Multiple Linear Regression Model: Using Education Condition, Unemployed, Number of Immigrant and Average of Shelter Costs to Predict Income

Youwen Xu(1006675493)

2020/12/22

Appendix

GitHub link: https://github.com/youwen1127/STA304-Finalreport

Abstract

Income is an indispensable thing in human lives. However, it will be affected or changed by many other factors. The purpose of this report is to predict what are the variables that affect income, by analyzing a **voluntary National Household Survey (NHS)** in 2011 by Statistics Canada(5). This data investigates the relationship between respondents' **educational background, work status, social trends, and income**. Through the cleaned data, 6 variables are decided to establish a **multiple linear regression model**.(5) Through analysis and research, it is shown that the complete education level or, there are more other potential factors in society. Therefore, in order to analyze this topic more comprehensively, future investigations and research are also essential.

Introduction

Income has played an important role since ancient times and is closely related to daily life. It can even represent the state and quality of life of a family or individual. From the **census data**, the **causal relationship** reflected in income is based on various factors such as **education level, work status and the number of immigrants** and so on. The government uses this data to infer and formulate some support plans or develop new services to balance families or families with different levels of living conditions (5). Therefore, according to different factors, the consumption levels in different regions can also be used to infer income differences. The **Canada Revenue Agency** can also adjust the taxation of each region through the data (5). This means that analyzing income will be an object worthy of research because the results of this research can provide support to families in need and improve the happiness of life.

In this case, according to the **National Household Survey (NHS)**, an average of one in six children lives in a low-income family or even poor (1). Then it can be shown that income will be directed as one of the factors that affect the ability of a family to choose a quality life. In this report, it will also reflect whether there is a causal relationship between family background and income level.

The 2011 Toronto census collected by the NHS was voluntarily, and a **Multiple Linear Regression Model** will be used to investigate the relationship between the **education level, employment status, average monthly housing costs, and income of each family, and inferences the causal relationship between income** and these variables (5). In the Methodology section, the model and data used in the analysis will be described. In the result section, the correlation between variables will be shown and analyzed in the form of tables or images. Finally, in the conclusion section, the significance of the results and the aspects in which this report can be improved will be discussed.

Methodology:

The data in this report was collected by a voluntary survey, then the **Multiple Linear Regression model** is using to analyze the data. A multiple linear regression model “is a statistical technique that uses several explanatory variables to predict the outcome of a response variable”(Kenton,2020).

Data

The data that has been used for analysis is part of the **voluntary National Household Survey (NHS)** from Statistics Canada (5). Among them, it is aimed at people aged 15 and above. So all eligible populations are the frame population, and the sample population is the people who actually responded. Besides, the case of inheritance or lottery prizes is excluded(2).

The collected data is according to the way of filling out online questionnaires. Respondents were asked permission to use income information in their tax files. This method is to increase the quality of the survey, but if the respondent does not give permission, they will fill out the questionnaire. The advantage of this method is that the respondent can check back their answers and will be notified if there are unanswered questions, which can ensure the completeness of the data. However, this method also has disadvantages. If the wrong income is entered, it will directly affect the error of the final data. For example, the salary of $30000 is incorrectly filled in like $3000. According to people who do not respond, the number of questionnaires will be adjusted to maintain a certain amount of response rate (2).

In the cleaned data set, data.csv, which contains 140 observations and 6 variables (5). The 6 variables are **NoDiploma, TotalImmigrant, Renter, AverageShelterCosts, Unemployed and Income**. The variable **NoDiploma** show the number of people who do not have a degree and diploma. The variable **TotalImmigrant** shows the number of immigrants. The variable **Renter** is the number of the renter. The variable **AverageShelterCosts** is the number of shelter costs on average, which is calculated together with the main payment of the house and the rent. The variable Unemployed is the nubmber of unemployed. The variable **Income** is the total number of income, which includes after-tax income and other income such as employment income (5).

