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

Guannan Shen

October 18, 2020

Contents

1 Abstract

It has been widely reported that more Canadians are living 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, which including never married single, divorced, separated and widowed cases. Our work demonstrated that age, gender, the place of birth (POB), and the correlation between age and gender 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 probability of being alone. Particularly, the marital status was dichotomized into single and non-single sub-groups. The "single" group was made up of people belong to divorced cases, separated cases, never-married single cases and widowed cases. The main reason of combining different types of aloneness is to tackle the unbalanced group problem. In details, different single groups are minority groups, ranging from 3.1% to 22.9% of whole population. Combination of those groups together can increase sample size in the minority group, and thus increase the power of the statistical tests as well. As a result, the two levels of the outcome, single and non-single, have relatively close sample sizes.

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 mixed model (GLMM) 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. In general, models fitted by the GLMM approach were compared using Akaike information criterion (AIC), and a lower AIC value indicates a better fit. Models which had a better performance in GLMM approach were fitted through Bayesian approach using "Stan" on the back-end (Gelman et al. 2020). Eventually, all Bayesian models were compared using LOOIC (Leave One Out Information Criterion) and results from better models were reported.

This study demonstrated that age, gender and POB are significantly associated with the singleness rate. Particularly, men and elderly are more likely to be alone, and people born outside Canada are more likely to be alone as well. Especially, the effect of age is significantly modified by gender, which means elder women are more likely to have partners and men are likely to be alone as growing old. 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). The sampling frame is a list of address, associated with one or several telephone numbers. 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 questionnaire of this survey was designed based on research and consultations. And qualitative testing of the questionnaire carried out by Statistics Canada's Questionnaire Design Resource Center (QDRC) was conducted to evaluate its quality. A two-stage sampling was applied to identify household first, then identify one eligible person per household. The data were collected using the computer-assisted telephone interviewing (CATI) system. This is a large national survey, where approximately 43,000 questionnaires were sent and more than 20,000 of them were completed. The cost consists of survey software cost, survey design cost and respondents cost, and the cost is proportional to number of respondents, response rate and mean time taken to complete the survey. A survey at this scale may cost several million dollars (Civil Service UK, 2015).

Overall, this is a well-designed robust survey, with quality evaluation and error evaluation applied across all steps. However, this survey still has problems such as questions relating to income had rather low response rates. This indicates that the questionnaire should be improved on this aspect, such as incentives might be used to encourage people to answer income related questions. To deal with the non-response of income

related question, this survey use Tax information linkage and missing value imputation. Another problem is that the 2017 GSS only has samples from 10 provinces, and its results will be generalized to whole Canada.

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. Throughout this work, the outcome was called as singleness or marital status and represented as "Marital" in the model, where the "Single" level was treated as the base level and coded as 0. The gender variable was called as "sex" in the model with. Meanwhile, the POB was called as "place" in the model.

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 mixed model (GLMM) is typically faster to fit compared with the Bayesian approach. Thus, the GLMM approach can serve as the preliminary study of the Bayesian approach, making the model selection stage faster. This type of multilevel model allows us to model the age effect at the population level, as well as varied age effects at different gender or POB group levels.

The Bayesian modeling in this study was done by the R package "brms" (Bürkner 2017). In general, this package uses "Stan" on the back-end. In general, weakly informative priors were used for Bayesian modeling. Particularly, normal distribution with zero mean and large standard deviation were used as priors for intercept and other coefficients, half Cauchy priors were used for standard deviations. Especially, LKJ-Correlation prior with parameter equals 2 was used for covariance as suggested in "brms" manuals (Bürkner 2017).

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. The GLMM models were listed and ordered as shown in **Table 3.** , and the Bayesian models were listed and ordered as shown in **Table 4.** . The convergence of Bayesian models was measured by Rhat. When Rhat equals one, it means good convergence, and Rhat should not greater than 1.1.

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

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

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

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

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

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

where the a is the intercept, b representing coefficients and the correlation (clustering) is represented by variables multiplication.

And the priors used in the Bayesian models are:

$$s.d. \sim HalfCauchy(0, 5) \tag{7}$$

$$a \sim Normal(0, 10) \tag{8}$$

$$b \sim Normal(0, 10) \tag{9}$$

$$cov. \sim LKJ(2) \tag{10}$$

where s.d. is the standard deviation and cov. is the correlation (covariance).

5 Methods

Code supporting this analysis is available at: https://github.com/Guannan-Shen/pathToMLJob/tree/master/Developer_Basics/R. 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. The Bayesian models were fitted using "brms" R-package, version 2.14.0 (Bürkner 2017).

