Survey of Health, Aging and Retirement - Life Satisfaction

Stefano Amadori
Virginia Pagliero

November 5, 2022

1 Research question

The aim of the project is to understand impacts on living condition of elderly people and their families. In order to do so, the dataset gathers socio-demographic and health-related information from the Survey of Health, Ageing and Retirement in Europe[1] via SHARE project 1 in three waves (i.e. 2011, 2013 and 2015) which enables the study of how economic, social and health variables may affect condition of happiness of people in the sample[2].

To measure the level of happiness the dataset provides a dummy variable $life\_sat$, that refers to the overall level of self perceived life satisfaction, whereas the main independent regressor is represented by *income* that collects the total household income in ten thousands of euros.

We present the model of specification with the variables of interest and the vector $\mathbf{X}$, that collects socio-demographic and health-related features and it is used as control variable. Later, we will define more precisely the vector $\mathbf{X}$ in order to include the most relevant information and explain a higher or lower level of life satisfaction.

$$life\_sat_{i,t} = \alpha + \gamma \cdot income_{i,t} + \mathbf{X'_{i,t}}\beta + \lambda_t \cdot \iota + \epsilon_{i,t} \tag{1}$$

The model is defined as follows: i represents the identification number of each observation, t is the time period of the observation and λ_t is a categorical variable that captures this time feature.

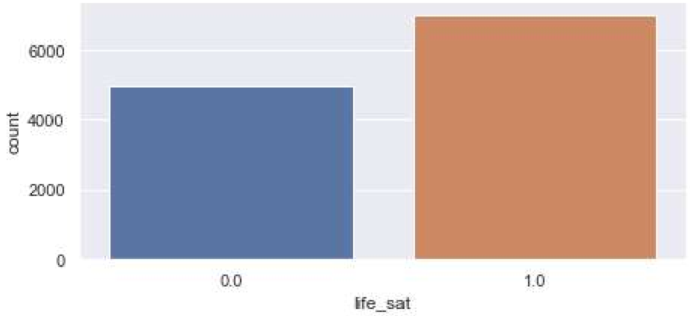

Figure 1: Distribution of life satisfaction in the overall sample

2 Exploratory Analysis

Recalling the main model (1), we begin by operating some preliminary analysis on the set in order to better understand its nature and the underlying relation of the dependent variables. In particular, we perform a Principal Component Analysis and a Discriminant Analysis.

[1] Information come from many European countries and Israel

[2] The survey contains information about several thousands of individuals aged 50 and over since ageing is considered one of the challenges of the century

2.1 Principal Component Analysis

The Principal Component Analysis consists in a dimensionality reduction technique able to embed in few components r a large part of the informational power of the original variables. We create d components as linear combination of the previous d variables: the first one explains the highest overall variability that must remains unchanged. The components are orthogonal and by looking at theirs coefficients (i.e. the loadings) we understand the variable that contributes the most to their definition and we can properly interpret: larger value of the loading for a given variable implies that its variability is highly explained by that component.
In order to find the loadings that satisfy the constraint of orthogonality and maximum explanation power, we analyse the correlation matrix: from an algebraic point of view, the loadings of the first component are the eigenvectors of the largest eigenvalue in the matrix considered.
We point out that we have conducted the PCA based only on the continuous variables in our dataset, hence we consider: $fahc$, $hprf$, $thexp$, $hnetw$, age, $yedu$, bmi and $income$[3].

In the literature, there are many criteria to choose the most appropriate number of components to further analyse and interpret: one of the most common hinges on the scree plot.

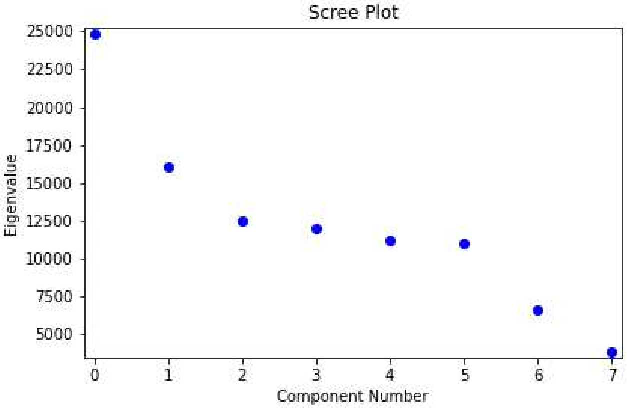

Figure 2: Scree plot of the eigenvalues

Looking at the figure (2), we observe on the y-axis the values of the eigenvalues and on the x-axis the associated component. Since we can interpret the eigenvalues as proxy of the total variability explained by the new variable created, the *elbow rule* suggests to choose three components because the addition of the fourth one would not add relevant explanatory power.

	comp 0	**comp** 1	**comp** 2
fahc	-0.534008	0.283135	0.211429
hprf	-0.058106	0.514289	-0.429427
thexp	-0.509411	0.434916	0.183523
hnetw	-0.280006	-0.086682	-0.681530
age	0.361821	0.349411	-0.304622
yedu	-0.419973	-0.471630	-0.002996
bmi	0.101380	0.326241	0.312363
income	-0.236614	-0.087860	-0.286971

Table 1: Loadings of the three principal components

To understand their interpretation, we examine the loadings: the first component exhibits remarkable negative values for $fahc$, $hnetw$, $yedu$ and positive for age. As a consequence of this, the component could catch the peculiarity of elderly people with low level of education and household net worth. The second component is characterized by positive weights for $hprf$, $thexp$ and negative for $yedu$ which means it represents a measure of high family lifestyle associated to low instruction level for the respondent.

