Your Age, Gender and Place of Birth May Explain Why You Are Alone in Canada

Xi Cheng, Shichao Feng and Zhitong Liu

October 18, 2020

Contents

1 Abstract

It has been widely reported that more Canadians live alone now than ever before. In this study, the Bayesian hierarchical logistic regression model was applied to the 2017 Canadian General Social Survey – family (GSS) data, to discover potential important factors related with the singleness rate, including never married, divorced, separated and widowed cases. Our work demonstrated that age, gender and the place of birth (POB), and the correlation between age and sex were significantly associated with the singleness rate in Canada. Our results confirmed the descriptive results reported by Statistics Canada and provided a deeper understanding of the singleness issue in Canada.

2 Introduction

The General Social Survey was designed to gather data on social trends in order to monitor changes of Canadians over time. The 2017 GSS focused changes in Canadian families, which contains information on marriages, family origins, child care and other socioeconomic characteristics (Statistics Canada, 2017). Descriptive results from this dataset have been published, which suggest different age groups have different

likelihood to be separated or divorced (Statistics Canada, 2019). Moreover, women are more likely to be separated or divorced, and Canadian born as well (Statistics Canada, 2019).

This work was built upon those findings, and was developed to test the associations between those factors such as age, gender and POB, and the likelihood of being alone. Particularly, the marital status was dichotomized into single and non-single. The single group was made up of divorced cases, separated cases, never-married single cases and widowed cases. The main reason of combining different sources of single group is to deal with the unbalanced group problem. Different single groups are minority groups, from 3.1% to 22.9% of whole population. Combination of groups together can increase sample size in the minority group, and increase the power of the statistical tests as well.

The characteristics of the subset of 2017 GSS dataset used in the model fitting was demonstrated in this work. With a basic understanding of the dataset, the generalized linear model (GLM) approach was applied first to generate preliminary results. The preliminary results can guide us during the Bayesian Hierarchical logistic regression model, especially the model comparison stage. The detailed information of model used was shown in the **Model** section.

This study demonstrated that age, gender and POB are significantly associated with the singleness rate. Particularly, women and elderly are more likely to be alone, and people born in Canada are more likely to be alone as well. Especially, the effect of age is significantly modified by gender, which means elder women are in the highest risk of being alone. Our results proved the descriptive reports published by Statistics Canada and might help in programs and policy making involving spousal support.

3 Data

The original dataset is the 2017 GSS focused changes in Canadian families (Statistics Canada, 2017). The survey is specially designed to gather information and impact areas involving spousal support, child care and parental benefits. The data collection through this survey was last from 2017-02-01 to 2017-11-30, which is from a sample survey with a cross-sectional design. The target population of this survey is all non-institutionalized persons 15 years of age or older, living in the 10 provinces of Canada (Statistics Canada, 2017). This is a large national survey, where approximately 43,000 questionnaires were sent and more than 20,000 of them were completed. The sampling frame is a list of address, associated with one or several telephone numbers. Moreover, GSS only selects one eligible person per household to be interviewed. The sampling is based on a stratified design employing probability sampling, where the stratification is done at province/census metropolitan area (CMA) level (Statistics Canada, 2017). The responding to this survey is voluntary, with overall response rate is 52.4%.

The

4 Model

Given that the predicted outcome estimated in this work is binary, either single or not, the natural modeling approach would be logistic regression. The outcome, singleness rate, was represented using Marital in this study. The predictors selected in this study contained age, gender(sex) and POB(place), the latter two are binary variables. Thus, the intercept of the model, and the coefficients of age might differ in different levels of gender and POB, which means the final model might have both population level and group level coefficients. All together, these suggested the final model might be a hierarchical logistic regression, and both Frequentist and Bayesian approach were used. The Frequentist approach using the generalized linear model (GLM) is typically faster to fit compared with the Bayesian approach. Thus, the GLM approach can serve as the preliminary study of the Bayesian approach, making the model selection stage faster.

