MLDS 401/IEMS 404-1 (Fall 2023): Lab 6 – 11/07/2023

7.1 (2×2 **contingency table**) Consider a clinical trial to compare a new treatment (coded as $x = 1$) with a control (coded as $x = 0$) for some disease. Denote their success probabilities by p_1 and p_0. Suppose that there are n_0 patients in the control group of whom s_0 are successes and n_1 patients in the treatment group of whom s_1 are successes. The MLEs of p_0 and p_1 can be shown to be the sample proportions of successes $\widehat{p}_0 = s_0/n_0$ and $\widehat{p}_1 = s_1/n_1$.

a) Show that
$$\widehat{\beta}_1 = \ln\widehat{\psi} = \ln[\{\widehat{p}_1/(1-\widehat{p}_1)\}/\{\widehat{p}_0/(1-\widehat{p}_0)\}],$$
where $\widehat{\beta}_1$ is the MLE of β_1 in the logistic response model and $\ln\widehat{\psi}$ is the sample log-odds ratio.

b) Use the information matrix (7.6) to derive the formula:
$$\widehat{\mathrm{Var}}(\ln\widehat{\psi}) \approx \frac{1}{n_0\widehat{p}_0(1-\widehat{p}_0)} + \frac{1}{n_1\widehat{p}_1(1-\widehat{p}_1)}.$$

7.2 (**Nonconvergence of MLEs in logistic regression**) This exercise is based on Allison (2008). Consider completely separated data in the following table.

x	-5	-4	-3	-2	-1	+1	+2	+3	+4	+5
y	0	0	0	0	0	1	1	1	1	1

Because these data are symmetric, it can be shown that β_0 in the simple logistic regression model can be taken to be zero. So the likelihood function can be treated as a function only of the slope parameter β_1.

a) Write the log-likelihood function and plot it versus β_1 for these data and check that it approaches the maximum value of 0 (i.e., the likelihood function approaches the maximum value of 1) as $\beta_1 \to \infty$. So the MLE of β_1 does not exist and the algorithm to find it does not converge.

b) Next consider quasi-separated data obtained by adding two observations $(x, y) = (0, 0)$ and $(x, y) = (0, 1)$ to the above data set and repeat the exercise. Check that the log-likelihood function approaches a number less than 0 as $\beta_1 \to \infty$. So again the MLE of β_1 does not exist and the algorithm to find it does not converge.

7.8 (Radiation therapy) Twenty four cancer patients were treated with radiation therapy for different number of days (x) and the presence ($y = 0$) or absence ($y = 1$) of tumor was observed.

Days (x)	Response (y)	Days (x)	Response (y)
21	1	51	1
24	1	55	1
25	1	25	0
26	1	29	0
28	1	43	0
31	1	44	0
33	1	46	0
34	1	46	0
35	1	51	0
37	1	55	0
43	1	56	0
49	1	58	0

Source: Tanner (1996), p. 28.

a) Fit a binary logistic regression model to the data.

b) Calculate a 95% confidence interval for the odds of absence of tumor vs. presence of tumor if the number of days of therapy is increased by 5 days.

c) Calculate the estimated success probabilities $\widehat{p}_i$ for the 24 patients in the sample. Find the optimum threshold p^* that maximizes the correct classification rate (CCR). Calculate sensitivity, specificity and the F_1-score for this p^*.