means that any solution can be expressed as

$$\mathbf{b} = (X^TX)^cX^T\mathbf{y} + [I - (X^TX)^cX^TX]\mathbf{z}$$

for *any* conditional inverse $(X^TX)^c$, and some $\mathbf{z}$.

Solving the normal equations

Example. In the two-class example, we found the solution

$$\mathbf{b}_1 = \begin{bmatrix} 8 \\ -2 \\ 0 \end{bmatrix}$$

using our original conditional inverse.
But we also noted that the conditional inverse

$$(X^TX)_2^c = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

produces the solution

$$\mathbf{b}_2 = \begin{bmatrix} 0 \\ 6 \\ 8 \end{bmatrix}.$$

Solving the normal equations

Using the theorem, the first solution can be written in terms of the second solution:

$$\begin{aligned}
\mathbf{b}_1 &= (X^TX)_2^cX^T\mathbf{y} + (I - (X^TX)_2^cX^TX)\mathbf{z} \\
&= \begin{bmatrix} 0 \\ 6 \\ 8 \end{bmatrix} + \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \right) \begin{bmatrix} 8 \\ -2 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} 0 \\ 6 \\ 8 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 8 \\ -2 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} 8 \\ -2 \\ 0 \end{bmatrix}.
\end{aligned}$$

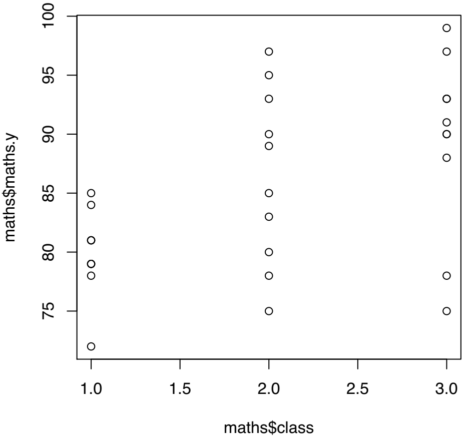
Exam marks example

We compare the marks of students in 3 different mathematics classes. There is another factor (IQ), but we ignore this for the time being.

```
> maths <- read.csv("../data/maths.csv")
> str(maths)

'data.frame':        30 obs. of  5 variables:
 $ X      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ maths.y: int  81 84 81 79 78 79 81 85 72 79 ...
 $ iq     : int  99 103 108 109 96 104 96 105 94 91 ...
 $ class  : int  1 1 1 1 1 1 1 1 1 1 ...
 $ class.f: int  1 1 1 1 1 1 1 1 1 1 ...

> maths$class.f <- factor(maths$class.f)
```

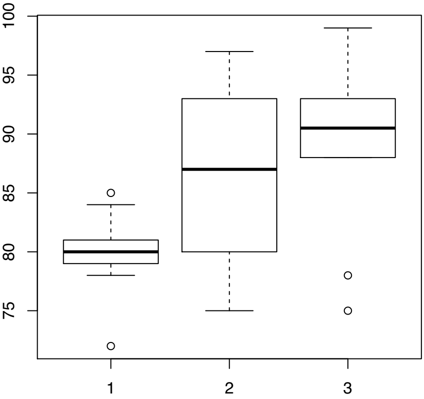
Exam marks example

```
> plot(maths$class, maths$maths.y)
```



Exam marks example

```
> plot(maths$class.f, maths$maths.y)
```



Exam marks example

```
> (y <- maths$maths.y)
 [1] 81 84 81 79 78 79 81 85 72 79 85 78 93 80 83 95 90 89
[26] 91 88 93 90 78
> n <- dim(maths)[1]
> k <- length(levels(maths$class.f))
> X <- matrix(0,n,k+1)
> X[,1] <- 1
> X[maths$class.f==1,2] <- 1
> X[maths$class.f==2,3] <- 1
> X[maths$class.f==3,4] <- 1
```

Exam marks example

```
> X
      [,1] [,2] [,3] [,4]
 [1,]    1    1    0    0
 [2,]    1    1    0    0
 [3,]    1    1    0    0
 [4,]    1    1    0    0
 [5,]    1    1    0    0
 [6,]    1    1    0    0
 [7,]    1    1    0    0
 [8,]    1    1    0    0
 [9,]    1    1    0    0
[10,]    1    1    0    0
[11,]    1    0    1    0
[12,]    1    0    1    0
[13,]    1    0    1    0
[14,]    1    0    1    0
[15,]    1    0    1    0
[16,]    1    0    1    0
[17,]    1    0    1    0
[18,]    1    0    1    0
[19,]    1    0    1    0
[20,]    1    0    1    0
```

Exam marks example: reparametrisation

```
> Xre <- X[,-1]
> (b <- solve(t(Xre) %*% Xre, t(Xre) %*% y))
     [,1]
[1,] 79.9
[2,] 86.5
[3,] 89.4
```

Exam marks example: reparametrisation

```
> modelre <- lm(y ~ 0 + X[,2] + X[,3] + X[,4])
> summary(modelre)

Call:
lm(formula = y ~ 0 + X[, 2] + X[, 3] + X[, 4])

Residuals:
   Min     1Q Median     3Q    Max
-14.40  -1.80   0.85   3.60  10.50

Coefficients:
       Estimate Std. Error t value Pr(>|t|)
X[, 2]   79.900      2.053   38.92   <2e-16 ***
X[, 3]   86.500      2.053   42.14   <2e-16 ***
X[, 4]   89.400      2.053   43.55   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.492 on 27 degrees of freedom
Multiple R-squared:  0.9948,       Adjusted R-squared:  0.9942
```

