Understanding How Household Income and Number of Children Correlate to Life Satisfaction

Generalized Linear Model Analysis of 2017 GSS Life Satisfaction data

David Wong, Iris Shao, Jingjing Zhan, Muhammad Tsany

October 19, 2020

1 Abstract

A family's feeling of life is heavily dependent on multiple variables. In this report, we developed a generalized linear regression model (GLM) in R to better understand the correlation between the predictor variables, family income and total number of children, and response variable, feelings of life. Our approach was to use this GLM, its estimates, and residuals and interpret them to create a conclusion. We found that there is strong correlation between the predictor variables and our response variable, which informs us that these predictor variables are informative of a family's overall feeling of life.

2 Introduction

Understanding life satisfaction has always been a perplexing subject. collecting and analyzing health data may prove to even be a more complex task. Notably, the study of the pursuit of happiness has been a concept that has been gaining traction in the scientific world (Positive Psychology, 2020). For Canadians, understanding how we can improve our life satisfaction will translate into helping our governing bodies better understand how to provide an infrastructure for Canadian life satisfaction. Studies have been conducted on Canadians' life satisfaction in relation to other relevant variables; we will further elaborate on our selected variables which have appeared within several scientific studies to improve our understanding of Canadian life satisfaction.

Given our GSS data, we apply a Generalized Linear Regression model (GLM) to understand the correlation between our chosen variables as well as attempt to draw upon larger conclusions from our data from our statistical expertise. Our hypothesis is that feelings of life (life satisfaction) is correlated with household income and total children in the household.

3 Data

3.1 Metadata We examine data from the Family cycle of Statistics Canada's General Social Survey (GSS), collected in 2017. The GSS is a long-running national survey established in 1985, with primary objectives to observe trends in the well-being of Canadians and to gain insight on social policy issues. According to the GSS User guide, the 2017 Family cycle concentrates on understanding the imperative role family plays in people's lives (CSS User Guide, 2017).

3.2 Survey and Sampling Technique For the 2017 Family cycle, the **target population** includes all persons aged 15 and over, living in all 10 provinces of Canada. The **sampling frame** combines a list of both landline and cellular telephone numbers in use available to Statistics Canada and a list of all dwellings contained in the Address Registrar. This combined list includes people who have telephone numbers and/or addresses, but excludes people who do not have telephone numbers, people who do not have access to a telephone at the time of contact, people who have missed or ignored calls from the interviewer, people who have changed numbers but did not update their respective government departments, and the homeless population, which alone comprises of at least 235,000 individuals in any given year (cite CAN encyclopedia). The **sampled population** (total number of respondents) is 20,602, reflecting a 52.4% response rate. Responses

are voluntary and collected by telephone interviews during which the interviewers enter the captured data into computers simultaneously.

Based on the metadata, there are adjustments made in the collection process to reduce sampling errors and to improve interpretability of the data. Imputations for the age field if the data was missing or if age is to be approximated to a whole number based off of survey timing. Personal income in the 2017 data wasn't self-reported but retrieved from a tax data linkage for respondents who didn't object from it. Lastly, the survey introduces a **stratified sampling method** which is classified based off of provincial-age-sex population distributions and are weighted according to their analysis of the population. Notable non-sampling errors in the survey is non-response (including partial non-response), in which weights for that stratum would be adjusted accordingly.

During the scientist's data collection process, some components of our chosen fields were impacted by the data decision they made. Notably, Age and Income (household and respondent) fields were not necessarily self-reported statistics but may have either been imputed or collected from other sources. Furthermore, only 91.8% of Statistics Canada's telephone repositories were considered valid households, meaning that they had a reachable telephone number (GSS User Guide, 2017); in addition, this doesn't factor in non-response which is a non-sampling error.

Sampling error can't be exactly determined from the data, however their confidence interval are as follows: "68 out of 100 that the difference between a sample estimate and the true population value would be less than one standard error, about 95 out of 100 that the difference would be less than two standard errors, and virtually certain that the differences would be less than three standard errors" (GSS User Guide, 2017).