[3] *fahc*: amount spent on food; *hprf*: annual home produced consumption; *thexp*: total household expenditure; *hnetw*: household net worth; *income*: total household revenue in ten thousand EUR; *bmi*: Body Mass Index

The last one has negative loadings for $hprf$ and $hnetw$ and can be interpret as an inverse measure of household expenditure capacity.
The magnitude of each variable's loadings also suggests us which one has the largest explanatory capacity: despite the first component remains the most representative by construction, the range of the loadings is quite confined in the interval $[-0.7, 0.5]$ for every components, meaning that it is possible to try to interpret them but these characterisations have little quantitative foundation.

In the end, we look at the scores to understand whether by using these components we are able to discriminate the observations' life satisfaction. The shape of points in both graphs is quite dense and well distributed around the 0 apart from some outliers that register low score for the first component; hence there are more skewness and variability in the left part of that score distribution. The units with high level of satisfaction (yellow points) are homogeneously distributed in the first and second plot. On the other hand, the purple points seem to take scores mostly in the upper bound of the points cloud of second plot, hence, on average, higher scores of second component .
Although, the graph representation is not able to highlight a clear pattern in the feature's description of units' satisfaction level: hence, data reduction to only three variables is not useful to investigate the research's phenomena.

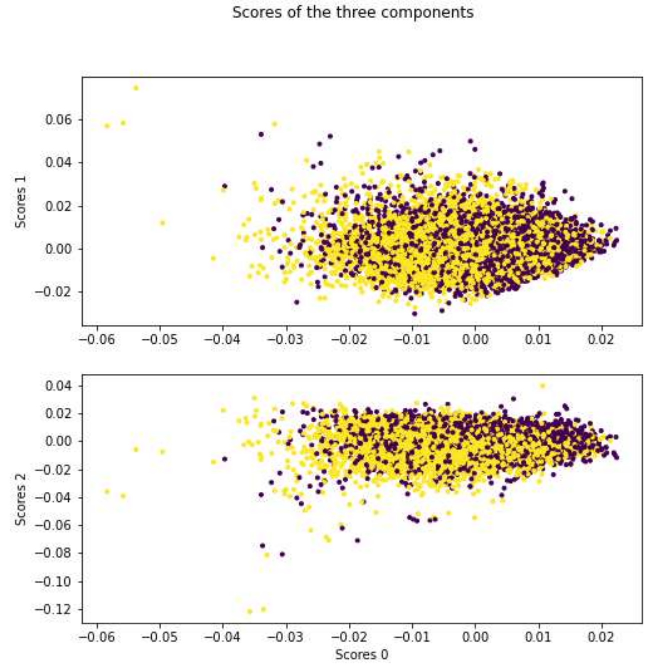

Figure 3: Scores of the three components

2.2 Linear Discriminant Analysis

Given that we did not obtain relevant results with the previous analysis, we decided to implement the Linear Discriminant Analysis. This method is applied by starting from data with known classification to find a specific partitioning rule that can be applied to group new observations in the two original classes[4].
This approach still aims to reduce the dimensionality by projecting the input to the most discriminative direction but, unlike the components analysis, we find a linear decision boundary and then use Bayes' rule to classify the observation[5].

[4] There are several approaches that can be used, some of them also make assumptions of normality and homoskedasticity of the data in order to classify the unit according to the likelihood functions

[5] Bayes' rule assigns the unit to a certain class according to the highest posterior probability

Since we are dealing with huge variability within the variables, we standardise our data and then apply the analysis to find one discriminant.

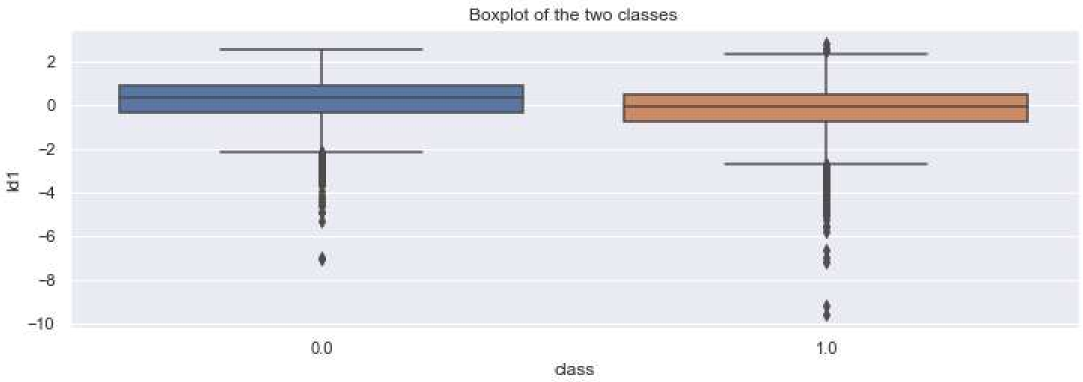

Figure 4: Box plot of the two classes discriminated

The figure (4) shows the distribution of our data: indeed, the analysis led to a slight separation of the two groups since we can see that the class zero (i.e. $life\_sat = 0$) matches with higher values and the median differs. As we expected, we can notice that satisfied people (i.e. $life\_sat = 1$) present more outliers than the others. However, we can't interpret the discriminating variable, since it is only an artificial tool to separate as much as possible the two groups.