The models fitted in this study were listed below, following the R package "brms" model demonstration convention. In logistic regression, the outcome follows the Bernoulli/Binomial distribution and the probability

in Bernoulli trial was transformed using Logit function.

$$y \sim Bernoulli(\frac{1}{1+exp(-(a+bx))}) \tag{1}$$

$$Marital \sim Bernoulli(\frac{1}{1+exp(-(a+b_{age}x_{age}))}) \tag{2}$$

$$Marital \sim Bernoulli(\frac{1}{1+exp(-(a+b_{sex}x_{sex}))}) \tag{3}$$

$$Marital \sim Bernoulli(\frac{1}{1+exp(-(a+b_{place}x_{place}))}) \tag{4}$$

$$Marital \sim Bernoulli(\frac{1}{1+exp(-(a+b_{age}x_{age}+b_{sex}x_{sex}+b_{place}x_{place}))}) \tag{5}$$

$$Marital \sim Bernoulli(\frac{1}{1+exp(-(a+b_{age}x_{age}+b_{sex}x_{sex}+b_{place}x_{place}+b_{age*sex}x_{age*sex}+b_{age*place}x_{age*place}))}) \tag{6}$$

5 Methods

Code supporting this analysis is available at: https://github.com/Chelsea-Cheng99/STA304/tree/master/ProbSet3. All datasets used in this study were kept locally, if you have the access of the 2017 GSS data, please check the gss_cleaning-1.R code to get the full dataset, followed by the getDf.R code to get the subset of data used in this work. All models were fit in the runModelBrms.R. Figures and tables were partly generated in mcmcplot.R. With the data.

All work were done in R (version 4.0.2) (R Core Team 2020) and Rstudio (version 1.3.1093). Tidyverse (version 1.3.0) was used for data wrangling and visualization (Wickham et al., 2019). R package "forcats" (version 0.5.0) was also used for data pre-processing (Hadley Wickham, 2020). All R packages used in this study

6 Results

Table. 1: Characteristics Summary of the 2017 GSS Data

	Overall (N=20499)
Age	
Mean (SD)	52.199 (17.748)
Range	15.000 - 80.000
Gender	
Female	11155 (54.4%)
Male	9344 (45.6%)
Place of Birth	
Born in Canada	16350 (79.8%)
Born outside Canada	4096 (20.0%)
Don't know	53 (0.3%)

	Overall (N=20499)
Marital Status	
Divorced	1760 (8.6%)
Living common-law	2066 (10.1%)
Married	9453 (46.1%)
Separated	640 (3.1%)
Single, never married	4688 (22.9%)
Widowed	1892 (9.2%)

Table. 2: Characteristics Summary of the Combined groups of 2017 GSS

	Overall (N=20446)
Age	
Mean (SD)	52.212 (17.752)
Range	15.000 - 80.000
Gender	
Female	11124 (54.4%)
Male	9322 (45.6%)
Place of Birth	
Born in Canada	16350 (80.0%)
Born outside Canada	4096 (20.0%)
Marital Status	
Nonsingle	11484 (56.2%)
Single	8962 (43.8%)

Table. 3: Summary of Logistic Regression Models (Generalized Linear Models)

Table. 4: Summary of LOOIC of Logistic Regression Models (Bayesian)

Number	Models	LOOIC
1	~ sex	27940.8
2	~ age + sex	27871.9
3	~ age + (1 + age | sex)	27506.4
4	~ age + (1 + age | sex) + (1 + age | place)	27469.6

Table. 5: Standard Deviation Test: Complete Bayesian Hierarchical Model

Hypothesis	Estimate	Error	Lower 95% CI	Upper 95% CI
s.d.: (Intercept-age) > 0 (Sex Group)	3.047317	3.050582	0.4499506	8.731504
s.d.: (Intercept-age) > 0 (POB Group)	2.461309	3.293193	0.1347284	8.133125

Figure. 1:

Figure. 2:

Table 3:

	Dependent variable:				
	factor(Marital)				
	(1)	(2)	(3)	(4)	(5)
age	-0.006^{***}			-0.007^{***}	0.006^{***}
	(0.001)			(0.001)	(0.001)
sexMale		-0.277^{***}		-0.288^{***}	1.323^{***}
		(0.028)		(0.029)	(0.089)
placeBorn outside Canada			-0.212^{***}	-0.218^{***}	-0.374^{***}
			(0.036)	(0.036)	(0.110)
age:sexMale					-0.031^{***}
					(0.002)
age:placeBorn outside Canada					0.003
					(0.002)
Constant	0.081^{*}	-0.123^{***}	-0.206^{***}	0.283^{***}	-0.410^{***}
	(0.044)	(0.019)	(0.016)	(0.047)	(0.064)
Observations	20,446	20,446	20,446	20,446	20,446
Log Likelihood	−13,984.510	−13,968.380	−13,998.250	−13,914.070	−13,728.670
Akaike Inf. Crit.	27,973.010	27,940.760	28,000.500	27,836.150	27,469.340
Note:					*p<0.1; **p<0.05; ***p<0.01

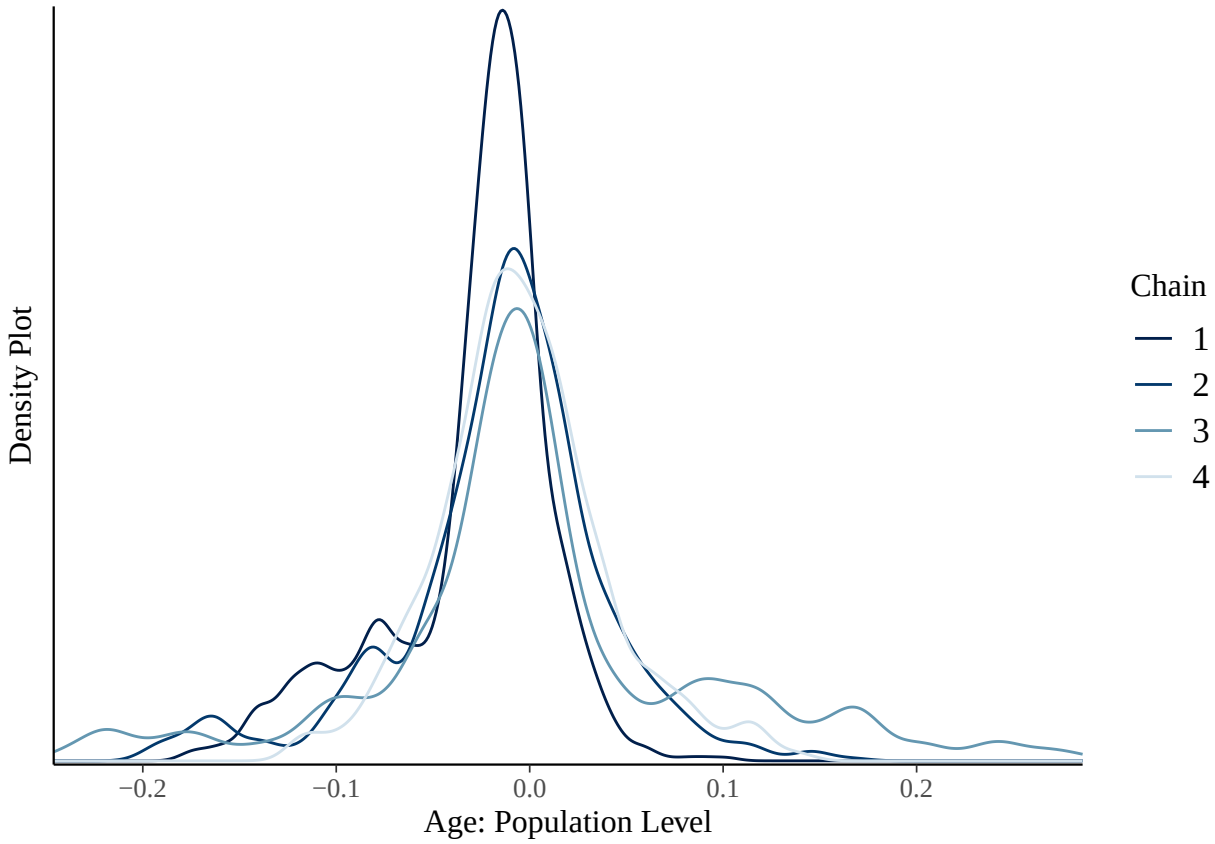


Figure 1: Table. 6: The Posterior Distribution of Age: Population Level (Bayesian Model3)

