Homework 3
Lincole Jiang (lj2575)
November 22, 2023

Background

For this homework, we implemented a Hamiltonion Monte Carlo (HMC) and an automatic differentiation variational inference (ADVI) as for a hierarchical Bayesian generalized linear model (GLM) to infer on the possible correlates or causations of the increasing occurence of hypertension amongst U.S. adults over the past decade. The dataset considered was available from the National Health and Nutrition Examination Survey (NHANES) through Centers for Disease Control (CDC), and was accessed from the `cardioStatsUSA` package in `R`. Since this exercise serves only as a preliminary dataset, we only considered the potential effects of demographic and comorbidity variables on hypertension. Specifically, the outcome variable is `htn_jnc7`, which is an indicator variable on whether the surveyed individual has hypertension defined by JNC7 guideline at time of survey. We consider 10 groups for the hierarchical structure, diffentiated by `svy_year`, or the survey periods, which runs biennially from 1999-2016 (i.e., 1999-2000, 2001-2002, etc.) with the last cycle from 2017-2020. For demographic variables, we considered age as a scaled continuous variable (`demo_age`), race and ethnicity as a categorical variable with five levels–non-Hispanic white, non-Hispanic black, non-Hispanic Asian, Hispanic, and other (`demo_race`); pregnancy status as an indicator variable (`demo_pregnant`), as well as gender (`demo_gender`). For comorbidity variables, we considered smoking status (`cc_smoke`, never, former, or current) and BMI index (`cc_bmi`, <25, 25 - <30, 30 - <35, and 35+) as categorical variables, diabetic status (`cc_diabetes`), prevalent chronic kidney disease (`cc_ckd`), history of myocardial infarction (`cc_cvd_mi`), history of coronary heart disease (`cc_cvd_chd`), history of stroke (`cc_cvd_stroke`), ASCVD (`cc_cvd_ASCVD`), heart failure (`cc_cvd_hf`), and CVD overall (`cc_cvd_any`) as indicator variables. There are a total of 56,017 data points, but for this preliminary analysis only a total of 5000 randomly data points were considered to fit the model and an additional 5000 data points were considered to evaluate the test error. The categorical variables were made into dummy one-hot vectors for each category, which produced a predictor matrix of 22 columns.

Methods

Model Specification

We considered a hierarchical Bayesian GLM for this exercise. Explicitly, denoting p as the number of input features ($p = 22$), n as the number of groups of data ($n = 10$), m_i as the number of datapoints in each group, λ as the hyperprior, $\pi(y), t(y), a_y(\eta)$ as the exponential family for the response, and η_μ as the mean mapping for the response family, f as the canonical link function, and $x_{ij} \in {}^p$ as feature vectors, we generate the model as follows:

1. Draw the prior parameter
$$\lambda \sim \text{gamma}(\eta, 1)$$

2. For each group i:

(a) Draw coefficients

$$\beta_i \mid \lambda \sim \mathrm{N}(0, \lambda^2)$$

(b) For each datapoint j, draw

$$y_{ij} \mid x_{ij}, \beta_i \sim \text{expfam}(\eta_{ij}) \qquad \eta_{ij} = \eta_\mu(f(\beta_i \cdot x_{ij}))$$

In particular, the exponential family we considered here is the Bernoulli-Beta exponential family, and denoting θ as the probability of outcome, i.e., having hypertension, we have $\eta = \log(\frac{\theta}{1-\theta}), t(x) = x, \pi(x) = 1, a(\eta) = -\log(1-\theta) = \log(1+e^\eta)$.

Algorithm

We implemented both ADVI and HMC for this model, using built-in features in `stan`. The algorithm for ADVI is specified in Algorithm 1 of the Kucukelbir 2017 paper, whereas the No-U-Turn sampler, which was the default MCMC algorithm implemented in `stan` and a variant of HMC, was detailed as Algorithm 2 in the 2014 Hoffman and Gelman paper (both cited in the reference page).

