Marriage of Equivalents Turns Into Norm

Group 130: Jie Huang(1004925156) & Qing Li(1005148010) & Xi Zheng(1005153628)

2020/10/19

Code and data supporting this analysis is available at: https://github.com/xixiaoguai727/sta304ps2

Abstract

Marriage of equivalents means people choose partners with similar education levels, earnings, values of the original family, and lifestyle. In this report, we build a statistical model and aim to estimate the education level of partners based on the education level and income of respondents from the 2017 Canadian General Social Survey–Family. By analyzing this model, we find that the education level and income of the respondents are positively related to the education level of their partners. In other words, our model justifies the idea of marriage of equivalents turns into the norm.

Introduction

Nowadays, more and more people put backgrounds in the first place to consider when they are choosing partners, "the equality of marriage becomes the norm"[7]. The change of the original criteria for choosing a spouse influences society a lot. As highly educated students, we are interested in how education level and income affect the assortative mating. In the step-by-step analysis, the first step, we sort the dataset, and then we build a multiple linear regression model based on sorted data. Finally, we make a discussion on the results, strengths, and weaknesses of the model.

Data

We download the datasets from the Chass Data Centre, and we get the related questionnaire from Canadian General Social Survey(GSS). We use R studio to run code to help us select the valid respondents and move away from the respondents who choose the option DK(do not know the question) and RF(reject to answer the problem).
In our project, the target population is all Canadians over fifteen years old; the frame is 20602 respondents; the sample is 12199 respondents. They are the people who provide valid information about their income and education level.
The survey asks for a lot of personal related information, and many people made responses to the survey, so the survey collects diverse data for us to analyze. However, some of the questions are too private, so some respondents may refuse to answer them. This action will cause invalid information and will further lead to low accuracy.

Table1: Numbers With Corresponding Education Level

Number	Education Level
1	Less than high school diploma or its equivalent
2	High school diploma or high school equivalency certificate
3	Trade certificate or diploma
4	College, CEGEP or other non-university certificate or di...
5	University certificate or diploma below the bachelor's level
6	Bachelor's degree (e.g. B.A., B.Sc., LL.B.)
7	University certificate, diploma or degree above the bach...

Table2: Numbers With Corresponding IncomeRange

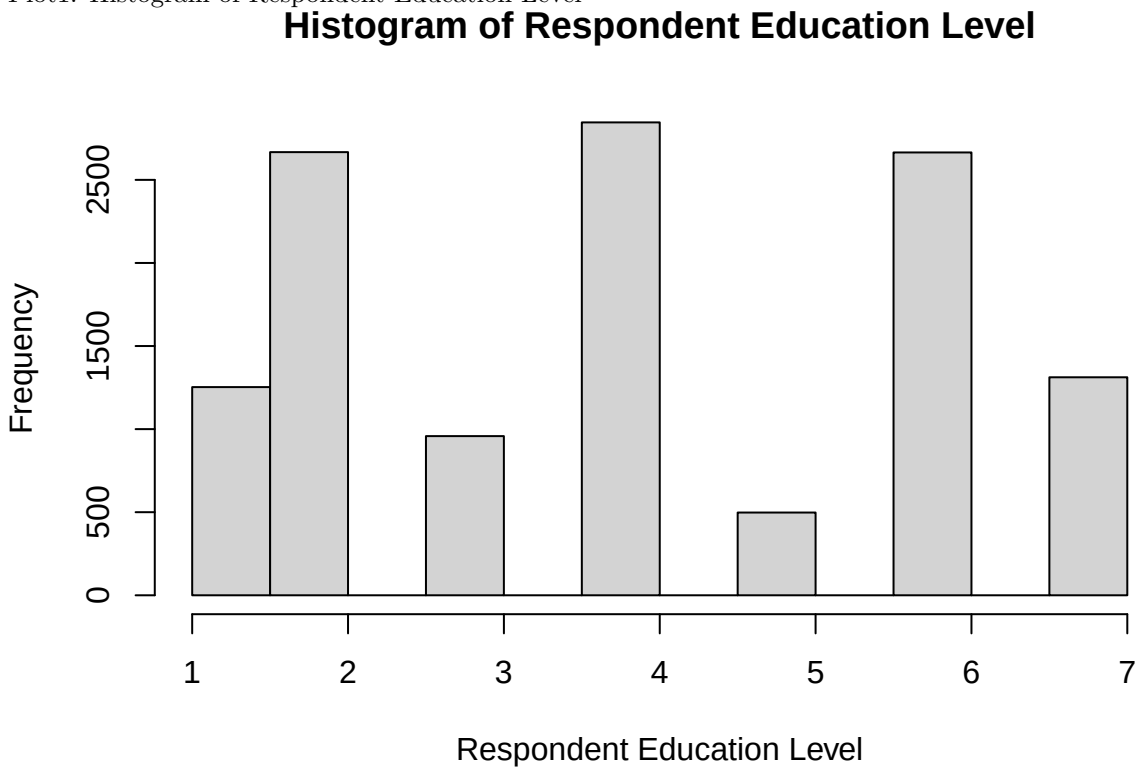
Number	Income before tax and unit in CAD
1	Less than 25000
2	25000 to 49999
3	50000 to 74999
4	75000 to 99999
5	100000 to 124999
6	125000 and more

The independent variable of the education level of respondents are categorical as shown in Table1. Still, we are going the treated them all as numerical variables since the numbers increase in a way which makes sense(bigger number represents higher level). The education level of their partners is a dependent variable. The other independent variable is the income of the respondent. It is a categorical variable, as shown in Table2. There is no similar variable we can use from the datasets, so that we will use only those three variables.

Below is the overall summary table for our dataset:
Table3: Summary Table for the Overall Dataset

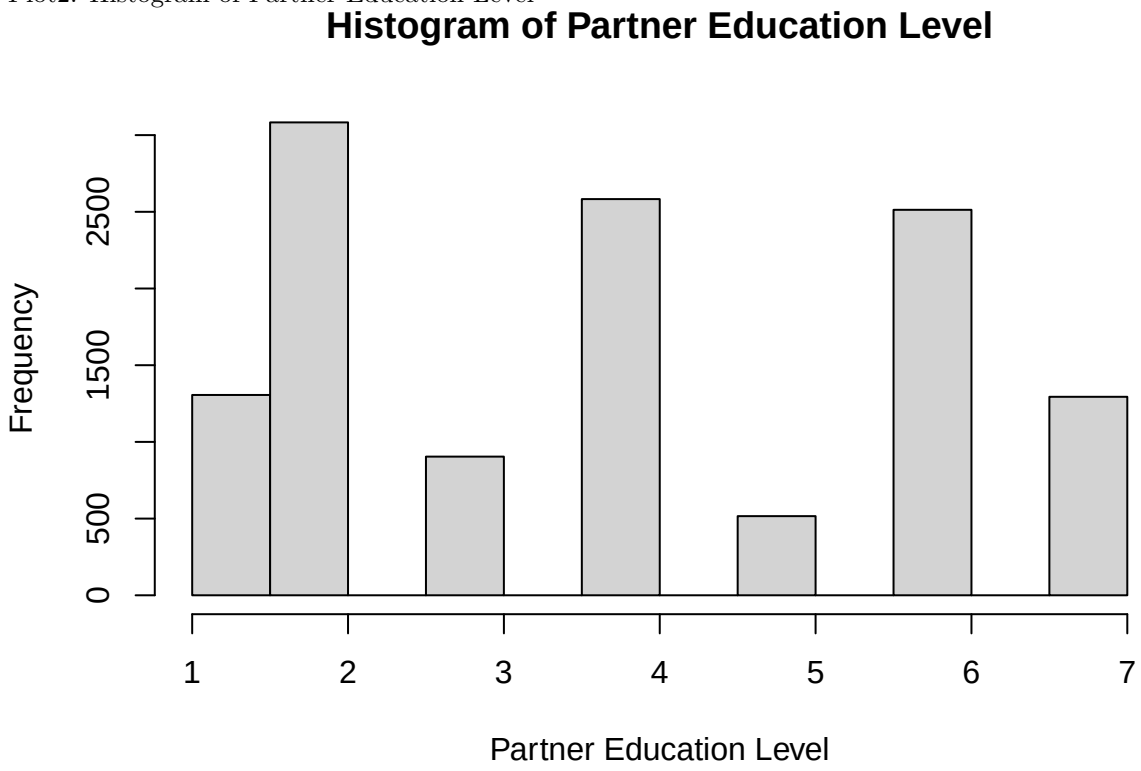
```
##      CaseID      RespondentEducationLevel PartnerEducationLevel
##  Min.   :    1   Min.   :1.000            Min.   :1.000
##  1st Qu.: 5227   1st Qu.:2.000            1st Qu.:2.000
##  Median :10352   Median :4.000            Median :4.000
##  Mean   :10337   Mean   :3.976            Mean   :3.872
##  3rd Qu.:15482   3rd Qu.:6.000            3rd Qu.:6.000
##  Max.   :20602   Max.   :7.000            Max.   :7.000
##  RespondentIncome
##  Min.   :1.000
##  1st Qu.:1.000
##  Median :2.000
##  Mean   :2.556
##  3rd Qu.:3.000
##  Max.   :6.000
```

In Plot1, the histogram of the education level of the respondent is symmetric since it has a prominent peak. From Table3, we know that the median and the mean of the asymmetric dataset are similar simultaneously, which are 4.000 and 3.976. We can also find that the first quarter is 2.000, and the third quarter is 6.00.
Plot1: Histogram of Respondent Education Level

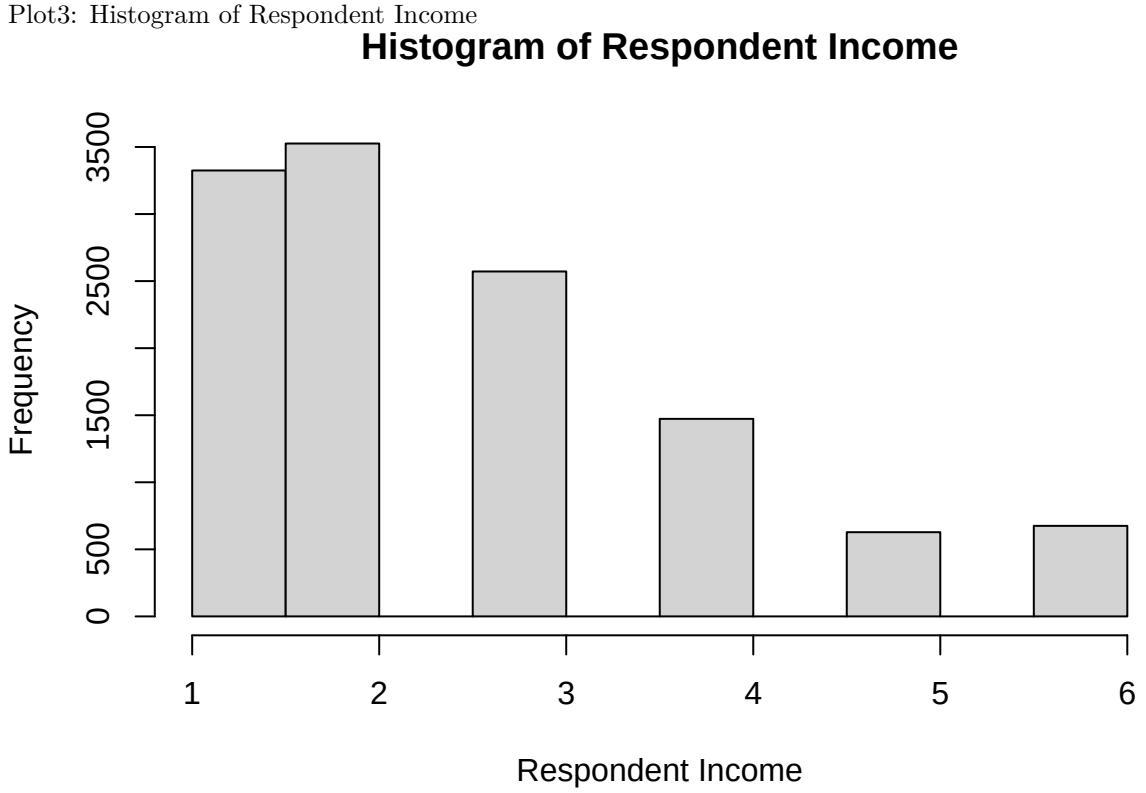


In Plot2, the histogram of the education level of the partner is a bimodal shape. There are two peaks from two different systems in the graph. The mean of the education level of the partner is 3.872, and the median of the education level is 4.000.

