Pseudo-likelihood

Patrick Breheny

December 1

Introduction

- Today, we'll be discussing the idea of constructing some function of the parameters, depending on the data, that is not the likelihood but nevertheless has properties similar to that of the likelihood
- These functions are known as "pseudo" likelihoods
- The term "pseudo-likelihood" is difficult to define precisely, as it is used by various authors to mean different things, but the goal today is to see a general overview of what it means and how it works

Why pseudo-likelihood?

- Broadly speaking, pseudo-likelihood is an attractive approach in situations where the actual likelihood is either very messy and difficult to work with, or requires knowledge or assumptions about unknown factors
- In such situations, it is sometimes possible to replace the complicated likelihood with a simpler likelihood, often involving some estimate of the unknown factors
- Obviously, there is no guarantee that doing so is valid, so each case must be evaluated individually, although we will discuss some general theory later on

Response-biased sampling

- A common situation in which pseudo-likelihoods often appear is that of response-biased sampling – i.e., 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

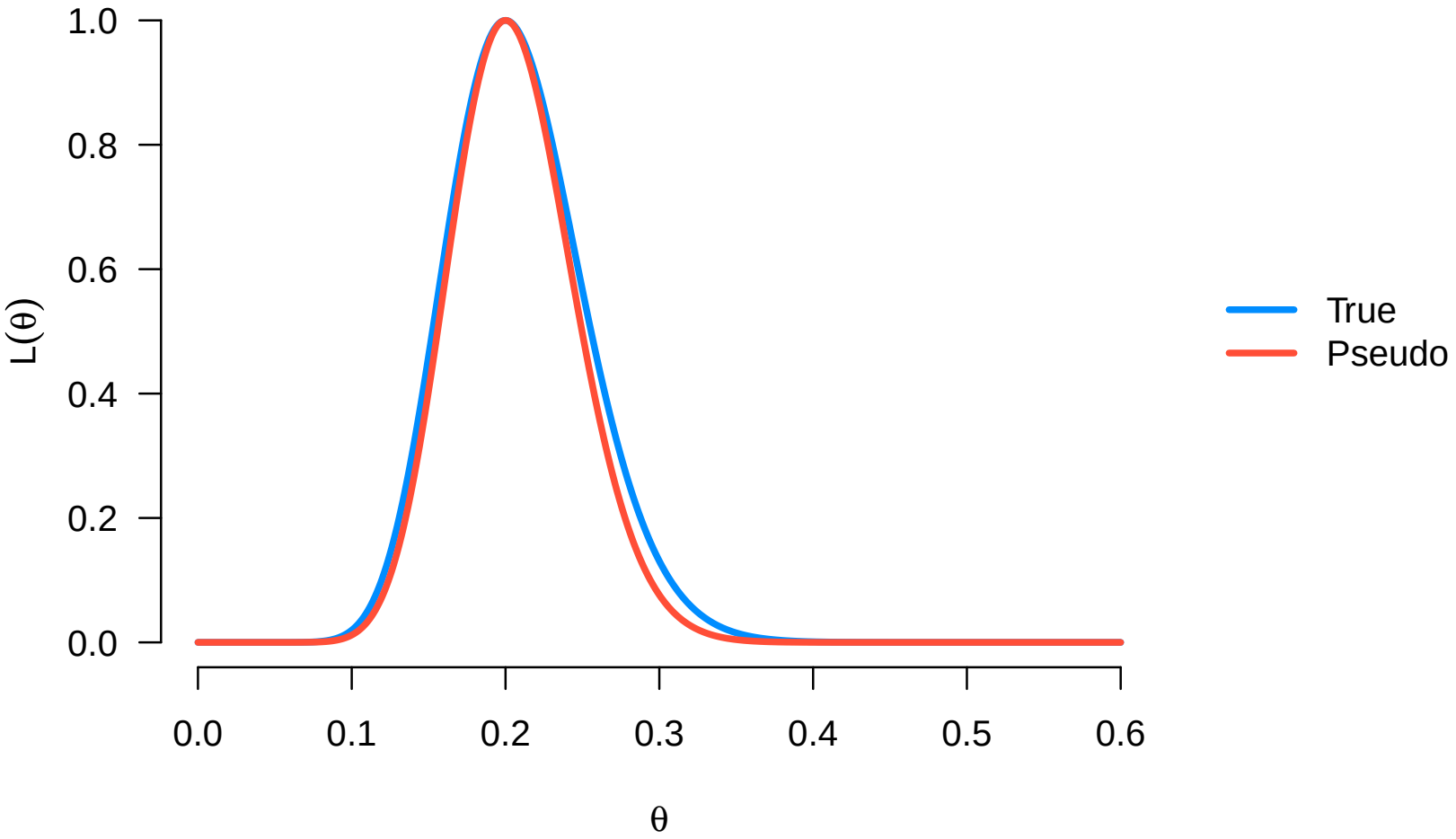
Inverse probability weighting

- 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

- Note, however, that the pseudo-score still has mean zero at π^*
- Thus, we have

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

  where $A = -\mathbb{E}\nabla^2 \ell_i(\pi^*)$ is the pseudo-information and $V = \mathbb{V} u_i(\pi^*)$ is the variance of the score statistic
- 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, it must be regarded as a pseudo-likelihood, as 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 (homework) 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

A general definition of pseudo-likelihood

- Finally, let's look at a general theory of pseudo-likelihood proposed by Gong and Samaniego (1981)
- As we have done in previous lectures, suppose that θ is the parameter of interest and $\boldsymbol{\eta}$ are nuisance parameters
- Further suppose that we have an estimate $\hat{\boldsymbol{\eta}}$ of $\boldsymbol{\eta}$ (could be the MLE, doesn't have to be)
- The *pseudo-likelihood* is then defined as

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

where $\hat{\boldsymbol{\eta}}$ is treated as a fixed constant

Pseudo-likelihood vs profile likelihood

- Note that this is 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

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

(b) 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

- Pseudo-likelihood is 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
- For example, suppose we had the regression model

$$\mathbb{E}Y = f(\mathbf{x}, \boldsymbol{\beta})$$
$$\mathbb{V}Y = g(\mathbf{x}, \boldsymbol{\beta}, \theta)$$

and our interest was in modeling the variance (g)

Remarks (cont'd)

- Although perhaps possible to consider the full likelihood, it would be simpler and more convenient to first estimate β (stage 1) and then use the residuals from that fit to model the variance (stage 2)
- However, as we have seen, drawing inferences about g will not have proper coverage (overconfident) unless we make adjustments
- The preceding theorem is useful for working out how the variance needs to be adjusted such settings, and has been applied to the variance modeling problem, among others