DS-GA 3001.009 Applied Statistics: Homework #2 Solutions

Due on Thursday, September 28, 2023

Please hand in your homework via Gradescope (entry code: RKXJN2) before 11:59 PM.

1. Prove Hoeffding's formula in class: let $y_1, \cdots, y_n$ be iid observations from an exponential family, and $\widehat{\theta}_n$ be the MLE of θ based on $(y_1, \cdots, y_n)$. Then

$$nD(\widehat{\theta}_n; \theta) = 2\log \frac{p_{\widehat{\theta}_n}(y_1, \cdots, y_n)}{p_\theta(y_1, \cdots, y_n)}$$

holds for any θ, where $D(\cdot; \cdot)$ is the deviance in the exponential family.

Solution: It holds that

$$\begin{aligned} 2\log \frac{p_{\widehat{\theta}_n}(y_1, \cdots, y_n)}{p_\theta(y_1, \cdots, y_n)} &= 2\sum_{i=1}^n \log \frac{p_{\widehat{\theta}_n}(y_i)}{p_\theta(y_i)} \\ &= 2\sum_{i=1}^n \log \frac{\exp(\langle \widehat{\theta}_n, T(y_i)\rangle - A(\widehat{\theta}_n))h(y_i)}{\exp(\langle \theta, T(y_i)\rangle - A(\theta))h(y_i)} \\ &= 2n\left(A(\theta) - A(\widehat{\theta}_n) - \left\langle \theta - \widehat{\theta}_n, \frac{1}{n}\sum_{i=1}^n T(y_i)\right\rangle\right) \\ &= 2n\left(A(\theta) - A(\widehat{\theta}_n) - \left\langle \theta - \widehat{\theta}_n, \nabla A(\widehat{\theta}_n)\right\rangle\right) = nD(\widehat{\theta}_n; \theta). \end{aligned}$$

2. Recall Fisher's 2×2 table in class, but this time we use a multinomial model $(X_1, \cdots, X_4) \sim \text{Multi}(N; (\pi_1, \cdots, \pi_4))$ to fit the data. It is easy to verify that the MLE $\widehat{\pi} = (\widehat{\pi}_1, \cdots, \widehat{\pi}_4)$ is given by $\widehat{\pi}_i = X_i/N$, for all $i = 1, 2, 3, 4$.

   (a) Based on the definition of $\widehat{\pi}_i$, directly verify that

$$\text{Cov}(\widehat{\pi}_i, \widehat{\pi}_j) = \begin{cases} \frac{\pi_i(1-\pi_i)}{N} & \text{if } i = j, \\ -\frac{\pi_i \pi_j}{N} & \text{if } i \neq j. \end{cases}$$

   (*Hint: Recall how to compute the variance of the Binomial distribution.*)

   (b) For the log odds estimator $\widehat{\theta} = \log \frac{\widehat{\pi}_1 \widehat{\pi}_4}{\widehat{\pi}_2 \widehat{\pi}_3}$, using the Delta method and the plug-in approach, show that

$$\mathsf{Var}(\widehat{\theta}) \approx \sum_{i=1}^4 \frac{1}{X_i}.$$

   (c) If $(X_1, X_2, X_3, X_4) = (9, 12, 7, 17)$, compute $\widehat{\theta}$ and the approximation of $\mathsf{Var}(\widehat{\theta})$ in (b). Compare your results with Fisher's hypergeometric modeling in class.

Solution:

(a) Let $Z_1, \cdots, Z_n \in \{1,2,3,4\}$ be iid observations from the distribution $(\pi_1, \cdots, \pi_4)$, then

$$\widehat{\pi}_i = \frac{X_i}{N} = \frac{1}{N}\sum_{k=1}^{N} \mathbb{1}(Z_k = i).$$

Consequently,

$$\begin{aligned}
\mathrm{Cov}(\widehat{\pi}_i, \widehat{\pi}_j) &= \frac{1}{N^2}\sum_{k=1}^{N}\sum_{\ell=1}^{N} \mathrm{Cov}(\mathbb{1}(Z_k = i), \mathbb{1}(Z_\ell = j)) \\
&= \frac{1}{N^2}\sum_{k=1}^{N} \mathrm{Cov}(\mathbb{1}(Z_k = i), \mathbb{1}(Z_k = j)) \\
&= \frac{1}{N^2}\sum_{k=1}^{N} (\mathbb{P}(Z_k = i, Z_k = j) - \mathbb{P}(Z_k = i)\mathbb{P}(Z_k = j)) \\
&= \frac{\pi_i \mathbb{1}(i = j) - \pi_i\pi_j}{N}.
\end{aligned}$$

