Notes accompanying BDA chapter 8

Modeling accounting for data collection

Jeff Miller

March 4, 2016

Contents

Many times, the construction of a data set involves some selection or collection process governing which samples we see. In many cases, we don't need to model this data collection process (it is "ignorable"), but sometimes there are biases in the data collection process that are important to account for. Recognizing when such issues arise, and properly accounting for them, is probably one of the more subtle aspects of using statistics in practice.

1 Introduction

1.1 A few motivating examples

- The local city government installed traffic cameras at the intersections that had the highest number of accidents in the previous year. This year, they noticed that the number of accidents at those intersections was lower, on average, compared to the previous year. Is this evidence that installing traffic cameras helps decrease the number of accidents? Not necessarily! The issue is that the number of accidents in any given year will exhibit some randomness, and by choosing intersections with a high number of accidents, we will tend to choose intersections that had higher numbers of accidents than their individual means. So, in fact, we would expect the number of accidents at those intersections to be less this year, even if the cameras were not there. Basically, there is a selection bias in how the intersections were chosen.

- For several years, observational studies suggested that women receiving hormone therapy have a lower risk of cardiovascular disease. However, this was more recently contradicted by randomized controlled trials, which suggested that in fact, the risk of cardiovascular disease was significantly increased. How could this happen? (By the way, in lecture I mistakenly said "human growth hormone" instead of "hormone therapy", which is different.) One possibility is that the observational studies did not properly account for all of the relevant "confounders" — for example, wealthier women are more likely to have hormone therapy, but they are also more likely to have a healthier lifestyle. Another, more subtle, possible explanation (Hernan et al., Epidemiology, 2008) is that it appears that the observational studies compared women who were currently receiving hormone therapy to women who had never received hormone therapy—however, this is missing is the subset of women who started taking hormone therapy and then stopped taking it before the observations were made (possibly even due to death). The underlying reason for both of these possible explanations is failure to correctly model the data collection process.

- During World War II, the US military found that the bomber planes returning from missions were being struck by bullets in certain parts of the plane more than other parts. They were planning to provide armor to protect these frequently-struck parts of the planes, however, statistician Abraham Wald was consulting for the military, and he realized that this was precisely the opposite of what they should do. Why? The key insight was that their observations were based on the planes that *returned* from their missions—they were not considering all of the planes that were shot down! He performed a careful statistical analysis and recommended that they reinforce the planes in the parts where bullet holes had not been observed. Basically, the idea is that these were critical regions, and the reason the observed planes returned was that they were not struck there.

1.2 Discussion

The basic point of this chapter is that sometimes, we need to model the biases in how the data were collected. A canonical case of this is handling missing data—in other words, handling the possibility that some potential samples were not included in the data set. It is important to realize that missing data can contain information about the parameters you are interested in, and ignoring it can bias your results. In some lucky cases we can ignore the data collection process—we will study this property, called "ignorability", and establish some sufficient conditions under which it holds. When ignorability does not hold, the good news is that if we model the data collection process correctly, we can just use standard Bayesian methods to perform inference.

2 Of airplanes and bullet holes

- To introduce some of the basic concepts and notation, let's consider a simplification of the Wald story.
- Suppose we want to know the mean number of bullet holes occurring in planes during a mission.
- Denote

$$\begin{aligned} y_j &= \#\text{ of bullet holes in plane } j \\ I_j &= \mathbb{1}(\text{plane } j \text{ returned from its mission}) \\ \text{obs} &= \{j : I_j = 1\} = \text{the set of indices of the observed values} \\ \text{mis} &= \{j : I_j = 0\} = \text{the set of indices of the missing values.} \end{aligned}$$

- Let's assume the following model:

$$\begin{aligned} Y_1, \ldots, Y_n|\theta \text{ i.i.d.} &\sim \text{Poisson}(\theta) \\ I_j|y, \theta &\sim \text{Bernoulli}(\phi_{y_j}), \end{aligned}$$

  and for simplicity, let's assume the parameters $\phi_0, \phi_1, \ldots$ are known, and are strictly decreasing in size.

