TBD

Arjun Dhatt, Benjamin Draskovic, Yiqu Ding, Gantavya Gupta

10/16/2020

Abstract

Introduction

Data

Model

We are interested in how variables such as age, family income, education level, and population center affect a woman's decision to have children. Due to the large sample size and the nature of GSS, the sample represents the population well. The model that we are using is logistic regression, it works well for our response variable, which is a categorical variable, and it incorporates both numerical and categorical explanatory variables.

Logistic regression estimates $\beta_0 ... \beta_k$ in the following equation:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

In our case, it estimates $\beta_{age}, \beta_{inc}, \beta_{edu}, \beta_{pop}$ in:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_{age} x_{age} + \beta_{inc} x_{inc} + \beta_{edu} x_{edu} + \beta_{pop} x_{pop}$$

We use `glm()` from `stats` package to fir the model to our data. We use `as.factor()` to incorporate dummy variables for all the categorical variables: family income, education level and the type of population center. For each categorical variable with n levels, we need n-1 dummy variables to fully study its influence on our response variable.

The dummy variables setting ups are stated in table x, y and z.

Table 1: Dummy Variable Coding Set Up for Income Levels

income.level	inc1	inc2	inc3	inc4	inc5	inc6
greater than $125,000	1	0	0	0	0	0
between $25,000 to $49,999	0	1	0	0	0	0
between $50,000 to $74,999	0	0	1	0	0	0
between $75,000 to $99,999	0	0	0	1	0	0
less than $25,000	0	0	0	0	1	0
between $100,000 tp $124,999	0	0	0	0	0	0

Table 2: Dummy Variable Coding Set Up for Education Levels

education.level	HS	graduate	under	college
high school or less education	1	0	0	0
University Graduate	0	1	0	0
University Undergraduate	0	0	1	0
college/trade	0	0	0	0

Table 3: Dummy Variable Coding Set Up for Population Center

population.center	PEI	rural	urban
Poplation centered at PEI	1	0	0
Rural areas and small population centres(non CMA/CA)	0	1	0
Larger urban population centres (CMA/CA)	0	0	0

Results

Table x summaries our model results:

Table 4: Summary of Losgistic Estimates

variable	estimate	pvalue
intercept	-1.513991	< 2e-16
age	0.060790	< 2e-16
income greater than $125,000	0.161424	0.0810
income between $25,000 to $49,999	-0.680099	8.24e-13
income between $50,000 to $74,999	-0.449230	3.49e-06
income between $75,000 to $99,999	-0.197017	0.0491
income less than $25,000	-0.872898	< 2e-16
high school or less education	-0.102521	0.0982
University Graduate	-0.779989	< 2e-16
University Undergraduate	-0.468140	9.25e-13
Poplation centered at PEI	0.276375	0.0409
Rural areas and small population centres(non CMA/CA)	0.526007	2.86e-14

In an equation, this means

$$log(\frac{\hat{p}}{1-\hat{p}}) = \hat{\beta}_0 + \hat{\beta}_{age}x_{age} + \hat{\beta}_{inc_1}x_{inc_1} + \hat{\beta}_{inc_2}x_{inc_2} + \hat{\beta}_{inc_3}x_{inc_3}$$

$$+\hat{\beta}_{inc_4}x_{inc_4} + \hat{\beta}_{inc_5}x_{inc_5} + \hat{\beta}_{HS}x_{HS} + \hat{\beta}_{graduate}x_{graduate} + \hat{\beta}_{under}x_{under}$$

$$+\hat{\beta}_{PEI}x_{PEI} + \hat{\beta}_{rural}x_{rural}$$

Notice that we have incorporated our categorical variables using dummy variables. “inc1”, “inc2”, “inc3”, “inc4”, “inc5” represent the six income categories; “HS”, “graduate” and “under” describe the four education levels; “PEI” and “rural” represent the three population center.

The model predicts the following result for the probability a woman has children p, rounded to three decimal places. Since log is an one to one function with p, we say the change on $log(\frac{\hat{p}}{1-\hat{p}})$ is isomorphic to any change on p, the probalbility that a woman in Canada has children.

$$log(\frac{\hat{p}}{1-\hat{p}}) = -1.51 + 0.061x_{age} + 0.161x_{inc_1} - 0.68x_{inc_2} - 0.449x_{inc_3}$$

$$-0.197x_{inc_4} - 0.873x_{inc_5} - 0.103x_{HS} - 0.78x_{graduate} - 0.468x_{under}$$

$$+0.276x_{PEI} + 0.526x_{rural}$$

The interpretation of the dummy variables' prediction results is comparing it to a certain level that is not in the equation above. A $\hat{\beta}_{HS} = -0.103$ does not mean the coefficient for a high school level education is -0.103. It represents the difference of influence (on childbirth decision) between a woman who has a college level of education and a high school level of education. -0.103 indicates that a woman who has a high school degree or below is less likely to have children than a woman who has a college degree if it is influential at all depending on the p-value.

Using $\alpha = 0.05, H_0 : \hat{\beta} = 0, \hat{\beta} \neq 0$, the p values indicate weak evidence that having a family income higher than 125,000 dollars/year affects a woman's probability of having children. At the same time, it is evident that other income levels do influence the probability. A p-value of $0.0982 > 0.05$ indicates weak evidence to reject H_0; therefore, we cannot say having high school or fewer education influences a woman's probability of having children.

There is evidence that both having a family income between 75,000 to 99,999 dollars per year and living in an area with a population center at PEI have influences on p because their p values are smaller but close to 0.05. There is strong evidence that the rest of the variables influences p.

All else the same, an older woman is more likely to have children than a young woman. A woman's family income affects her decision to have children when the income is below 125,000 dollars per year. Having a university graduate or undergraduate degree reduces a woman's probability of having children comparing to having a college degree while having high school or less education does not affect that decision. A woman living in PEI or rural areas and smaller population centers are more likely to have children than a woman living in urban areas and larger population centers.

Discussion

References

- https://mc-stan.org/rstanarm/articles/mrp.html
- https://www.monicaalexander.com/posts/2019-08-07-mrp/