Estimation of sparse data using bayesian regression techniques

William Pang[1]

[1]Department of Epidemiology of Microbial Diseases, Yale University

Introduction

Consider AIDS, a disease that affects 0.36% of the US population [1]. Imagine we are given the EHR data of *n = 1000 patients* at the Yale New Haven Hospital, and we might want to estimate the odds ratio (OR) of different predictors that could lead to AIDS. One such predictor is having a prior HIV diagnosis, an arguably strong predictor for AIDS. Since maximum likelihood regression techniques require balanced data, we would expect our effect estimates to be heavily biased [2]. Bayesian regression techniques, on the other hand, can penalize predictors based on prior information. This project aims to examine whether bayesian methods can perform better under sparse data scenarios.

Data and Statistical Model

We will be simulating our sparse dataset with 1000 samples (*n = 1000)* and 100 predictors (*p = 100)*. Predictors were generated with `mvnnorm` $N \sim (0,I_p)$.

The regression coefficients were generated with `rnorm` $N \sim (0,0.25)$; then, we set $\beta_1 = 5$. To make the data more sparse, we tune the intercept to be more negative. We then transform the linear model $\tilde{y} = X\beta$ into logistic regression using the logit link, and then use `rbinom` to transform $\tilde{y}$ into binary outcomes.

We model: $Y_i \mid p_i \sim \text{Bernoulli}(p_i)$

$\text{logit}(p_i) = X_{i,j}\beta_j$

IID Prior: $\beta_j \sim N(0,100^2)$

Ridge Prior: $\beta_j \sim N(0,\lambda), \lambda \sim IG(0.01,0.01)$

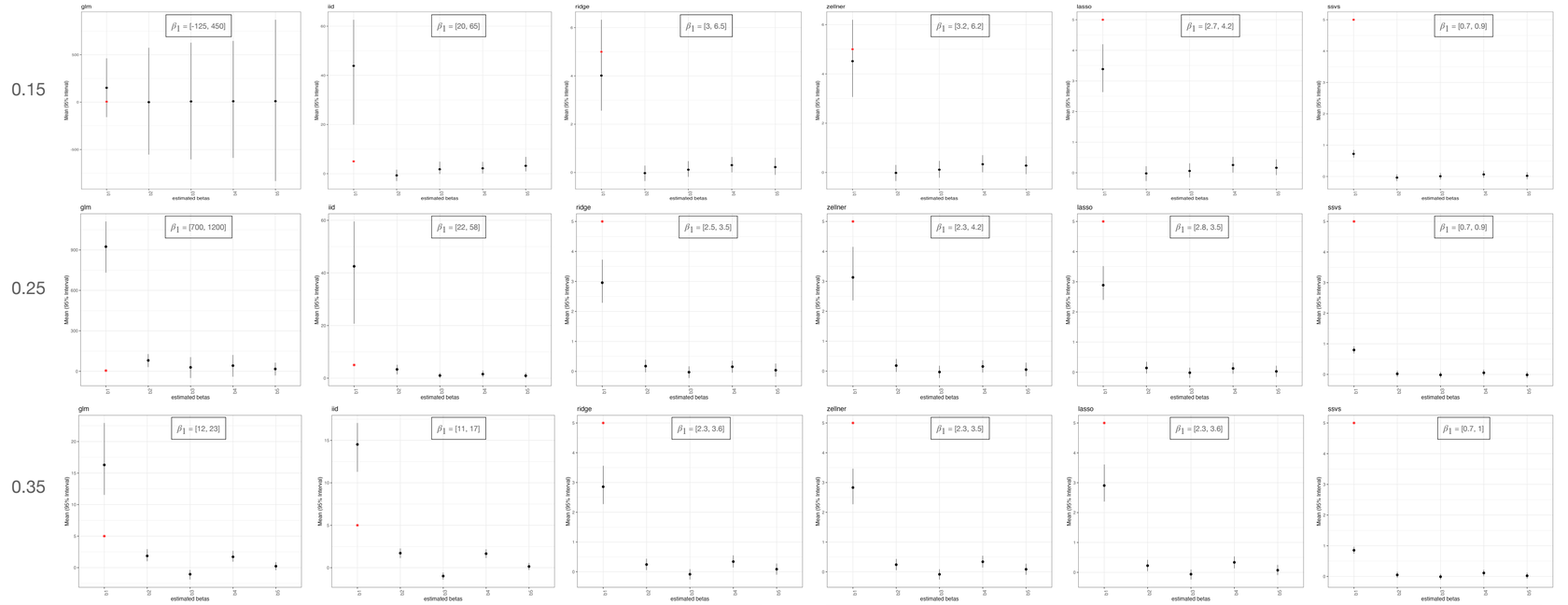
Zellner Prior: $\beta \sim MVN(0,\lambda(X^TX)^{-1}), \lambda \sim IG(0.01,0.01)$

Lasso Prior: $\beta_j \sim f(\beta_j); f(\beta_j) \propto \frac{1}{\lambda}e^{-\frac{|\beta_j|}{2\lambda}}, IG \sim (0.01,0.01)$

SSVS Prior:

$\beta_j \mid \delta j \sim N(0,0.01^2(1-\delta_j + 0.10^2\delta_j), \delta_j \sim Bern(\pi), \pi \sim Unif(0,1)$

Results



Conclusions and Significance

- We find that glm struggles the most with estimating the beta coefficients, often failing to converge and reporting coefficients that significantly biases the OR away from the null. Ridge, Zellner, and Lasso seem to perform fairly well, and either one of these approaches could be adopted. SSVS seems to consistently underestimate the true OR.
- This project demonstrated that traditional regression tools can lead to widely inaccurate estimates on sparse data. With the wide adoption of machine learning (ML) techniques to make predictions which rely on frequentist estimations, an overinflated OR could lead to predictions that significantly bias against certain groups (in our example, it would be patients with prior HIV diagnosis).

Future Work and References

For the next step, I would like to use regression coefficients generated by glm and bayesian methods to predict the outcome. I would then use standard ML metrics (i.e., accuracy, F1, precision, accuracy) to propose the best strategy.

[1] Centers for Disease Control and Prevention (CDC. "HIV prevalence estimates--United States, 2006." MMWR. Morbidity and mortality weekly report 57.39 (2008): 1073-1076.

[2] Greenland, Sander, Mohammad Ali Mansournia, and Douglas G. Altman. "Sparse data bias: a problem hiding in plain sight." bmj 352 (2016).

Acknowledgements: I thank Professor Joshua Warren (Biostatistics) as well as Biqing Zhu for their technical expertise and advice in this project.

Note: This is a poster presentation for Fall 2022 — Bayesian Statistics (BIS567A)