Exercises for MA 413 – Statistics for Data Science

This sheet will cover lecture material regarding GLMs.

1. Assume $Y_i|x_i$ is modelled using a generalized linear model with link function $g(x)$ that is either chosen as probit or logit, and where we set $\phi_i = \beta_0 + \beta_1 x_i$. Describe how to fit this model and discuss why probit or logit might be more appropriate for a given data set.

2. Refer to the given artificial data set:

x	no of trials	no of successes
0	4	1
1	4	2
2	4	4

   Denote by M_0 the logistic model of only an intercept and by M_1 the model which has a linear predictor. Denote the maximized log likelihood values by L_0 for M_0 and L_1 for M_1 as well as L_s for the saturated model. Recall that the deviance is twice the difference of the log-likelihood from the saturated log-likelihood. Create a data file in two ways, entering the data as i) ungrouped, 12 individual binary observations, ii) grouped data, three summary binomials with sample size four. The saturated model has 12 parameters, but the model with three groups has three parameters.

   (a) Fit M_0 and M_1 for each of the two data structures. Report L_0 and L_1 in both cases.

   (b) Show that the deviances depend on the form of the data entry.

3. According to the independent newspaper (London, March 8, 1994) the Metropolitan police reported 30475 people as missing in the year ending March 1993. For those aged 13 or less 33 of 3271 missing males, and 38 of 2486 missing females were still missing a year later. For ages 14-18 the values were 63 of 7256 males and 108 of 8877 females, while for ages 19 and above the values were 157 of 5065 males and 159 of 3520 females. Analyse this data set, check model fit and interpret the estimated parameters.

4. The following are true/false questions.

   (a) A model for a binary response has a continuous predictor. If the model truly holds the deviance statistic for the model has an asymptotic chi-squared distribution as the sample size increases. It can be used to test for goodness of fit.

   (b) For the model $\text{logit}(\pi(x)) = \alpha + \beta x$ suppose $y = 1$ for all $x \leq 50$ and $y = 0$ for all $x > 50$. The MLE estimate of β tends to negative infinity.