- Then

$$p(y, I|\theta) = \prod_{j=1}^{n} p(y_j|\theta)p(I_j|y_j) = p(y_{\text{obs}}|\theta)\Big(\prod_{j\in\text{obs}} \phi_{y_j}\Big)\Big(\prod_{j\in\text{mis}} p(y_j|\theta)(1-\phi_{y_j})\Big).$$

- It would be incorrect to use $p(\theta|y_{\text{obs}})$ for posterior inferences about θ, since this would not take into account the bias in the data collection process due to the fact that planes with more bullet holes are less likely to return.

- Instead, the correct distribution to use is

$$\begin{aligned} p(\theta|y_{\text{obs}}, I) &\propto p(y_{\text{obs}}, I|\theta)p(\theta) = p(\theta)\sum_{y_{\text{mis}}} p(y, I|\theta) \\ &= p(\theta)p(y_{\text{obs}}|\theta)\Big(\prod_{j\in\text{obs}} \phi_{y_j}\Big)\sum_{y_{\text{mis}}}\Big(\prod_{j\in\text{mis}} p(y_j|\theta)(1-\phi_{y_j})\Big) \\ &\propto p(\theta)p(y_{\text{obs}}|\theta)\prod_{j\in\text{mis}}\sum_{y_j} p(y_j|\theta)(1-\phi_{y_j}) \\ &= p(\theta)p(y_{\text{obs}}|\theta)\mathbb{P}(I_1 = 0 \mid \theta)^{|\text{mis}|}. \end{aligned}$$

  where the sum $\sum_{y_{\text{mis}}}$ is overall possible values of the vector $y_{\text{mis}} = (y_j : j \in \text{mis})$.

- Note that (in this example) the correct posterior is proportional to the naive incorrect posterior $p(\theta|y_{\text{obs}})$ times the factor $\mathbb{P}(I_1 = 0 \mid \theta)^{|\text{mis}|}$ accounting for the missing data.

2.1 Ignorability

- This example (airplanes and bullet holes) is one in which the data collection process is not "ignorable". On the other hand, if it were the case that $\phi_0 = \phi_1 = \cdots = c$ for some constant $c \in (0, 1)$, then $I \perp (Y, \theta)$, and it turns out that when this is so, the data collection process is "ignorable" in the sense that $p(\theta|y_{\text{obs}}, I) = p(\theta|y_{\text{obs}})$. Actually, this is easy to see in this example since if $\phi_0 = \phi_1 = \cdots = c$ then $\mathbb{P}(I_1 = 0 \mid \theta)^{|\text{mis}|} = (1-c)^{|\text{mis}|} \propto_\theta 1$.

- More generally, the parameters $\phi_0, \phi_1, \ldots$ might be unknown, in which case we would put a prior on them. Also, we might have covariates x_j associated with plane j. In this more general setting, the appropriate distribution to use for inferences about θ is

$$p(\theta \mid x, y_{\text{obs}}, I) = \int p(\theta, \phi \mid x, y_{\text{obs}}, I)d\phi.$$

- In general, the data collection process is said to be *ignorable* if

$$p(\theta \mid x, y_{\text{obs}}, I) = p(\theta \mid x, y_{\text{obs}}).$$

- Some care is needed to properly interpret the quantity on the right-hand side of this equation, since the convention of using the same letter for both a random variable and its value makes this expression ambiguous. The short explanation is that the right-hand side should be interpreted formally as

$$p(\theta|x, y_{\text{obs}}) \propto p(y_{\text{obs}}|x, \theta)p(\theta|x).$$

This might seem obvious, but consider the following thought process: "if I know the observed values y_{obs}, then I must know which subset of variables I observed, so conditioning on y_{obs} should be the same as conditioning on both y_{obs} and I." To clear up the confusion, we need to make the notation a little more precise. Let $\mathbb{I}$ denote the random variable taking values I, and write y_I instead of y_{obs} to denote the values of the observed variables. The reason why the thought process above is invalid is that the expression $p(\theta \mid x, y_{\text{obs}})$ should be interpreted as

$$p(\theta \mid x, y_{\text{obs}}) = p(\theta \mid x,\, Y_I = y_I),$$

and *not* $p(\theta \mid x,\, Y_{\mathbb{I}} = y_I)$, which *would* be equal to $p(\theta \mid x,\, Y_{\mathbb{I}} = y_I, \mathbb{I} = I) = p(\theta \mid x, y_{\text{obs}}, I)$.

- To further understand the distinction, here's a simple example. Suppose your model is $Y_1, Y_2|\theta$ i.i.d. $\sim N(\theta, 1)$ and you have a prior on θ. Consider the following two scenarios:

  (a) I tell you that $y_1 = 2.4$. What is your posterior on θ?

