CVEN9407
Transport Modelling
Final Report

Yan Zhou
z5107253

Table of Content

1. Data analysis

1.1 Data Characteristics

The student number is z5107253 and the dependent variable is losat (how satisfied are you with your life).

The data frame is a table with 5099 rows and 114 columns, covering from family member, big events in the past year, gender, education, occupations, etc. The data for 2006 is complete and there is no missing data. So the total number of 5099 people's information can be used for analysis. Large tables are attached as Appendix.

1.1.1 Variable Definitions

The definitions of each column in the dataset is shown in the Table 1 in Appendix.

1.1.2 Continuous variables mean and standard deviations

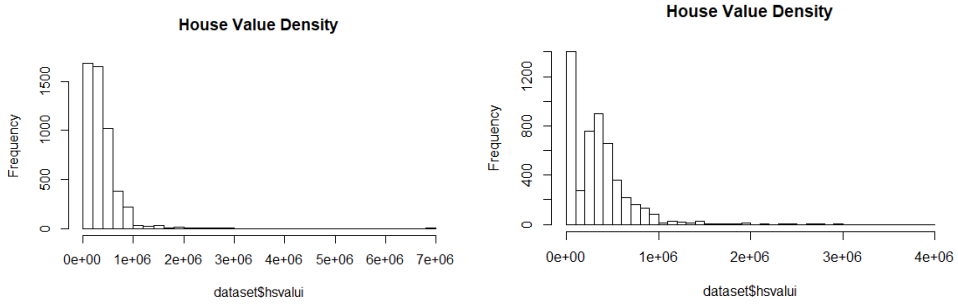
In this dataset, the continuous variables are ages, time spent on activities, household income, and expenditures. Mean value of each column represents the average value of this column, and standard deviation represents how concentrated are the variables. The results are shown in Table 2 in Appendix.

1.1.3 Discrete Variables Frequencies

In this dataset, the discrete variables are gender, family types occupations, education. These variables are mainly used to describe the person's identity. The definitions of the variables are given in the project outline. The discrete variables aim at analyzing whether factors including big events and choices in the past year, relative socio-economic standard, education and occupation and economic resources have close relationship with people's life satisfaction in the past year. The results are shown in Table 3 in Appendix.

1.1.4 Variables Variation

According to the mean and standard deviation table above, the standard deviations of different variables show great differences. Among all the variables, the standard deviation for hsvalui (house value) is the highest. The density of house value in this project is shown in the figures below. The figure on the right is the histogram of house value density without zero objects.



As is shown above, the variations in the two variables are large, and most people do not have own house. After deleting the data of zero and only considering those who have their own

property, majority of participants have properties that are cheaper than 1 million dollars.

Apart from the house value, the financial year disposable regular income and household financial year disposable income also have large standard deviation values. This indicates that among all the variables provided, income and spending for a family and for a single person show the greatest differences among people. The income gap between the rich and the poor is still wide.

This problem is similar with the range of variables. The variables having wide ranges are always expected to have large variances in comparison with the mean value, which also indicates that this variable varies significantly among different participants.

1.1.5 Potential Issue of the Dataset

For most variables included in the project, the values vary from 0 to 10. But for some monetary variables, the mean values and standard deviations are much greater. This may have an impact on establishing linear or discrete choice model, and it should be log-transformed to reduce the impact.

1.2 Variables Correlation

In this project, the number of total variables is more than 5000, indicating it difficult to discuss all the variables in analysis. The correlation coefficients are taken into account in choosing the independent variables, for the variables with strong correlation coefficients should not be used in modelling at the same time.

1.2.1 Variable Correlation Matrix

According to the correlation matrix, several pairs of variables have relatively strong relationships. In this project, variables with relation coefficients over 0.5 or less than 0.5 are considered to have strong correlation. The results are shown in the table below, and the strong coefficients are marked in red.

	hh0_4	hhadult	hifdip	hsvalui	hgage	tifdip	hhfty	losatlc	hhyng	hhold	lschd	Married	Le_bth	Le_prg	CoupleWd	CoupleW	Single	Renter
hh0_4	1	-0.09	-0.02	-0.01	-0.12	0.05	-0.16	-0.01	-0.45	-0.24	0.53	0.21	0.49	0.34	-0.21	0.33	-0.14	0.00
hhadult	-0.09	1	0.39	0.22	-0.14	-0.12	-0.29	0.05	-0.29	0.29	-0.11	0.08	-0.05	-0.06	-0.16	0.50	-0.51	-0.22
hifdip	-0.02	0.39	1	0.37	0.02	0.55	-0.26	0.05	-0.09	0.15	-0.02	0.16	0.01	0.02	0.03	0.25	-0.31	-0.22
hsvalui	-0.01	0.22	0.37	1	0.18	0.20	-0.18	0.13	-0.03	0.30	0.03	0.13	-0.02	-0.05	-0.05	0.22	-0.19	-0.55
hgage	-0.12	-0.14	0.02	0.18	1	0.29	-0.06	0.11	0.44	0.59	-0.03	0.30	-0.10	-0.12	0.13	-0.07	0.03	-0.24
tifdip	0.05	-0.12	0.55	0.20	0.29	1	-0.06	0.00	0.05	0.02	0.02	0.20	0.04	0.04	0.06	-0.01	0.01	-0.08
hhfty	-0.16	-0.29	-0.26	-0.18	-0.06	-0.06	1	-0.07	0.18	-0.04	-0.15	-0.68	-0.09	-0.12	-0.51	-0.32	0.74	0.25
losatlc	-0.01	0.05	0.05	0.13	0.11	0.00	-0.07	1	0.00	0.12	0.03	0.07	-0.01	-0.03	-0.03	0.10	-0.05	-0.18
hhyng	-0.45	-0.29	-0.09	-0.03	0.44	0.05	0.18	0.00	1	0.39	-0.46	-0.15	-0.24	-0.18	0.55	-0.67	0.35	0.01
hhold	-0.24	0.29	0.15	0.30	0.59	0.02	-0.04	0.12	0.39	1	-0.18	-0.08	-0.17	-0.20	-0.01	0.08	-0.14	-0.37
lschd	0.53	-0.11	-0.02	0.03	-0.03	0.02	-0.15	0.03	-0.46	-0.18	1	0.21	0.28	0.20	-0.29	0.36	-0.14	-0.06
Married	0.21	0.08	0.16	0.13	0.30	0.20	-0.68	0.07	-0.15	-0.08	0.21	1	0.14	0.14	0.34	0.27	-0.46	-0.17
Le_bth	0.49	-0.05	0.01	-0.02	-0.10	0.04	-0.09	-0.01	-0.24	-0.17	0.28	0.14	1	0.54	-0.11	0.18	-0.07	0.03
Le_prg	0.34	-0.06	0.02	-0.05	-0.12	0.04	-0.12	-0.03	-0.18	-0.20	0.20	0.14	0.54	1	0.00	0.09	-0.07	0.03
CoupleWd	-0.21	-0.16	0.03	-0.05	0.13	0.06	-0.51	-0.03	0.55	-0.01	-0.29	0.34	-0.11	0.00	1	-0.61	-0.22	0.04
CoupleW	0.33	0.50	0.25	0.22	-0.07	-0.01	-0.32	0.10	-0.67	0.08	0.36	0.27	0.18	0.09	-0.61	1	-0.37	-0.27
Single	-0.14	-0.51	-0.31	-0.19	0.03	0.01	0.74	-0.05	0.35	-0.14	-0.14	-0.46	-0.07	-0.07	-0.22	-0.37	1	0.24
Renter	0.00	-0.22	-0.22	-0.55	-0.24	-0.08	0.25	-0.18	0.01	-0.37	-0.06	-0.17	0.03	0.03	0.04	-0.27	0.24	1

1.2.2 Potential reasons behind strong correlations

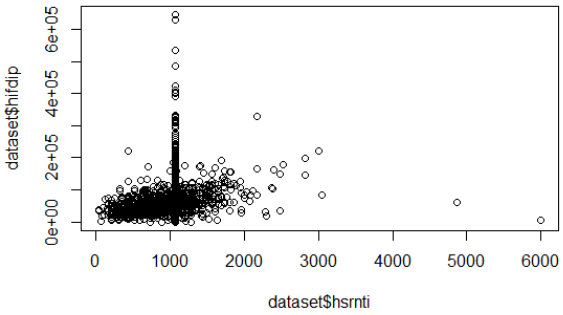
Take two pairs of variables in this part as examples. From the table above, the number of children between 0-4 years old is closely connected with the time parents spent playing with children. Although the number of kids between 0-4 years old is not a typical continuous variable, the increase in the number of young kids and infants in a family is expected to be related to the increase in the time parents spent on playing with kids.

The strongest correlation in this project is the relationship between Single and hhfty (family type). However, "single" is also a variable indicating family structure, so these two variables

are actually similar ones. When considering the independent variables to discuss in modelling, these pair of variables should not be chosen at the same time.

1.2.3 Discussion

According to my own experience, house rent spending may be closely connected to the money a family earn. Those who have higher income are always considered successful and may be more willing to have a better place to live when it is affordable.



However, as is shown from the scatter plot, a large number of people spend around $1000 for the accommodation. There are several people spending over $3000/month which is relatively high, but most pay less than $2000/month.

A potential reason for this phenomenon is that the house loan interest in Australia is low in comparison with other countries, so it will be more reasonable and considerate to apply for buying a house instead of spending large amount of money for renting even if it is affordable.

Another pair is jompf (I get paid fairly for the things I do in my job) and jbmsall (Overall job satisfaction). It is not hard to imagine that those who considered themselves fairly-paid will have more passion in working, and they will be more satisfied for their jobs. However, the correlation coefficient is less than 0.4 which is not a strong relationship. The frequencies of each pair are shown in the table below.

		jbmsall										
		0	1	2	3	4	5	6	7	8	9	10
jompf	1	1	5	4	10	11	34	27	43	46	26	19
	2	3	2	6	18	22	44	54	112	101	51	19
	3	3	1	6	7	13	61	88	163	185	75	35
	4	4	3	5	6	19	60	87	215	273	125	52
	5	0	1	1	11	19	38	98	270	407	196	66
	6	1	3	3	7	9	35	63	173	483	441	134
	7	0	1	2	3	3	17	17	51	107	144	166

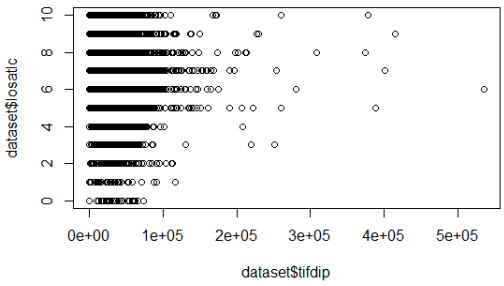
According to the table, there is a general trend that people tend to feel more satisfied about their jobs when they consider themselves fairly-payed. However, this is not applicable for all the cases. Job satisfaction reaches a peak when jompf is 4 or 5, and then decreases.