Results

We first implemented the ADVI algorithm, which gave very poor performances with Pareto k-values ranging from 3.6 to 4.4, way above the desired range of within 1. This problem persisted despite dramatically increasing the number of iterations (to 10^5) and decreasing convergence tolerance on the relative norm of the objective (to 10^{-6}). Therefore, we followed the advice of the program and moved to MCMC. The No-U-Turn sampler swiftly converged, as both indicated by the traceplot shown in Figures 1-4 of λ and selected β's (3 out of 220) and the $\hat{R}$ or Gelman-Rubin statistic, which ranged from 0.9995385 to 1.0019473 across all parameters, indicating good convergence. On the other hand, in terms of model fitness, we calculated the posterior predictive for each training data point using the formula

$$p(x \mid \mathbf{x}) = \pi_l(x) \exp a_c(\hat{\lambda}_1 + t(x), \hat{\lambda}_2 + 1 - a_c(\hat{\lambda}_1, \hat{\lambda}_2))$$

which was then classified as 0 if the expected value <0.5 and >= 0 and classified to 1 if >= 0.5 but <= 1. In this way, the classification error rate was calculated as the number of times the model has misclassified the response, which came out as 0.471662, giving a model training accuracy of 0.528338 and indicating a very poor fit. Finally, For interpretations of the β, we listed the β_{ij} values for the most positive and negative predictors over survey periods (chosen according to mean β values across survey periods) in Table 1. As we can see, the most significant positive predictors of hypertension status is hypertension awareness, whereas the most negative predictor for hypertension status is being non-Hispanic white. More in-dept analysis will be performed in the final project to illustrate the changes in the hypertension predictors over the survey period.

Discussion

For the choice of priors and hyperparameters, the β's are chosen as 0-mean normal with unknown prior on variance, which then is assumed to have a gamma distribution with parameters η and 1, in which η is solely informed by the data from other groups/survey periods, while the β_i's are specific to each grouping. Perhaps the assumed-fixed hyperparameter η still warrants some experimentation to select the optimal value that produces the least test and training prediction error. Furthermore, we should consider models that accounts for the time-series heterogeneity in hopes for better model performance.

References

Byron Casey Jaeger (2022). NHANES Data, 1999 - 2020

Hoffman, Matthew D., and Andrew Gelman. "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." J. Mach. Learn. Res. 15.1 (2014): 1593-1623.

Kucukelbir, Alp, et al. "Automatic differentiation variational inference." Journal of machine learning research (2017).

Blei. Lecture Notes. (2023)

Appendix: Figures & Tables

Table 1

Survey Period	Hypertension Status	Non-Hispanic White
1999-2000	3.2352373639	-2.6116734895
2001-2002	3.5026493575	-1.9636719069
2003-2004	3.3102904474	-2.4422967916
2005-2006	3.8844325237	-2.7791472028
2007-2008	3.3758437574	-2.0226292324
2009-2010	3.0503616559	-2.5703432902
2011-2012	2.7907967919	-1.9897728653
2013-2014	4.3315637200	-1.7796450273
2015-2016	3.3683797613	-2.3094789503
2017-2020	3.8097639783	-2.7573642535