(b) Write $\theta = f(\pi_1, \cdots, \pi_4) = \log \frac{\pi_1\pi_4}{\pi_2\pi_3}$, it holds that

$$\nabla_\pi f = \left(\frac{1}{\pi_1}, -\frac{1}{\pi_2}, -\frac{1}{\pi_3}, \frac{1}{\pi_4}\right).$$

In (a) we have shown that $\mathrm{Cov}(\widehat{\pi}) = (\mathrm{diag}(\pi) - \pi\pi^\top)/N$, the Delta method gives

$$\begin{aligned}
\mathsf{Var}(\widehat{\theta}) &\approx (\nabla_\pi f)\mathrm{Cov}(\widehat{\pi})(\nabla_\pi f)^\top \\
&= \sum_{i=1}^{4}\sum_{j=1}^{4} Z_{ij}\left(\frac{1}{\pi_i\pi_j}\cdot\frac{\pi_i\mathbb{1}(i=j) - \pi_i\pi_j}{N}\right) \\
&= \sum_{i=1}^{4}\frac{1}{N\pi_i} \approx \sum_{i=1}^{4}\frac{1}{N\widehat{\pi}_i} = \sum_{i=1}^{4}\frac{1}{X_i},
\end{aligned}$$

where $Z_{ij} = 1$ if i, j both belong to $\{1,4\}$ or $\{2,3\}$, and $Z_{ij} = -1$ otherwise.

(c) Plugging in the numbers, we have

$$\begin{aligned}
\widehat{\theta} &= \log\frac{9\cdot 17}{12\cdot 7} \approx 0.600, \\
\mathsf{Var}(\widehat{\theta}) &\approx \frac{1}{9} + \frac{1}{12} + \frac{1}{7} + \frac{1}{17} \approx 0.396.
\end{aligned}$$

The answer is very similar to the hypergeometric modeling results in class.

3. Coding I: we will verify numerically that in Poisson models, the deviance residual looks more normally distributed than the Pearson residual. Based on the inline instructions, fill in the missing codes in `https://tinyurl.com/nhezb6cu`. Be sure to submit a pdf with your codes, outputs, and colab link.

   Solution: see `https://tinyurl.com/y88hvfap`.

4. Coding II: in this problem, we investigate if a newly discovered poem is indeed written by Shakespeare, replicating the paper "Did Shakespeare write a newly discovered poem?" by Thisted and Efron in 1987. To this end, we collect 884,647 total words of known Shakespeare, and count the following:

   - there are 9 words in the poem which never appears in known Shakespeare;
   - there are 7 words in the poem which exactly appears once in known Shakespeare;
   - there are 5 words in the poem which exactly appears twice in known Shakespeare;
   - $\cdots$

   We use a data vector $y = (9, 7, 5, 8, 11, 10, 21, 16, 18, 8, 5)$ to record these numbers.

   On the other hand, using a theory based on empirical Bayes (we will cover it in Lecture 7), a statistician may predict the number of new words in the poem if Shakespeare indeed wrote it, and similarly the number of words which appear once in known Shakespeare, etc. These predictions are presented by a vector

   $$\theta = (6.97, 4.21, 3.33, 5.36, 10.24, 13.96, 10.77, 8.87, 13.77, 9.99, 7.48).$$

   The key assumption here is that, if Shakespeare indeed wrote this poem, then

   $$y_i \overset{\text{ind.}}{\sim} \text{Poi}(\theta_i), \quad \forall i = 0, 1, \cdots, 10.$$

   This problem aims to test this null hypothesis using the tests we learned in class. Based on the inline instructions, fill in the missing codes in `https://tinyurl.com/yn7tjtxp`. Be sure to submit a pdf with your codes, outputs, and colab link.

   Solution: see `https://tinyurl.com/yc3r7fd5`.