Life Satisfaction Analysis using Linear Regression Model

Peiyu Li

October 19, 2020

Code and data supporting this analysis is available at:
https://github.com/PeiyuBellaLi/Peiyu-Li-STA304-PS2

Abstract

I used three variables(feelings_life, income_family, self_rated_mental_health) from the 2017 GSS data to fit a linear model based on their survey method. The result showed people's feeling about life was positively correlated with their family income and negatively correlated with their mental health status. Although this model was not very good, it was still very significant and meaningful. This analysis could help providing more information on some specific social policy issues that related with this topic, and also helped people to be aware of the factors that could potentially lower their life satisfaction.

Introduction

The General Social Survey(GSS) conducted in 2017 collected data about family which included the family income, mental health, and the respondents' feelings about life as a whole. The objectives for the GSS to collect those data were to modify the living conditions of Canadians and helped to provide information on specific policy. The sample was randomly selected from all non-institutionalized persons 15 years of age and older within 10 provinces of Canada.

The goal for this analysis was to see if people's feelings about their life are related with their mental health and family income. This topic would be benefit to monitor changes in order to increase life satisfaction for people living in Canada. Although, life satisfaction depends not only on the income and mental health, but also depends on other factors such as education level, marriages, and social relationships. It is still interesting to learn about how the income level and mental health are related with life satisfaction.

A linear model was built based on stratified sampling method. The model analysis as well as the model assumptions check will be explained in the following section. In general, this model is useful but not perfect, and all the predictors are in did correlated with people's life satisfaction.

Data

About the survey

The 2017 GSS data I got from the CHASS website contains over 20,000 observations and 81 variables. The data was collected by asking people to take a computer assisted telephone interviews, and those who were selected to be the respondents could choose their preference languages. If someone refuse to do the interview or did not answer the phone, they would be contacted several times later. Non-response occurred when respondents chose to valid skip or not stated or refuse to answer except for questions that were required for weighting (age, sex, etc.), and those value would be imputed or became missing values.

The target population for this survey included people who are 15 years old or older and except for full-time residents of institutions or people who lived in Yukon, Northwest Territories, and Nunavut. The frame was a list contained telephone numbers(both landline and cellular) and address for the target population. Also, the sample included people who were selected from each subgroups of the target population frame by their records.

In general, they used the stratified sampling method. They first assigned the records to strata within each province, and randomly selected samples within each stratum. The advantage of using stratified sampling could be that the sample selected from each stratum were representatives of the target population, since it made sure that people from all different regions in Canada had the chance to be selected and take this survey. It contained lots of information that could be use to do many analysis from different aspects.

However, their were 14% of telephone numbers not linked to the Address Register in the frame, it was hard to divide those records into different strata based on province. This survey took about 10 month to be conducted, which was a very long period of time that also cost lots of people to do it. Since the target population is very large, they had to use some other data such as the Census of population, which made the data to be messy before cleaning.

About the data

I used two categorical variables(income_family, self_rated_mental_health) both contained 6 levels to predict the numerical variable(feeling_life) ranged from 0 to 10. There was another variable(self_rated_health) that were very similar to self_rated_mental_health, but I nonetheless did not use it since I want the predictors to be different and mental health should affect life satisfaction more in my mind.

Figure 4 in Appendix shows an overview of the response variable. The information I got from there was that most respondents had their life satisfaction value greater than 5, and majority of people were around 8. There must be some different for people in different groups such as people with different family income level or mental health status in this case.

Model

In order to get an better idea on how i can explained the relationships on those three variables, I built a multiple linear regression model based on the stratified sampling method using R. A multiple linear model means I could use more than one predictors both numerical or categorical to predict a numerical response.

The Finite Population Correction(FPC) is needed when the sample is selected without replacement from a finite population. Since the GSS used stratified sampling (ie. sampled based on 10 provinces), i needed to adjust FPC by adding a new variable called 'fpc' that specified the population in the province that the observations sampled from. The GSS data user guide said their target population is 30,302,287, and I calculate the target population within each province by multiplied this number to different proportions for each province of the Canadian population.