With the full 2017 GSS data, the final dataset used in the model fitting was generated by selecting only age, gender, POB and marital variables. Then null values and non-informative values such as "Don't Know" were removed. Finally, the marital variable was collapsed into a binary variable, either single or non-single.

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 were cited in the **References**.

6 Results

The characteristics summary of the subset of GSS dataset used in this work was shown in **Table. 1** and **Table. 2**. The **Table. 1** demonstrated that the marital status was divided into several sub-groups such as divorced, living common-law, married, separated, never married single and widowed. However, different single groups are minority groups, ranging from 3.1% to 22.9% of whole population. Thus, those groups were combined together to tackle unbalanced group issued. As shown in **Table. 2**, by levels combination, the single and non-single subgroups have roughly close sample size. Eventually, 20446 samples were used for model fitting.

Table. 1: Characteristics Summary of the 2017 GSS Data

	Overall (N=20499)
Age	
Mean (SD)	52.199 (17.748)
Range	15.000 - 80.000

	Overall (N=20499)
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%)
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) (Large table shown at the last.)

Models shown in the **Model** section with equation (2) – (6) were fitted using the GLMM approach. The results were summarized in **Table. 3** and were listed as GLMM model 1 – model 5. According to the AIC, model 5 has the lowest AIC, 27,469.34 and model 4 has the second lowest AIC, 27,836.15. The differences of AICs are larger than 2. By the empirical rule of AIC, the model 5 has the significantly best fit, the model 4 has the second significantly best fit. The model 1 – model 3 are simple univariate regression. The model 5 is a hierarchical logistic regression. Thus, the GLMM results demonstrated that there was significant clustered correlation effects in the dataset, and a Bayesian hierarchical logistic regression is necessary.

Negative coefficients in the logistic regression means that as the value of the predictor increase, the non-single (coded as 1 in the model) is less likely than the single group (coded as 0 in the model). Thus, from the model 5, people born outside Canada is significantly more likely to live alone with coefficient -0.374 ($p < 0.01$). Interestingly, if you are female and born in Canada, as you growing old, you are significantly less likely to live alone with coefficient 0.006 ($p < 0.01$).

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

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

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: Age Effects Differ in Gender Group: Bayesian Hierarchical Model 3

Group	Coef.	Estimates	Lower 95% CI	Upper 95% CI
Female	age	0.006879	0.004745	0.008978
Female	Intercept	-0.667685	-5.695875	4.096554
Male	age	-0.024183	-0.026633	-0.021682
Male	Intercept	0.648814	-4.404998	5.400791

Table. 6: Standard Deviation Test: Bayesian Hierarchical Model 4

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

Similarly, Bayesian models 1 – 4 were shown in **Table. 4**, where model 1 – 2 did not have random effects (cluster) and model 3 – 4 had multilevel effects. As a result, model 3 and model 4 has lower LOOIC, 27469.6 and 27506.4 respectively, indicating the hierarchical models had better fits. According to Bayesian model 3, hypothesis estimation was made and shown in **Table. 5**. It demonstrated that as age increasing, women are significantly more likely to be non-single (0.006879, $p < 0.05$) and men are significantly more likely to be single (-0.024183, $p < 0.05$).

Bayesian approach can make hypothesis estimation which can not easily be conducted with GLMM approach. As shown in **Table. 6**, Bayesian approach can test the difference of standard deviation of coefficients. It shown that, in model 4, the standard deviations of the intercepts in both gender and POB group are significantly greater than the standard deviations of the coefficients of the age factor in those groups.

More summary information and diagnostic plots of Bayesian approach can be found in **Appendix**. In general, all Bayesian models have good convergence with Rhat approximately 1, and as model becoming more complex, the convergence getting worse.

7 Discussion

7.1 Choices and results

In our work, hierarchical logistic regression approach was used to test the associations between singleness and variables in 2017 GSS data, which is a type of linear mixed model. When it comes to regression, a multilevel model pools information across cluster, allowing variations and correlations among clusters. Particularly, it allows the effects of the same factor to vary across subgroups, such as the effects of age. Thus, a hierarchical model has better estimations for clustered samples and imbalanced subsamples.

In linear mixed (hierarchical) model, the intraclass correlation coefficient (ICC), which is the ratio of the between-cluster variance to the total variance, is important in determination of whether a linear mixed model is necessary. However, in the context of binomial distribution, this is no variance parameter. Thus, there is

no direct estimation of residuals at the population level. Therefore, the LOOIC was used for model selection. As suggested by "brms" manual, LOOIC is more accurate than another criterion WAIC (Bürkner 2017).