Exam marks example

Let's look at the normal equations.

```
> t(X) %*% X

     [,1] [,2] [,3] [,4]
[1,]   30   10   10   10
[2,]   10   10    0    0
[3,]   10    0   10    0
[4,]   10    0    0   10
```

```
> t(X) %*% y

     [,1]
[1,] 2558
[2,]  799
[3,]  865
[4,]  894
```

Exam marks example

```
> XtXc <- matrix(0,4,4)
> XtXc[2:4,2:4] <- solve((t(X) %*% X)[2:4,2:4])
> (b <- XtXc %*% t(X) %*% y)

     [,1]
[1,]  0.0
[2,] 79.9
[3,] 86.5
[4,] 89.4

> round(t(X) %*% X %*% b - t(X) %*% y, 3)

     [,1]
[1,]    0
[2,]    0
[3,]    0
[4,]    0
```

Exam marks example

```
> (b2 <- ginv(t(X) %*% X) %*% t(X) %*% y)
      [,1]
[1,] 63.95
[2,] 15.95
[3,] 22.55
[4,] 25.45
> round(t(X) %*% X %*% b2 - t(X) %*% y, 3)
     [,1]
[1,]    0
[2,]    0
[3,]    0
[4,]    0
```

Exam marks example

```
> I4 <- diag(4)
> z <- c(2,8,-2,1)
> (b3 <- b + (I4 - XtXc %*% t(X) %*% X) %*% z)

     [,1]
[1,]  2.0
[2,] 77.9
[3,] 84.5
[4,] 87.4

> round(t(X) %*% X %*% b3 - t(X) %*% y, 3)

     [,1]
[1,]    0
[2,]    0
[3,]    0
[4,]    0
```

Exam marks example

```
> b + (I4 - XtXc %*% t(X) %*% X) %*% b3
     [,1]
[1,]  2.0
[2,] 77.9
[3,] 84.5
[4,] 87.4
> b3
     [,1]
[1,]  2.0
[2,] 77.9
[3,] 84.5
[4,] 87.4
```

Estimability

Now we know how to solve the normal equations; furthermore, we know how to find *all* solutions for them.

But which solution(s) do we want?

Or rather, which solutions can we find?

Estimability

Some quantities do not change no matter what solutions we use for the normal equations. We call these quantities *estimable*.

A trivial example is the responses $\mathbf{y}$.

Definition 6.8

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

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

Estimability

Theorem 6.9

In the general linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\mathbf{t}^T\boldsymbol{\beta}$ is estimable if and only if there is a solution to the linear system $X^TX\mathbf{z} = \mathbf{t}$.

Proof. ($\Leftarrow$) Let $\mathbf{z}_0$ be a solution to $X^TX\mathbf{z} = \mathbf{t}$ and put $\mathbf{c} = X\mathbf{z}_0$.

Then

$$E[\mathbf{c}^T\mathbf{y}] = E[\mathbf{z}_0^TX^T\mathbf{y}] = \mathbf{z}_0^TX^TE[\mathbf{y}] = \mathbf{z}_0^TX^TX\boldsymbol{\beta} = \mathbf{t}^T\boldsymbol{\beta},$$

so $\mathbf{t}^T\boldsymbol{\beta}$ is estimable.

Estimability

Example. Consider our two-class example. We had

$$X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \quad X^T X = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}.$$

Consider the combination of parameters $\beta_1 - \beta_2$. This corresponds to $\mathbf{t}^T \boldsymbol{\beta}$ where

$$\mathbf{t} = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}.$$

Estimability

We look for a solution to the system

$$\begin{bmatrix} 2 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}.$$

This system has solution $z_1 = 0, z_2 = 1, z_3 = -1$, so $\beta_1 - \beta_2$ is estimable.

Estimability

Theorem 6.10

In the general linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\mathbf{t}^T\boldsymbol{\beta}$ is estimable if and only if

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

for some (and thus all) conditional inverse of (X^TX).

Proof. ($\Leftarrow$) Assume that $\mathbf{t}^T(X^TX)^cX^TX = \mathbf{t}^T$, so

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

This means that $(X^TX)^c\mathbf{t}$ is a solution to the system $X^TX\mathbf{z} = \mathbf{t}$, and Theorem 6.9 implies that $\mathbf{t}^T\boldsymbol{\beta}$ is estimable.

Estimability

($\Rightarrow$) Suppose that $\mathbf{t}^T\boldsymbol{\beta}$ is estimable. By Theorem 6.9, there exists a solution to the system $X^TX\mathbf{z} = \mathbf{t}$.

We know that a solution is $\mathbf{z} = (X^TX)^c\mathbf{t}$. (Note that the conditional inverse is arbitrary.)

In other words,

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

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

Estimability

Example. Consider the previous example. Let us take the conditional inverse

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

and consider again the quantity $\beta_1 - \beta_2$, which corresponds to $\mathbf{t} = \begin{bmatrix} 0 & 1 & -1 \end{bmatrix}^T$.

Estimability

Then

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

so again we see that $\beta_1 - \beta_2$ is estimable.