After fitted a linear model based on the survey, I got the following formula,

$$\textit{feelings\_life} = \beta_0 + \beta_1\textit{income\_family \$25,000 to \$49,999} + \beta_2\textit{income\_family \$50,000 to \$74,999} + \beta_3\textit{income\_family \$75,000 to \$99,999} + \beta_4\textit{income\_family \$100,000 to \$124,000} + \beta_5\textit{income\_family \$125,000 and more} + \beta_6\textit{self\_rated\_mental\_health Very Good} + \beta_7\textit{self\_rated\_mental\_health Good} + \beta_8\textit{self\_rated\_mental\_health Fair} + \beta_9\textit{self\_rated\_mental\_health Poor} + \beta_{10}\textit{self\_rated\_mental\_health Don't Know}$$

All the β's in this model represents the estimates of the coefficients, and the specific value is shown in the result section. The estimate intercept (β_0) represents the life satisfaction value when the family_income is

Less than Less than $25,000, the self_rated_mental_health is 'Excellent'. This model can be explained in the following way: when one's family income is between $25,000 and $49,999 and with very good mental health status, then this person's feeling about life are likely to be $\beta_0 + \beta_1 + \beta_6$.

The reason why income_family was a categorical variable was that people might not want to provide specific income information, and for people who refused to provide this information, the income_family values were imputed and that was why it could not be specific numbers. Also, it was more clear for respondents to self rate their mental health level so this variable was also categorical.

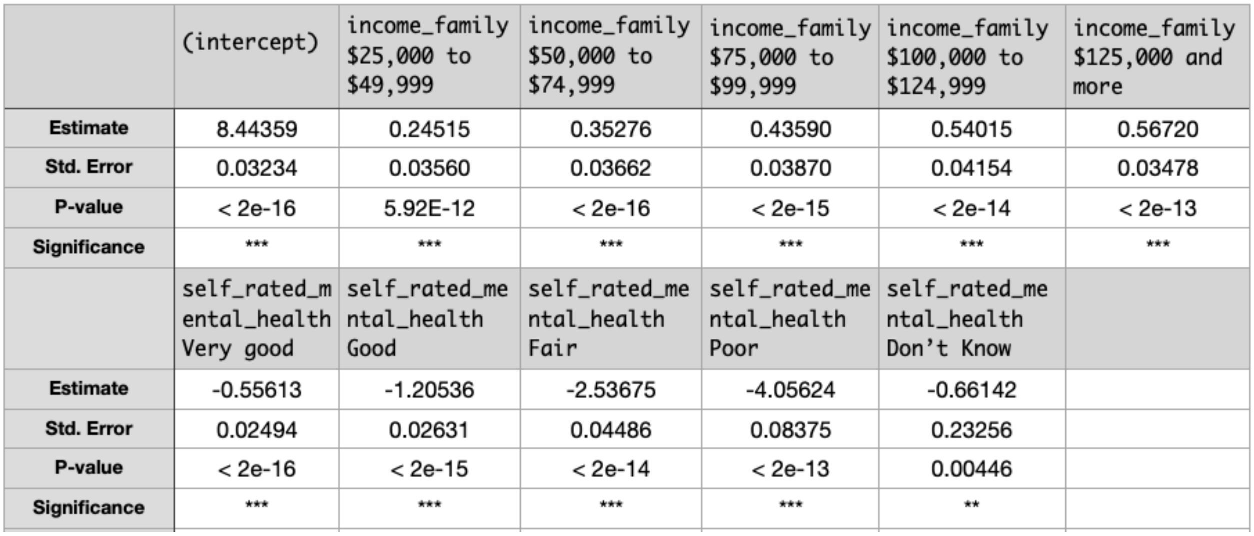
The drawbacks for this model is that it could only be used with people who live in Canada based on the target population. It is also insufficient to predict life satisfaction just based on these two predictors. This model can be used to predict one's life satisfaction using family income and mental health level, but there must have other factors affect feelings about life.

The model assumptions and diagnostic issues will be discussed in details in the next section, but this model didn't actually meet all the assumptions perfectly. However, it ended up being passed the hypothesis test which mean it had a very small p-value that suggested this model is meaningful. All the variables were also significant, and 25% of variations could be explained by this model.

The alternative model can be done by dividing the values in feelings_life into binary case and doing a logistic model. The advantage is the data can fit the model much better than the linear one I did, but it loses many important features in the response variable. Despite there are other models can be chose, they should all based on the stratified survey sampling method to reflect how this survey was done.

Results

After fitting a linear model, I got the following summary table(Figure 1) for each coefficient.