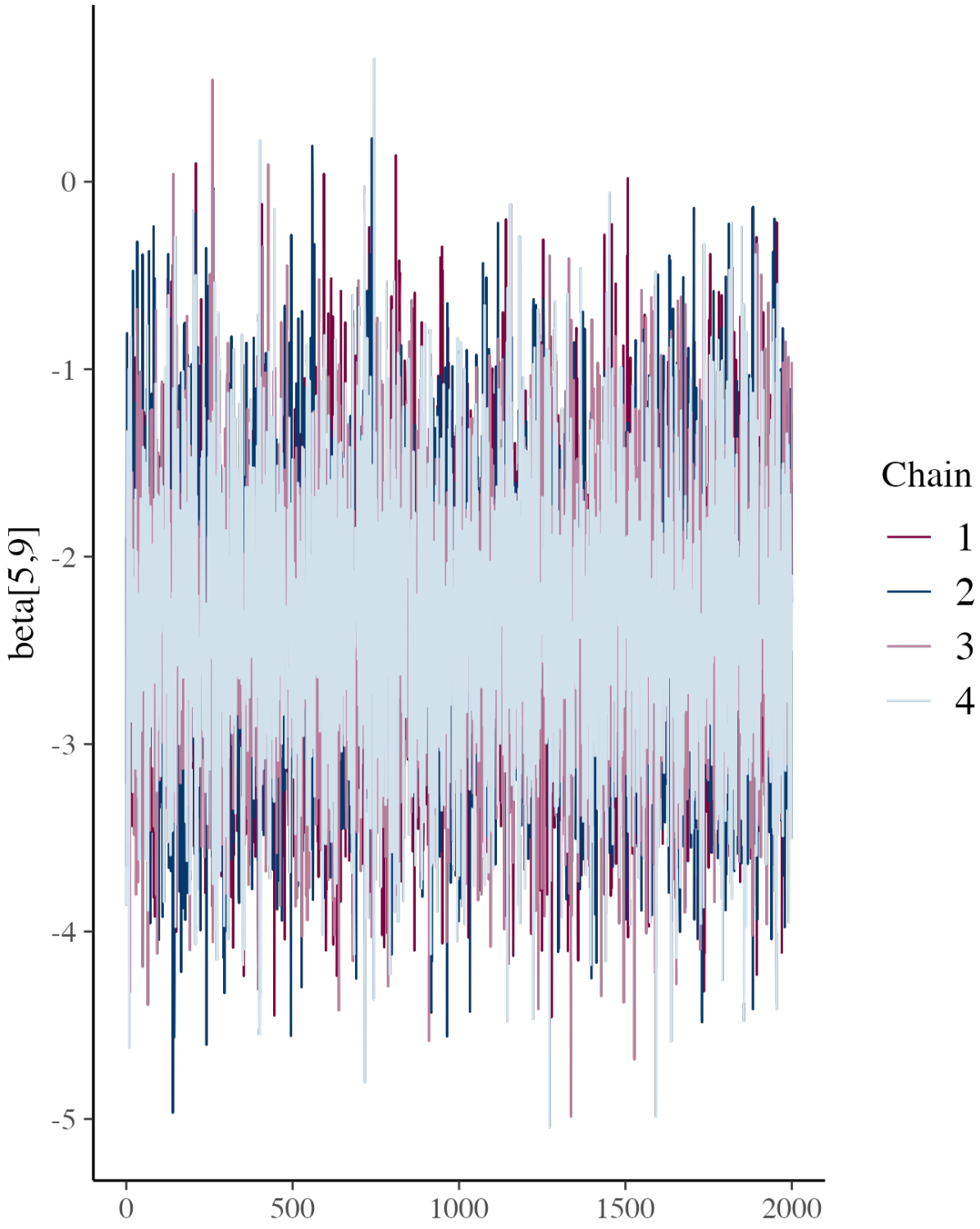

Figure 1: Beta for survey cycle 2015-2016, race-others over number of iterations

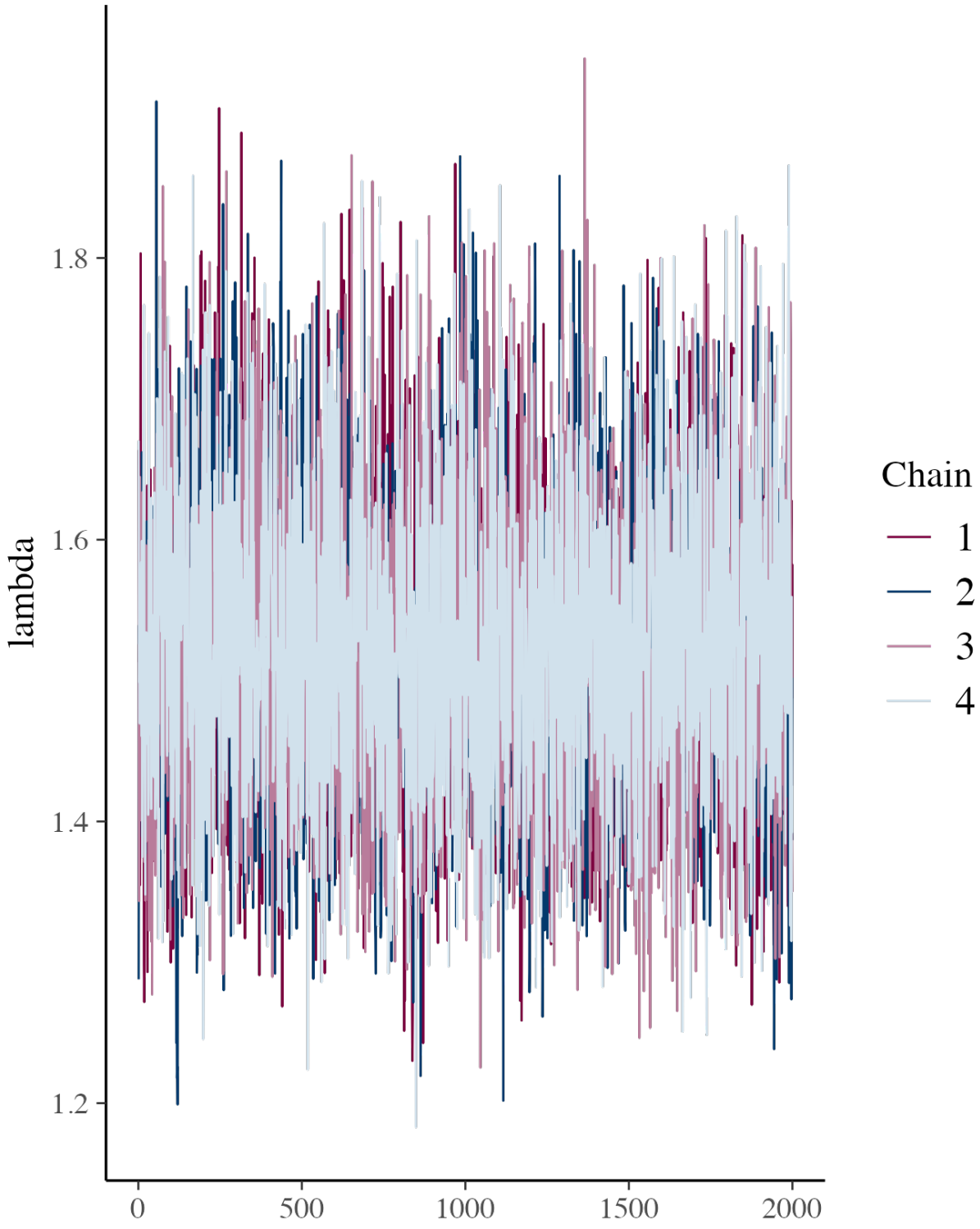