Plot2: Histogram of Partner Education Level



In Plot3, the histogram of the income of the respondent has a right-skewed tail, which means the distribution is a positive-skew distribution. Additionally, the IQR is Q3-Q1=3.000-1.000=2.000.
Plot3: Histogram of Respondent Income



Model

We use R studio to build a multiple linear regression model to see an association between the two independent variables and the response variable. Additionally, it is appropriate to use a multiple regression model to analyze the datasets because we want to predict the value of a variable based on the amount of two or more other variables. Besides, we set up a null hypothesis that the coefficients of two predictor variables are equal to zero, indicating no association between the two predictor variables and the response variable.
Furthermore, we choose the significance level of 0.05. The significance level is "a measure of the strength of the evidence that must be present in our sample before we will reject the null hypothesis and conclude that the effect is statistically significant"[3], and we will compare it to the p-value we got to see if we should reject the null hypothesis or not, "p-value is a measure of the probability that an observed difference could have occurred just by random chance and the lower the p-value, the greater the statistical significance of the observed difference"[2].
This is the equation produced by our model:
Equation1: Predict Equation

$$y = 1.736 + 0.519X_{RespondentEducationLevel} - 0.023X_{RespondentIncome2} + 0.095X_{RespondentIncome3}$$
$$+0.212X_{RespondentIncome4} + 0.152X_{RespondentIncome5} + 0.445X_{RespondentIncome6}$$

The interpretation of Equation1 is the education level of a partner equals 1.736 plus 0.519 times the education level of the respondent plus the specific coefficient times the corresponding income range.
Where the intercept parameter 1.736 is the education level when the education level and income range for the respondent are both zero, which is impossible since the lower limit for those two variables is 1.
0.519 is the slope parameter for education level for the respondent, and another coefficient is the slope parameter for income range for the respondent.

One possible alternative model to replace the multiple linear regression model is a simple linear model. The advantage of using simple linear regression is that we can quickly check if there is a relationship between the education level of respondents and the education level of their partners. However, having only one variable is not accurate enough to estimate the response variable.

Moreover, to build the model, we treat the income of respondents as a categorical variable because the survey did not ask for a specific income number, so we keep this dataset as categorical. Next, ideally, education level should be categorical, but we treat the education level as a numerical variable in this model. Furthermore, the weakness of treating education level as numerical is that the number from 1 to 7 can not summarize all education levels.

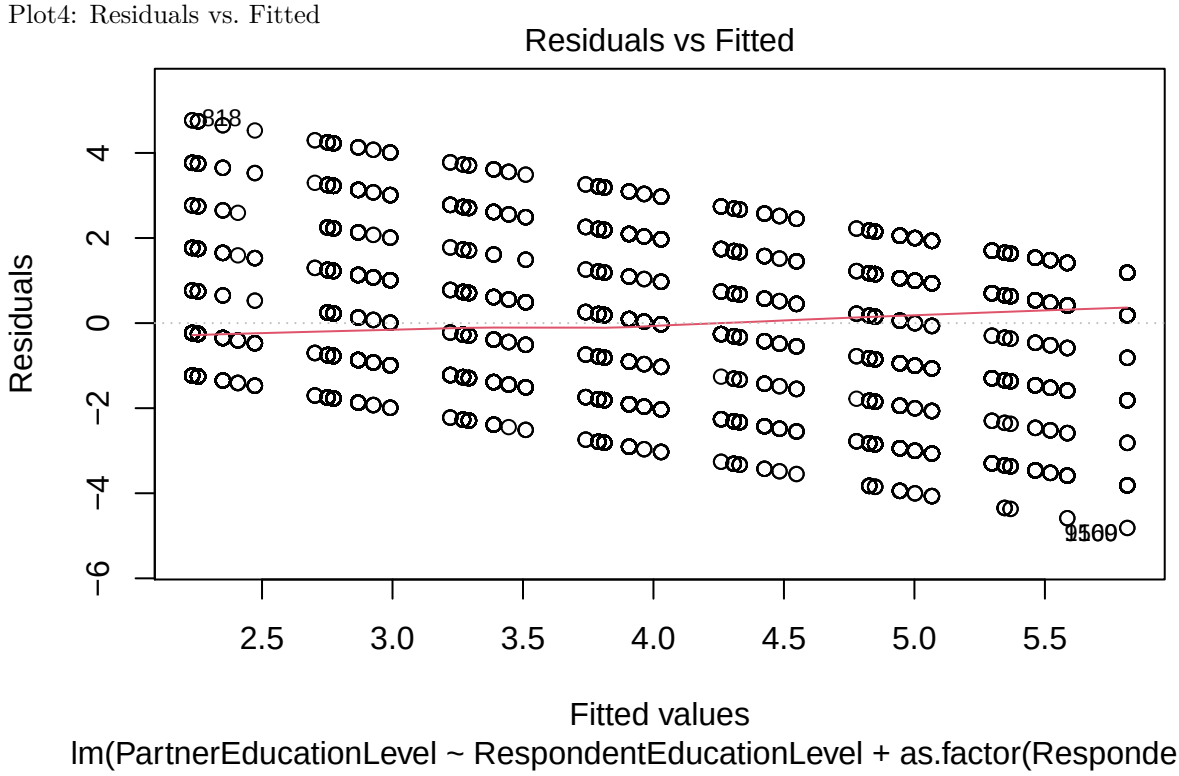
Results

By fitting the linear regression model, we find that the p-value in Table4 is $< 2.2e\text{-}16$, which is significantly smaller than 0.05. Thus we reject our null hypothesis. Therefore, Equation1 in the Model section makes sense.

Table4: Summary Table for Regression Model

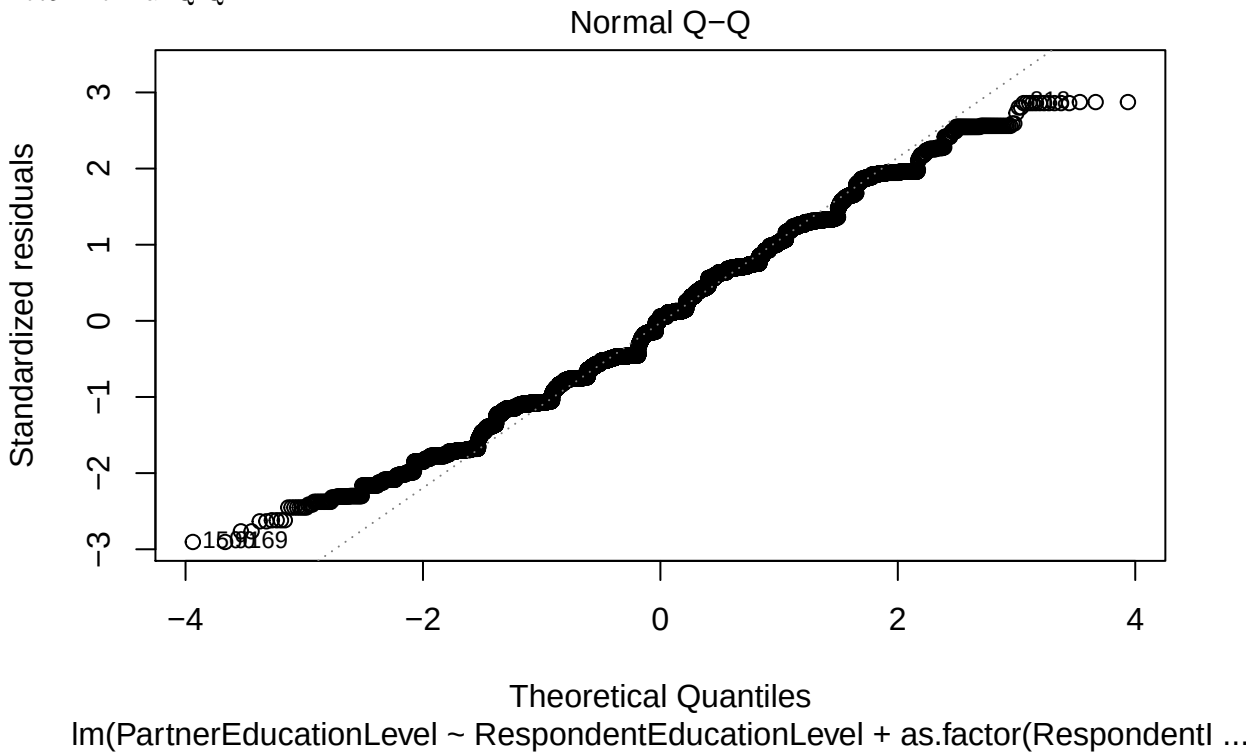
```
##
## Call:
## lm(formula = PartnerEducationLevel ~ RespondentEducationLevel +
##     as.factor(RespondentIncome), data = EducationData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8158 -1.2550  0.0736  1.1732  4.7677
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.736156   0.039317  44.158  < 2e-16 ***
## RespondentEducationLevel      0.518887   0.008234  63.014  < 2e-16 ***
## as.factor(RespondentIncome)2 -0.022695   0.040235  -0.564   0.5727
## as.factor(RespondentIncome)3  0.094801   0.044488   2.131   0.0331 *
## as.factor(RespondentIncome)4  0.217504   0.053571   4.060 4.94e-05 ***
## as.factor(RespondentIncome)5  0.152450   0.073515   2.074   0.0381 *
## as.factor(RespondentIncome)6  0.447434   0.071915   6.222 5.08e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.66 on 12192 degrees of freedom
## Multiple R-squared:  0.2856, Adjusted R-squared:  0.2852
## F-statistic: 812.3 on 6 and 12192 DF,  p-value: < 2.2e-16
```

Now we are going to do some diagnostic checks.
–Diagonostic check:
Plot4: Residuals vs. Fitted



In Plot 4, the horizontal line in this plot indicates a linear relationship between the two independent variables and the dependent variable.

Plot5: Normal Q-Q

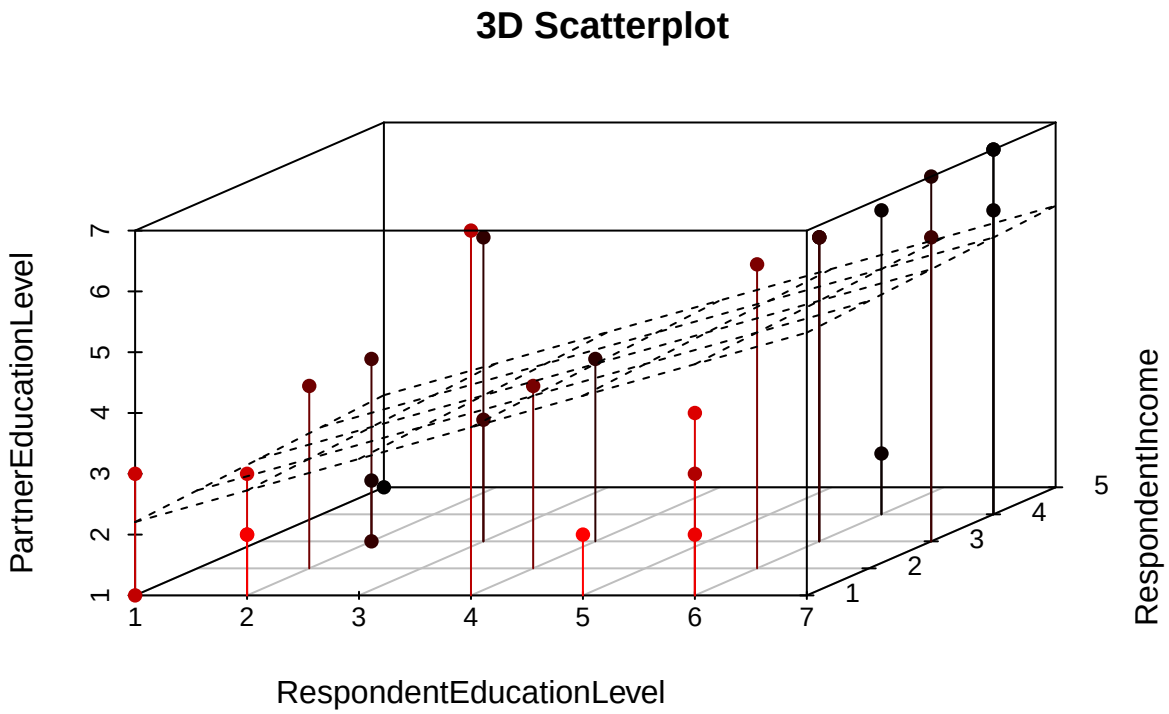


In Plot 5(Normal Q-Q), we can see that residuals points follow the straight dashed line to conclude that model meets the normality assumption.