	(intercept)	income_family $25,000 to $49,999	income_family $50,000 to $74,999	income_family $75,000 to $99,999	income_family $100,000 to $124,999	income_family $125,000 and more
Estimate	8.44359	0.24515	0.35276	0.43590	0.54015	0.56720
Std. Error	0.03234	0.03560	0.03662	0.03870	0.04154	0.03478
P-value	< 2e-16	5.92E-12	< 2e-16	< 2e-15	< 2e-14	< 2e-13
Significance	***	***	***	***	***	***
	self_rated_mental_health Very good	self_rated_mental_health Good	self_rated_mental_health Fair	self_rated_mental_health Poor	self_rated_mental_health Don't Know	
Estimate	-0.55613	-1.20536	-2.53675	-4.05624	-0.66142	
Std. Error	0.02494	0.02631	0.04486	0.08375	0.23256	
P-value	< 2e-16	< 2e-15	< 2e-14	< 2e-13	0.00446	
Significance	***	***	***	***	**	

Figure 1: Summary of the model

The estimate intercept (β_0) means when the family_income is Less than Less than $25,000, the self_rated_mental_health is 'Excellent', then the feelings_life is around 8.44. The model treated the two predictor variables as dummy variables, and the baselines are income_family less than $25,000 and the self_rated_mental_health is 'Excellent'. As a result, the coefficient for 'family_income $25,000 to $49,000' can be interpreted as when the family income increases from less than $25,000 to between $25,000 to $49,000, the life satisfaction increases by 0.245. In a similar way, the coefficient for self_rated_mental_health 'Very

Good' can be interpreted as when the self rated mental health level changes from 'Excellent' to 'Very Good', the life satisfaction of that person will be decreased by -0.556.

The overall p-value for this model is less than 2.2e-16 (<0.05) which is very significant. This means I have enough evidence to say that this model is useful and meaningful in this case. Also notice that the p-value for each predictors are also very significant, which indicates as the predictor changes the response variable will also change. In other words, all the variables in this model are meaningful and strong enough to affect one's life satisfaction. The R-squared for this model is 0.2497, which means this model can explain 24.97% variation in feelings about life. This is good enough though it is not too high.

To check the model assumptions, I plot the following residual plot(Figure 2) and the Normal Q-Q plot(Figure 3). The linearity was proved by Figure 2, since the red line was horizontal at 0, I could assume there was a linear relationship between the predictors and the response. Figure 3 could be used to check normality assumption, the points at the end of each side did not fall around the line perfectly so this assumption seemed not passed.

