Regression Analysis on the Age of Having First Child

Cheng Chen, Jiachen Lu, Yingyu Li

2020-10-18

The age of a person at first birth has great impact on the society level, therefore it is interesting to analyze the factors effects on the age at first birth. This study fit a Bayesian Regression Model to the Canadian Social Survey data on family in 2017 and presents the associated results and analysis to have a deeper understanding of both the Bayesian model and the underlying decisional features and pattern for the age of a person at first birth in Canada.

Introduction

The age of having first child is an interesting and important topic, which brings impact on both social profits and company policy and gender equality (Miles *et al* (2011)). People often expect their first child depends on their statues in life, working situations and midsetting. Therefore it is interesting to investigate the influential factors for people to have their first kids and ideally building an mathematical model to analysis the current patterns as well as predict future trend.

This work focus on using regression models to analysis the data obtained from the Canadian General Social Survey (GSS) on family. The data set is cleaned up and interested features will be abstract for the study. A regression model is explained and applied to the pre-processed data. The results is discussed with emphasize on the performance of the model as well as the selected features.

Data

The data is obtained from the Canadian General Social Survey and presents the survey from the year 2017. The data was collected from February 2 to November 30, 2017. The data was collected through computer assisted telephone interviews (CATI). A simple random sampling is applied. The overall response rate of the survey was 52.4%. The key features it contains covers basic information, origins, conjugal history, children situation, labour market, education and more. The strong strength of this data set certainly lies on the fact that it is very detailed and has the whole coverage of the whole country. Yet it might be hard to utilize for certain investigations as it might be hard to select the key factor features. On the other hand, there is missing values as N/A in this data set, therefore after cleaning the data size shrinks.

To prepare the data for usage, the clean-up is firstly applied on the data set. The preprocessing of the data is twofold. Firstly the potential contributing features are selected, it includes age,age_at_fist_birth, education, income_respondent. Secondly, all the rows contains NA are removed. The preprocessing results in data with 20602 items with 4 features, including the target variable of the age at first birth.

Model

We are interested in explaining the whether the age at first at birth depends on sex, nationality, family income, education, and work situation. The modeling approach uses the Bayesian Linear Regression for two reasons. Firstly, the Bayesian model is promising as the uncertainty of the model is expected to perform well following a probability distribution (Wang, S. (2017)). As we do not have a concrete mathematical model to describe the formula, we approach it with a simple linear model with a slightly adjustment on the factor of age. Research have shown that women are having their first child at an older age now than in history (Eurostat, 2019), therefore we expect a reciprocal pattern in between age and age at giving first birth. We use 1/age in the linear model. The model can be described as follow:

$$yy \sim NN(\beta\beta^{TT}XX, \delta\delta^{2}II)$$

where the yy is the target, as the age at first birth in our data. XX are the selected features, $\beta\beta$ are the regression coefficients. The output follows a normal distribution with mean as $\beta\beta^{TT}XX$ and standards deviation as $\delta\delta$ a. To find the fitting values of the model parameters, the posterior distribution for the model parameters are determined by the Bayes rule, i.e. the Prior multiply by likelihood divided by normalization. This is described by the mathematic formula as:

$$PP(\beta\beta|yy, XX) = \frac{PP(yy|\beta\beta, XX) * PP(\beta\beta|XX)}{PP(yy|XX)}$$

In this work, we use the brms package in R to fit the bayes linear regression model. The code can be found at: https://github.com/corallyy/304.git

Results

The fitting results of the Bayes linear regression model is presented as follow:

```
> summary(brm1)
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: age_at_first_birth ~ (1 | age) + education + income_respondent
   Data: feature (Number of observations: 12533)
Samples: 4 chains, each with iter = 3000; warmup = 1500; thin = 1;
         total post-warmup samples = 6000

Group-Level Effects:
~age (Number of levels: 584)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.98      0.07     0.84     1.13 1.00     2084     3240
```

Population-Level Effects:

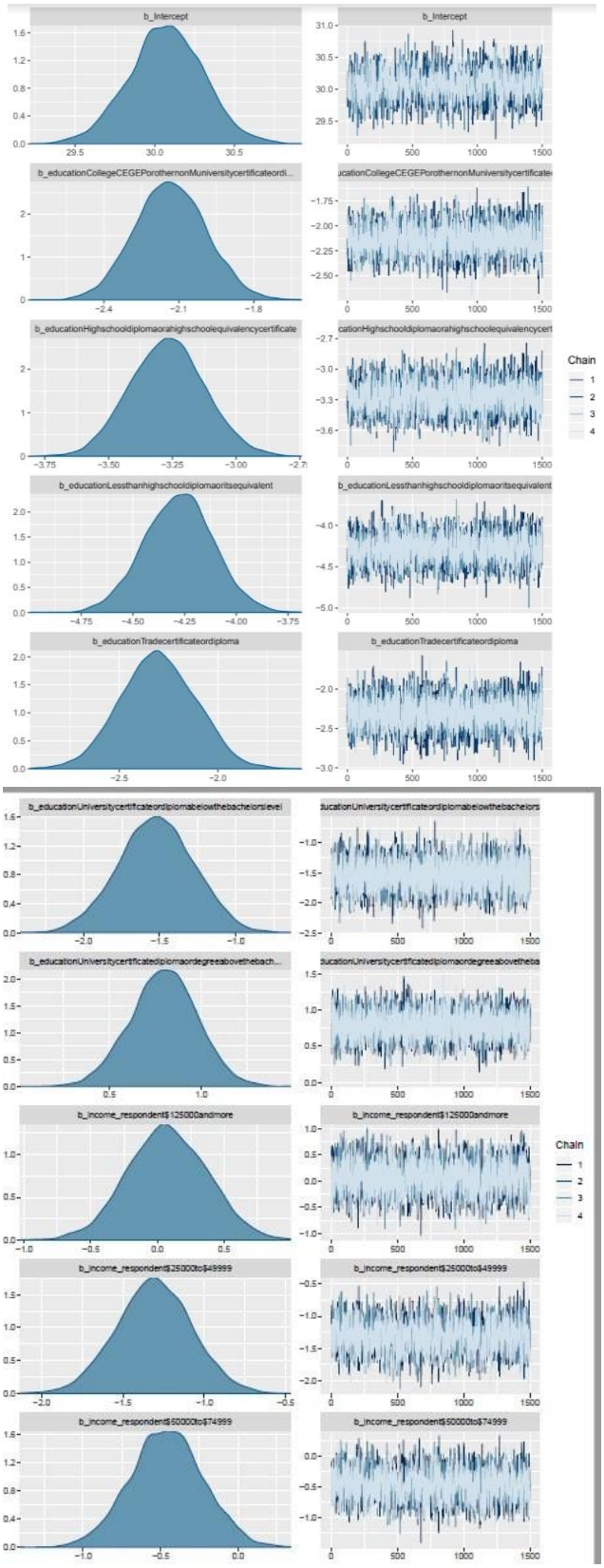
	Estimate	Est.Error	l-95% CI	u-95% CI
Intercept	30.07	0.23	29.61	30.53
educationCollegeCEGEPorothernonMuniversitycertificateordi...	-2.13	0.14	-2.41	-1.85
educationHighschooldiplomaorahighschoolequivalencycertificate	-3.27	0.15	-3.55	-2.98
educationLessthanhighschooldiplomaoritsequivalent	-4.28	0.17	-4.61	-3.95
educationTradecertificateordiploma	-2.30	0.19	-2.67	-1.93
educationUniversitycertificateordiplomabelowthebachelorslevel	-1.52	0.25	-2.01	-1.05
educationUniversitycertificatediplomaordegreeabovethebach...	0.80	0.18	0.44	1.15
income_respondent$125000andmore	0.08	0.30	-0.50	0.66
income_respondent$25000to$49999	-1.31	0.23	-1.76	-0.86
income_respondent$50000to$74999	-0.46	0.24	-0.93	0.01
income_respondent$75000to$99999	-0.31	0.26	-0.80	0.19
income_respondentLessthan$25000	-1.89	0.23	-2.34	-1.43

