Ex 4.4

We will begin by extending the log-likelihood to the K-class case. We will find a general version of 4.20.

$$\ell(\theta) = \sum_{i=1}^{N} \log P_{g_i}(x_i;\theta)$$

Where $P_K(x_i;\theta) = P(G=K \mid X=x_i;\theta)$

$$\Rightarrow \ell(\theta) = \sum_{i=1}^{N} y_1 \log P_1(x_i;\theta) + \dots + y_{K-1} \log P_{K-1}(x_i;\theta) + (1 - y_1 - \dots - y_{K-1}) \log P_K(x_i;\theta)$$

Where $\begin{cases} y_K = 1 & \text{if observation } i \text{ in } g_K \\ y_K = 0 & \text{otherwise} \end{cases}$

Notice: $y_j \log P_j(x_i;\theta) - y_j \log P_K(x_i;\theta) = y_j \log \frac{P_j(x_i;\theta)}{P_K(x_i;\theta)}$

$$\Rightarrow \ell(\theta) = \sum_{i=1}^{N} \left(\sum_{j=1}^{K-1} \left(y_j \log \frac{P_j(x_i;\theta)}{P_K(x_i;\theta)} \right) + \log P_K(x_i;\theta) \right)$$

And from the definition of logistic regression this gives:

$$\ell(\beta) = \sum_{i=1}^{N} \left(\sum_{j=1}^{K-1} \left(y_i \beta_j^T x_i \right) + \log \left(\quad \right) P_K(x_i;\beta) \right)$$

Where $\beta_j = \{\beta_{j,0}, \beta_{j,1}, \dots, \beta_{j,p}\}$ and each observation x_i has a first element 1 corresponding to an intercept.

$$\ell(\beta) = \sum_{i=1}^{N} \left(\sum_{j=1}^{K-1} \left(y_i \beta_i^T x_i \right) + \log \left(\quad \right) P_K(x_i;\beta) \right)$$

Now, as in the two class case, we wish to apply Newton-Raphson. The first term can be written in terms of a $(K-1)(P+1)$ β vector, however the second term cannot. Thus, unlike the 2-class case we will need to take partial derivatives and write in

non-diagonal Hessian Form.

To calculate the derivatives we will need to know the derivatives of the logistic log Probabilities:

① $$\frac{\partial \log P_k(x_i;\beta)}{\partial \beta_m} = \frac{\partial}{\partial \beta_m} \log\left(\frac{1}{1+\sum_{\ell=1}^{k-1} \exp(\beta_\ell^T x_i)}\right)$$

$m \in 1, 2, \ldots, K-1$

$$= \frac{-x_i \exp(\beta_m^T x_i)}{1+\sum_{\ell=1}^{k-1} \exp(\beta_\ell^T x_i)} = -x_i P_m(x_i;\beta)$$

② $$\frac{\partial P_m(x_i;\beta)}{\partial \beta_m} = \frac{\partial}{\partial \beta_m}\left(\frac{\exp(\beta_m^T x_i)}{1+\sum_{\ell=1}^{k-1} \exp(\beta_\ell^T x_i)}\right)$$

$$= \frac{\exp(\beta_m^T x_i)\, x_i}{1+\sum_{\ell=1}^{k-1} \exp(\beta_\ell^T x_i)} - \frac{\exp(\beta_m^T x_i)^2\, x_i}{\left[1+\sum_{\ell=1}^{k-1} \exp(\beta_\ell^T x_i)\right]^2}$$

$$= P_m(x_i;\beta)\left[1 - P_m(x_i;\beta)\right] x_i$$

③ $$\frac{\partial P_N(x_i;\beta)}{\partial \beta_m} = \frac{\partial}{\partial \beta_m}\left(\frac{\exp(\beta_N^T x_i)}{1+\sum_{\ell=1}^{K-1} \exp(\beta_\ell^T x_i)}\right)$$

where $m \neq N$

$$= \frac{-\exp(\beta_N^T x_i)\cdot \exp(\beta_m^T x_i)\, x_i}{\left[1+\sum_{\ell=1}^{K-1} \exp(\beta_\ell^T x_i)\right]^2}$$

$$= -P_N(x_i;\beta) P_m(x_i;\beta) x_i$$

Now we may calculate the First and Second order derivatives of $\ell(\beta)$ as in the two-class case. We can use these (Partial) derivitives to calculate the Hessian matrix to Fill in Newton-Raphson.

