MAST90104: A First Course in Statistical Learning

Week 9 Practical and Workshop

1 Practical questions

1. We revisit the milk data last week. We study the effect of various breeds and diets on the milk yield of cows. A study is conducted on 9 cows and the following data obtained:

	Diet		
Breed	1	2	3
1	18.8	16.7	19.8
	21.2		23.9
2	22.3	15.9	21.8
		19.2	

   (a) Input this data into R.

   (b) Test for the presence of interaction.

   (c) What is the degrees of freedom used for the interaction test?

   (d) From the interaction model, what is the estimated amount of milk produced from breed 2 and diet 3?

   (e) Find a 95% confidence interval under the interaction model, for the amount of milk produced from breed 2 and diet 3.

2. The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on 768 adult female Pima Indians living near Phoenix. The purpose of the study was to investigate factors related to diabetes. The data may be found in the the dataset `pima`.

   (a) This question use a data set in package `faraway`. Load the package and read the help file (`?pima`) to get a description of the predictor and response variables, then use `pairs` and `summary` to perform simple graphical and numerical summaries of the data.

   There are some obvious irregularities in the data. Take appropriate steps to correct the problems.

   (b) Fit a model with `test` as the response and all the other variables as predictors.

   Can you tell whether this model fits the data?

   Odds are sometimes a better scale than probability to represent chance. The odds o and probability p are related by

$$o = \frac{p}{1-p} \quad p = \frac{o}{1+o}$$

   In a binomial regression model with a logit link we have

$$\text{logit}(p_j) = \log\left(\frac{p_j}{1-p_j}\right) = \eta_j = \beta_0 + \beta_1 x_{1,j} + \cdots + \beta_q x_{q,j}.$$

   That is $\log o_j = \eta_j$, where o_j are the odds for the j-th observation.

   (c) By what proportion do the odds of testing positive for diabetes change for a woman with a BMI at the first quartile compared with a woman at the third quartile, assuming that all other factors are held constant? Give a confidence interval for this difference.

   (d) Do women who test positive have higher diastolic blood pressures? Is the diastolic blood pressure significant in the regression model? Explain the distinction between the two questions and discuss why the answers are only apparently contradictory.

   (e) Predict the outcome for a woman with predictor values 1, 99, 64, 22, 76, 27, 0.25, 25 (same order as in the dataset). Give a confidence interval for your prediction.

2 Workshop questions

1. Verify that for the binomial regression model with logistic link

$$\begin{aligned} \mathbb{E}\frac{\partial l(\boldsymbol{\theta};\mathbf{Y})}{\partial \theta_i} &= 0 \\ -\mathbb{E}\frac{\partial^2 l(\boldsymbol{\theta};\mathbf{Y})}{\partial \theta_i \partial \theta_j} &= \mathbb{E}\left(\frac{\partial l(\boldsymbol{\theta};\mathbf{Y})}{\partial \theta_i}\frac{\partial l(\boldsymbol{\theta};\mathbf{Y})}{\partial \theta_j}\right) \end{aligned}$$

2. Show that the Gamma density, f, in the form

$$f(y;\lambda,\alpha) = \frac{1}{\Gamma(\alpha)}\lambda^\alpha y^{\alpha-1}e^{-\lambda y}$$

is an exponential family with $\theta = -\frac{\lambda}{\alpha}, \phi = \frac{1}{\alpha}$. Identify the functions a, b, c and find the mean and variance functions as functions of θ.

3. Show that the inverse Gaussian density, f, in the form

$$f(y;\mu,\lambda) = \frac{\lambda}{\sqrt{2\pi y^3}}e^{-\frac{\lambda(y-\mu)^2}{2\mu^2 y}}$$

is an exponential family with $\theta = \frac{1}{\mu^2}, \phi = \frac{1}{\lambda}$. Identify the functions a, b, c and find the mean and variance functions as functions of μ, λ.