One of the reasons of this phenomenon might be that those who consider themselves fairly payed do the well-payed but challenging work for a living, either physically or mentally. They work hard to earn reasonable and decent salary, but feel unsatisfied by the toil and pressure that come with the challenging job.

1.2.4 Variable Transformation

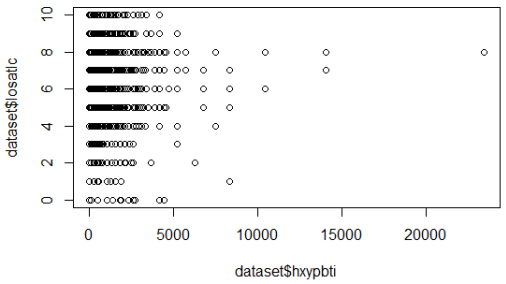
From the correlation matrix, the correlation coefficients of tifdip (Financial year disposable regular income) and the dependent variable, although there is expected to have a relationship between the two variables.

As is shown in the figure, people tend to feel more satisfied with higher income before 100 thousand, but to achieve high life satisfaction, high income is not necessary. The relationship between the two factors is expected to be $losatlc \sim ln(tifdip)$.



.

Similarly, hxypbti (Household annual expenditure - Public transport and taxis) is expected to have negative relationship with life satisfaction, for more money spent on travelling indicates longer commuting time and less disposable regular income. The relationship between the two variables is similar to the one above, but much flatter. The relationship is expected to be $losatlc \sim \sqrt{hxypbti}$.



1.2.5 Updated Correlation Coefficients

According to the discussion above, the new correlation matrix is attached as appendix. The changes of correlation coefficients of the two pairs are shown in the table below, and the whole matrix is attached as Appendix.

Item	Before	After
losatlc ~ tifdip	7.58E-05	-0.02015
losatlc ~ hxypbti	-0.04725	-0.0131

The absolute values of correlation coefficients of losatlc ~ tifdip increased after the step, but that of losatlc~hxypbti decreased. Although the steps taken for the first pair makes the

relationship between the two variables more obvious, the coefficient is not enough for determining a strong relationship between them.

2. Linear Regression Modelling

2.1 Sub-datasets Division

In this project, dataset is divided into test dataset accounting for 80% of the observations, and test dataset for the remaining 20%. Train dataset is used for determining parameters in the models, while the test dataset is used for validating the accuracy of the model.

In order to avoid a random selection, the student number 5107253 is used as the seed number, and the confidence level is 95%. And the dependent variable is losatlc standing for how satisfied are you with your life.

2.2 One-variable Regression Modelling

In order to determine the relationship between transport related variables and the level of satisfaction, one-variable regression models of transport related variables are run firstly to see whether the relationship between the dependent variable and a certain transport related variable is significant.

The transport related variables in this study are lscom (Combined hrs/mins per week - Travelling to/from paid employment), hxymvfi (Household annual expenditure - Motor vehicle fuel ($)), hxymvri (Household annual expenditure - Motor vehicle repairs/maintenance ($)), hxyncri (Household annual expenditure - New motor vehicles, motorbikes or other vehicles ($)), hxypbti (Household annual expenditure - Public transport and taxis ($)). Linear models are run separately for each of the dependent variables with the independent variable and the coefficients and significances are recorded and compared.

In the table below, R Square stands for how well the model fits, while adjusted R Square taking the number of coefficients into account. P-value indicates the possibility of the coefficients being zero which indicates this variable has no impact on the dependent variable.

The comparison of the five models is shown in the table below.

One-variable regression							
Variables	coefficients		Standard Error	t-Stat	p-value	R Square	Adjusted R Square
lscom	(Intercept)	6.887989	0.043761	157.401	<2e-16	0.008478	0.008235
	lscom	-0.047251	0.008003	-5.904	3.83E-09		
hxymvfi	(Intercept)	6.65E+00	4.03E-02	165.007	<2e-16	0.001304	0.001059
	hxymvfi	1.98E-05	8.59E-06	2.308	0.0211		
hxymvri	(Intercept)	6.63E+00	4.21E-02	157.499	<2e-16	0.001489	0.001244
	hxymvri	6.70E-05	2.72E-05	2.466	0.0137		
hxyncri	(Intercept)	6.70E+00	3.22E-02	208.214	<2e-16	0.0001122	-0.0001331
	hxyncri	2.30E-06	3.40E-06	0.676	0.499		
hxypbti	(Intercept)	6.75E+00	3.35E-02	201.149	<2e-16	0.002386	0.002142
	hxypbti	-1.06E-04	3.41E-05	-3.123	0.0018		

According to the table above, the increase in life satisfaction in this model is related to the decrease of commuting time and household expenditure on public transport and taxis. The increase in household annual expenditure on motor vehicle fuel, motor vehicle repairs/maintenance and new motor vehicles, motorbikes or other vehicles also makes people feel more satisfied.

As for the goodness-of-fit, for each transport related independent variable, the R^2 and R^2_{adj} value are both very small and we can hardly determine that there is a significant linear relationship between any single independent variable and the dependent variable.

We can also tell from the table that hxyncri may have little relationship with the dependent variable, for the p-value of the scaling is 0.499, which is too big to reject the null hypothesis that the coefficient is zero, which means that the independent variable has no effect on the dependent variable.

2.3 Multiple-variable Regression Modelling

There are 31 combinations of the transport related variables, five of which have been discussed in the part above. The model that fits the best among the 31 models should have the highest goodness-of-fit and all of the coefficients included should be statistically significant. According to the criterion, the best model in this step is:

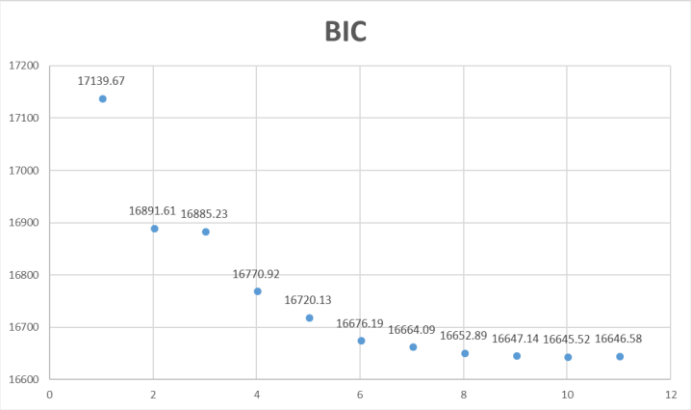
Variables	coefficients		Standard Error	t-Stat	p-value	R Square	Adjusted R Square
lscom + hxymvri + hxypbti	(Intercept)	6.83E+00	5.22E-02	130.85	< 2e-16	0.01147	0.01074
	lscom	-4.50E-02	8.09E-03	-5.562	2.84E-08		
	hxymvri	7.28E-05	2.71E-05	2.692	0.00713		
	hxypbti	-7.93E-05	3.43E-05	-2.308	0.02107		

2.4 Add new independent variables

There are two types of binary variables among the independent variables one type of which is referred to as Socio Economic Indexes for Areas (SEIFA). Instead of including them in the model as we do to other independent variables, SEIFA variables tend to indicate a cut-off point for each category which has the most significant impact on the regression model, so only one variable in one category can be chosen in the model.

The selection of new variables using stepwise forward method for each step is based on Bayesian Information Criterion (BIC) index. One independent variable is added into the model in each step which improves the BIC the most and is statistical significant as well. Finally, the best model we select is the model which minimize BIC and maximizes R^2_{adj}.

Before adding new variables, the BIC index of the transport related model is 17139.67, and the selection of SEIFA variables is conducted after the selection of other variables. The steps we had and the change of BIC after each step are shown in the graph and table below.



	Step	BIC
0	-	17139.67
1	+ jbmsall	16891.61
2	- hxypbti -hxymvri	16885.23
3	+ Renter	16770.92
4	+ lsvol	16720.13
5	+ gh1	16676.19
6	+ hgage	16664.09
7	+ CoupleW	16652.89
8	+ Le_fnw	16647.14
9	+ Mltpljob	16645.52
10	+ hsbedrm	16646.58

After the continuous decreasing of BIC, there is an increase in BIC after adding the variable of hsbedrm (the number of bedrooms in the house), so this is the place where we stop adding new variables, and the regression model after adding these variables is:

variables	coefficients		Std. Error	t value	Pr(>|t|)	R Squared	Adjusted R Squared
lscom + jbmsall + Renter + lscol + gh1 + hgage + CoupleW + Le_fnw + Mltpljob	(Intercept)	4.957949	0.19843	24.986	< 2e-16	0.1349	0.133
	lscom	-0.039622	0.007511	-5.275	1.40E-07		
	jbmsall	0.259339	0.018391	14.102	< 2e-16		
	Renter	-0.563234	0.071298	-7.9	3.57E-15		
	lsvol	0.084099	0.011719	7.176	8.46E-13		
	gh1	-0.260205	0.034331	-7.579	4.27E-14		
	hgage	0.012273	0.002357	5.207	2.02E-07		
	CoupleW	0.265236	0.060438	4.389	1.17E-05		
	Le_fnw	-0.823375	0.218195	-3.774	0.000163		
	Mltpljob	0.316138	0.10036	3.15	0.001644		

After this, SEIFA variables are added into the model. There are four kinds of SEIFA variables which are hhad10_X, hhda10_X, hhed10_X and hhec10_X. For each column of SEIFA variables, 0 and 1 are used to indicate specific characteristic of the observation, so using one column in one category at a time naturally separates the dataset into two parts. SEIFA variables are added into the model in a similar process which is used before, but once one variable in a category is chosen, all of the other variables of this category are not to be considered any more.

