Sum of squares.

Normal linear model

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + \varepsilon \]

Simpler model

\[ y = \beta_0 + \varepsilon \]

\[ \hat{\beta} = \bar{y} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta} \cdot \bar{x} \quad SS_0 = \sum (y_i - \bar{y})^2 \quad df = v = m - 1 \]

Suppose bring in \( x' \) successively

\[ y - \hat{y}_0 = (y - \hat{y}_m) + (\hat{y}_m - \hat{y}_{m-1}) + \ldots + (\hat{y}_1 - \hat{y}_0) \]

Terms orthogonal

\[ SS_0 = SS_m + (SS_m - SS_{m-1}) + \ldots + (SS_0 - SS_1) \]

\( SS_r = SS_m - SS_r : \) reduction in residual \( SS \) due to adding \( x_r \)

\[ \| y - \hat{y}_0 \|^2 = \| y - \hat{y}_m \|^2 + \| \hat{y}_m - \hat{y}_{m-1} \|^2 + \ldots + \| \hat{y}_1 - \hat{y}_0 \|^2 \]

\( y \) normal \( \Rightarrow y_r - \hat{y}_{r-1}, y - \hat{y}_m \) normal

\[ SS_m = SS_{m-1} - SS_r \text{ independent} \]
AN OVA Table based on $\mathbf{y}$

<table>
<thead>
<tr>
<th>Terms added</th>
<th>df</th>
<th>$\text{Reduction in SS}$</th>
<th>$\text{MS} = \frac{\text{SS}}{\text{df}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>$m-1-n_1$</td>
<td>$SS_0 - SS_1$</td>
<td></td>
</tr>
<tr>
<td>$X_2$</td>
<td>$n_1-n_2$</td>
<td>$SS_1 - SS_2$</td>
<td></td>
</tr>
<tr>
<td>$X_3$</td>
<td>$n_2-n_3$</td>
<td>$SS_2 - SS_3$</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>$X_m$</td>
<td>$n_{m-1}-n_m$</td>
<td>$SS_m - SS_m$</td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>$n_m$</td>
<td>$SS_m$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$m-1$</td>
<td>$SS_0$</td>
<td></td>
</tr>
</tbody>
</table>

F ratios: $\frac{(SS_{m-1} - SS_r)}{(SS_m / n_m)} / (y_{r-1} - y_r)$
Cement data.

\[ y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \beta_4 x_{4j} + \epsilon_j \]

\( m = 13 \quad \hat{\sigma} = 5.1 \)

### 8.1 - Introduction

Table 8.1 Cement data

<table>
<thead>
<tr>
<th>Case</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( x_4 )</th>
<th>( y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>26</td>
<td>6</td>
<td>60</td>
<td>78.5</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>29</td>
<td>15</td>
<td>52</td>
<td>74.3</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>56</td>
<td>8</td>
<td>20</td>
<td>104.3</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>31</td>
<td>8</td>
<td>47</td>
<td>87.6</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>52</td>
<td>6</td>
<td>33</td>
<td>95.9</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td>55</td>
<td>9</td>
<td>22</td>
<td>109.2</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>71</td>
<td>17</td>
<td>6</td>
<td>102.7</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>31</td>
<td>22</td>
<td>44</td>
<td>72.5</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>54</td>
<td>18</td>
<td>22</td>
<td>93.1</td>
</tr>
<tr>
<td>10</td>
<td>21</td>
<td>47</td>
<td>4</td>
<td>26</td>
<td>115.9</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>40</td>
<td>23</td>
<td>34</td>
<td>83.8</td>
</tr>
<tr>
<td>12</td>
<td>11</td>
<td>66</td>
<td>9</td>
<td>12</td>
<td>113.3</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
<td>68</td>
<td>8</td>
<td>12</td>
<td>109.4</td>
</tr>
</tbody>
</table>

Bring in \( x_2 \) successively.
```r
df <- matrix(scan("cement"), byrow=T, ncol=5)
x1 <- df[,1]; x2 <- df[,2]; x3 <- df[,3]; x4 <- df[,4]; y <- df[,5]
m0 <- lm(y ~ 1)
m1 <- lm(y ~ x1)
m2 <- lm(y ~ x1 + x2)
m3 <- lm(y ~ x1 + x2 + x3)
m4 <- lm(y ~ x1 + x2 + x3 + x4)
anova(m0, m1, m2, m3, m4)
```

