MULTINOMIAL LOGISTIC REGRESSION

Logistic Regression with SAS

Project II prepared by Jiyoung Kim (110075) and Artur Pososhko (110554)

This work is carried out as part of a Logistic Regression course at Warsaw School of Economics in the academic year 2020/2021 under care of PhD Adam Korczyński

MASTER'S DEGREE

Advanced Analytics – Big Data

Table of Contents

1. Introduction

Historically, the 'income' has been closely associated with the economy. In the past, people used to regard 'income' as the earnings and compensations they get by offering their labour as a laborer in the labour market. However, the world has been changed and there are so many diverse source of income. Some people focus on "Making money while you sleep", which is making Capital Income, by investing on the stocks and properties or having their own business while others only focus on working as an employee (Labour Income) or receive social grants or benefits (Social Income) as their main source of income. So, we started to become curious what kind of factors may influence the people' main source of income. Especially among those possible factors, we want to focus on 'People' satisfaction level on their country's current state of economy'.

The main hypothesis that we try to prove or disprove in this project is as follows: "People who have lesser satisfaction with present state of economy in their country are more likely to have Social Grants/Benefits Income as their main source of income than the "Labour Income". In order to check our main hypothesis, with European Social Survey (ESS 9 - 2018) data, we will use SAS studio to proceed descriptive analysis, create a multinomial logistic regression model and interpret the results in detail.

2. Descriptive Statistics

2.1. Frequency distribution and missing data

Since ESS dataset is a survey data, it has values such as 'Not Applicable (66)', 'Refusal (77)', 'Don't know (88)', 'No answer (99)' (sometimes expressed in 7, 9). Those values should be regarded as missing values and need to be dealt with imputation or dropping. Before doing so, it's better to check the proportion of missing values per each variable and if there is any missing data pattern.

There was one thing we found before checking missing data pattern. As checking the distribution of this variable, it seemed that '0' and 'Not applicable' is an outlier because '0' has only 2 observations which are nearly 0% and 'Not applicable' shows 29.91%. Possible explanation behind this is that the respondents just simply checked 'Not applicable' if they don't have any child. So, we assume that it is acceptable to group these two levels (0 and Not applicable) into the '0' category.

For checking missing data pattern, we used PROC MI statement and it turns out there is no specific pattern and most variables have very small percentage of them(around 0-1%) except for variable hinctnta(total net income, 16.29% of missing values). In here, we only show the part of our result table since most of the rows show very low percent which equal to almost 0. You can find the original

whole table in the Appendix. Since it doesn't show any specific missing pattern, we decided to drop all those missing values including ones in the variable hinctnta.

Missing Data Patterns

Group	Y	age	inc	mar	_nbthcld	_stfeco	_stfgov	Freq	Percent	Group Means						
										Y	age	inc	mar	_nbthcld	_stfeco	_stfgov
1	X	X	X	X	X	X	X	38141	77.02	1.828505	53.980179	1.946121	1.298865	19.701214	2.129415	1.882567
2	X	X	X	X	X	X	.	783	1.58	1.865900	54.655172	1.770115	1.325670	21.798212	2.107280	.
3	X	X	X	X	X	.	X	522	1.05	2.312261	60.582375	1.519157	1.272031	17.624521	.	1.783525
4	X	X	X	X	X	.	.	241	0.49	2.315353	53.763485	1.443983	1.365145	24.497925	.	.
5	X	X	X	X	.	X	X	48	0.10	1.895833	95.270833	1.645833	1.125000	.	2.166667	1.770833
6	X	X	X	X	.	X	.	1	0.00	1.000000	16.000000	2.000000	2.000000	.	3.000000	.
7	X	X	X	X	.	.	X	1	0.00	1.000000	72.000000	2.000000	1.000000	.	.	2.000000
8	X	X	X	.	X	X	X	35	0.07	1.885714	79.685714	1.771429	.	36.685714	1.600000	1.742857
9	X	X	X	.	X	X	.	2	0.00	1.000000	43.000000	2.000000	.	66.000000	1.500000	.
10	X	X	.	X	X	X	X	8065	16.29	1.818599	60.083447	.	1.362678	25.845009	2.032734	1.865716

2.2. Target Variable hincsrca - Main source of household income

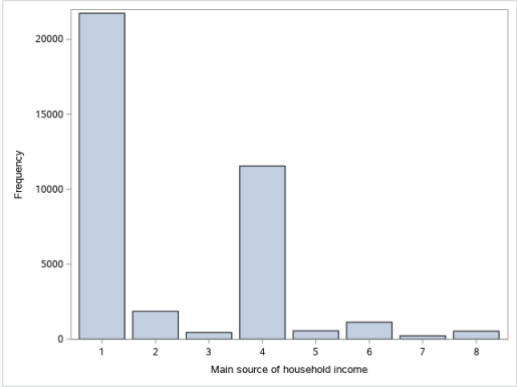
The question respondents were asked is as follows: "Please consider the income of all household members and any income which may be received by the household as a whole. What is the main source of income in your household?". You can see possible answers to this question in the table below.

Levels	Categories
1	Wages or salaries
2	Income from self-employment (excluding farming)
3	Income from farming
4	Pensions
5	Unemployment/redundancy benefit
6	Any other social benefits or grants
7	Income from investments, savings etc.
8	Income from other sources
77	Refusal
88	Don't know
99	No answer

You can see the distribution of our target variable on the plot below. For majority of respondents the main source of income is either "Wages or salaries" (57.12%) or "Pensions" (30.35%). The distribution between all other main income sources except two mentioned before is relatively same.