In this step, hhda10_4 and hhec10_5 are chosen, and the regression model after this step is:

Variables	coefficients		Std. Error	t-Stat	p-value	R Square	Adj R Square
lscom + jbmsall + Renter + lsvol + gh1 + hgage + CoupleW + Le_fnw + Mltpljob + hhda10_4 + hhec10_5	(Intercept)	4.987603	0.198579	25.116	< 2e-16	0.1431	0.1408
	lscom	-0.038311	0.0075	-5.108	3.4E-07		
	jbmsall	0.257585	0.018315	14.064	< 2e-16		
	Renter	-0.534978	0.071122	-7.522	6.6E-14		
	lsvol	0.083498	0.011668	7.156	9.81E-13		
	gh1	-0.24336	0.034396	-7.075	1.75E-12		
	hgage	0.011389	0.002351	4.845	1.32E-06		
	CoupleW	0.256883	0.060275	4.262	0.0000207		
	Le_fnw	-0.823588	0.217303	-3.79	0.000153		
	Mltpljob	0.31974	0.09991	3.2	0.001384		
	hhda10_4	-0.494979	0.079396	-6.234	5E-10		
	hhec10_5	0.277338	0.07403	3.746	0.000182		

Although the values of R^2 and R^2_{adj} are still very small, the model has improved greatly in comparison with that in former parts. BIC value for the model reduces to 16623.33.

It is worth mentioning that two of the transport related variables, hxymvri and hxypbti are deleted due to the stepwise variable selection criterion. Because of the necessity of being cautious in deleting transported-related variables, whether these two variables could be added into the model again should be discussed. The comparison of the three models (containing one of the two variables and two of them) is shown in the table below.

Variable to Add	p-value		R Square	Adj R Square	BIC
-	-	-	0.1431	0.1408	16623.33
hxymvri	hxymvri	0.774136	0.1431	0.1406	16631.56
hxypbti	hxypbti	0.096744	0.1437	0.1411	16628.88
hxymvri + hxypbti	hxymvri	0.097708	0.1437	0.141	16637.12
	hxypbti	0.795546			

According to the table above, R^2 and R^2_{adj} values do not increase greatly after adding the variables, BIC values for the three models all increase, and the p-values for the two variables in all of the three models are too big, indicating they are not statistically significant. So none of the two variables should be included into the final model.

2.5 Variable Justification

The model shown above includes 11 significant variables the definitions of which are illustrated in the former report. Justification of the variables is discussed below.

lscom: Combined hrs/mins per week - Travelling to/from paid employment. The longer time a person spends on commuting, the farther the person lives from work. Spending long time on commuting indicates less time for entertainment even sleeping, and people tend to feel tired. So the increase in commuting time is expected to cause the decrease in life satisfaction.

jbmsall: overall job satisfaction. Feeling satisfied about a person's job indicates that working makes the person feel enriched and satisfied, which contributes to the person's positive attitude towards life.

Renter: Binary variable indicating if the individual is renting his/her living place. Spending money on house rent indicates less money on entertainment or for saving. So it has negative impact on people's life satisfaction.

lsvol: Combined hrs/mins per week - Volunteer/Charity work. Doing charity and volunteer work possibly makes a person feel enriched and happy. Being willing to do charity volunteer work also indicates that the person is willing to devote himself/herself to others, and they are also capable of doing this, so this person tends to be health, has no financial stress. So the increase in time spent on volunteer/charity work is expected to be related to the increase in life satisfaction.

gh1: self-assessed health. In this project, the lower this value is, the heathier consider himself/herself to be. Being unhealthy makes it hard for people to have any severe physical activities, and they may have to spend more money on medical care, so people feel less satisfied with the increase of the value.

hgage: Age last birthday at June 30 2016. It is hard to define how age influence life satisfaction. It is possible that as a person gets older, it is easier for them to feel content with what he/she has for the abundant experience, and after working for years, they tend to have less financial pressure. So the increase in age is related to the increase in life satisfaction.

CoupleW: Binary variable indicating if family structure is couple with children. Being in this family indicates that this person does not have to support a family himself/herself, and people can also feel content from family life, which makes them more satisfied with their lives.

Le_fnw: Binary variable indicating if the individual has experienced worsening in finance last year. Having financial problems is expected to cause great pressure during people's daily lives as well as the reduction in quality of lives, so this experience is expected to reduce people's life satisfaction toward life.

Mltpljob: Binary variable indicating if the individual is employed in multiple jobs. It is hard to illustrate the reason why being employed by multiple jobs tends to make people feel happy. One possible explanation for it is that for those who have multiple jobs without financial problems, the second or even third job is more like a part-time job which is chosen because of their own hobbits and this makes them happier.

hhda10_4: Binary variable indicating if the 'IRSD' index of the home zone is less than 4.

According to the definition, the Index of Relative Socio-economic Disadvantage (IRSD) is a general socio-economic index that summarises a range of information about the economic and social conditions of people and households within an area [1]. A low value indicates that there are many households with low income, many people with no qualifications, or many people in low skill occupations. The higher the value is, the better the neighborhood is. However, in this model, being in the group of higher value of this index have an impact on the decrease in life satisfaction. It is hard to find a convincing explanation, but one possible way to explain is, those who live in a better place need more spending to maintain their present lives, and they tend to have more daily pressure.

hhec10_5: Binary variable indicating if the 'IER' index of the home zone is less than 5. The Index of Economic Resources (IER) focuses on the financial aspects of relative socio-economic advantage and disadvantage, by summarising variables related to income and wealth [2]. A lower value indicates more households have low income or pay low rent and a higher value shows the opposite. The higher the value is, the relatively greater access the households have to economic resources. Being in the group with higher value of this index indicates that it is relatively easier for this household to earn their livings, and they tend to live in a more developed area, and they may feel more comfortable and satisfied with life.

Discussion

Personally, I consider hgage (age) and Mltpljob (whether employed by multiple jobs) to be unnecessary to be included in the model. In different stages in life, there are different factors that have an impact on people feeling about life-satisfaction such as study, employment or marriage, but age should not influence people's life satisfaction directly, and the absolute value of the coefficient for age is also the smallest among all coefficients, so it has less impact on life satisfaction than other coefficients.

In addition, the impact of being employed by multiple jobs and the impact of the categorical variable hhda10_4 are completely opposite to what we expected and deleting them from the model is considered. But according to the model we had, the absolute value of the coefficient of the latter variable is relatively large in comparison to other variables. Without further information, we cannot decide on a convincing explanation for the impact, but it should be better to include the hhda10_4. Since the p-value for Mtpljob is much higher, it is deleted from the model.

After this step, the final model is decided to be:

losatlc =	5.39044	-0.03664 lscom	+ 0.263 jbmsall	- 0.62605 Renter
Std. Error	(0.1849)	(0.00752)	(0.01831)	(0.06895)
t value	(29.153)	(-4.872)	(14.363)	(-9.08)
p-value	(0.0000)	(0.0000)	(0.0000)	(0.0000)
	- 0.22053 gh1	+ 0.213 CoupleW	-0.79633 Le_fnw	- 0.51677 hhda10_4
Std. Error	(0.0341)	(0.05988)	(0.21809)	(0.07956)
t value	(-6.467)	(3.557)	(-3.651)	(-6.496)
p-value	(0.0000)	(0.0004)	(0.0003)	(0.0000)

2.6 Classic Linear Regression Model Assumptions and Assumptions Testing

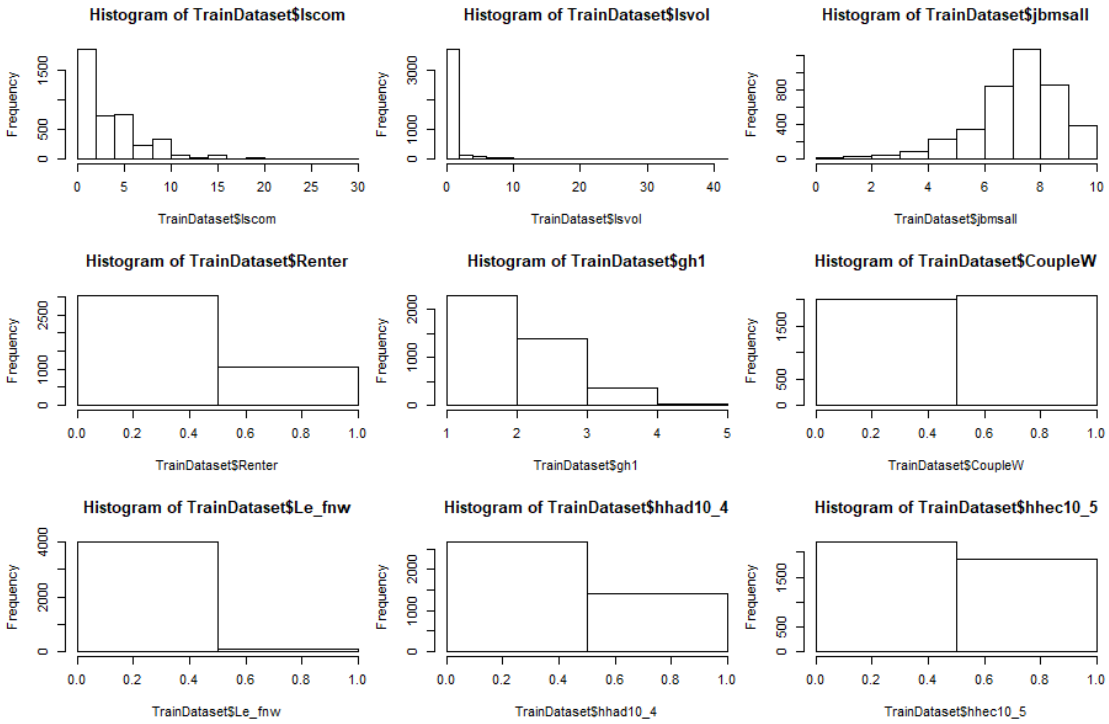
There are 10 assumptions of CLRM that are to be tested.

- *Assumption 1: The regression model is linear in the parameters.*

According the modeling processed conducted in modeling, the model is linear in the parameters and this assumption holds true for this model.

- *Assumption 2: The values of independent variables are fixed in repeated sample, which means the independent variables are assumed to be non-estochastic.*

In this model, most of the independent variables are discrete variables with integer values, so the values of independent variables are fixed in all of the observations. The only two variables that are not discrete variables are lscom and lsvol, and the distributions of all variables are shown in the graph below.