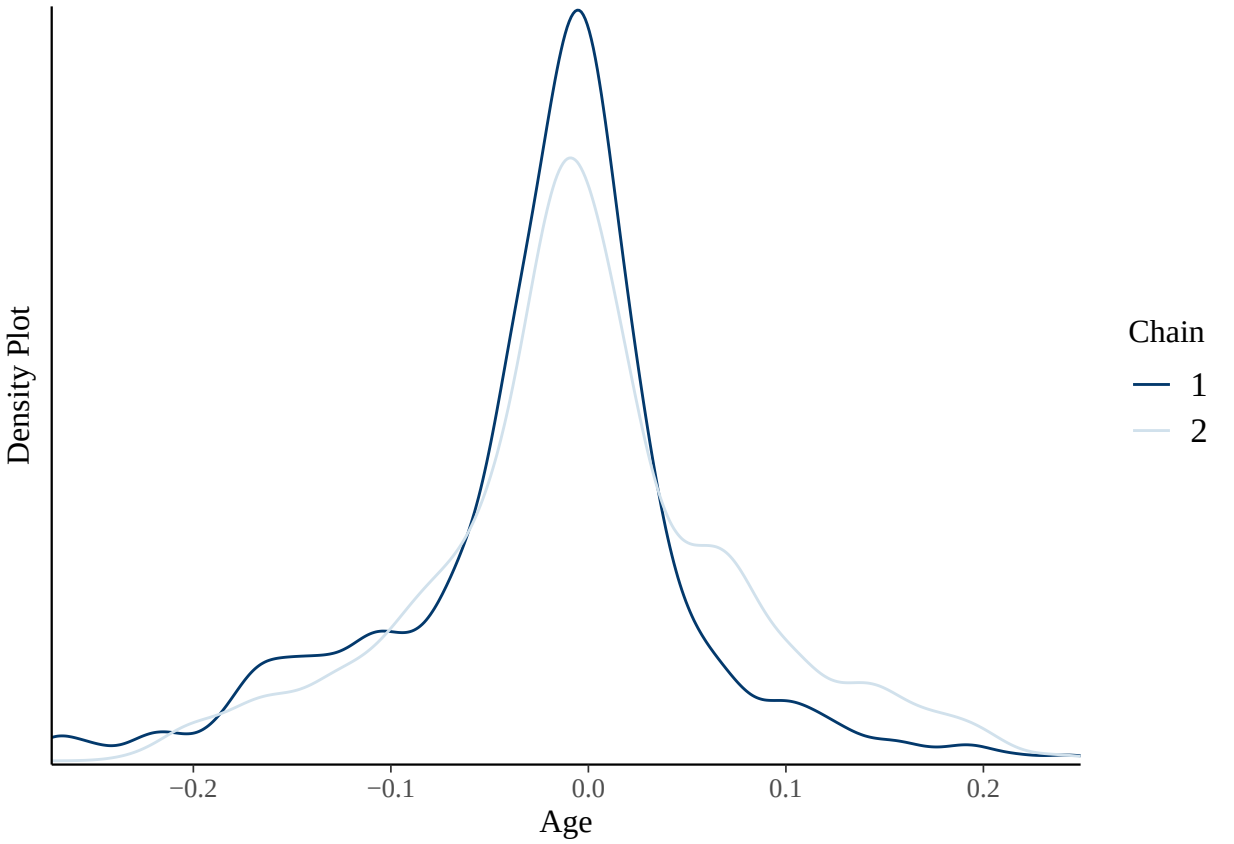


Figure 2: Table. 7: The Posterior Distribution of Age: Population Level (Bayesian Model4)

7 Discussion

In the 2017 GSS dataset.