Not all values in our selected fields are true values and some chosen fields have approximated values which were carefully determined via other trends and responses in the survey. This approximation would increase sampling errors. Regardless, the reported confidence interval seems to be relatively strong within at least 2 standard errors based off of the scientist's calculations.

3.3 Selected Variables

We extracted information on three concepts: 1) total number of children, 2) total family income, and 3) feelings about life as a whole (life satisfaction). To investigate how family size and family income affects life satisfaction, the predictor variables are total number of children and total family income, and the response variable is self-reported life satisfaction score. Details on each variable is as follows.

Our response variable, "life satisfaction" corresponds to the survey question: "using a scale of 0 to 10 where 0 means 'Very dissatisfied' and 10 means 'Very satisfied, how do you feel about your life as a whole right now?" Answer categories include integer values from 0 to 10, each presents a state of satisfaction, thus making this a discrete numerical variable. A strong national and global interest, this particular survey question on life satisfaction has been included in the GSS for over three decades, with fairly consistent wording and response categories, which allows for in-depth scrutiny on trends from year to year (Bonikowska et al, 2013).

The predictor variable, "total number of children" reported by respondent, is another discrete numerical variable, consisting of integer values ranging from 0 to 7, with 7 representing 7 children and more. Some related variables in the dataset are total number of children intending to have, number of children by age, number of children by union type, and number of children by current residence, all of which are disregarded as they are subsets of our chosen variable. In the raw data, around 30% families have no child, around 14% have 1 child, around 30% have 2 children, around 15% have 3 children, around 6% have 4 children, around 2% have 5 children, around 0.8% have 6 children and around 0.9% have 7 children.

Response data for our predictor variable "total family income" (and all income-related concepts) are automatically derived from the respondent's income tax file for 2016. All numerical income values are then organized into 6 ordered categories: "less than $25,000", "$25,000 to $49,999", "$50,000 to $74,999", "$75,000 to $99,999", "$100,000 to $124,999", and "$125,000 and more". This ordinal categorical variable represents the sum of all incomes (before taxes and inductions) of a family. A family is defined as a married couple, a common-law couple, or a lone parent of any marital status, with at least one child. For the family income predictor variable, we categorized each group using 1 to 6 for family income in an ascending order. In the raw data, around 13% families have a total income less than $25,000, around 21% have $25,000 to $49,999,

around 18% have $50,000 to $74,999, around 14% have $75,000 to %99,999, around 10% have $100,000 to $124,999 and 22% have more than $125,000.

4 Model

We used generalized linear regression in R statistical software using life satisfaction as the response variable and total number of children and family income as predictor variables. These predictor variables were chosen because we believe these variables are the most impactful to an individual's feeling of life.

The generalized linear regression is:

$$Y = \beta_0 + \beta_{child} x_{childi} + \beta_{incfam} x_{incfami} + \epsilon_i, \text{where } i = 1, 2, ...n$$

Our model converges according to our function svyglm() because the step-halving method, to determine if Iterated Reweighted Least Squares (IRLS) converges for the Generalized Linear model, is valid (Marschner, 2011). Precisely according to our GLM output, it took 2 steps to converge because the Fisher Scoring iterations = 2.

We chose not to implement Finite Population Correction (FPC) to our sample because it falls under 5% of the total population. Our total population was the population of all of Canada minus the territories (Stats Can. Census, 2016). The total population is 35,038,124 whereas our sample size was 20602; our sample is ~ 0.06% of the total population whereas the ratio threshold to apply FPC is for your sample population to be 5% or more of your total population. Furthermore, our FPC calculation is as follows:

$((35038124 - 20602)/(35038124 - 1))^{0.5} = 0.99970$

Since this value is very close to 1, we can see why FPC would have minimal effect on our sample statistics:

$$\sigma_{sample} = (\frac{\sigma_{sample}}{\sqrt{n}}) * \frac{\sqrt{(N-n)}}{(N-1)}$$

We can see that our normalizing FPC calculation, since it's close to 1, can simplify our sample statistics (in this case, sample error) to be approximately just the normal uncorrected sample standard error. Thus, FPC is not needed for our Generalized Linear Model (StatisticsHowTo, 2016). This is also further evidence for our model convergence, as the CLT normality assumption does roughly apply to our GLM terms. Our sample size relative to the population is minuscule and Finite Population Correction would have minimal impact on our outcomes.