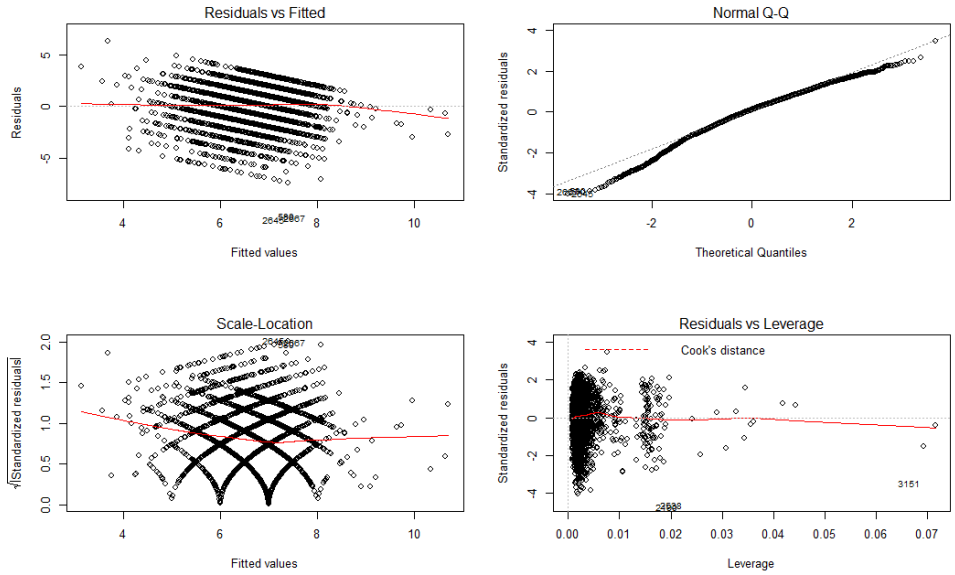
We can tell from the graph that all variables included in the model does not show a random distribution, and we can decide that this assumption holds true in this model.

- *Assumption 3: Zero mean value of disturbance.*

The mean of residual of the final model is -2.550239e-17, which is small enough and the assumption holds true for this model.

- *Assumption 4: Homoscedasticity of equal value of* $\mathbf{u_i}$.

In order to determine whether there is a heteroscedasticity in this model, the change of residuals as fitted value increases is plotted.



According to the plot of residuals vs fitter values the pattern of which is indicated by the red line, the residuals are basically flat. We can also tell this from the scale-location graph. The relationship between the square root of standardized residual and fitted value is basically flat as well, so there is no heteroscedasticity in this model and the condition of homoscedasticity can be accepted.

- *Assumption 5: No autocorrelation between disturbance.*

Durbin-Watson test is run with the model to test the existence of autocorrelation. The null hypothesis of the test is "There is no correlation among residuals", and the alternative hypothesis is "Residuals are autocorrelated".

With DW = 2.0356, and p-value = 0.872, the possibility of null hypothesis is higher than 0.05, so we cannot reject the null hypothesis that residuals are independent. The DW value that is close to 2 also indicates the same finding.

Apart from the Durbin-Watson test, a runs test is conducted with the residuals to test for randomness. The null hypothesis is "the order of the data is random", and the alternative hypothesis is "the order of the data is not random".

Standardized Runs Statistic = 1.6442, p-value = 0.1001. The p-value is higher than 0.05, so we cannot reject the null hypothesis, and the order of the data should be random, so the residuals of the model is independent, and the assumption holds true for this model.

- *Assumption 6: Zero covariance between u_i and X_i.*

Variables	p-value
lscom	1
jbmsall	1
Renter	1
lsvol	1
gh1	1
CoupleW	1
Le_fnw	1
hhda10_4	0.8353
hhec10_5	1

A correlation test is conducted to each independent variable with residuals. The null hypothesis is "true correlation is equal to 0" and the alternative hypothesis is "true correlation is not equal to 0". The p-values of the correlation tests are shown in the table on the left.

We can tell from the table that for each of the independent variables, p-value is larger than 0.05, so null hypothesis that true correlation is 0 can't be rejected, and the assumption holds true for this model.

- *Assumption 7: The number of observations should be greater than the number of parameters to be estimated.*

The number of observations in the train dataset is 4079 and the number of parameters independent variables included in the model is 11. So this assumption holds true for this model.

- *Assumption 8: $var(X)$ must be a finite positive number.*

The variability of all the independent variable is tested and the results are shown in the table below.

Variables	lscom	jbmsall	Renter
var(X)	14.87874	2.574832	0.1901243
Variables	lsvol	gh1	CoupleW
var(X)	6.176746	0.7466248	0.2499627
Variables	Le_fnw	hhda10_4	hhec10_5
var(X)	0.017817	0.2171856	0.2483341

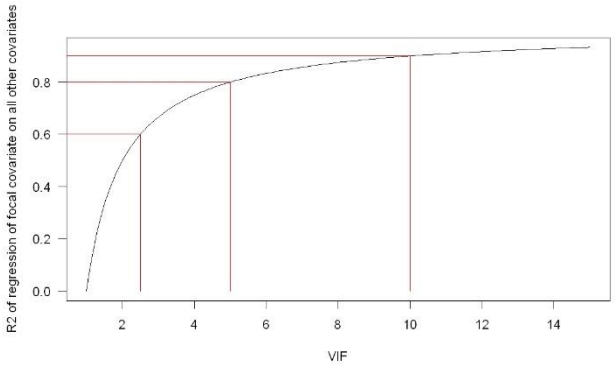
We can tell from the table that the variances of all independent variables in this model are finite positive numbers, so this assumption holds true in this model.

- *Assumption 9: There is no specification bias or error in the regression model.*

In this project, only Classic Linear Regression Model is considered, so we assume this assumption holds true for this model.

- *Assumption 10: There is no perfect multicollinearity between explanatory variables.*

Variance Inflation factor (VIF) is used for testing the multicollinearity between explanatory variables. Since VIF is closely connected with R^2, the strict cut-off point should be decided first.



In this model, the R^2 for the final model is approximately 0.15, and the VIF cutoff point should be approximately 1.5. We can tell from the table below that all of the VIF values in this model are close or less than 1.5, which basically meet the requirement of this assumption.

Variables	lscom	jbmsall	Renter
VIF	1.011228	1.037698	1.086379
Variables	lsvol	gh1	CoupleW
VIF	1.005306	1.043562	1.07729
Variables	Le_fnw	hhda10_4	hhec10_5
VIF	1.01858	1.652257	1.646031

In conclusion, the model we choose for analysis generally meet the requirements for assumptions. However, both R^2 and R^2_{adj} are very small, so the model may not be able to account for majority of the dataset.

2.7 Model Validation

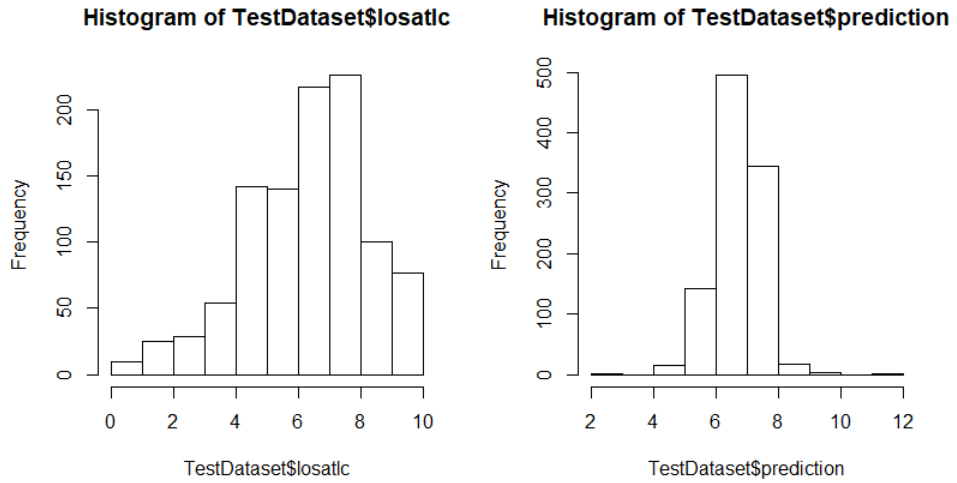
Test dataset which is the remaining 20% of the origin dataset is used for validation. The dependent variable in the test dataset is calculated based on the model we chose. In this part, the methods used to determine how good the model is mean value, comparison of

independent variable distribution with simulation, distribution of predicted residuals and the comparison of variances of simulation and observation.

2.7.1 Mean

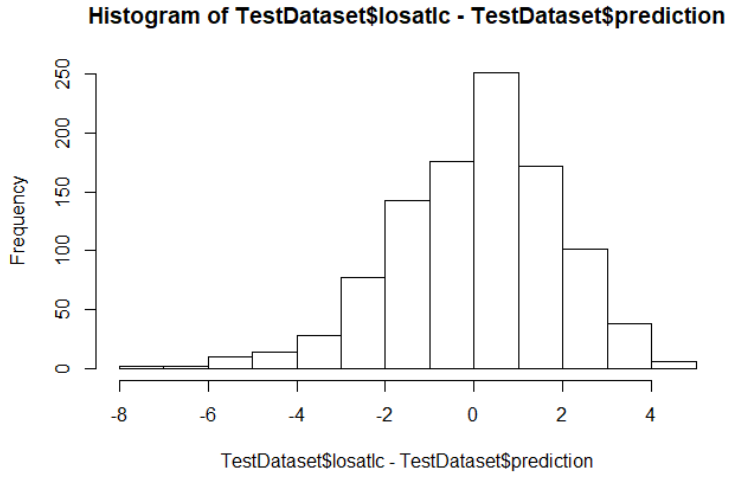
The mean of the dependent variable losatlc of the test dataset is 6.769608, and the mean of the predicted value is 6.713451. The mean value of the predicted independent variable is close to the mean of observations, which indicates the potential application of predicting the mean life satisfaction.

2.7.2 Comparison of Dependent Variable Distribution with Simulation



The graph above shows the comparison of the distribution of the dependent variable of observations and of prediction. The distribution of the prediction is more concentrated than the observations, and the proportion of simulation within the range (6, 8) is 82.25%, while that for observations is only 21.27%. This indicates that the model is not a good one and can hardly explain the majority of the dataset.

2.7.3 Distribution of Predicted Residuals



We can tell from the graph above that the distribution of the simulated residuals is not

uniformly distributed, which indicates that the model is not a good one and using it for prediction results in large errors.