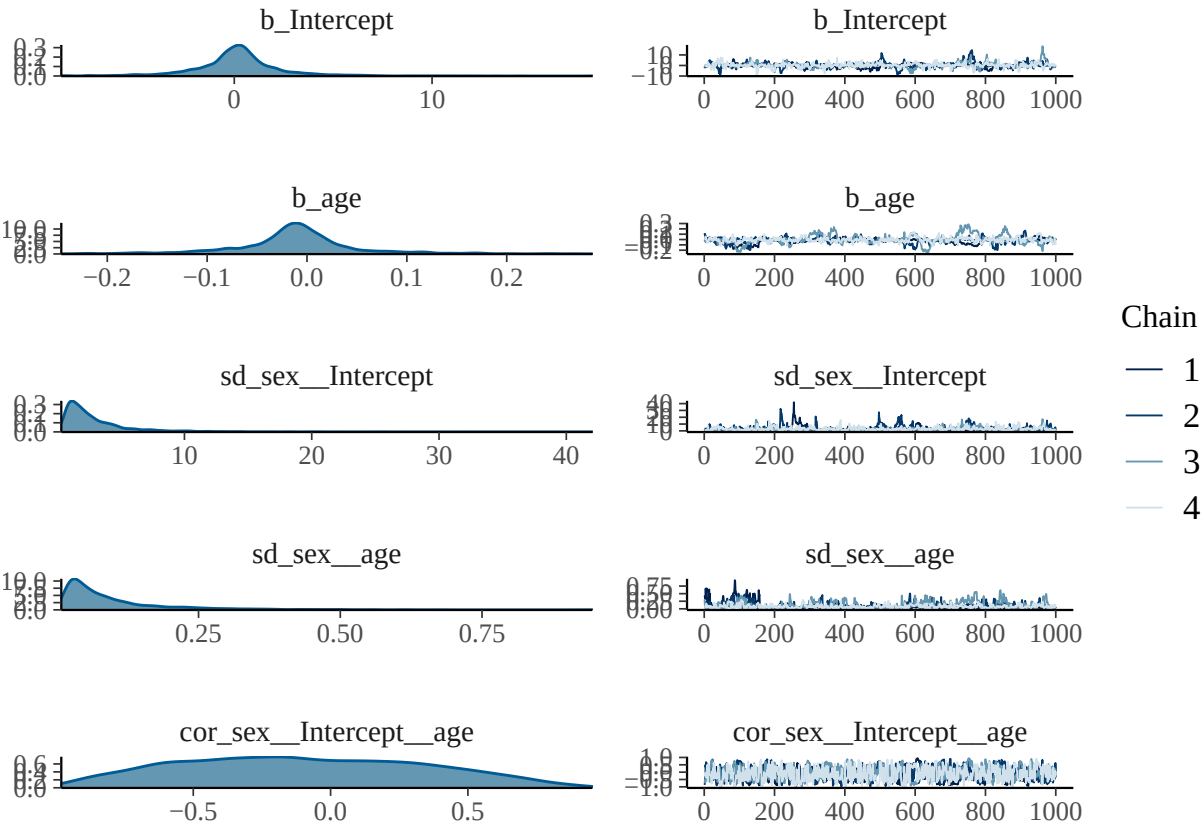
7.1 Weakness and next steps

Among all 81 variables in the 2017 GSS data, besides the age, gender and POB used in this work, other variables such as province of residence, education level, household income level and having children or not might also be closely related to the singleness in Canada. A more completed modeling approach might