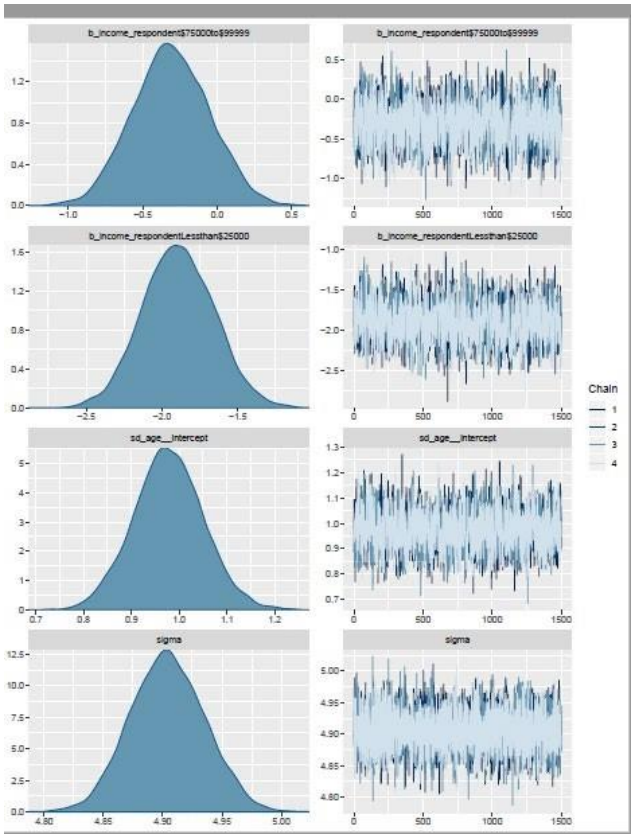
```
                                                                    Rhat Bulk_ESS
Intercept                                                           1.00     1905
educationCollegeCEGEPorothernonMuniversitycertificateordi...        1.00     3494
educationHighschooldiplomaorahighschoolequivalencycertificate       1.00     3425
educationLessthanhighschooldiplomaoritsequivalent                   1.00     3688
educationTradecertificateordiploma                                  1.00     4484
educationUniversitycertificateordiplomabelowthebachelorslevel       1.00     5529
educationUniversitycertificatediplomaordegreeabovethebach...        1.00     4759
income_respondent$125000andmore                                     1.00     2358
income_respondent$25000to$49999                                     1.00     1961
income_respondent$50000to$74999                                     1.00     2021
income_respondent$75000to$99999                                     1.00     2124
income_respondentLessthan$25000                                     1.00     2006
                                                                    Tail_ESS
Intercept                                                               2851
educationCollegeCEGEPorothernonMuniversitycertificateordi...            3965
educationHighschooldiplomaorahighschoolequivalencycertificate           4095
educationLessthanhighschooldiplomaoritsequivalent                       4342
educationTradecertificateordiploma                                      4327
educationUniversitycertificateordiplomabelowthebachelorslevel           4112
educationUniversitycertificatediplomaordegreeabovethebach...            4560
income_respondent$125000andmore                                         3821
income_respondent$25000to$49999                                         3525
income_respondent$50000to$74999                                         3187
income_respondent$75000to$99999                                         3570
income_respondentLessthan$25000                                         3146

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     4.90      0.03     4.84     4.96 1.00     8678     3895
```

Figure: Bayes Regression results.

The plot shows the convergence:


b_Intercept
b_Intercept
b_educationCollegeCEGEPorothernonMuniversitycertificateordi...
b_educationHighschooldiplomaorahighschoolequivalencycertificate
b_educationLessthanhighschooldiplomaoritsequivalent
b_educationTradecertificateordiploma
b_educationUniversitycertificateordiplomabelowthebachelorslevel
b_educationUniversitycertificatediplomaordegreeabovethebach...
b_income_respondent$125000andmore
b_income_respondent$25000to$49999
b_income_respondent$50000to$74999
Chain
1
2
3
4



Figures: Plots of the BRM fitting results.

Discussion

The results give the fitting model of the Bayes regression model. The plot shows the normal distribution of the parameters and the convergence of the model. Taking a close look at the Population-level Effects, we could see the estimated mean of the each parameters, We could see that the education level has a negative relation with the age of first birth and it has the maximal negative contribution with the education less than high school. It implies that within this regression model, the less educated, the early people will give their first birth. On the other hand, the income of the respondent people has a positive encouraging of late birth when the salary is highest, followed by negative patterns by the income lower than this level. Interestingly, the lowest income is early in the birth age. Relatively, the middle-income level has the latest age of giving first birth. The estimation error is within the acceptable level. Yet there are still many reasons for the big room of the estimation model. First of all, although Bayes model seem quite capable, it still might not be the best fit for the unknown underlying logic problems. Many features have a resporithal relation with the target features. For example, it is hard to say whether people who has higher income will have a late age of giving first birth, or the delayed birth contributes to their financial advances. Still, we could know it is positively related within each other though. There are many more features not selected which will have equally big

influence of the target variable. For example, as the statues of the partners are equally important as they should contribute to a similar level, However, this is hard to abstract from the dataset as it is not very clear whether the current partners stay the same as the partner at the birth of first children, therefore this feature has been dropped out. Lastly, tunning the model will lead to different results, whether it follows gaussian distribution could be debatable and the priors should be more carefully set with feature selected techniques, for example, histograms could be used to gain certain information of priors.

References

References Mills, M., Rindfuss, R. R., McDonald, P., & Te Velde, E. (2011). Why do people postpone parenthood? Reasons and social policy incentives. *Human reproduction update*, 17(6), 848-860.

Wang, S., Sun, X., & Lall, U. (2017). A hierarchical Bayesian regression model for predicting summer residential electricity demand across the USA. *Energy*, 140, 601-611.

Eurostat(2019). Women are having their first child at an older age. Eurostat. Retrived from:https://ec.europa.eu/eurostat/web/products-eurostat-news/-/DDN-20190318-1