Assignment 3

Stat 623

Due on Oct 15, 2020

Turn in problems 1–3. Problem 5 is extra credit.

1. Plot the following data, i.e. y against x. Suppose you want to fit a logistic regression model to these data set. Superimpose how the fitted probabilities might look like on the plot. Next, using R command glm, fit a logistic regression to these observations. Explain the output carefully.

x	0	.5	1	2	3	4	5	5	5.5	6
y	0	0	0	0	0	1	1	1	1	1

2. When a family of average income was looking to buy a house, the family identified eight criteria which they thought they had to look for in a house. These criteria fall into three categories: economic, geographic and physical. Although one may have begun by examining the relative importance of these clusters, the family felt they wanted to prioritize the relative importance of all the criteria without working with clusters.

   The criteria important to the individual family were: (1) Size of the house, (2) Location to the bus lines, (3) Neighborhood, (4) Age of the house, (5) Yard space, (6) Modern facilities, (7) General condition (Repairs needed, walls, carpet, drapes, cleanliness, wiring) and (8) Financing available.

   As the next step, they elicited pairwise comparison judgments. Specifically, they elicited judgments from the people who had been in a similar situation before and understood the relative importance of these eight criteria with respect to the overall goal: 'Satisfaction with the House'. The data on pairwise comparison judgments are given in Table 1.

   Thus, Table 1 provides an 8×8 pairwise comparison matrix $R = r_{i,j}$ such that

   $$r_{j,i} = 1/r_{i,j}.$$

   Let $w_1, \ldots, w_8$ ($w_1 > 0, \ldots w_m > 0$) respectively be the true weights (i.e., importance) of the 8 criteria which are normalized so that

   $$w_1 + w_2 + \ldots + w_8 = 1.$$

Table 1: Pairwise comparison matrix

	1	2	3	4	5	6	7	8
1	1	5	3	7	6	6	1/3	1/4
2	1/5	1	1/3	5	3	3	1/5	1/7
3	1/3	3	1	6	3	4	6	1/5
4	1/7	1/5	1/6	1	1/3	1/4	1/7	1/8
5	1/6	1/3	1/3	3	1	1/2	1/5	1/6
6	1/6	1/3	1/4	4	2	1	1/5	1/6
7	3	5	1/6	7	5	5	1	1/2
8	4	7	5	8	6	6	2	1

It is obvious that the pairwise comparison $r_{i,j}$ provides information about w_i/w_j. Now write down the following model

$$y_{i,j} = \log r_{i,j} = \log w_i - \log w_j + \epsilon_{i,j} = \alpha_i - \alpha_j + \epsilon_{i,j} \quad 1 \leq i < j \leq 8 \tag{1}$$

with $\alpha_i = \log w_i$. Furthermore, assume that $\epsilon_{i,j}$ are i.i.d. $N(0, \sigma^2)$. Overall, the model that we consider here is the same model that we used for analysis of our electro-chemical metal voltage data. However, our purpose of inference is very different.

(a) Explain why $(\alpha_1, \ldots, \alpha_8)$ and $(\alpha_1+c, \ldots, \alpha_8+c)$ are equivalent as parameter points in the model. However, argue why we can still estimate the weights $w_1, \ldots, w_8$.

(b) Fit the model in (1) to the pairwise comparison data in Table 1. Report the estimates that you obtain from calling the appropriate cran-R function.

(c) Report the estimates of the weights $w_1, \ldots, w_8$.

(d) Assess the evidence for and against the hypothesis that the all 8 criteria are equally important.

(e) Perform residual analysis to check validity of model assumptions in (1) and summarize your conclusions.

3. Toxicity of insecticides: Flour beetles Tribolium castaneum were sprayed with one of

three insecticides in solution at different doses. The number of insects killed after a six-day period is recorded below:

	Deposit of insecticide ($mg/10cm^2$)					
Insecticide	2.00	2.64	3.48	4.59	6.06	8.00
DDT	3/50	5/49	19/47	19/38	24/49	35/50
γ-BHC	2/50	14/49	20/50	27/50	41/50	40/50
DDT +γ-BHC	28/50	37/50	46/50	48/50	48/50	50/50

(a) Investigate graphically the relationship between the dose, either in original units or in log units, and the kill rate.

(b) On the graph for part (a), plot the linear logistic fitted curve for each of the insecticides plus the combination.

(c) Consider the two models, one in which the relationship is described by three parallel straight lines in the log dose and and one in which the three lines are straight but not parallel. Assess the evidence against the hypothesis of parallelism.

(d) Let chem be a 3-level factor, and let ldose be the log dose. Explain the relationship between the regression coefficients in the model formulae chem + ldose and chem + ldose - 1. Explain the relationship between the two covariance matrices.

(e) On the assumption that three parallel straight lines suffice, estimate the potency of the combination relative to each of the components. Use FiellerÕs method to obtain a 90% confidence interval for each of these relative potencies.

(f) Check to see if one of the alternative link functions probit, c-log log or log log, gives an appreciably better fit. Give the answer to part (e) for the c-log log model.

(g) Under the linear logistic model, estimate the combination dose required to give a 99% kill rate, and obtain a 90% confidence interval for this dose.

(h) Give a brief summary of your conclusions regarding the effectiveness of these three insecticides.

4. Consider the probit regression model where response variables $y_1, \ldots y_n$ are independent with

$$y_i \sim Ber(p_i), \quad p_i = \Phi(x_i^T \beta).$$

Show that the MLE of β can be obtained using an iteratively reweighed least square method. In other words, the Fisher scoring method will take the form of

$$X^T W^{(k)} X \beta^{(k+1)} = X^T W^{(k)} z^{(k)}.$$

Identify the weight matrix $W^{(k)}$ and the linearized vector $z^{(k)}$.

5. Consider a logistic regression with binary response variable y and a single covariate x. Let $\hat{\beta}$ be the maximum likelihood estimate of the slope in a logistic regression of y on x with intercept. Further let $\bar{x}_0$ and $\bar{x}_1$ be the average of sample x values for cases with $y = 0$ and $y = 1$, respectively. Can you say that

$$sign(\hat{\beta}) = sign(\bar{x}_1 - \bar{x}_0)?$$

Justify your answer numerically and analytically.