2.7.4 Comparison with Variance

The variance of prediction is 0.559 while the variance of observations is 3.899. This indicates that the distribution of the prediction is much more concentrated than the observations, which is in accordance with our former findings.

2.8 Conclusion

Generally speaking, although this model meets the requirements of assumption testing, it does not fit the data well and can hardly be used for future prediction. The relationship between the dependent variable and the independent variable is illustrated completely by personal experience, and it may not be critical and may not be in accordance with the reality. Apart from that, linear regression cannot simulate discrete dependent variables which is exactly the case in this project. It is hard to decide whether using linear regression model is proper in this project.

In short, this model does not fit the data well. More information is needed for improvement.

3. Discrete Choice Modelling

As is discussed in the former part, linear model is not suitable for modelling and estimating in this project, thus in this part, discrete choice model is used for estimation. Without further information, the decision of the best model in this part is chosen from ordered logit model and multinomial logit model.

3.1 Choice Definition and Utility Function

To simplify the model, the dependent variable is aggregated into three categories indicating dissatisfaction, medium and complete satisfaction based on the value of the dependent variable. Dissatisfaction covers the dependent variable from 0 to 4, medium covers from 5 to 7 and complete satisfaction covers the rest from 8 to 10.

Train dataset is used for estimating the parameters, and test dataset is used for model validation, and this is consistent with what we did for the linear model. Before the modelling process, whether the train dataset and test dataset are representative enough should be confirmed. To be clearer, the comparison of the proportions of each satisfaction level and the difference in proportions in comparison with the whole dataset are shown in the table below.

Satisfaction Level	Whole Dataset	Train Dataset	Test Dataset	$\frac{\lvert\text{Train Dataset} - \text{Whole Dataset}\rvert}{\text{Whole Dataset}}$	$\frac{\lvert\text{Test Dataset} - \text{Whole Dataset}\rvert}{\text{Whole Dataset}}$
Complete Satisfaction (2)	38.18	37.85	39.51	0.86%	3.48%
Medium Satisfaction (1)	50.26	50.6	48.92	0.68%	2.67%
Dissatisfaction (0)	11.55	11.55	11.57	0.00%	0.17%

We can tell from the table above that in regardless of slight differences, the proportions of the three categories in three datasets are similar, which means that both the train dataset and the test dataset represent the whole dataset well. The differences between the distribution of three satisfaction levels of train dataset and test dataset in comparison of whole dataset show that the differences are all within 5%, so using the Train Dataset for modelling and using Test Dataset for validation are acceptable.

3.2 Model Specification and Utility Function

The models used for comparison in this part are ordered logit model and multinomial logit model. The utility functions and probabilities are discussed in following parts.

3.2.1 Ordered Logit Model

In ordered logit model, people 's choices are ordinal and based on the values of cutoffs and utility. If $U > k_1$, people choose "complete". If $k_1 \geq U \geq k_2$, people choose "medium". If $U < k_2$, people choose "dissatisfaction".

In the ordered logit model, the utility function is defined as:

$$U = \beta_1 x_1 + \beta_1 x_2 + \beta_1 x_3 + \cdots + \varepsilon$$

Where ε is assumed to be distributed logistic, and the cumulative distribution of ε is:

$$F(\varepsilon) = \frac{e^{\varepsilon}}{1 + e^{\varepsilon}}$$

3.2.2 Multinomial Logit Model

The variables included in logit model used in mlogit are categorized into three groups: alternative specific variables with generic coefficients, individual specific variables with alternative specific coefficients, alternative specific variables with alternative specific coefficients.

It is considered reasonable at first that monetary variables are considered to be variables with generic variables, indicating that the variables have the same impact on all the individuals. However, the nature of the project that people have different level of life satisfaction is different from making choice. Without further information, all variables in the project are considered to be individual specific variables showing variables have different impact on people having different levels of satisfaction, and the final model is considered to be multinomial logit model.

According to this, the utility functions for three groups are:

$$U_{complete} = \beta_0^{complete} + \beta_1^{complete} x_1 + \beta_2^{complete} x_2 + \cdots + \varepsilon$$

$$U_{medium} = \beta_0^{medium} + \beta_1^{medium} x_1 + \beta_2^{medium} x_2 + \cdots + \varepsilon$$

$$U_{dissatisfaction} = \beta_0^{dissatisfaction} + \beta_1^{dissatisfaction} x_1 + \beta_2^{dissatisfaction} x_2 + \cdots + \varepsilon$$

Probability for an individual to be in a group is calculated based on the utility.

$$P_{ni} = \frac{e^{U_{ni}}}{\sum_j e^{U_{nj}}}$$

3.3 Independent Variable Preprocesses

Since the discrete choice model is adapted based on the independent variables used for the linear regression model, the model used for discrete choice model should be decided based on the independent variables discussed in Section 2.5. In the ordered logit modelling process, in order to avoid the error of "infinite or missing value", the independent variables which have large variances and large mean values are log-transformed (hifdip - Household financial year disposable regular income ($), hsrnti - Rent usual payments $ per month, hsmgi - Mortgage usual repayments $ per month, hxypbti - Household annual expenditure - Public transport and taxis ($), hxymvfi - Household annual expenditure - Motor vehicle fuel ($), hxymvri - Household annual expenditure - Motor vehicle repairs/maintenance ($), hxyedci - Household annual expenditure - Education fees ($), hxyhlpi - Household annual expenditure - Fees paid to health practitioners ($), hxyncri - Household annual expenditure - New motor vehicles, motorbikes or other vehicles ($), hsvalui - Home value ($), tifdip - Financial year disposable regular income ($)).

3.4 Model Comparison

In this part, ordered logit model and multinomial logit model are performed with the train dataset, and the final models are compared before deciding on the most suitable model for this project before assumption testing and validation. The modelling processes are performed by R packages of "MASS" and "mlogit" respectively. At the beginning of each variable selection process, the independent variables included in the model are those included in the

final model of Section 2. The confidence level is consistent with that for linear regression model as 95%.

In order to avoid including insignificant variables or adding variables blindly, AIC, BIC or McFadden R^2 are used as variable selection criteria.

3.4.1 Ordered Logit Model

- **Variable Selection**

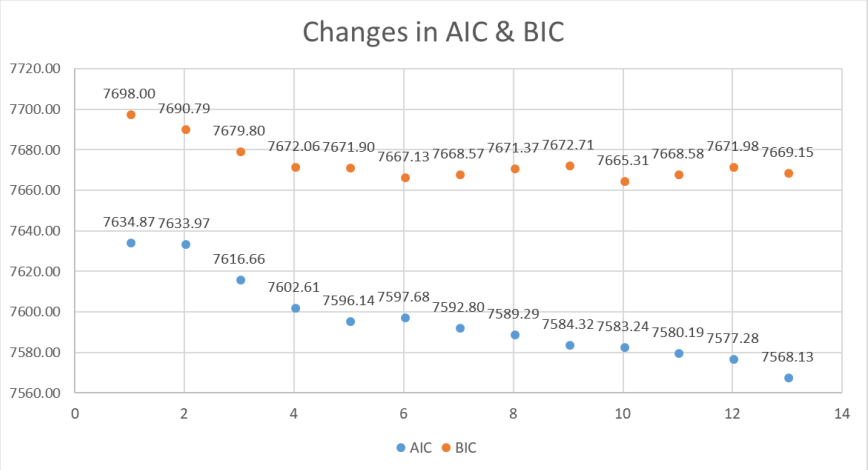
The basic ordered logit model using the independent variables decided on Section 2 is:

Formula	variables	Coefficients	Std. Error	t value	p value	AIC
satisfaction ~ lscom + lsvol + jbmsall + Renter + gh1 + CoupleW + Le_fnw + hhad10_4 + hhec10_5	lscom	0.03	0.01	3.11	0.0019	7636.807
	lsvol	-0.11	0.02	-7.05	0.0000	
	jbmsall	-0.20	0.02	-9.80	0.0000	
	Renter	0.30	0.07	4.17	0.0000	
	gh1	0.08	0.04	2.32	0.0205	
	CoupleW	-0.07	0.06	-1.04	0.2972	
	Le_fnw	-0.44	0.21	-2.07	0.0387	
	hhad10_4	-0.02	0.10	-0.24	0.8097	
	hhec10_5	-0.14	0.09	-1.56	0.1195	
	complete|dissatisfaction	-1.82	0.20	-9.05	0.0000	
	dissatisfaction|medium	-1.32	0.20	-6.61	0.0000	

According to the table above, independent variables CoupleW, hhad10_4 and hhec10_5 are not statistically significant. These variables are deleted from the model one by one, and the significance of all included variables is checked every time after deleting a variable.

After this step, new independent variables are added into the model one by one. The selection of new variables is based on the improvement of AIC, BIC and p-value of the variables. So the independent variables that are added into the model minimize the AIC and are statistically significant. The significance of all included variables is also checked after each step of adding or deleting a variable, and insignificant variables are deleted. To make it simple, the whole process of variable selection is summarized in the table and graph below.

Action	- hhad10_4	- CoupleW	+ hsmgi	+ hh10_14	+ hhold	- Le_fnw	+ Le_sep
AIC	7634.87	7633.97	7616.66	7602.61	7596.14	7597.68	7592.80
BIC	7698.00	7690.79	7679.80	7672.06	7671.90	7667.13	7668.57
Action	+ tifdip	+ hgage	- hhold	+ Mltpljob	+ hhda10_4	+ hsrnti	
AIC	7589.29	7584.32	7583.24	7580.19	7577.28	7568.13	
BIC	7671.37	7672.71	7665.31	7668.58	7671.98	7669.15	



We can tell from the table and graph above that the changes of BIC and AIC in this process show different trends. AIC keeps decreasing but BIC reaches the lowest point after deleting "hhold". In order to avoid overfitting, variable selection stops after deleting "hhold". After the process, the final ordered logit model is:

Formula	Variable	Value	Std. Error	t value	p value	AIC
satisfaction ~ lscom + lsvol + jbmsall + Renter + gh1 + hhec10_4 + hh10_14 + hsmgi + Le_sep + tifdip + hgage	lscom	0.018	0.008	2.252	0.024	7583.235
	lsvol	-0.101	0.015	-6.603	0.000	
	jbmsall	-0.187	0.020	-9.318	0.000	
	Renter	0.397	0.087	4.534	0.000	
	gh1	0.104	0.037	2.843	0.004	
	hhec10_4	-0.136	0.062	-2.185	0.029	
	hh10_14	-0.208	0.053	-3.882	0.000	
	hsmgi	0.039	0.010	3.694	0.000	
	Le_sep	-0.435	0.154	-2.822	0.005	
	tifdip	0.091	0.025	3.571	0.000	
	hgage	-0.010	0.003	-3.832	0.000	
	complete|dissatisfaction	-1.062	0.316	-3.361	0.001	
	dissatisfaction|medium	-0.559	0.316	-1.770	0.077	

The utility function after this step is:

$$U = 0.018lscom - 0.101lsvol - 0.187jbmsall + 0.397Renter + 0.104gh1 - 0.136hhec10\_4 - 0.208hh10\_14 + 0.039log(hsmgi) - 0.435Le\_sep + 0.091log(tifdip) - 0.010hgage$$

- **Interpretation**

The order defined in the model is "complete, dissatisfaction, medium". There are two intercepts, which indicate the places that the latent variables are cut into three groups. The coefficients and p-value are used to determine the utility function and variable significance respectively. AIC shows the goodness-of-fit, and a lower AIC value indicates better fitness.

