Pseudo-likelihood

Patrick Breheny

December 11, 2024

Introduction

- For our final lecture, we'll take a look at "pseudo" likelihoods
- Unlike the other variants, pseudo-likelihood is somewhat vague term with no single theoretical framework
- Rather, the term is used to describe functions of the parameters that depend on the data which are not the likelihood but nevertheless have properties similar to that of the likelihood

Why pseudo-likelihood?

- Pseudo-likelihoods arise is three main contexts:
    - Response-biased sampling
    - Two-stage (“plug-in”) likelihoods
    - Composite likelihoods
- In all of these scenarios, the true likelihood is complicated; to make analyzing the data feasible, we are going to replace it with something simpler

Response-biased sampling

- We'll start with response-biased sampling: instead of a simple random sample, observations are sampled conditional on the outcome, with the case-control study being the most common
- In such situations, the prospective likelihood (the one based on the simple random sample) is usually straightforward and easy to work with, but isn't the actual likelihood based on the study design . . . is it OK to use it anyway?

Binomial example: Setup

- Let's start with the simplest case: $Y_i \overset{\text{iid}}{\sim} \text{Bern}(\pi)$ for $i = 1, \ldots, N$
- However, we do not get to observe all N observations; instead, if $Y_i = 1$, the observation is sampled with (known) probability p_1, while if $Y_i = 0$, it is sampled with (known) probability p_0
- Introducing some extra notation, let N_1 and N_0 denote the unobserved number of events, with n_1 and n_0 the observed number of cases and controls in our sample

Binomial example (cont'd)

- As a concrete example, let's suppose $\pi = 0.2$, $p_1 = 1$, and $p_0 = 1/2$ (we get to see all the cases, but only half of the controls)
- In this scenario, if $N = 100$, we would expect to see $n_1 = 20$ cases and $n_0 = 40$ controls; the naïve estimate $n_1/(n_1 + n_0)$ would produce the biased estimate $\hat{\pi} = 0.333$
- Clearly, we must make adjustments for the sampling frequencies p_1 and p_0

Likelihood?

- Let's say we attempted to carry out a likelihood-based analysis of this problem with

$$\begin{aligned} L_i &= \mathbb{P}(Y_i \cap S_i) \\ &= \begin{cases} \pi p_1 & \text{if } Y_i = 1 \\ (1-\pi)p_0 & \text{if } Y_i = 0 \end{cases} \end{aligned}$$

  where S_i denotes the event that the observation was sampled
- Unfortunately, this produces the "MLE" of $\hat{\pi} = n_1/(n_1 + n_0)$, exactly what we said we didn't want
- What went wrong?

Correct likelihood

- This likelihood is incorrect, as we have ignored the unsampled data
- The correct likelihood is $\mathbb{P}(Y_i \cap S_i | S_i)$, the probability of Y_i *conditional* on the fact that the observation made it into the sample
- With this likelihood, the score is now

$$u(\pi) = \frac{n_1}{\pi} - \frac{n_0}{1-\pi} - \frac{(n_0+n_1)(p_1-p_0)}{\pi p_1 + (1-\pi)p_0}$$

- The good news is that this score is now "correct", in that the MLE is now sensibly adjusted for sampling fraction:

$$\hat{\pi} = \frac{n_1 p_0}{n_1 p_0 + n_0 p_1}$$

Remarks

- The bad news is that the likelihood is far more complicated and difficult to work with
- In this simplest of scenarios, it is still possible to work through the algebra, but messy enough that I chose to skip it during class time
- One can imagine that this approach is not going to scale up particularly well with more complex probability models

An "estimated" likelihood

- Perhaps there's a simpler way
- In terms of N_1 and N_0, the likelihood for π is simply that of a binomial distribution
- Unfortunately, N_1 and N_0 are unobserved; however, they can easily be *estimated*: $\hat{N}_j = n_j/p_j$
- Thus, perhaps a reasonable way to proceed is to simply plug in these estimates into the binomial likelihood

