Lecture 24 — The Bradley-Terry model

In the remaining lectures, we will explore how the methods of statistical inference that we developed for the setting of n IID observations $X_1, \ldots, X_n \overset{IID}{\sim} f(x|\theta)$ may be applied to other types of data and statistical models. Each of the next five lectures will introduce a statistical model using a different motivating example, and then use our tools from Unit 2 to solve several inference questions in this example.

A parametric model for a data vector $\mathbf{Y}$ (not necessarily consisting of IID coordinates) is a specification of the joint distribution of $\mathbf{Y}$ in terms of a small number of parameters θ. The likelihood $\text{lik}(\theta) = f(\mathbf{Y}|\theta)$ is the joint PMF or PDF of $\mathbf{Y}$ viewed as a function of θ. The log-likelihood is $l(\theta) = \log \text{lik}(\theta)$, and the MLE $\hat{\theta}$ is the value of θ that maximizes $\text{lik}(\theta)$.

To extend the theory of maximum likelihood and Fisher information to the non-IID setting, note that for IID data $X_1, \ldots, X_n \overset{IID}{\sim} f(x|\theta)$, we may introduce the notation

$$I_{\mathbf{X}}(\theta) := nI(\theta) = \sum_{i=1}^{n} -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f(X_i|\theta) \right] = -\mathbb{E}_\theta[l''(\theta)],$$

which represents the total Fisher information of all n observations $\mathbf{X} = (X_1, \ldots, X_n)$. Our main theorem regarding the MLE $\hat{\theta}$ states that it is approximately distributed as $\mathcal{N}(\theta_0, \frac{1}{n} I(\theta_0)^{-1}) = \mathcal{N}(\theta_0, I_{\mathbf{X}}(\theta_0)^{-1})$ for large n if the parametric model is correct and the true parameter is θ_0. For non-IID data and the general log-likelihood $l(\theta) = \log f(\mathbf{Y}|\theta)$, let us define

$$I_{\mathbf{Y}}(\theta) = -\mathbb{E}_\theta[l''(\theta)]$$

in the single-parameter case $\theta \in \mathbb{R}$ and

$$I_{\mathbf{Y}}(\theta) = -\mathbb{E}_\theta[\nabla^2 l(\theta)]$$

in the multi-parameter case $\theta \in \mathbb{R}^k$, where

$$\nabla^2 l(\theta) = \left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} l(\theta) \right)_{1 \le i,j \le k}$$

is the second-derivative (Hessian) matrix for $l(\theta)$. In all of the non-IID settings we will consider, under appropriate asymptotic conditions, the approximate sampling distribution of $\hat{\theta}$ is still given by the (multivariate) normal distribution $\mathcal{N}(\theta_0, I_{\mathbf{Y}}(\theta_0)^{-1})$ when the total sample size is large. We will appeal to this approximation without proof in our examples.

Start Here

24.1 The Bradley-Terry model

Example 24.1. There are 30 basketball teams in the NBA, each playing 82 games in the regular season (so there are 1230 total games). We observe, at the end of the regular season,

which two teams (i, j) played in each game, and whether team i or team j won. How can we rank the teams and/or determine the strength of each team?

The simplest strategy might be to compare the number of games won by each team. However, the NBA season is structured so that every team plays every other team a different number of times (between 2 and 4). So the teams have different "strengths of schedule", meaning that some teams play stronger opponents more frequently than do other teams. These teams might have worse win-loss records, but in fact be better than other teams that won more games against weaker opponents.

A model-based approach to address this problem is the following: Let $\beta_i \in \mathbb{R}$ represent the "strength" of team i, and let the outcome of a game between teams (i, j) be determined by $\beta_i - \beta_j$. The **Bradley-Terry model** treats this outcome as an independent Bernoulli random variable with distribution Bernoulli(p_{ij}), where the log-odds corresponding to the probability p_{ij} that team i beats team j is modeled as

$$\log \frac{p_{ij}}{1 - p_{ij}} = \beta_i - \beta_j.$$

Equivalently, solving for p_{ij} yields

$$p_{ij} = \frac{e^{\beta_i - \beta_j}}{1 + e^{\beta_i - \beta_j}} = \frac{e^{\beta_i}}{e^{\beta_i} + e^{\beta_j}}.$$

This model is over-parametrized in the sense that it is exactly the same if we add a fixed constant c to all values β_i, because the differences $\beta_i - \beta_j$ remain unchanged. We may fix this problem by setting $\beta_i \equiv 0$ for a particular team, for example $\beta_{\text{Warriors}} \equiv 0$. Then for every other team j, $\beta_j = \beta_j - 0$ represents the log-odds that team j beats the Warriors.