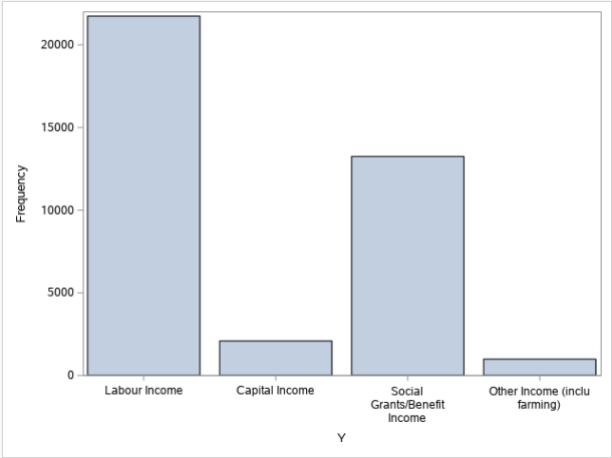
The FREQ Procedure

Main source of household income				
hincsrca	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	21737	57.12	21737	57.12
2	1858	4.88	23595	62.01
3	448	1.18	24043	63.18
4	11548	30.35	35591	93.53
5	563	1.48	36154	95.01
6	1138	2.99	37292	98.00
7	225	0.59	37517	98.59
8	536	1.41	38053	100.00



For better result of Multinomial Logistic Regression analysis, we grouped this target variable into 4 groups as below. We included 'Income from self-employment (excluding farming)(2)' into the Capital Income category since we regarded it as the income generated through the possession of wealth which is their own business as an entrepreneur. And for the 'Income from farming(3)', we put it into the Other Income category even though it can be regarded as Labour Income. It is because that it has such a small portion as 1.18% out of whole and didn't want to contribute on increasing the obs of Labour Income which already counts the largest portion within the variable.

Previous Levels	New Levels	Categories
1	1	Labour Income
2, 7	2	Capital Income
4, 5, 6	3	Social Grants/ Benefit Income
3, 8	4	Other Income



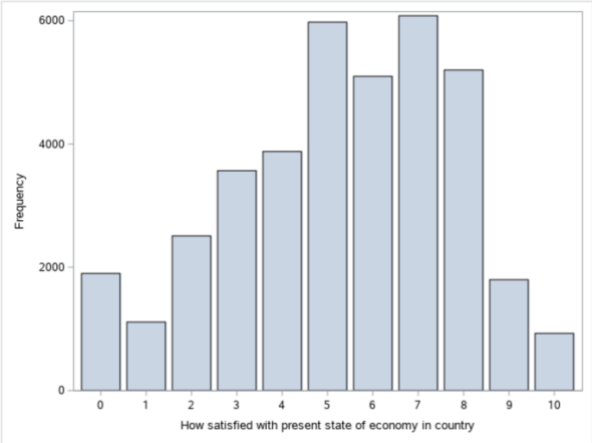
2.3. Explanatory Variables

Now we will describe 6 explanatory variables that we choose for our model. Those 6 variables are *stfeco*, *stfgov*, *evmar*, *nbthcld*, *hinctnta*, and *agea.*

2.3.1. Variable stfeco: How satisfied with present state of economy in country

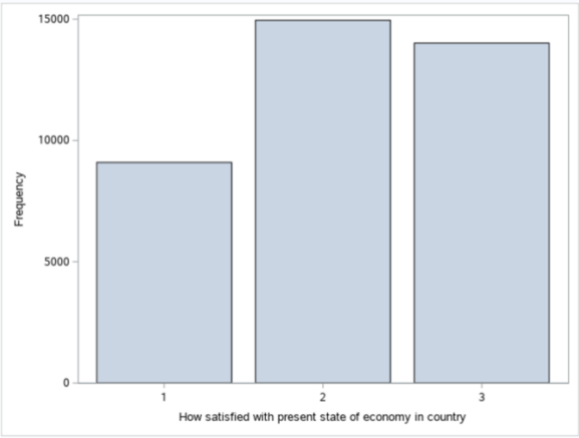
Possible answers for this variable are from 0 to 10 with "0" standing for "Extremely dissatisfied" and "10" standing for "Extremely satisfied". You can see on the left plot below that distribution of variable *stfeco*. To have better interpretation of modelling result later, we decided to group the levels of this variable into 3 following categories: "Low Satisfaction (1)", "Mid Satisfaction (2)", "High Satisfaction (3)". As per the table below, levels 0, 1, 2, and 3 are now equal to level 1 (low

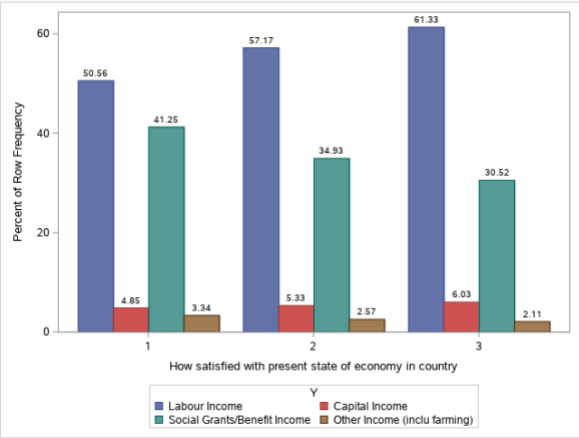
satisfaction); levels 4, 5, and 6 are now equal to level 2 (mid satisfaction); and levels 7, 8, 9, and 10 are now equal to level 3 (high satisfaction).



Previous Levels	New Levels	Categories
0~3	1	Low Satisfaction
4~6	2	Mid Satisfaction
7~10	3	High Satisfaction

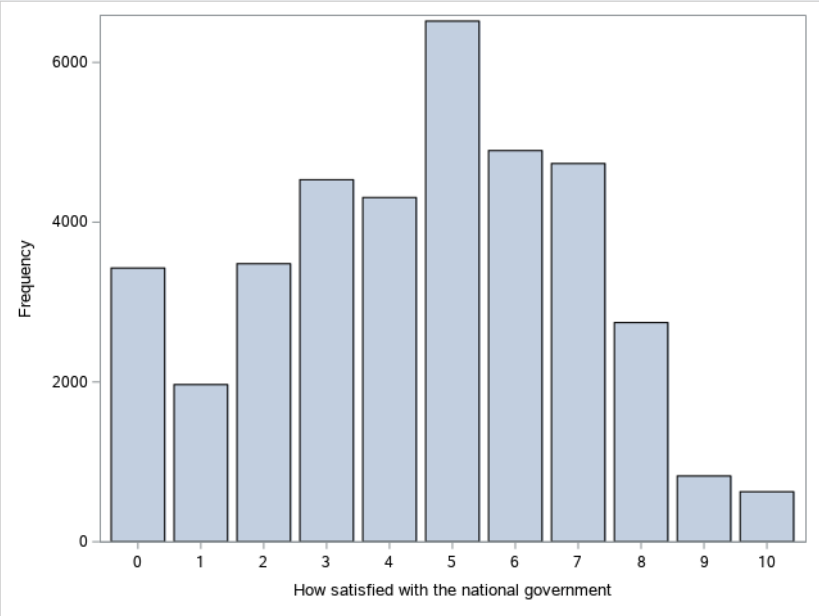
You can see the distribution of variable *stfeco* after categorizing it on the left plot below. The biggest number of respondents – around 15000 respondents – have mid satisfaction of present state of economy in their country and the lowest number of respondents – around 9000 respondents – have low satisfaction. However, if we look on the right plot below with distribution of our target variable *hincsrca* on variable *stfeco,* it can be noted that the group with low satisfaction with present state of economy in their country has the biggest percent of income from Social Grants/Benefit Income. So , we can say that the proportion of social grants/benefit income as the main source is higher in the group of people who have lesser satisfaction on their country's economy.



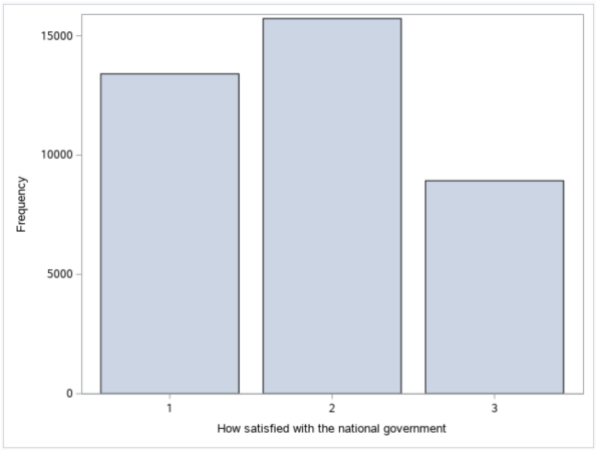


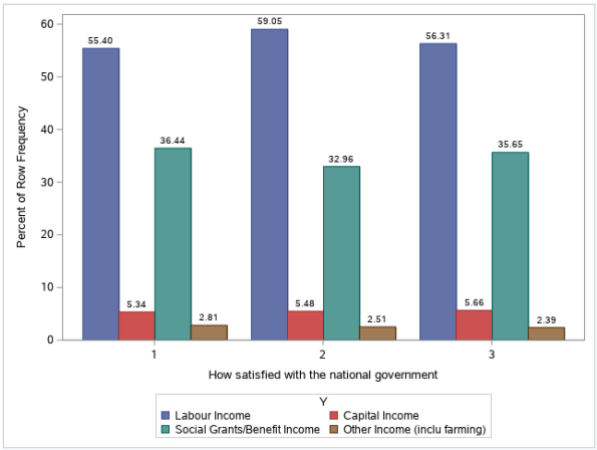
2.3.2. Variable stfgov: How satisfied with the national government

You can see on the plot below the distribution of variable *stfgov* with *"*0" standing for "Extremely dissatisfied" and "10" standing for "Extremely satisfied".



As in the previous case, after checking the distribution of variable *stfgov*, we decided to group the levels of this variable into same 3 categories: "Low Satisfaction", "Mid Satisfaction", "High Satisfaction". Levels 0 - 3 are equal to level 1 (low satisfaction); levels 4 - 6 are equal to level 2 (mid satisfaction); and levels 7 - 10 are equal to level 3 (high satisfaction). As per the distribution on the left plot below majority of respondents fall into either mid satisfaction or low satisfaction category. If we make discriminatory performance analysis (see right plot below), we can see that the distribution of income sources within these three groups is stable and with relatively small changes in proportion.



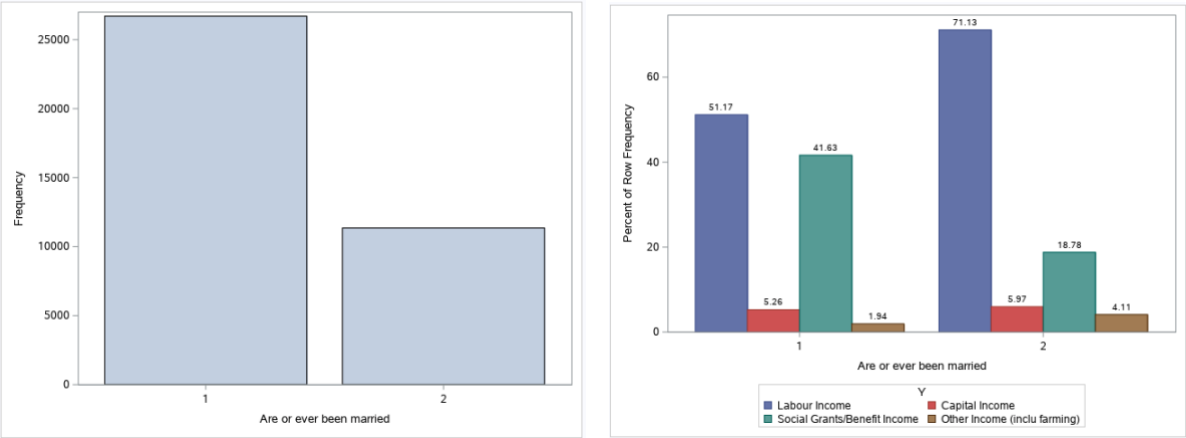


2.3.3. Variable evmar: Are you or have you ever been married

Possible answers for this variable are "1" and "2" standing for "Yes" and "No" respectively.

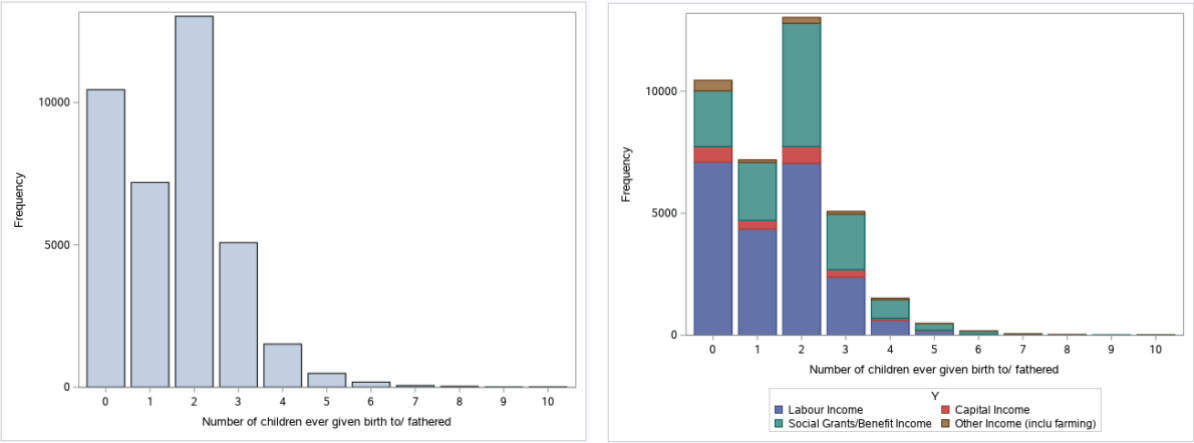
This analysis gives us the idea of distribution of respondents that have been married and distribution of income sources for respondents that have never been married and those that are married or were

married. According to the left plot below overwhelming majority of respondents are or were married, whereas number of not married respondents is significantly less (around 26500 respondents versus around 12000 respondents respectively). When looking at the distribution of our target variable on the right plot below we can see that not married people have significantly higher proportion of "Labour income" to "Social Grants/Benefit Income" than married people. However, for married people or people that have been married this distribution looks different. Percentage of income from social grants or benefits for married people is more than twice higher than for not married and close to percentage of income from labour (41.63% vs 51.17% respectively).



2.3.4. Variable nbthcld: Number of children ever given birth to/ fathered

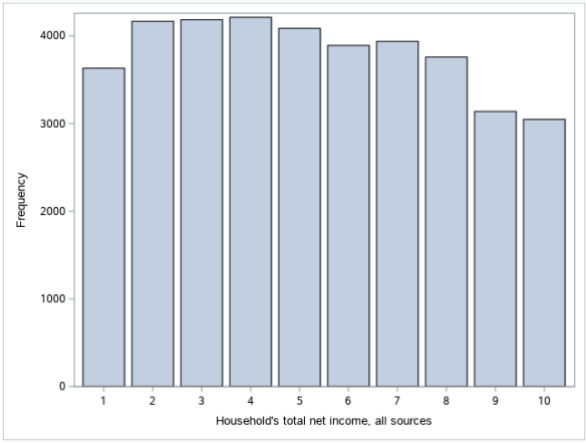
As can be seen on the graphs below, the majority of respondents have two children. Respondents with two children are almost twice as frequent as respondents with one or three children. Interestingly, the higher the number of children the higher the proportion of income from social grants or benefits. People with no children have much lower proportion of income from social grants or benefits than people with children.

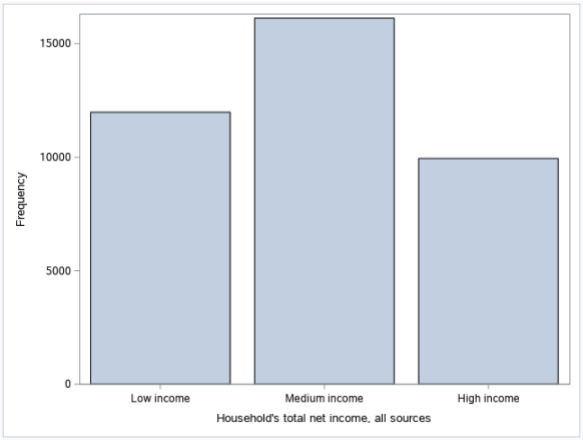


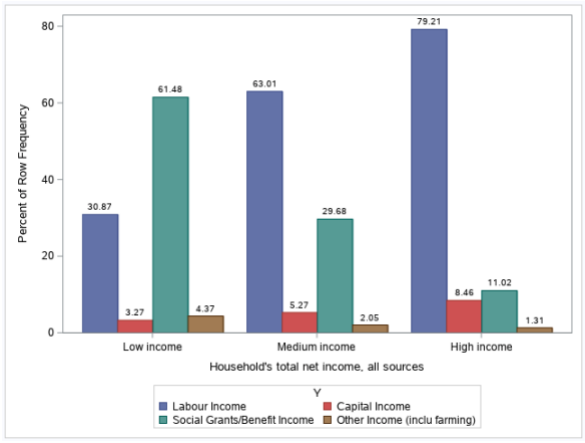
2.3.5. Variable hinctnta: Household's total net income, all sources?

Possible answers are from J – 1st decile (1 on the plot) to H – 10th decile (10 on the plot) representing smallest and highest total net income respectively. We decided to group these values into three categories: low income (1st, 2nd, 3rd), medium income (4th, 5th, 6th, and 7th deciles), and high income (8th, 9th, and 10th deciles).

As can be seen on the graphs below, majority of respondents fall into "Medium income" category. Interestingly, for people with low income "Social Grants/Benefit Income" is higher than income from labour (61.48% vs 30.87% respectively). The higher the income group, the lower percent of income comes from social grants or benefits.

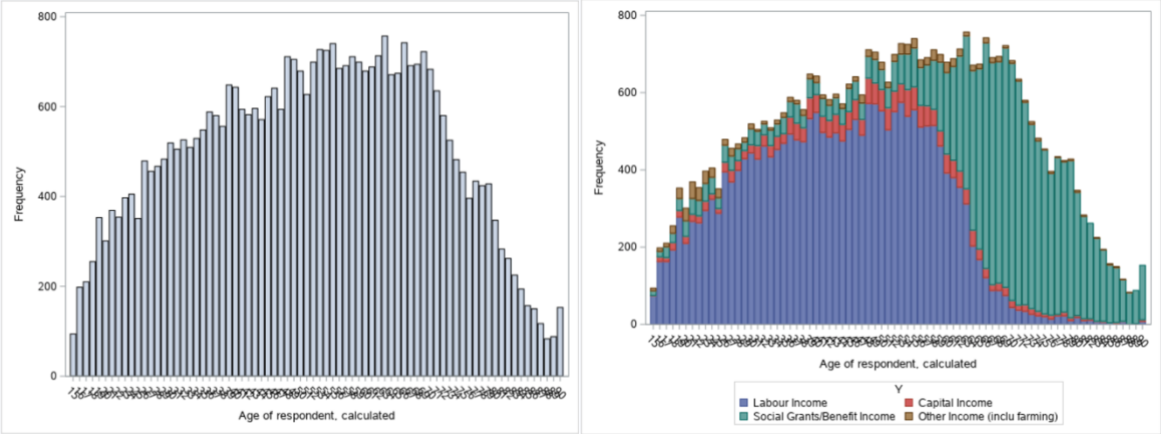






2.3.6. Variable agea: Age of respondent, calculated

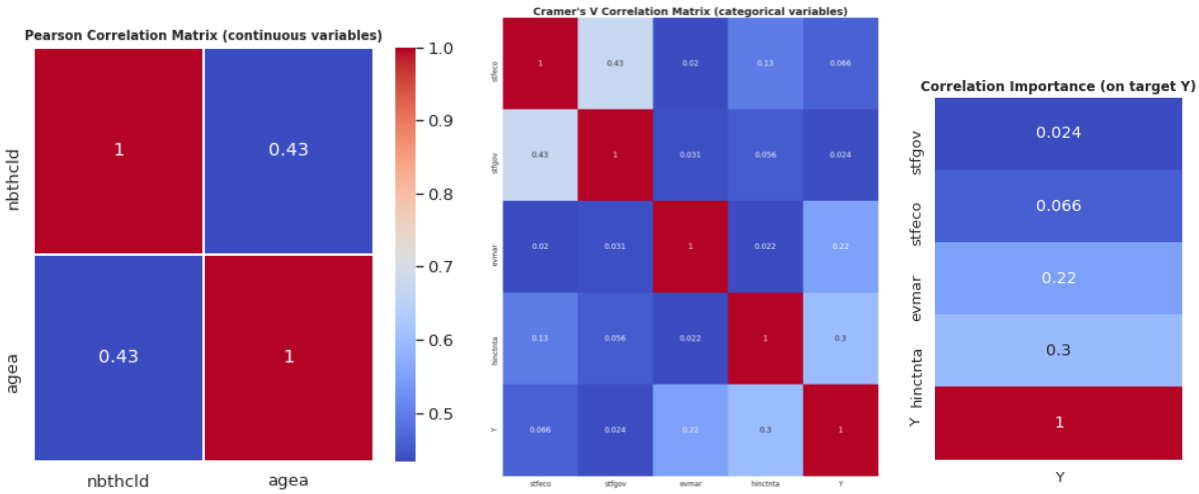
The distribution of age can be seen below. Percentage of income from social grants or benefits starts slowly increasing for respondents in their fifties and continue to increase even more drastically in their sixties and above. For respondents that are seventy years old and older the main source of income is almost solely income from social grants or benefits.



3. Substantive Analysis

3.1. Collinearity assessment

Before we build our model, we need to check if there is any possible correlation between our X explanatory variables. It is important since collinearity in logistic regression may lead to overestimation of standard errors and regression coefficients. Since we have both numerical and categorical variables in our dataset, we will use 2 different coefficients: 1) numerical variables' correlation with Pearson Correlation Coefficients; 2) categorical variables' correlation with Cramer's V coefficients. For this Collinearity assessment part, we utilized 'Python' to see more intuitive correlation matrix as below (code can be found in the appendix).



As we can see in the plot above, there is no any severe correlation both in numerical and categorical variables. Even though it's low correlation with 0.43, stfgov and stfeco also shows a bit of correlation. It is natural that they are correlated since a national government literally influence a lot on economy of the country based on their policy(and vice versa). Among categorical variables, the hinctnta (total net income) is mostly related to our target variable Y. It is a matter of course that different level of

total net income is most likely to be correlated with main source of household income among all our X variables.

3.2. Stepwise Variable Selection

As building a model, we used stepwise variable selection method. All the explanatory variables pass through the stepwise selection procedure with very low p-value (<0.05). Moreover, the result table of global null hypothesis test shows all p-values lower than 0.05 which means that we reject the null hypothesis that "all the explanatory effects can be eliminated and the model can contain only intercepts". So, our model doesn't need to drop any X variables.

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	agea		3	1	15113.5594		<.0001	Age of respondent, calculated
2	hinctnta		6	2	4404.9287		<.0001	Household's total net income, all sources
3	evmar		3	3	287.8038		<.0001	Are or ever been married
4	stfeco		6	4	20.9975		0.0018	How satisfied with present state of economy in country
5	stfgov		6	5	16.6527		0.0106	How satisfied with the national government
6	nbthcld		3	6	10.9764		0.0119	Number of children ever given birth to/ fathered

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	23639.7787	27	<.0001
Score	18996.3239	27	<.0001
Wald	10711.1738	27	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
evmar	3	244.7939	<.0001
hinctnta	6	3357.1523	<.0001
stfeco	6	26.3546	0.0002
stfgov	6	16.8194	0.0100
nbthcld	3	10.9770	0.0119
agea	3	7587.5352	<.0001

3.3. Model estimates

For Maximum Likelihood Estimate, there are some variables with higher p-value than 0.05, which means they are statistically insignificant. In the result table below, the variables with red box are insignificant ones. Unfortunately, most categories of our variable stfeco and stfgov have been rejected. However, we can see that the variable stfeco's 1st category which represents 'Low satisfaction on country's present economy state' survived. And it literally proves our hypothesis by showing the positive estimates. Let's take a look into the Odds Ratio for detailed interpretation.

Analysis of Maximum Likelihood Estimates								
Parameter		Y	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		Capital Income	1	-3.8091	0.1159	1080.7501	<.0001	0.022
Intercept		Other Income (inclu farming)	1	-4.3778	0.1643	709.7288	<.0001	0.013
Intercept		Social Grants/Benefit Income	1	-8.2283	0.1003	6736.6096	<.0001	0.000
evmar	2	Capital Income	1	0.2812	0.0647	18.9149	<.0001	1.325
evmar	2	Other Income (inclu farming)	1	0.5491	0.0904	36.8773	<.0001	1.732
evmar	2	Social Grants/Benefit Income	1	0.7212	0.0479	227.0390	<.0001	2.057
hinctnta	High income	Capital Income	1	0.2698	0.0516	27.3609	<.0001	1.310
hinctnta	High income	Other Income (inclu farming)	1	-0.6248	0.1051	35.3152	<.0001	0.535
hinctnta	High income	Social Grants/Benefit Income	1	-0.9305	0.0442	443.5885	<.0001	0.394
hinctnta	Low income	Capital Income	1	0.2905	0.0649	20.0375	<.0001	1.337
hinctnta	Low income	Other Income (inclu farming)	1	1.4285	0.0740	372.2969	<.0001	4.172
hinctnta	Low income	Social Grants/Benefit Income	1	1.5377	0.0371	1716.4303	<.0001	4.654
stfeco	1	Capital Income	1	-0.00914	0.0685	0.0178	0.8939	0.991
stfeco	1	Other Income (inclu farming)	1	0.2686	0.0905	8.8108	0.0030	1.308
stfeco	1	Social Grants/Benefit Income	1	0.1009	0.0445	5.1377	0.0234	1.106
stfeco	3	Capital Income	1	0.0218	0.0559	0.1521	0.6966	1.022
stfeco	3	Other Income (inclu farming)	1	-0.1618	0.0865	3.4983	0.0614	0.851
stfeco	3	Social Grants/Benefit Income	1	-0.0678	0.0398	2.9015	0.0885	0.934
stfgov	1	Capital Income	1	0.0313	0.0594	0.2777	0.5982	1.032
stfgov	1	Other Income (inclu farming)	1	-0.0780	0.0855	0.8311	0.3620	0.925
stfgov	1	Social Grants/Benefit Income	1	0.0103	0.0406	0.0647	0.7992	1.010
stfgov	3	Capital Income	1	0.0616	0.0624	0.9757	0.3233	1.064
stfgov	3	Other Income (inclu farming)	1	0.1497	0.0941	2.5301	0.1117	1.161
stfgov	3	Social Grants/Benefit Income	1	0.1569	0.0440	12.6896	0.0004	1.170
nbthcld		Capital Income	1	0.0177	0.0221	0.6394	0.4239	1.018
nbthcld		Other Income (inclu farming)	1	0.0765	0.0303	6.3682	0.0116	1.080
nbthcld		Social Grants/Benefit Income	1	-0.0206	0.0141	2.1137	0.1460	0.980
agea		Capital Income	1	0.0254	0.00191	176.4333	<.0001	1.026
agea		Other Income (inclu farming)	1	0.0138	0.00273	25.7351	<.0001	1.014
agea		Social Grants/Benefit Income	1	0.1295	0.00149	7530.6920	<.0001	1.138

3.1. Odds ratios with CI & Interpretations

In this odds ratios table, we also need to disregard the variables with red box since they are statistically insignificant because we checked their p-value is higher than 0.05 in the previous ML estimates table and they are including '1' in their Confidence Intervals.

For the variable 'evmar'(ever been married?, 1= married, 2 = not married), the people who didn't ever get married are twice (2.057) more likely to have social grants/benefit income as their main source of household income (than the labour income which was our parameter for Y variable) comparing to the people who ever got married. This result is unexpected, but makes sense considering that EU countries have various type of social funding and grants for the youth and people who didn't ever get married are highly likely to be youth in general.

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals					
Effect	Y	Unit	Estimate	95% Confidence Limits	
evmar 2 vs 1	Capital Income	1.0000	1.325	1.167	1.503
evmar 2 vs 1	Other Income (inclu farming)	1.0000	1.732	1.450	2.067
evmar 2 vs 1	Social Grants/Benefit Income	1.0000	2.057	1.873	2.259
hinctnta High income vs Medium income	Capital Income	1.0000	1.310	1.184	1.449
hinctnta High income vs Medium income	Other Income (inclu farming)	1.0000	0.535	0.434	0.656
hinctnta High income vs Medium income	Social Grants/Benefit Income	1.0000	0.394	0.362	0.430
hinctnta Low income vs Medium income	Capital Income	1.0000	1.337	1.176	1.517
hinctnta Low income vs Medium income	Other Income (inclu farming)	1.0000	4.172	3.611	4.827
hinctnta Low income vs Medium income	Social Grants/Benefit Income	1.0000	4.654	4.328	5.006
stfeco 1 vs 2	Capital Income	1.0000	0.991	0.866	1.133
stfeco 1 vs 2	Other Income (inclu farming)	1.0000	1.308	1.095	1.562
stfeco 1 vs 2	Social Grants/Benefit Income	1.0000	1.106	1.014	1.207
stfeco 3 vs 2	Capital Income	1.0000	1.022	0.916	1.140
stfeco 3 vs 2	Other Income (inclu farming)	1.0000	0.851	0.718	1.007
stfeco 3 vs 2	Social Grants/Benefit Income	1.0000	0.934	0.864	1.010
stfgov 1 vs 2	Capital Income	1.0000	1.032	0.918	1.159
stfgov 1 vs 2	Other Income (inclu farming)	1.0000	0.925	0.782	1.093
stfgov 1 vs 2	Social Grants/Benefit Income	1.0000	1.010	0.933	1.094
stfgov 3 vs 2	Capital Income	1.0000	1.064	0.941	1.201
stfgov 3 vs 2	Other Income (inclu farming)	1.0000	1.161	0.965	1.395
stfgov 3 vs 2	Social Grants/Benefit Income	1.0000	1.170	1.073	1.275
nbthcld	Capital Income	1.0000	1.018	0.974	1.063
nbthcld	Other Income (inclu farming)	1.0000	1.080	1.016	1.145
nbthcld	Social Grants/Benefit Income	1.0000	0.980	0.953	1.007
agea	Capital Income	1.0000	1.026	1.022	1.030
agea	Other Income (inclu farming)	1.0000	1.014	1.009	1.019
agea	Social Grants/Benefit Income	1.0000	1.138	1.135	1.142

For the variable 'hinctnta'(total net income grouped as 3 levels), we can see the interesting result that people with Low Income are 4~4.6 times more likely to receive other income or social grants/benefit income as their main source (than the labour income) comparing to the people with Medium Income level. Moreover, we can see that people with high income comparing to the ones with medium

income are 1.3 times more likely to have Capital Income than the labor income, and 60% less likely to have social grants/benefit income. This result also can be interpretated that the people who have higher income are more prone to invest or have more entrepreneurship to think of new business to increase their income more exponentially and efficiently.

For the variable 'stfeco(satisfaction level on current economy of their country)', the people with low satisfaction on their economy comparing to the people with medium satisfaction are 10.6% more likely to have social grants/benefit income and 30.8% more likely to have other income including farming (than the labour income) as their main source of income. These results prove our hypothesis "People who have lesser satisfaction with present state of economy in their country are more likely to have Social Grants/Benefits Income as their main source of income" is true.

For the variable 'agea (age)', it shows very similar outcome for 3 categories of our Y target variable. Still, the people who are 1 year older are 13.8% more likely to have social grants/benefit income than the labour income.

3.2. Model diagnostics

The metrics such as AIC, SC, -2 Log L will be useful when we compare this model to the other models. Usually, the lower these values are, the better the model is. For R-Square, the higher this value is, the better the model is (in terms of prediction power). In our project, all these metrics are not utilized that much as we didn't build other model to compare, but they were actually used in the process of Variable Selection with Step wise method to compare models by adding and deleting explanatory variables.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	71602.996	48017.218
SC	71628.637	48273.620
-2 Log L	71596.996	47957.218

R-Square	0.4627	Max-rescaled R-Square	0.5459

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	20909.8349	29E3	0.7263	1.0000
Pearson	42788.5199	29E3	1.4862	<.0001

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
582.3904	24	<.0001

For our model, we cannot trust the result of Deviance and Pearson Goodness-of-Fit Statistics since our model contains quite many explanatory variables, even have 3 categories per each variable in average, and includes 2 continuous variables (age and number of child). Therefore, we moved on to check Hosmer and Lemeshow Goodness-of-Fit test result. However, our model actually didn't pass the Hosmer-Lemeshow test with showing very small p-value, which means our model doesn't have a good fit. However, according to one of academic papers, this H-L test has some issue in that it is highly likely to cause the rejection of the hypothesis of perfect fit with very low p-value in large data

(Nattino, Pennell and Lemeshow, 2020). Considering that our data sample is 38053 obs, we can say the test result showing rejection may have influenced from size of our sample. In this case, to pass through this test, we can do the test again with decreased subsample of our data or use a parameter which doesn't depend on the sample size as proposed in the above paper. Even though we didn't proceed additional test again this time due to lack of skills and time, it was at least meaningful to know the limit of H-L test in specific cases and what we can do to deal with it.

4. Extended research on the topic (Innovative aspect)

Throughout the project, we focused on identifying the relation between subjective opinion toward the current state of country's economy and (type of) main source of household income. Our main X variable 'stfeco' represents the subjective opinion in that it's about 'how people perceive and feel about the present economy in their own country'. Taking things further, we started to be curious if the result from our multinomial logistic regression model and our hypothesis (which turns out to be true) also applies same in all countries regardless of 'objective' state of their economy.

Therefore, we decided to compare the countries with relatively higher GDP per capita and lower GDP per capita. To have enough amount of obs (samples) and balance in our dataset, we grouped some countries into 2 groups based on Rank of real GDP per capita 2019 data from Eurostat (https://ec.europa.eu/eurostat/databrowser/view/sdg_08_10/default/table?lang=en) as follows:

(relatively) Top GDP per capita countries *- Austria, Switzerland, Denmark, Norway*

(relatively) Low GDP per capita countries *- Italy, Czechia, Spain, Estonia*

We tried to balance the sample size of both dataset with 5637 obs for top countries group and 5933 obs for low groups. For top country group's multinomial logistic regression, variable stfgov has been dropped and other 5 variables are included in the model based on the result of stepwise variable selection. For low country group's multinomial logistic regression, variable stfgov and nbthcld have been dropped and other 4 variables are included in the model based on the result of stepwise variable selection.

For the result (ML Coefficients, Odds Ratio) comparison, we will only focus on the 'stfeco' part of result table. Original table which contains all variables can be found by running code in the Appendix.

In top countries group, the people with low satisfaction(1) toward their country's present economy are 60% more likely to have social grants/benefit income as their main source than the labour income comparing to the medium satisfaction(2) group (which was our reference parameter). At the same time, the people with higher satisfaction toward economy in top countries group are 27% less likely

to have social grants/benefit income than the labour income comparing to the people with medium satisfaction.

In low countries group, the people with low satisfaction(1) showed 27% more chance to have social grants/benefit income than the labour income comparing to the medium satisfied people.

Based on these outcomes, we can conclude that our hypothesis "People who have lesser satisfaction with present state of economy in their country are more likely to have Social Grants/Benefits Income as their main source of income" is true regardless of the countries' different and objective economic level (state) measured by real GDP per capita.

Result of Top Countries Group

Analysis of Maximum Likelihood Estimates

Parameter		Y	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		Capital Income	1	-5.2283	0.3540	218.1150	<.0001	0.005
Intercept		Other Income (inclu farming)	1	-4.0901	0.5051	65.5762	<.0001	0.017
Intercept		Social Grants/Benefit Income	1	-8.0149	0.2583	963.0408	<.0001	0.000
stfeco	1	Capital Income	1	0.3071	0.3378	0.8260	0.3634	1.359
stfeco	1	Other Income (inclu farming)	1	-0.5340	0.5474	0.9517	0.3293	0.586
stfeco	1	Social Grants/Benefit Income	1	0.4716	0.2022	5.4382	0.0197	1.602
stfeco	3	Capital Income	1	-0.1296	0.1645	0.6209	0.4307	0.878
stfeco	3	Other Income (inclu farming)	1	-0.1668	0.2199	0.5756	0.4480	0.846
stfeco	3	Social Grants/Benefit Income	1	-0.3171	0.0993	10.1890	0.0014	0.728

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals

Effect	Y	Unit	Estimate	95% Confidence Limits	
stfeco 1 vs 2	Capital Income	1.0000	1.359	0.672	2.554
stfeco 1 vs 2	Other Income (inclu farming)	1.0000	0.586	0.170	1.540
stfeco 1 vs 2	Social Grants/Benefit Income	1.0000	1.602	1.077	2.381
stfeco 3 vs 2	Capital Income	1.0000	0.878	0.641	1.223
stfeco 3 vs 2	Other Income (inclu farming)	1.0000	0.846	0.556	1.320
stfeco 3 vs 2	Social Grants/Benefit Income	1.0000	0.728	0.599	0.885

Result of Low Countries Group

Analysis of Maximum Likelihood Estimates

Parameter		Y	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		Capital Income	1	-2.7619	0.2503	121.7643	<.0001	0.063
Intercept		Other Income (inclu farming)	1	-3.9514	0.4443	79.1085	<.0001	0.019
Intercept		Social Grants/Benefit Income	1	-8.7343	0.2590	1137.0575	<.0001	0.000
stfeco	1	Capital Income	1	0.2793	0.1237	5.1001	0.0239	1.322
stfeco	1	Other Income (inclu farming)	1	0.4651	0.2187	4.5217	0.0335	1.592
stfeco	1	Social Grants/Benefit Income	1	0.2406	0.1004	5.7419	0.0166	1.272
stfeco	3	Capital Income	1	-0.1295	0.1249	1.0739	0.3001	0.879
stfeco	3	Other Income (inclu farming)	1	0.2116	0.2258	0.8778	0.3488	1.236
stfeco	3	Social Grants/Benefit Income	1	-0.00896	0.1031	0.0075	0.9308	0.991

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals

Effect	Y	Unit	Estimate	95% Confidence Limits	
stfeco 1 vs 2	Capital Income	1.0000	1.322	1.036	1.682
stfeco 1 vs 2	Other Income (inclu farming)	1.0000	1.592	1.032	2.440
stfeco 1 vs 2	Social Grants/Benefit Income	1.0000	1.272	1.045	1.549
stfeco 3 vs 2	Capital Income	1.0000	0.879	0.686	1.120
stfeco 3 vs 2	Other Income (inclu farming)	1.0000	1.236	0.788	1.916
stfeco 3 vs 2	Social Grants/Benefit Income	1.0000	0.991	0.809	1.213

5. Conclusion

In conclusion, we have discussed the influence factors on main source of household income throughout EDA, checked multicollinearity using two different method, Pearson coefficient for numeric variables and Cramer's V for categorical variables, using Python. Based on all those steps, we created the multinomial logistic regression model that supported our hypothesis (true) about influence of 'satisfaction on one's own country's present economy' on 'type of main source of household income'. As an extension, we also experimented if this hypothesis can be applied regardless of the countries' objective state of present economy by comparing two groups of countries selected using the economic indicator 'GDP per capita'. Our hypothesis also turns out to be true in both country groups. Therefore, we can conclude that both subjective and objective aspect regarding the satisfaction level on country's current state of economy have significant influence on having 'Social Grants/Benefit income' as the main source of household income.

6. Bibliography

Nattino, G., Pennell, M. and Lemeshow, S., 2020. Assessing the goodness of fit of logistic regression models in large samples: A modification of the Hosmer-Lemeshow test. Biometrics, 76(2), pp.549-560.

Analyticsvidhya. 2015. What to do when Hosmer lemeshow test fails during Logistic regression?. [online] Available at: <https://discuss.analyticsvidhya.com/t/what-to-do-when-hosmer-lemeshow-test-fails-during-logistic-regression/2304/2> [Accessed 10 June 2021].

Statalist (The STATA forum). 2021. Hosmer Lemeshow test for large data. [online] Available at: <https://www.statalist.org/forums/forum/general-stata-discussion/general/1596105-hosmer-lemeshow-test-for-large-data> [Accessed 10 June 2021].

7. Appendix

https://github.com/jiyoungkimcr/SAS-Multinomial-Logistic-Regression-Project