Conditional likelihood

Patrick Breheny

November 8

Introduction

- Today we're going to discuss an alternative approach to likelihood-based inference called conditional likelihood
- The main idea is that while the data may depend on both our parameters of interest $\boldsymbol{\theta}$ and nuisance parameters $\boldsymbol{\eta}$, perhaps we can transform the data in such a way that we can factor the likelihood into a conditional distribution depending only on $\boldsymbol{\theta}$

Conditional likelihood: Definition

- Specifically, suppose we can transform the data x into v and w such that

$$p(x|\boldsymbol{\theta}, \boldsymbol{\eta}) = p(v|w, \boldsymbol{\theta})p(w|\boldsymbol{\theta}, \boldsymbol{\eta})$$

- The first term, $L(\boldsymbol{\theta}) = p(v|w, \boldsymbol{\theta})$, is known as the *conditional likelihood*; note that this term is free of nuisance parameters
- Note that, unlike the profile likelihood, the conditional likelihood *is* an actual likelihood, in the sense that it corresponds to an actual distribution of observed data

Factorization

- Note that in our partition of the probability model, we have

$$p(x|\boldsymbol{\theta}, \boldsymbol{\eta}) = L_1(\boldsymbol{\theta})L_2(\boldsymbol{\theta}, \boldsymbol{\eta})$$

- With conditional likelihood, we are proposing to use only L_1 for inference, even though our parameter of interest $\boldsymbol{\theta}$ also shows up in L_2
- Is this valid?
- Absolutely; there is no requirement that we use all of the data in order for likelihood-based inference to be valid
- Is it a good idea, though?

When conditional likelihood is appealing

- This depends on how much information we are losing (not always easy to measure)
- In general, conditional likelihood is appealing when either of the following conditions are met:
  - The conditional likelihood is simpler than the original or profile likelihood
  - The original or profile likelihood leads to biased or unstable estimates
- No matter how much simpler the conditional likelihood is, however, conditional likelihood is not going to be attractive if substantial information is being lost

Regression

- For example, the most widespread use of conditional likelihood is probably in regression analysis
- It is often the case that both the predictor $\mathbf{X}$ and the outcome $\mathbf{y}$ are random variables
- We could, in principle, try to specify the joint distribution of $\mathbf{X}$ and $\mathbf{y}$, but there would be many parameters involved in defining the distribution of $\mathbf{X}$ and these parameters are not of interest in regression
- By considering instead the conditional distribution (i.e., the conditional likelihood) of $\mathbf{y}|\mathbf{X}$, these nuisance parameters are eliminated

Poisson model

- To get a sense of how conditional likelihood works, let's consider the case of two independent Poisson random variables:

$$X \sim \text{Pois}(\lambda)$$
$$Y \sim \text{Pois}(\mu)$$

and suppose that we are interested in the relative risk $\theta = \mu/\lambda$

- One way of approaching this problem would be to derive the full likelihood $L(\lambda, \mu)$, then use likelihood theory and the delta method to derive the distribution of θ:

$$\frac{\hat{\theta} - \theta}{\text{SE}} \xrightarrow{\text{d}} \text{N}(0, 1),$$

where $\text{SE}^2 = (\mu^2 + \mu\lambda)/\lambda^3$, as $\mu, \lambda \to \infty$

Conditional likelihood

- However, suppose we instead let $t = x + y$ and then proceeded along these lines:

$$\begin{aligned} p(x, y|\lambda, \mu) &= p(y, t|\lambda, \mu) \\ &= p(y|t, \lambda, \mu)p(t|\lambda, \mu) \end{aligned}$$

- The second term, we will just ignore; the first term is the conditional likelihood
- Writing the conditional likelihood in terms of θ, we have

$$L(\theta) = \left(\frac{1}{1+\theta}\right)^x \left(\frac{\theta}{1+\theta}\right)^y;$$

note that this likelihood is free of nuisance parameters

Orthogonal parameters

- Are we losing information about θ?
- In this particular case, we are losing nothing: letting $\eta = \lambda + \mu$, we can write

$$L(\theta, \eta) = L_1(\theta) L_2(\eta)$$

- In other words, θ does not show up in the part of the likelihood that we are ignoring
- When such a factorization exists, the parameters θ and η are said to be *orthogonal parameters*

