Potential Reasons for Decreasing Fertility Identified from GSS 2017 Family Survey

Yuting Ge, Yifu Guo, Ziqin Zhou, Ying Cao

2020/10/18

Code and data supporting this analysis is available at: https://github.com/ziqin10086/STA304-ProblemSet-2-from-Group-90

Abstract

The fertility rate has been gradually dropping these years ever since 2008. Such decreasing in fertility rate has been observed throughout the world and greatly triggered socialist interests. Therefore, we want to know more about people's intention on giving birth/being father to come to some effective policies in stimulating fertility rate. This answer can be found in the survey, Public Use Microdata File (PUMF) of the 2017 General Social Survey (GSS) on the Family. The survey covers a quite large topic in FAMILY field, from the conjugal history to current spouse information, and also records respondents' personal information that might be helpful to analyze these topics. Fortunately, the information of respondents' dream children number (the existing plus future intention) is included in these topics. In this way, we can fit the processed data into a linear regression model. This model may help us find that people with older age and live in low-population density area would like to have children most and people with high level education would not like to have children most. Then, policies aiming at stimulating fertility rate can be investigated from this direction.

Introduction

Canada, as one of the most powerful developed countries in the world, cannot avoid the aging society social problem which puzzles most countries in the world. This problem is generally thought to be caused by the low fertility rate. The more serious problem is that, the fertility rate is still going down. In 2017, 1496 births were given per 1000 women compared with 1681 births in 2008 (CanadaStatistics, 2019). However, in the common sense, 2000 births per 1000 women can settle for keeping the society in a balance state. Therefore, some policies should be made as soon as possible to stimulate the fertility – at least, we cannot afford it to keep going down.

To make policies appropriately stimulate the fertility rate in the positive direction, we need to first figure out the important factors that push or impede people from giving birth or fathering. To answer this question, we utilize the survey conducted in 2017, Public Use Microdata File (PUMF) of the 2017 General Social Survey (GSS) on the Family (GSS, 2019). This survey, as its name indicating, focusing on the family problems, which also includes people's dream children number. This type of questions, are defined as "core questions" in the survey. There are also another set of questions called "classification question", such as age, education, income, etc., to help analyze the core questions. We will fit variables derived from the classification questions and the dream children number into a linear regression model to find out what factors influence the dream children number most within the given factor range.

This report is organized as follows. Section 2 will first introduce the survey data we utilize, including the survey itself and the data pre-processing steps. Section 3 will introduce the method we adopt to analyze

the survey data. Section 4 will present the results of the model we fitted. After that, section 5 will discuss results presented in section 5. In the end, section 6 & 7 will talk about the weakness and the next step for this case in respective.

Data

Public Use Microdata File (PUMF) of the 2017 General Social Survey (GSS) on the Family, is a typical survey conducted by General Social Survey (GSS), an organization designing and implementing social surveys through telephones across ten provinces in Canada ever since 1985. It aims at monitoring the living conditions as well as the well-beings of Canadians, and providing information for some social issues (GSS, 2019). For this specific 2017 family survey, GSS worked on it between Feb. 2nd and Nov. 30th, 2017 via a 2013 frame, which combines landlines and cellulars with Statistics Canada's Address Register, which are also the sample frames. The frame here means the method for the sampling implementation. Under this sample frame, the survey's target population covers all persons over 15 years old in Canada except for residents of the Yukon, Northwest Territories, and Nunavut, and full-time residents of institutions. This "target population" can be understood as what the sample is trying to represent. Stratified sampling is adopted in process, which means the sampling is based on the sub-population which are partitioned from the main population. The strata are based on the geographic area. The major advantages for this method is can best represent the whole population from the specific strata view. However, it cannot be applied to every data set. For example, if there exists some overlapping issues on the possible strata, we cannot apply stratified sample method to the data.

We utilize this data set is because: 1) its topic covers what we want to know, thus the variables it contains should also be highly related with our interests; 2) the question we want to investigate is of highly time-based, and this survey data is the newest data we can find. However, there are still some limitations of this data: 1) though it is the newest one, it still has been 3 years ever since the survey was conducted, and three years can change a lot of things due to the rapid development of technology; 2) this survey covers a rather large topic – family, and the fertility rate is just a small topic in this field, therefore, it does not contain specific questions directly related to the possible policy the government might make; 3) this survey is a one-time survey, which does not contain time series information that might be important to our question (e.g., will female be more willing to give birth when they getting oldder).