Estimability

On the other hand, suppose we take $\mathbf{t} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T$ so that $\mathbf{t}^T\boldsymbol{\beta} = \beta_0$.

Then we have

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

so β_0 is not estimable.

Estimability

Example. We return to the carbon removal example. We are interested in seeing if the three carbon removal treatments have (significantly) different means.

To test this, we look at the quantities $\tau_1 - \tau_2$ and $\tau_1 - \tau_3$.

If both of these are (close to) 0, then the treatments are not significantly different.

Estimability

We have

$$X^T X = 3\begin{bmatrix} 3 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \quad (X^T X)^c X^T X = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

and the coefficient vectors

$$\mathbf{t}_1 = \begin{bmatrix} 0 \\ 1 \\ -1 \\ 0 \end{bmatrix}, \quad \mathbf{t}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ -1 \end{bmatrix}.$$

Estimability

$$\mathbf{t}_1^T(X^TX)^cX^TX = \left[\begin{array}{cccc} 0 & 1 & -1 & 0 \end{array}\right] \left[\begin{array}{cccc} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{array}\right] = \left[\begin{array}{cccc} 0 & 1 & -1 & 0 \end{array}\right]$$

so $\mathbf{t}_1^T\boldsymbol{\beta} = \tau_1 - \tau_2$ is estimable.

$$\mathbf{t}_1^T(X^TX)^cX^TX = \left[\begin{array}{cccc} 0 & 1 & 0 & -1 \end{array}\right] \left[\begin{array}{cccc} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{array}\right] = \left[\begin{array}{cccc} 0 & 1 & 0 & -1 \end{array}\right]$$

so $\mathbf{t}_2^T\boldsymbol{\beta} = \tau_1 - \tau_3$ is also estimable.

Estimability

Next we will prove that no matter what conditional inverse we use, we will still generate the same estimate for an estimable quantity.

Theorem 6.11 (A Gauss-Markov Theorem)

In the general linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$*, suppose* $\mathbf{t}^T\boldsymbol{\beta}$ *is estimable. Then the best linear unbiased estimator (BLUE) for* $\mathbf{t}^T\boldsymbol{\beta}$ *is* $\mathbf{z}^T X^T\mathbf{y}$*, 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* $\mathbf{t}^T\mathbf{b}$*, where* $\mathbf{b}$ *is any solution to the normal equations.*

Estimability

Proof. We first show unbiasedness of the estimator.

$$\begin{array}{rcl} E[\mathbf{z}^T X^T \mathbf{y}] & = & \mathbf{z}^T X^T E[\mathbf{y}] \\ & = & \mathbf{z}^T X^T X \boldsymbol{\beta} \\ & = & \mathbf{t}^T \boldsymbol{\beta}. \end{array}$$

BLUEness is more involved, but is similar to the proof of Theorem 4.4.

Now suppose we have two solutions to the system $X^T X\mathbf{z} = \mathbf{t}$, called $\mathbf{z}_0$ and $\mathbf{z}_1$. Let $\mathbf{b}$ be any solution to the normal equations:

$$X^T X\mathbf{b} = X^T \mathbf{y}.$$

Estimability

The best linear unbiased estimator of $\mathbf{t}^T\boldsymbol{\beta}$ is

$$\mathbf{z}_0^T X^T \mathbf{y} = \mathbf{z}_0^T X^T X \mathbf{b} = (X^T X \mathbf{z}_0)^T \mathbf{b} = \mathbf{t}^T \mathbf{b}.$$

Since $\mathbf{b}$ is an arbitrary solution, this is the same no matter what solution we choose.

Similarly,

$$\mathbf{z}_1^T X^T \mathbf{y} = \mathbf{t}^T \mathbf{b} = \mathbf{z}_0^T X^T \mathbf{y}.$$

Thus the best linear unbiased estimator is unique, and equal to $\mathbf{t}^T\mathbf{b}$.

Estimability

Example. Let's look again at the two-class example. We know that $\beta_1 - \beta_2$ is estimable. We also know that solutions to the normal equations include

$$\mathbf{b} = \begin{bmatrix} 8 \\ -2 \\ 0 \end{bmatrix}, \quad \mathbf{b}_0 = \begin{bmatrix} 0 \\ 6 \\ 8 \end{bmatrix}.$$

To estimate $\beta_1 - \beta_2$, we can use

$$\mathbf{t}^T\mathbf{b} = \begin{bmatrix} 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 8 \\ -2 \\ 0 \end{bmatrix} = -2.$$

Estimability

However, from Theorem 6.11, we can also use

$$\mathbf{t}^T\mathbf{b}_0 = \begin{bmatrix} 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 6 \\ 8 \end{bmatrix} = -2.$$

This estimate is the same as the previous one, which follows from the theorem: *any* solution to the normal equation, using *any* conditional inverse, will produce exactly the same estimate.

In other words, the estimator is unique.

Estimability

Example. Back to the carbon removal example. We have shown that $\tau_1 - \tau_2$ and $\tau_1 - \tau_3$ are estimable. We estimate them by

$$\mathbf{t}_1^T\mathbf{b} = \begin{bmatrix} 0 & 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 35 \\ 39.3 \\ 26.8 \end{bmatrix} = -4.3$$

and

$$\mathbf{t}_2^T\mathbf{b} = \begin{bmatrix} 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 35 \\ 39.3 \\ 26.8 \end{bmatrix} = 8.2$$

respectively.