The following table summarizes the basic characteristics of the data

Table 1: The Summary of Variables

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
NoDiploma	140	2,721.214	1,856.839	340	1,442.5	3,620	9,295
TotalImmigrant	140	1,546.786	1,290.024	130	531.2	2,156.2	6,650
Renter	140	3,400.679	2,396.059	200	1,808.8	4,166.2	13,640
AverageShelterCosts	140	1,019.793	219.622	631	878.5	1,124.8	2,388
Unemployed	140	934.464	531.798	215	565	1,240	3,310
Income	140	62,730.460	21,441.670	32,172	50,096	69,553.8	208,674

The **mean** number of people without a diploma or degree is about 2721, the **minimum** number is 340, the **maximum** number is 9295, the **standard deviation** is 1857. The **mean** value of the total immigrant is 1547, the **minimum** value is 340, the **maximum** value is 6650, the **standard deviation** is 1290. The **mean** value of the renter is 3401, the **minimum** value is 200, the **maximum** value is 13640, the **standard deviation** is 2396. The **mean** value of average shelter costs is 1020, the **maximum** value is 3301, the **minimum** value is 641. The ****mean value for the unemployed is 934, the **minimum** value is 215, the **maximum** value is 3310. The **mean** value for youth is 2382, the **minimum** value is 730, the **maximum**value is 7580. The **mean** income is 62730, the **minimum** income is 32172, the **maximum** income is 208674 (5).

Model

The data is from the website of **Toronto Open Data Portal**, and the **census data** is about **Wellbeing-Toronto**, and it was collected in **2011** by NHS (5). The data is cleaned and analyzed by **RStudio**. After cleaning the data, there are **six** variables are left, which are **NoDiploma, TotalImmigrant, Renter, AverageShelterCosts, Unemployed, and Income**. To explain the variables, income is the explanatory variable and others are the response variable.

The reason for choosing this model is that there is a **linear relationship** between the two types of variables, and the results of income can be predicted clearly based on variables such as **education levels or the average shelter costs** and so on. Before the analysis, most people believe that if a person does not have a degree or diploma, the chance of low income will increase, and unemployment will also directly affect income.

The formula is $\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_{i,NoDiploma} + \hat{\beta}_2 x_{i,TotalImmigrant} + \hat{\beta}_3 x_{i,Renter} + \hat{\beta}_4 x_{i,AverageShelterCosts} + \hat{\beta}_5 x_{i,Unemployed}$. The P-values and R^2 are very important and it will be analyzed in detail.

The performance of this model is significant since the **P-value** is small, therefore, it can be used to explain the relationship between the **NoDiploma, TotalImmigrant, Renter, AverageShelterCosts, Unemployed, and Income**.

The caveats are:

1. This survey is voluntary and is not mandatory, so there will be non-respondents. Therefore, there may not be enough data, resulting in unreliability.
2. The data is only classified according to different regions of Toronto, so it does not represent other regions. The data does not apply to the world.
3. Since the voluntary survey itself is biased, the comparability with other data may be weak.

Result

Table 2: Multiple Linear Regression

	Dependent variable:
	Income
NoDiploma	-0.756
	(1.031)
TotalImmigrant	-7.248^{***}
	(1.452)
Renter	-3.384^{***}
	(0.575)
AverageShelterCosts	70.674^{***}
	(5.293)
Unemployed	17.128^{***}
	(4.915)
Constant	-572.332
	(6,250.646)
Observations	140
R^2	0.685
Adjusted R^2	0.673
Residual Std. Error	12,259.810 (df = 134)
F Statistic	58.234^{***} (df = 5; 134)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Coefficients	t value	P-value
(intercept)	-0.092	0.927181
NoDiploma	-0.733	0.464933
TotalImmigrant	-4.991	1.84e-06
Renter	-5.885	0.000164
AverageShelterCosts	13.353	< 2e-16
Unemployed	3.485	0.000666

