6. Goodness of Fit

(1) In ordinary regression, a single *outlier* can be of interest, and/or suggest that the model fit is poor. Not so in logistic regression.
(2) Suppose $Y_i \sim Bin(n_i, p_i(\beta)), i = 1, \ldots, n$ are independent and the n_i's sufficiently large (e.g. so that $E[Y_i] \geq 5$), then components of χ^2-distributed test statistics can be used to assess goodness of fit. The distribution of standardized *Pearson residuals* is

$$R_i \equiv \frac{y_i - n_i p_i(\hat{\beta})}{\sqrt{n_i p_i(\hat{\beta})(1 - p_i(\hat{\beta}))}} \sim N(0,1), \quad i = 1, \ldots, n,$$

asymptotically.
(3) In this case the goodness of fit of the whole model can be tested with the statistic

$$X_P^2 \equiv \sum_{i=1}^{n} R_i^2 \sim \chi^2_{n-(k+1)}.$$

(4) Under the same assumptions on expected counts, deviance

$$D = \sum_{i=1}^{n} d_i^2 \sim \chi^2_{n-(k+1)},$$

(5) where the summands are based on the *deviance residuals*

$$d_i \equiv sgn_i \times \sqrt{2}\{y_i log \frac{y_i}{n_i p_i(\hat{\beta})} + (n_i - y_i) log \frac{n_i - y_i}{n_i - n_i p_i(\hat{\beta})}\}^{1/2}$$

where $sgn_i = 1$, if $y_i - n_i p_i(\hat{\beta}) \geq 0$, otherwise $sgn_i = -1$.
(6) In addition, $d_i \sim N(0, 1)$ asymptotically.

(7) When $n_i \equiv 1$, single residuals do not have a clear interpretation, so residual checks are based on *data grouping*. Any substantively meaningful grouping (that does not depend on the observed y_i's!) can be informative.

(8) *Hosmer-Lemeshow tests* are based on groupings based on the estimated probabilities.

(9) In addition to studying the responses y_i, it is important to screen the explanatory variables for high *leverage*.

7. Conditional Logistic Regression

(1) In a cohort study, the subjects are sometimes matched, e.g., according to place of residence, if it is supected that there are spatially varying exposures that are hard to measure or unknown.
(2) Similarly, in a case-control study one may select cases and controls from the same areas.
(3) In both cases, logistic regression is applied the same way, by conditioning on the matched sets. Consider the simplest case, the pair-matched case-control study.
(4) Index pairs by $J = 1, \ldots, J$ and the individuals by $i = 2j - 1$ for the case in pair j, and $i = 2j$ for the control.
(5) Suppose the values of the exposure varibles are generated randomly, with probabilities (or density) $P(\mathbf{x})$. The response probability is $P(Y = 1 \mid \mathbf{x})$, so the joint probability of the responses and exposures is $P(Y = 1, \mathbf{x}) = P(Y = 1 \mid \mathbf{x})P(\mathbf{x})$.

(6) Suppose that a logistic regression model for the response in pair j, given that the individual has characteristics x, is

$$logit(P(Y = 1 \mid \mathsf{x})) = \alpha_j + \mathsf{x}^T\beta, \quad j = 1, \dots, J.$$

(7) Here α_j represents unmeasured characteristics shared by the pair j.
(8) The problem is that as the number of pairs J increases, so does the number of parameters, and the usual asymptotic properties of the MLEs do not hold.
(9) A solution is to *condition* on the outcomes of the pairs. The probability that the case is the one with exposures x_{2j-1}, given the exposures $\{\mathsf{x}_{2j-1}, \mathsf{x}_{2j}\}$, equals

$$\frac{P(Y = 1, \mathsf{x}_{2j-1})P(Y = 0, \mathsf{x}_{2j})}{P(Y = 1, \mathsf{x}_{2j-1})P(Y = 0, \mathsf{x}_{2j}) + P(Y = 0, \mathsf{x}_{2j-1})P(Y = 1, \mathsf{x}_{2j})}.$$

(10) This conditional probability equals

$$\frac{exp((\mathbf{x}_{2j-1} - \mathbf{x}_{2j})^T \beta)}{1 + exp((\mathbf{x}_{2j-1} - \mathbf{x}_{2j})^T \beta)}.$$

(11) Computation: J Bernoulli trials, all successes, explanatory variables $\mathbf{x}_{2j-1} - \mathbf{x}_{2j}$, no constant term.
(12) As in case-control studies in general, absolute risk cannot be estimated.
(13) The argument extends to multiple cases, and multiple controls. The resulting likelihood is the *same* as in Cox regression!
(14) If applied in cohort studies, matched sets with all cases, or all non-cases, are not informative.
(15) *Random effects* can also be used to represent hard to measure group characteristics, due to neighborhood or family, for example. So-called *generalized linear mixed models* incorporate such effects.