With these basic understanding of raw data in mind, the data pre-processing steps are presented as follows. Firstly, we clean the raw data which were directly downloaded from the website (GSS, 2019) via the provided Rcode (GSS, 2019)to transfer the raw data to the format the coding software, R, which we use for the further data analysis, can understand. Then, we manually select some variables which highly possibly influence the response variable and can be treated as "classification variables" in (GSS, 2019). They are 'province, 'marital_status', 'education', 'religion_has_affiliation', 'occupation', 'ever_married', 'age', 'total_children', 'region', 'income_family', 'sex', 'pop_center'

Model

We applied linear regression to analyzing the dataset. The logistic regression model is not performed for this data set since it focus on data contains mostly catagorical variables. To analyze the selected sample, we use the stratified sampling method under the linear regression since the response variable is a numerical variable instead of a categorical one. The strata here is the province instead of the specific strata defined in the manuscript because we cannot find the corresponding variable in the cleaned data set. Fortunately, the province itself can also be treated as a "large" strata as well. We use the finite population correction (fpc) to adjust the variance when the total population of interests are sampled. The value of finite population correction in our model is based on the population number presented on Wikipedia (Wikipedia, 2020)

We construct the Survey-weighted generalised linear model (svyglm) for analyzing since our survey data obtained stratified sampling method. Such model uses imformation from the survey design to correct variance estimates, which can be shown as advantage to our analysis. The alternative model is to apply only the

generalised linear regression model (glm). However, the difference between svyglm and glm are the treatment of weights. The weights in glm function simply adjust the weight given to the errors in the least squares estimation, therefore the standard errors are not correct and not suitable for our data and model analysis; while the survey weights are used to ensure that unbiased estimates of the finite survey population parameters are produced.

Thus, the survey-weighted linear model here can be written as:

$N = \beta_0 + \beta_1 Prov_B + \beta_2 ProvM + \beta_3 Prov_{NB} \cdots + \beta_{48} Pop_n$, where N represents the number of total intention children; β_0 is the intercept of the linear regression; $\beta_1 \ldots \beta_{48}$ are the coefficients for different input variables; and $Prov_B$ means the dummy variable of the province variable: if it equals to 1, represents the province is British Columbia, otherwise, it is not the British Columbia, if all dummy variables for province equals to zero, it means the province is the first level – the Alberta; and so does dummy variables for other inputs.

We use R studio to preform survey-weighted linear model. Since we are using a linear regression here with the stratified sampling, we are assuming that residuals are following normal distribution with zero mean and constant variance. This assumption will be tested later.

Results

Summary Tables - avaliable at github: (https://github.com/ziqin10086/STA304-ProblemSet-2-from-Group-90/blob/main/PS2-summary-tables.pdf)

From survey-weighted linear model fitting results, we can come to the final model to be:

$N = 2.818750 - 0.007525 Prov_B + 0.080638 Prov_M + -0.188980 Prov_{NB} \cdots + 0.255177 mar_{CL} \cdots + -0.061452 age + 0.039243 child_1 + -0.585303 child_2 \cdots - 0.653410 child_7 \cdots + 0.241521 Sex_m - 0.030570 Pop_N$

All the notations can be interpreted as discussed in the Model section. Since there are 48 variables, we cannot write down all of them. But basically, the β value for each variable is the corresponding estimate value showed in summary table. Noted that the linear regression processed pivots to dummy variables, for example, in province variable, our data set contains 10 provinces from 'Alberta' to 'Saskatchewan', while povits only have 9 dummy variables from β_1 to $\beta 9$. In this case 'Alberta' are consider to be the baseline references for province variables, similarly for all other variables.

The second summary table contains Residual standard error which measure of the quality of a linear regression fit; Multiple R-squared and Adjust R-squared show how well the model is fitting the actual data; and the p-value for the entire model indicates the significance of our model.