  (b) I tell you that $y_1 = 2.4$ and $y_1 < y_2$. What is your posterior on θ?

  It should be clear that these two scenarios lead to different posteriors for θ. Now, to make the connection with y_{obs} and I, suppose $p(I|y, \theta)$ is such that $I_1 = 1, I_2 = 0$ whenever $y_1 < y_2$, and otherwise, $I_1 = 0, I_2 = 1$. When $y_1 < y_2$, scenario (a) above corresponds to using $p(\theta|y_{\text{obs}})$, and scenario (b) corresponds to using $p(\theta|y_{\text{obs}}, I)$. The difference is that in scenario (a), I didn't tell you why you were only seeing y_1.

3 Medical treatment example

- Before considering the general setup and establishing some general conditions under which ignorability is guaranteed, let's look at another example.

- Suppose a doctor has n patients with some disease, and each patient is given one of two treatments, A or B.

- Denote by y the matrix of "potential outcomes",

$$y = \begin{bmatrix} y_1^{\text{A}} & y_1^{\text{B}} \\ y_2^{\text{A}} & y_2^{\text{B}} \\ \vdots & \vdots \\ y_n^{\text{A}} & y_n^{\text{B}} \end{bmatrix}$$

  where y_j^T denotes the outcome patient j would exhibit if given treatment T.

- Likewise, denote by I the matrix of observation indicators,

$$I = \begin{bmatrix} I_1^{\mathrm{A}} & I_1^{\mathrm{B}} \\ \vdots & \vdots \\ I_n^{\mathrm{A}} & I_n^{\mathrm{B}} \end{bmatrix}$$

  where $I_j^T = \mathbb{1}(\text{patient } j \text{ is given treatment } T)$. It is assumed that each patient is given exactly one treatment, A or B; in other words, $I_j^{\mathrm{B}} = 1 - I_j^{\mathrm{A}}$.

- Suppose we have a vector of covariates $x_j = (x_{j1}, \ldots, x_{jd})^{\intercal} \in \mathbb{R}^d$ for patient j.

- Consider the following simple model for y and I:

$$\begin{aligned} Y_j^{\mathrm{A}} | x, \theta &\sim \mathcal{N}(\theta_{\mathrm{A}} + \beta^{\intercal} x_j,\, \sigma^2) \\ Y_j^{\mathrm{B}} | x, \theta &\sim \mathcal{N}(\theta_{\mathrm{B}} + \beta^{\intercal} x_j,\, \sigma^2) \\ I_j^{\mathrm{A}} | x, y, \phi &\sim \text{Bernoulli}(\text{logit}^{-1}(\phi^{\intercal} x_j)) \end{aligned}$$

  where $\theta = (\theta_{\mathrm{A}}, \theta_{\mathrm{B}})^{\intercal} \in \mathbb{R}^2$ and $\phi = (\phi_1, \ldots, \phi_d)^{\intercal} \in \mathbb{R}^d$. Assume a prior $p(\theta, \phi | x)$ on these parameters. For simplicity, let's assume that the vector of regression coefficients β and the variance σ^2 are known, but of course more generally we could put priors on them.

- We will revisit this example to illustrate various concepts.

4 General framework

4.1 Setup

- The framework described here is used in a large number of situations, including models for missing data, censored data, surveys/polls, randomized experiments, and causal inference.

- Let y denote a matrix of "potential outcomes" y_{ij}, some of which will be observed, and some of which will not be observed.

- Let $I_{ij} = \mathbb{1}(\text{entry } i, j \text{ of } y \text{ is observed})$.

- Let $\text{obs} = \{(i, j) : I_{ij} = 1\}$ and $\text{mis} = \{(i, j) : I_{ij} = 0\}$.

- Let x denote some accompanying collection of covariates.

- Let θ denote parameters governing the distribution of the potential outcomes y, and let ϕ denote parameters governing the distribution of the data collection process I.

- Typically, we get to see x, y_{obs}, and I (but not y_{mis}).
- Assume the following factorization holds:

$$p(y, I|x, \theta, \phi) = p(y|x, \theta)p(I|x, y, \phi).$$

This is referred to as the *complete-data likelihood*.

