Research topic & topic Parts:

Our topic will investigate the happiness and wellbeing for Thailand meaning that we want to predict whether a person is happy, quite happy or unhappy based on 5 explanatory variables to have a total of 6 variables as shown in the next pages.

The target population for the 7th wave of World Values Survey was eligible voters over the age of 18, where The sampling technique used in this survey was multi-stage systematic random sampling.

This Topic will consist of 2 sections as follows:

1. 1st section will be about performing Binary logistic regression model for our **binary** response variable (feeling of happiness) .
2. 2nd section will be about performing multinomial logistic regression model for our **Muli categorical** response variable (feeling of happiness).

To perform the binary logistic regression model I decided to combine 1 category with the other one to finally have 2 categories of the response variable and we will observe the effect of that merge on the outcome of the analysis.

Regarding the missing values or unspecified answers like (Don’t Know, Didn’t answer,…) I removed them since their frequency was around 3-15 times which wouldn’t affect the analysis significantly which leaves us with n = 1454.

The Analysis was done with Stata, R and Python.

1st section (Binary logistic regression)

Variables & Distributions:

1. Q46: Feeling Happy (Binary Response variable):

The variable initially had 3 categories (Happy , Quite Happy , Unhappy) but I will combine The categories (Quite Happy & Unhappy) in a new category (Unhappy) to have the following categories and the distribution of each:

With the following categories:

1- Happy.
2- Unhappy.

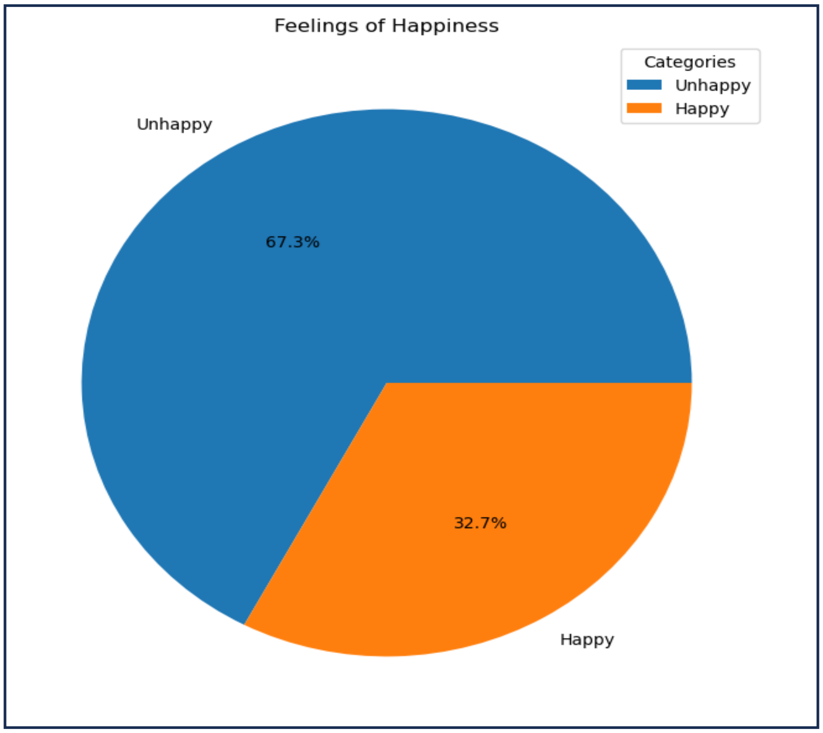


Figure 1: Feeling of Happiness Distribution

2. Q48: How much freedom of choice and control (Binary Variable, Explanatory Variable)

   With the following categories:

   1-Good control

   2-Poor control

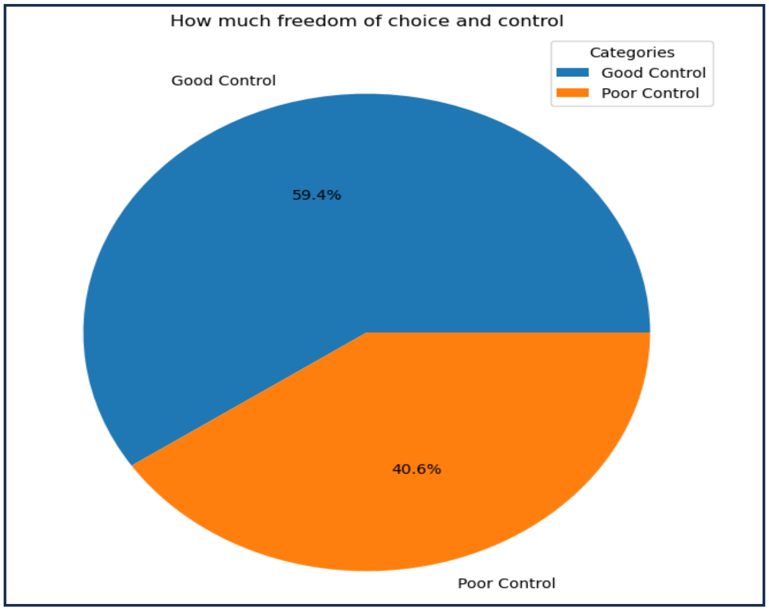


Figure 2: Freedom of Choice & Control Distribution

3. Q50: Satisfaction with financial situation of household (Binary variable, explanatory variable)

   With the following values:

   1-Satisfied

   2-Dissatisfied

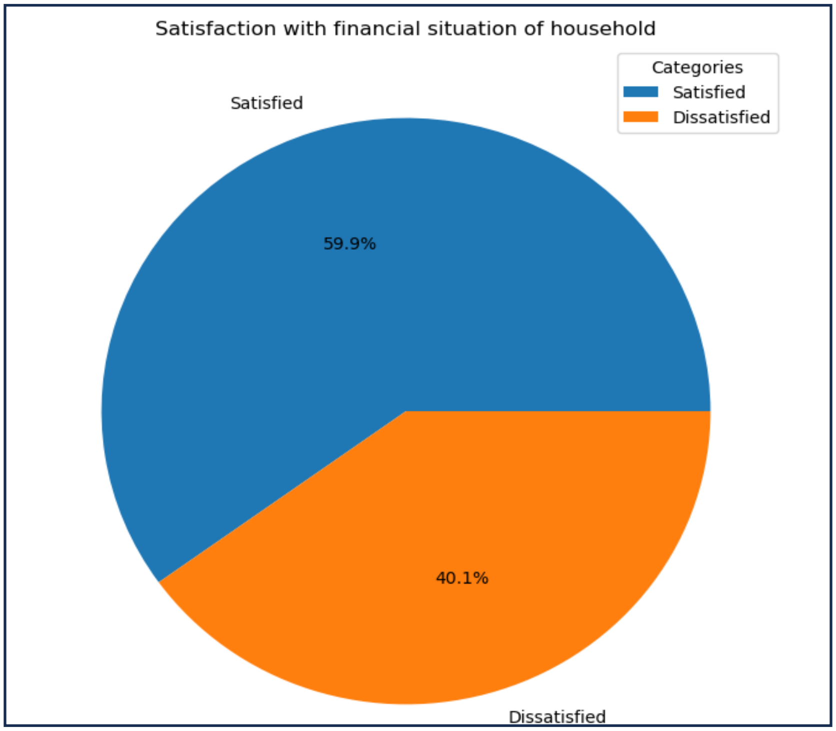


Figure 3: Satisfaction with financial situation of household Distribution

4. Q54: Frequency you/family (last 12 month): Gone without a cash income (Nominal variable, explanatory variable)

With the following categories:

1-Often

2-Somtimes

3-Never

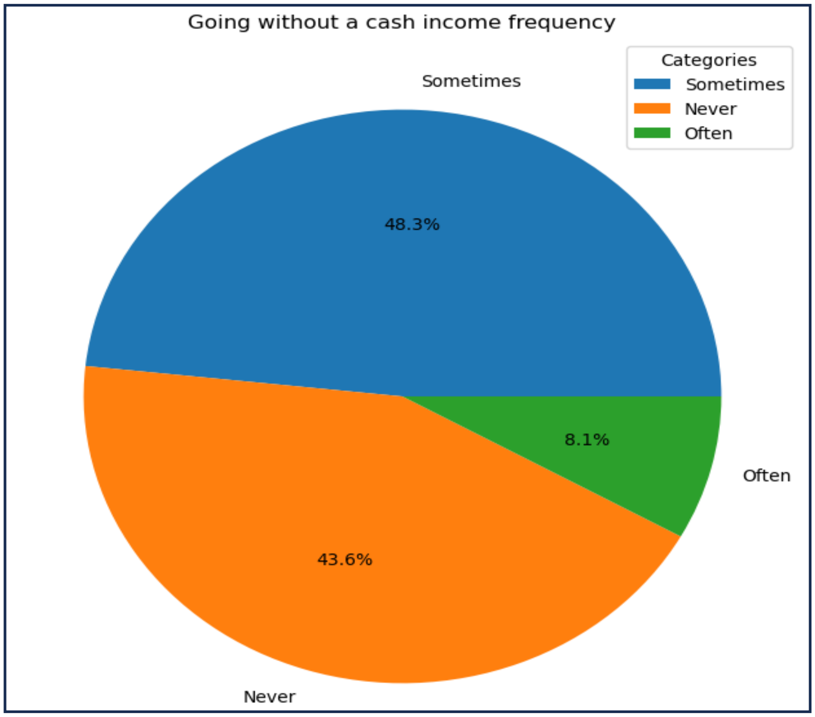


Figure 4: Going without a cash income Frequency.

