Homework 4

Due Thursday, 2/18/20 on Canvas.

Treat the data analysis problem as a report to a client, and do NOT turn in any R, SAS or python code. For example, when reporting coefficient estimates, your answer should just be $\hat{\beta} = 2.47$ and NOT

```
> fit$coefficients[2]
[1] 2.469829453
```

It is up to you as to how many decimal places to report, but please be reasonable and consistent. Two or three decimal places is sufficient for most applications. You should only report what is necessary, i.e. estimates, confidence intervals, P values, plots, etc., in a clear and concise manner.

1. In the Bradley-Terry model for ranking k competitors, parameters $\theta_1, \ldots, \theta_k$ representing 'abilities' are introduced in such a way that $\pi_{ij} = P(\text{competitor } i \text{ beats } j)$ is a function of the difference in their abilities. In the logit model, we have

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \theta_i - \theta_j, \quad i \neq j \text{ and } i, j = 1, \ldots, k. \tag{1}$$

   Suppose that $k = 7$ teams compete in a round-robin tournament, with each pair competing at least once. Assume all of the games are independent. Let m_{ij} be the number of games in which teams i and j play each other, and define the observations Y_{ij} to be the number of games in which team i beats team j.

   (a) What is the distribution of Y_{ij} in terms of m_{ij} and π_{ij}?

   (b) Let $\ell_{ij}\left(\theta_i, \theta_j; Y_{ij}\right)$ be the log-likelihood for the distribution defined in part (a). Show that

$$\ell_{ij}\left(\theta_i, \theta_j; Y_{ij}\right) = Y_{ij}\left(\theta_i - \theta_j\right) - m_{ij}K\left(\theta_i - \theta_j\right) + h\left(Y_{ij}\right)$$

   for some functions $K(t)$ and $h(t)$. Find an expression for $K(t)$.

   (c) Write out the 21×7 design matrix $\boldsymbol{X}$ for the Bradley-Terry model.

   (d) Now suppose each team plays all of the other teams at their home ballpark at least once. Extend the Bradley-Terry model in (1) to include a home field advantage effect (that is the same for each team). What is the new design matrix $\boldsymbol{X}$?

2. Consider the general exponential family with density

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

   where ϕ is called the dispersion parameter and θ is the canonical paramter.

   (a) What are θ, ϕ and $K(\theta)$ if $Y \sim \text{Poisson}(\mu)$ for $\mu > 0$? How about if $Y \sim N\left(\mu, \sigma^2\right)$?

(b) Suppose $Y \sim f(y; \theta, \phi)$. Find expressions for $\mathbb{E}(Y)$ and $\text{Var}(Y)$ in terms of $K(\theta)$ and ϕ.

(c) Suppose $Y_i \sim f\left(y; \boldsymbol{x}_i^T \boldsymbol{\beta}, \phi\right)$, where both $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\phi > 0$ are unknown. If $Y_1, \ldots, Y_n$ are independent, show that the MLE for $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, satisfies

$$\boldsymbol{X}^T \left\{ \boldsymbol{Y} - \mathbb{E}_{\hat{\boldsymbol{\beta}}}(\boldsymbol{Y}) \right\} = \boldsymbol{0}_p$$

where $\mathbb{E}_{\hat{\boldsymbol{\beta}}}(Y_i)$ is the expectation of Y_i assuming $Y_i \sim f\left(y; \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}, 1\right)$.

(d) What is the asymptotic variance of $\hat{\boldsymbol{\beta}}$ (i.e. the inverse of the Fisher information matrix) in terms of $K, \boldsymbol{\beta}, \boldsymbol{X}$ and ϕ?

3. You observe an outcome Y_i that can take on values 1,2, and 3 and a continuous covariate x_i. Consider fitting the proportional odds model where:

$$P(Y_i = j) = \pi_{ij} > 0, \quad j = 1, 2, 3, \quad i = 1, \ldots, n.$$

$$\log\left(\frac{\pi_{i1}}{\pi_{i2} + \pi_{i3}}\right) = \alpha_1 + \beta x_i,$$

$$\log\left(\frac{\pi_{i1} + \pi_{i2}}{\pi_{i3}}\right) = \alpha_2 + \beta x_i.$$

(a) What is π_{i2} in terms of $\alpha_1, \alpha_2, \beta$ and x_i?

(b) Show that $\alpha_1 \leq \alpha_2$.

(c) The file Q1c.txt on Canvas contains a sample data set with $n = 200$. Using the data, fit two separate logistic regression models:

(Mi) $logit(\pi_{i1}) = \beta_{01} + \beta_{11} x_i$,

(Mii) $logit(\pi_{i1} + \pi_{i2}) = \beta_{02} + \beta_{12} x_i$.

Using R, make two plots:

(i) On the same axes, plot the estimated probabilities from (Mi) and the estimated probabilities from (Mii) as a function of x_i.

(ii) On the same axes, plot the logit of the estimated probabilities from (Mi) and the logit of the estimated probabilities from (Mii) as a function of x_i.

(d) The file Q1d.txt on the Courseweb also contains a sample data set with $n = 200$. Repeat part (c) using this data set.

(e) From the plots in parts (c) and (d), are you more conformable fitting the proportional odds model to Q1c.txt or Q1d.txt and why?