Bayesian Data Analysis

Part I: Fundamentals of Bayesian Inference

정지원

2020.09.15

Content of the Book

Part I: Fundamentals of Bayesian Inference

1.6 Example of probability assignment: football point spreads

- Assigning Probabilities for football game result:

  1. **Subjective assessments**
  2. **Empirical probabilities using observed data**
  3. **Parametric probability model**

1.6 Example of probability assignment: football point spreads

- **Point Spread**
  - Measure of the difference in ability between the two teams
  - E.g. team A is 3.5 point favorite to defeat team B
    - Pr(A wins by more than 3.5 points) is $\frac{1}{2}$
    - Pr(B wins outright or lose less than the point) is $\frac{1}{2}$

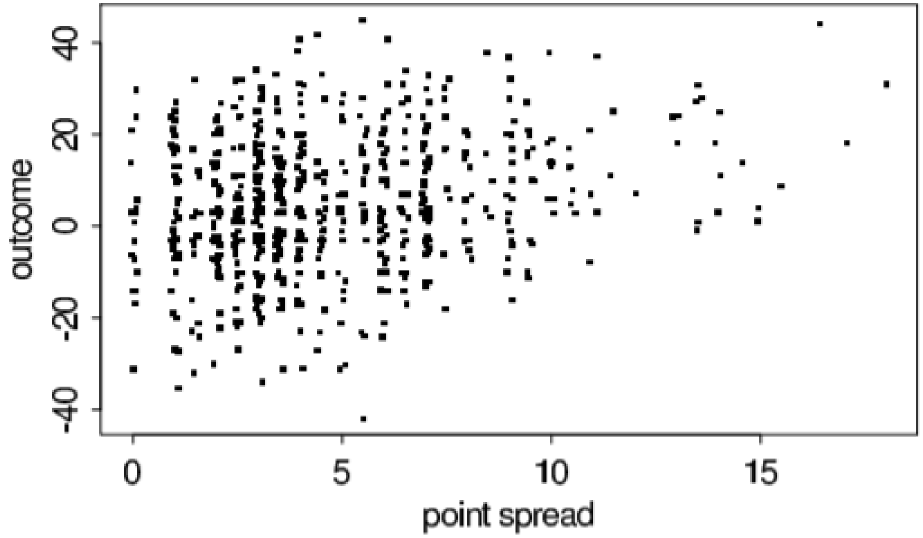
1.6 Example of probability assignment: football point spreads

1. Subjective assessments

- Reading the newspaper and watching football games
- E.g. Pr(favorite wins)
  - Between 0.6 and 0.75 ?
- **Problem:**
  - Requires more intuition or knowledge for more complex events

2. Empirical probabilities using observed data

- Pr(favorite wins) = $\frac{410.5}{655}$= 0.63
- Pr(favorite wins | x = 3.5) = $\frac{36}{59}$ = 0.61
- Pr(favorite wins by more than the point spread) = $\frac{308}{655}$ = 0.47
- Pr(favorite wins by more than the point spread | x = 3.5)= $\frac{32}{59}$ = 0.54