5. Q158: Science and technology are making our lives healthier, easier, and more comfortable (Binary variable, explanatory variable)

With the following categories:

1- Agree.
2- Disagree.

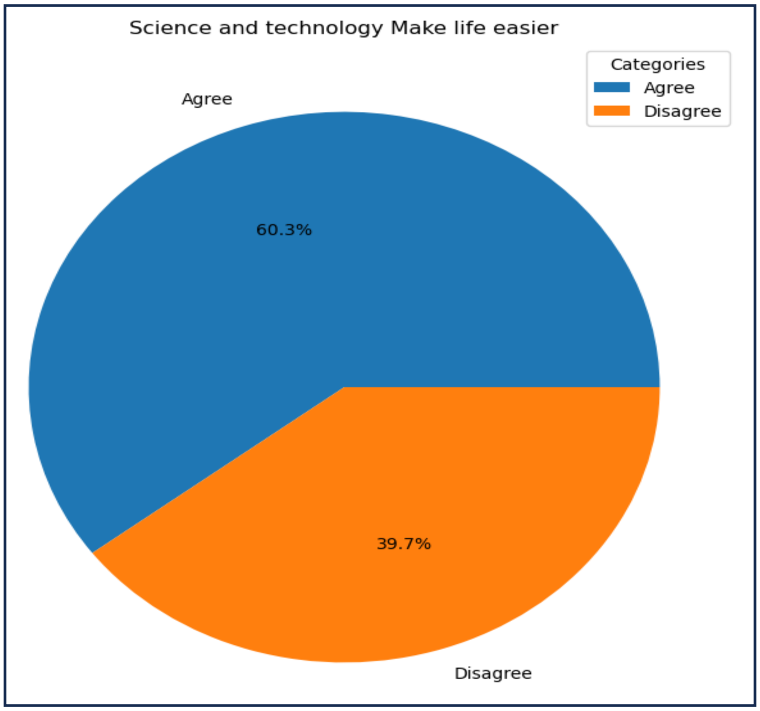


Figure 5: Science & technology make life easier Distribution.

6. Q260: Sex (Nominal variable, explanatory variable)

With values:

1- Male

2- Female

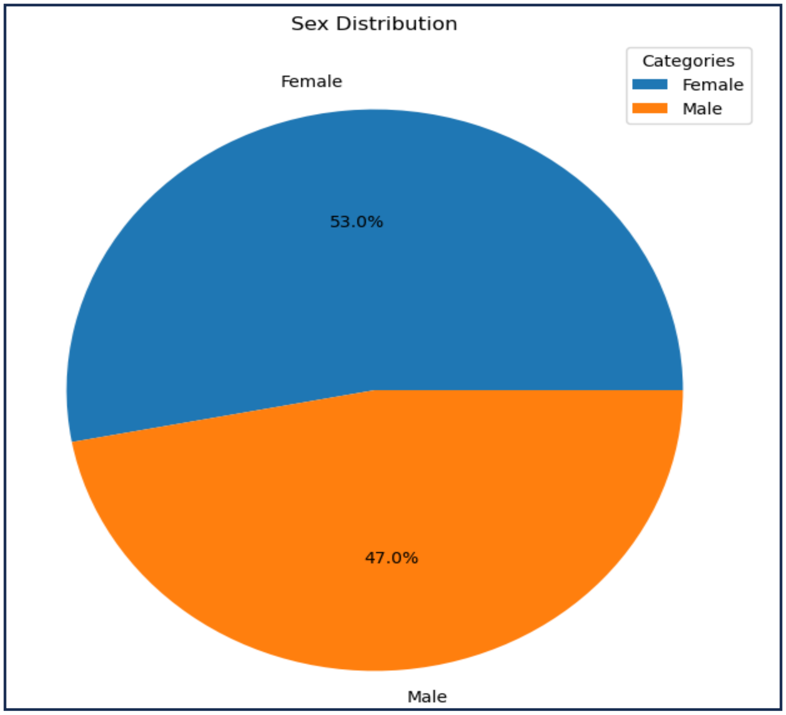


Figure 6: Sex Distribution

Fitting The Binary Logistic regression

The dataset is split into 2 groups:

- Training set: it's a portion of the entire dataset (70% of the data) used to fit and estimate the coefficients of the model.
- Test set: it's a portion of the entire dataset (30% of the data) used to validate the fitted model by driving the confusion matrix & the ROC curve.