In this model, the increase in the utility function value is associated with the increase of commuting time to work, decrease of volunteer/charity working time, decrease of job satisfaction, renting the accommodation, being less healthy, not being the group of hhec10_4, having less family members aged between 10-14, the increase of mortgage usual repayments per month, not being separated with partner, having higher financial year

disposable regular income ($) and lower age.

Since the utility function is related to the probability of being in each group with a cumulative distribution of ε $F(\varepsilon) = \frac{e^{\varepsilon}}{1+e^{\varepsilon}}$, the increase of utility function does not necessary result in the increase in being in a higher ranked group.

3.4.2 Multinomial Logit Model

- **Variable Selection**

With the reference level of the group of complete, the basic multinomial logit model using the independent variables decided on Section 2 is:

Formula	Variables	Estimate	Std. Error	z-value	Pr(>|z|)	McFadden R^2	AIC	Log-Likelihood
choice ~ lscom + lsvol + jbmsall + Renter + gh1 + CoupleW + Le_fnw + hhad10_4 + hhec10_5	dissatisfaction:(intercept)	0.4917	0.3459	1.4215	0.1552	0.0686	7346.918	-3653.5
	medium:(intercept)	2.1897	0.2444	8.9594	< 2.2e-16			
	dissatisfaction:lscom	0.0534	0.0139	3.8482	0.0001			
	medium:lscom	0.0336	0.0094	3.5526	0.0004			
	dissatisfaction:lsvol	-0.1474	0.0333	-4.4292	0.0000			
	medium:lsvol	-0.1185	0.0169	-7.0333	0.0000			
	dissatisfaction:jbmsall	-0.3949	0.0346	-11.4112	< 2.2e-16			
	medium:jbmsall	-0.2861	0.0252	-11.3698	< 2.2e-16			
	dissatisfaction:Renter	1.0434	0.1235	8.4517	< 2.2e-16			
	medium:Renter	0.4763	0.0875	5.4427	0.0000			
	dissatisfaction:gh1	0.4230	0.0651	6.4942	0.0000			
	medium:gh1	0.1367	0.0419	3.2622	0.0011			
	dissatisfaction:CoupleW	-0.4127	0.1168	-3.5334	0.0004			
	medium:CoupleW	-0.1043	0.0722	-1.4450	0.1484			
	dissatisfaction:Le_fnw	0.9103	0.3327	2.7359	0.0062			
	medium:Le_fnw	-0.2983	0.3145	-0.9484	0.3429			
	dissatisfaction:hhad10_4	0.6746	0.1851	3.6447	0.0003			
	medium:hhad10_4	0.0510	0.1092	0.4674	0.6402			
	dissatisfaction:hhec10_5	-0.4489	0.1822	-2.4641	0.0137			
	medium:hhec10_5	-0.1913	0.1034	-1.8496	0.0644			

In comparison with linear model, there is no t-value or p-value in multinomial logit model, but z and Pr(>|z|). The test statistic z is the ratio of the coefficient to the standard error of the respective predictor, and the Pr>|z| is the probability the z test statistic would be observed under the null hypothesis. Both z and Pr(>|z|) show the significance of variables.

In this model, the variables CoupleW for group medium, Le_fnw for group medium, hhad10_4 for group medium and hhec10_5 for medium are not statistically significant. These variables should be deleted from the model one by one. And after each step of deleting a variable, the significance of remaining variables is checked.

After this step, new variables are added into the model. AIC is used as the criterion for variable selection. The best model chosen in this part, which is the same as ordered logit model, minimizes the AIC and Log-Likelihood, but also maximise the $McFadden\ R^2$. At the same time, all variables included in the model should be statistically significant. To be brief, all the steps of adding and deleting variables are concluded in the graph and table below.

Action	AIC	R	Log-Likelihood
- hhec10_5	7350.189	0.067719	-3657.1
- Le_fnw	7362.762	0.065606	-3665.4
- hhad10_4	7375.372	0.063489	-3673.7
- CoupleW	7383.644	0.061925	-3679.8
+ hgage	7363.147	0.065048	-3667.6
+ hh10_14	7348.918	0.067371	-3658.5

The final model in this step is:

Formula	Variables	Estimate	Std. Error	z-value	Pr(>|z|)	McFadden R^2	AIC	Log-Likelihood
choice ~ lscom+ lsvol + jbmsall + Renter + gh1 + hgage + hh10_14	dissatisfaction:(intercept)	1.0093	0.3636	2.7758	0.0055	0.067371	7348.918	-3658.5
	medium:(intercept)	2.4581	0.2519	9.7582	< 2.2e-16			
	dissatisfaction:lscom	0.0596	0.0138	4.3117	0.0000			
	medium:lscom	0.0368	0.0095	3.8681	0.0001			
	dissatisfaction:lsvol	-0.1322	0.0323	-4.0903	0.0000			
	medium:lsvol	-0.1099	0.0167	-6.5971	0.0000			
	dissatisfaction:jbmsall	-0.3947	0.0346	-11.4202	< 2.2e-16			
	medium:jbmsall	-0.2786	0.0251	-11.0876	< 2.2e-16			
	dissatisfaction:Renter	0.9960	0.1232	8.0831	0.0000			
	medium:Renter	0.3887	0.0878	4.4291	0.0000			
	dissatisfaction:gh1	0.5021	0.0653	7.6875	0.0000			
	medium:gh1	0.1559	0.0423	3.6872	0.0002			
	dissatisfaction:hgage	-0.0209	0.0046	-4.5369	0.0000			
	medium:hgage	-0.0111	0.0028	-3.9406	0.0001			
	dissatisfaction:hh10_14	-0.2652	0.1023	-2.5912	0.0096			
	medium:hh10_14	-0.2394	0.0590	-4.0594	0.0000			

The utility functions for the three groups are:

$$U_{complete} = 0$$

$$U_{dissatisfaction} = 1.0093 + 0.0596lscom - 0.1322lsvol - 0.3947jbmsall + 0.9960Renter + 0.5021gh1 - 0.0209hgage - 0.2652hh10\_14$$

$$U_{medium} = 2.4581 + 0.0368lscom - 0.1099lsvol - 0.2786jbmsall + 0.3887Renter + 0.1559gh1 - 0.0111hgage - 0.2394hh10\_14$$

- **Interpretation**

The goodness-of-fit in this model is shown by $McFadden\ R^2$, AIC and Log-likelihood. After adding significant variables, although the values for these criteria are close to the original model, insignificant variables are deleted.

The impact of the factors included in the model is consistent for all groups. The increase in utility of "dissatisfaction" and "medium" groups in this model is related to the increase in commuting time for work, the decrease in volunteer/charity work time, the decrease in job satisfaction, the decrease in age, renting the accommodation, being less healthy, and have less family members aged 10-14.

3.5 Model Selection

After the selection, both of the models meet the criteria of all variables statistically significant and minimizing AIC value. The variables included in both models are similar, and so as the impact of included variables on utility function.

However, the cut-off points in the ordered logit model are not as we expected, which may indicate that using ordered logit model is not a good choice. Apart from this, the AIC value for the final multinomial logit model is also less than that for final ordered logit model. So the multinomial logit model is selected for this part for assumption testing and validation.

3.6 Assumption Testing

The assumptions that should be tested in the project are variable significance and IIA hypothesis (independence of irrelevant alternatives) which indicates that probability ratio for any two alternatives depends only on the characteristics of these two alternatives and not on those of other alternatives [3].

- **Variable Significance**

Apart from the z-values provided in 3.4.2 where the z-values of all included variables in the final model are less than 0.05, the confidence intervals for each variable included in the model is also tested.

Coefficients	2.50%	97.50%
dissatisfaction:(intercept)	0.2966	1.7220
medium:(intercept)	1.9644	2.9518
dissatisfaction:lscom	0.0325	0.0867
medium:lscom	0.0181	0.0554
dissatisfaction:lsvol	-0.1955	-0.0688
medium:lsvol	-0.1426	-0.0773
dissatisfaction:jbmsall	-0.4625	-0.3270
medium:jbmsall	-0.3278	-0.2293
dissatisfaction:Renter	0.7545	1.2376
medium:Renter	0.2167	0.5607
dissatisfaction:gh1	0.3741	0.6301
medium:gh1	0.0730	0.2388
dissatisfaction:hgage	-0.0299	-0.0119
medium:hgage	-0.0166	-0.0056
dissatisfaction:hh10_14	-0.4658	-0.0646
medium:hh10_14	-0.3550	-0.1238

We can tell from the table above that none of the confidence intervals of the included variables cover 0 with a level of 95% confident, so all included variables are significant.

- Independence of Irrelevant Alternatives Testing

The IIA hypothesis is tested with Hausman-McFadden test. The alternative hypothesis that IIA is rejected is accepted with the p-value lower than 0.05 or the null hypothesis that IIA holds is accepted.

Variables	p-value
lscom	1
lsvol	1
jbmsall	1
Renter	1
gh1	1
hgage	1
hh10_14	1

With one variable tested at a time, the p-values are all over 0.05, so the null hypothesis can not be rejected, and the assumption holds true for this model.

3.7 Model Simulation

After the comparison and assumption testing, the model chosen for the project is the multinomial logit model. In order to determine the accuracy of the model, test dataset is used for simulation and validation. For each individual, the group that has the largest predicted possibility is considered to be the predicted choice for that individual.

The probabilities in three groups are calculated based of the results of utilities:

$$P_{complete} = \frac{e^{U_{complete}}}{e^{U_{complete}} + e^{U_{medium}} + e^{U_{dissatisfaction}}}$$

$$P_{medium} = \frac{e^{U_{medium}}}{e^{U_{complete}} + e^{U_{medium}} + e^{U_{dissatisfaction}}}$$

$$P_{dissatisfaction} = \frac{e^{U_{dissatisfaction}}}{e^{U_{complete}} + e^{U_{medium}} + e^{U_{dissatisfaction}}}$$

3.7.1 Example of Choice Estimation

Take the first observation in the test dataset for example. The values of variables included in the model are shown in the table below.

Variable	lscom	lsvol	jbmsall	Renter	gh1	hgage	hh10_14
Value	4.6667	0	6	0	4	52	0

The probability is calculated based on the probability formula. Since the model is performed with the reference level of "complete", the utility of the group of "complete" is zero.

$$U_{complete} = 0$$

$$U_{dissatisfaction} = 1.0093 + 0.0596 \times 4.6667 - 0.1322 \times 0 - 0.3947 \times 6 + 0.9960 \times 0 + 0.5021 \times 4 - 0.0209 \times 52 - 0.2652 \times 0 = -0.15767$$

$$U_{medium} = 2.4581 + 0.0368 \times 4.6667 - 0.1099 \times 0 - 0.2786 \times 6 + 0.3887 \times 0 + 0.1559 \times 4 - 0.0111 \times 52 - 0.2394 \times 0 = 1.00609$$

The predicted probabilities of this individual being in the three groups are calculated afterwards.

$$e^{U_{complete}} = e^{0} = 1$$

$$e^{U_{dissatisfaction}} = e^{-0.15767} = 0.85413$$

$$e^{U_{medium}} = e^{1.00609} = 2.73489$$

$$e^{U_{complete}} + e^{U_{dissatisfaction}} + e^{U_{medium}} = 4.58902$$

$$P_{complete} = \frac{e^{U_{complete}}}{e^{U_{complete}} + e^{U_{dissatisfaction}} + e^{U_{medium}}} = \frac{1}{4.58902} = 0.21791$$

$$P_{medium} = \frac{e^{U_{medium}}}{e^{U_{complete}} + e^{U_{medium}} + e^{U_{dissatisfaction}}} = \frac{2.73489}{4.58902} = 0.59596$$

$$P_{dissatisfaction} = \frac{e^{U_{dissatisfaction}}}{e^{U_{complete}} + e^{U_{medium}} + e^{U_{dissatisfaction}}} = \frac{0.85413}{4.58902} = 0.18612$$

Since the maximum value of the three estimated probabilities is the probability of choosing "medium", "medium" is considered to be the estimated group for this individual.

3.7.2 Comparison of Observations and Estimations

In comparison with the observations, 576 out of 1020 observations are estimated correctly, taking up of 56.42% of the total observations. The comparison of observations with estimated results of each satisfaction group are shown in the table below.

	complete	medium	dissatisfaction
Estimated	286	732	2
Observation	499	403	118
Correct	171	403	2

The proportions of the three groups are 28.04%, 71.76% and 0.20% for complete, medium and dissatisfaction respectively. In comparison with the proportions provided in previous parts, the proportions for dissatisfaction and complete decreased greatly, and medium takes the majority of the whole test dataset.

For observation groups, the estimation of “complete” group only covers 34.27% of the total “complete” group observations, but it covers 100% of the “medium” group. As for “dissatisfaction” group, only 2 out of 118 observations are predicted correctly.

For estimation groups, however, the prediction of “dissatisfaction” group is accurate, while the accuracy for “complete” group and “medium” group are 59.79% and 55.05% respectively.

3.8 Conclusion

Multinomial logit model and ordered logit model are used in this part for discrete choice modelling, and we consider multinomial logit model to be more suitable for this project.

The prediction of the multinomial logit model shows that the accuracy of prediction is still not satisfactory enough, and the distribution of the three groups is also far from the reality.

In short, multinomial logit model explains only about a half of the total dataset, and it may not be suitable for modelling and estimation in this project.

4. Conclusion

In this report, data from the survey of Household Income and Labour Dynamics in Australia (HILDA) is used, and linear regression model and discrete choice model are performed for modelling and estimating the potential choice of the individuals. The main conclusion of this report covers the following factors:

4.1 Differences between Linear Regression Model and Discrete Choice Model

Linear regression model considers there is a linear relationship between the dependent variable and independent variables. Included variables have a direct impact on the independent variable, and the estimations of the model is not discrete values as the observations. The comparison of the distribution of estimation and observation shows that linear regression estimations are more concentrated than the observations, and there are even some individuals the estimations of which are out of the range.

Discrete choice models in this project are ordered logit model and multinomial logit model, and both of them estimate the group of the dependent variables as probabilities for each group. The coefficients indicate the impact of independent variables on utility function instead of on choices. In this report, the estimations are simply the group with highest probabilities, and the distribution of estimations are not close to the observations.

4.2 Potential Implement

Although both linear and discrete choice model are not suitable for predicting the dependent variables, there is similarity between the models. The variable selection in the three models are all based on criteria of AIC, BIC, variable significance, and variables included in the final three models (even their impact on dependent variables) show great similarity, which indicates the potential implement of the process could be determining the key factors that have significant impact on the dependent variable.

In addition, although the distribution of discrete choice modelling is far from observations, it caught two individuals in dissatisfaction group perfectly. This provides with a potential implement that this model may contribute to catching individuals with extreme low life satisfaction.

4.3 Future Study

In the future, more information should be gathered before updating the model. Having adequate information contributes to the significance of the impact of independent variables. In discrete choice model, including variables with different types of coefficients (generic or alternative specific) should be considered.

Reference

[1] Australian Bureau of Statisitcs, "2033.0.55.001 - Census of Population and Housing: Socio-Economic Indexes for Areas (SEIFA), Australia, 2011," Australian Bureau of Statistics, 23 March 2018. [Online]. Available: http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/2033.0.55.001~2011~Main%20Features~IRSD~10005. [Accessed: 10 May 2018].

[2] Australian Bureau of Statistics, "2033.0.55.001 - Census of Population and Housing: Socio-Economic Indexes for Areas (SEIFA), Australia, 2011," Australian Bureau of Statistics, 23 March 2018. [Online]. Available: http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/2033.0.55.001~2011~Main%20Features~IER~10006. [Accessed: 10 May 2018].

[3] Y. Croissant, "Estimation of multinomial logit models in R: The mlogit Packages," Facult ´e de Droit et d'Economie.

Appendix

Table 1 Variable Definition

Waveid	IDENTIFIERS - XW Cross Wave ID
hh0_4	Number of persons aged 0-4 years at June 30 2016
hh5_9	Number of persons aged 5-9 years at June 30 2016
hh10_14	Number of persons aged 10-14 years at June 30 2016
Hhadult	Number of persons aged 15+ years at June 30 2016
Hifdip	Household financial year disposable regular income ($)
Hsrnti	Rent usual payments $ per month
Hsmgi	Mortgage usual repayments $ per month
Hxypbti	Household annual expenditure - Public transport and taxis ($)
Hxymvfi	Household annual expenditure - Motor vehicle fuel ($)
Hxymvri	Household annual expenditure - Motor vehicle repairs/maintenance ($)
Hxyedci	Household annual expenditure - Education fees ($)
Hxyhlpi	Household annual expenditure - Fees paid to health practitioners ($)
Hxyncri	Household annual expenditure - New motor vehicles, motorbikes or other vehicles ($)
Hsvalui	Home value ($) [weighted topcode]
Hsbedrm	Number of bedrooms
Hgage	Age last birthday at June 30 2016
tifdip	Financial year disposable regular income ($) [weighted topcode]
hhfty	Family type
jbmmth	Number of days usually worked in a 4 week period
jbmsall	Overall job satisfaction
losat	How satisfied are you with your life
hhyng	Age of youngest person in household
hhold	Age of oldest person in household
gh1	Self-assessed health
jompf	I get paid fairly for the things I do in my job
lsemp	Combined hrs/mins per week - Paid employment
lscom	Combined hrs/mins per week - Travelling to/from paid employment
lshw	Combined hrs/mins per week – Housework
lsod	Combined hrs/mins per week - Outdoor tasks
lschd	Combined hrs/mins per week - Playing with your children
lsvol	Combined hrs/mins per week - Volunteer/Charity work
bmi	Body Mass Index (BMI)

Table 2 Mean and Standard Deviation for Continuous Variables

	hh0_4	hh5_9	hh10_14	hhadult	hifdip	hsrnti	hsmgi	hxypbti
Mean	0.19	0.23	0.28	2.39	80425.56	1038.43	675.04	385.02
Standard Deviation	0.51	0.55	0.59	0.99	47610.22	243.90	1015.27	945.56
	hxymvfi	hxymvri	Hxyedci	hxyhlpi	hxyncri	hsvalui	hsbedrm	hgage
Mean	3022.13	1058.40	1444.32	934.59	2589.28	354282.20	3.35	38.42
Standard Deviation	3671.61	1138.22	3903.91	1515.52	9137.14	374556.97	0.99	13.01
	tifdip	hhfty	Jbmmth	jbmsall	losat	hhyng	hhold	gh1
Mean	39359.51	7.88	16.04	7.65	7.87	21.72	44.62	2.40
Standard Deviation	29576.80	7.98	2.11	1.62	1.24	16.88	12.04	0.86
	jompf	lsemp	Lscom	lshw	lsod	lschd	lsvol	bmi
Mean	4.63	36.70	3.89	9.12	3.63	6.32	0.78	26.19
Standard Deviation	1.62	15.50	3.88	8.94	4.95	12.25	2.54	5.10

Table 3 Discrete Variable Frequencies

	Female	Married	ESL	Le_mar	Le_sep	Le_job	Le_bth	Le_prg
1	2446	3226	5078	208	204	1017	183	288
0	2653	1873	21	4891	4895	4082	4916	4811
	Le_death	Le_fni	Le_fnw	Le_frd	Le_prm	Le_rtr	Le_ins	Mltpljob
1	864	160	91	120	552	15	306	456
0	4235	4939	5008	4979	4547	5084	4793	4643
	Manager	Professional	Technician	Service Worker	Administrative	SalesWorker	Machinery Operator	Labour
1	631	1289	667	504	768	496	280	464
0	4468	3810	4432	4595	4331	4603	4819	4635
	FlxWork	HmWork	PrtStudy	FullStudy	Postgrad	Bachelor	hhad10_1	hhad10_2
1	2656	1033	449	249	595	872	334	800
0	2443	4066	4650	4850	4504	4227	4765	4299
	hhad10_3	hhad10_4	hhad10_5	hhad10_6	hhad10_7	hhad10_8	hhad10_9	hhda10_1
1	1294	1761	2237	2722	3268	3826	4472	311
0	3805	3338	2862	2377	1831	1273	627	4788
	hhda10_2	hhda10_3	hhda10_4	hhda10_5	hhda10_6	hhda10_7	hhda10_8	hhda10_9
1	712	1217	1627	2126	2664	3265	3851	4484
0	4387	3882	3472	2973	2435	1834	1248	615
	hhec10_1	hhec10_2	hhec10_3	hhec10_4	hhec10_5	hhec10_6	hhec10_7	hhec10_8
1	362	824	1341	1808	2325	2811	3386	3944
0	4737	4275	3758	3291	2774	2288	1713	1155
	hhec10_9	hhed10_1	hhed10_2	hhed10_3	hhed10_4	hhed10_5	hhed10_6	hhed10_7
1	4520	358	841	1325	1760	2253	2712	3223
0	579	4741	4258	3774	3339	2846	2387	1876
	hhed10_8	hhed10_9	CoupleWo	CoupleW	LoneW	Single	Renter	
1	3820	4449	1336	2589	422	604	1278	
0	1279	650	3763	2510	4677	4495	3821	

Table 4 Comparison of Correlation Coefficients

	hh0_4	hhadult	hifdip	hsvalui	hgage	tifdip	hhfty	losatic	hhyng	hhold	lschd	Married	Le bth	Le prg	CoupleWo	CoupleW	Single	Renter	hxypbti
hh0_4	1.000	-0.087	-0.020	-0.014	-0.122	0.061	-0.165	-0.014	-0.454	-0.236	0.526	0.215	0.489	0.342	-0.214	0.333	-0.138	0.003	0.060
hhadult	-0.087	1.000	0.392	0.219	-0.138	-0.213	-0.285	0.054	-0.292	0.289	-0.114	0.076	-0.055	-0.065	-0.164	0.500	-0.513	-0.223	-0.191
hifdip	-0.020	0.392	1.000	0.370	0.021	0.382	-0.258	0.054	-0.094	0.146	-0.015	0.161	0.006	0.015	0.032	0.247	-0.312	-0.218	0.442
hsvalui	-0.014	0.219	0.370	1.000	0.181	0.117	-0.182	0.132	-0.028	0.302	0.029	0.130	-0.021	-0.045	-0.051	0.220	-0.192	-0.547	0.149
hgage	-0.122	-0.213	0.382	0.181	1.000	0.374	-0.063	-0.109	0.444	0.589	-0.026	0.265	0.051	0.061	0.095	-0.071	0.044	-0.044	0.364
tifdip	0.061	-0.285	-0.258	-0.182	-0.063	1.000	-0.034	-0.020	0.079	-0.025	0.050	-0.685	-0.087	-0.120	-0.511	-0.320	0.741	0.250	-0.043
hhfty	-0.165	-0.285	-0.258	-0.182	-0.063	-0.055	1.000	-0.071	0.180	-0.042	-0.145	1.000	-0.087	-0.120	-0.511	-0.320	0.741	0.250	-0.043
losatic	-0.014	0.054	0.054	0.132	0.109	0.001	-0.071	1.000	0.005	0.117	0.029	0.070	-0.015	-0.031	-0.027	0.098	-0.051	-0.180	-0.013
hhyng	-0.454	-0.292	-0.094	-0.028	0.444	0.046	0.180	0.005	1.000	0.387	-0.459	-0.152	-0.241	-0.175	0.552	-0.675	0.352	0.010	0.071
hhold	-0.236	0.289	0.146	0.302	0.589	0.020	-0.042	0.117	0.387	1.000	-0.182	-0.082	-0.171	-0.201	-0.014	0.082	-0.136	-0.370	-0.013
lschd	0.526	-0.114	-0.015	0.029	-0.026	0.022	-0.145	0.029	-0.459	-0.182	1.000	0.212	0.282	0.202	-0.287	0.355	-0.144	-0.056	0.042
Married	0.215	0.076	0.161	0.130	0.301	0.198	-0.685	0.070	-0.152	-0.082	0.212	1.000	0.136	0.141	0.344	0.273	-0.463	-0.171	0.255
Le bth	0.489	-0.055	0.006	-0.021	-0.099	0.044	-0.087	-0.015	-0.241	-0.171	0.282	0.136	1.000	0.537	-0.115	0.184	-0.067	0.032	0.051
Le prg	0.342	-0.065	0.015	-0.045	-0.125	0.042	-0.120	-0.031	-0.175	-0.201	0.202	0.141	0.537	1.000	0.003	0.091	-0.074	0.033	0.058
CoupleWo	-0.214	-0.164	0.032	-0.051	0.130	0.095	-0.511	-0.027	0.552	-0.014	-0.287	0.344	-0.115	0.003	1.000	-0.605	-0.218	0.035	0.087
CoupleW	0.333	0.500	0.247	0.220	-0.067	-0.071	-0.320	0.098	-0.675	0.082	0.355	0.273	0.184	0.091	-0.605	1.000	-0.372	0.236	0.034
Single	-0.138	-0.513	-0.312	0.192	0.031	0.044	0.741	-0.051	0.352	-0.136	0.144	-0.463	-0.067	-0.074	-0.218	-0.372	1.000	0.236	-0.056
Renter	0.003	-0.223	-0.218	-0.547	-0.044	0.044	0.250	-0.180	0.010	-0.370	-0.056	-0.171	0.032	0.033	0.035	-0.267	0.236	1.000	-0.060
hxypbti	0.060	-0.191	0.442	0.149	0.364	0.986	-0.043	-0.013	0.071	-0.013	0.042	0.255	0.051	0.068	0.087	-0.056	0.034	-0.060	1.000

	hh0_4	hhadult	hifdip	hsvalui	hgage	tifdip	hhfty	losatic	hhyng	hhold	lschd	Married	Le bth	Le prg	CoupleWo	CoupleW	Single	Renter	hxypbti
hh0_4	1.000	-0.087	-0.020	-0.014	-0.122	0.045	-0.165	-0.014	-0.454	-0.236	0.526	0.215	0.489	0.342	-0.214	0.333	-0.138	0.003	-0.069
hhadult	-0.087	1.000	0.392	0.219	-0.138	-0.116	-0.285	0.054	-0.292	0.289	-0.114	0.076	-0.055	-0.065	-0.164	0.500	-0.513	-0.223	0.112
hifdip	-0.020	0.392	1.000	0.370	0.021	0.552	-0.258	0.054	-0.094	0.146	-0.015	0.161	0.006	0.015	0.032	0.247	-0.312	-0.218	0.121
hsvalui	-0.014	0.219	0.370	1.000	0.181	0.199	-0.182	0.132	-0.028	0.302	0.029	0.130	-0.021	-0.045	-0.051	0.220	-0.192	-0.547	0.031
hgage	-0.122	-0.138	0.021	0.181	1.000	0.294	-0.063	0.109	0.444	0.589	-0.026	0.301	-0.099	-0.125	0.130	-0.067	0.031	-0.242	-0.073
hhfty	-0.165	-0.285	-0.258	-0.182	-0.063	-0.055	1.000	-0.071	0.180	-0.042	-0.145	-0.685	-0.087	-0.120	-0.511	-0.320	0.741	0.250	0.004
losatic	-0.014	0.054	0.054	0.132	0.109	0.001	-0.071	1.000	0.005	0.117	0.029	0.070	-0.015	-0.031	-0.027	0.098	-0.051	-0.180	-0.047
hhyng	-0.454	-0.292	-0.094	-0.028	0.444	0.046	0.180	0.005	1.000	0.387	-0.459	-0.152	-0.241	-0.175	0.552	-0.675	0.352	0.010	0.003
hhold	-0.236	0.289	0.146	0.302	0.589	0.020	-0.042	0.117	0.387	1.000	-0.182	-0.082	-0.171	-0.201	-0.014	0.082	-0.136	-0.370	-0.018
lschd	0.526	-0.114	-0.015	0.029	-0.026	0.022	-0.145	0.029	-0.459	-0.182	1.000	0.212	0.282	0.202	-0.287	0.355	-0.144	-0.056	-0.060
Married	0.215	0.076	0.161	0.130	0.301	0.198	-0.685	0.070	-0.152	-0.082	0.212	1.000	0.136	0.141	0.344	0.273	-0.463	-0.171	-0.013
Le bth	0.489	-0.055	0.006	-0.021	-0.099	0.044	-0.087	-0.015	-0.241	-0.171	0.282	0.136	1.000	0.537	-0.115	0.184	-0.067	0.032	-0.049
Le prg	0.342	-0.065	0.015	-0.045	-0.125	0.042	-0.120	-0.031	-0.175	-0.201	0.202	0.141	0.537	1.000	0.003	0.091	-0.074	0.033	-0.028
CoupleWo	-0.214	-0.164	0.032	-0.051	0.130	0.095	-0.511	-0.027	0.552	-0.014	-0.287	0.344	-0.115	0.003	1.000	-0.605	-0.218	0.035	0.040
CoupleW	0.333	0.500	0.247	0.220	-0.067	-0.013	-0.320	0.098	-0.675	0.082	0.355	0.273	0.184	0.091	-0.605	1.000	-0.372	-0.267	-0.034
Single	-0.138	-0.513	-0.312	-0.192	0.031	0.007	0.741	-0.051	0.352	-0.136	-0.144	-0.463	-0.067	-0.074	-0.218	-0.372	1.000	0.236	-0.024
Renter	0.003	-0.223	-0.218	-0.547	-0.242	-0.083	0.250	-0.180	0.010	-0.370	-0.056	-0.171	0.032	0.033	0.035	-0.267	0.236	1.000	0.071
tifdip	0.045	-0.116	0.552	0.199	0.294	1.000	-0.055	0.001	0.046	0.020	0.022	0.198	0.044	0.042	0.056	-0.013	0.007	-0.083	0.047
hxypbti	-0.069	0.112	0.121	0.031	-0.073	0.047	0.004	-0.047	0.003	-0.018	-0.060	-0.013	-0.049	-0.028	0.040	-0.034	-0.024	0.071	1.000