To tackle the unbalanced data issue in the 2017 GSS dataset, we combined different "single" sub-groups together. Thus, the model fit was simplified. As a result, the correlations in our data may become more complex, due to the singleness is caused by different reasons. Moreover, the interpretation of the results generated from the model can be very complex. A simple but interesting result is that as a woman born in Canada getting older, she is less likely to live alone. It may sound counter-intuitive and this result was not widely reported.

7.2 Weakness and next steps

In future works, the divorce/separate rate and never married single rate might be estimated separately. A undersampling or oversampling approach can be applied together with a hierarchical model to fit the data.

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 take more factors into considerations.

In our preliminary study, the age factor may have significant polynomial coefficients in GLMM models, as well as its correlations with gender and POB. In next steps, better computing power can be used to model the polynomial effects with Bayesian approach.

8 Appendix

8.1 Supplementary results

The examples of posterior distributions and trace plots from the Bayesian model 3.

```
##  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).
```

The examples of posterior distributions and trace plots from the Bayesian model 4.

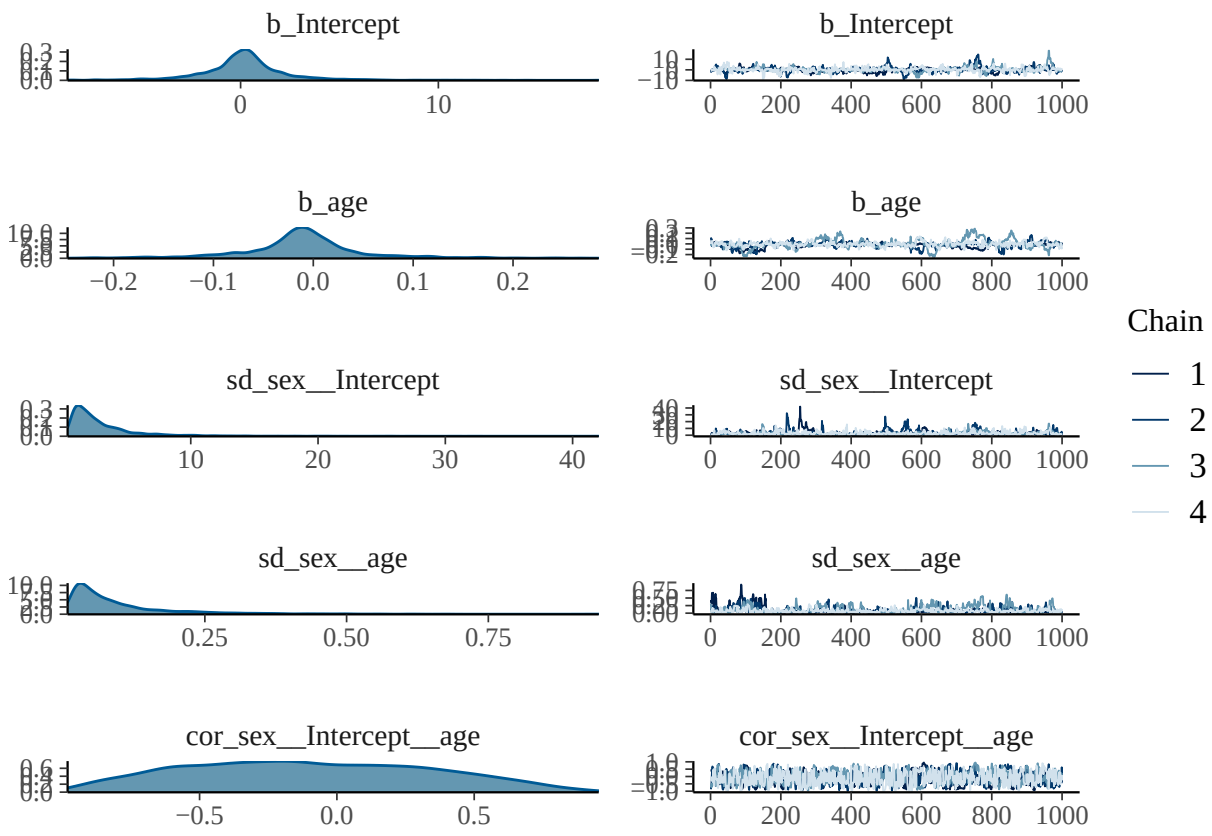
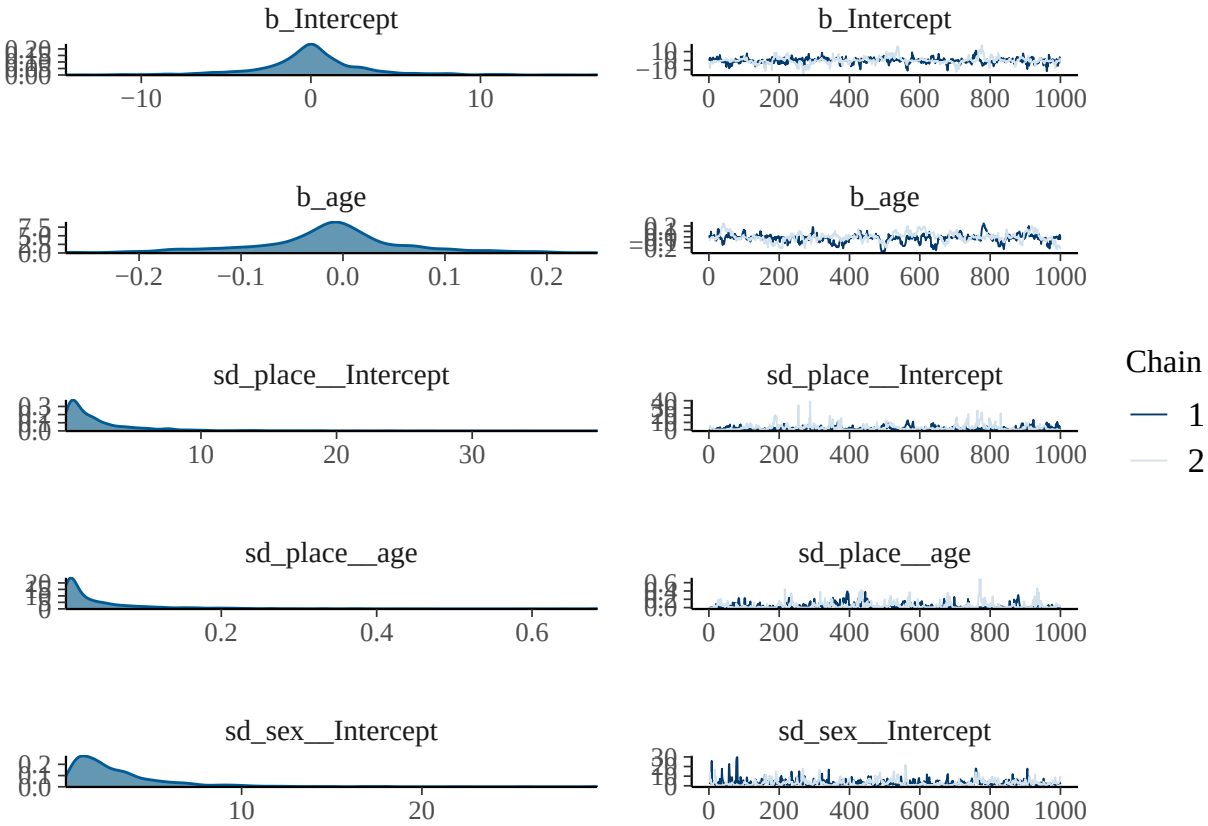