An **estimated multiple linear regression model** between response variable and explanatory variables is:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,NoDiploma} + \hat{\beta}_2 x_{i,TotalImmigrant} + \hat{\beta}_3 x_{i,Renter} + \hat{\beta}_4 x_{i,AverageShelterCosts} + \hat{\beta}_5 x_{i,Unemployed}$$

$$= -572.3323 - 0.7558 x_{i,NoDiploma} - 7.2484 x_{i,TotalImmigrant} - 3.3836 x_{i,Renter} + 70.6735 x_{i,AverageShelterCosts} + 17.128 x_{i,Unemployed}.$$

The **standard error** of $\hat{\beta}_0$ is **6250.6462**.

The **standard error** of $\hat{\beta}_1$ is **1.0314**.

The **standard error** of $\hat{\beta}_2$ is **1.4524**.

The **standard error** of $\hat{\beta}_3$ is **0.575**.

The **standard error** of $\hat{\beta}_4$ is **5.2925**.

The **standard error** of $\hat{\beta}_5$ is **4.9149**.

For β_0:

Assume $H_0 : \beta_0 = 0$, $H_a : \beta_0 \neq 0$. The **p-value**= 0.927181, which is **larger** than $\alpha = 0.05$. Therefore, it has very weak evidence against H_0 , which means **fail to reject** H_0. Thus, $\beta_0 = 0$.

For β_1:

Assume $H_0 : \beta_1 = 0$, $H_a : \beta_1 \neq 0$. The **p-value**= 0.464933, which is **larger** than $\alpha = 0.05$. Therefore, it has very weak evidence against H_0 , which means **fail to reject** H_0. Thus, $\beta_1 \neq 0$, which means there is not a relation between x_1 and y.

For β_2:

Assume $H_0 : \beta_2 = 0$, $H_a : \beta_2 \neq 0$. The **p-value**= 1.84e-06, which is **smaller** than $\alpha = 0.05$. Therefore, it has very strong evidence against H_0 , which means **reject** H_0. Thus, $\beta_2 \neq 0$, which means there is a relation between x_2 and y.

For β_3:

Assume $H_0 : \beta_3 = 0$, $H_a : \beta_3 \neq 0$. The **p-value**= 3.03e-08, which is **smaller** than $\alpha = 0.05$. Therefore, it has very strong evidence against H_0 , which means **reject** H_0. Thus, $\beta_3 \neq 0$, which means there is a relation between x_2 and y.

For β_4:

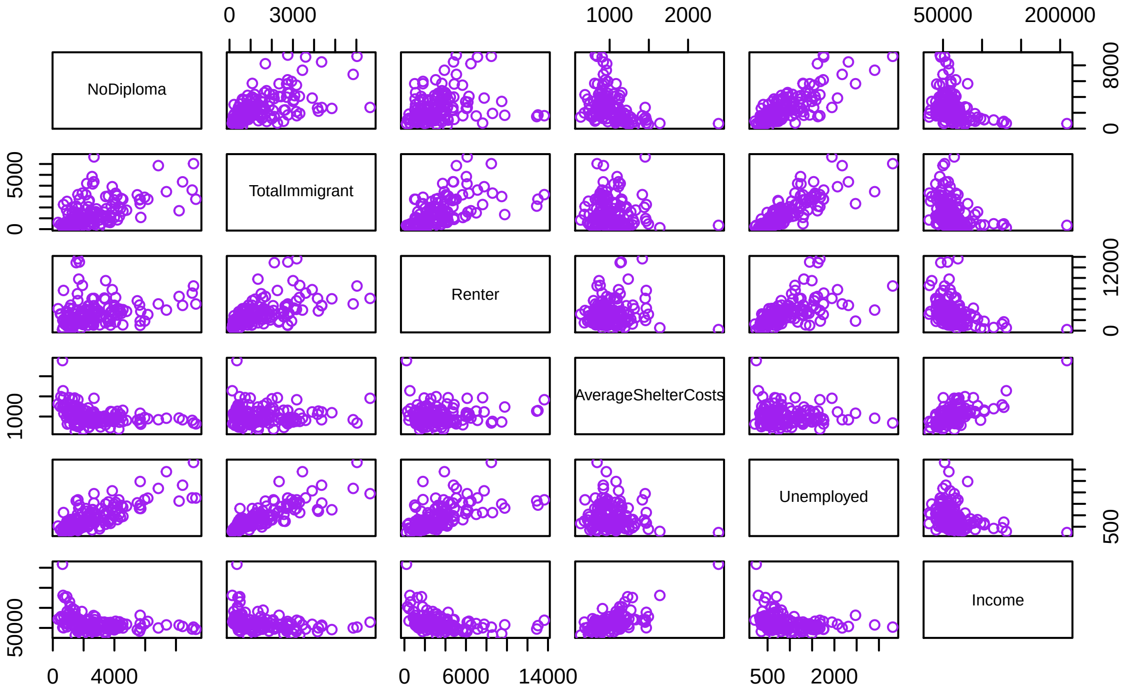
Assume $H_0 : \beta_4 = 0$, $H_a : \beta_4 \neq 0$. The **p-value** < 2e-16, which is **smaller** than $\alpha = 0.05$. Therefore, it has very strong evidence against H_0 , which means **reject** H_0. Thus, $\beta_4 \neq 0$, which means there is a relation between x_2 and y.