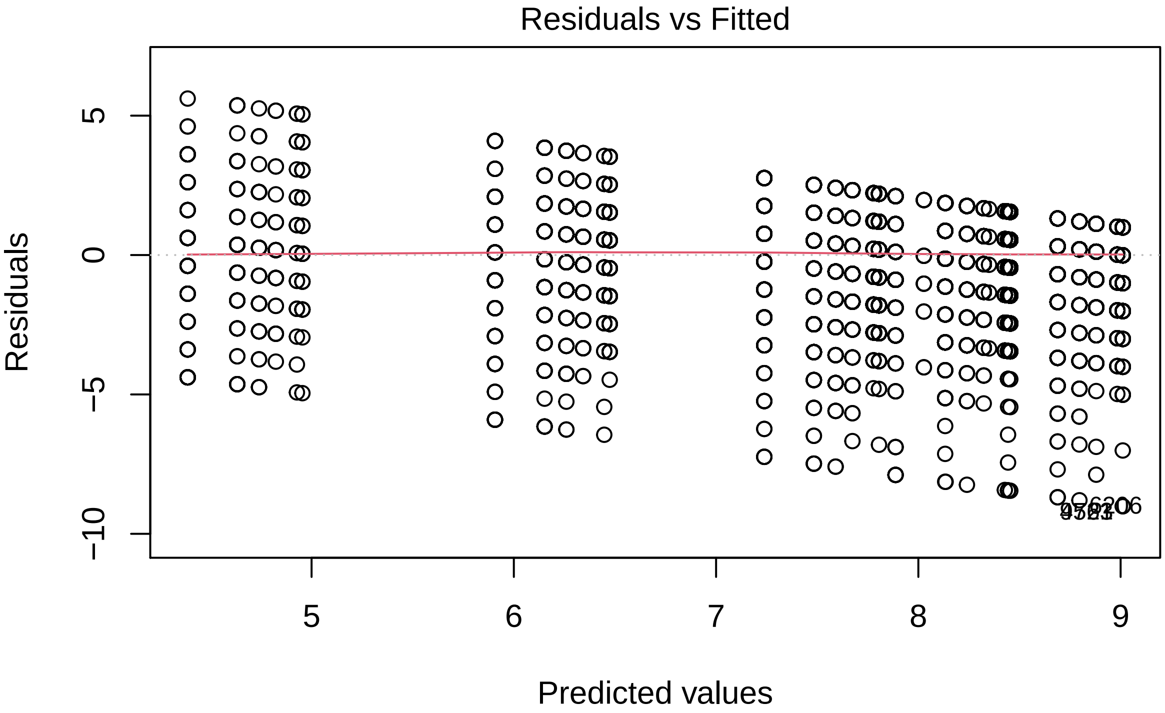
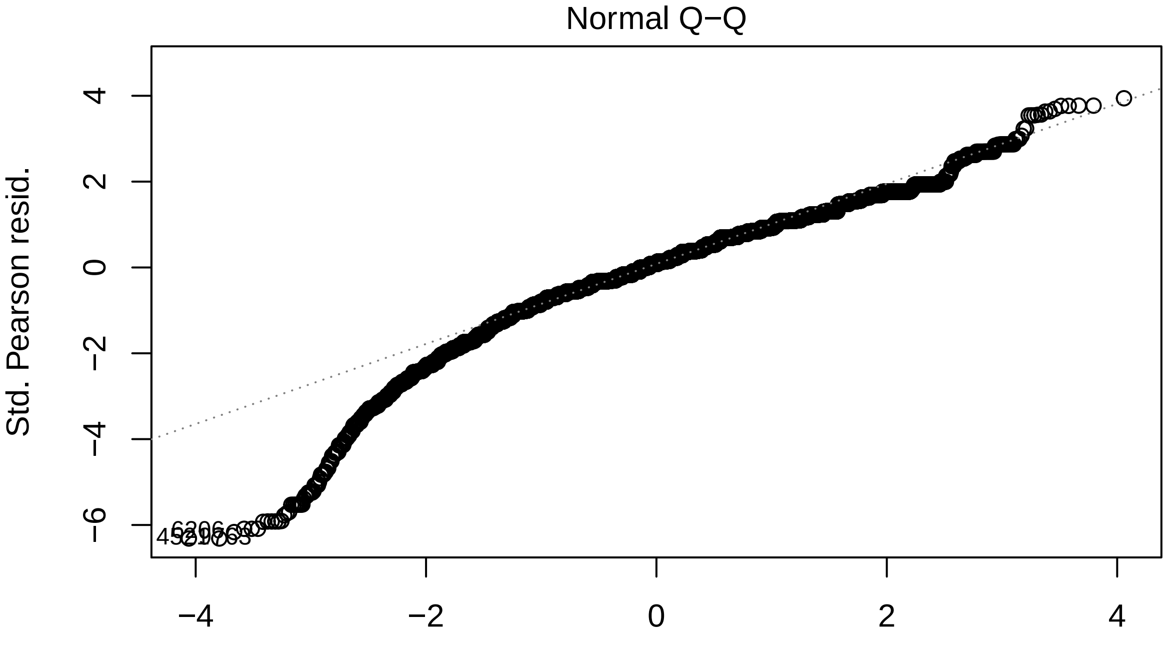


Figure 2



Theoretical Quantiles
Figure 3

Discussion

From the linear model i built, it shows as the family income level increases, the feelings about life will also increases, and as the self rated mental health level decreases, the feelings about life will also decreases. The box plots in Appendic (Figure 5, Figure 6) further verified this results. In Figure 5, we can see the medians for life satisfaction level for people with different family income are the same, and there is not a clear relationship between family income and feelings about life but for people with the highest income level, the overall satisfaction is obviously higher. In Figure 6, as people's mental health level decreases, the median life satisfaction level decreases, except for people who don't know about their mental health status.

The small p-values suggests this model is meaningful, and the R-squared indicates it can explain around 25% of the variation. It is good enough though the value is small. However, the linearity assumption for the model passed while the normality was not good. The result suggest that maybe I should try to transform the data in order to use linear model better, or I could just use another model to fit the data.