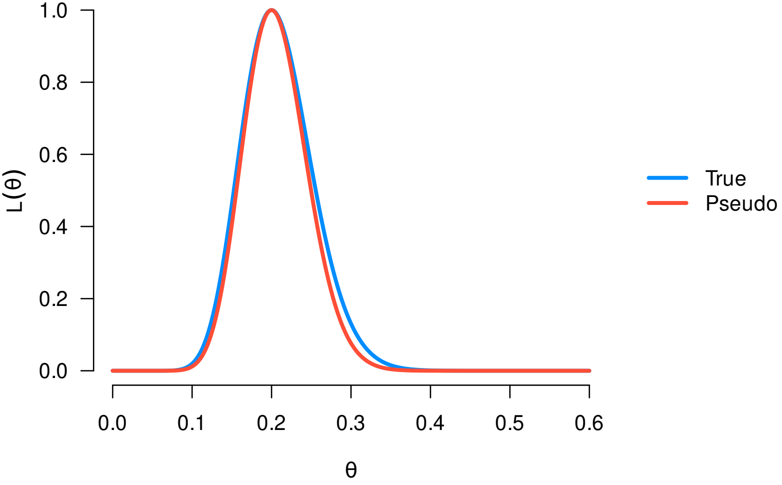
Inverse probability weighting

- Doing so, we obtain the log-likelihood

$$\ell(\pi) = \frac{n_1}{p_1} \log \pi + \frac{n_0}{p_0} \log(1 - \pi)$$

- Note that this is the original, "naïve" likelihood, but where the observations have been weighted by $1/p_1$ and $1/p_0$
- This idea, known as inverse probability weighting, comes up often in statistics, in a variety of contexts

Connection with true likelihood



Remarks

- As the figure illustrates, the pseudo-likelihood is roughly similar to the true likelihood, and the pseudo-MLE is the same as the true MLE
- However, the likelihoods are not the same – in particular, the pseudo-likelihood is narrower
- Treating the pseudo-likelihood as an ordinary likelihood, therefore, is going to produce variance estimates that are too small

Variance estimation

- This is exactly the kind of thing that one would use a sandwich estimator for:

$$\sqrt{n}(\hat{\pi} - \pi^*) \xrightarrow{\mathrm{d}} \mathrm{N}(0, A^{-1}BA^{-1}),$$

  where $A = -\mathbb{E}\nabla^2 \ell_i(\pi^*)$ is the pseudo-information and $B = \mathbb{V}u_i(\pi^*)$ is the variance of the pseudo-score
- These approaches yield the following 95% Wald CIs for π:
  - True likelihood: [0.114, 0.286]
  - Pseudo-likelihood (no adjustment): [0.122, 0.278]
  - Pseudo-likelihood (corrected): [0.114, 0.286]

Case-control studies

- The most common scenario in which response-biased sampling arises is in the application of logistic regression to case-control studies
- In this experimental design, a fixed number of cases (n_1) and controls (n_0) are sampled
- The disease status, therefore, is not random; rather it is the exposure(s) that are random
- The true likelihood, therefore, is

$$L = \prod_i p(\mathbf{x}_i | y_i)$$

A pseudo-likelihood

- This is an inconvenient likelihood for several reasons; perhaps most importantly, it requires us to specify a (multivariate) distribution on the predictors, something that is not required in regression approaches
- Suppose we instead treat the data as prospectively acquired, with the likelihood

$$L = \prod_i p(y_i|\mathbf{x}_i);$$

  this is obviously much more convenient, as this is just the usual likelihood from a logistic regression model
- However, this is a pseudo-likelihood in the sense that it does not correspond to the actual likelihood from the experiment

Inference

- In terms of estimating the intercept, the kinds of adjustments we just worked through for response-biased sampling are necessary in order to obtain consistent estimates and correct standard errors
- However, in the special case of logistic regression, it can be shown that simply treating the pseudo-likelihood as the true likelihood yields the correct MLEs and standard errors (i.e., those of the true likelihood) for all parameters except the intercept
- Since the regression coefficients and their associated odds ratios are typically the only parameters of interest, this means that regular logistic regression can be applied; no adjustments for the retrospective design are necessary

Composite likelihood

- Another type of pseudo-likelihood arises from multiplying together separate small components of the likelihood; this is known as *composite likelihood*:

$$L_{\mathsf{comp}}(\boldsymbol{\theta}|\mathbf{y}) = \prod_{k=1}^{K} L_k(\boldsymbol{\theta}|\mathbf{y})$$

- Typically, this is done when the components are simple to derive but the full likelihood is very complicated

One-dimensional lattice

- For example, suppose we have ordered observations $y_1, y_2, \ldots, y_n$ (perhaps ordered with respect to time, or along a genome)
- We might specify a model for how each observation depends on its neighbors: $p(y_k | y_{k-1}, y_{k+1})$
- Multiplying these probabilities together, however

$$p(y_2|y_1, y_3) \times p(y_3|y_2, y_4) \ldots$$

does not actually result in the correct likelihood:

$$p(y_2) \times p(y_3|y_2) \times p(y_4|y_2, y_3) \ldots$$

Ising model

- For example, suppose $y_k \in \{0, 1\}$ and let $n_k = y_{k-1} + y_{k+1}$
- One way to model the dependence of a point on its neighbors is with the *Ising model*

$$p(y_2, \ldots, y_{n-1} | y_1, y_n) = \exp\left\{\alpha \sum_{k=2}^{n-1} y_k + \beta \sum_{k=2}^{n-1} y_k n_k - h(\alpha, \beta)\right\},$$

  where positive values of β reflect positive dependence (1s and 0s tend to cluster together)
- This true likelihood is intractable, however, since the normalizing constant $h(\alpha, \beta)$ is very complicated