Again, no matter what conditional inverse we use, these estimates remain the same.

Estimability theorems

Theorem 6.12

In the linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, elements of $X\boldsymbol{\beta}$ are estimable.

Proof. We know that $E[\mathbf{y}] = X\boldsymbol{\beta}$. Now take $\mathbf{e}_i$ to be the ith standard basis vector.

We have

$$\begin{aligned}(X\boldsymbol{\beta})_i &= \mathbf{e}_i^T X\boldsymbol{\beta} \\ &= \mathbf{e}_i^T E[\mathbf{y}] \\ &= E[\mathbf{e}_i^T \mathbf{y}]\end{aligned}$$

and so the ith element of $X\boldsymbol{\beta}$ is estimable.

Estimability theorems

Example. Consider the carbon removal example. We have

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix}.$$

We know that we cannot estimate the parameter vector $\boldsymbol{\beta}$, because it is not uniquely determined.

Estimability theorems

However, the real quantities of interest are the mean responses from the three treatments. These are:

$$\begin{aligned}
\mu + \tau_1 &= \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix} \boldsymbol{\beta} \\
\mu + \tau_2 &= \begin{bmatrix} 1 & 0 & 1 & 0 \end{bmatrix} \boldsymbol{\beta} \\
\mu + \tau_3 &= \begin{bmatrix} 1 & 0 & 0 & 1 \end{bmatrix} \boldsymbol{\beta}
\end{aligned}$$

and each of these are elements of $X\boldsymbol{\beta}$. Therefore, they are estimable.

In a one-way classification model with any number of levels, $\mu + \tau_i$ is always estimable.

Estimability theorems

Theorem 6.13

Let $\mathbf{t}_1^T\boldsymbol{\beta}, \mathbf{t}_2^T\boldsymbol{\beta}, \ldots, \mathbf{t}_k^T\boldsymbol{\beta}$ be estimable functions, and let

$$z = a_1\mathbf{t}_1^T\boldsymbol{\beta} + a_2\mathbf{t}_2^T\boldsymbol{\beta} + \ldots + a_k\mathbf{t}_k^T\boldsymbol{\beta}.$$

Then z is estimable, and the best linear unbiased estimator for z is

$$a_1\mathbf{t}_1^T\mathbf{b} + a_2\mathbf{t}_2^T\mathbf{b} + \ldots + a_k\mathbf{t}_k^T\mathbf{b}.$$

Estimability theorems

Proof. By definition,

$$z = (a_1\mathbf{t}_1 + a_2\mathbf{t}_2 + \ldots + a_k\mathbf{t}_k)^T\boldsymbol{\beta}.$$

From Theorem 6.10,

$$\begin{aligned}
&(a_1\mathbf{t}_1 + a_2\mathbf{t}_2 + \ldots + a_k\mathbf{t}_k)^T(X^TX)^cX^TX \\
&\quad= a_1\mathbf{t}_1^T(X^TX)^cX^TX + a_2\mathbf{t}_2^T(X^TX)^cX^TX + \ldots + a_k\mathbf{t}_k^T(X^TX)^cX^TX \\
&\quad= a_1\mathbf{t}_1^T + a_2\mathbf{t}_2^T \ldots + a_k\mathbf{t}_k^T \\
&\quad= (a_1\mathbf{t}_1 + a_2\mathbf{t}_2 + \ldots + a_k\mathbf{t}_k)^T.
\end{aligned}$$

Therefore z is estimable, with BLUE

$$(a_1\mathbf{t}_1 + a_2\mathbf{t}_2 + \ldots + a_k\mathbf{t}_k)^T\mathbf{b}.$$

Estimability theorems

Of particular interest in many studies is the way different populations compare against each other. To attach a numerical value to these comparisons, we form linear combinations

$$a_1\tau_1 + a_2\tau_2 + \ldots + a_k\tau_k,$$

where $\sum_{i=1}^{k} a_i = 0$.

These *treatment contrasts* wipe out the effect of the overall mean response, to describe the differences between populations.

Estimability theorems

In a one-way classification model, any treatment contrast is estimable.

If

$$z = a_1\tau_1 + a_2\tau_2 + \ldots + a_k\tau_k$$

is a treatment contrast, then

$$\begin{aligned} z &= \sum_{i=1}^{k} a_k\mu + a_1\tau_1 + a_2\tau_2 + \ldots + a_k\tau_k \\ &= a_1(\mu + \tau_1) + a_2(\mu + \tau_2) + \ldots + a_k(\mu + \tau_k) \end{aligned}$$

is a linear combination of the estimable functions $\mu + \tau_i$, and is therefore estimable.

Estimability theorems

Of particular interest among treatment contrasts is the contrast of the form $\tau_i - \tau_j$, for some $i \neq j$. This is because

$$\tau_i - \tau_j = (\mu + \tau_i) - (\mu + \tau_j)$$

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

We would expect to estimate this contrast by the corresponding difference in sample means, $\bar{y}_i - \bar{y}_j$. We can show using the theory we have developed that this is in fact the case.