If we always order each pair (i, j) so that team i is the home team and j is the away team, then we may incorporate a home-court advantage by including an intercept term α:

$$\log \frac{p_{ij}}{1 - p_{ij}} = \alpha + \beta_i - \beta_j,$$

or equivalently

$$p_{ij} = \frac{e^{\alpha + \beta_i - \beta_j}}{1 + e^{\alpha + \beta_i - \beta_j}}. \tag{24.1}$$

This increases the log-odds of the home team winning in every game by a constant value α.

More generally, the Bradley-Terry model assigns scores to a fixed set of items based on pairwise comparisons of these items, where the log-odds of item i "beating" item j is given by the difference of their scores. An intercept term may be included to account for a systematic difference between the first and second item of each comparison.

24.2 Statistical inference

Let $k = 30$ be the number of NBA teams, and denote the Warriors as team 1. We might be interested in the following inferential tasks:

- Estimate the home-court advantage α and the team strengths $\beta_1, \ldots, \beta_k$ (constraining, say, $\beta_1 = \beta_{\text{Warriors}} \equiv 0$)
- Test the null hypothesis of no home-court effect, $\alpha = 0$
- Obtain a confidence interval for $\beta_i - \beta_j$ for two particular teams (i, j)

Suppose we observe n total games $(i_1, j_1), \ldots, (i_n, j_n)$ between these k teams, where each (i, j) is a pair of distinct teams in $\{1, \ldots, k\}$ and the home team is team i. Let $Y_1, \ldots, Y_n \in \{0, 1\}$ be such that $Y_m = 1$ if i_m beat j_m in the mth game and $Y_m = 0$ otherwise. The likelihood for the parameters $\theta = (\alpha, \beta_2, \ldots, \beta_k)$ is then given by

$$\text{lik}(\alpha, \beta_2, \ldots, \beta_k) = \prod_{m=1}^{n} p_{i_m j_m}^{Y_m} (1 - p_{i_m j_m})^{1 - Y_m} = \prod_{m=1}^{n} (1 - p_{i_m j_m}) \left(\frac{p_{i_m j_m}}{1 - p_{i_m j_m}} \right)^{Y_m},$$

where p_{ij} is given as a function of α, β_i, and β_j by equation (24.1) and we set $\beta_1 \equiv 0$. The log-likelihood is

$$\begin{aligned} l(\alpha, \beta_2, \ldots, \beta_k) &= \sum_{m=1}^{n} Y_m \log \left(\frac{p_{i_m j_m}}{1 - p_{i_m j_m}} \right) + \log(1 - p_{i_m j_m}) \\ &= \sum_{m=1}^{n} Y_m (\alpha + \beta_{i_m} - \beta_{j_m}) - \log(1 + e^{\alpha + \beta_{i_m} - \beta_{j_m}}). \end{aligned} \tag{24.2}$$

To estimate the parameters $\theta = (\alpha, \beta_2, \ldots, \beta_k)$ using the MLE, we set the partial derivative with respect to each parameter $\alpha, \beta_2, \ldots, \beta_k$ equal to 0:

$$0 = \frac{\partial l}{\partial \alpha} = \sum_{m=1}^{n} Y_m - \frac{e^{\alpha + \beta_{i_m} - \beta_{j_m}}}{1 + e^{\alpha + \beta_{i_m} - \beta_{j_m}}} \tag{24.3}$$

$$0 = \frac{\partial l}{\partial \beta_i} = \sum_{m : i_m = i} \left(Y_m - \frac{e^{\alpha + \beta_{i_m} - \beta_{j_m}}}{1 + e^{\alpha + \beta_{i_m} - \beta_{j_m}}} \right) + \sum_{m : j_m = i} \left(-Y_m + \frac{e^{\alpha + \beta_{i_m} - \beta_{j_m}}}{1 + e^{\alpha + \beta_{i_m} - \beta_{j_m}}} \right). \tag{24.4}$$

This yields a system of k equations in the k unknowns $\alpha, \beta_2, \ldots, \beta_k$, which may be solved numerically using the Newton-Raphson algorithm. The solution is the MLE $\hat{\theta} = (\hat{\alpha}, \hat{\beta}_2, \ldots, \hat{\beta}_k)$. end here