Initially, I started with the main effects model (no interactions) to extract the significant variables from the 5 explanatory variables we have using **Backward** & **stepwise** regression with no interaction after which I will add the interactions for the extracted significant variables.

The initial model to estimate(1):

$$Log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \alpha + \beta_1\ Freedom\ control + \beta_2\ Financial\ satisfaction + \beta_3\ Frequency\ no\ cash\ income + \beta_4\ Science\ and\ technology + \beta_5\ Gender$$

Variable	Estimate (β_i)	Std.error	P-value
intercept	-1.148288	0.277840	0.0000358
Freedom control (Poor control)	-0.026389	0.150938	0.8612
Financial satisfaction(Dissatisfied)	0.278337	0.148784	0.0614
Frequency no cash income (Somtimes)	0.008482	0.261221	0.9741
Frequency no cash income (Never)	-0.672446	0.266224	0.0115
Science and technology(Disagree)	-0.570795	0.146075	0.0000932
Gender(Female)	-0.101031	0.135898	0.4572
AIC	1280		

Table 1: Initial model summary

In the model above we start with all the variables and then apply **Backward** & **stepwise** selection to it.

Main effects model to estimate (2):

After running both **Backward** & **stepwise** selection we reach the same result as follows:

$$Log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \alpha + \beta_1\ Financial\ satisfaction + \beta_2\ Frequency\ of\ no\ cash\ income + \beta_3\ Science\ and\ technology$$

Variable	Estimate (β_i)	Odds ratio(e^{β_i})	Std.error	P-value
intercept	-1.09238	0.3354	0.26619	0.0000407
***Financial satisfaction*(Dissatisfied)**	-0.26428	0.7677	0.14032	0.0596
***Frequency no cash income*(Somtimes)**	-0.01059	0.989	0.26100	0.9676
***Frequency no cash income*(Never)**	0.67611	1.966	0.26581	0.0110
***Science and technology*(Disagree)**	0.57613	1.77913	0.14342	0.0000589
AIC	1276			
McFadden Pseudo R Squared	0.02916542			

Table 2: refined model summary

Variable Name	Generalized variance inflation factor (GVIF)
Financial Satisfaction	1.029805
Frequency no cash income	1.090924
Science and technology	1.104018

Table 3: GVIF values

Table 2: Freedom control was removed as well as gender while financial satisfaction (Dissatisfied) is significant at ($\alpha = 0.1$) and therefore can be kept in the model but frequency no cash income (Sometimes) is insignificant but it will be kept in the model since one of the variables' categories is significant which is frequency no cash income(Never). We can also notice that McFadden Pseudo R squared value is relatively low as most of the Pseudo R squared measures are.

Table 3 is constructed to indicate whether we have significant multicollinearity or not, in our case all of the values are around 1 which indicate that we have little multicollinearity for all the 3 retained variables.

Interpretation:

1. (0.7677): The estimated odds for being happy is lower for financially dissatisfied people compared with financially satisfied people by almost 23.23% holding frequency no cash income & science and technology constant.
2. (0.989): The estimated odds for being happy isn't statistically significantly different for people who have sometimes gone without cash income compared with people who have usually gone without cash income.
3. (1.966): The estimated odds for being happy is greater for people who have never gone without cash income compared with people who have usually gone without cash income by almost 97% holding Financial satisfaction & science and technology constant.
4. (1.77913): The estimated odds for being happy is greater for people who disagree that Science and technology are making lives better compared with the people who agree by almost 78% holding Financial satisfaction & frequency no cash income constant.

After selecting the significant variables through stepwise selection, we can create interaction coefficients as shown in model (3).

Notice that gender was insignificant in the model (hence, it's removed) which goes along with what we have found in the 2-way contingency tables where gender was independent of feeling of happiness, therefore, we can say that knowing the gender till us no info about whether a person is happy or unhappy.