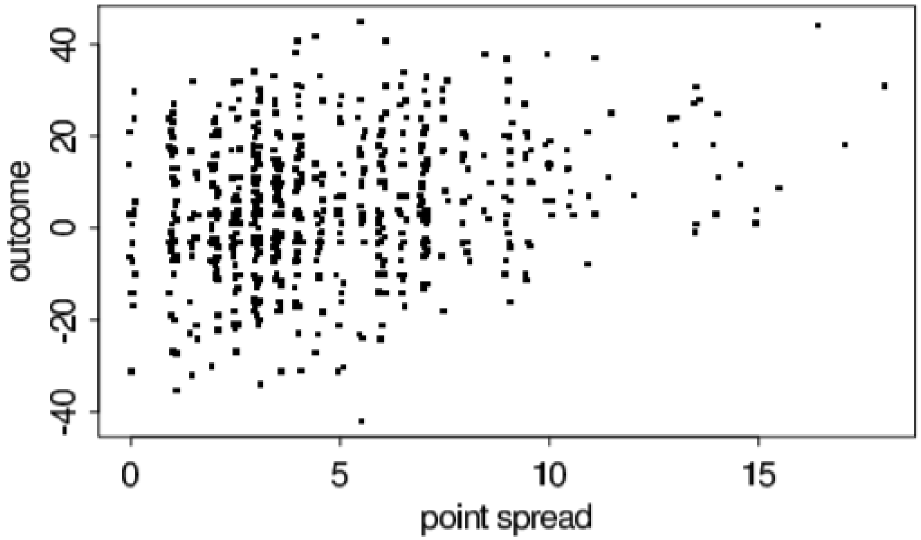


Problem: when there are events with few directly relevant data points

1.6 Example of probability assignment: football point spreads

2. Empirical probabilities using observed data

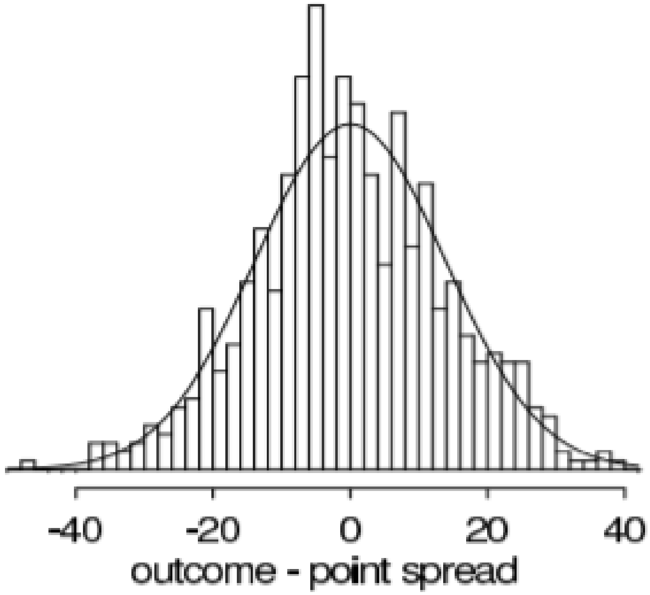
- Problematic example:
  - Favorite wins 5/5 (spread point = 8.5)
  - Favorite wins 13/20 (spread point =9)



- **The small sample size leads to imprecise probability assignment**

3. Parametric probability model

- Model: distribution of d = (y − x) as independent of x



$d|x \sim N(\mu, \sigma^2)$ where $\mu \approx 0$ and $\sigma \approx 14$

1.6 Example of probability assignment: football point spreads

3. Parametric probability model

- Probability that favorite wins for x spread point

$$Pr_{norm}(y > 0 \mid x) = Pr_{norm}(d > -x \mid x) = 1 - \Phi(-\frac{x}{14})$$

- E.g.
  - $Pr_{norm}(favorite\ wins \mid x = 3.5) = 0.60$
  - $Pr_{norm}(favorite\ wins \mid x = 8.5) = 0.73$
  - $Pr_{norm}(favorite\ wins \mid x = 9.0) = 0.74$

More intuitive sense than the empirical values based on small samples.

1.7 Example: estimating the accuracy of record linkage

- Purpose: to emphasize empirical nature of probabilities estimated from data

- Methods to estimate the accuracy of record linkage

  1. **Existing methods for assigning scores to potential matches**
  2. **Estimating match probabilities empirically**
  3. **External validation of the probabilities using test data**

1.7 Example: estimating the accuracy of record linkage

- **Record Linkage**
  - algorithmic technique to identify records from different DB that correspond to the same individual

- **Importance of accuracy of record linkage:**
  - To declare as many records as possible ‘matched’ without an excessive rate of error
  - To avoid the cost of manual processing

1. Existing methods for assigning scores to potential matches

- y - score that measures closeness b/w two records
- Cutoff score - whether the pair is 'matched' or 'falsely matched'
- False-matched rate - $\frac{\#\, of\ falsely\ matched\ pairs}{\#\, of\ declared\ matched\ pairs}$
- **Accurate method for assessing the probability that a candidate matched pair is correct match?**
  - ➢ Convert scores into probabilities (x)
  - ➢ Empirically estimate the probability of a match as a function of y using Bayesian Method

2. Estimating match probabilities empirically

- To obtain accurate match probabilities:
  - ➢ Use mixture modeling with parameters estimated from data

$$\mathbf{p(y) = Pr(match)\ p(y|match) + Pr(non\text{-}match)\ p(y|non\text{-}match)}$$