For β_5:

Assume $H_0 : \beta_5 = 0$, $H_a : \beta_5 \neq 0$. The **p-value** = 0.000666, which is **smaller** than $\alpha = 0.05$. Therefore, it has very strong evidence against H_0 , which means **reject** H_0. Thus, $\beta_5 \neq 0$, which means there is a relation between x_5 and y.

The R^2 of this estimated model is **0.6848**, which means 68.48% of the total variation in y is explained by the regression model.

Figure 1: Scatter Plot of All the Variables and Income



The first graph is **The Scatter Plot of NoDiploma, TotalImmigrant, Renter, AverageShelterCosts, Unemployed and Income**. From the overall graph, the plots show that **NoDiploma** is positive associated with the Unemployed, the **TotalImmigrant** is positive assocaited with the **Unemployed**, Income is positive **assocaited** with **AverageShelterCosts**. On the other hand, the variable **NoDiploma**, **TotalImmigrant**, **Renter** and **Unemployed** are relatively negative associated with the **Income**. All choosing variables and income are relatively negatively correlated except for **AverageShelterCosts**. To explain, when the above variables such as the number of **NoDiploma** increase, **income** will fall.

It can be analyzed from this that if people's **education level** is not high enough, it will affect the employment situation, and if it is serious, it may lead to **unemployment**. Then the income situation will not be optimistic. If a person has a family, he or she may have a heavier burden. Since to a certain extent, she or he needs to pay monthly rent or other places where the money is needed.

In a more in-depth analysis, when such low-income families are caused by education or employment problems, the probability of creating an ideal living environment for their children will also decrease.

Discussion

Overall, the final multiple linear regression model is:
$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_{i,NoDiploma} + \hat{\beta}_2 x_{i,TotalImmigrant} + \hat{\beta}_3 x_{i,Renter} + \hat{\beta}_4 x_{i,AverageShelterCosts} + \hat{\beta}_5 x_{i,Unemployed}$
$= -572.3323 - 0.7558 x_{i,NoDiploma} - 7.2484 x_{i,TotalImmigrant} - 3.3836 x_{i,Renter} + 70.6735 x_{i,AverageShelterCosts} + 17.128 x_{i,Unemployed}$. The $\hat{\beta}_0$ intercept here is negative, therefore, it does not have real meaning in this formula. $\hat{\beta}_1$ is -0.7558 which means the amount of income will be decreasing 0.7558 as the number of NoDiploma increasing by 1 if other variables are maintain the same value. $\hat{\beta}_2$ is -7.2484 which means the amount of income will be decreasing 7.2484 as the number of TotalImmigrant increasing by 1 if other variables are maintain the same value. $\hat{\beta}_3$ is -3.3836 which means the amount of income will be decreasing 3.3836 as the number of Renter increasing by 1 if other variables are maintain the same value. $\hat{\beta}_4$ is 70.6735 which