$$\ell(\beta) = \sum_{i=1}^{N} \left(\sum_{j=1}^{K-1} (y_j \beta_j^T x_i) + \log P_K(x_i;\beta) \right)$$

1st derivative

$$\frac{\partial \ell(\beta)}{\partial \beta_m} = \sum_{i=1}^{N} y_m x_i - P_m(x_i;\beta) x_i \qquad \text{(using ①)}$$

2nd derivative

(a) $$\frac{\partial^2 \ell(\beta)}{\partial \beta_m \partial \beta_m^T} = - P_m(x_i;\beta)[1 - P_m(x_i;\beta)] x_i^T x_i \qquad \text{(using ②)}$$

(b) $$\frac{\partial^2 \ell(\beta)}{\partial \beta_m \partial \beta_N^T} = P_N(x_i;\beta) P_m(x_i;\beta) x_i^T x_i \qquad \text{(using ③)}$$

where $M \neq N$

Now we have all the pieces, so we now can put them in matrix notation.

1st derivative

(*) $$\underset{p \times 1}{\frac{\partial \ell(\beta)}{\partial \beta_m}} = \underset{p \times N}{X^T} \underset{N \times 1}{(\bar{y}_m} - \underset{N \times 1}{\bar{P}_m)}$$ (Note this does the $\sum_{i=1}^{N}$ for each column as required)

Where $\bar{y}_m = \begin{bmatrix} y_{m,1} \\ \vdots \\ y_{m,N} \end{bmatrix}$ $y_{m,i} = 1$ if obs i is of class m

and $\bar{P}_m = \begin{bmatrix} P_m(x_1;\beta) \\ \vdots \\ P_m(x_N;\beta) \end{bmatrix}$

This is for a single class m. We can write in block format to later multiply by Hessian.

$$\frac{\partial \ell(\beta)}{\partial \beta} = \begin{bmatrix} X^T(\bar{y}_1 - \bar{P}_1) \\ \vdots \\ X^T(\bar{y}_{K-1} - \bar{P}_{K-1}) \end{bmatrix}$$

2nd derivative

(a) $$\frac{\partial^2 \ell(\beta)}{\partial \beta_m \partial \beta_m} = -X^T P_m^* X$$

where P_m^* is an $N \times N$ diagonal matrix with the i^{th} diagonal element equal to: $P_m(x_i;\beta)[1 - P_m(x_i;\beta)]$

(b) $$\frac{\partial^2 \ell(\beta)}{\partial \beta_m \partial \beta_r} = X^T P_{m,r}^{**} X$$

where $P_{m,r}^{**}$ is an $N \times N$ diagonal matrix with the i^{th} diagonal element equal to: $P_m(x_i;\beta) P_r(x_i;\beta)$

Now we can write in Hessian form

(**) $$\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} = \begin{bmatrix} -X^T P_1^* X & X^T P_{1,2}^{**} X & \cdots & X^T P_{1,K-1}^{**} X \\ X^T P_{2,1}^{**} X & -X^T P_2^* X & \cdots & X^T P_{2,K-1}^{**} X \\ \vdots & \vdots & \ddots & \vdots \\ X^T P_{K-1,1}^{**} X & X^T P_{K-1,2}^{**} X & \cdots & -X^T P_{K-1}^* X \end{bmatrix}$$

Finally, combining (*) and (**) we can solve (iteratively) the Newton-Raphson equation.

$$\beta^{new} = \beta^{old} - \left(\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} \right)^{-1} \frac{\partial \ell(\beta)}{\partial \beta}$$