- Curve giving the false-match rate as a function of the decision threshold
  - ➢ Decision maker can determine the threshold from the curve

3. External validation of the probabilities using test data

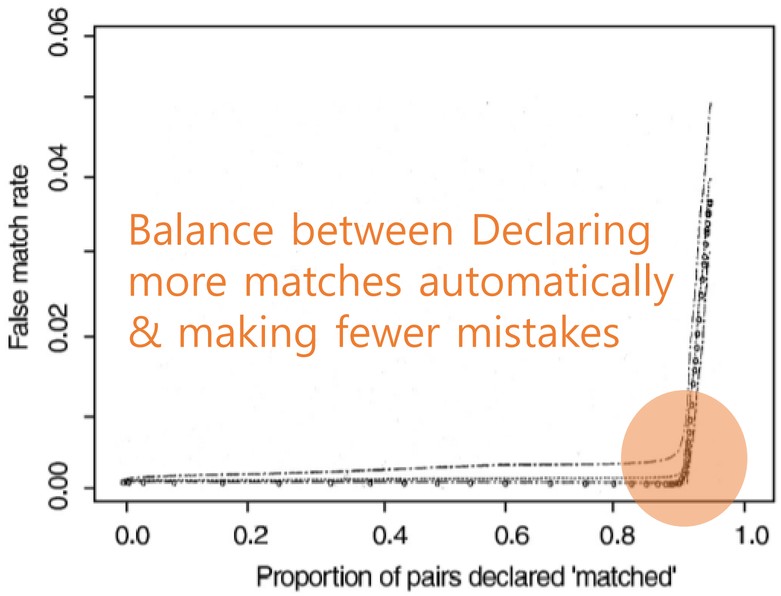
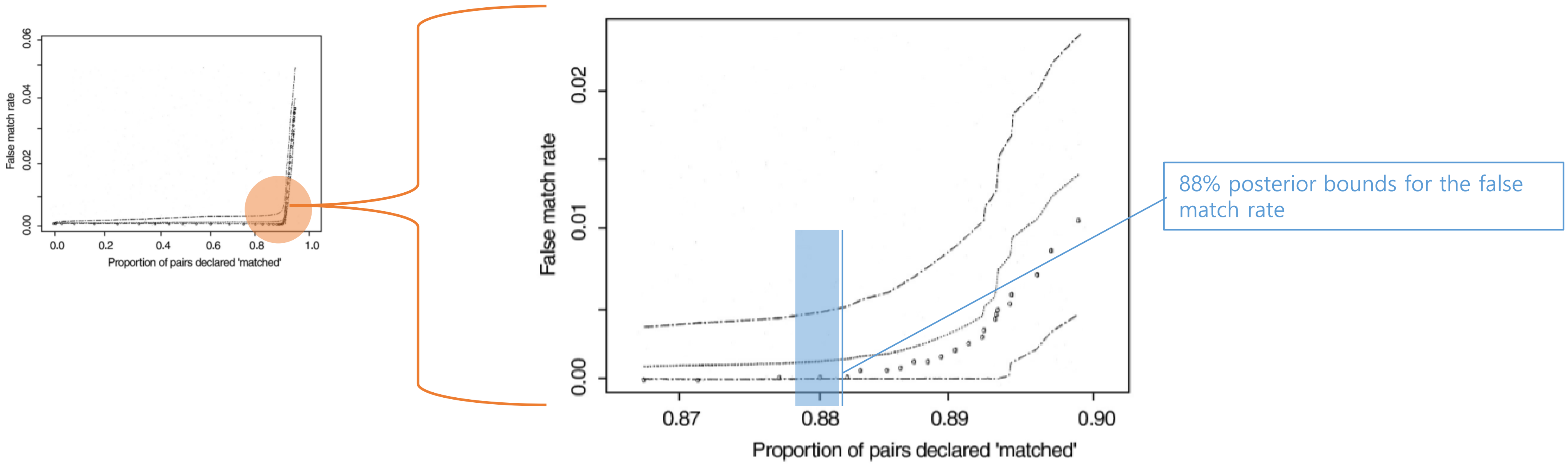


Figure shows the expected proportion of false matches and 95% posterior bounds for the false-match rate