Interaction model to estimate (3):

$$Log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \alpha + \beta_1\ Financial\ satisfaction + \beta_2\ Frequency\ of\ no\ cash\ income + \beta_3\ Science\ and\ technology + \beta_4(Financial\ satisfaction) * (Frequency\ of\ no\ cash\ income)$$

Variable	Estimate (β_i)	Std.error	P-value
Intercept	-1.4713	0.3701	0.0000703
$\boldsymbol{Financial\ satisfaction}$(**Dissatisfied**)	0.4678	0.4796	0.32932
$\boldsymbol{Frequency\ no\ cash\ income}$(**Somtimes)**	0.4225	0.3799	0.26614
$\boldsymbol{Frequency\ no\ cash\ income}$(**Never)**	1.0519	0.3808	0.00574
$\boldsymbol{Science\ and\ technology}$(**Disagree**)	0.5854	0.1439	0.0000471
Financial_Satisfaction(Dissatisfied)* Frequency_No_Cash_Income(Somtimes)	-0.8701	0.5219	0.09550
Financial_Satisfaction(Dissatisfied)* Frequency_No_Cash_Income(Never)	-0.7310	0.5236	0.16269
AIC	1277.3		

Table 4: Interaction model summary

After trying different combinations of interaction terms I found that this is the best model so far in terms of:

- Lower AIC.
- More significant interaction coefficients.
- Simpler model (less parameters).

We can notice from the table above that Financial_Satisfaction(Dissatisfied)* Frequency_No_Cash_Income(Never) coefficient is insignificant while at ($\alpha = 0.05$) Financial_Satisfaction(Dissatisfied)* Frequency_No_Cash_Income(Somtimes) at ($\alpha = 0.1$) as well as financial satisfaction(Dissatisfied) which has become more insignificant.

Note to mention, after running stepwise selection I ended up with the main effects model which shows how better the main effects model(2) is.

Comparing the main effects model & interaction model:

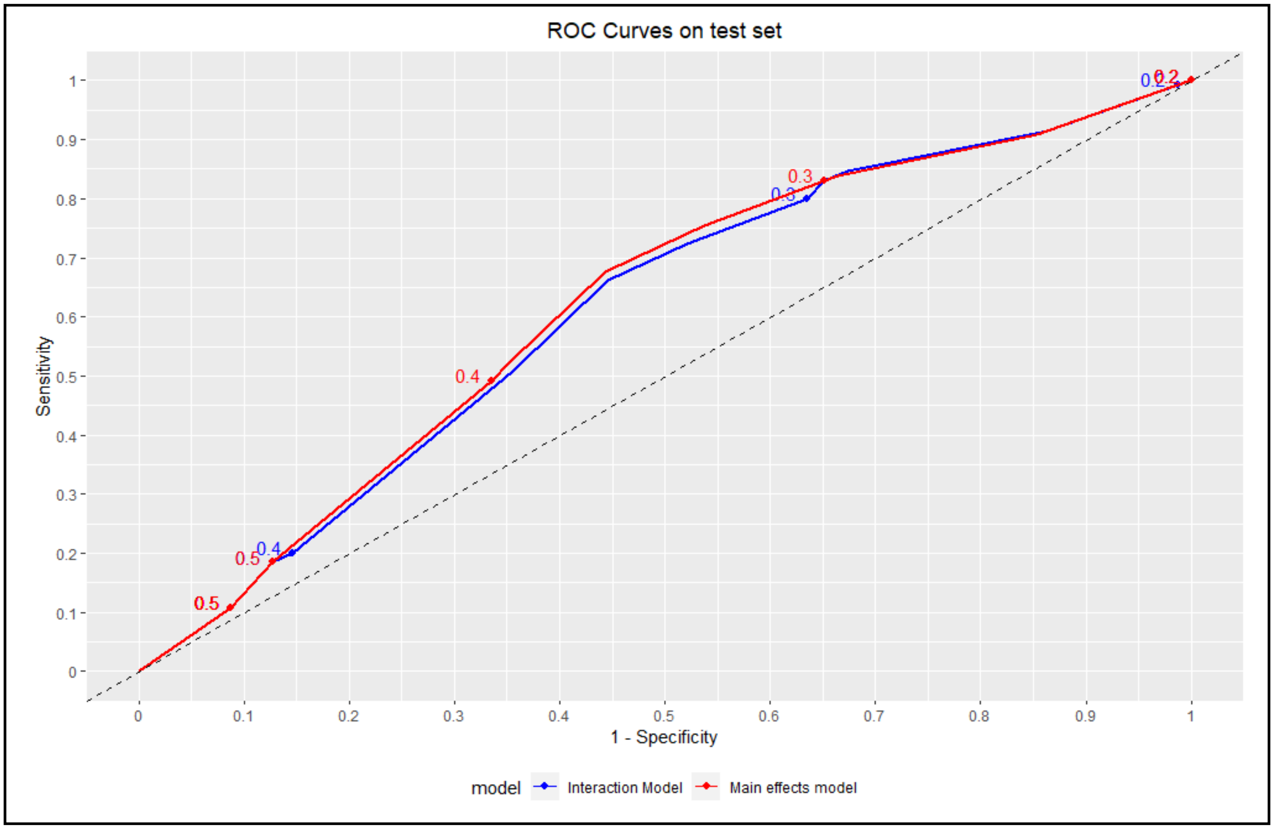


Figure 7:Main effects model & Interaction model ROC curves

Model	AUC
Main effects	0.6176
interaction	0.6098

Table 5: Area Under ROC curves for main effects model & interaction model

Figure 7 shows that the ROC curve for main effects model is slightly closer to the perfect model(0,1) with 1 for sensitivity and 0 specificity **than** the interaction model, and that was reflected on the area under both ROC curves in table 4 where AUC for main effects model is also slightly greater than AUC for interaction model, **hence**, we can say that both the interaction & main effects models are better than random assignment (Null model).

The main effects model prove to be the best out of all interactions models and that was based on the ROC curve & the AUC, hence, I will use the main effects model for prediction since it’s simpler and better in prediction compared with the interaction model.

Actual	Predicted	
	Happy (1)	Unhappy (0)
Happy (1)	88	42
Unhappy (0)	136	171
Accuracy	0.6	
Sensitivity	0.677	
Specificity	0.557	
Cutoff point	0.353695	

Table 6: Confusion matrix for main effects model on the test set

After comparing the classification table between the 2 models, The main effects model had higher accuracy, Sensitivity, Specificity than the interaction model.

Based on the best Cutoff point, the model correctly classified 60% of the observations correctly which is a low value indicating that we might need more explanatory variables to increase that proportion.

2^{nd} section (Multinomial logistic regression)

Variables & Distributions:

Q46: Feeling Happy (Response variable):

With the following categories:

1- Happy.

2- Quite happy.

3- Not happy.

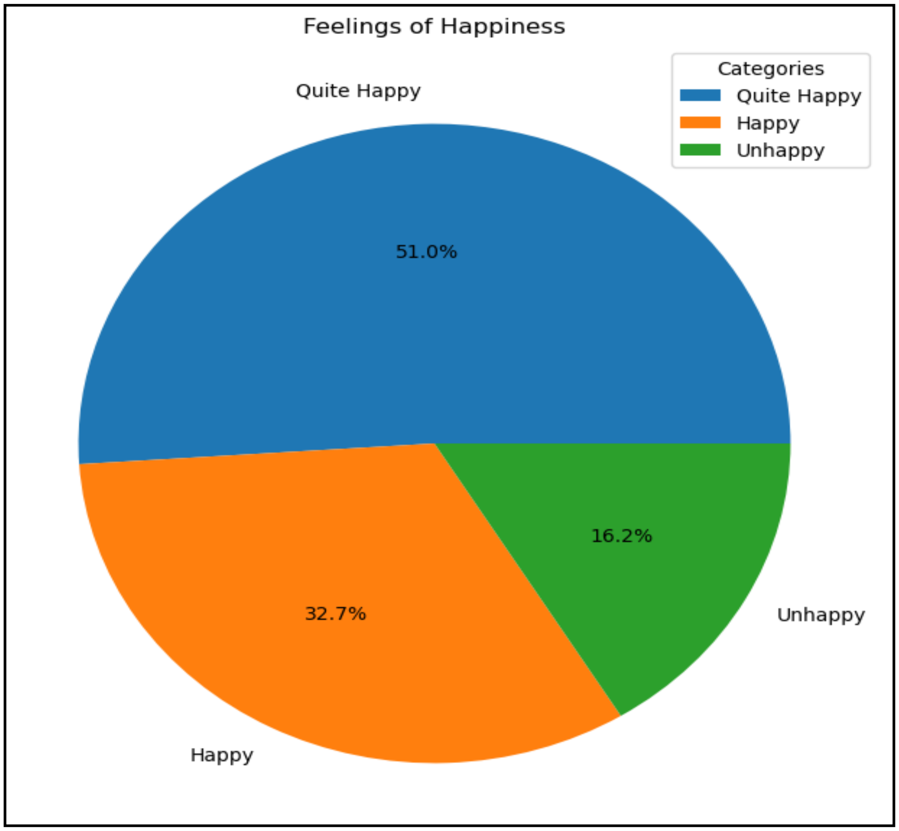


Figure 8: Feeling of Happiness Distribution

While the rest of the variables remain the same.