Model 1: $y \sim 1$
Model 2: $y \sim x_1$
Model 3: $y \sim x_1 + x_2$
Model 4: $y \sim x_1 + x_2 + x_3$
Model 5: $y \sim x_1 + x_2 + x_3 + x_4$

<table>
<thead>
<tr>
<th>Res.Df</th>
<th>RSS</th>
<th>Df Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>2715.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>1265.69</td>
<td>1</td>
<td>1450.08</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>57.90</td>
<td>1</td>
<td>1207.78</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>48.11</td>
<td>1</td>
<td>9.79</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>47.86</td>
<td>1</td>
<td>0.25</td>
</tr>
</tbody>
</table>

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Orthogonality.

The order of inserting terms affects their reduction in sum of squares, generally

Suppose there are $X_1$ and $X_2$ and

$$y = 1, \beta_0 + X_1 \beta_1 + X_2 \beta_2 + \varepsilon$$

The normal equations are

$$
\begin{bmatrix}
I \\ X_1^T \\ X_2^T
\end{bmatrix}
\begin{bmatrix}
1 \\ X_1 \\ X_2
\end{bmatrix}
\begin{bmatrix}
\beta_0 \\ \beta_1 \\ \beta_2
\end{bmatrix}
= 
\begin{bmatrix}
y \\ X_1 y \\ X_2 y
\end{bmatrix}
$$

Suppose $X_1^T 1, X_2^T 1, X_1^T X_2 = 0$ orthogonal

Then

$$
\begin{bmatrix}
I \\ 0 \\ 0 \\
0 \\ X_1^T X_1 \\ 0 \\
0 \\ 0 \\ X_2^T X_2
\end{bmatrix}
\begin{bmatrix}
\beta_0 \\ \beta_1 \\ \beta_2
\end{bmatrix}
= 
\begin{bmatrix}
y \\ X_1 y \\ X_2 y
\end{bmatrix}
$$

If inverses exist

$$
\beta_0 = \bar{y}, \quad \beta_r = (X_r^T X_r)^{-1} X_r y, \quad r = 1, 2
$$
Residual sum of squares

\[(y - X \hat{\beta})^T (y - X \hat{\beta}) = \hat{\beta}_0 - \frac{\hat{\beta}_1}{\hat{\beta}_2} X^T X \hat{\beta}_2\]

Order of fitting does not matter

If \( \varepsilon \)'s are \( \text{IN}(0, \sigma^2) \), then

\[\hat{\beta}_0 \sim N(\beta_0, \frac{\sigma^2}{n})\]

\[\hat{\beta}_1 \sim N(\beta_1, (X^T X)^{-1} \sigma^2)\]

\[\hat{\beta}_2 \sim N(\beta_2, (X^T X)^{-1} \sigma^2)\]
8.1 - Introduction

Table 8.2 Data and experimental setup for bicycle experiment (Box et al., 1978, pp. 360-372). The lower part of the table shows the average times for each of the eight combinations of settings of seat height, tyre pressure, and dynamo, and the average times for the eight observations at each setting, considered separately.

<table>
<thead>
<tr>
<th>Setup</th>
<th>Day</th>
<th>Run</th>
<th>Seat height (inches)</th>
<th>Dynamo</th>
<th>Tyre pressure (psi)</th>
<th>Time (secs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>51</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>54</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>41</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>3</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>43</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>3</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>54</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>1</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>60</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>1</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>44</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>3</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>43</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>50</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>48</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>5</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>39</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>2</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>39</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>4</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>53</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>3</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>51</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>2</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>41</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>4</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>44</td>
</tr>
</tbody>
</table>