3. External validation of the probabilities using test data



1.8 Some useful results from probability theory

Conditional distribution

- Conditional Distribution > Complicated unconditional distribution

- Conditional Mean and Variance

  - $E(u) = E(E(u|v)) = \int\int u\, p(u,v)dudv = \int\int u\, p(u|v)du\, p(v)dv = \int E(u|v)p(v)dv$

  - $Var(u) = E\big(Var(u|v)\big) + Var(E(u|v))$

1.8 Some useful results from probability theory

Transformation of variables

- **Parameter transformation** (v = f(u))
  - If p_u is a **discrete dist**.
    - $p_v(v) = p_u\left(f^{-1}(v)\right)$ (f is one to one function)
    - $p_v(v) = \sum p_u\left(f^{-1}(v)\right)$ (f is many to one function)
  - If p_u is a **continuous dist**.
    - $p_v(v) = |J|p_u\left(f^{-1}(v)\right)$ (f is one to one function)
    - $p_v(v) = \int |J|p_u\left(f^{-1}(v)\right)dv$ (f is many to one function)

- **Logarithm transformation**
  - $logit(u) = \log\left(\frac{u}{1-u}\right)$
  - $logit^{-1}(v) = \frac{e^v}{1+e^v}$

1.9 Computation and software

- Computational tasks arise in Bayesian data analysis
  - Vector and matrix manipulations
  - Computing probability density functions
  - Drawing simulations from probability distributions
  - Structured programming
  - Calculating the linear regression estimate and variance matrix
  - Graphics, including scatterplots with overlain lines and multiple graphs per page

1.9 Computation and software

- **Simulation**
  - Central part of much applied Bayesian analysis,
    - Relative ease that samples can often be generated from a probability distribution
    - Extremely large or small simulated values often flag a problem with model specification or parameterization
  - Sampling using inverse CDF
    - CDF: $$F(v_*) = \Pr(v \le v_*) = \begin{cases} \sum_{v \le v_*} p(v) \ \ if\ p\ is\ discrete \\ \int_{-\infty}^{v_*} p(v)dv\ if\ p\ is\ continuous \end{cases}$$

1.9 Computation and software

- Simulation of posterior and posterior quantities

Simulation draw	Parameters			Predictive quantities		
	θ_1	$\ldots$	θ_k	$\tilde{y}_1$	$\ldots$	$\tilde{y}_n$
1	θ_1^1	$\ldots$	θ_k^1	$\tilde{y}_1^1$	$\ldots$	$\tilde{y}_n^1$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
S	θ_1^S	$\ldots$	θ_k^S	$\tilde{y}_1^S$	$\ldots$	$\tilde{y}_n^S$

Result of a set of S simulation

Estimate the posterior distribution of any quantity of interest

1.10 Bayesian inference in applied statistics

- **Pragmatic reasons for the use of Bayesian method**
  - Inherent flexibility
  - Ability to incorporate all reasonable sources of uncertainty in inferential summaries
  - Psychological reason
  - Conditional on probability models containing approximations in attempt to represent complicated real-world relationships
- **Strength of the Bayesian approach**
  - ability to combine information from multiple sources
  - more encompassing of uncertainty about the unknowns in a statistical problem

1.10 Bayesian inference in applied statistics

- **Important themes that are common to modern applied statistical practice**
  - **willingness to use many parameters**
  - **hierarchical structuring of models**, which is the essential tool for achieving partial pooling of estimates and compromising in a scientific way between alternative sources of information
  - **model checking**—not only by examining the internal goodness of fit of models to observed and possible future data, but also by comparing inferences about estimands and predictions of interest to substantive knowledge
  - **an emphasis on inference in the form of distributions** or at least interval estimates rather than simple point estimates
  - **the use of simulation as the primary method of computation**; the modern computational counterpart to a 'joint probability distribution' is a set of randomly drawn values, and a key tool for dealing with missing data is the method of multiple imputation (computation and multiple imputation are discussed in more detail in later chapters)
  - **the use of probability models as tools for understanding and possibly improving dataanalytic techniques** that may not explicitly invoke a Bayesian model
  - **the importance of including in the analysis as much background information as possible**, so as to approximate the goal that data can be viewed as a random sample, conditional on all the variables in the model
  - **the importance of designing studies to have the property that inferences for estimands of interest will be robust to model assumptions**.