We can evaluate our method through the *Cross-Validation Criterion* that splits the dataset into the train and the test set: the former will be used to define the rule that discriminates and the latter will evaluate it[6]. Again, we apply the Linear Discriminant Analysis and we use the Logistic Regression to classify the predictions: as a result of this, we can compute the confusion matrix that compares the true values with the predicted ones and gives us an idea of the goodness of the rule. From the table (2) we can see that the accuracy rate is 0.63[7] and the misclassification rate is 0.37.

	Predicted 0	**Predicted 1**
True 0	518	991
True 1	382	1783

Table 2: Confusion matrix for the Discriminant Analysis

To conclude, neither the Principal Component nor the Linear Discriminant Analysis gave satisfactory results to effectively investigate life satisfaction dummy. According to us, this is due to the variables chosen: they all consider an economic condition that may be redundant. To overcome this issue, we decided to keep *income* as representative of the financial status and include the categorical variables accounting for social condition that may have a relevant impact for our research question.

3 Descriptive analysis

We start the analysis through some descriptive statistics that help us to better understand the nature of some variables.
Given their numerosity, we create another dataset which collects only the ones that we consider more relevant for our purpose and therefore will be used to implement our regressions in the following sections.

[6] In this way, we avoid an optimistic result given by the judgement of the rule through the dataset used to built it

[7] Given by the ratio of the well predicted over the total units, implies that the model predicts correctly the 63% of the overall sample

The dataset is composed by:

- $life\_sat$: outcome variable that describes the self-perceived life satisfaction[8]
- $income$: total household revenue in ten thousand of EUR (main explanatory regressor)
- $yedu$: years of education (part of the control variables)
- $mstat$: marital status of the individual[9] (part of the control variables)
- $hstatus$: self perceived health status[10] (part of the control variables)
- $gali$: physical limitation in activities[11] (part of the control variables)
- $male$: gender of the individual[12](part of the control variables)
- $otrf$: proxy of housing living condition[13] (part of the control variables)
- bmi: Body Mass Index[14](part of the control variables)
- cjs: proxy of the current job situation[15] (part of the control variables)

To begin, we proceed with some measures to clean our dataset: indeed, we evaluate the number of missing values and manage those observations. Since we find out a negligible amount of missing (48 units), we assess that dropping them will not affect our results and therefore we obtain a final dataset of 11993 observations.
Afterward, we operate a preliminary descriptive analysis about the chosen variables in the overall sample grouped by year, to detect significant changes over time.

Starting by the dependant variable, we recall that its distribution is quite balanced (1), with a slight majority toward people that self perceived as satisfied. As we expected, the sample mean value is constant across years and slightly greater than 0.5 (3) in all waves, meaning that individuals tend to perceive themselves as satisfied.

Year	Mean	Std. Dev.	Min	50%	Max
2011	0.60	0.49	1.0	1.0	1.0
2013	0.56	0.50	1.0	1.0	1.0
2015	0.59	0.49	0.0	1.0	1.0

Table 3: Descriptive table of the dependent variable life satisfaction grouped by years

From the analysis of the variable *income* we can see that the average value fluctuates over the years: indeed, it increases until 2013 when it reaches the peak and it declines steeply afterwards. As we can see from the table (4), the highest value reached is 0.39 which is almost halved to 0.22 the following year.
If we consider the European economic framework in the years analyzed here, we can state that the households faced a critic period that led to a rapid decline in families' income: the economic history may provide us some information about its causes, however, the current analysis' tools prevent us to define with precision the reasons underlying this dramatic shrinks.

[8] dummy variable: takes values 1 if the interviewed is really satisfied with her/his life and 0 otherwise. It is built from a life satisfaction index on a 10-point scale which takes 0 if completely dissatisfied person and 10 if completely satisfied: as a consequence, $life\_sat$ takes value 1 if the index takes value 7, 8, 9 or 10.

[9] categorical variable: 1 if married and living with spouse, 2 if registered partnership, 3 if married and not living with spouse, 4 if never married, 5 if divorce and 6 if widowed

[10] categorical variable: 1 if poor health, 2 if fair, 3 if good, 4 if very good and 5 if excellent

[11] dummy variable: 1 if limitation with activities and 0 if not

[12] dummy variable: 1 if male and 0 if female

[13] categorical variable: 1 if owner of the house where lives, 2 if member of a cooperative, 3 if tenant, 4 if subtenant, 5 if rent free

[14] According to the World Health Organization a bmi higher or equal to 30 determines an obese condition among adults (age-standardize estimate)

[15] categorical variable: 1 if retired, 2 if employed/self-employed, 3 if unemployed, 4 if permanently sick, 5 if homemaker, 6 if other

Year	Mean	Std. Dev.	Min	50%	Max
2011	0.35	0.47	0.05	0.20	3.36
2013	0.39	0.59	0.04	0.20	4.20
2015	0.22	0.11	0.06	0.19	0.60

Table 4: Descriptive table for the total household income grouped by years (in ten thousands)

Regarding the years of education, it is interesting to notice (6) that the distribution of *yedu* has significant peaks in correspondence to 5, 8 and 13 years of education, which could represent the end of school cycles (considering that the countries have very different educational systems). Overall, the sample mean is regular across the years, with a slight increase from 8.25 to 9.05 and a higher distribution on the intermediate values (11).

