Statistical Modelling and Methods: Homework 4

Due by 5pm on May 29, online through Blackboard

Homework format: all homework must be written in latex. You must turn in both your tex and pdf files. Attach your code and computer output if there is any programming.

1. Use the technique of high-order Bartlett's identity to derive a recursion formula to calculate $E(Y-\mu)^k$ for any integer $k \geq 1$, where Y has a canonical exponential family distribution.

2. For each of the following probability distributions, determine if it is a member of the canonical exponential family. If yes, rewrite in canonical form, i.e., specify $\theta, b(\theta), a(\phi), c(y,\phi)$, and find EY, $\text{var}(Y)$ and variance function $v(\cdot)$ expressed with EY. If not, explain why.

   (a) Gamma distribution ($\gamma > 0$ is a nuisance parameter):

   $$f(y;\mu,\gamma) = \frac{1}{\Gamma(\gamma)y}(\frac{\gamma y}{\mu})^{\gamma}\exp(-\frac{\gamma y}{\mu}), \qquad y>0,\ \mu>0.$$

   (b) Inverse Gaussian distribution ($\sigma^2 > 0$ is a nuisance parameter):

   $$f(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2 y^3}}\exp(-\frac{(\mu-y)^2}{2\sigma^2\mu^2 y}).$$

   (c) Pareto distribution:

   $$f(y;\theta) = \theta y^{-\theta-1}, \qquad y>1,\ \theta>0.$$

   (d) Negative Binomial distribution (n is known):

   $$f(y;\theta) = \begin{pmatrix} n+y-1 \\ \\ y \end{pmatrix}(1-\theta)^y\theta^n, \qquad y=1,2,\ldots.$$

3. Consider the probability density function

   $$f(y) = \frac{\exp(y)}{\{1+\exp(y)\}^2}, \qquad -\infty < y < \infty.$$

(a) Show that its moment generating function is

$$M(t) = \frac{t\pi}{\sin(t\pi)}.$$

(Hint: you may need to use the identity $\Gamma(-x) = -\pi \csc(\pi x)/(x\Gamma(x))$.)

(b) Apply the *exponential tilting* technique to $f(y)$ to generate a "new" distribution family.

(c) Find the mean and variance of the generated distribution, expressed in terms of the canonical parameter θ.

(d) How do $\mu(\theta)$ and $v(\mu(\theta))$ behave as θ approaches zero, where $\mu(\theta) = E_\theta Y$ and $v(\cdot)$ is the variance function of the generated distribution?

4. Derive the observed and expected Fisher's information matrices with respect to β for a Poisson model with log link, where $\eta_i = \sum_{k=1}^{p} x_{ik}\beta_k$, $i = 1, \ldots, n$.

5. Derive the observed and expected Fisher's information matrices with respect to β for a Binomial model with logistic and probit links, respectively, where $\eta_i = \sum_{k=1}^{p} x_{ik}\beta_k$, $i = 1, \ldots, n$, and use the observed frequency as response.

6. In a dose-response experiment with dose level denoted by x, suppose a logistic link is used, find the explicit expression based on delta-method of the $100(1-\alpha)$th confidence interval for the dose level x_π such that the probability of response is a known value $\pi \in (0, 1)$ under x_π. What if a Probit link is used instead?

7. Iteratively weighted least squares programming:

   Write a program in R (or other software you are familiar with) to implement the iteratively weighted least squares algorithm. Include at least the predictor variables, dependent variable, choice of families and link functions as input arguments. You can program the algorithm only for Binomial family and logistic link function in this assignment. The variance function, link function and its derivative shall be obtained from self-written subroutines. Provide at least the estimates of the regression coefficients, the expected information matrix, as well as the Pearson, deviance residuals as output options. Provide a description of your algorithm with key formulas, and submit your code together.

   Use your own program to fit a logistic model with all predictor variables in LUNG.dat as described in Problem 4, not considering interaction and dispersion. *You only need to fit the full model using your own algorithm here. In problem 4, you can use the built-in GLM functions in R to do the analysis.*

8. Fit a logistic regression model to the data set LUNG.dat. Byssinosis is a lung disease of cotton workers and is related to the dustiness of the workplace. The data set has 7 columns, byssinosis yes (number), byssinosis no (number), dust level (1=High, 2=Middle, 3=Low), race, sex (1=Female, 2=Male), smoking (1=Smoker, 2=Non-smoker), length of employment (coded 1-3 for Short/Middle/Long). Each row of the data corresponds to one workplace containing a number of workers who are grouped according to their covariate levels. You do not need to consider interactions in this analysis. Select relevant predictors, discuss the influence of dispersion on the model selection, interpret your findings and the model estimates in the application context.

   Note: You are expected to write a concise data analysis report: JUSTIFY and summarize important steps in your analysis, and INTERPRET parameter estimates of the final model. Please only insert relevant outputs in the text and attach computer codes at the end. Do not exceed two pages, including inserted outputs but excluding computer codes. The report writing is taken into account for credit.