STA 711 Homework 3

Due: Friday, February 9, 11:00am on Canvas.

Instructions: Submit your work as a single PDF. Your document should be created using LaTeX; see the course website for a homework template file and instructions on getting started with LaTeX and Overleaf.

Looking towards Generalized Linear Models

So far, you have seen linear regression and logistic regression models, which each deal with different types of response variables. The same ideas can be generalized to a wider variety of response distributions; this is exactly what a *generalized linear model* does. Linear regression (with normal errors), logistic regression, Poisson regression, gamma regression, and other models are all examples of generalized linear models.

In this part of the assignment, we will work towards defining a generalized linear model, with linear and logistic regression as special cases.

1. To define a generalized linear model, we need a general form for the distribution of our response variable. Consider a random variable (either discrete or continuous) with probability function

$$f(y;\theta,\phi) = a(y,\phi)\exp\left\{\frac{y\theta-\kappa(\theta)}{\phi}\right\},$$

where $\theta \in \mathbb{R}$ is called the *canonical parameter*, and $\phi > 0$ is called the *dispersion parameter*. The function $a(y,\phi)$ depends only on y and ϕ (not on θ). The function $\kappa(\theta)$ depends only on θ (not on y or ϕ).

A probability function which can be written in this way is called an *exponential dispersion model* (EDM). We will describe each piece of $f(y;\theta,\phi)$ in the problems below.

   (a) Show that the $N(\mu,\sigma^2)$ pdf can be written in the EDM form: identify θ, ϕ, and κ.

   (b) Show that the $Bernoulli(p)$ pmf can be written in the EDM form: identify θ, ϕ, and κ.

   (c) Show that the $Poisson(\lambda)$ pmf can be written in the EDM form: identify θ, ϕ, and κ.

2. Suppose that $Y \sim f(y;\theta,\phi)$ comes from an exponential dispersion model. Let $\mu = \mathbb{E}[Y]$; then $\theta = g(\mu)$ for some function g.

   (a) If $Y \sim N(\mu,\sigma^2)$, find the function g (in terms of μ).

   (b) If $Y \sim Bernoulli(p)$, find the function g (in terms of p).

   (c) If $Y \sim Poisson(\lambda)$, find the function g (in terms of λ).

3. The *cumulant generating function* (CGF) for a random variable Y is given by $C(t) = \log M(t) = \log \mathbb{E}[e^{tY}]$ (i.e., the log of the moment generating function).

(a) Using properties of the MGF, show that

$$\frac{d}{dt}C(t)\Big|_{t=0} = \mathbb{E}[Y]$$
$$\frac{d^2}{dt^2}C(t)\Big|_{t=0} = Var(Y)$$

(b) Show that if the distribution of Y is an exponential dispersion model (with probability function $f(y;\theta,\phi)$ above), then

$$C(t) = \frac{\kappa(\theta + t\phi) - \kappa(\theta)}{\phi}$$

(c) Suppose that the distribution of Y is an exponential dispersion model (with probability function $f(y;\theta,\phi)$ above), and let $\mu = \mathbb{E}[Y]$. Show that

$$\mu := \mathbb{E}[Y] = \frac{d}{d\theta}\kappa(\theta)$$

and

$$Var(Y) = \phi\frac{d^2}{d\theta^2}\kappa(\theta)$$

(For this reason, κ is called the *cumulant* function, because derivatives of κ are related to the cumulants of Y).

4. Suppose we observe $Y_1, ..., Y_n \overset{iid}{\sim} f(y;\theta,\phi)$ from an exponential dispersion model. Let $\mu = \mathbb{E}[Y]$; then $\theta = g(\mu)$ for some continuous, monotone increasing g. Find the maximum likelihood estimates $\widehat{\mu}$ and $\widehat{\theta}$ (one or both of your answers might involve g).

5. Suppose that $(X_1, Y_1), ..., (X_n, Y_n)$ are an iid sample from the following model:

$$Y_i \sim EDM(\theta_i, \phi)$$
$$\theta_i = g(\mu_i) = \beta^T X_i,$$

where $Y_i \in \mathbb{R}$ is the ith response, $X_i \in \mathbb{R}^p$ is the vector of observed covariates for the ith individual, and $\beta \in \mathbb{R}^p$ is the vector of regression coefficients.

So: the distribution of Y_i is an exponential dispersion model; the dispersion parameter ϕ is the same for all Y_i; and the parameter θ_i (and so also the mean $\mu_i = \mathbb{E}[Y_i|X_i]$) depends on X_i.

(a) Using your answers to questions 1 and 2, explain how linear regression (with normal errors) and logistic regression are both special cases of the model described here.

(b) To estimate the coefficient vector β, we will use maximum likelihood estimation. Generally, there is no closed-form solution, so we must resort to an iterative algorithm like Newton's method. Show that

$$U(\beta|X,Y) = \frac{1}{\phi}X^T(Y - \mu)$$
$$\mathbf{H}(\beta|X,Y) = -X^T W X$$

where X is the design matrix, $Y = (Y_1, ..., Y_n)^T \in \mathbb{R}^n$ is the vector of observed responses, $\mu = (\mu_1, ..., \mu_n)^T$ with $\mu_i = g^{-1}(\beta^T X_i)$, and $W = \frac{1}{\phi^2}\text{diag}(Var(Y_i))$.