Then, if we look at the distribution of *hstatus* (7), it reveals a very high concentration of people who perceive a good health status for themselves (value 3.0)[16] but with an evident skewness toward the left tail, which implies that there is higher distribution of individuals with a bad perception of their health status.

Looking at the variable of current job situation *cjs*[17], we observe that a large majority of units are retired (since the sample is composed by people over 50 this is not surprising) and most of the observations are either employed or homemaker (5).

cjs	Relative frequence (%)
Retired (1)	52.40
Employed (2)	22.26
Unemployed (3)	2.82
Permanently sick (4)	2.31
Homemaker (5)	20.21

Table 5: Relative frequence of the current job situation (*cjs*)

The variable accounting for the housing condition, i.e. $otrf$, reveals that most of the sample units (83.71%) own their house and an irrelevant percentage belongs either to *Member of a cooperative* or *Subtenant* (6).

otrf	Relative frequence (%)
Owner of the house (1)	83.80
Member of cooperative (2)	0.13
Tenant (3)	9.80
Subtenant (4)	0.03
Rent free (5)	6.25

Table 6: Relative frequence of the housing condition ($otrf$)

Finally, the average of the Body Mass Index evaluated in the sample (13) is stable across the years at a value that is lightly higher than 26 which is the limit to be considered as a physically healthy individual according to the World Health Organization (10).

4 Regressions

4.1 OLS Regression

We now estimate the model (1) using the OLS estimator and a vector of control variables $\mathbf{X}$ composed by: *age*, *male*, *yedu*, *mstat*, *hstatus*, *gali*, $otrf$, *bmi* and cjs. We expect that the aforementioned regressors will substantially influence the self perception of life satisfaction: for instance, the ownership of the living house ($otrf$) as the current job situation (cjs) represent huge concerns from the financial and stability point of view. Despite the unquestionable effect of some of these relevant variables, we will mainly focus on how the *income* affects the dependant variable.

[16] As we expected, the sample mean is around 2.75 and it is stable across the years (12)

[17] The distribution of the variable can be seen in Appendix (9)

Since $life\_sat$ is a dummy, the model estimated can be interpreted as a Linear Probability Model (LPM) which assumes a Bernoulli distribution of the dependent variable and enable us to estimate the probability level that $life\_sat = 1$ as follows:

$$P(life\_sat_{i,t} = 1|\mathbf{X}) = E(life\_sat_{i,t} = 1|\mathbf{X}) = \alpha + \gamma \cdot income_{i,t} + \mathbf{X}'_{\mathbf{i,t}}\beta + \lambda_t \cdot \iota \tag{2}$$

However, since some categorical variables have not a clear (numerical) interpretation, we evaluated the regression by considering them independently from the overall control vector of $\mathbf{X}$.

The table results (7) shows that an increase of ten thousands in the household income can rise the probability of life satisfaction by 2.74%. Moreover, the test for the coefficient provides a p-value of 0.006 which implies that the regressor is statistically significant[18].

Overall, the marital status categories have a negative effect compared to the base one (i.e. married and living with spouse) and are all significant with the exception of the second one, for which we cannot reject the null hypothesis (i.e. the regressor is statistically equal to zero).

Similarly, the variable cjs has a negative effect compared to the base category (i.e. retired), hence, it can be considered as the optimal status. In addition it is almost always significant apart from the case of employed unit ($cjs = 2$).

	Coeff	t	p-value
Intercept	0.225	3.773	0.000
C(mstat)[T.2]	-0.054	-1.498	0.134
C(mstat)[T.3]	-0.136	-2.655	0.008
C(mstat)[T.4]	-0.166	-8.672	0.000
C(mstat)[T.5]	-0.165	-6.160	0.000
C(mstat)[T.6]	-0.137	-9.472	0.000
C(otrf)[T.2.0]	-0.173	-1.443	0.149
C(otrf)[T.3.0]	-0.095	-6.483	0.000
C(otrf)[T.4.0]	-0.513	-1.910	0.056
C(otrf)[T.5.0]	-0.038	-2.126	0.034
C(cjs)[T.2.0]	-0.004	-0.297	0.767
C(cjs)[T.3.0]	-0.194	-7.065	0.000
C(cjs)[T.4.0]	-0.103	-3.485	0.000
C(cjs)[T.5.0]	-0.059	-4.526	0.000
C(year)[T.2013]	-0.037	-3.344	0.001
C(year)[T.2015]	-0.009	-0.790	0.429
income	0.027	2.740	0.006

Table 7: Coefficients, t-test and p-values of the OLS regression

Referring to $otrf$, we can observe that the estimated coefficients have a negative impact on the dependent variable with respect to the baseline category (ownership of the living house), which can be considered as a coherent result with the economic theory. Although, only the value of tenant ($otrf = 3$) which is -0.095 is statistically significant.

