Applied Logistic Regression - Brief Notes

1. Introduction to Causal Modeling
2. Probability Models for Binary Responses
3. Logistic Regression
4. Interpretation of Coefficients
5. Model Choice
6. Goodness of Fit
7. Conditional Logistic Regression

1. Introduction to Causal Modeling

Statistical Model is a *simplified description* of a real phenomenon:

$$Y = f(X, \beta) + \epsilon$$

Or, there is *systematic part* and a *random residual*.

Response Y:
(1) dichotomous $Y = 0, 1$
(2) waiting time $Y \geq 0$
(3) real valued $Y \in \mathbf{R}$

Explanatory variables X:
(a) causal factors (is $\beta = 0$?)
(b) other characteristics

Theoretical validity of a causal model $\neq$ statistical fit

Example. Response Y, exposure A, confounder K

Experimental research: K is known $\rightarrow$ standardize; K unknown $\rightarrow$ randomize
Observational research: identify K from theory, measure K, condition on K

Main types of causal research:
(1) Cohort study
Def. *Cohort* is a group of individuals who have experienced the same phenomenon simultaneously
(2) Case-control study
Controls can be chosen from among those who would have become cases, had they fallen ill

Cohort design for rare exposures
Case-control design for rare illnesses

2. Probability Models for Binary Responses

(1) *Bernoulli* distribution: $Y \sim Ber(p)$ if $P(Y = 1) = p$ and $P(Y = 0) = 1 - p$, where $0 < p < 1$
Expectation: $E[Y] = p \cdot 1 + (1 - p) \cdot 0 = p$
Variance: $Var(Y) = E[Y^2] - E[Y]^2 = p(1 - p)$

(2) In general, $E[Y_1 + Y_2] = E[Y_1] + E[Y_2]$
and if Y_1 and Y_2 are *independent*, then
$Var(Y_1 + Y_2) = Var(Y_1) + Var(Y_2)$

(3) *Binomial* distribution: if $Y_i \sim Ber(p), i = 1, \ldots, n$ are independent and $Y = Y_1 + \ldots + Y_n$, then $Y \sim Bin(n, p)$ and
$P(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y}$ for $y = 0, \ldots, n$
Expectation: $E[Y] = np$
Variance: $Var(Y) = np(1 - p)$

(4) *Maximum likelihood estimation* based on $Y = y$: *likelihood function* is $L(p) = \binom{n}{y} p^y (1-p)^{n-y}$

(5) Maximize this (or its logarithm) as a function of p to get $\hat{p} = y/n$.

(6) This has *sampling distribution* $\hat{p} \sim N(p, p(1-p)/n)$.

(7) Testing equality of two independent populations:
exposed $(A = 1)$, $Y_1 \sim Bin(n_1, p_1)$
non-exposed $(A = 0)$, $Y_0 \sim Bin(n_0, p_0)$
Null hypothesis is $H_0 : p_1 = p_0$
Alternative hypothesis is $H_A : p_1 \neq p_0$

Confidence interval for $p_1 - p_0$ is $\hat{p}_1 - \hat{p}_0 \pm 1.96SE$, where *standard error* of the difference is $SE = \sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_0(1-\hat{p}_0)/n_0}$.

(8) The data can be presented in a 2×2 table:

$$\begin{array}{cc|c} a & b & n_1 \\ c & d & n_0 \\ \hline m_1 & m_0 & n \end{array}$$

where
$a = Y_1, b = n_1 - Y_1, c = Y_0, d = n_0 - Y_0, n = n_1 + n_0, m_0 = n - m_1$.
If we *condition* on $Y_0 + Y_1 = m_1$, then also m_0 is fixed, and a determines all other cells.

(9) Under the null hypothesis a has a *hypergeometric* distribution,
$P(a) = \binom{n_1}{a}\binom{n_0}{m_1 - a}/\binom{n}{m_1}$, for $max\{0, n_1 - m_0\} \leq a \leq min\{n_1, m_1\}$
Expectation: $E[a] = n_1 m_1 / n$
Variance: $Var(a) = n_1 n_0 m_1 m_0 / [n^2(n-1)]$
Normal *approximation*: $(a - E[a])/Var(a)^{1/2} \sim N(0, 1)$ under H_0

(10) In cohort analysis m_1 and m_0 are random, in case-control analysis n_1 and n_0 are random. Conditioning on all marginals leads to the same statistical analysis for both!

(11) *Odds* for the exposed are $= a/b$ and for the non-exposed they are $= c/d$. *Odds Ratio* is then $OR = ad/bc$.

(12) Partition the data into strata according to values $k = 1, \ldots, K$ of a possible confounder. The corresponding 2×2 tables are then

$$\begin{array}{cc|c} a_k & b_k & n_{1k} \\ c_k & d_k & n_{0k} \\ \hline m_{1k} & m_{0k} & n_k \end{array}$$

(13) Assume homogeneous odds ratios, $OR_k = OR, k = 1, \ldots, K$. The *Mantel-Haenszel estimator* for the common odds ratio is

$$OR_{MH} = \sum_{k=1}^{K} a_k d_k / n_k \Big/ \sum_{k=1}^{K} b_k c_k / n_k$$

(14) Suppose $H_0 : OR = 1$. The *Cochran-Mantel-Haenszel* test statistic is

$$X^2 = [\sum_{k=1}^{K} (a_k - E[a_k])]^2 \Big/ \sum_{k=1}^{K} Var(a_k) \sim \chi^2_1$$

when H_0 is true. (N.B. Often a *continuity correction* is used)

(15) What if K is big, or confounder is continuous?