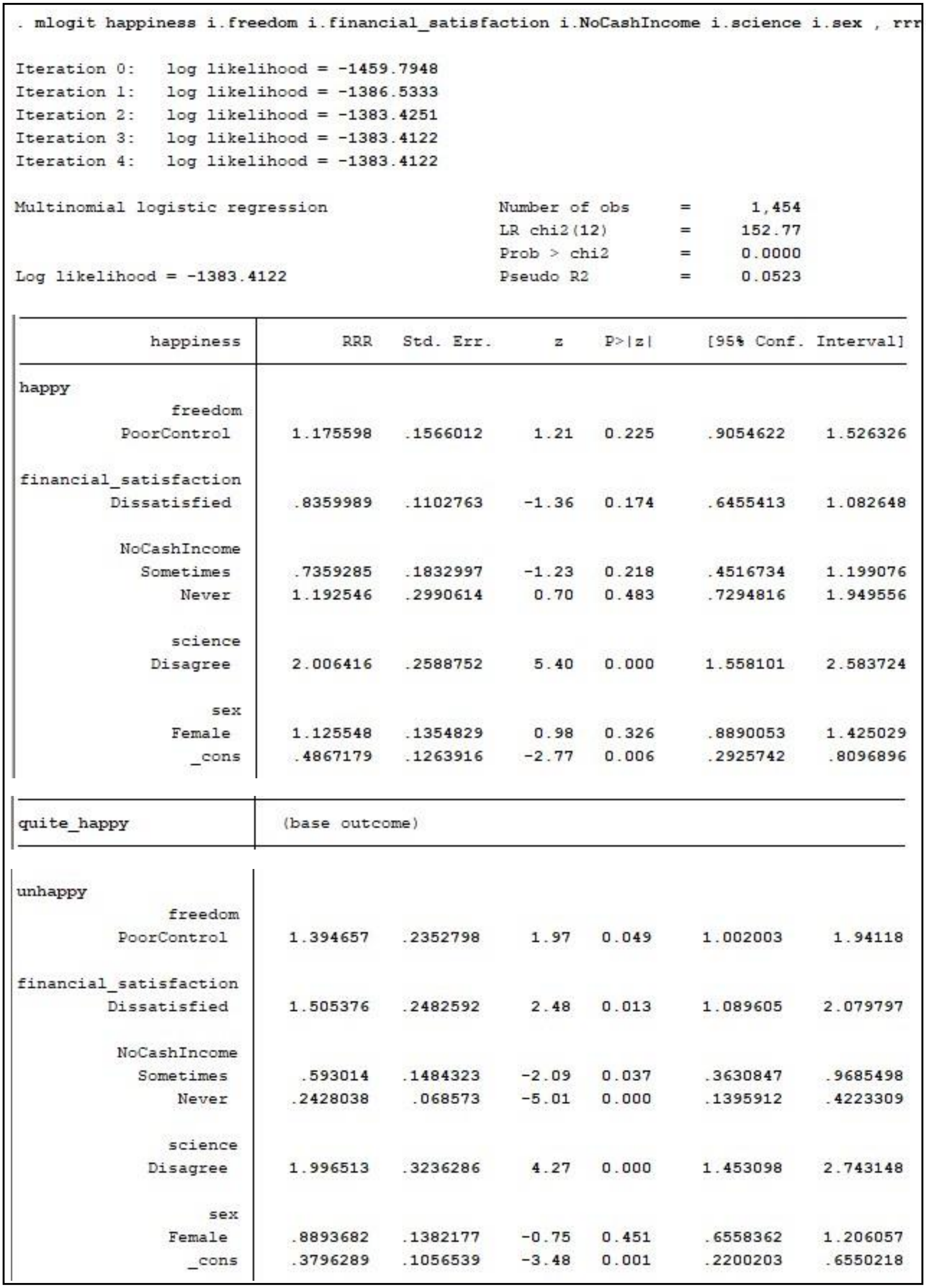
Testing proportional odds assumption

$$H_0: The\ coefficients\ are\ all\ equal\ across\ categories$$

P-value
2.315e-15

Therefore, we reject H_0 at $(\alpha = 0.05)$, we have no evidence to support the proportional odds assumption and we must use the multinomial logistic regression model.

