2.1 Vectors in $\mathbb{R}^3$

- vector $y$ in $\mathbb{R}^3$: a list of 3 numbers, eg $y = \begin{bmatrix} 1 \\ 5 \\ 6 \end{bmatrix} \in \mathbb{R}^3$

- squared length of $y = y_1^2 + y_2^2 + y_3^2 = y' y$

- space spanned by the vectors $X_1$ and $X_2$:
  - the (2D) plane that includes the vectors $X_1$ and $X_2$
  - the set of linear combinations of $X_1$ and $X_2$; i.e., vectors of the form $aX_1 + bX_2 \quad \forall \ a, b$
  - referred to as the linear manifold generated by $X_1$ and $X_2$; denoted by $\mathcal{M}(X_1, X_2)$

- vector in $\mathcal{M}(X_1, X_2)$ closest to $y$:
  - the orthogonal projection of $y$ onto $\mathcal{M}$, denoted by $\hat{y}$
  - $\hat{y}$ minimizes the squared length of $(y - \hat{y}) = \sum_{i=1}^{3} (y_i - \hat{y}_i)^2$
2.1 Vectors in $\mathbb{R}^3$

Example

\[ y \sim = \begin{bmatrix} 1 \\ 5 \\ 6 \end{bmatrix} ; \ X_1 \sim = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \ X_2 = \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix} \]

\[ \hat{y} \sim = \begin{bmatrix} 1.8 \\ 4.0 \\ 6.4 \end{bmatrix} = 2.2 \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} - 0.2 \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix} \]

2.2 Least squares in $\mathbb{R}^n$

Response data:

\[ y \sim = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n \]

Predictors: (columns of the design matrix $X$)

\[ X = [X_1 \ X_2 \ \ldots \ X_k] \text{ where each } X_j \in \mathbb{R}^n \]

Estimation space $[M(X)]$: space spanned by the ($k$) column vectors of $X$

\[ = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k = [X_1 \ \ldots \ X_k] \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} = X \hat{\beta} \text{ for all possible } \hat{\beta}. \]
2.2 Least squares in $\mathbb{R}^n$

Dimension of $\mathcal{M}(X) = r \leq k$, with $<$ if and only if some $X_i$ lie in the space generated by the other columns of $X$.

$$r = \text{rank}(X) = \text{rank}(X'X)$$

Squared length of $\hat{y} = y_1^2 + \ldots + y_n^2 = \sum_{i=1}^n y_i^2 = y'\hat{y}$

Least squares estimation

Find $\hat{y}$ belonging to $\mathcal{M}(X)$ that minimizes

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = (y - \hat{y})'(y - \hat{y})$$

$\hat{y}$ is the orthogonal projection of $y$ onto $\mathcal{M}(X)$

Since $\hat{y} \in \mathcal{M}(X)$,

$$\hat{y} = \beta_1 X_1 + \ldots + \beta_k X_k$$

$$= X \hat{\beta}$$ for some $\hat{\beta}$

Sums of squares, and ANOVA

Various sums of squares are usually arranged in an Analysis of Variance (ANOVA) table:

<table>
<thead>
<tr>
<th>Source (vector)</th>
<th>df (dimension)</th>
<th>SS (length$^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model ($\hat{y}$)</td>
<td>$r$</td>
<td>$\sum_{i=1}^n \hat{y}_i^2$</td>
</tr>
<tr>
<td>Residual ($y - \hat{y}$)</td>
<td>$n - r$</td>
<td>$\sum_{i=1}^n (y_i - \hat{y}_i)^2$</td>
</tr>
<tr>
<td>Total ($y$)</td>
<td>$n$</td>
<td>$\sum_{i=1}^n y_i^2$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source (vector)</th>
<th>df (dimension)</th>
<th>SS (length$^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model* ($\hat{y} - \bar{y}$)</td>
<td>$r - 1$</td>
<td>$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$</td>
</tr>
<tr>
<td>Residual* ($y - \hat{y}$)</td>
<td>$n - r$</td>
<td>$\sum_{i=1}^n (y_i - \hat{y}_i)^2$</td>
</tr>
<tr>
<td>Total* ($y - \bar{y}$)</td>
<td>$n - 1$</td>
<td>$\sum_{i=1}^n (y_i - \bar{y})^2$</td>
</tr>
</tbody>
</table>

The model must include an intercept for the RHS to sum appropriately.

By Pythagoras’s theorem, the squared lengths of orthogonal vectors sum to the squared length of the sum of the vectors.