Estimability theorems

Example. We do this for $k = 3$ and the contrast $\tau_1 - \tau_2$. Our matrices are

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ y_{31} \\ \vdots \\ y_{3n_3} \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix}.$$

Estimability theorems

Direct multiplication gives

$$X^T\mathbf{y} = \begin{bmatrix} \sum_{i=1}^{3}\sum_{j=1}^{n_i} y_{ij} \\ \sum_j y_{1j} \\ \sum_j y_{2j} \\ \sum_j y_{3j} \end{bmatrix}, \quad X^TX = \begin{bmatrix} n & n_1 & n_2 & n_3 \\ n_1 & n_1 & 0 & 0 \\ n_2 & 0 & n_2 & 0 \\ n_3 & 0 & 0 & n_3 \end{bmatrix}.$$

Using the conditional inverse algorithm on the lower right corner of X^TX gives

$$(X^TX)^c = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{n_1} & 0 & 0 \\ 0 & 0 & \frac{1}{n_2} & 0 \\ 0 & 0 & 0 & \frac{1}{n_3} \end{bmatrix}.$$

Estimability theorems

Therefore a solution to the normal equations is

$$\mathbf{b} = (X^TX)^cX^T\mathbf{y} = \begin{bmatrix} 0 \\ \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \end{bmatrix}.$$

We have $\tau_1 - \tau_2 = \begin{bmatrix} 0 & 1 & -1 & 0 \end{bmatrix} \boldsymbol{\beta}$, so the best linear unbiased estimator for $\tau_1 - \tau_2$ is

$$\begin{bmatrix} 0 & 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \end{bmatrix} = \bar{y}_1 - \bar{y}_2.$$

If we took any conditional inverse, we would get the same result.

Exam marks example

We return to the maths dataset. Recall that b, b2 and b3 are all solutions to the normal equations.

```
> (tt <- c(0,1,-1,0))

[1]  0  1 -1  0

> round(tt %*% XtXc %*% t(X) %*% X, 5) # estimable

     [,1] [,2] [,3] [,4]
[1,]    0    1   -1    0

> (tt2 <- c(1,1,1,1))

[1] 1 1 1 1

> tt2 %*% XtXc %*% t(X) %*% X # not estimable

     [,1] [,2] [,3] [,4]
[1,]    3    1    1    1
```

Exam marks example

```
> tt %*% b
     [,1]
[1,] -6.6
> tt %*% b2
     [,1]
[1,] -6.6
> tt %*% b3
     [,1]
[1,] -6.6
> mean(maths$maths.y[maths$class.f==1]) -
+      mean(maths$maths.y[maths$class.f==2])
[1] -6.6
```

Exam marks example

```
> tt2 %*% b
      [,1]
[1,] 255.8
> tt2 %*% b2
      [,1]
[1,] 127.9
> tt2 %*% b3
      [,1]
[1,] 251.8
```

Exam marks example

For the less than full rank model, R uses contrasts for its tests. The two main contrast sets are `contr.treatment` and `contr.sum`. For the one-way classification model:

Label	`contr.treatment`	`contr.sum`
Intercept	μ_1	$\bar{\mu}$
factor1		$\mu_1 - \bar{\mu}$
factor2	$\mu_2 - \mu_1$	$\mu_2 - \bar{\mu}$
factor3	$\mu_3 - \mu_1$	$\mu_3 - \bar{\mu}$
$\vdots$	$\vdots$	$\vdots$
factor(k-1)	$\mu_{k-1} - \mu_1$	$\mu_{k-1} - \bar{\mu}$
factor(k)	$\mu_k - \mu_1$	

Exam marks example

In terms of our parameters:

Label	`contr.treatment`	`contr.sum`
Intercept	$\mu + \tau_1$	$\mu + \frac{1}{k}\sum \tau_i$
factor1		$\tau_1 - \frac{1}{k}\sum \tau_i$
factor2	$\tau_2 - \tau_1$	$\tau_2 - \frac{1}{k}\sum \tau_i$
factor3	$\tau_3 - \tau_1$	$\tau_3 - \frac{1}{k}\sum \tau_i$
$\vdots$	$\vdots$	$\vdots$
factor(k-1)	$\tau_{k-1} - \tau_1$	$\tau_{k-1} - \frac{1}{k}\sum \tau_i$
factor(k)	$\tau_k - \tau_1$	

Exam marks example

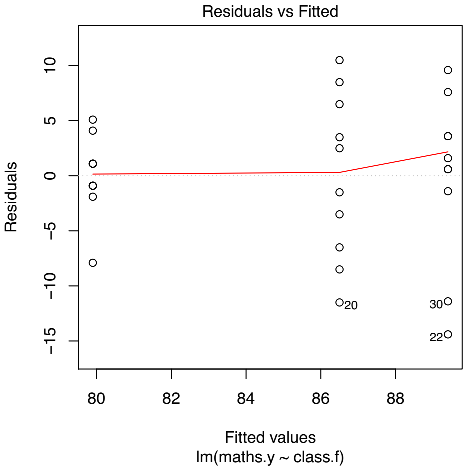
```
> contrasts(maths$class.f) <- contr.treatment(k)
> model <- lm(maths.y ~ class.f, data = maths)
> summary(model)
Call:
lm(formula = maths.y ~ class.f, data = maths)

Residuals:
   Min     1Q Median     3Q    Max
-14.40  -1.80   0.85   3.60  10.50

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   79.900      2.053  38.922  < 2e-16 ***
class.f2       6.600      2.903   2.273  0.03117 *
class.f3       9.500      2.903   3.272  0.00292 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.492 on 27 degrees of freedom
```

