Universiteit Utrecht

Methodology & Statistics for the Behavioural, Biomedical and Social Sciences

Bayesian Statistics

Bayesian Data Integration using Metropolis Sampler with a Normal Approximation of the Posterior Distribution

Author:

Florian Metwaly

Student-Nr.: 0778265

Lecturers:

Noémi Schuurman

Herbert Hoijtink

July 19, 2024

Contents

1 Introduction

In all fields of statistics, probability survey data is the gold standard of estimation and causal inference. Probability survey data describes a dataset, that is acquired by a sampling procedure in which the probabilities of drawing a person are known. Non probability samples (NPS) describe samples in which this probability is not known for every member of the population. While probability samples (PS) count as the gold standard, there are many good reasons not to use probability survey data. Non probability surveys are often cheaper, and it is possible to gain larger samples with less expense. In this report, I will explore a Bayesian data integration approach, which uses NPS data as prior information to enrich PS data to gain a better estimate of the population estimates. The approach is shown with the example of a logistic regression with two predictors, and the samples are drawn from a simulated population. The procedure follows the process as described in Salvatore, Biffignandi, Sakshaug, Wiśniowski, and Struminskaya (2024).

The following report consists of four parts. First, I'll provide arguments why a Bayesian approach to data integration could be superior over a frequentist approach, next I will explain the simulation process, then I will explain how the model was estimated and lastly I will do a hypothesis test to test the estimated parameters against the population parameters.

2 Bayesian Data Integration

Data integration is a big and recent topic in statistics. Frequentist procedures are most often based on weighting, which requires a range of covariates to be able to accurately describe the population and estimate their propensity scores. These covariates need to be present in both datasets that are supposed to be combined. This Bayesian approach does not need these covariates, as the data integration process is solely based on variables that are part of the analysis. This makes this approach more flexible and less prone to errors that occur when the covariates to create weights are selected, by not selecting the right variables to accurately describe the population.

3 Simulation

To be able to evaluate the estimation, a population is first simulated, then a PS sample is drawn with simple random sampling (n = 500) without replacement (srswr) and finally a NPS sample is drawn (n = 2000), with first giving a higher weight (1.3) to population members with the variable X1 = 1, to create a bias. This is a realistic scenario, for example, if a NPS is distributed through the internet. Having internet then is a prerequisite of being able to be drawn into the NPS, therefore not everyone has the same probability of being in the sample. The simulated population has a size of 1 million and the population parameters $\beta = (\beta_0, \beta_1, \beta_2)$ are defined as follows:

$$\beta \in (0.5, 0.8, 0.3) \tag{1}$$

All parameters and the population size were chosen somewhat arbitrarily. The sample sizes for the PS and NPS were picked to represent a realistic scenario, with a smaller PS and a larger but biased NPS.

4 Estimation

For Bayesian estimation, normally a sampler is needed, because of the complexity of the posterior distribution. The high complexity prevents an analytical solution.

Typical samplers are Gibbs, MCMC or Metropolis Hastings. Here, a Metropolis Sampler was used. A metropolis sampler is a special case of a Metropolis Hastings sampler, I will explain that later. Metropolis Hastings (MH) sampler first use a Markov-Chain-Monte-Carlo (MCMC) procedure to sample values from a proposal distribution, then apply a decision making rule if that proposed value is accepted or rejected. All accepted values form the posterior distribution Chib and Greenberg (1995).

Following, an MH is introduced which forms a posterior distribution out of the likelihood of the binary outcome variable and the so-called power prior. The power prior is a informative prior, using historical information to form a posterior distribution (Ibrahim, Chen, Gwon, & Chen, 2015). As proposed by Salvatore et al. (2024), here the NPS data is used as historical information in the power prior to integrate the data of the NPS into the PS.