Ising model with composite likelihood

- The composite likelihood, however, is quite convenient:

$$p(y_k|y_{k-1}, y_{k+1}) = \frac{\exp(\alpha + \beta n_k)}{1 + \exp(\alpha + \beta n_k)},$$

  in other words, simple logistic regression
- The parameters α and β are then estimated by maximizing

$$\ell_{\textsf{comp}}(\alpha, \beta) = \sum_k \ell_k(\alpha, \beta|y_k);$$

  derivatives, Hessians, etc., are straightforward
- The same idea can be extended to higher dimensions as well as continuous outcomes

Standard errors

- In general, composite likelihoods may be seen as misspecified likelihoods, and the sandwich estimator $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$ can be used to obtain standard errors
- However, the dependence among observations can make it difficult to estimate $\mathbf{B}$, the "meat" of the sandwich estimator; the empirical estimator

$$\hat{\mathbf{B}} = \frac{1}{K}\sum_{k=2}^{k-1} u(\hat{\boldsymbol{\theta}}|y_k)u(\hat{\boldsymbol{\theta}}|y_k)^\top$$

can be biased for correlated data because $u(\hat{\boldsymbol{\theta}})$ is systematically closer to zero than $u(\boldsymbol{\theta}^*)$ (many alternative estimators have been proposed)

Remarks

- Composite likelihood methods have found many uses in analyzing longitudinal, time series, genetic, and spatio-temporal data
- They are also used in network analysis, where it is (relatively) easy to model how an individual depends on their neighbors, but hard to specify the full likelihood of an entire network
- The idea of taking a valid likelihood for an individual observation but then combining these likelihoods in a way that is *not* the full likelihood also appears in a variant called *partial likelihood*, which is used extensively in survival analysis

The plug-in likelihood

- In previous lectures, we have discussed many approaches for handling the scenario where θ is a parameter of interest and $\boldsymbol{\eta}$ are nuisance parameters
- Consider the following pseudo-likelihood, where $\hat{\boldsymbol{\eta}}$ is an estimate of $\boldsymbol{\eta}$ (not necessarily the MLE):

  $$L(\theta) = L(\theta, \hat{\boldsymbol{\eta}}),$$

  where $\hat{\boldsymbol{\eta}}$ is treated as a fixed constant
- This is sometimes referred to as the "two-stage" likelihood, the "plug-in" likelihood, or the "estimated" likelihood

Pseudo-likelihood vs profile likelihood

- Note that this is quite different from the profile likelihood
- In a profile likelihood, $\hat{\boldsymbol{\eta}}(\theta)$ is a function of θ
- In the pseudo-likelihood, we have simply plugged in $\hat{\boldsymbol{\eta}}$ for $\boldsymbol{\eta}$ and are not accounting for its potential dependence on θ in any way
- Because of this, as we saw in the earlier response-biased sampling approach, adjustments must be made to the variance in order to compensate for the failure to account for this dependence

Theoretical behavior of pseudo-likelihood

Theorem (Gong & Samaniego): Suppose assumptions (A)-(C) from the consistency of MLE lecture are met. Then

- If $\hat{\boldsymbol{\eta}}$ is consistent, there exists a sequence of consistent roots $\hat{\theta}$
- If

$$\begin{bmatrix} \frac{1}{\sqrt{n}} u_1(\theta^*, \boldsymbol{\eta}^*) \\ \sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*) \end{bmatrix} \xrightarrow{\mathrm{d}} \mathrm{N}\left(\mathbf{0}, \begin{bmatrix} \Sigma_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right),$$

  then $\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{\mathrm{d}} \mathrm{N}(0, \sigma^2)$, where

$$\sigma^2 = \mathscr{I}_{11}^{-1} + \mathscr{I}_{11}^{-2} \boldsymbol{\mathscr{I}}_{12}(\boldsymbol{\Sigma}_{22} \boldsymbol{\mathscr{I}}_{21} - 2\boldsymbol{\Sigma}_{21}),$$

  where the Fisher information matrices are for a single observation and evaluated at $(\theta^*, \boldsymbol{\eta}^*)$

Remarks

- This can be a useful framework for studying "two-stage" procedures, in which some analysis is done in stage one and results/estimates from that step are fed into a second stage
- However, the Gong & Samaniego approach is considerably more difficult to apply in practice than the sandwich estimator, as empirical estimators for $\boldsymbol{\Sigma}$ are not straightforward

Some final thoughts

- Hopefully by this point in the course you feel that you've seen the wide applicability of likelihood, along with many useful extensions, modifications, and applications
- Certainly, there are others we didn't cover, but hopefully you've gained enough experience and familiarity with the tools we have derived and used that you could read and understand how they work on your own