Exam marks example

```
> contrasts(maths$class.f) <- contr.sum(k)
> model2 <- lm(maths.y ~ class.f, data = maths)
> summary(model2)

Call:
lm(formula = maths.y ~ class.f, data = maths)

Residuals:
   Min     1Q Median     3Q    Max
-14.40  -1.80   0.85   3.60  10.50

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   85.267      1.185  71.943  < 2e-16 ***
class.f1      -5.367      1.676  -3.202  0.00348 **
class.f2       1.233      1.676   0.736  0.46818
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.492 on 27 degrees of freedom
```

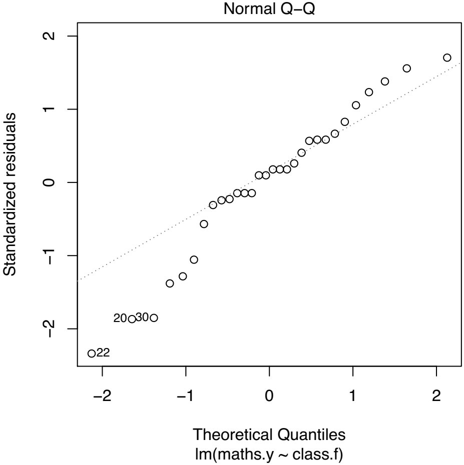
Exam marks example

```
> plot(model, which=1)
```



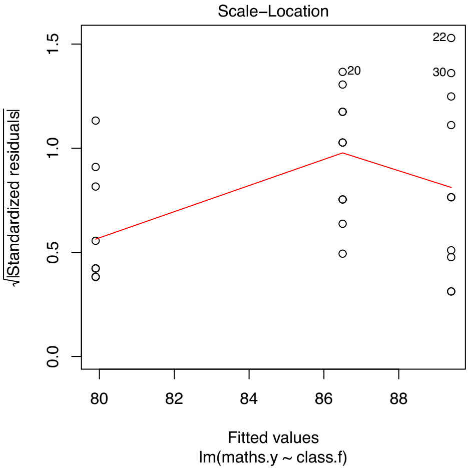
Exam marks example

```
> plot(model, which=2)
```



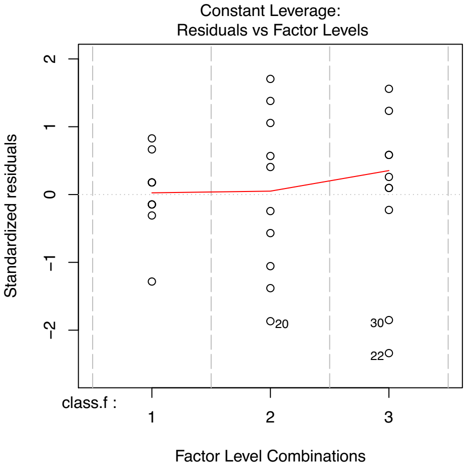
Exam marks example

```
> plot(model, which=3)
```



Exam marks example

```
> plot(model, which=5)
```



Estimating σ^2 in the less than full rank model

In the full rank model, we estimated σ^2 by

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

where n is the sample size, p is the number of parameters, and SS_{Res} is the sum of squares of the residuals:

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

Estimating σ^2 in the less than full rank model

For the less than full rank model we can still define the residual sum of squares as

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

where $\mathbf{b}$ is any solution to the normal equations.

Although $\mathbf{b}$ can vary, $X\mathbf{b}$ will not, because $X\boldsymbol{\beta}$ is estimable. Therefore SS_{Res} is invariant to the choice of $\mathbf{b}$.

Estimating σ^2 in the less than full rank model

Theorem 6.14

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

Proof. Let $\mathbf{b} = (X^TX)^cX^T\mathbf{y}$. Then

$$\begin{aligned}
SS_{Res} &= (\mathbf{y}^T - \mathbf{b}^TX^T)(\mathbf{y} - X\mathbf{b}) \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^TX\mathbf{b} + \mathbf{b}^TX^TX\mathbf{b} \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^TX(X^TX)^cX^T\mathbf{y} + \mathbf{y}^TX(X^TX)^cX^TX(X^TX)^cX^T\mathbf{y} \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^TX(X^TX)^cX^T\mathbf{y} + \mathbf{y}^TX(X^TX)^cX^T\mathbf{y} \\
&= \mathbf{y}^T[I - X(X^TX)^cX^T]\mathbf{y}.
\end{aligned}$$

Estimating σ^2 in the less than full rank model

How do we find an estimator for σ^2?

Let's consider SS_{Res} again. Take $H = X(X^TX)^cX^T$ and remember that $HX = X$.

$$\begin{aligned} E[SS_{Res}] &= E[\mathbf{y}^T(I-H)\mathbf{y}] \\ &= tr(I-H)\sigma^2 + (X\boldsymbol{\beta})^T(I-H)X\boldsymbol{\beta} \\ &= tr(I-H)\sigma^2 + \boldsymbol{\beta}^TX^TX\boldsymbol{\beta} - \boldsymbol{\beta}^TX^THX\boldsymbol{\beta} \\ &= tr(I-H)\sigma^2 + \boldsymbol{\beta}^TX^TX\boldsymbol{\beta} - \boldsymbol{\beta}^TX^TX\boldsymbol{\beta} \\ &= tr(I-H)\sigma^2. \end{aligned}$$

Estimating σ^2 in the less than full rank model

Since $I - H$ is symmetric and idempotent, we have

$$E[SS_{Res}] = r(I - H)\sigma^2 = (n - r)\sigma^2,$$

where $r = r(X)$, the rank of X.