Caveats for our Generalized Linear Regression model mainly stem from sampling and non-sampling errors. Since our calculated FPC is near 1, we can probably even fit a Multiple Regression Model because we can approximately say that the CLT does apply to our terms. With the assumptions made by the data scientists, it would result in more favorable error terms and a smoother fit for our data since some data points are not true data points that were surveyed from participants

5 Results

Table 1: Model coefficients

Coefficients	Estimate	Std. Error	T-Statistic	p-value
(Intercept)	7.356	0.032	233.391	0
total_children	0.092	0.008	11.879	0
income_family	0.164	0.007	24.661	0

Table 1 contains the GLR model coefficients of the sample population. Estimates of intercept predictor variables are statistically significant with p-values $<$ 2e-16. We conducted a T-test for the fitted model described above. The T-test can tell us whether the differences of life satisfaction just "happened by chance" or is statistically correlated with our explanatory variables (Glen, 2020). The standard error measures how

Table 2: Residual values of GLM fit

total_children	feelings_life	Residuals	Std. Residuals	Leverage	Cook's Distance
0	7.879	-0.126	-1.862	0.417	1.238
1	8.045	-0.011	-0.145	0.274	0.004
2	8.213	0.107	1.328	0.179	0.192
3	8.197	0.039	0.474	0.131	0.017
4	8.337	0.128	1.554	0.131	0.182
5	8.196	-0.064	-0.796	0.179	0.069
6	8.311	0.001	0.010	0.274	0.000
7	8.287	-0.074	-1.094	0.417	0.427
income_family	feelings_life	Residuals	Std. Residuals	Leverage	Cook's Distance
1	7.479	-0.160	-2.124	0.524	2.481
2	7.902	0.086	0.938	0.295	0.184
3	8.112	0.120	1.209	0.181	0.162
4	8.210	0.041	0.416	0.181	0.019
5	8.361	0.015	0.168	0.295	0.006
6	8.420	-0.102	-1.354	0.524	1.008

precise our estimates are. If it is small compared to the coefficients, then our estimates are close to the true value (Princeton University Library, 2007).

By definition, the "t statistic is the coefficient estimate divided by its standard error" (Princeton University Library, 2007) and "p-value is the probability that the result from your sample data occurred by chance" (Glen, 2020). So when we have a small p-value, usually smaller than 0.05, we can say that the result is "statistically significant", which means that our estimates are close to the true population value (Glen, 2020). As such, the intercept β_0 has an estimate of 7.356, standard error of 0.032, test-statistic of 233.391. The p-value is less than 2e-16, which is statistically significant. The coefficient total_children β_{child} has an estimate of 0.092, standard error of 0.008, t-statistics of 11.879. The p-value is less than 2e-16, which is statistically significant. The coefficient income_family β_{incfam} has an estimate of 0.164, standard error of 0.007, t-statistics of 24.661. The p-value is less than 2e-16, which is statistically significant.

