AMS 274 – Generalized Linear Models (Fall 2018)

Homework 1 (due Tuesday October 16)

1. The list below includes a number of distributions, providing in each case the probability density or mass function, support, and parameter space. Determine whether each of the distributions belongs to the exponential dispersion family. Similarly for the two-parameter exponential family of distributions. In both cases, justify your answers.

   (a) *Double exponential* (or *Laplace*) distribution.

   $$f(y \mid \theta, \sigma) = \frac{1}{2\sigma} \exp\left(-\frac{|y-\theta|}{\sigma}\right) \qquad y \in \mathbb{R},\ \ \theta \in \mathbb{R},\ \sigma > 0.$$

   (b) *Uniform* distribution.

   $$f(y \mid \theta, \sigma) = \frac{1}{2\sigma} \qquad \theta - \sigma < y < \theta + \sigma,\ \ \theta \in \mathbb{R},\ \sigma > 0.$$

   (c) *Logistic* distribution.

   $$f(y \mid \theta, \sigma) = \frac{\exp((y-\theta)/\sigma)}{\sigma \left\{1 + \exp((y-\theta)/\sigma)\right\}^2} \qquad y \in \mathbb{R},\ \ \theta \in \mathbb{R},\ \sigma > 0.$$

   (d) *Cauchy* distribution.

   $$f(y \mid \theta, \sigma) = \frac{1}{\pi\sigma \left\{1 + ((y-\theta)/\sigma)^2\right\}} \qquad y \in \mathbb{R},\ \ \theta \in \mathbb{R},\ \sigma > 0.$$

   (e) *Pareto* distribution.

   $$f(y \mid \alpha, \beta) = \frac{\beta \alpha^\beta}{y^{\beta+1}} \qquad y \geq \alpha,\ \ \alpha > 0,\ \beta > 0.$$

   (f) *Beta* distribution.

   $$f(y \mid \alpha, \beta) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{\mathrm{B}(\alpha, \beta)} \qquad 0 \leq y \leq 1,\ \ \alpha > 0,\ \beta > 0,$$

   where $\mathrm{B}(\alpha, \beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\,\mathrm{d}u$ is the beta function.

   (g) *Negative binomial* distribution.

   $$f(y \mid \alpha, p) = \frac{\Gamma(y+\alpha)}{\Gamma(\alpha) y!} p^\alpha (1-p)^y \qquad y \in \{0, 1, 2, ...\},\ \ \alpha > 0,\ 0 < p < 1,$$

   where $\Gamma(c) = \int_0^\infty u^{c-1} \exp(-u)\,\mathrm{d}u$ is the gamma function.

2. Consider the linear regression setting where the responses Y_i, $i = 1, ..., n$, are assumed independent with means $\mu_i = \mathrm{E}(Y_i) = \boldsymbol{x}_i^T\boldsymbol{\beta} = \sum_{j=1}^{p} x_{ij}\beta_j$ for (known) covariates x_{ij} and (unknown) regression coefficients $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$.

(i) Show that if the response distribution is normal,

$$Y_i \overset{ind.}{\sim} f(y_i \mid \mu_i, \sigma) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right), \quad i = 1, ..., n,$$

then the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$ is obtained by minimizing the L_2-norm,

$$S_2(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2.$$

(ii) Show that if the response distribution is double exponential,

$$Y_i \overset{ind.}{\sim} f(y_i \mid \mu_i, \sigma) = (2\sigma)^{-1} \exp\left(-\frac{|y_i - \mu_i|}{\sigma}\right), \quad i = 1, ..., n,$$

then the MLE of $\boldsymbol{\beta}$ is obtained by minimizing the L_1-norm,

$$S_1(\boldsymbol{\beta}) = \sum_{i=1}^{n} |y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}|.$$

(iii) Show that if the response distribution is uniform over the range $[\mu_i - \sigma, \mu_i + \sigma]$,

$$Y_i \overset{ind.}{\sim} f(y_i \mid \mu_i, \sigma) = (2\sigma)^{-1}, \text{ for } \mu_i - \sigma \leq y_i \leq \mu_i + \sigma, \quad i = 1, ..., n,$$

then the MLE of $\boldsymbol{\beta}$ is obtained by minimizing the L_∞-norm,

$$S_\infty(\boldsymbol{\beta}) = \max_i |y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}|.$$

(iv) Obtain the MLE of σ under each one of the response distributions in (i) – (iii) and show that, in all cases, it is a function of the minimized norm.

3. Consider the special case of the Cauchy distribution, $C(\theta, 1)$, with scale parameter $\sigma = 1$, and density function

$$f(y \mid \theta) = \frac{1}{\pi\{1 + (y - \theta)^2\}} \qquad y \in \mathbb{R}, \;\; \theta \in \mathbb{R},$$

where θ is the median of the distribution.

(a) Let $\boldsymbol{y} = (y_1, ..., y_n)$ be a random sample from the $C(\theta, 1)$ distribution. Develop the Newton-Raphson method and the method of scoring to approximate the maximum likelihood estimate of θ based on the sample $\boldsymbol{y}$. (For the method of scoring, you can use the result $\int_0^\infty (1 - x^2)/\{(1 + x^2)^3\}\, \mathrm{d}x = \pi/8$.)

(b) Consider a sample, assumed to arise from the $C(\theta, 1)$ distribution, with $n = 9$ and $\boldsymbol{y} = (-0.774,\ 0.597,\ 7.575,\ 0.397,\ -0.865,\ -0.318,\ -0.125,\ 0.961,\ 1.039)$. Apply both methods from (a) to estimate θ. To check your results, try a few different starting values and also plot the likelihood function for θ.

(c) Now consider a sample (again, assumed to arise from the $C(\theta, 1)$ distribution) with $n = 3$ and $\boldsymbol{y} = (0,\ 5,\ 9)$. Apply again the methods from (a) to estimate θ, using three different starting values, $\theta^0 = -1$, $\theta^0 = 4.67$, $\theta^0 = 10$. Comment on the results.

4. The data in the table below show the number of cases of AIDS in Australia by date of diagnosis for successive 3-months periods from 1984 to 1988.

	Quarter			
Year	1	2	3	4
1984	1	6	16	23
1985	27	39	31	30
1986	43	51	63	70
1987	88	97	91	104
1988	110	113	149	159

Let $x_i = \log(i)$, where i denotes the time period for $i = 1, ..., 20$. Consider a GLM for this data set based on a Poisson response distribution with mean μ, systematic component $\beta_1 + \beta_2 x_i$, and logarithmic link function $g(\mu) = \log(\mu)$.

(a) Fit this GLM to the data working from first principles, that is, derive the expressions that are needed for the scoring method, and implement the algorithm to obtain the maximum likelihood estimates for β_1 and β_2.

(b) Use function "glm" in R to verify your results from part (a).