Figure 1: The Posterior Distribution and Trace Plots: Bayesian Model3

```
##  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).
```



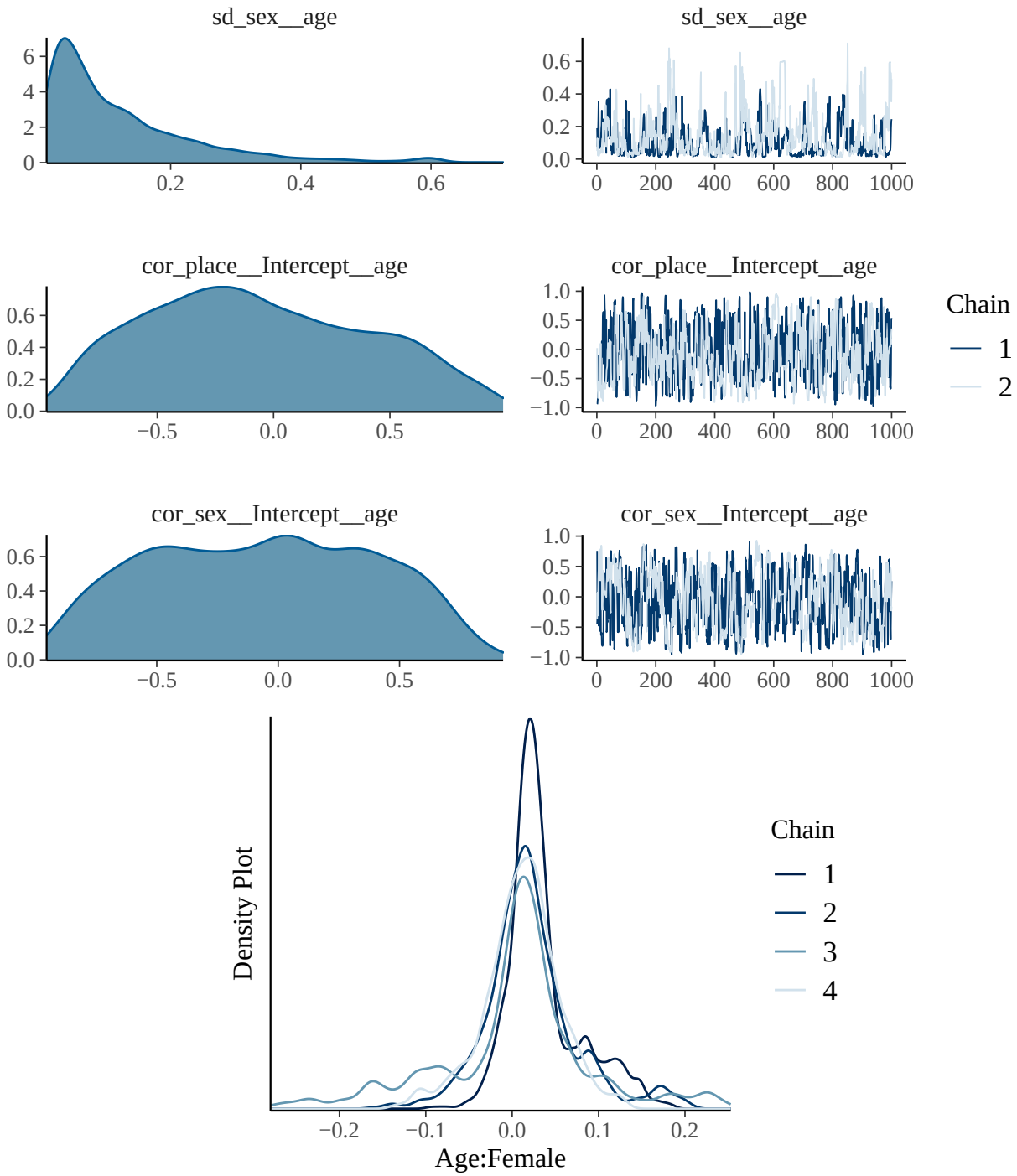


Figure 2: The Posterior Distribution of Age:Female Group Level (Bayesian Model3)

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.

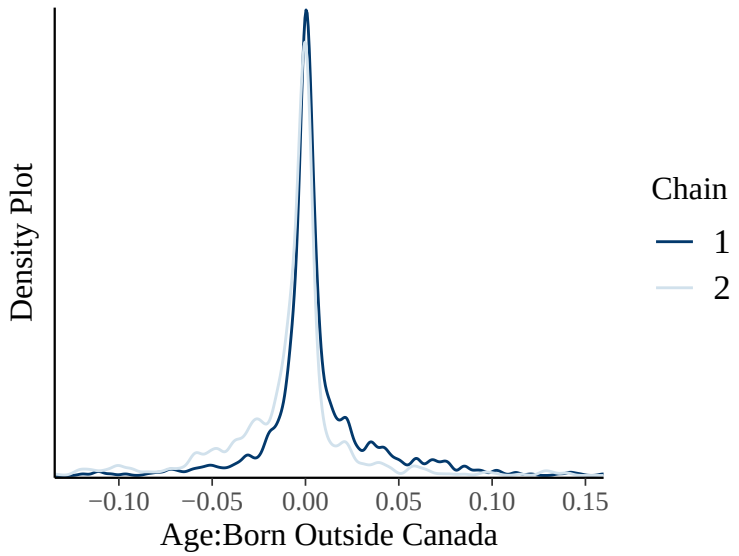


Figure 3: The Posterior Distribution of Age:Born Outside Canada Group Level (Bayesian Model4)

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

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

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

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

Civil Service UK (2015). Guidance on Calculating Compliance Costs. https://gss.civilservice.gov.uk/wp-content/uploads/2015/12/Guidance-on-Calculating-Compliance-Costs.pdf

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

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/

Kay M (2020). *tidybayes: Tidy Data and Geoms for Bayesian Models.* doi: 10.5281/zenodo.1308151 (URL: https://doi.org/10.5281/zenodo.1308151), R package version 2.1.1

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/.

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

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.