The likelihood given the binary outcome variable is defined as

$$L(Y|\beta_p) = p_i^{yi}(1 - p_i^{yi}) \tag{2}$$

As the outcome is binary, a link function used to fit a generalized linear model (McCullagh, 2019). The link function defines

$$p_i = \frac{1}{1 + e^{\beta_0 + x1\beta_1 + x2\beta_2}} \tag{3}$$

Following the example of Salvatore et al. (2024) the power prior

$$\pi(\beta, \alpha | Y_{nps}) \propto (L(Y_{nps}|\beta_p))^{\alpha} \pi_0(\beta_p | Y, \nu_0, \mu_0, \sigma_0) \tag{4}$$

with the baseline prior π_0 being a t-prior distribution which is simplified to

$$\pi_0(\beta_p | Y, \nu_0, \mu_0, \sigma_0) = (1 + \frac{1}{\nu_0}(\frac{\beta_p - \mu_0}{\sigma_0})^2)^{-\frac{\nu_0+1}{2}} \tag{5}$$

is used.

The α parameter describes the weight that is given to the historical information. Here, a Hotellings T^2 is computed to compare the maximum likelihood estimates of the PS and the NPS. The p-value from that test describes the α parameter. A p-value closer to 1 describes strong similarity of the estimates, therefore borrowing more data from the NPS, a p-value closer to 0 describes a large difference between the estimates and therefore leads to borrowing less data from the NPS (Salvatore et al., 2024).

Using Bayes theorem, the posterior distribution therefore is proportional to

$$\pi(\beta_p, \nu_0, \mu_0, \sigma_0 | Y) \propto L(Y_{ps}|\beta_p)(L(Y_{nps}|\beta_p))^{\alpha} \pi_0(\beta_p | Y, \nu_0, \mu_0, \sigma_0) \tag{6}$$

.

For the proposal distribution, a normal approximation of the posterior distribution was chosen. The normal approximation is defined as follows:

$$q(\beta^*) = \mathcal{N}(\hat{\beta}_p, \hat{\sigma}^2_{\beta}) \tag{7}$$

$\hat{\beta}_p$ is the estimated mode of the conditional posterior distribution. This mode was approximated numerically by using the optimise function in R. The variance $\hat{\sigma}^2_{\beta}$ is minus the inverse of the fisher information $\mathcal{I}(\hat{\beta}_p)$, estimated by first taking the logarithm of the conditional posterior,

$$\begin{aligned}
&\sum \left[y_{i_{ps}} \ln \left(\frac{1}{1 + \mathrm{e}^{-(\beta_{0_{ps}} + \beta_{1_{ps}} x_{1i_{ps}} + \beta_{2_{ps}} x_{2i_{ps}})}} \right) + (1 - y_{i_{ps}}) \ln \left(1 - \frac{1}{1 + \mathrm{e}^{-(\beta_{0_{ps}} + \beta_{1_{ps}} x_{1i_{ps}} + \beta_{2_{ps}} x_{2i_{ps}})}} \right) \right] \\
&+ \alpha \sum \left[y_{i_{nps}} \ln \left(\frac{1}{1 + \mathrm{e}^{-(\beta_{0_{nps}} + \beta_{1_{nps}} x_{1i_{nps}} + \beta_{2_{nps}} x_{2i_{nps}})}} \right) + (1 - y_{i_{nps}}) \ln \left(1 - \frac{1}{1 + \mathrm{e}^{-(\beta_{0_{nps}} + \beta_{1_{nps}} x_{1i_{nps}} + \beta_{2_{nps}} x_{2i_{nps}})}} \right) \right) \\
&+ \frac{-(\nu_0 + 1)}{2} \ln \left(1 + \frac{(\beta_p - \mu_0)^2}{\nu_0 \sigma_0^2} \right)
\end{aligned} \tag{8}$$