\[
\begin{pmatrix}
Y_1 \\
Y_2 \\
Y_3 \\
Y_4 \\
Y_5 \\
Y_6 \\
Y_7 \\
Y_8 \\
Y_9 \\
Y_{10} \\
Y_{11} \\
Y_{12} \\
Y_{13} \\
Y_{14} \\
Y_{15} \\
Y_{16}
\end{pmatrix} = \begin{pmatrix}
1 \\
1 \\
1 \\
1 \\
1 \\
1 \\
1 \\
1 \\
1 \\
1 \\
1 \\
1 \\
1 \\
1 \\
1 \\
1
\end{pmatrix}\begin{pmatrix}
\beta_0 \\
\beta_1 \\
\beta_2 \\
\beta_3
\end{pmatrix} + \begin{pmatrix}
\epsilon_1 \\
\epsilon_2 \\
\epsilon_3 \\
\epsilon_4 \\
\epsilon_5 \\
\epsilon_6 \\
\epsilon_7 \\
\epsilon_8 \\
\epsilon_9 \\
\epsilon_{10} \\
\epsilon_{11} \\
\epsilon_{12} \\
\epsilon_{13} \\
\epsilon_{14} \\
\epsilon_{15} \\
\epsilon_{16}
\end{pmatrix} \quad (8.2)
\]
SMILE IF YOU BUNT

Brian D. Wang, an associate professor of statistics at Skidmore College, analyzed baseball managers from the 2007 season to determine which facial features are significant. He found that managers with higher batting averages tend to have faces with more prominent noses, larger mouths, and wider chins. The study suggests that managers might use their facial features as a form of visual communication to influence their players.

**Percentage of players who had the advantage of being against an ineffective-batted player at the end of the game**

Note: Because different teams and managers use different strategies, the effectiveness of each manager's tactics was compared with the team's overall batting average.

---

**Number of different managers used**

- 1

**Percentage of players used**

- 100%

**Percentage of players used in the game**

- 100%

**Percentage of players used in the batter’s box**

- 100%

**Number of different managers used**

- 1

**Percentage of players used**

- 100%

**Percentage of players used in the game**

- 100%

**Percentage of players used in the batter’s box**

- 100%
Model checking

What are the assumptions of the normal/Gaussian regression model?

\[ y = X \beta + \epsilon \quad \epsilon \sim N(0, \sigma^2) \]

1. Uniqueness in \( \beta, X \)
2. Constant variance \( \sigma^2 \)
3. Independence \( \epsilon \)'s (uncorrelated)
4. Normality

What to check for?
- Isolated discrepancies, outliers
- Systematic dependency
- Transformation of dependent variable needed
- Transformation of covariates
- Omitted variable
- Correlated errors

How?
- Residuals
Raw residuals

\[ e = y - \hat{y} = y - X(X'X)^{-1}X'y \]
\[ = (I - H)y \quad H = X(X'X)^{-1}X' \]
\[ = (I - H)e \]

Properties

\[ E(e) = 0 \]
\[ \text{var}(e) = \sigma^2(I - H) \]
\[ \text{var}(e_j) = \sigma^2 (1 - h_{jj}) \]
\[ \text{cov}(e_j, e_k) = -\sigma^2 h_{jk} \quad j \neq k \]
Standardized residuals

\[ r_j = \frac{\hat{e}_j}{\sqrt{1-h_{jj}}} \]

\[ = \frac{(y_j - x_j^T \hat{\beta})}{\sigma \sqrt{1-h_{jj}}} \]

\[ E(r_j) = 0 \]

\[ \text{Var}(r_j) \approx 1 \]

Check on linearity

Plot \( y_j \) on \( x_l \), \( l=1, \ldots, p \)

Plot \( r_j \) on \( x_l \), \( l=1, \ldots, p \)

Plot \( r \) on \( x_l \), omitted variable

Look for pattern
Cycling data

8 · Linear Regression Models

Figure 8.4: Residual plots for data on cycling up a hill. The panels showing residuals plots against levels of day and run, and against fitted values, would show random variation if the model is adequate, as seems to be the case. To normal scores plot show that the errors appear close to normal.

