

Data Task - RA position for Professor Dan Benjamin

Giacomo Ricciardi

August 25, 2021

This short paper analyzes the relationship between well-being and factors that could be influencing it. The software used was Stata (the attached Do-file is also extensively commented, question by question) and this document was produced using Latex.

1 Question 1

- a. Imported the ratings.csv file in Stata
- b. The number of unique (distinct) respondents is 1,056. The number of unique (distinct) aspects of life is 17. By using the *codebook* command in Stata we can have a sense of the most important metrics of the specified variable (e.g. the number of unique values). Compact just produces a compact report. If needed, a short code was implemented to save these two values as well.
- c. Each respondent did not rate each aspect only once (by sorting by worker, aspect and time, we have a clear overview of this). Using the *by* command we can create a variable summarizing whether the answer to aspects for each respondent is unique or not. The variable duplicate could be 0 if the answer is unique, or greater than 0 if there are more than one answer. After checking no respondent rated more than twice the same aspect, the most recent values were kept. 237 Observations were dropped.
- d. Generate a variable describing the average of the rating per respondent, now that the dataset is clean. The summarized results are the following:

Table 1: Summary of the subjective ratings statistics, Question 1.d).

	Min	25th	50th	75th	Max
Subjective ratings	5.7647	49.0588	61.4412	75.0882	100.0000

2 Question 2

- Imported the demographics.csv file in Stata
- The number of rows in this dataset is 1,056 (*count* command), and it coincides with the number of respondents computed in question 1.a).
- To have the subjective ratings I merged the 1-to-many datasets by maintaining the subjective ratings values only (then sorted by worker and dropped duplicated values).
- By regressing the subjective ratings variable on the respondent's household income we notice how a unitary increase in income does not seem to influence the average rating in a major way. Indeed, the coefficient is statistically significantly different from 0 (p-value=0), and the p-value of the F-Test is also null ($F(1,1,054) = 41.57$, p-value=0), meaning that the income reliably predicts subjective ratings. However, the estimated value of the beta is actually close to 0 (table below), even though being slightly positive (as possibly expected ex-ante, i.e., the higher the income, the better the perception of personal well-being). In economic terms, the amount of money families earn does not seem to play a crucial role in how respondents rate aspects of their well-being ("money does not buy happiness"?).

Finally, it must also be noticed how the R-squared of the regression is very weak (3.79%), symptom that this model does not fully explain the variance of the subjective ratings.

Table 2: Univariate regression, Question 2.d).

Variables	subjective_ratings
income	0.000*** (0.000)
Constant	56.279*** (1.011)
Observations	1,056
R-squared	0.038
<i>Standard errors in parentheses</i>	
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$	

e. The multivariate regressions was then run by including the five control variables¹, to understand if the model is correctly specified. These introductions do not alter the statistical significance of the income coefficient, which remains positive but basically null. The p-value of the F-Test is zero ($F(6, 1,048) = 9.03$) but the R-squared (and the adjusted R-squared) is still small (c. 5%). In addition, three out of five of the control variables do not seem to be statistically significant from zero, as their p-values are all greater than commonly accepted values (age, age2 and race, hence these can be excluded to maintain a more parsimonious model).

Instead, education level and gender's p-value are less than 5%, with negative and positive coefficients respectively. Males seem to provide greater ratings than female, while the interpretation for education seems to be that the less educated people tend to answer these questions providing ratings closer to 100.

From the economic point of view, it could be said that age (both linear and quadratic approximations) and race are not likely to explain the average rating given by each

¹First age-squared was created as a variable. Moreover, education and race were string variables, hence they were transformed into numeric variables using the encode-generate command. Notice how interpretation for these results must be careful, as the encode command alphabetically orders the values for the strings.

worker, whereas further analysis is needed for educational level and gender. Furthermore, other variables that are not present in this dataset could be impacting these metric, as this represents just a minor study of the survey.

Table 3: Multivariate regression, Question 2.e).

Variables	subjective_ratings
income	0.000*** (0.000)
age	-0.374 (0.291)
age2	0.004 (0.003)
male	2.319** (1.170)
education_level	-0.535** (0.213)
race_type	0.295 (0.348)
Constant	63.591*** (6.104)
Observations	1,055
R-squared	0.049
<i>Standard errors in parentheses</i>	
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$	

- f. Should the household size be given, I would compute the individual income by dividing the total household income by its size. This could represent a better proxy of the well-being of the individual answering the survey. By inserting the total income rather than the personal one, we can only study the relationship between what the family earns as a whole and the average rating given, and not between each individual's income

and how happy he / she is. This does not mean that the total income would be completely excluded from the analysis, but it just means that further "more personal and subjective" conclusions could be drawn.