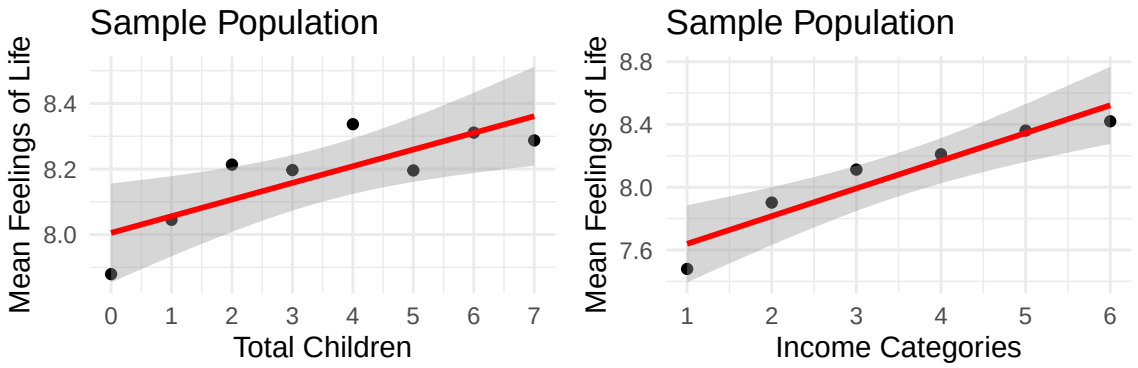

Figure 1: Best fit line using each subgroups

In **Figure 1**, the left graph depicts a scatter plot for each subgroup in total children and average life satisfaction. The right graph shows income categories and average life satisfaction. This combined graph makes it easier to interpret and compare our two explanatory variables. As stated previously, both predictor variables have a positive correlation with the response variable. So, there is a positive correlation in both graphs.

Table 2 Depicts the values of residuals, standard residuals, leverage, and Cook's distance for our GLR model.

The residuals are relatively low. The standard residuals have values $|r_i| < 2$ except for income_family = 1. Both of these values implies that there is one possible outlier point when using income_family as the predictor variable to feelings_life. Leverage measures the squared distance (in the x-direction) of x_i from the mean. No points in both GLM fit stands out. However, it is notable that the point where total_children == 0 and income_family == 1 has the highest value in its respective GLM fits. Cook's distance is used to identify influential data points when using linear regression analysis. It states the magnitude of change when the ith point, (x_i, y_i) is removed from the model. It can be seen in Table 2 that Cook's distance is highest when total_children == 0 and income_family == 1. Its respective standard residuals and leverage values also support the evidence that these points could be an influential point (outlier). This implies that this point is influential and may need further analysis. According to Sta302 Lec 5 slide 51, a point is noteworthy if $D_i > \frac{4}{n-2}$.

That is: $D_i > \frac{4}{8-2} = 0.67$ for total children and $D_i > \frac{4}{6-2} = 1$ for family income.

Thus, for our notable points for the family income graph are $(6, 8.420)$ and $(1, 7.479)$ and $(0, 7.879)$ for total children graph.

6 Discussion

The results show that there is a very strong correlation in our GLM. $beta_0$, $beta_{child}$, *and* $beta_{incfam}$ are all statistically significant, given a benchmark $\alpha = 0.05$ significance level. So, the null hypotheses, $H_0 : \beta_i = 0$ is rejected in favor of the alternative hypotheses, $H_a : \beta_i \neq 0$, where $i = 0$, child, incfam. The intercept and regressors' respective test statistics also support this position. This implies that our result is very unlikely to have been caused by chance or a sampling error. Furthermore, the sample size is large enough to further support this claim. Therefore, it can be confidently concluded that there is a strong positive correlation between both predictor variables and life satisfaction.

However, it is important to note that this is not sufficient enough to state that family income and number of children are causal effects to life satisfaction. The presence of unseen confounding variables, such as age and physical health, may also affect the value of life satisfaction. These confounders have not been taken into account in our model.

6.1 Weaknesses

Our findings have a number of limitations. First, sampling errors and nonsampling errors are unavoidable in GSS's complex stratification survey method, which were thoroughly discussed in the Data section. Second, the arbitrary ranges of income categories in our predictor family income, as determined by Statistics Canada, may result in misrepresenting some individual family income amounts, especially those at the fringes of a particular income category. For example, a household income of 25,001 is in the same category as a household income of 49,999. There is a large discrepancy between these two households, but it is determined to be equivocally the same in category 2 by the survey approach. Rather, 49,999 is much closer to category 3 than category 2 and 25,001 is much closer to category 1 than the rest of the data points in category 2, these arbitrary ranges cause errors at the fringes.