Theorem 6.15

In the general linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, suppose X has rank r and $\boldsymbol{\varepsilon}$ has mean $\mathbf{0}$ and variance $\sigma^2 I$. Then an unbiased estimator for σ^2 is

$$\frac{SS_{Res}}{n - r}.$$

Estimating σ^2 in the less than full rank model

Example. We return to the carbon removal example.

```
> y <- c(34.6,35.1,35.3,38.8,39.0,40.1,26.7,26.7,27.0)
> X <- matrix(c(rep(1,9),rep(0,27)),9,4)
> X[1:3,2] <- 1
> X[4:6,3] <- 1
> X[7:9,4] <- 1
> X
      [,1] [,2] [,3] [,4]
 [1,]    1    1    0    0
 [2,]    1    1    0    0
 [3,]    1    1    0    0
 [4,]    1    0    1    0
 [5,]    1    0    1    0
 [6,]    1    0    1    0
 [7,]    1    0    0    1
 [8,]    1    0    0    1
```

Estimating σ^2 in the less than full rank model

```
> (b <- ginv(t(X)%*%X)%*%t(X)%*%y)

       [,1]
[1,] 25.275
[2,]  9.725
[3,] 14.025
[4,]  1.525

> e <- y - X%*%b
> (SSRes <- sum(e^2))

[1] 1.3

> (s2 <- SSRes/(9-3))

[1] 0.2166667
```

Exam marks example

```
> library(Matrix)
> (SSRes <- sum((y-X%*%b)^2))

[1] 1137.8

> sum(y^2) - t(y) %*% X %*% XtXc %*% t(X) %*% y

       [,1]
[1,] 1137.8

> (s2 <- SSRes/(n - rankMatrix(X)[1]))

[1] 42.14074
```

Exam marks example

```
> deviance(model)
[1] 1137.8
> deviance(model)/model$df.residual
[1] 42.14074
```

Interval estimation in the less than full rank model

We can find point estimates for estimable quantities. The next step is to try and find confidence intervals for them.

For the Gauss-Markov theorem we only required that ε has mean $\mathbf{0}$ and variance $\sigma^2 I$. However, to find confidence intervals, we need some idea of the distribution of the variables, so we suppose that $\boldsymbol{\varepsilon} \sim MVN(\mathbf{0}, \sigma^2 I)$.

Interval estimation in the less than full rank model

Recall that in the full rank model, we generated confidence intervals by finding a t-distributed quantity, which was created by dividing a normal variable by (the square root of) a χ^2 variable.

The χ^2 variable was

$$\frac{SS_{Res}}{\sigma^2},$$

which had $n - p$ degrees of freedom.

The σ^2 term was not known, but cancelled out another σ^2 term in the numerator to leave us with something that we could calculate.

We can proceed in a similar manner for the less than full rank model.

Interval estimation in the less than full rank model

Theorem 6.16

In the general linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$*, assume* $\boldsymbol{\varepsilon} \sim MVN(\mathbf{0}, \sigma^2 I)$*. Then*

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

has a χ^2 *distribution with* $n-r$ *degrees of freedom.*

Theorem 6.17

In the general linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$*, assume* $\boldsymbol{\varepsilon} \sim MVN(\mathbf{0}, \sigma^2 I)$*. If* $\mathbf{t}^T\boldsymbol{\beta}$ *is estimable, then* $\mathbf{t}^T\mathbf{b}$ *is independent of* s^2*.*

Interval estimation in the less than full rank model

The steps to derive a confidence interval are very similar to that for the full rank case, but with two small differences. Firstly, we can only find confidence intervals for quantities that are estimable!

Secondly, we replace the inverse $(X^TX)^{-1}$ by the conditional inverse $(X^TX)^c$.

All other steps are the same.

Interval estimation in the less than full rank model

We have

$$\begin{aligned}\text{Var } \mathbf{t}^T\mathbf{b} &= \text{Var } \mathbf{t}^T(X^TX)^cX^T\mathbf{y}\\ &= \mathbf{t}^T(X^TX)^cX^T\sigma^2IX(X^TX)^c\mathbf{t}\\ &= \sigma^2\mathbf{t}^T(X^TX)^c\mathbf{t}.\end{aligned}$$

Thus

$$\frac{(\mathbf{t}^T\mathbf{b}-\mathbf{t}^T\boldsymbol{\beta})/\sigma\sqrt{\mathbf{t}^T(X^TX)^c\mathbf{t}}}{\sqrt{s^2/\sigma^2}}$$

has a t distribution with $n-r$ degrees of freedom.

Interval estimation in the less than full rank model

This gives us the confidence interval for the (estimable) quantity $\mathbf{t}^T\boldsymbol{\beta}$, using a t distribution with $n - r$ degrees of freedom:

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

This formula can also be used to find confidence intervals for the individual parameters, if they are estimable.