Two vectors $u$ and $v$ are orthogonal iff $u'v = \sum_{i=1}^n u_i v_i = 0$. 
2.3 The Normal Equations

Since $\hat{y} \in \mathcal{M}(X)$, we can write $\hat{y} = X\hat{\beta}$
Further, $(y - \hat{y}) \perp \mathcal{M}(X)$, hence

$$X'(y - \hat{y}) = 0, \quad i = 1, \ldots, k$$

$\Rightarrow X'(y - \hat{y}) = 0$

Substituting $\hat{y} = X\hat{\beta}$ then gives

$$X'(y - X\hat{\beta}) = 0$$
$$\Rightarrow X'y - X'X\hat{\beta} = 0$$
$$\Rightarrow X'\hat{\beta} = X'y \text{ the normal equations}$$

2.3 The Normal Equations

If $X'X$ is non-singular, then $\hat{\beta} = (X'X)^{-1}X'y$.

Hence $\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y = Hy$, where

$H = X(X'X)^{-1}X'$

is known as the 'hat' matrix — $H$ transforms $y$ into $\hat{y}$.

2.4 Estimability

What if some $X_i$ lie in the space spanned by the other $X_i$?
For example, what if $X_3$ can be expressed as a linear combination of $X_1$ and $X_2$?

Then:
- the dimension of $\mathcal{M}(X) = r < k$, where $k$ is the number of columns of $X$
- there are (infinitely) many ways to express $\hat{y}$ as a linear combination of the $X_i$'s, i.e., there are infinitely many values of $\hat{\beta}$, infinitely many solutions of the normal equations
- $X'X$ is singular, hence $(X'X)^{-1}$ does not exist
2.4 Estimability

- **BUT** the following are unique:
  - \( \hat{y} \), the fitted values
  - \((y - \hat{y})\), the residuals
  - \(d(\hat{\beta})\), the deviance (residual sum of squares)
  - some linear functions of \( \hat{\beta}; (P\hat{\beta}) \); such linear functions \( (P\hat{\beta}) \) are said to be **estimable**

If \( X \) is of **full** rank, then \( \hat{\beta} = (X'X)^{-1}X'y \) is unique and hence \( (P\hat{\beta}) \) is estimable for all \( P \).

---

**Exercise**

Give the rank of the matrix \( X \):

(a) \[
\begin{bmatrix}
1 & 2 \\
1 & 8 \\
1 & -1 \\
1 & 3 \\
1 & 6
\end{bmatrix}
\]

(b) \[
\begin{bmatrix}
1 & 0 \\
1 & 0 \\
0 & 1 \\
0 & 1 \\
0 & 1
\end{bmatrix}
\]

(c) \[
\begin{bmatrix}
1 & 1 & 0 \\
1 & 1 & 0 \\
1 & 1 & 0 \\
1 & 0 & 1 \\
1 & 0 & 1 \\
1 & 0 & 1
\end{bmatrix}
\]

---

**Exercise**

One-way ANOVA with 2 groups.

<table>
<thead>
<tr>
<th>Group</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>5</td>
</tr>
</tbody>
</table>

\[ y \sim \begin{bmatrix} 0 \\ 4 \\ 7 \\ 5 \end{bmatrix} \]

**Model 1:** \( y_i = \begin{cases} 
\mu_1 \text{ in group 1} \\
\mu_2 \text{ in group 2} + e_i 
\end{cases} \)

\[ X = \begin{bmatrix} \end{bmatrix} \]

\[ \hat{\mu}_1 = \hat{\mu}_2 = \]
2.4 Estimability

Model 2: \[ y_i = \mu + \begin{cases} 
\alpha_1 & \text{in group 1} \\
\alpha_2 & \text{in group 2} \\
e_i 
\end{cases} \]

\[ X = \begin{pmatrix} 
\ldots \end{pmatrix} \]

\[ \begin{bmatrix} 
\hat{\beta} 
\end{bmatrix} \sim \begin{bmatrix} 
\ldots 
\end{bmatrix} \text{ or } \begin{bmatrix} 
\ldots 
\end{bmatrix} \text{ or } \begin{bmatrix} 
\ldots 
\end{bmatrix} \text{ or } \ldots \]

all of which give

\[ \hat{y} = X\hat{\beta} = \begin{pmatrix} 
2 \\
2 \\
6 \\
6 
\end{pmatrix} \]

2.4 Estimability

Which of the following functions of the parameters are estimable?