Since the model is still significant, it can provide information to both policy makers and Canadians. They can get an better idea on how their feelings about life as a whole could be affected by their family income and mental health status. The income in a family did have an effect on their life satisfaction as suggest in this model, but from Figure 1 we can see most people are satisfied and have a high median life satisfaction for all income levels. Also, it remind people that mental health issues are very important to their lives. If people feel their mental health status is low, they should better go to see their psychologists or try other methods to deal with the issues.

Weaknesses

According to the model assumption check, this model is not good enough though all the variables are significant. In the small world, life satisfaction have strong correlation with both family income and mental health problem. However, in the big world, there must be some other factors that are related with life satisfaction and maybe a third variable is taking place. The information provided here is limited because life satisfaction is a really general thing that can't be explained perfectly just using two variables.

The data itself is also not perfect due to various reasons. According to the GSS user guide, they only collected 82.6% of households' family income value. The family income for respondents who did not answer was imputed. The overall response rate for the survey was 52.4% which is not high, and therefore many missing value in the raw data set.

Next Steps

Since the normality assumption was not satisfied, the next step could be using data transformation method to see if it improve the performance of linear model. Another option is to use other models such as Bayesian that may fit better to the data. It's also a good idea to include more predictors from the GSS data and do a more complicated model, and this can help to explain how people's life satisfaction change based on more factors.

In order to make this topic more meaningful and useful, another survey could be done. The survey questions can narrow down to some more specific questions that are related to feelings about life. It can be implemented not only in Canada but also in some other countries if the cost is affordable. In this way, it can further benefit to monitor changes worldwild in order to increase people's life satisfaction.

References

Alexander,R.,& Caetano, S. (2020). gss_cleaning.R. https://q.utoronto.ca/courses/184060/modules/items/1867317

Bürkner, P. (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

Kassambara. (2018). Linear Regression Assumptions and Diagnostics in R: Essentials. http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/#linearity-of-the-data

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

Pareto, A. (2015). How to add an image to markdown. https://rpubs.com/RatherBit/90926

Statistics Canada. (2020). 2017 General Social Survey: Families Cycle 31: Public Use Microdata File. Using CHASS (distributor). https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_Codebook.pdf

Statistics Canada. (2020). General Social Survey, Cycle 31: Families, Public Use Microdata File Documentation and User's Guide. Using CHASS (distributor). https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/sdaweb/dli2/gss/gss31/gss31/more_doc/GSS31_User_Guide.pdf

Statistics Canada. (2017). Genral social survey on Family (cycle 31), 2017. Using CHASS(distributor). https://sda-artsci-utoronto-ca.myaccess.library.utoronto.ca/cgi-bin/sda/hsda?harcsda4+gss31

Statistics Canada. (2020). Table 051-0005:Estimates of population, Canada, provinces and territories. http://www5.statcan.gc.ca/cansim/a26?lang=eng&retrLang=eng&id=0510005&paSer=&pattern=&stByVal=1&p1=1&p2=31&tabMode=dataTable&csid=

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

Appendix

Code and data supporting this analysis is available at:

https://github.com/PeiyuBellaLi/Peiyu-Li-STA304-PS2

Figure 4

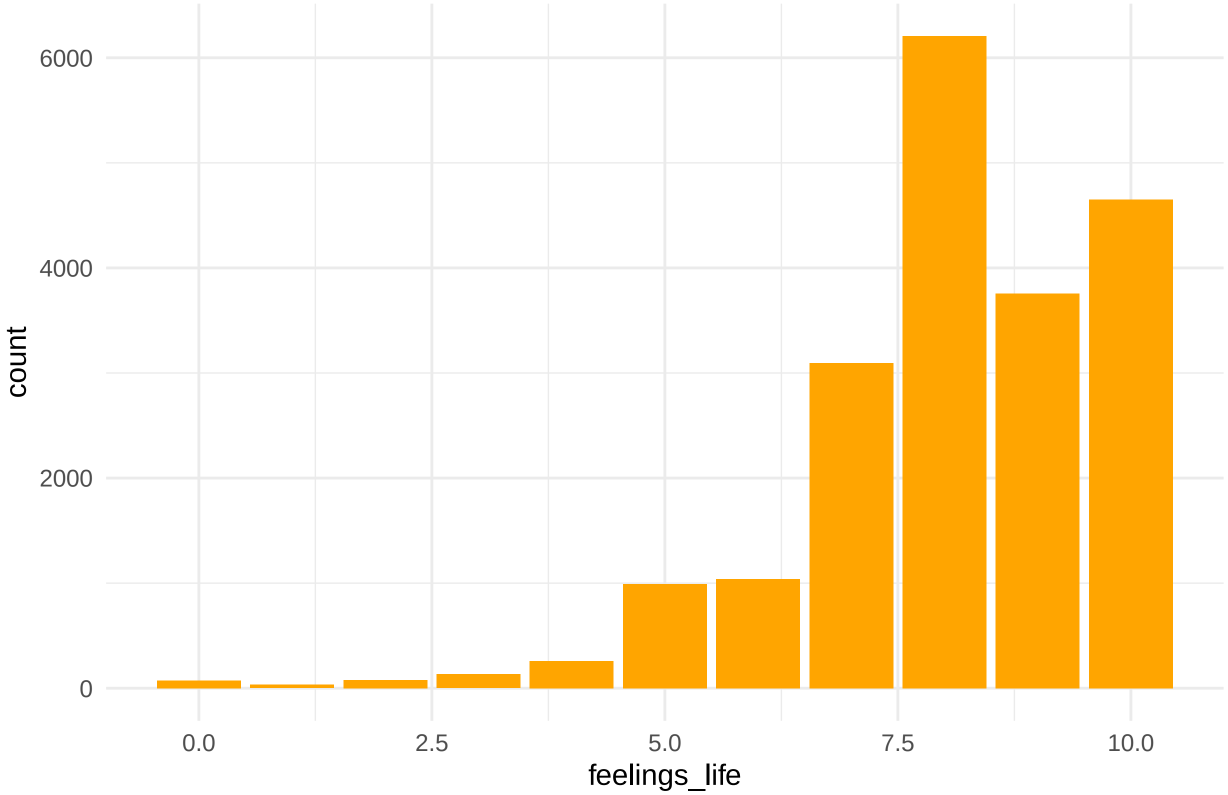


Figure 5

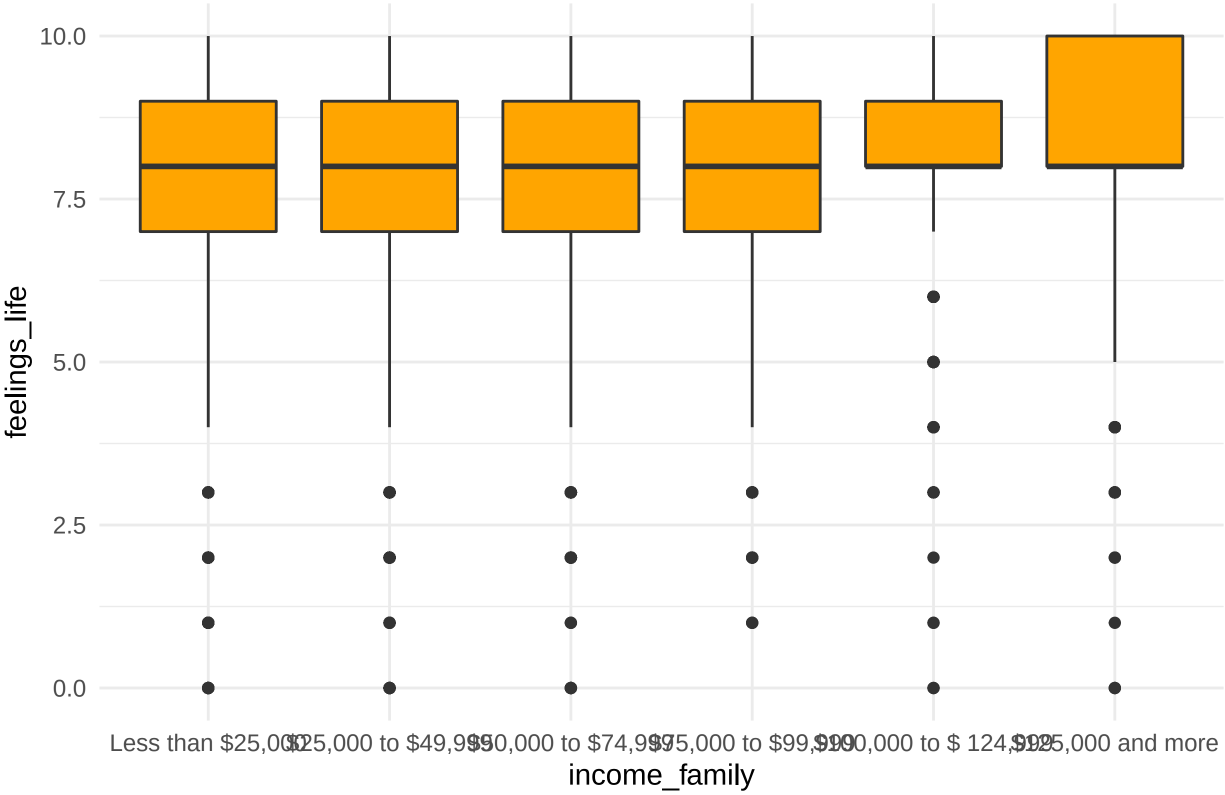