Interval estimation in the less than full rank model

Example. We return again to the carbon removal example. Suppose we want to find a 95% confidence interval for $\tau_1 - \tau_2$.

```
> (tt <- c(0,1,-1,0))

[1]  0  1 -1  0

> ta <- qt(0.975,9-3)
> halfwidth <- ta*sqrt(s2*t(tt)%*%ginv(t(X)%*%X)%*%tt)
> tt%*%b + c(-1,1)*halfwidth

[1] -5.22997 -3.37003
```

In particular, we can say with 95% confidence that the the first carbon removal treatment is not as effective as the second.

Interval estimation in the less than full rank model

Example. We showed earlier that in a 3-level 1-way classification model, the contrast $\tau_1 - \tau_2$ can be estimated by the difference in the respective population means, $\bar{y}_1 - \bar{y}_2$.

We also had

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

Interval estimation in the less than full rank model

Therefore we have

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

and the confidence interval is

$$\bar{y}_1 - \bar{y}_2 \pm t_{\alpha/2} s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

You may have seen this formula before. The linear models framework has allowed us to derive it from first principles.

Exam marks example

We find a confidence interval for the estimable quantity $\mu + \tau_1$, the mean mark of class 1.

```
> tt <- as.vector(c(1,1,0,0))
> halfwidth <- qt(0.975,df=n-k)*sqrt(s2*t(tt)%*%XtXc%*%tt)
> tt %*% b + c(-1,1)*halfwidth

[1] 75.68796 84.11204
```

```
> newdata <- data.frame(class.f=factor(1))
> predict(model, newdata, interval="confidence", level=0.95)

   fit      lwr      upr
1 79.9 75.68796 84.11204
```

Exam marks example

We find a *prediction* interval for a new student from class 1.

```
> tt <- as.vector(c(1,1,0,0))
> halfwidth <- qt(0.975,df=n-k)*sqrt(s2)*
+       sqrt(1+t(tt)%*%XtXc%*%tt)
> tt %*% b + c(-1,1)*halfwidth

[1] 65.93024 93.86976
```

```
> newdata <- data.frame(class.f=factor(1))
> predict(model, newdata, interval="prediction", level=0.95)

   fit      lwr      upr
1 79.9 65.93024 93.86976
```

Exam marks example

We now find a confidence interval for the estimable quantity $\tau_1 - \tau_2$, the difference between the first two classes.

```
> tt <- as.vector(c(0,1,-1,0))
> halfwidth <- qt(0.975,df=n-k)*sqrt(s2*t(tt)%*%XtXc%*%tt)
> tt %*% b + c(-1,1)*halfwidth

[1] -12.5567252  -0.6432748
```

```
> confint(model, level=0.95)

                2.5 %   97.5 %
(Intercept) 75.6879592 84.11204
class.f2     0.6432748 12.55673
class.f3     3.5432748 15.45673
```

Exam marks example

We have to express more obscure parameter combinations relative to the treatment contrasts used. Remember:

Label	contr.treatment	contr.sum
Intercept	$\mu + \tau_1$	$\mu + \frac{1}{3}\sum \tau_i$
class.f1		$\tau_1 - \frac{1}{3}\sum \tau_i$
class.f2	$\tau_2 - \tau_1$	$\tau_2 - \frac{1}{3}\sum \tau_i$
class.f3	$\tau_3 - \tau_1$	

So for contr.treatment, $\tau_1 - \tau_2 = -$class.f2.

```
> library(gmodels)
> ci <- estimable(model, c(0,-1,0), conf.int=0.95)
> c(ci$Lower, ci$Upper)

[1] -12.5567252  -0.6432748
```

Exam marks example

For the `contr.sum` model, we have

$$\begin{aligned}
\texttt{Intercept} &= \mu + \frac{1}{3}(\tau_1 + \tau_2 + \tau_3) \\
\texttt{class.f1} &= \frac{2}{3}\tau_1 - \frac{1}{3}(\tau_2 + \tau_3) \\
\texttt{class.f2} &= \frac{2}{3}\tau_2 - \frac{1}{3}(\tau_1 + \tau_3) \\
\tau_1 - \tau_2 &= \texttt{class.f1} - \texttt{class.f2}
\end{aligned}$$

```
> ci2 <- estimable(model2, c(0,1,-1), conf.int=0.95)
> c(ci2$Lower, ci2$Upper)

[1] -12.5567252  -0.6432748
```

Exam marks example

To find the difference between class 3 and the average of the other two classes, we need

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

```
> ci3 <- estimable(model, c(0,-0.5,1), conf.int=0.95)
> c(ci3$Lower, ci3$Upper)

[1]  1.041325 11.358675
```

Exam marks example

For contr.sum:

$$\begin{aligned}\texttt{class.f1} + \texttt{class.f2} &= \frac{1}{3}\tau_1 + \frac{1}{3}\tau_2 - \frac{2}{3}\tau_3 \\ \tau_3 - \frac{1}{2}\tau_2 - \frac{1}{2}\tau_1 &= -\frac{3}{2}\texttt{class.f1} - \frac{3}{2}\texttt{class.f2}\end{aligned}$$

```
> ci4 <- estimable(model2, c(0,-1.5,-1.5), conf.int=0.95)
> c(ci4$Lower, ci4$Upper)

[1]  1.041325 11.358675
```