The temporal changes highlight a decreasing and negative trend on the probability of life satisfaction meaning that, ceteris paribus, in 2013 the average satisfaction level is significantly smaller by 3.69% whereas in 2015 by 0.93% (not statistically significant). Furthermore, as we can see from the overall table of the coefficients (14), the control variable of $\mathbf{X}$ is not significant for $male$ and bmi[19].

However, the model shows an estimated intercept equals to 0.225 that represents the marginal impact on life satisfaction in case of female individual and null income level (given the categorical variables at baseline category) which is a counter-intuitive result and not consistent with the underlying economic theory.

To conclude, since the fitted values are not bounded to lie in the [0,1] interval, we may end up predicting probabilities of life satisfaction which are greater than 1 or smaller than 0, giving us meaningless results.

[18] We consider a level of significance of 5%

[19] Which are respectively $xols\_reg2.T[1]$ and $xols\_reg2.T[5]$

By analyzing the probabilities predicted by this model we see that this is actually happening for 9 values out of 11993 units observed.
Hereafter, we perform an analysis regarding the accuracy of the OLS model in predicting the values of $life\_sat$ by comparing the fitted values and the observed data for each unit. We have decided to drop the 9 wrongly predicted observations in order to conduct the aforementioned analysis considering the set of units properly predicted. To classify the fitted value in two classes, we choose a threshold value equal to 0.5 and we find the number of match and mismatch.

	Classification	
Life sat	0	1
0	2245	2738
1	1304	5697

Table 8: OLS classification table

The bi-variate table (8) shows on the main diagonal the number of matches and out of the diagonal the mismatches: we can state that the 66.27% of observations are correctly predicted and we can use this figure as proxy of the accuracy of the OLS estimation model.

In order to radically copy with the issue of values outside our admissible range, we can use an index model, in particular a Probit one, depending on the distributional assumption of the dependent variable. Moreover, in this framework we are interested in estimating the marginal effect on the probability of life satisfaction. However, to benefit of a model whose predicted probabilities are completely contained in the [0,1] range there are some costs: for instance, we lose the direct interpretation of the coefficients' magnitude and we need to impose stricter assumptions on the data generating process.

4.2 Probit Regression

In this section we implement a different model framework in order to overcome the theoretical and empirical issues exhibited in the previous subsection. In this sense, we built up a model specification as follows:

$$Pr[life\_sat_{i,t} = 1|I_{i,t}] = \phi(\alpha+\gamma income_{i,t}+\beta_1 mstat_{i,t}+\beta_2 otrf_{i,t}+\beta_3 cjs_{i,t}+\beta\mathbf{X_{i,t}}+\lambda_t\iota+\gamma[income_{i,t}\cdot year_i] \quad (3)$$

where I represents the overall information set and $\phi(.)$ is the cumulative standard normal distribution function of the probit model: we use the mapping function $\phi(x)$ that provides a result in the admissible set $[0, 1]$.
It is important to point out that in a probit regression the magnitude of the coefficient estimated cannot be interpreted as the marginal effect of the variable. Indeed, the partial effect is not constant since it always depends on the characteristics of the observation itself and it can be found for a continuous variable by computing the derivative of the dependent variable:

$$\frac{\partial p(x)}{\partial x_j} = \frac{\partial\Phi(x\beta)}{\partial x\beta}\frac{\partial(x\beta)}{\partial x_j} = \phi(x\beta)\beta_j$$

Thus, the coefficients cannot be thought of as the partial effects but are still informative: since $\phi(x\beta) > 0$ by definition, the sign of β_j has information on the direction of the partial effect.

The table (9) reports the average partial effect of each variable and its statistical significance. We can observe that, on average, large part of the effect has a negative sign, hence provokes a reduction in the probability to register $life\_sat = 1$. In particular, the categorical variable of marital status shows a significant negative impact slightly above 10% for all the categories with respect to the baseline one of being married and living with the spouse which seems to be the most preferable situation. On the other hand, the variable $otrf$ is statistically significant with respect to the baseline situation of house ownership only for the case of tenant and free rent. Both variables have a negative impact on the probability of life satisfaction.
Taking into account the current job situation, we notice that the only cases that have a significant negative impact are the one that referred to unemployment (19%), sickness (9.6%) and homemaker (5.7%). Indeed, being retired or being employed has no significant difference which implies that the inability to work has the highest effect on the life satisfaction dummy.

Finally, our main regressor (i.e. *income*) presents a positive marginal effect equal to 2.7% given a increase of income by ten thousands euro.

