STAT 668
Homework 3
Assigned: February 25, 2019
Due: March 11, 2019

1. The dataset "dog-racing.csv" contains information about the outcomes of greyhound races at different racetracks. The columns are indexed as follows:

   - startix: a unique index for each row (not relevant to this analysis)
   - ndate: date of the race in YYYYMMDD format
   - trk: abbreviation of the racetrack where the race took place
   - meet: 'A' or 'P' indicates whether the race was during the morning meet (A) or afternoon meet (P) at the stated track
   - race: the race number during the meet
   - raceix: unique identifier of each race
   - hid: unique identifier for the dog whose results are recorded in the row
   - fin: final position of dog in the race
   - post0: post position
   - nh1: number of dogs in the race

   For now consider only whether a dog finished in 1st place ('fin = 1') or not. Write $Y_{d,r}$ for the binary variable (0 or 1) indicating whether dog d finished in 1st place in race r. When fitting your model, only include races with 'raceix' less than or equal to 195000.

   (a) Write down a model for $Y_{d,r}$ in which all dogs and races are treated as exchangeable. Fit this model and save the output for future use.

   (b) Write down a hierarchical model that attributes a different effect to each post position. Fit the model. In a single figure, plot the posterior distributions of probability of winning from each post position.

   (c) Write down a model that treats the post position as a linear effect. In particular, specify a Bayesian model in which the likelihood is given by

   $$\text{logit}(\Pr(\text{win} \mid \text{post0})) = \beta \times \text{post0}.$$

   Fit this model.

   (d) Finally, fit a model which allows the effect of post position to vary with track.

   (e) Compare the above 4 models by assigning a win probability to dogs in the left out races (those with raceix 195001 and greater). Which model performs best? How might this model be improved further?

   Now consider each race as a unit. The Plackett–Luce model is a family of probability distributions on the top-k rankings of a set. For this dataset, we take $k = 3$ and fit the top 3 finishers in each race as follows. To each dog i, we assign a weight w_i and assume the probability that

the first three finishers are $i_1 < i_2 < i_3$ in a race between dogs with weights $\mathbf{w} = (w_1, \ldots, w_N)$ is

$$\Pr(i_1 < i_2 < i_3 \mid \mathbf{w}) = \left(\frac{e^{w_{i_1}}}{\sum_{j=1}^{N} e^{w_j}}\right)\left(\frac{e^{w_{i_2}}}{\sum_{j \neq i_1} e^{w_j}}\right)\left(\frac{e^{w_{i_3}}}{\sum_{j \neq i_1, i_2} e^{w_j}}\right).$$

In particular, this model assumes that the ranking of the remaining finishers in a race, given the top finishers, is independent of the weights of the top finishers.

(a) Fit a hierarchical model in which the weight w_i of each dog depends only on its post position.

(b) Fit a model in which the weight w_i depends linearly on post position and is allowed to vary depending on the track t. In particular, specify a model with likelihood depending on the weights w_i by

$$w_{i,r,t} \mid \text{post0}_{i,r}, \beta_t = \beta_t \times \text{post0}_{i,r},$$

where $w_{i,r,t}$ is the weight of dog i in race r at track t.

(c) Compare these models based on the logarithmic scoring rule fit to the left out races.

(d) Are there any other models you would propose based on the given data? Fit this model to the data and compare based on predictive accuracy.

Conclude this problem with a discussion of the models fit above and their relative performances for predicting race outcomes. Discuss possible extensions or ways to improve the model. What additional data would be helpful for improving predictive accuracy?

]