Figure 2: Parameter lambda over number of iterations

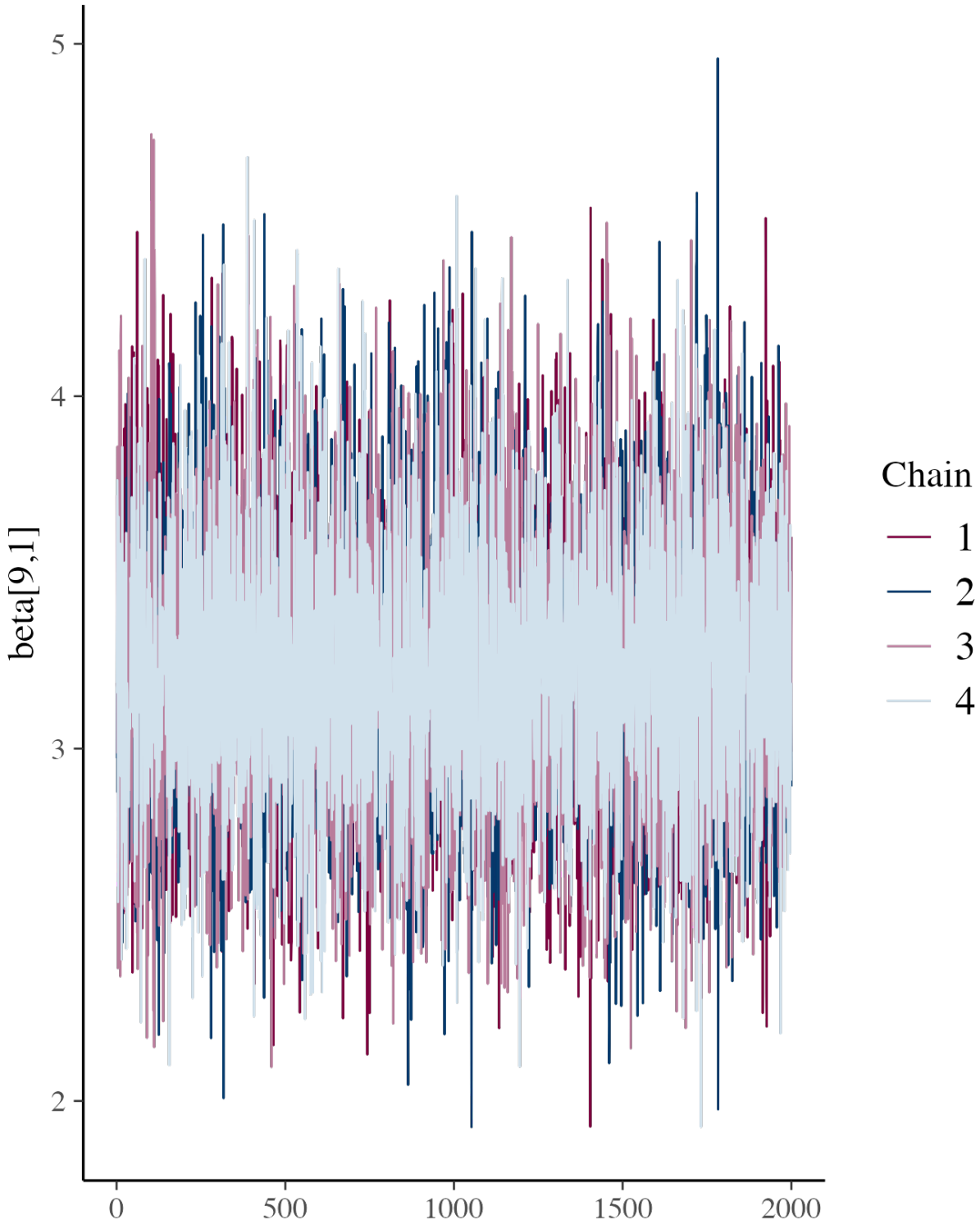

Figure 3: Beta for survey cycle 2000-2001, hypertension awareness, over number of iterations