	dy/dx	std. err	z	P> |z|	[0.025]	[0.975]
C(mstat)[T.2]	-0.0555	0.035	-1.569	0.117	-0.125	0.014
C(mstat)[T.3]	-0.1331	0.051	-2.605	0.009	-0.233	-0.033
C(mstat)[T.4]	-0.1617	0.019	-8.583	0.000	-0.199	-0.125
C(mstat)[T.5]	-0.1604	0.026	-6.108	0.000	-0.212	-0.109
C(mstat)[T.6]	-0.1303	0.014	-9.217	0.000	-0.158	-0.103
C(otrf)[T.2.0]	-0.1766	0.121	-1.461	0.144	-0.414	0.060
C(otrf)[T.3.0]	-0.0928	0.014	-6.434	0.000	-0.121	-0.065
C(otrf)[T.4.0]	-2.3964	2542.284	-0.001	0.999	-4985.182	4980.389
C(otrf)[T.5.0]	-0.0376	0.018	-2.143	0.032	-0.072	-0.003
C(cjs)[T.2.0]	-0.0022	0.014	-0.161	0.872	-0.030	0.025
C(cjs)[T.3.0]	-0.1888	0.027	-6.926	0.000	-0.242	-0.135
C(cjs)[T.4.0]	-0.0985	0.030	-3.284	0.001	-0.157	-0.040
C(cjs)[T.5.0]	-0.0570	0.013	-4.431	0.000	-0.082	-0.032
C(year)[T.2013]	-0.0365	0.011	-3.347	0.001	-0.058	-0.015
C(year)[T.2015]	-0.0094	0.011	-0.862	0.389	-0.031	0.012
income	0.0271	0.010	2.677	0.007	0.007	0.047
xreg.T[0]	0.0028	0.001	4.549	0.000	0.002	0.004
xreg.T[1]	-0.0070	0.010	-0.705	0.481	-0.026	0.012
xreg.T[2]	0.0050	0.001	4.509	0.000	0.003	0.007
xreg.T[3]	0.1013	0.005	20.790	0.000	0.092	0.111
xreg.T[4]	-0.0624	0.010	-6.012	0.000	-0.083	-0.042
xreg.T[5]	-0.0017	0.001	-1.641	0.101	-0.004	0.000

Table 9: Margins of the Probit regression

In addition, we evaluated the model including the interaction between *income* and *year* and we run some tests to better understand if the value of the main independent variable is statistically different over the years: indeed, since we are considering the period following the global crisis, it is interesting to understand if this had a substantial impact on self perceived life satisfaction of respondents. However, our results show that the average partial effect of income is not significant and it is statistically equal only between 2011 and 2013[20], meaning that the value of additional income is different between 2011-2015 and 2013-2015.

	Classification	
Life sat	0	1
0	2278	2713
1	1309	5693

Table 10: Probit classification table

Recalling the previous accuracy analysis performed for OLS model, we evaluate the Probit ability to correctly predict the dependent variable $life\_sat$. From the classification matrix (10) we can state that the model is able to accurately approximate 66.46% of observations: the confusion matrix forecasts that 70.09% of people are satisfied ($life\_sat = 1$) which is still very distant to the value obtained from the real dataset[21].

To conclude, despite the accuracy of the probit model is not much higher than the one obtained from the ols model, we think this is an acceptable result since the former solved the theoretical problem of predicting probability values outside the admissible range.

[20] The p-value obtained is 0.596

[21] Not accounting for time variability, the dataset presents the 58.38% of individuals classified as satisfied