then taking the first derivative in respect to $\hat{\beta}_p$

$$\sum \left[y_{i_{ps}} \frac{1}{1+\mathrm{e}^{\beta_{0_{ps}}+\beta_{1_{ps}}x_{1i_{ps}}+\beta_{2_{ps}}x_{2i_{ps}}}} + (y_{i_{ps}}-1)\frac{\mathrm{e}^{\beta_{0_{ps}}+\beta_{1_{ps}}x_{1i_{ps}}+\beta_{2_{ps}}x_{2i_{ps}}}}{1+\mathrm{e}^{\beta_{0_{ps}}+\beta_{1_{ps}}x_{1i_{ps}}+\beta_{2_{ps}}x_{2i_{ps}}}} \right]$$
$$+\alpha \sum \left[y_{i_{nps}} \frac{1}{1+\mathrm{e}^{\beta_{0_{nps}}+\beta_{1_{nps}}x_{1i_{nps}}+\beta_{2_{nps}}x_{2i_{nps}}}} + (y_{i_{nps}}-1)\frac{\mathrm{e}^{\beta_{0_{nps}}+\beta_{1_{nps}}x_{1i_{nps}}+\beta_{2_{nps}}x_{2i_{nps}}}}{1+\mathrm{e}^{\beta_{0_{nps}}+\beta_{1_{nps}}x_{1i_{nps}}+\beta_{2_{nps}}x_{2i_{nps}}}} \right]$$
$$+\frac{(\nu_0+1)(\beta_p^2-\mu_0)}{(\beta_p-\mu_0)^2+\nu_0\sigma_0^2} \tag{9}$$

$$\sum \left[y_{i_{ps}} x_{\frac{1}{2}i_{ps}} \frac{1}{1+\mathrm{e}^{\beta_{0_{ps}}+\beta_{1_{ps}}x_{1i_{ps}}+\beta_{2_{ps}}x_{2i_{ps}}}} + (y_{i_{ps}}-1)x_{\frac{1}{2}i_{ps}}\frac{\mathrm{e}^{\beta_{0_{ps}}+\beta_{1_{ps}}x_{1i_{ps}}+\beta_{2_{ps}}x_{2i_{ps}}}}{1+\mathrm{e}^{\beta_{0_{ps}}+\beta_{1_{ps}}x_{1i_{ps}}+\beta_{2_{ps}}x_{2i_{ps}}}} \right]$$
$$+\alpha \sum \left[y_{i_{nps}} x_{\frac{1}{2}i_{nps}} \frac{1}{1+\mathrm{e}^{\beta_{0_{nps}}+\beta_{1_{nps}}x_{1i_{nps}}+\beta_{2_{nps}}x_{2i_{nps}}}} + (y_{i_{nps}}-1)x_{\frac{1}{2}i_{nps}}\frac{\mathrm{e}^{\beta_{0_{nps}}+\beta_{1_{nps}}x_{1i_{nps}}+\beta_{2_{nps}}x_{2i_{nps}}}}{1+\mathrm{e}^{\beta_{0_{nps}}+\beta_{1_{nps}}x_{1i_{nps}}+\beta_{2_{nps}}x_{2i_{nps}}}} \right]$$
$$+\frac{(\nu_0+1)(\beta_p^2-\mu_0)}{(\beta_p-\mu_0)^2+\nu_0\sigma_0^2} \tag{10}$$

