5. Model Choice

(1) Basic asymptotic result of maximum likelihood estimation. Suppose the observations $Y_i \sim Bin(n_i, p_i(\beta)), i = 1, \ldots, n$ are independent, then (under some further assumptions) the MLE $\hat{\beta}$ is consistent, and satisfies

$$\hat{\beta} \sim N(\beta, (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}),$$

as $n \to +\infty$. Here $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$ is the $n \times (k+1)$ design matrix, and $\mathbf{W} = Cov(\mathbf{Y}) = diag(Var(Y_1), \ldots, Var(Y_n))$.
(2) The value of $(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$ is computed as an intermediate term in Newton's method.
(3) The diagonals yield the estimates of $Var(\hat{\beta}_j), j = 0, 1 \ldots, k$.
(4) 95 % confidence intervals for individual coefficients can be computed approximately as

$$\hat{\beta}_j \pm 1.96 \sqrt{Var(\hat{\beta}_j)}$$

(5) 95 % confidence interval for an arbitrary linear combination of the coefficients: $\mathbf{L}^T\beta = L_0\beta_0 + \ldots + L_k\beta_k$, is approximately

$$\mathbf{L}^T\hat{\beta} \pm 1.96\sqrt{\mathbf{L}^T Cov(\hat{\beta})\mathbf{L})}$$

(6) Further formulas are available for approximate *simultaneous* confidence intervals.
(7) *Saturated model* has as many parameters as there are units of observation ($= n$). It fits the data perfectly. N.B. there is a distinction here between Bernoulli and Binomial observations.
(8) Write the resulting probability estimates as $p_i^*, i = 1, \ldots, n$, and the corresponding loglikelihood as ℓ^*
(9) Write the loglikelihood corresponding to the MLE as $\ell(\hat{\beta})$.
(10) Def. $D(\hat{\beta}) = 2(\ell^* - \ell(\hat{\beta}))$ is the so-called *deviance*.

(11) Deviance does not have a known distribution, in general. An exception: the n_i's are large.
(12) On the other hand, write $\beta = (\beta_{(0)}^T, \beta_{(1)}^T)^T$, where $\beta_{(0)} = (\beta_0, \dots, \beta_r)^T$, and $\beta_{(1)} = (\beta_{r+1}, \dots, \beta_k)^T$. Consider the hypothesis $H_0 : \beta_{(1)} = \mathbf{0}$.
(13) When H_0 is true, then (under some regularity conditions)

$$D(\tilde{\beta}, \mathbf{0}) - D(\hat{\beta}) \sim \chi^2_{k-r}$$

asymptotically, as $n \to +\infty$.
(14) Change in deviance, or the *likelihood ratio test*, is the preferred way of testing adequacy of *nested* submodels.

(15) Causal research: keep variables of scientific interest in the model at least as long as to see what their maximal influence in the data at hand are. Report this.
(16) If there is a small number of candidate variables, one can put them all in the model to see what their combined explanatory power might be.
(17) An essential question: in what scale would the explanatory variables to exert an influence that conforms to the theory.
(18) There are non-parametric formulations for logistic regression, but they require a lot of data.
(19) In descriptive studies automated, stepwise procedures can be useful.
(20) *Information criteria* (AIC, BIC, DIC, NIC,...) can be used to distinguish models that are not nested, but only as rough guides, because their performance is inadequately known.