- Constancy of variance: $\sigma_j$ (or $1/s_j$) as $j$
- Wedging
- Independence: $r_j$, $s_j$, $t_j$, or run number
- Distribution of errors: normal prob plot on $r_j$
Nonlinearity

\[ a(y) = x^T \beta + \varepsilon \]

\[ y = a^{-1}(x^T \beta + \varepsilon) \]

\[ y = b(x^T \beta) + \varepsilon \]

Box-Cox transform

\[ u^{(\lambda)} = \alpha(x) = \frac{u^{^\lambda} - 1}{\lambda}, \quad \lambda \neq 0 \]

\[ \log u, \quad \lambda = 0 \]

\[ y^{(\lambda)} = x^T \beta + \varepsilon \]

Jacobian

\[ \frac{dy^{(\lambda)}}{dy} = \lambda y^{\lambda - 1} \]

\[ g(y^{(\lambda)}) dy^{(\lambda)} = g(y^{(\lambda)}) \lambda y^{\lambda - 1} dy \]
n fixed

log likelihood

$$-\frac{1}{2} \sum \log \sigma^2 + \frac{1}{\sigma^2} \sum_{j=1}^{m} (y_j^{(a)} - \hat{x}_j^T \hat{\beta})^2$$

$$+ (\alpha - 1) \sum \log y_j$$

$$\hat{\beta} = (X^T X)^{-1} X^T y^{(a)}$$

$$SS(\hat{\beta}_a)/m = \hat{\sigma}^2$$

profile log likelihood

$$l_p(\alpha) = -\frac{m}{2} \left[ \log SS(\hat{\beta}_a) - \log \alpha^{2(\alpha-1)} \right]$$

$$g = (\prod y_j)^{1/n}$$
Generalized additive model

\[ y = b(x^T \beta) + \varepsilon \]

\( b \): smooth
- spline

\text{lowess}(), \text{loess}(), \text{gam}()
$\hat{y}_j = X \hat{\beta} = X (X^T X)^{-1} X^T y = h_{jj} y_j$

$\hat{y}_j = h_{jj} y_j + \sum_{j \neq i} h_{ij} y_i$

$\hat{y}_j$ will be dominated by $y_j$ if $y_j$ is an outlier.

$0 \leq h_{jj} \leq 1$

If $h_{jj}$ is large, changing $y_j$ will move $\hat{y}_j$ a lot.

$tr(H) = \sum_{i} h_{ii} = p$

Cases with $h_{jj} > 2p/n$ deserve close inspection.
we obtain \( \hat{\beta} = (X'X)^{-1}X'y \). The fitted value \( \hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y = Hy \) is the orthogonal projection of \( y \) onto the plane spanned by the columns of \( X \), and the matrix representing that projection is \( H \). Notice that \( \hat{y} \) is unique whether or not \( X'X \) is invertible.

Figure 8.2 shows that the vector of residuals, \( e = y - \hat{y} = (I_n - H)y \), and the vector of fitted values, \( \hat{y} = Hy \), are orthogonal. To see this algebraically, note that

\[
\hat{y}'e = y'Hy = (1_n - H'y) = y'H'Hy = 0, \tag{8.6}
\]

because \( H' = H \) and \( HH = H \), that is, the projection matrix \( H \) is symmetric and idempotent (Exercise 8.2.5). The close link between orthogonality and independence for normally distributed vectors means that (8.6) has important consequences, as we shall see in Section 8.3. For now, notice that (8.6) implies that

\[
y'y = (y - \hat{y} + \hat{y})'y = (y - \hat{y} + \hat{y})' = e'e + \hat{y}'\hat{y}, \tag{8.7}
\]

as is clear from Figure 8.2 by Pythagoras’ theorem. That is, the overall sum of squares of the data, \( \sum y_j^2 = y'y \), equals the sum of the residual sum of squares, \( SS(\hat{\beta}) = \sum (y_j - \hat{y}_j)^2 = e'e \), and the sum of squares for the fitted model, \( \sum \hat{y}_j^2 = \hat{y}'\hat{y} \).

Such decompositions are central to analysis of variance, discussed below.