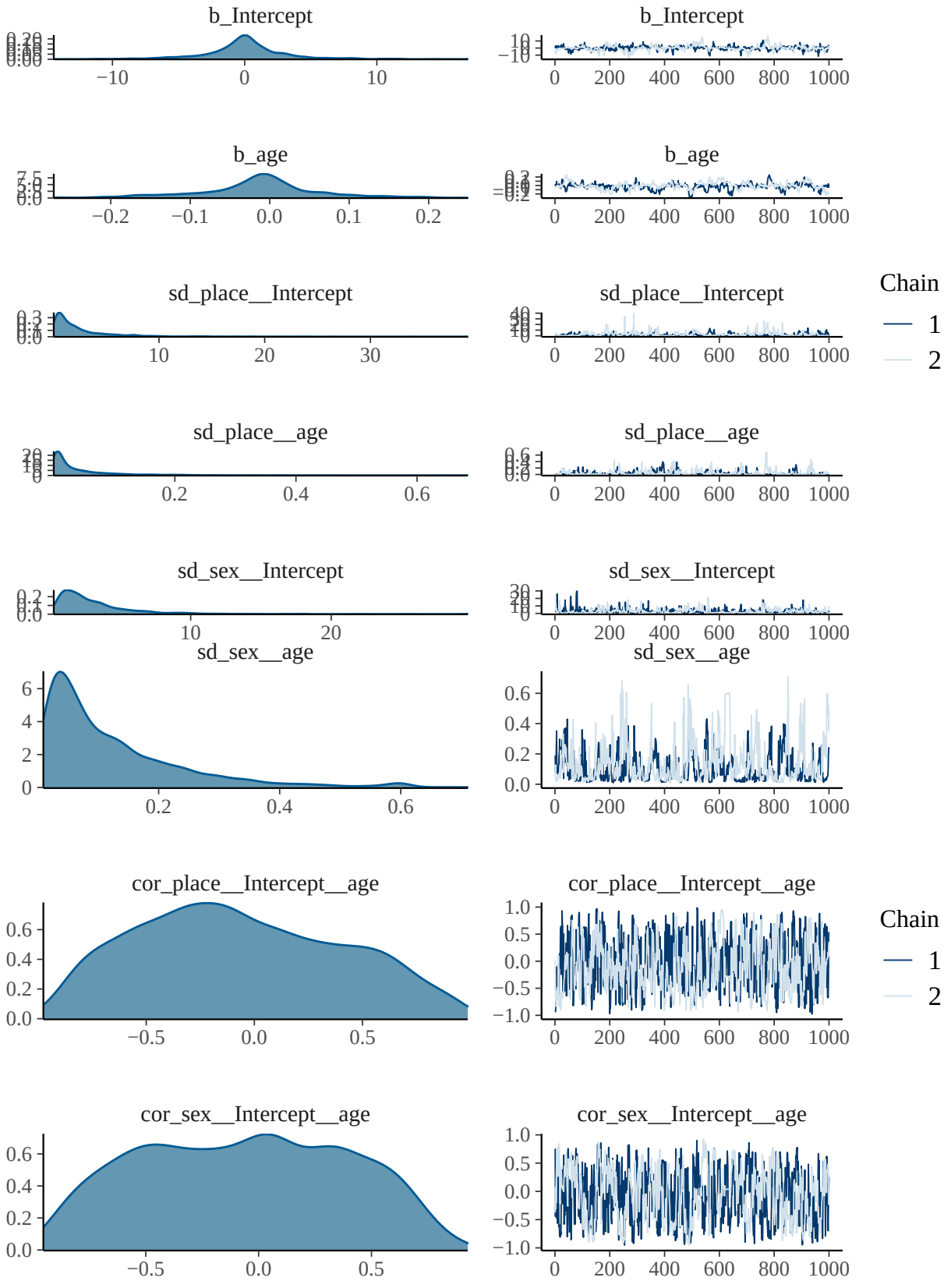
8 Appendix

8.1 Supplementary results

```
##  Family: bernoulli
##   Links: mu = logit
## Formula: factor(Marital) ~ age + (1 + age | sex)
##    Data: df (Number of observations: 20446)
## Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup samples = 4000
##
## Group-Level Effects:
## ~sex (Number of levels: 2)
##                     Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)           2.97      2.96     0.58    10.70 1.02      225      379
## sd(age)                 0.10      0.10     0.01     0.37 1.03      120      262
## cor(Intercept,age)     -0.10      0.44    -0.86     0.72 1.01      464      874
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     0.19      2.25    -4.54     5.20 1.01      211      374
## age          -0.01      0.06    -0.15     0.13 1.03      119      105
##
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```




