MAST90139: Statistical Modelling for Data Science
Assignment 2

Due time: 5 pm Friday May 7, 2021.

1. The data below come from a study of the relationship between beetle mortality and the concentration of an insecticide they were sprayed with. [10]

Group i	Dosage [$= \log_{10}$(dose)] x_i	Number of beetles in group n_i	Number killed y_i
1	1.69	59	6
2	1.72	60	13
3	1.76	62	18
4	1.78	56	28
5	1.81	63	52
6	1.84	59	53
7	1.86	62	61
8	1.88	60	60

   (a) Plot the empirical logits against dosage. Does the plot look linear?

   (b) Fit a (straight line) logistic regression model to the data, and give estimates of the intercept and slope.

   (c) Find an approximate 95% confidence interval for the slope of the logistic regression model.

   (d) Obtain an estimate of the dosage that will kill 50% of the beetles.

   (e) Obtain an estimate of, and an approximate 95% confidence interval for, the factor by which the odds of being killed increase for a **0.1** increase in dosage.

   (f) Obtain an estimate of, and an approximate 95% confidence interval for, the probability that a dosage of 1.8 is fatal.

   (g) Use both the (residual) deviance and the (Pearson) X^2 test to test the adequacy of fit of the model.

   (h) Plot the deviance residuals against dosage, and comment on the plot.

   (i) Carry out a formal test of whether a quadratic logistic regression model provides a better fit to the data than the straight line model.

More questions on next pages.

2. The following data were compiled from two general social surveys carried out in 1974 and 1975 by the National Opinion Research Center, University of Chicago. Part of each survey was concerned with the relationship of education and gender to attitudes towards the role of women in society, and each respondent was asked if they agreed or disagreed with the statement:
"Women should take care of running their homes and leave the running of the country up to men".

[10]

	Sex			
	Male		Female	
	Attitude			
Education	Agree	Disagree	Agree	Disagree
6	25	9	17	5
7	27	15	26	16
8	75	49	91	36
9	29	29	30	35
10	32	45	55	67
11	36	59	50	62
12	115	245	190	403
13	31	70	17	92
14	28	79	18	81
15	9	23	7	34
16	15	110	13	115
17	3	29	3	28

(a) Fit the strictly additive logistic regression model with nominal main effects (ie both education and gender treated as factors), and carry out a formal test of whether the effects of gender and years of education are (strictly) additive (on the logit scale).

(b) Find the "best" logistic regression model that uses the ordinal nature of years of education, and interpret your model in terms of odds ratios.

[**Hint:** start with the biggest model that includes gender and education (as a linear covariate).]

(c) Based on your findings in parts (a) and (b), and by carrying out any further analyses that may be useful, what model do you consider to be the most appropriate/most useful for these data.

One more question on next page.

3. The following data were published in a paper by Radelet (1981), *Amer Sociol. Rev. 918–927*, who studied the effect of racial characteristics on whether individuals convicted of homicide receive the death penalty. The 326 subjects were defendants in homicide cases in 20 Florida counties during 1976 – 1977. [10]

Defendant's Race	Victim's Race	Death Penalty Yes	No
White	White	19	132
	Black	0	9
Black	White	11	52
	Black	6	97

(a) Carry out tests of the following hypotheses using log-linear models:

i. All three factors are mutually independent.

ii. Sentence is independent of both the defendant's and the victim's race.

iii. Given the defendant's race, sentence is independent of the victim's race.

iv. Given the victim's race, sentence is independent of the defendant's race.

(b) Use logistic regression to test those hypotheses in (a) which can be tested using logistic regression models.

Total marks = 30