\[
\begin{array}{cccc}
\mu & \mu + \alpha_1 & \mu - \alpha_1 & 2\mu + \alpha_1 + \alpha_2 \\
\alpha_1 & \mu + \alpha_2 & \mu - \alpha_2 & 2\mu + \alpha_1 - \alpha_2 \\
\alpha_2 & \alpha_1 + \alpha_2 & \alpha_1 - \alpha_2 & 7\mu + 9\alpha_1 - 2\alpha_2 \\
\end{array}
\]

Non-full rank \( X \) and non-estimable functions

▶ If \( X \) is not of full rank, omit columns (impose linear restrictions on parameters) until it is of full rank \( (X^* \text{ say}) \), then use

\[ \hat{\beta}^* = (X^*X^*)^{-1}X^*y \]

where \( \hat{\beta}^* \) denotes the parameters that remain after the restrictions have been imposed.

\textbf{R rule:} Fit terms in the model in the order listed. If the term (factor) adds one or more columns too many \( (\Rightarrow \text{singular } X'X) \), the following constraints are applied to the factor effects:

− sum to zero \quad \text{for } \text{contrasts} = "contr.sum"
− put first level = zero \quad \text{for } \text{contrasts} = "contr.treatment"
Non-full rank $X$ and non-estimable functions

- Constraints must be expressed in terms of non-estimable functions of the $\beta_i$s.
  Care needed if more than one constraint is required; a function that is non-estimable in the original formulation of a model may become estimable once a constraint is imposed.
- $P'\beta$ is estimable iff $\exists L$ such that $E(L'y) = P'\beta$
- In practice use experience and/or trial and error.

2.5 Model specification

\[ \chi = X\beta + e \]

The model depends on $X$ only in as much as $X$ determines the estimation space $M(X)$.

Any two models that have the same estimation space are equivalent.

This equivalence leads to the following result:
Let $M(X)$ be an $r$-dimensional sub-space of $\mathbb{R}^n$.
Let $V = [V_1, V_2, \ldots, V_r]$ be a set of $r$ vectors in $\mathbb{R}^n$ such that $M(V) = M(X)$.
Then the projection of $\chi$ onto $M(X)$ can be found as
\[ \hat{\chi} = V(V'V)^{-1}V'\chi \]

Exercise

i. Show that $\hat{\chi} = V(V'V)^{-1}V'y \in M(V)$

ii. Show that $V(V'V)^{-1}V'$ is a symmetric idempotent matrix.
    ($H$ is idempotent if $H^2 = H$)

iii. Show that $\hat{\chi}$ (as defined above) is orthogonal to $(\chi - \hat{\chi})$
2.6 Deviance (residual sum of squares)

\[
\text{deviance} = d(\hat{\beta}) = (y - \hat{X}\hat{\beta})'(y - \hat{X}\hat{\beta}) = \hat{\beta}'\hat{\beta}
\]
\[
= y'y - 2\hat{\beta}'\hat{y} + \hat{\beta}'\hat{X}'\hat{y}
\]
\[
= y'y - \hat{\beta}'\hat{X}'\hat{y}
\]

Remarks:
- Deviance measures how well the model fits the data.
- Deviance is small for 'good' models.
- \(\hat{\sigma}^2 = d(\hat{\beta})/(n - r)\) where \(n\) is the number of obsn and \(r = \text{rank}(X)\)
- If \(E \sim N(0, \sigma^2 I)\), then \(d(\hat{\beta})/\sigma^2 \sim \chi^2_{n - r}\), which can be used to test the adequacy of fit of the model if \(\sigma^2\) is known.

2.7 Inference I: F-tests

For comparing models (testing hypotheses).

'Big model' (H1):
Assume \(y \sim = X_1\beta \sim 1 + e\sim\) where \(E \sim N(0, \sigma^2 I)\), \(\text{rank}(X_1) = r_1\) and deviance = \(d(\hat{\beta}_1)\).

Null Hypothesis (H0):
Must be expressible in terms of (linear) constraints on (estimable) functions of \(\beta\).
That is, \(L\hat{\beta} = c\sim\) where (usually) \(c\sim = 0\sim\), but need not be.
Eg \(\beta_1 = 0; \beta_2 = \beta_3; \beta_4 = 2\beta_5 □ 3\beta_6 = 2\).