Thirdly, during the classification of our household income fields, we treated them as continuous values in our model. Treating the categories of incomes as continuous means that we can comparatively determine that for example, income categorized in 4 is larger than if a participant were to be categorized in 2. However, this doesn't necessarily mean that there is a mathematical correlation between the income of two given families in different categories. For example, in category 2 with an income of 25,001 and someone in category 4 with an income of 98,001, the difference between these values is 73,000 and not necessarily double the income given double the categorization group which would be the assumption of a continuous value in our model (Princeton University Library, 2007).

Lastly, the presence of confounders, as stated previously, can affect the life satisfaction scores. In a future study, there should be an incorporation of more predictor variables in the GLM. This will allow the model to be more informative. That is, life satisfaction scores are a subjective measure and there's no true standardization of the variable. If an unfortunate event happened to the survey participant before the collection of data, it may affect the life satisfaction score given to scientists.

6.2 Next Steps

Family income and children are strongly linked with life satisfaction. Further studies could be useful to track life satisfaction scores as family income and family size change over time. Results may be useful for all levels of governments and policy makers to better understand the needs of Canadian families, as well as to implement relevant programs and initiatives that enhance life satisfaction in the general Canadian population.

References

About Reaching Home: Canada's Homelessness Strategy. (2020, June 09). Retrieved October 20, 2020, from https://www.canada.ca/en/employment-social-development/programs/homelessness.html

Auguie, B. (2017). *gridExtra: Miscellaneous Functions for "Grid" Graphics.* R package version 2.3. https://CRAN.R-project.org/package=gridExtra

Bonikowska, A., Helliwell, J. F., Hou, F., & Schellenberg, G. (2013). An Assessment of Life Satisfaction Responses on Recent Statistics Canada Surveys. *Social Indicators Research*, 118(2), 617-643. doi: 10.1007/s11205-013-0437-1

DSS - Interpreting Regression Output. (2007). Retrieved October 19, 2020, from https://dss.princeton.edu/online_help/analysis/interpreting_regression.htm

DSS - Working with Dummy Variables. (2007). Retrieved October 19, 2020, from https://dss.princeton.edu/online_help/analysis/dummy_variables.htm

Lumley, T. (2020) *survey: analysis of complex survey samples.* R package version 4.0.

Marschner, I. C. (2011). Glm2: Fitting Generalized Linear Models with Convergence Problems. *The R Journal*, 3(2), 12-15. doi:10.32614/rj-2011-012

R Core Team (2019b). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rech, N. (2019, April 29). Homelessness in Canada. Retrieved October 19, 2020, from https://thecanadianencyclopedia.ca/en/article/homelessness-in-canada

Robinson, D., Hayes, A., and Couch, S. (2020). *broom: Convert Statistical Objects into Tidy Tibbles.* R package version 0.7.0. https://CRAN.R-project.org/package=broom

Rohan, Alexander. 2020. "GSS_Cleaning". STA304

Sheather, S. J. (2010). *A Modern approach to regression with R.* New York: Springer.

Statistics Canada. (2017, March 31). Canada at a Glance 2017. Retrieved October 19, 2020, from https://www150.statcan.gc.ca/n1/pub/12-581-x/2017000/pop-eng.htm

Statistics Canada. (2020, April 30). General Social Survey – Family (GSS). Retrieved October 19, 2020, from https://www.statcan.gc.ca/eng/survey/household/4501

Statistics Canada. (2020, May 26). Frequently asked questions. Retrieved October 20, 2020, from https://www.statcan.gc.ca/eng/survey/faq

Sun, J., Harris, K., & Vazire, S. (2019). Is well-being associated with the quantity and quality of social interactions? *Journal of Personality and Social Psychology.* doi:10.1037/pspp0000272

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

Wickham, H., François, R., Henry, L., and Müller, K. (2020). *dplyr: A Grammar of Data Manipulation.* R

package version 1.0.0. https://CRAN.R-project.org/package=dplyr

Wu, Changbao, and Mary E. Thompson. *Sampling Theory and Practice*. Springer, 2020.