(Hoff p. 152ff.)

Let $\mathbf{y}$ be the n-dimensional column vector $(y_1, ..., y_n)^T$ and let $\mathbf{X}$ be the $n \times p$ matrix whose ith row is $\mathbf{x}_i = \{x_{i,1}, x_{i,2}, ..., x_{i,p}\}$. Then the normal regression model is

$$\{\mathbf{y}|\mathbf{X}, \beta, \sigma^2\} \sim \text{multivariate normal}\left(\mathbf{X}\beta, \sigma^2\mathbf{I}\right),$$

where $\mathbf{I}$ is the $p \times p$ identity matrix and

$$\mathbf{X}\beta = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} \beta_1 x_{1,1} + \cdots + \beta_p x_{1,p} \\ \vdots \\ \beta_1 x_{n,1} + \cdots + \beta_p x_{n,p} \end{pmatrix} = \begin{pmatrix} E\left[Y_1|\beta, \mathbf{x}_1\right] \\ \vdots \\ E\left[Y_n|\beta, \mathbf{x}_n\right] \end{pmatrix}$$

We compute the ordinary least squares estimates

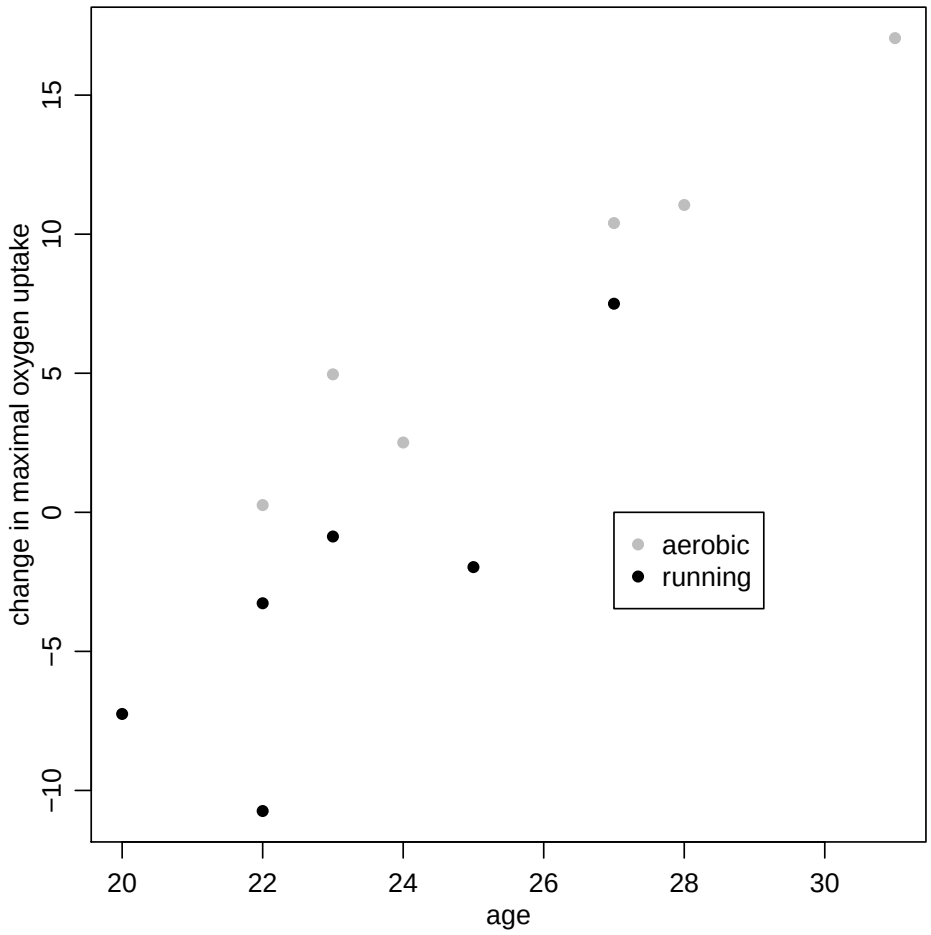
$$\hat{\beta}_{ols} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

and

$$\hat{\sigma}^2_{ols} = \frac{SSR\left(\hat{\beta}_{ols}\right)}{(n-p)} = \frac{\sum\left(y_i - \hat{\beta}^T_{ols}x_i\right)^2}{(n-p)}.$$

Example: Oxygen uptake (from Kuehl (2000), Hoff p. 149ff)

Twelve healthy men who did not exercise regularly were recruited to take part in a study of the effects of two different exercise regimens on oxygen uptake. Six of the twelve men were randomly assigned to a 12-week flat-terrain running program, and the remaining six were assigned to a 12-week step aerobics program. The maximum oxygen uptake of each subject was measured (in liters per minute) while running on an inclined treadmill, both before and after the 12-week program. Of interest is how a subject's change in maximal oxygen uptake may depend on which program they were assigned to. However, other factors, such as age, are expected to affect the change in maximal uptake as well. The results are shown here:



Hoff's regression model:

$$
\begin{aligned}
Y_i &= \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \epsilon_i, \text{ where} \\
x_{i,1} &= 1 \text{ for each subject } i \\
x_{i,2} &= 0 \text{ if subject } j \text{ is on the running program, 1 if on aerobic} \\
x_{i,3} &= \text{ age of subject } i \\
x_{i,4} &= x_{i,2} \times x_{i,3}
\end{aligned}
$$

Under this model the conditional expectations of Y for the two different levels of $x_{i,1}$ are

$$
\begin{aligned}
E[Y|\mathbf{x}] &= \beta_1 + \beta_3 \times \text{ age if } x_1 = 0, \text{ and} \\
E[Y|\mathbf{x}] &= (\beta_1 + \beta_2) + (\beta_3 + \beta_4) \times \text{ age if } x_1 = 1
\end{aligned}
$$

In other words, the model assumes that the relationship is linear in age for both exercise groups, with the difference in intercepts given by β_2 and the difference in slopes given by β_4.

The normal linear regression model specifies that, in addition to $E[Y|\mathbf{x}]$ being linear, the sampling variability around the mean is i.i.d. normal: