Finger Exercises To Be Discussed Fri Nov 1, 2002

1. **Logistic regression / theory.** One formulation of the logistic regression model is as follows. Given data $y_1, \ldots, y_n$ with corresponding covariate (row) vectors $x_1^T, \ldots, x_n^T$, the model is that

$$y_i | x_i \overset{indep}{\sim} \text{Bin}\left(1, \frac{1}{1+\exp[-x_i^T\theta]}\right) \quad , \; i = 1, \ldots, n$$

Do the following, starting from our discussion of scoring in generalized linear models in class. Compare with Monahan, p. 218.

(a) Show that the gradient of the log-likelihood (the score function) for logistic regression is

$$\nabla \ell_n(\theta) = X^T[Y - \mu(\theta)]$$

where

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}_{n \times p}$$

is the matrix of covariates (model matrix or design matrix as in linear regression), $Y_{n\times 1} = (y_1, \ldots, y_n)^T$ is the data (a column vector of 0's and 1's),

$$\mu(\theta) = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} \frac{1}{1+\exp[-x_1^T\theta]} \\ \frac{1}{1+\exp[-x_2^T\theta]} \\ \vdots \\ \frac{1}{1+\exp[-x_n^T\theta]} \end{bmatrix}_{n\times 1}$$

is the vector of means ($p_i = E[y_i|\theta]$), and $\theta = (\theta_1, \ldots, \theta_p)^T$ is the vector of "regression coefficients".

(b) Show that the negative Hessian and the Fisher Information for logistic regression are the same value, $X^T\Sigma X$:

$$-\nabla^2 \ell_n(\theta) = I(\theta) = X^T \Sigma X$$

where X is as before and now

$$\Sigma = \begin{bmatrix} p_1(1-p_1) & 0 & \cdots & 0 \\ 0 & p_2(1-p_2) & \cdots & 0 \\ \vdots & \ddots & \vdots & 0 \\ 0 & \cdots & 0 & p_n(1-p_n) \end{bmatrix}$$

2. **Logistic regression / example.** Cox (1970) presented the following data on ingot manufacture.

Condition c	Number of trials n_c	Number of successes y_c	Covariate x_{c1}	Covariate x_{c2}
1	55	0	1	7
2	157	2	1	14
3	159	7	1	27
4	16	3	1	57

   (a) Formulate the problem as a logistic regression problem (trials are independent of each other within and between conditions; covariates are the same for all trials within a single condition. What is the total number of trials?).

   (b) Find the MLE's of the regression coefficeints, and their standard errors, using either the methods of problem 2 or using a general optimizer like `ms()` in Splus or R.

   (c) [*Optional.*] Compare your results to what you would get from `glm()` in Splus or R.

3. **Delta Method.** The *multivariate delta method* is a way of obtaining approximate standard errors for transformed MLE's. Suppose $\hat{\theta}$ is the MLE of the p-dimensional parameter θ, and suppose $\tau = g(\theta)$ is some r-dimensional transformation of θ. Then by Taylor's theorem,

$$\hat{\tau} - \tau = g(\hat{\theta}) - g(\theta) \approx \nabla g(\theta)(\hat{\theta} - \theta)$$

and so, approximately, the variance-covariance matrix of $\hat{\tau}$ may be related to the variance-covariance matrix of $\hat{\theta}$:

$$\text{Var}_\tau(\hat{\tau}) = \nabla g(\theta)\text{Var}_\theta(\hat{\theta})\nabla g(\theta)^T .$$

Multiplying the Taylor expansion by $\sqrt{n}$, we get a similar result relating asymptotic normality of $\sqrt{n}(\hat{\tau} - \tau)$ to $\sqrt{n}(\hat{\theta} - \theta)$. (*Note:* If $g(\theta)$ is an $r \times 1$ vector valued function of the $p \times 1$ vector variable θ, then $\nabla g(\theta) = \left[\frac{\partial g_i}{\partial g_j}\right]_{r \times p}$.)

   (a) The fitted model $\hat{P}[Y = 1|x_1, x_2] = e^{\hat{\theta}_1 x_1 + \hat{\theta}_2 x_2}$ can be used to estimate the probability of success at various levels (x_1, x_2) of the covariates. Compute this probability, for $x_1 = 1$, $x_2 = 42$ in problem 3. Use the multivariate delta method to compute an approximate 95% confidence interval for this probability.

   (b) Plot the fitted response curve $\hat{P}(x) = \hat{P}[Y = 1|x_1 = 1, x_2 = x]$, over the range $x = 0$ to $x = 100$. On the same graph, plot an approximate 95% confidence envelope (that is, plot the curves corresponding to the upper and lower edpoints of approximate 95% confidence intervals for $P(x)$ at each value of x, using the multivariate delta method. (This can be done with a single `matplot(...,type="l")` command, or with several `plot` and `lines` commands).

4. **Exponential families / theory.** The following items were needed in the discussion of the scoring algorithm for exponential family models.

   (a) Review assumptions and argument needed to assert that, when $\theta = (\theta_1, \ldots, \theta_p)$,

   $$I(\theta) = E_\theta[-\nabla^2 \ell_n(\theta)] = \text{Var}_\theta[\nabla \ell_n(\theta)] = E_\theta[\nabla_\theta \ell_n(\theta)^2]$$

   where $\ell_n(\theta) = \log \prod_i f(y_i|\theta)$ is the log-likelihood of iid data $y_1, \ldots, y_n$.
   (*Notation:* $E_\theta[\cdot] = E[\cdot|\theta]$, i.e. it is the expectation over the data $y_1, \ldots, y_n$ for fixed parameter value θ.)

   (b) Let $L(\theta) = g(Y) \exp[B(\theta) + K(Y)^T \gamma(\theta)]$ be an exponential family likelihood, with $\mu(\theta) = E_\theta[K(Y)]$ and $\Sigma(\theta) = \text{Var}_\theta[K(Y)]$. Complete the argument (assuming that integration and differentiation can be interchanged) that shows

   $$\begin{aligned} \nabla \mu(\theta) &= \nabla \int K(Y) L(\theta) d\nu(Y) \\ &= \cdots \\ &= \Sigma(\theta) \nabla \gamma(\theta) \end{aligned}$$

   Note: If $\mu(\theta)$ is an $r \times 1$ vector valued function of the $p \times 1$ vector variable θ, then

   $$\nabla \mu(\theta) = \left[\frac{\partial \mu_i}{\partial \theta_j} \right]_{r \times p}$$