and second derivative

$$\sum \left[y_{i_{ps}} \frac{\mathrm{e}^{\beta_{0_{ps}}+\beta_{1_{ps}}x_{1i_{ps}}+\beta_{2_{ps}}x_{2i_{ps}}}}{(1+\mathrm{e}^{\beta_{0_{ps}}+\beta_{1_{ps}}x_{1i_{ps}}+\beta_{2_{ps}}x_{2i_{ps}}})^2} + (y_{i_{ps}}-1)\frac{\mathrm{e}^{\beta_{0_{ps}}+\beta_{1_{ps}}x_{1i_{ps}}+\beta_{2_{ps}}x_{2i_{ps}}}}{(1+\mathrm{e}^{\beta_{0_{ps}}+\beta_{1_{ps}}x_{1i_{ps}}+\beta_{2_{ps}}x_{2i_{ps}}})^2} \right]$$
$$+\alpha \sum \left[y_{i_{nps}} \frac{\mathrm{e}^{\beta_{0_{nps}}+\beta_{1_{nps}}x_{1i_{nps}}+\beta_{2_{nps}}x_{2i_{nps}}}}{(1+\mathrm{e}^{\beta_{0_{nps}}+\beta_{1_{nps}}x_{1i_{nps}}+\beta_{2_{nps}}x_{2i_{nps}}})^2} + (y_{i_{nps}}-1)\frac{\mathrm{e}^{\beta_{0_{nps}}+\beta_{1_{nps}}x_{1i_{nps}}+\beta_{2_{nps}}x_{2i_{nps}}}}{(1+\mathrm{e}^{\beta_{0_{nps}}+\beta_{1_{nps}}x_{1i_{nps}}+\beta_{2_{nps}}x_{2i_{nps}}})^2} \right]$$
$$+\frac{(\nu_0+1)((\beta_p-\mu_0)^2-\nu_0\sigma_0^2)}{\left((\beta_p-\mu_0)^2+\nu_0\sigma_0^2\right)^2} \tag{11}$$

$$\begin{aligned}
&\sum \left[y_{i_{ps}} x_{1/2i_{ps}} \frac{\mathrm{e}^{\beta_{0_{ps}}+\beta_{1_{ps}}x_{1i_{ps}}+\beta_{2_{ps}}x_{2i_{ps}}}}{\left(1+\mathrm{e}^{\beta_{0_{ps}}+\beta_{1_{ps}}x_{1i_{ps}}+\beta_{2_{ps}}x_{2i_{ps}}}\right)^2} \right. \\
&\left. +(y_{i_{ps}}-1)x_{1/2i_{ps}} \frac{\mathrm{e}^{\beta_{0_{ps}}+\beta_{1_{ps}}x_{1i_{ps}}+\beta_{2_{ps}}x_{2i_{ps}}}}{\left(1+\mathrm{e}^{\beta_{0_{ps}}+\beta_{1_{ps}}x_{1i_{ps}}+\beta_{2_{ps}}x_{2i_{ps}}}\right)^2} \right] \\
&+\alpha \sum \left[y_{i_{nps}} x_{1/2i_{nps}} \frac{\mathrm{e}^{\beta_{0_{nps}}+\beta_{1_{nps}}x_{1i_{nps}}+\beta_{2_{nps}}x_{2i_{nps}}}}{\left(1+\mathrm{e}^{\beta_{0_{nps}}+\beta_{1_{nps}}x_{1i_{nps}}+\beta_{2_{nps}}x_{2i_{nps}}}\right)^2} \right. \\
&\left. +(y_{i_{nps}}-1)x_{1/2i_{nps}} \frac{\mathrm{e}^{\beta_{0_{nps}}+\beta_{1_{nps}}x_{1i_{nps}}+\beta_{2_{nps}}x_{2i_{nps}}}}{\left(1+\mathrm{e}^{\beta_{0_{nps}}+\beta_{1_{nps}}x_{1i_{nps}}+\beta_{2_{nps}}x_{2i_{nps}}}\right)^2} \right] \\
&+\frac{(\nu_0+1)\left((\beta_p-\mu_0)^2-\nu_0\sigma_0^2\right)}{\left((\beta_p-\mu_0)^2+\nu_0\sigma_0^2\right)^2}
\end{aligned} \tag{12}$$

and finally by taking minus the inverse, resulting in

$$\hat{\sigma}_{\hat{\beta}}^2 = -\frac{1}{\mathcal{I}(\hat{\beta}_p)}$$

.

Given that the proposal distribution is symmetric, the acceptance ratio can be simplified to

$$r = \frac{\pi(\beta^*)}{\pi(\beta_{t-1})} \frac{q(\beta^*)}{q(\beta_t - 1)} = \frac{\pi(\beta^*)}{\pi(\beta_{t-1})} \tag{13}$$

because

$$q(\beta^*) = q(\beta_{t-1}) \tag{14}$$

This special case is simply called a metropolis sampler.