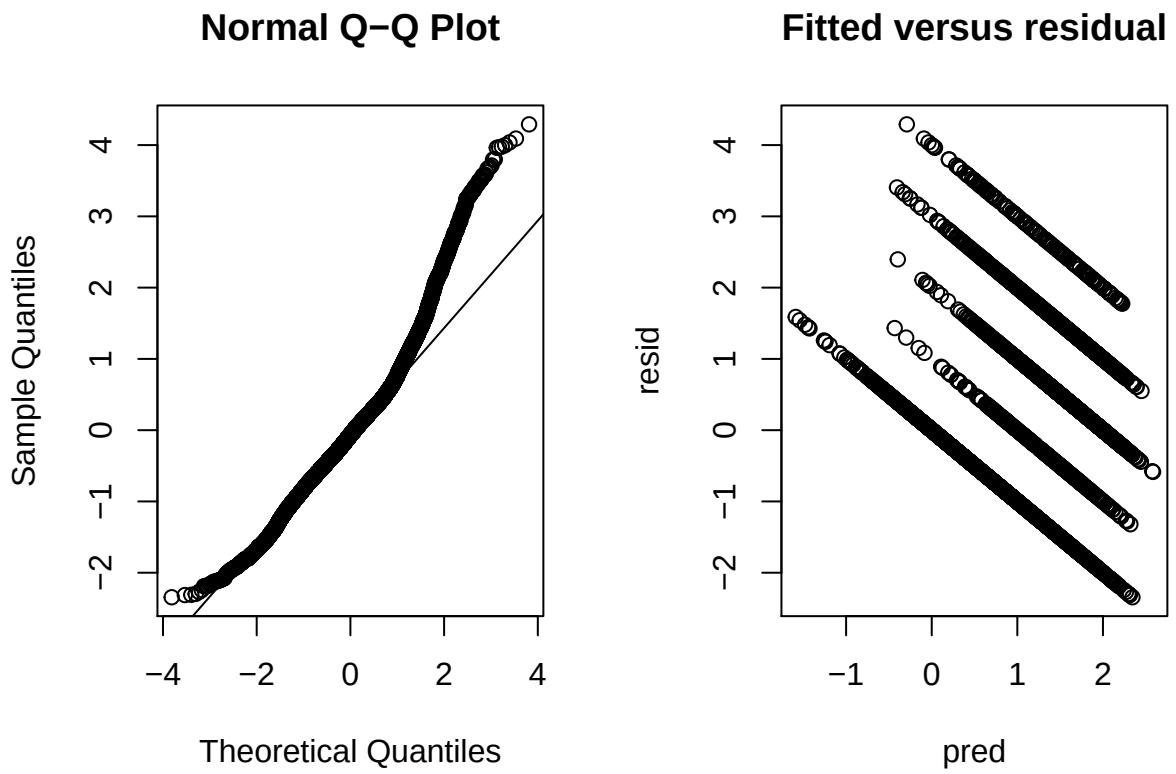

Figure 1. Diagnostics on the fitted model

We perform a diagnostics plots. The Normal Q-Q plot shows if residuals are normally distributed and the fitted versus residual plot is used for the observation of linear relationship between predictor variables and an outcome variable in our regression.

Discussion

From the Summary table and the fitted model, we can know following things: 1) the intercept value here does not have any practical meaning, but from the numerical points of view, when all numerical variables equal to zero (e.g., the age), and all categorical variables are set to be level 1, the number of intention babies equals to 2.818750; 2) for the coefficient of dummy variables, for example, when the province is changed from Brithish Columbia to Manitoba, the birth intention is increased by $\beta_2 - \beta_1 = 0.080638 + 0.007525 = 0.088163$, but if the province is changed from British Columbia to Alberta, the intention number is increased by $-\beta_1 = 0.007525$ since Alberta is considered to be the baseline reference of provinces ; 3) for numerical variables, like age, it means when other condition remains age is increased by one year, the intention number of babies is likely to decrease by 0.061452.

Noted that even though all of the occupation status have positive influences on having children, most of them have p-value greater than 0.05 which indicate those variables are unsignificant to our model. It may possibly become a weakness in our model and with no need to be included in analysis instead.

From the summary table of lm function and the fitted model, we can know following things: 1) the residual standard error is 0.9285 which means 92.85% variation can be explained in the model 2) the multiple R-squared is 0.3835 which states 38% of the variance found in the response variable can be explained by the predictor variable(intention of babies), also the adjust R-square is 0.3786 means when adjusts the number of variables, we have also 38% of the variances that can be explained by the predictors 3) the model have

p-value of $< 2.2e\text{-}16$, whcih is way smaller than 0.05, it tells us our model is significant and can be use to explain the y value's variation.

Besides these information, from the policy making point of view, we can find following observations as well: 1) people in Manitoba has the highest intention children number and people in New Brunswick has the lowest one; 2)all occupation status all have positive influences on having children, this may be weird and might caused by outliers or other issue 3) people with education level in bachelor or above shows the high interest in giving birth intention; 4)for family already have 1 child, they would like to give a another one, while for family have 2 or more childern, their intention are droped.

In the meantime, we can also get some information from the diagnostics plots. The residual does not follow the normal distribution due to the heavy head, but unfortunately, there exists strong pattern in the fitted versus residual plots, which indicates the residual is violating the assumption.

Weaknesses

Basically, for the given data set, we have limitations on: 1) the residual assumption is violated; 2) the outliers are not kicked out, which may influence the model performance; 3)Some of the varibles have unsignificants from p-value evaluations, which should not be included in our linear model; 4) the sampled data points are limited, which may also bring unstableness for the fitted model. 5)The absolute size of the sampling error is less important than its relative size, so the standard error is not always the best measure of sampling error. The size of the sampling error is sometimes related to the size of the estimate. To better estimate the sampling variability, standard error and the coefficient of variation should be used to specific analysis.

Next Steps

For the next steps, if we assume the above results are valid, the first thing we should do is to comprehensively identify the possible directions that may influence people's intention on giving birth. Some experiments should be conducted again with detailed design of experiments like factorial design to identify the most significant factors. Besides, it can also update the data to make the policy fit better to the current situation.

References

CanadaStatistics. (2019). Births, 2017. Retrived on Oct. 16th, 2020, from https://www150.statcan.gc.ca/n1/daily-quotidien/180928/dq180928c-eng.htm.

GSS. (2019). Public Use Microdata File Documentation and User's Guide.

GSS. (2019). Data. Retrived on Oct. 15th, 2020 from https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/html/gss.htm

Wikipedia. (2020). Population of Canada by province and territory. Retrived on Oct. 17th, 2020 from https://en.wikipedia.org/wiki/Population_of_Canada_by_province_and_territory.