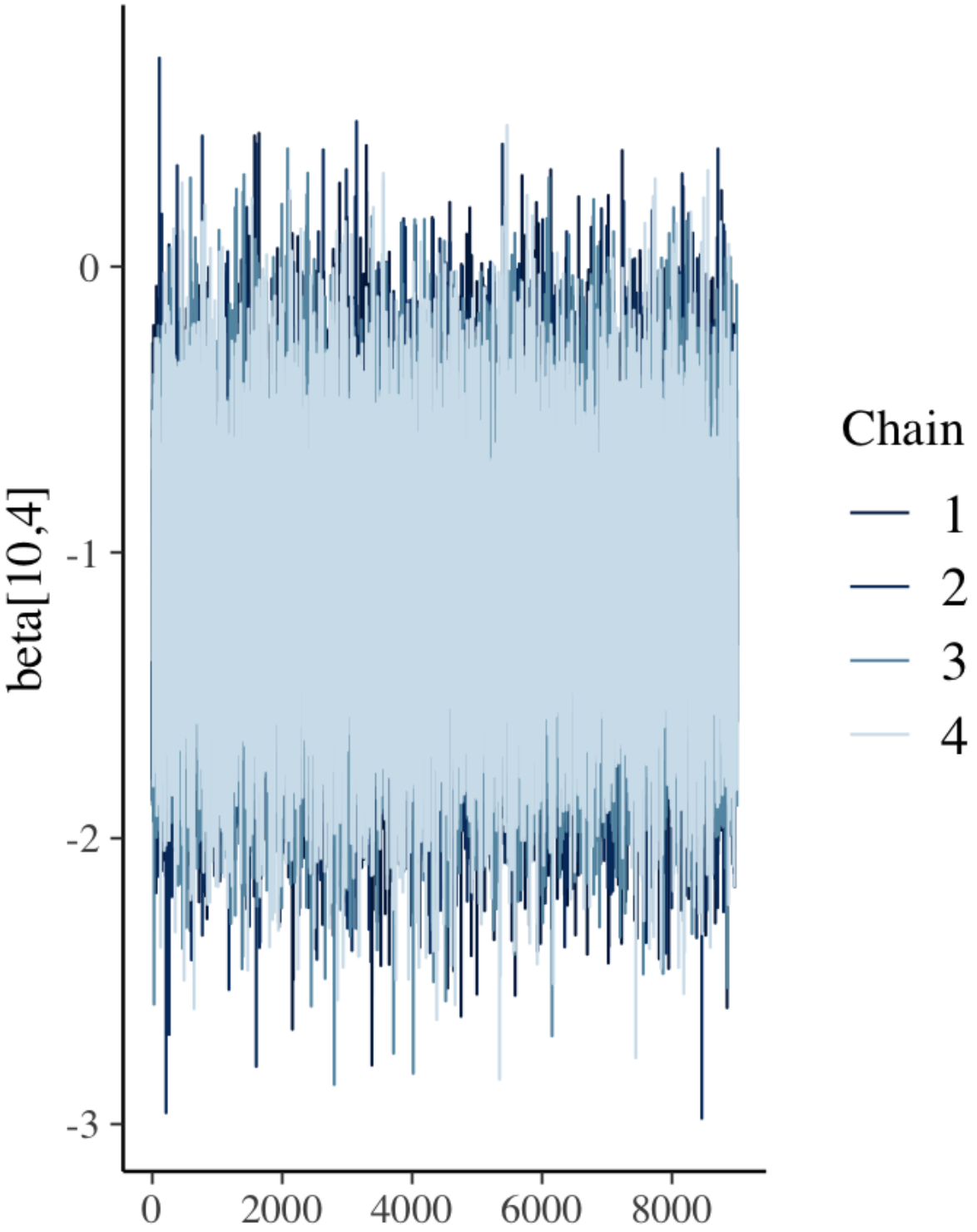

Figure 4: Beta for survey cycle 2003-2004, history of coronary heart disease over number of iterations