If \(H_0\) is true (which implies that the model under \(H_1\) is also true), then
\[
\frac{d(\hat{\beta}_0) □ d(\hat{\beta}_1)}{(r_1 □ r_0)}\frac{d(\hat{\beta}_1)}{d(\hat{\beta}_0)} = F(r_1 □ r_0, n □ r_1)
\]
Exercise

One-way ANOVA with 2 groups (continued).

\[
y = \begin{bmatrix}
0 \\
4 \\
7 \\
5
\end{bmatrix}
= \begin{bmatrix}
1 & 0 \\
1 & 0 \\
0 & 1 \\
0 & 1
\end{bmatrix} \begin{bmatrix}
\mu_1 \\
\mu_2
\end{bmatrix} + e
\]

Test \( H_0 : \mu_1 = \mu_2 \).

R analysis

\[
\begin{verbatim}
> y <- c(0, 4, 7, 5)
> group <- c(1, 1, 2, 2)
> group.f <- factor(group)
> anova(lm(y ~ group.f))
\end{verbatim}
\]

Analysis of Variance Table

Response: y

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>group</td>
<td>1</td>
<td>16</td>
<td>16</td>
<td>3.2</td>
</tr>
<tr>
<td>Residuals</td>
<td>2</td>
<td>10</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

2.8 Inference II: t-tests and confidence intervals

Applicable to (a single) linear function of the parameters:

\[
p_1\beta_1 + p_2\beta_2 + \ldots + p_k\beta_k = \beta'
\]

For example:

\[
\beta_1 \Rightarrow \beta' = (1, 0, 0, \ldots, 0)
\]

\[
\beta_2 - \beta_1 \Rightarrow \beta' = (-1, 1, 0, \ldots, 0)
\]

\[
\beta_3 - \frac{1}{2}(\beta_1 + \beta_2) \Rightarrow \beta' = (-\frac{1}{2}, -\frac{1}{2}, 1, \ldots, 0)
\]

Under the usual assumptions,

\[
\frac{\beta' - c}{\text{se}(\beta')} \sim t_{n-r} \quad \text{where} \quad c = \beta'
\]
Notes

▶ Usually $c = 0$, (eg to test $H_0 : \beta_1 = 0$, ) or $c = P\beta$ for confidence intervals.
▶ $t_{n-r}^2 \equiv F_{1, n-r}$. An $F$-test with 1 and $\nu$ df is equivalent to a (2-sided) $t$-test with $\nu$ df.
▶ All of the model assumptions are needed: normality, independence, constant variance, but the results are not sensitive to small departures from the assumptions.
▶ $se(P\beta) = \sqrt{\text{var}(P\hat{\beta})}$
▶ $\text{var}(P\hat{\beta}) = P\hat{\beta} D(\hat{\beta}) P = \sigma^2 P(X'X)^{-1} P$, (if $X'X$ is non-singular).

Notes

▶ Standard errors (se) are the key. If we collect many sets of data, compute $P\hat{\beta}$ for each, and make a histogram, then mean $\approx P\hat{\beta}$; standard deviation $\approx se(P\hat{\beta})$
▶ (95%) confidence interval for $P\hat{\beta}$:

$$P\hat{\beta} \pm t_{n-r}(0.975) se(P\hat{\beta})$$

point estimate $\pm$ (table value) $\times$ se

▶ Two important standard errors.
  ▶ Average (of $n$ independent observations): $se(\bar{Y}) = \frac{\hat{\sigma}}{\sqrt{n}}$
  ▶ Difference between two averages: $se(\bar{Y}_1 - \bar{Y}_2) = \frac{\hat{\sigma}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

Exercise

One-way ANOVA with 2 groups (continued).

$\begin{bmatrix} 0 \\ 4 \\ 7 \\ 5 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + e$

1. Test $H_0 : \mu_1 = \mu_2$.

2. Find a 95% CI for $\mu_2 - \mu_1$. 

Exercise

One-way ANOVA with 2 groups (continued).
R analysis

```r
> coef(summary(lm.1 <- lm(y ~ group.f)))

   Estimate Std. Error  t value Pr(>|t|)
(Intercept)   2.000000   1.581139  1.26491   0.3333
group.f2      4.000000   2.236068  1.78885   0.2155

> TukeyHSD(aov(lm(y ~ group.f)))

Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = lm(y ~ group.f))

group.f
     diff lwr upr p adj
2-1  4.000000 -5.612751 13.61275 0.2155025

> confint(lm.1)

     2.5 % 97.5 %
(Intercept) -4.803091  8.803091
group.f2    -5.621024  13.621024
```