3. Logistic Regression

(1) Units of observation $i = 1, \ldots, n$, with binary responses $Y_i = y_i$, where $y_i = 0$ or $y_i = 1$.
One explanatory variable with values $x_i, i = 1, \ldots, n$.
(2) Example. $Y_i = 1$, if i is ill, $Y_i = 0$ if not, and x_i is the level of exposure (amount smoked; dose of medicine etc.)
(3) Dependence of response $P(Y_i = 1)$ on exposure x_i?
(4) Definition. $logit(p) = log(p/(1-p))$ for $0 < p < 1$.
(5) Write $P(Y_i = 1) = p_i$. *Logistic regression model*, or a *logit-model* is

$$logit(p_i) = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, n$$

(6) Using the inverse function we get

$$p_i = logit^{-1}(\beta_0 + \beta_1 x_i) = exp(\beta_0 + \beta_1 x_i)/(1 + exp(\beta_0 + \beta_1 x_i))$$

(7) Properties of function $p = logit^{-1}(x)$: (i) $0 < p < 1$, (ii) $p \to 1$ as $x \to +\infty$, (iii) $p \to 0$ as $x \to -\infty$, (iv) strictly increasing.
(8) Data generation using a computer. (i) Choose the number of observations, say, $n = 30$. (ii) Generate the values of explanatory varibles, say, $x_i = (i - 15.5)/14.5$. (iii) Select values for regression coefficients, say, $\beta_0 = logit(0.3)$ and $\beta_1 = 1$. (iv) Compute the probabilities $p_i = logit^{-1}(\beta_0 + \beta_1 x_i)$.(v) Generate the responses: first pick independent $U_i \sim U[0, 1], i = 1, \ldots, n$, and then define $Y_i = 1$ if $U_i < p_i$, otherwise $Y_i = 0$.

(9) Estimation of coefficients using maximum likelihood. In the binomial case $Y_i \sim Bin(n_i, p_i(\beta_0, \beta_1)), i = 1, \dots, n$ independent, with observed values $Y_i = y_i$ where $y_i \in \{0, \dots, n_i\}$. The likelihood function is (omitting binomial coefficients that do not depend on the parameters)

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p_i(\beta_0, \beta_1))^{y_i}(1 - p_i(\beta_0, \beta_1))^{n_i - y_i}$$

(10) Differentiating with respect to the parameters we get the *likelihood* equations

$$\sum_{i=1}^{n}(y_i - n_i p_i(\beta_0, \beta_1)) = 0$$
$$\sum_{i=1}^{n}(y_i - n_i p_i(\beta_0, \beta_1))x_i = 0$$

Equations are *nonlinear*, to be solved numerically using e.g. Newton's method.

(11) Generalization: model has a constant term and k explanatory variables

$$logit(p_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}, \quad i = 1, \ldots, n$$

(12) This can be written in vector form. Define $\beta = (\beta_0, \ldots, \beta_k)^T$ and $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ik})^T$, so

$$logit(p_i) = \mathbf{x}_i^T \beta \quad i = 1, \ldots, n$$

(13) Now the likelihood equations ($k+1$ nonlinear equations, $k+1$ unknowns) are (binomial case!) in vector form

$$\sum_{i=1}^{n} (y_i - n_i p_i(\beta)) \mathbf{x}_i = \mathbf{0},$$

where $p_i(\beta) = exp(\mathbf{x}_i^T \beta)/(1 + exp(\mathbf{x}_i^T \beta))$. These equations are also usually solved using some variant of Newton's method.