8.2.3 Likelihood quantities

Chapter 4 shows how the observed and expected information matrices play a central role in likelihood inference, by providing approximate variances for maximum likelihood estimates. To obtain these matrices for the normal linear model, note that the

\begin{table}[h]
\centering
\begin{tabular}{cccccccccccc}
\hline
\textbf{Setup} & \textbf{Seat height} & \textbf{Dynamo} & \textbf{Tyre pressure} & \textbf{Time} & \textbf{\( \hat{y} \)} & \textbf{\( e \)} & \textbf{\( r \)} & \textbf{\( r' \)} & \textbf{\( h \)} & \textbf{\( C \)} \\
\hline
1 & -1 & -1 & -1 & 51 & 52.62 & -1.625 & -0.99 & -0.99 & 0.25 & 0.08 \\
2 & -1 & -1 & -1 & 54 & 52.62 & 1.375 & -0.84 & 0.83 & 0.25 & 0.06 \\
3 & 1 & -1 & -1 & 41 & 41.75 & -0.750 & -0.46 & -0.44 & 0.25 & 0.02 \\
4 & -1 & 1 & -1 & 43 & 41.75 & 1.250 & 0.75 & 0.75 & 0.25 & 0.05 \\
5 & -1 & 1 & -1 & 54 & 55.75 & -1.750 & -1.06 & -1.07 & 0.25 & 0.09 \\
6 & -1 & 1 & 1 & 60 & 55.75 & 4.250 & 2.59 & 3.72 & 0.25 & 0.56 \\
7 & 1 & 1 & -1 & 44 & 44.87 & -0.875 & -0.53 & -0.52 & 0.25 & 0.02 \\
8 & 1 & 1 & 1 & 43 & 44.87 & -1.875 & -1.14 & -1.16 & 0.25 & 0.11 \\
9 & -1 & -1 & 1 & 50 & 49.50 & 0.500 & 0.30 & 0.29 & 0.25 & 0.01 \\
10 & -1 & 1 & 1 & 48 & 49.50 & 1.500 & 0.91 & 0.91 & 0.25 & 0.07 \\
11 & 1 & -1 & 1 & 39 & 38.62 & 0.375 & 0.23 & 0.22 & 0.25 & 0.00 \\
12 & 1 & -1 & 1 & 39 & 38.62 & 0.375 & 0.23 & 0.22 & 0.25 & 0.00 \\
13 & -1 & 1 & 1 & 53 & 52.62 & 0.375 & 0.23 & 0.22 & 0.25 & 0.00 \\
14 & -1 & 1 & 1 & 51 & 52.62 & -1.625 & -0.99 & -0.99 & 0.25 & 0.08 \\
15 & 1 & 1 & 1 & 41 & 41.75 & -0.750 & -0.46 & -0.44 & 0.25 & 0.02 \\
16 & 1 & 1 & 1 & 44 & 41.75 & 2.250 & 1.37 & 1.43 & 0.25 & 0.16 \\
\hline
\end{tabular}
\caption{Data from bicycle experiment, together with fitted values \( \hat{y} \), raw residuals \( e \), standardized residuals \( r \), deletion residuals \( r' \), leverage \( h \) and Cook distances \( C \).}
\end{table}
Simple linear regression

\[ y_j = \gamma_0 + (\gamma_1 - \bar{x})x_j + \epsilon_j \]

\[
\begin{pmatrix}
  y_1 \\
  \vdots \\
  y_m
\end{pmatrix}
= \begin{pmatrix}
  1 & x_1 - \bar{x} \\
  \vdots & \vdots \\
  1 & x_m - \bar{x}
\end{pmatrix}
\begin{pmatrix}
  \gamma_0 \\
  \gamma_1
\end{pmatrix}
+ \begin{pmatrix}
  \epsilon_1 \\
  \vdots \\
  \epsilon_m
\end{pmatrix}
\]

\[ \hat{\beta} = (X^T X)^{-1} X^T \begin{pmatrix}
  y_1 \\
  \vdots \\
  y_m
\end{pmatrix} = \begin{pmatrix}
  \sum (x_j - \bar{x})^2 \\
  \sum (x_j - \bar{x})
\end{pmatrix}^{-1} \begin{pmatrix}
  \sum y_j (x_j - \bar{x}) \\
  \sum (x_j - \bar{x})^2
\end{pmatrix} = \begin{pmatrix}
  \frac{1}{m} + \frac{(x_j - \bar{x})^2}{\sum (x_k - \bar{x})^2} \\
  \frac{1}{m} + \frac{(x_j - \bar{x})^2}{\sum (x_k - \bar{x})^2}
\end{pmatrix} \]