Estimation and inference

- Now we can just carry out all the usual likelihood operations on the conditional likelihood
- The score is

$$u(\theta) = y/\theta - t/(1+\theta),$$

so $\hat{\theta} = y/x$, which seems like the obvious estimator

- The information, in this case, yields the same approximate variance as the delta method

$$\mathcal{I}(\theta) = \frac{y}{\theta^2} - \frac{t}{(1+\theta)^2},$$

Exact inference

- In the Poisson case, however, we don't really need asymptotic approximations, as we can carry out exact inference based on the conditional relationship

$$Y|T \sim \text{Binom}(T, \tfrac{\theta}{1+\theta})$$

- Exact tests and confidence intervals for the binomial proportion could then be constructed and transformed to give confidence intervals for θ
- This is often true, generally speaking, for conditional likelihood approaches: non-asymptotic methods are often available, albeit not always so easily calculated

Profile likelihood

- Yet another way of approaching this problem is to derive the profile likelihood of θ
- In this case, we end up with the same likelihood as the conditional approach:

$$L(\theta) = \left(\frac{1}{1+\theta}\right)^x \left(\frac{\theta}{1+\theta}\right)^y$$

- This is only true in the case of orthogonal parameters, however (i.e., only if the nuisance parameters can be factored out does the profile likelihood automatically produce a conditional likelihood)

Binomial proportions

- Another very common application of conditional likelihood is for comparing two binomial proportions: $X \sim \text{Binom}(n_1, \pi_1)$ and $Y \sim \text{Binom}(n_2, \pi_2)$, and our interest is in the odds ratio θ
- By conditioning on the total $T = X + Y$, we arrive at a conditional distribution for $X|T$ containing only the odds ratio that we can use as our conditional likelihood:

$$p(x|t) = \frac{\binom{n_1}{x}\binom{n_2}{t-x}\theta^x}{\sum_{s=0}^{t}\binom{n_1}{s}\binom{n_2}{t-s}\theta^s}$$

Information loss

- Unlike the earlier Poisson case, however, here the parameters are not orthogonal (the parameter of interest cannot be entirely factored apart from other parameters)
- Thus, there is the possibility of information loss
- Assessing the information loss would depend on how π_1 and π_2 are related to one another
- Intuitively, however, it seems unlikely that the total of X and Y can carry much meaningful information about the odds ratio unless we are willing to make very strong assumptions

Connection with hypergeometric distribution

- Returning to the conditional likelihood, at $\theta = 1$ the conditional distribution is the hypergeometric distribution
- Thus, we could carry out non-asymptotic inference on the basis of this distribution; this is known as Fisher's exact test
- We could also use any of our asymptotic likelihood approaches

Score test

- The score test is particularly convenient to apply, since the likelihood is simplified considerably at the null hypothesis $\theta = 1$
- Letting μ and σ denote the mean and standard deviation of the (n_1, n_2, t) hypergeometric distribution, the score test statistic is

$$z = \frac{x - \mu}{\sigma}$$

- Confidence intervals would involve the use of noncentral hypergeometric distributions

Matched pairs, binary outcome

- On a related note, let's consider the question of matched pairs of subjects with a binary outcome (essentially, this is a discrete version of the Neyman-Scott problem)
- Suppose we have n pairs of observations with Y_{i1} and Y_{i2} representing independent binary outcomes, and our probability model is

$$\begin{aligned}\text{logit}(\pi_{i1}) &= \alpha_i \\ \text{logit}(\pi_{i2}) &= \alpha_i + \beta;\end{aligned}$$

  this would arise, for example, in a study of identical twins where one was exposed to a risk factor and the other was not

Profile likelihood bias

- Our interest is the odds ratio e^β, but as in the Neyman-Scott problem, the number of nuisance parameters is growing with n
- This causes problems with the profile likelihood: letting a denote with number of $\{Y_{i1} = 1, Y_{i2} = 0\}$ pairs and b denote with number of $\{Y_{i1} = 0, Y_{i2} = 1\}$ pairs,

$$\begin{aligned} \hat{\alpha}_i(\beta) &= -\beta/2 \\ \widehat{\beta} &= 2\log\frac{b}{a} \\ \widehat{\mathrm{OR}} &= \left(\frac{b}{a}\right)^2 \end{aligned}$$

- The estimator (b/a) is known to be consistent, so the MLE here converges to OR^2, highly biased if $\mathrm{OR} \neq 1$

Conditional likelihood to the rescue

- Using conditional likelihood, however, this problem is avoided
- Within each table, we can condition on $y_{i1} + y_{i2}$, arriving at a Bernoulli distribution if the pair is informative
- Since pairs are independent of each other, the total likelihood is then

$$\ell(\theta) = \sum_i \ell_i(\theta)$$

- The result is that b has a binomial likelihood conditional on $a + b$ and the MLE is now consistent
- In this context, the score test is known as McNemar's test