4.1 Results

The sampler is run with 3 chains, with each chain sampling 10000 times. The baseline prior is defined as mildly informative as proposed by Salvatore et al. (2024) with the parameters $\mu_0 = 0, \sigma_0 = 2.5, \nu_0 = 3$. The first 3000 samples were removed as burn in period. Starting values β_p are $beta_0 = 1, \beta_1 = 2, \beta_2 = 3$. Through all sampling processes, a seed is set to ensure reproducibility. Finally, running the sampler and estimating the parameters $beta_0, \beta_1$ and β_2 leads to the results shown in Table 1. The trace plots in Figure 1 to 3 show that all

Table 1: Estimates, 95% Credible Intervals, and MSE for Logistic Regression Coefficients

	Estimate	95% Credible Interval	MSE($\hat{\beta_p}$)
β_0	0.47	[0.21; 0.73]	0.0183
β_1	0.83	[0.60; 1.07]	0.0154
β_2	0.40	[0.18; 0.62]	0.0225

three chains converged well.

Further, the sampler has an acceptance rate of 0.9996, 0.9993 and 0.9998 for the three estimates respectively and each has an autocorrelation of 0.299.

Next to the estimates, Table 1 also shows the mean squared error (MSE), a measure quantifying the total survey error TSE as described in Biemer (2010). The MSE is calculated taking bias and variance of the sample into account.

$$MSE(\hat{\beta_p}) = Bias^2(\hat{\beta_p}) + Var(\hat{\beta_p}) \tag{15}$$

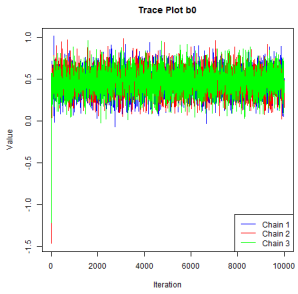

Figure 1: Traceplot β_0

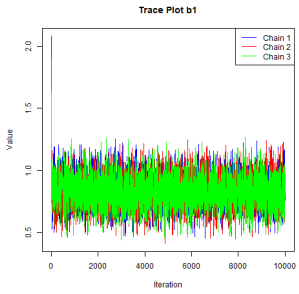

Figure 2: Traceplot β_1

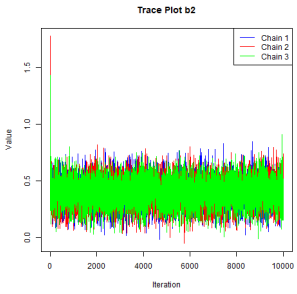

Figure 3: Traceplot β_2

A lower MSE can be interpreted as a better estimate.

The intercept $\hat{\beta}_0$ is estimated to be 0.47 with a 95% CI from 0.21 to 0.73. The MSE is 0.0183, showing that the estimate is very close to the population parameter $\beta_0 = 0.5$. $\hat{\beta}_1$ is estimated with 0.83 with a 95% CI from 0.6 to 1.07 and an MSE of 0.0154, again confirming a good estimate close to the population parameter $\beta_1 = 0.8$. Lastly, $\hat{\beta}_2$ is estimated to be 0.4, with a 95% CI of 0.18 to 0.62 and an MSE of 0.0225. The population parameter $\beta_2 = 0.3$ was again well estimated. All three parameters show somewhat large credible intervals. Further analyses could investigate if a different baseline prior or a different estimation of the α parameter of the power prior can lead to a closer estimation of the credible intervals.

In the following section a Bayesian hypothesis testing framework is used to compare the estimates to the population parameters.