– 3D-Plot

To test our results that there is a linear relationship between the education level of partner, the education level and income of the respondent, we set up the same regression model as we did in the model section, but treated the variable RespondentIncome as a numerical variable to build a 3D-Plot(we can not build 3D graph by using categorical variables). We randomly select 30 samples from the sample population to see if the model is fitted.

Plot6: 3D Graph for 30 Samples With Regression Plane



As we can see, in Plot6, the dashed plane represents our estimated values of the response, and we can see that points locate around the dashed plane, thus our predicted model should be applicatable.

Disscussion

From the summary table we can write this model as a recall:
Equation1: Predict Equation

$$y = 1.736 + 0.519X_{RespondentEducationLevel} - 0.023X_{RespondentIncome2} + 0.095X_{RespondentIncome3}$$
$$+0.212X_{RespondentIncome4} + 0.152X_{RespondentIncome5} + 0.445X_{RespondentIncome6}$$

In addition, the p-value is smaller than 2.2e-16, which means we have strong evidence against the null hypothesis. In other words, we can prove that there exists a relationship between two predictor variables and the response variable.

We download the Chass Data Centre datasets, and we get the related information from the Canadian General Social Survey(GSS). The datasets are not perfect because some respondents refuse to answer some of the questions(RF) or choose the option DK(don't know)as the answer to the question. Based on these data, we cannot find that the education level of respondents and their income is positively related to the education level of partners since these data cannot generate a table to analyze. Therefore, we remove these data, which is about DK and RF.

The survey questions cover a wide range of information. We can get all kinds of data from the survey. Furthermore, many respondents participated in this survey, so we can collect diverse data to analyze. However, the problem is that some of the questions are too private, so many respondents are unwilling to fill in or lie on the answer; this will further affect the overall accuracy of the data.

Weakness of Model and Possible Future Improvements

First of all, the data we are using is not perfectly accurate. However, we cannot fix this problem. Another weakness of the model is that the two predictors may have correlation, which means we should keep only one independent variable. To fix this problem, we should remember to check for a correlation between the selected independent variables in future projects. Besides, to make the model more accurate, we can take more predictors in the model.

References

[1]"Add Text to a Plot in R Software." STHDA, www.sthda.com/english/wiki/add-text-to-a-plot-in-r-software.
[2]Beers, Brian. "What P-Value Tells Us." Investopedia, Investopedia, 13 Sept. 2020, www.investopedia.com/terms/p/p-value.asp.
[3]Frost, Jim. "Significance Level." Statistics By Jim, 5 May 2017, statisticsbyjim.com/glossary/significance-level/.
[4]Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation.https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.
[5]Kassambara, et al. "Linear Regression Assumptions and Diagnostics in R: Essentials." STHDA, 11 Mar. 2018, www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/.
[6]Ligges, U. and Mächler, M. (2003). Scatterplot3d - an R Package for Visualizing Multivariate Data. Journal of Statistical Software 8(11), 1-20.
[7]Marriage of Equals Becomes the Norm, epaper.bostonglobe.com/BostonGlobe/article_popover.aspx?guid=b6de7202-f21b-4ff5-8cf6-1294595a7cbd.
[8]robk@statmethods.net, Robert Kabacoff -. "Axes and Text." Quick-R: Axes and Text, www.statmethods.net/advgraphs/axes.html.
[9]Statistics Canada: Canada's National Statistical Agency/Statistique Canada : Organisme Statistique National Du Canada, Government of Canada, Statistics Canada, 20 Feb. 2019, www150.statcan.gc.ca/n1/pub/89f0115x/89f0115x2019001-eng.htm.
[10]Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686/newline

[2], [3], [7] are references of context
[9], [10] are references of data source
[1], [4], [5], [6], [8] are references for r codes