Measurements of degradation of heme with different concentrations of hydrogen peroxide (H$_2$O$_2$), for different species of heme.
The regression model

Let $X$ be the predictor and $Y$ be the response. Assume we have $n$ observations $(x_1, y_1), \ldots, (x_n, y_n)$ from $X$ and $Y$. The simple linear regression model is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim \text{iid } N(0, \sigma^2).$$

How do we estimate $\beta_0$, $\beta_1$, $\sigma^2$?
We can write

\[ \epsilon_i = y_i - \beta_0 - \beta_1 x_i \]

For a pair of estimates \((\hat{\beta}_0, \hat{\beta}_1)\) for \((\beta_0, \beta_1)\) we define the fitted values as

\[ \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \]

The residuals are

\[ \hat{\epsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \]
Residual sum of squares

For every pair of values for $\beta_0$ and $\beta_1$ we get a different value for the residual sum of squares.

$$\text{RSS}(\beta_0, \beta_1) = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

We can look at RSS as a function of $\beta_0$ and $\beta_1$. We try to minimize this function, i.e. we try to find

$$\hat{(\beta_0, \beta_1)} = \min_{\beta_0, \beta_1} \text{RSS}(\beta_0, \beta_1)$$

Hardly surprising, this method is called least squares estimation.
Residual sum of squares

Notation

Assume we have \( n \) observations: \((x_1, y_1), \ldots, (x_n, y_n)\).

\[
\bar{x} = \frac{\sum_i x_i}{n}
\]
\[
\bar{y} = \frac{\sum_i y_i}{n}
\]
\[
SXX = \sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - n(\bar{x})^2
\]
\[
SYY = \sum_i (y_i - \bar{y})^2 = \sum_i y_i^2 - n(\bar{y})^2
\]
\[
SXY = \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i y_i - \bar{x}\bar{y}n
\]
\[
RSS = \sum_i (y_i - \hat{y}_i)^2 = \sum_i \epsilon_i^2
\]
Parameter estimates

The function

$$\text{RSS}(\beta_0, \beta_1) = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

is minimized by

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Useful to know

Using the parameter estimates, our best guess for any $y$ given $x$ is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

Hence

$$\hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} = \bar{y}$$

That means every regression line goes through the point $(\bar{x}, \bar{y})$. 
Variance estimates

As variance estimate we use
\[ \hat{\sigma}^2 = \frac{\text{RSS}}{n - 2} \]

This quantity is called the residual mean square. It has the property
\[ (n - 2) \times \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n - 2} \]

In particular, this implies
\[ E(\hat{\sigma}^2) = \sigma^2 \]

Example

<table>
<thead>
<tr>
<th>H\textsubscript{2}O\textsubscript{2} concentration</th>
<th>0</th>
<th>10</th>
<th>25</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3399</td>
<td>0.3168</td>
<td>0.2460</td>
<td>0.1535</td>
<td></td>
</tr>
<tr>
<td>0.3563</td>
<td>0.3054</td>
<td>0.2618</td>
<td>0.1613</td>
<td></td>
</tr>
<tr>
<td>0.3538</td>
<td>0.3174</td>
<td>0.2848</td>
<td>0.1525</td>
<td></td>
</tr>
</tbody>
</table>

We get
\[ \bar{x} = 21.25, \quad \bar{y} = 0.27, \quad SXX = 4256.25, \quad SXY = -16.48, \quad \text{RSS} = 0.0013. \]

Therefore
\[ \hat{\beta}_1 = \frac{-16.48}{4256.25} = -0.0039, \quad \hat{\beta}_0 = 0.27 - (-0.0039) \times 21.25 = 0.353, \]
\[ \hat{\sigma} = \sqrt{\frac{0.0013}{12 - 2}} = 0.0115. \]
The R function `lm()` does all these calculations for you. And more!

Comparing models

We want to test whether $\beta_1 = 0$:

$$H_0 : y_i = \beta_0 + \epsilon_i \quad \text{versus} \quad H_a : y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
Sum of squares

Under $H_a$:

$$RSS = \sum_i (y_i - \hat{y}_i)^2 = SYY - \frac{(SXY)^2}{SXX} = SYY - \beta_1^2 \times SXX$$

Under $H_0$:

$$\sum_i (y_i - \hat{\beta}_0)^2 = \sum_i (y_i - \bar{y})^2 = SYY$$

Hence

$$SS_{reg} = SYY - RSS = \frac{(SXY)^2}{SXX}$$

ANOVA

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>regression on X</td>
<td>1</td>
<td>$SS_{reg}$</td>
<td>$MS_{reg} = \frac{SS_{reg}}{1}$</td>
<td>$\frac{MS_{reg}}{MSE}$</td>
</tr>
<tr>
<td>residuals for full model</td>
<td>$n - 2$</td>
<td>RSS</td>
<td>MSE = $\frac{RSS}{n - 2}$</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>$n - 1$</td>
<td>SYY</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
David Sullivan’s pf3d7 data

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>regression on X</td>
<td>1</td>
<td>0.06378</td>
<td>0.06378</td>
<td>484.1</td>
</tr>
<tr>
<td>residuals for full model</td>
<td>10</td>
<td>0.00131</td>
<td>0.00013</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>11</td>
<td>0.06509</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Remember: The R function `lm()` does the calculations for you!