```
##  Family: bernoulli
##   Links: mu = logit
## Formula: factor(Marital) ~ age + (1 + age | sex) + (1 + age | place)
##    Data: df (Number of observations: 20446)
## Samples: 2 chains, each with iter = 1500; warmup = 500; thin = 1;
##          total post-warmup samples = 2000
##
## Group-Level Effects:
## ~place (Number of levels: 2)
##                     Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)           2.51      3.30     0.13    11.20 1.01       92      173
## sd(age)                 0.05      0.07     0.00     0.23 1.01      163      324
## cor(Intercept,age)     -0.06      0.47    -0.84     0.84 1.01      377      561
##
## ~sex (Number of levels: 2)
##                     Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)           3.17      3.04     0.49    10.98 1.01      170      257
## sd(age)                 0.12      0.12     0.01     0.46 1.03      126      258
## cor(Intercept,age)     -0.06      0.46    -0.86     0.72 1.00      193      373
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     0.03      3.26    -6.89     7.62 1.01      125      158
## age          -0.01      0.08    -0.18     0.15 1.02       96      152
##
## Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```


b_Intercept
b_Intercept
b_age
b_age
sd_place__Intercept
sd_place__Intercept
sd_place__age
sd_place__age
sd_sex__Intercept
sd_sex__Intercept
sd_sex__age
sd_sex__age
cor_place__Intercept__age
cor_place__Intercept__age
cor_sex__Intercept__age
cor_sex__Intercept__age
Chain
1
2
Chain
1
2

8.2 Acknowledgement

The code used to clean the 2017 GSS dataset was from Dr. Rohan Alexander and Dr. Sam Caetano, please contact rohan.alexander@utoronto.ca for more information. The code was distributed under the MIT License.

8.3 References

Alathea, Letaw (2015). captioner: Numbers Figures and Creates Simple Captions. R package version 2.2.3. https://CRAN.R-project.org/package=captioner

Bürkner, Paul-Christian (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software, 80(1), 1-28. doi:10.18637/jss.v080.i01

Bürkner, Paul-Christian (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. The R Journal, 10(1), 395-411. doi:10.32614/RJ-2018-017

Firke, Sam (2020). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.0.1. https://CRAN.R-project.org/package=janitor

Gelman Andrew (2019). Model building and expansion for golf putting. https://mc-stan.org/users/documentation/case-studies/golf.html.

Gelman et al., (2020). Regression and Other Stories, Cambridge University Press, Ch 22.

Heberer Ray (2019).Bayesian Priors and Regularization Penalties. https://towardsdatascience.com/bayesian-priors-and-regularization-penalties-6d0054d9747b.

Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.1. https://CRAN.R-project.org/package=stargazer

Kur, A Solomon (2019). Doing Bayesian Data Analysis in brms and the tidyverse version 0.0.5. https://bookdown.org/ajkurz/DBDA_recoded/

Gabry J, Mahr T (2020). "bayesplot: Plotting for Bayesian Models." R package version 1.7.2, https://mc-stan.org/bayesplot.

Guo et al., (2020). RStan: R interface to Stan. https://mc-stan.org/rstan/.

R Core Team (2020). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Statistics Canada (2017). General Social Survey - Family (GSS). https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=335816

Statistics Canada (2019). Family matters: Being separated or divorced in Canada. https://www150.statcan.gc.ca/n1/pub/11-627-m/11-627-m2019033-eng.htm

Statistics Canada (2019). Family matters: Being separated or divorced and aged 55 or older. https://www150.statcan.gc.ca/n1/pub/11-627-m/11-627-m2019036-eng.htm

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Wickham, Hadley (2020). forcats: Tools for Working with Categorical Variables (Factors). R package version 0.5.0. https://CRAN.R-project.org/package=forcats

Xie, Yihui (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29.