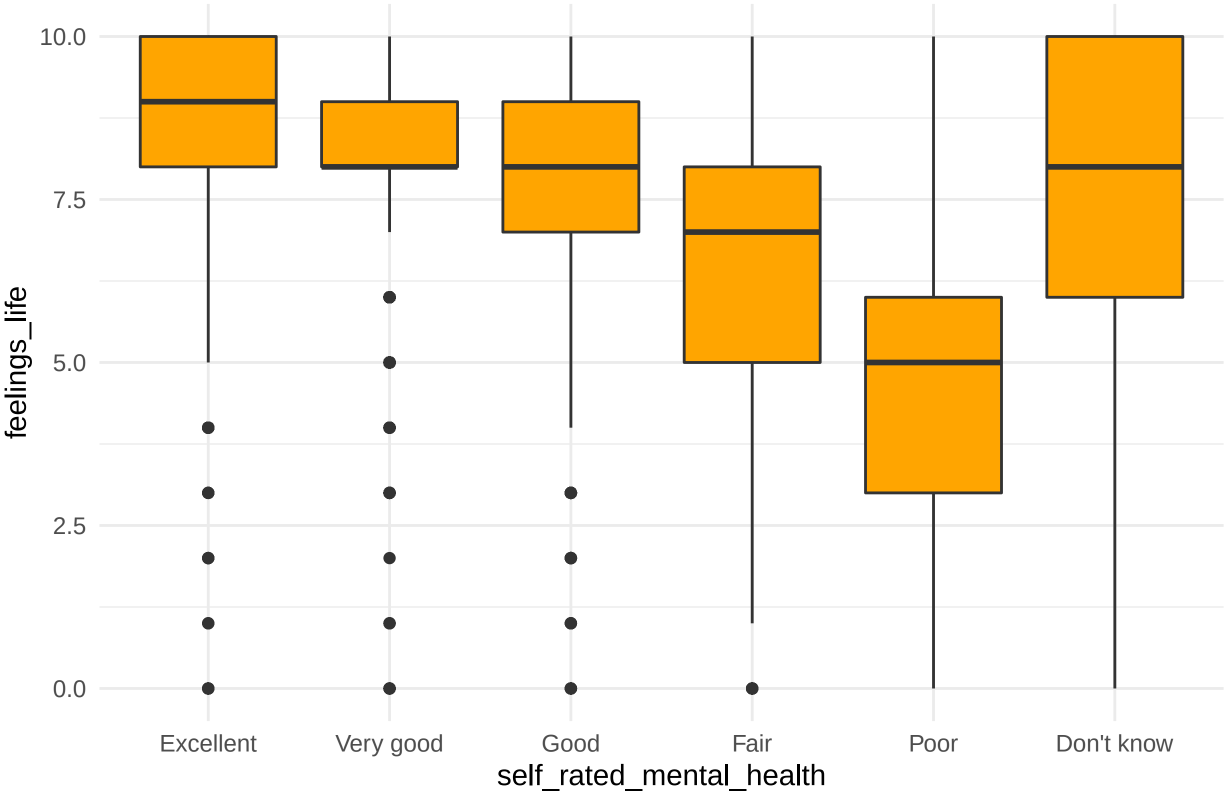


Figure 6


feelings_life
Excellent
Very good
Good
Fair
Poor
Don't know
self_rated_mental_health