MLDS 401/IEMS 404-1 (Fall 2023): Lab 7 – 11/14/2023

Question 1

You are working in sales. You keep calling you clients and track the number of calls you make until you get a sale:

```
12 19 33 17 36 22 11  2 75 13
```

You are going to fit a Geometric distribution, which has a probability mass function of

$$p(z) = (1-\theta)^{z-1}\theta \text{ if } z \geq 1.$$

What is the maximum likelihood estimate of θ?

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^*.

Question 4 Program Choice Example

The goal is to model the program choices of high school students based on a variety of attributes.

- Dependent variable: prog. The program choices are general program, vocational program and academic program.
- Attributes:
  - gender: female, male
  - ses: social/economic status (low, middle, high)
  - schtype: school type (private, public)
  - read
  - write
  - math
  - science
  - honors: enrolled, not enrolled
  - awards: number of awards received
- IDs: id, cid

1. Fit a multinomial regression with ses and write. Choose "academic" as the reference class.
2. If write is increased by 1 unit, how will the log-odds of being in general program vs. academic program change? What about the log-odds of being in vocation program vs. academic program?
3. If moving from ses = "low" to "high", how will the log-odds of being in general program vs. academic program change? What about moving from ses = "low" to "middle"?
4. Compute the z-statistics and p-values for the variables in the model. Are there any significant terms?
5. Split the data set into training (odd index) and test (even index) sets. Fit a multinomial regression using all the variables and the training set. Compute the trainning and test accuracies.