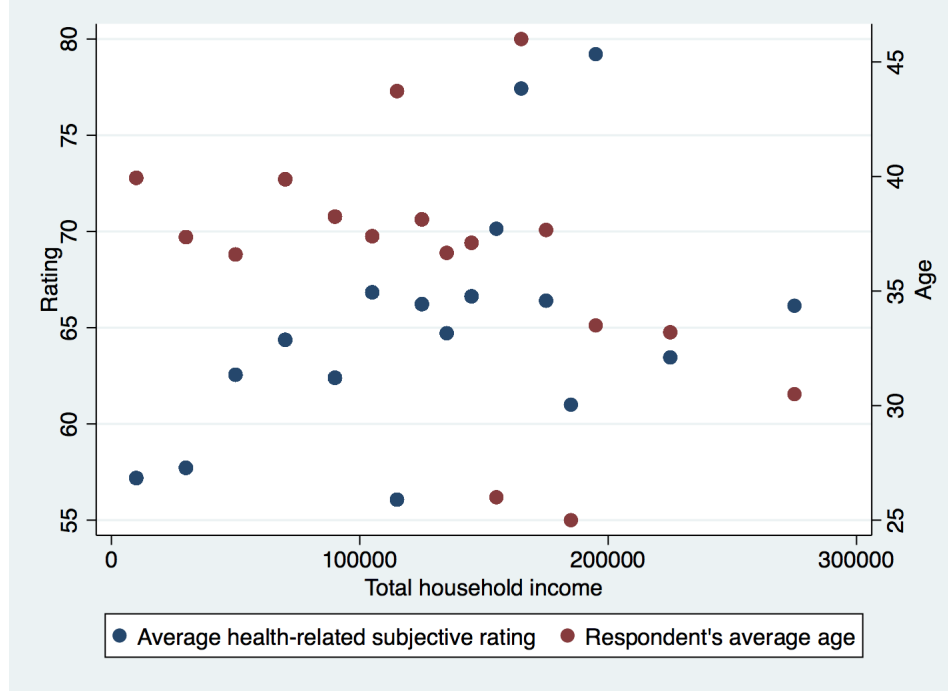
In addition, the size itself could serve as a control variable in the multivariate regression, to check whether having kids and / or a partner does influence the average rating associated to well-being aspects of people's life.²

3 Question 3

- a. This represents the step-by-step process I would implement (and I did implement):
 - (a) Identify the aspects that capture health among the 17 of the surveys. These were (1) "the quality of your sleep", (2) "you not feeling anxious", (3) "your emotional stability", (4) "your health", (5) "your mental health", (6) "your physical fitness", (7) "your physical safety and security". Of course, depending on personal choice other aspects could be included / excluded in this health-only filter.
 - (b) Compute the subjective ratings related to health-only aspects.
 - (c) Re-merge the two datasets, so to have 1,056 observations for health-only subjective ratings, one per worker.
 - (d) Compute the average of the health-only subjective ratings per each value of income. Compute the average of the ages per each value of income.
 - (e) Plot both average health-related ratings and average age on a 2-axes scatter plot having the income as x-axis.
- b. Indeed, the following graph was produced (related to average values):

²Note how specific details on the size must be gathered to do this. For instance, having a size of 4 could be 2 partners and 2 kids, but also 1 partner and 3 kids. This difference must be included in some way, otherwise the analysis could be biased.

Figure 1: Average health-related subjective ratings and age per income level.



- c. Understanding the factors influencing personal well-being is crucial when it comes to policy making. This analysis helped clarifying how specific drivers could be impacting proxies for people's well-being. First of all, we noticed how an increase in the total household income does not generate a significant (better, a huge) increase in how happy respondents perceived their lives. This result is indeed controversial³, but not new to the economics sphere (even if not exactly the same, the Easterlin Paradox could be mentioned, that is that over time happiness does not trend upward as income grows⁴). The educational level and the gender could both play a role in this matter (p-value less than 0.05 in the multivariate regression), but further studies could be needed: people having higher degrees usually face greater responsibilities, which could mean higher stress and lower levels of those ratings. By the same logic, women's lives often involve looking after kids in a more prominent way than what happens with men, inevitably impacting this study. In any case, an interesting further research could better investigate these topics, maybe deepening the scope of the questions in the

³The expectation is usually that the higher the money one owns, the better his / her life.

⁴Topic faced during personal studies.

survey to adapt for these specific details.

Happiness does not seem to be necessarily bought by money. However, the situation is slightly different when facing aspects such as health, mental well-being or anxiety. The scatterplot shows how the average ratings people give when health is involved does increase as income increases. This should not come as a surprise, as the greater the financial means of a family, the greater the possibility to stay healthy in general terms. This does not seem to be true for the upper end of income levels.

When including age in this topic, we notice how, even if it is not always the case, the older the average respondent, the lower the score registered: health problems could be faced as people grow old, even though it must be underlined how, per each income level, the average age of respondents is not exaggerated (circa between 25 and 45). Extending the analysis to younger and older generations could be crucial to clarify this relationship in a more comprehensive way⁵.

⁵This could be done either by increasing the percentage of very young and very old people answering the survey, or simply re-running the analysis / scatterplot for maximum and minimum levels of the considered variables.