4.2 Basic properties

- The joint posterior on (θ, ϕ) is then

$$\begin{aligned} p(\theta, \phi|x, y_{\text{obs}}, I) &\propto p(\theta, \phi|x)p(y_{\text{obs}}, I|x, \theta, \phi) \\ &= p(\theta, \phi|x) \int p(y, I|x, \theta, \phi)dy_{\text{mis}} \\ &= p(\theta|x)p(\phi|x, \theta) \int p(y|x, \theta)p(I|x, y, \phi)dy_{\text{mis}}. \end{aligned}$$

Note that here we are assuming the y's are continuous, but of course if they were discrete, the integral above would be replaced by a sum. Also, the proportionality here is with respect to both θ and ϕ.

- The posterior on θ is then obtained by just integrating this over ϕ, i.e.,

$$\begin{aligned} p(\theta|x, y_{\text{obs}}, I) &= \int p(\theta, \phi|x, y_{\text{obs}}, I)d\phi \\ &\propto_\theta p(\theta|x) \int p(\phi|x, \theta) \left(\int p(y|x, \theta)p(I|x, y, \phi)dy_{\text{mis}} \right) d\phi. \end{aligned} \tag{4.1}$$

4.3 Ignorability and related concepts

- We say that *ignorability* holds if

$$p(\theta|x, y_{\text{obs}}, I) = p(\theta|x, y_{\text{obs}}).$$

- When ignorability holds, we don't have to worry about modeling the data collection process. This makes life easier, and also makes our inferences more robust (since there are fewer modeling assumptions to possibly get wrong).
- Note that ignorability, as defined here, is a property of the assumed model—but of course your model might not actually be a good representation of the true distribution. If you assume a model in which ignorability holds, but the assumptions underlying your model are invalid, then obviously your resulting inferences will be compromised. In some cases, fortunately, the true data collection process is directly under our control (for example, in randomized controlled trials), so we can guarantee that it has a particular distribution—and thus, in such cases we can be confident that our model for the data collection process is correct.

- We say that data is *missing at random* (MAR) if

$$p(I|x, y, \phi) = p(I|x, y_{\text{obs}}, \phi).$$

- We say that data is *missing completely at random* (MCAR) if

$$p(I|x, y, \phi) = p(I|x, \phi).$$

- We say that *strong ignorability* holds if

$$p(I|x, y, \phi) = p(I|x).$$

- We say that the condition of *distinct parameters* holds if

$$p(\theta, \phi|x) = p(\theta|x)p(\phi|x).$$

  (This terminology is really bad, but unfortunately it seems to be standard.)

- Based on these definitions, we can derive the following results:

$$\begin{aligned}
&\text{strong ignorability} \Longrightarrow \text{MCAR} \Longrightarrow \text{MAR} \\
&\text{strong ignorability} \Longrightarrow \text{ignorability} \\
&\text{MAR} + \text{distinct parameters} \Longrightarrow \text{ignorability}.
\end{aligned}$$

  The first line can be seen directly from the definitions (if this is not obvious to you, review conditional independence). The second and third lines can be derived by plugging the definitions into equation 4.1.

5 Medical treatment example, revisited

- Here we illustrate the concepts above in the context of this example.

- In the original setup for this example, the data collection process was modeled as

$$I_j^{\text{A}}|x, y, \phi \sim \text{Bernoulli}(\text{logit}^{-1}(\phi^{\intercal} x_j))$$

  independently for $j = 1, \ldots, n$. This satisfies MCAR.

- If ϕ were known exactly, a priori, then we would also have strong ignorability. But if ϕ is unknown and we need to put a prior on it, then strong ignorability does not hold and we only have MCAR.

- To illustrate a situation in which MAR holds, but MCAR does not hold, suppose the treatment of patient j is adaptively chosen based on the observed outcomes of patients $1, \ldots, j-1$, in addition to x and ϕ. This would be the case if the doctor adjusts her treatment decisions based on what she has learned from previous patients. In this case, MAR holds, but not MCAR.

- If θ and ϕ are independent in the prior (given x), then we have "distinct parameters". However, assuming such independence might not be reasonable in this example, since if $\theta_{\mathrm{B}} - \theta_{\mathrm{A}}$ is large then it is plausible that the doctor would have domain knowledge indicating this (perhaps from her background knowledge or from the medical literature), and so the doctor would be more likely to assign one treatment over the other (and thus ϕ would depend on θ).