

(Hoff p. 149ff.)

Regression modeling is concerned with describing how the sampling distribution of one random variable  $Y$  varies with another variable or set of variables  $\mathbf{x} = (x_1, \dots, x_p)$ . Specifically, a regression model postulates a form for  $p(y|\mathbf{x})$ , the conditional distribution of  $Y$  given  $\mathbf{x}$ . Estimation of  $p(y|\mathbf{x})$  is made using data  $y_1, \dots, y_n$  that are gathered under a variety of conditions  $\mathbf{x}_1, \dots, \mathbf{x}_n$ .

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:

$$\begin{aligned}\epsilon_1, \dots, \epsilon_n &\stackrel{\text{i.i.d.}}{\sim} \text{normal}(0, \sigma^2) \\ Y_i &= \beta^T \mathbf{x}_i + \epsilon_i\end{aligned}$$

This model provides a complete specification of the joint probability density of observed data  $y_1, \dots, y_n$  conditional upon  $\mathbf{x}_1, \dots, \mathbf{x}_n$  and values of  $\beta$  and  $\sigma^2$ :

$$\begin{aligned}p(y_1, \dots, y_n | \mathbf{x}_1, \dots, \mathbf{x}_n, \beta, \sigma^2) &= \prod_{i=1}^n p(y_i | \mathbf{x}_i, \beta, \sigma^2) \\ &= (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta^T \mathbf{x}_i)^2 \right\}\end{aligned}$$

Another way to write this joint probability density is in terms of the multivariate normal distribution: Let  $\mathbf{y}$  be the  $n$ -dimensional column vector  $(y_1, \dots, y_n)^T$  and let  $\mathbf{X}$  be the  $n \times p$  matrix whose  $i$ th row is  $\mathbf{x}_i = \{x_{i,1}, x_{i,2}, \dots, x_{i,p}\}$ . Then the normal regression model is

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

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

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

We compute the ordinary least squares estimates

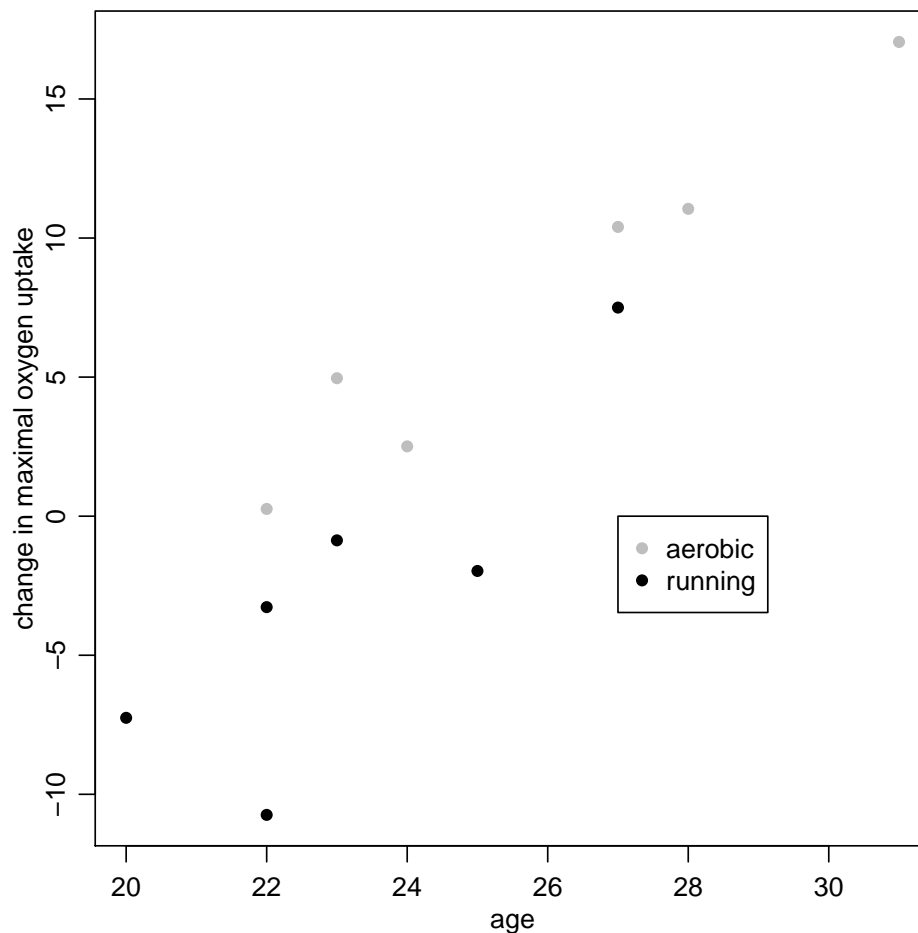
$$\hat{\beta}_{ols} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

and

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

*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:

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \epsilon_i, \text{ where}$$

$$x_{i,1} = 1 \text{ for each subject } i$$

$$x_{i,2} = 0 \text{ if subject } i \text{ is on the running program, } 1 \text{ if on aerobic}$$

$$x_{i,3} = \text{age of subject } i$$

$$x_{i,4} = x_{i,2} \times x_{i,3}$$

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

$$\frac{dx}{dt} = \sigma(y - x) \quad (2a)$$

$$\frac{dy}{dt} = x(\rho - z) - y \quad (2b)$$

$$\frac{dz}{dt} = xy - \beta z \quad (2c)$$

Look at the first line 2a, and now look at the second line 2b. They are both part of the whole system 2.

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  $\beta_2$  and the difference in slopes given by  $\beta_4$ .