Stata output for multinomial logistic regression



```
. mlogit happiness i.freedom i.financial_satisfaction i.NoCashIncome i.science i.sex , rrr

Iteration 0:   log likelihood = -1459.7948
Iteration 1:   log likelihood = -1386.5333
Iteration 2:   log likelihood = -1383.4251
Iteration 3:   log likelihood = -1383.4122
Iteration 4:   log likelihood = -1383.4122

Multinomial logistic regression                 Number of obs     =      1,454
                                                LR chi2(12)       =     152.77
                                                Prob > chi2       =     0.0000
Log likelihood = -1383.4122                     Pseudo R2         =     0.0523

------------------------------------------------------------------------------
             happiness |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------------+------------------------------------------------------
happy                  |
               freedom |
           PoorControl |   1.175598   .1566012     1.21   0.225     .9054622    1.526326
                       |
financial_satisfaction |
          Dissatisfied |   .8359989   .1102763    -1.36   0.174     .6455413    1.082648
                       |
          NoCashIncome |
             Sometimes |   .7359285   .1832997    -1.23   0.218     .4516734    1.199076
                 Never |   1.192546   .2990614     0.70   0.483     .7294816    1.949556
                       |
               science |
              Disagree |   2.006416   .2588752     5.40   0.000     1.558101    2.583724
                       |
                   sex |
                Female |   1.125548   .1354829     0.98   0.326     .8890053    1.425029
                 _cons |   .4867179   .1263916    -2.77   0.006     .2925742    .8096896
-----------------------+------------------------------------------------------
quite_happy            |  (base outcome)
-----------------------+------------------------------------------------------
unhappy                |
               freedom |
           PoorControl |   1.394657   .2352798     1.97   0.049     1.002003     1.94118
                       |
financial_satisfaction |
          Dissatisfied |   1.505376   .2482592     2.48   0.013     1.089605    2.079797
                       |
          NoCashIncome |
             Sometimes |    .593014   .1484323    -2.09   0.037     .3630847    .9685498
                 Never |   .2428038    .068573    -5.01   0.000     .1395912    .4223309
                       |
               science |
              Disagree |   1.996513   .3236286     4.27   0.000     1.453098    2.743148
                       |
                   sex |
                Female |   .8893682   .1382177    -0.75   0.451     .6558362    1.206057
                 _cons |   .3796289   .1056539    -3.48   0.001     .2200203    .6550218
```

Interpretation:

- The model is statistically significantly better than the null model since p-value is less than ($\alpha = 0.05$), where the pseudo r squared is low and almost equal 0.1
- Quite happy is taken as a baseline response category to compare with because it has the highest frequency as shown in the pie chart above (figure 8).
- Notice that Sex isn't statistically significant in all panels.
- The output shows the Relative Risk Ratio(RRR) which compares each category with both the reference explanatory variable category and the baseline response category(2 comparisons are made here)

1- In the first panel (Happy):

- The only significant coefficient is for the science variable which means that the estimated probability of being happy for people who disagree that science & technology make lives better is greater than (more likely) being quite happy by 100% compared to people who agree that science & technology make lives, holding other variables constant.

2- In the third panel (Unhappy):

- Freedom (poor control)(1.39): is barely significant at ($\alpha = 0.05$) which means that the estimated probability of being unhappy for people who have poor freedom of choice& control is greater than(more likely) being quite happy by 40% compared to people who have good freedom of choice& control, holding other variables constant.
- Financial_satisfaction (Dissatisfied)(1.50): the estimated probability of being unhappy for people who are financially dissatisfied is greater than (more likely) being quite happy by 50% compared to people who are financially satisfied, holding other variables constant.
- NoCashIncome (Sometimes) (0.59): the estimated probability of being unhappy for people who sometimes go without cash income is less than (less likely) being quite happy by 41% compared to people who usually go without cash income, holding other variables constant.

- NoCashIncome (Never) (0.242): the estimated probability of being unhappy for people who Never go without cash income is less than (less likely) being quite happy by 75% compared to people who usually go without cash income, holding other variables constant.
- Science (Disagree) (1.996): the estimated probability of being unhappy for people who Disagree that science & technology make lives better is greater than (more likely) being quite happy by almost 100% compared to people who agree that science & technology make lives better, holding other variables constant.

Conclusion:

To sum up, we have studied the relationship between the feeling of happiness in Thailand as a response variable based on the other explanatory variables (Freedom of choice & control, Satisfaction with Financial situation, Frequency of going without cash income , Science& technology make lives better and Sex) by using different measures of associations which showed that Sex has no statistically significant effect on feeling of happiness (with 95% confidence level) while the other variables showed influence on feeling of happiness.

Therefore, we found that the response variable is dependent on the previously mentioned variables. In addition, we studied the relationship between the feeling of happiness in Thailand with the mentioned explanator variables simultaneously through a Binary & Multinomial Logistic Regression Model.