means the amount of income will be increasing 70.6735 as the number of AverageShelterCosts increasing by 1 if other variables are maintain the same value. $\hat{\beta}_5$ is 17.128 which means the amount of income will be increasing 17.128 as the number of Unemployed increasing by 1 if other variables are maintain the same value.

Through data and model analysis, the coefficient of **NoDipoma**, the coefficient of the **total number of immigrants**, and the coefficient of the number of the **renter** are negative respectively. It shows that these variables will have a **negative** impact on Income. From the very beginning, we predicted that the lack of **education** would negatively affect the income situation. Observed from the data, this is indeed the case. The increase in the number of immigrants may lead to a decline in income because when a city has a large number of new populations, **social competitiveness** may increase. For example, many companies are not worried about not being able to find the employees they need, so they will not give a lot of salaries. Likely, it is just the minimum salary set by the government. Finally, concerning the increase in the number of renters, for some owners, the bills or property fees that need to be paid may increase. Or due to some personal problems of the renters. For example, When forced to be unemployed, as the impact of **Covid-19** nowadays, a large number of renters may demand that the rent be halved, so that part of the income that the owner should get is **reduced**. Based on the above analysis combined with the actual situation, we can predict what factors will cause low-income families or individuals in the future. Then the government can formulate new programs or benefits for those in need. For example, some subsidies are given to low-income families with children, so that the children can grow up **healthily**. And the minimum salary may be increased according to the consumption level of a different region. Through this process, it improves people's well-being, and then it will be meaningful for the development of society.

Weaknesses

Because the questionnaire is for people over **15 years old** in the Toronto area. Many people between 15 and 18 years old may not have income information, so this part of the data may not be the final purpose of the survey. Then the surveyed area is limited, so it cannot be compared with other countries. In the subsequent analysis, **improvements** can be made in terms of age, and data that may have no income under the age of 18 are removed from the sample. So as to make the data more accurate and then achieve the purpose of the **investigation**. Finally, because only a part of the many variables is selected for analysis. Perhaps there are other potential variables that have a more direct relationship with income that can be analyzed.

Next Steps

In the future, some additional investigations can be conducted based on the contents of this report, such as investigating high-income households or individuals which factors make them become this. Or to investigate the needs of high-income or low-income people in terms of quality of life and happiness. This will make it easier to build a **happy and beautiful** society for any group of people in the world.

References

1. Canada, P. (2019, April 24). Government of Canada. Retrieved December 22, 2020, from https://www.canada.ca/en/public-health/services/publications/science-research-data/inequalities-children-low-income-families-infographic.html

2. Government of Canada, S. (2018, July 25). Income Reference Guide, National Household Survey, 2011. Retrieved December 22, 2020, from https://www12.statcan.gc.ca/nhs-enm/2011/ref/guides/99-014-x/99-014-x2011006-eng.cfm

3. Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

4.Kenton, W. (2020, September 21). How Multiple Linear Regression Works. Retrieved December 22, 2020, from https://www.investopedia.com/terms/m/mlr.asp

5. Open Data Dataset. (n.d.). Retrieved December 22, 2020, from https://open.toronto.ca/dataset/wellbeing-toronto-demographics-nhs-indicators/

6.Schork, J. (2020, December 01). R pairs & ggpairs Plot Function: 5 Examples (Color, Labels, by Group). Retrieved December 22, 2020, from https://statisticsglobe.com/r-pairs-plot-example/

7.Shelter-cost-to-income ratio. (2016, January 04). Retrieved December 22, 2020, from https://www12.statcan.gc.ca/nhs-enm/2011/ref/dict/households-menage028-eng.cfm