General 2×2 tables

- The same logic works for more general 2×2 tables
- Here, each table's conditional likelihood corresponds to the hypergeometric distribution and the log-likelihood from these tables are again additive
- Again, the score test is particularly convenient:

$$z = \frac{\sum_i (x_i - \mu_i)}{\sqrt{\sum_i \sigma_i^2}},$$

  where μ_i and σ_i^2 are the mean and variance of the hypergeometric distribution for table i
- This is known as the Mantel-Haenzel test

Generality of conditional likelihood

- So, is conditional likelihood a general method, or only available in specialized cases?
- To some extent, both
- On the one hand, it is always possible to derive a conditional likelihood for exponential families; however, the resulting likelihood is often rather complicated

Exponential family: Setup

- Letting $\mathbf{v} = \mathbf{s}_1(x)$ and $\mathbf{w} = \mathbf{s}_2(x)$ denote the sufficient statistics of the exponential family,

$$p(\mathbf{v}, \mathbf{w}) = \exp\{\boldsymbol{\theta}^\top \mathbf{v} + \boldsymbol{\eta}^\top \mathbf{w} - \psi(\boldsymbol{\theta}, \boldsymbol{\eta})\} f_0(x)$$

- To derive the conditional likelihood, we first need to derive the marginal distribution of $\mathbf{w}$
- We can obtain this by summing (or integrating) $p(\mathbf{v}, \mathbf{w})$ over the set $\{x : \mathbf{s}_2(x) = \mathbf{w}\}$

Exponential family: Conditional likelihood

The conditional likelihood then arises from

$$
\begin{aligned}
p(\mathbf{v}|\mathbf{w}) &= p(\mathbf{v},\mathbf{w})/p(\mathbf{w}) \\
&= \frac{\sum_{x:\mathbf{s}_1(x)=\mathbf{v},\mathbf{s}_2(x)=\mathbf{w}} \exp\{\boldsymbol{\theta}^\top \mathbf{v} + \boldsymbol{\eta}^\top \mathbf{w} - \psi(\boldsymbol{\theta},\boldsymbol{\eta})\} f_0(x)}{\sum_{x:\mathbf{s}_2(x)=\mathbf{w}} \exp\{\boldsymbol{\theta}^\top \mathbf{s}_1(x) + \boldsymbol{\eta}^\top \mathbf{w} - \psi(\boldsymbol{\theta},\boldsymbol{\eta})\} f_0(x)} \\
&= \frac{\exp\{\boldsymbol{\theta}^\top \mathbf{v}\} \sum_{x:\mathbf{s}_1(x)=\mathbf{v},\mathbf{s}_2(x)=\mathbf{w}} f_0(x)}{\sum_{x:\mathbf{s}_2(x)=\mathbf{w}} \exp\{\boldsymbol{\theta}^\top \mathbf{s}_1(x)\} f_0(x)}
\end{aligned}
$$

- The likelihood is free of $\boldsymbol{\eta}$
- The expression is considerably simplified if $f_0(x) = 1$
- Sums would be replaced by integrals if x was continuous

Conditional logistic regression

- A common application of this idea is the logistic regression setting
- Consider the model $Y_i \sim \mathrm{Bern}(\pi_i)$ with

$$\log \frac{\pi_i}{1-\pi_i} = \alpha + \beta x_i$$

- The probability model is therefore

$$\log p(\mathbf{y}) = \alpha \sum_i y_i + \beta \sum_i x_i y_i - \sum_i \log(1 + \exp\{\alpha + \beta x_i\})$$

Conditional logistic regression (cont'd)

- Letting $v = \sum x_i y_i$ and $w = \sum y_i$, this is an exponential family, and we have the conditional likelihood

$$L(\beta) = \frac{\exp(\beta v)}{\sum_u \exp(\beta u)},$$

where the sum in the denominator is over all values of $u = \sum x_i y_i^*$ such that $\sum y_i^* = w$, where y_i^* represents potential values that the random variable Y_i could have taken

- Since the y_i^* values are all 0 or 1, this corresponds to the permutations of $\mathbf{y}$
- Similar to what we've seen before, this is particularly appealing when the data is matched or paired; this is probably the most common use of conditional logistic regression

Remarks

- The usual likelihood-based approaches to inference can now be applied, although we face a computational challenge in terms of evaluating $\sum \exp(\beta x_i y_i)$ over all possible permutations of $\mathbf{y}$
- Nevertheless, fast algorithms have been developed to tackle this problem and the method (known as *conditional logistic regression*) is widely implemented in statistical software
- We focused on the simple regression case here, but the idea can be extended to multivariate settings as well
- Futhermore, exact approaches to inference are possible using permutation tests (as in our earlier examples)