A Appendix - Descriptive statistics

A.1 Figures

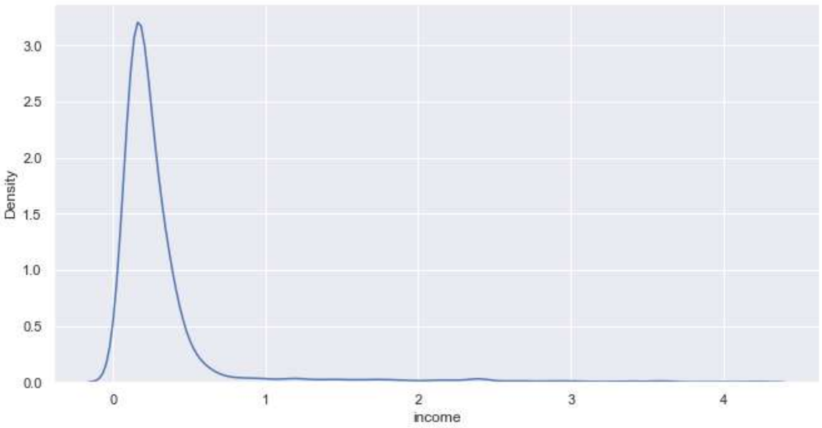

Figure 5: Distribution of income for the overall sample

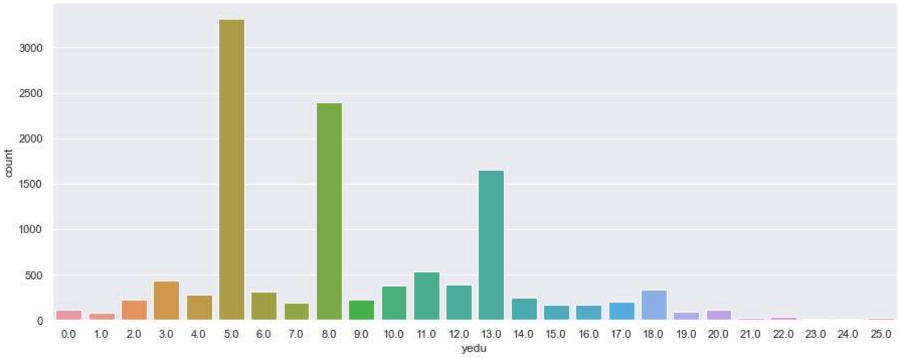

Figure 6: Distribution of years of education

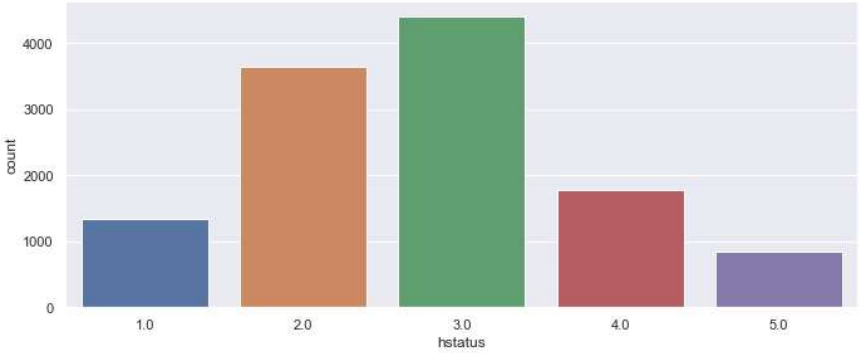

Figure 7: Distribution of health status perception in the overall sample

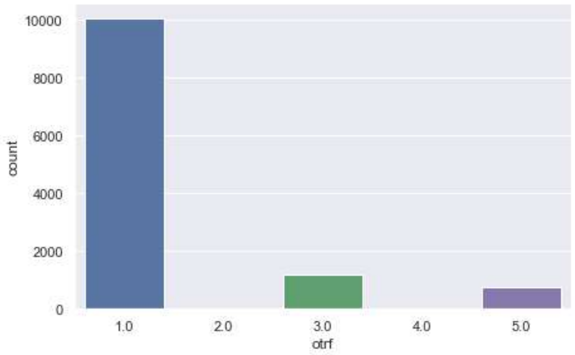

Figure 8: Distribution of the housing condition for the overall sample

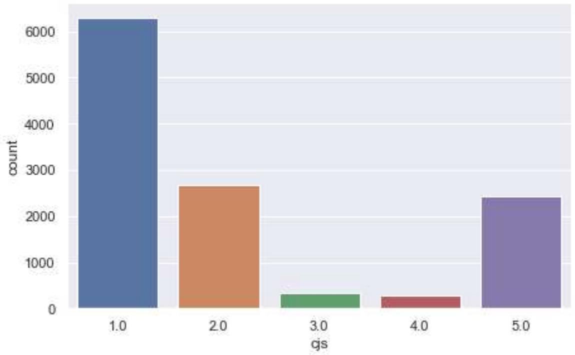

Figure 9: Distribution of the current job situation in the overall sample

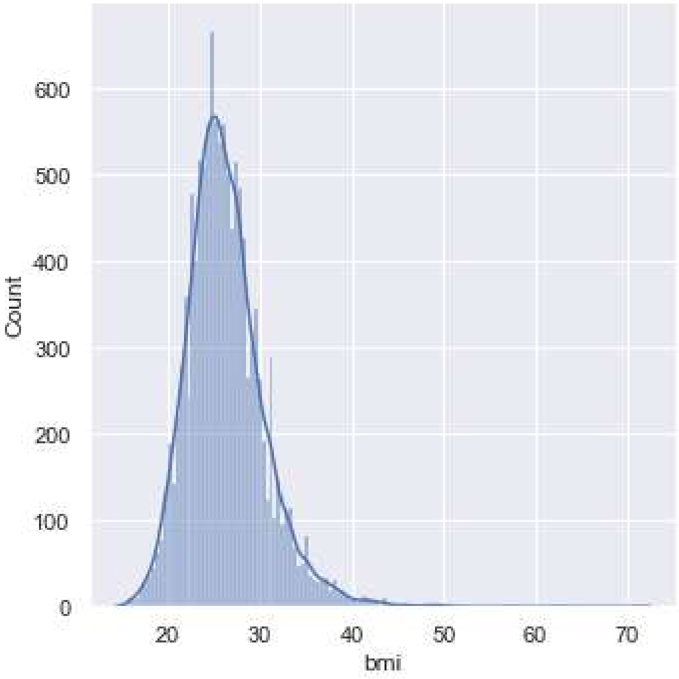

Figure 10: Distribution of the Body Mass Index in the overall sample

A.2 Tables

Year	Mean	Std. Dev.	Min	50%	Max
2011	8.25	4.29	0.0	8.0	25.0
2013	8.78	4.55	0.0	8.0	25.0
2015	9.05	4.58	0.0	8.0	25.0

Table 11: Descriptive table of years of education (*yedu*) grouped by years

Year	Mean	Std. Dev.	Min	50%	Max
2011	2.77	1.07	1.0	3.0	5.0
2013	2.72	1.08	1.0	3.0	5.0
2015	2.79	1.03	1.0	3.0	5.0

Table 12: Descriptive table of the perceived health status (*hstat*) grouped by years

Year	Mean	Std. Dev.	Min	50%	Max
2011	26.48	4.33	15.62	26.12	72.50
2013	26.20	4.23	15.52	25.64	66.60
2015	26.23	4.15	14.34	25.71	50.71