To test the null hypothesis $H_0 : \alpha = 0$, we may use the generalized likelihood ratio test (GLRT): Under the sub-model where $\alpha = 0$, the log-likelihood function is

$$l(\beta_2, \ldots, \beta_k) = \sum_{m=1}^{n} Y_m (\beta_{i_m} - \beta_{j_m}) - \log(1 + e^{\beta_{i_m} - \beta_{j_m}}),$$

and the system of score equations satisfied by the sub-model MLE is

$$0 = \frac{\partial l}{\partial \beta_i} = \sum_{m : i_m = i} \left(Y_m - \frac{e^{\beta_{i_m} - \beta_{j_m}}}{1 + e^{\beta_{i_m} - \beta_{j_m}}} \right) + \sum_{m : j_m = i} \left(-Y_m + \frac{e^{\beta_{i_m} - \beta_{j_m}}}{1 + e^{\beta_{i_m} - \beta_{j_m}}} \right)$$

for $i = 2, \ldots, k$. We may solve these equations using Newton-Raphson to obtain the submodel MLEs $\hat{\beta}_{2,0}, \ldots, \hat{\beta}_{k,0}$. The GLRT of $\alpha = 0$ is based on the test statistic

$$-2 \log \Lambda = -2 \log \frac{\text{lik}(0, \hat{\beta}_{2,0}, \ldots, \hat{\beta}_{k,0})}{\text{lik}(\hat{\alpha}, \hat{\beta}_2, \ldots, \hat{\beta}_k)},$$

and an approximate level-0.05 test rejects H_0 when $-2 \log \Lambda > \chi^2_1(0.05)$. (The number of degrees of freedom is 1 because the full model has one more parameter, α, than the submodel.)

We may obtain a confidence interval for $\beta_i - \beta_j$ by centering it around $\hat{\beta}_i - \hat{\beta}_j$, and estimating the standard error of $\hat{\beta}_i - \hat{\beta}_j$. Let us first consider the sampling distribution of the entire vector of MLE estimates $\hat{\theta} = (\hat{\alpha}, \hat{\beta}_2, \ldots, \hat{\beta}_k)$. When the number of total games n is large, this is approximately $\mathcal{N}(\theta, I_{\mathbf{Y}}(\theta)^{-1})$, where $I_{\mathbf{Y}}(\theta) = -\mathbb{E}_\theta[\nabla^2 l(\theta)]$. The Hessian matrix $\nabla^2 l(\theta)$ may be computed by differentiating the right sides of the score equations (24.3) and (24.4) a second time with respect to the variables $\alpha, \beta_2, \ldots, \beta_k$. (We will do this explicitly for the more general logistic regression model two lectures from now.) It is easy to see that $\nabla^2 l(\theta)$ is a constant quantity that does not involve $Y_1, \ldots, Y_n$, so $I_{\mathbf{Y}}(\theta) = -\nabla^2 l(\theta)$.

Finally, since $\hat{\beta}_i - \hat{\beta}_j$ is a linear combination of the coordinates of $\hat{\theta}$, it is approximately normal when $\hat{\theta}$ is approximately multivariate normal. Its mean is $\mathbb{E}[\hat{\beta}_i - \hat{\beta}_j] \approx \beta_i - \beta_j$, and its variance is

$$\begin{aligned}
\text{Var}[\hat{\beta}_i - \hat{\beta}_j] &= \text{Cov}[\hat{\beta}_i - \hat{\beta}_j, \hat{\beta}_i - \hat{\beta}_j] \\
&= \text{Var}[\hat{\beta}_i] + \text{Var}[\hat{\beta}_j] - 2\,\text{Cov}[\hat{\beta}_i, \hat{\beta}_j] \\
&\approx (I_{\mathbf{Y}}^{-1}(\theta))_{ii} + (I_{\mathbf{Y}}^{-1}(\theta))_{jj} - 2(I_{\mathbf{Y}}^{-1}(\theta))_{ij}.
\end{aligned}$$

We may estimate the standard error of $\hat{\beta}_i - \hat{\beta}_j$ by the plug-in estimate

$$\hat{\text{se}}_{ij} = \sqrt{(I_{\mathbf{Y}}^{-1}(\hat{\theta}))_{ii} + (I_{\mathbf{Y}}^{-1}(\hat{\theta}))_{jj} - 2(I_{\mathbf{Y}}^{-1}(\hat{\theta}))_{ij}}.$$

A 95% confidence interval for $\beta_i - \beta_j$, assuming correctness of the Bradley-Terry model, is then given by $\hat{\beta}_i - \hat{\beta}_j \pm z(0.025)\hat{\text{se}}_{ij}$.

We will discuss, two lectures from now, some alternative estimates of the standard error that are robust to model misspecification.