5 Bayesian Hypothesis Testing

In the following section the R-package $bain$ (Gu, Hoijtink, Mulder, & Rosseel, 2019) is used to perform a test of the estimated parameter against the population parameter. A Bayes Factor is computed to test the hypothesis $\hat{\beta}_p = \beta_p$. The Bayes Factor is a quantification of the support for H_1 over the unconstrained hypothesis H_u. It is computed by dividing the marginal likelihoods the hypotheses

$$BF_{1u} = \frac{m_1}{m_u} = \frac{fit_1}{complexity_1} \tag{16}$$

when the hypotheses are nested, the Bayes Factor can be computed solely through the fit and complexity of H_1 (Hoijtink, 2011). Table 2 shows the results of the Bayesian hypothesis

tests.

	Fit	Com	BF.u	BF.c	PMPa	PMPb	PMPc
$H_1 : \hat{\beta}_0 = 0.5$							
H_1	2.947	0.061	48.657	48.657	1.000	0.980	0.980
H_u					0.020		
H_c						0.020	
$H_1 : \hat{\beta}_1 = 0.8$							
H_1	3.206	0.067	47.963	47.963	1.000	0.980	0.980
H_u					0.020		
H_c						0.020	
$H_1 : \hat{\beta}_2 = 0.3$							
H_1	2.417	0.070	34.333	34.333	1.000	0.972	0.972
H_u					0.028		
H_c						0.028	

Table 2: Bayesian Hypothesis Test Results

The first part show the hypothesis test for the hypothesis $\hat{\beta}_0 = 0.5$. The Bayes Factor is 48.66, saying that there is 48.66 times the amount of support in favour of the hypothesis against the unconstrained hypothesis. The posterior model probability is 0.98. The posterior model probabilities can be interpreted as Bayesian error probabilities, as counterparts to the frequentist Type I and Type II error rates. With a certainty of 98% the hypothesis $\hat{\beta}_0 = 0.5$ can be accepted. The second part shows the BF and PMP for hypothesis $\hat{\beta}_1 = 0.8$. The Bayes Factor is 47.96 and the PMP again is 0.98. The same conclusion as with the previous hypothesis can be drawn. Lastly, $\hat{\beta}_2 = 0.3$ shows a smaller Bayes Factor of 34.33 and a PMP of 0.97. Although slightly less supported, there is still a large amount of evidence that in all three cases the hypothesis $\hat{\beta}_p = \beta_p$ can be accepted.

6 Reflection

To show a Bayesian data integration approach, a bayesian logistic regression model estimating a binary outcome variable with two predictors was run on two integrated data sets, drawn from the same population, using a Metropolis sampler. Further, the viability of the estimates was tested using a Bayesian hypothesis testing framework.

The estimation and the hypothesis test do indicate that this approach to data integration is viable and gives good estimates to the population parameters. The slightly worse estimate for the β_2 parameter probably results from the x_2 variable being used to create the weight with which the biased NPS was sampled. Further research can compare differently strong biased non probability samples and different sample sizes, as well as different amounts of parameters.

Bibliography

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public opinion quarterly*, *74*(5), 817–848.

Chib, S., & Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The american statistician*, *49*(4), 327–335.

Gu, X., Hoijtink, H., Mulder, J., & Rosseel, Y. (2019). Bain: A program for bayesian testing of order constrained hypotheses in structural equation models. *Journal of Statistical Computation and Simulation*, *89*(8), 1526–1553.

Hoijtink, H. (2011). *Informative hypotheses: Theory and practice for behavioral and social scientists*. CRC Press.

Ibrahim, J. G., Chen, M.-H., Gwon, Y., & Chen, F. (2015). The power prior: theory and applications. *Statistics in medicine*, *34*(28), 3724–3749.

McCullagh, P. (2019). *Generalized linear models*. Routledge.

Salvatore, C., Biffignandi, S., Sakshaug, J. W., Wiśniowski, A., & Struminskaya, B. (2024). Bayesian integration of probability and nonprobability samples for logistic regression. *Journal of Survey Statistics and Methodology*, *12*(2), 458–492.

Appendices