Table 13: Descriptive table of the Body Mass Index (*bmi*) grouped by years

B Appendix: Regression

Dep. Variable:	life_sat	**R-squared:**	0.114
Model:	OLS	**Adj. R-squared:**	0.113
Method:	Least Squares	**F-statistic:**	70.32
Date:	Tue, 18 Oct 2022	**Prob (F-statistic):**	7.50e-295
Time:	20:00:35	**Log-Likelihood:**	-7804.5
No. Observations:	11993	**AIC:**	1.566e+04
Df Residuals:	11970	**BIC:**	1.583e+04
Df Model:	22		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	0.2246	0.060	3.773	0.000	0.108	0.341
C(mstat)[T.2]	-0.0536	0.036	-1.498	0.134	-0.124	0.017
C(mstat)[T.3]	-0.1357	0.051	-2.655	0.008	-0.236	-0.036
C(mstat)[T.4]	-0.1665	0.019	-8.672	0.000	-0.204	-0.129
C(mstat)[T.5]	-0.1655	0.027	-6.160	0.000	-0.218	-0.113
C(mstat)[T.6]	-0.1373	0.014	-9.472	0.000	-0.166	-0.109
C(otrf)[T.2.0]	-0.1733	0.120	-1.443	0.149	-0.409	0.062
C(otrf)[T.3.0]	-0.0952	0.015	-6.483	0.000	-0.124	-0.066
C(otrf)[T.4.0]	-0.5126	0.268	-1.910	0.056	-1.039	0.014
C(otrf)[T.5.0]	-0.0379	0.018	-2.126	0.034	-0.073	-0.003
C(cjs)[T.2.0]	-0.0041	0.014	-0.297	0.767	-0.031	0.023
C(cjs)[T.3.0]	-0.1943	0.027	-7.065	0.000	-0.248	-0.140
C(cjs)[T.4.0]	-0.1026	0.029	-3.485	0.000	-0.160	-0.045
C(cjs)[T.5.0]	-0.0590	0.013	-4.526	0.000	-0.084	-0.033
C(year)[T.2013]	-0.0366	0.011	-3.344	0.001	-0.058	-0.015
C(year)[T.2015]	-0.0086	0.011	-0.790	0.429	-0.030	0.013
income	0.0274	0.010	2.740	0.006	0.008	0.047
xols_reg2.T[0]	0.0028	0.001	4.457	0.000	0.002	0.004
xols_reg2.T[1]	-0.0063	0.010	-0.637	0.524	-0.026	0.013
xols_reg2.T[2]	0.0050	0.001	4.541	0.000	0.003	0.007
xols_reg2.T[3]	0.1005	0.005	20.146	0.000	0.091	0.110
xols_reg2.T[4]	-0.0692	0.011	-6.498	0.000	-0.090	-0.048
xols_reg2.T[5]	-0.0017	0.001	-1.621	0.105	-0.004	0.000

Omnibus:	85456.725	**Durbin-Watson:**	2.010
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1234.242
Skew:	-0.300	**Prob(JB):**	9.72e-269
Kurtosis:	1.548	**Cond. No.**	4.62e+03

Table 14: OLS Regression Results

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.62e+03. This might indicate that there are strong multicollinearity or other numerical problems.

	Coeff	t	p-value
Intercept	-0.814	-4.836	0.000
C(mstat)[T.2]	-0.157	-1.569	0.117
C(mstat)[T.3]	-0.377	-2.603	0.009
C(mstat)[T.4]	-0.458	-8.507	0.000
C(mstat)[T.5]	-0.454	-6.081	0.000
C(mstat)[T.6]	-0.369	-9.123	0.000
C(otrf)[T.2.0]	-0.500	-1.461	0.144
C(otrf)[T.3.0]	-0.263	-6.401	0.000
C(otrf)[T.4.0]	-6.785	-0.001	0.999
C(otrf)[T.5.0]	-0.106	-2.141	0.032
C(cjs)[T.2.0]	-0.006	-0.161	0.872
C(cjs)[T.3.0]	-0.534	-6.886	0.000
C(cjs)[T.4.0]	-0.279	-3.280	0.001
C(cjs)[T.5.0]	-0.161	-4.420	0.000
C(year)[T.2013]	-0.103	-3.342	0.001
C(year)[T.2015]	-0.027	-0.862	0.389
income	0.077	2.675	0.007

Table 15: Coefficients, t-test and p-values of the Probit regression

	Coeff	t	p-value
Intercept	-0.818	-4.837	0.000
C(mstat)[T.2]	-0.148	-1.480	0.139
C(mstat)[T.3]	-0.350	-2.409	0.016
C(mstat)[T.4]	-0.434	-8.040	0.000
C(mstat)[T.5]	-0.431	-5.751	0.000
C(mstat)[T.6]	-0.352	-8.687	0.000
C(otrf)[T.2.0]	-0.447	-1.306	0.192
C(otrf)[T.3.0]	-0.257	-6.247	0.000
C(otrf)[T.4.0]	-6.472	-0.003	0.998
C(otrf)[T.5.0]	-0.103	-2.071	0.038
C(cjs)[T.2.0]	-0.016	-0.399	0.690
C(cjs)[T.3.0]	-0.509	-6.546	0.000
C(cjs)[T.4.0]	-0.265	-3.119	0.002
C(cjs)[T.5.0]	-0.145	-3.955	0.000
C(year)[T.2013]	-0.102	-2.719	0.007
C(year)[T.2015]	-0.233	-4.330	0.000
income	0.059	1.192	0.233
income:C(year)[T.2013]	0.004	0.058	0.954
income:C(year)[T.2015]	0.949	4.797	0.000

Table 16: Coefficients, t-test and p-values of the Probit regression with the interaction term