Predictors of income

Naiara F. V. Castello

Introduction and data exploration

To perform this analysis, I used simulated data based on a dataset from the Survey of Health, Ageing and Retirement in Europe (SHARE). As a preliminary examination, I investigated the relationship between variables in the dataset.

The number of people in the household was thought to be a relevant predictor of household income, based on the moderate linear association between these variables (r = 0.44, see Table 1) and on logical reasoning. Other potentially important predictors were age and years of education, with negative and positive association, respectively (see Figure 1). Being retired was also potentially related to household income, with retired participants having lower income (see Figure 2). However, since age and retirement seem to be strongly related (see Figure 3), it is possible that they account for the same variance in the data and that only one of them would remain in the final model. Number of books, language abilities, and math abilities at age 10 also seemed to be positively associated with household income (see figures 4, 5 and 6).

Although living with a partner seemed to be strongly associated with income (see Figure 7), the variable doesn't seem to convey much additional information about household income when n_household is accounted for (see Figure 8). Moreover, partner's age was strongly correlated with participant's age (r = 0.76, see Table 1), which might cause issues with multicollinearity. In addition to that, age_partner showed a great number of missing values (117 observations, approximately 28% of the total sample), corresponding to people not living with a partner. Thus, I decided not to include lives_with_partner and age_partner in the model.

Regression modelling and results

Succeeding the exploratory stage, I followed a backward selection process to fit models to the data. The first model included all variables in the dataset, except for the ones removed at the exploratory stage. Variables were then progressively removed, one at a time, based on their significance level (highest p-value at each step = removal). This procedure was repeated until all the variables in the model were statistically significant ($p < .05$) and the highest possible R-squared, under this condition, was achieved.

The final model included the variables n_household, retired, books_age_10, and relative_language_ability_at_age_10. The adjusted R^2 of the final model was 0.235, which suggests that these variables, together, explain 23.5% of the variance in household income. The most expressive predictor was n_household. For each extra person in the household, household income increases by 14462 SEK/year, when other predictors are kept constant.

Having less than 25 books at the age of 10 was also an important predictor in the model, with a corresponding estimated coefficient of -5413. Coefficients' estimates, standard errors, *t*-statistic, and significance levels can be found in Table 2. VIF values ranged between 1.024 and 1.072 (see Figure 9), far below the recommended threshold of 5.

Discussion

The model has shown a good fit to the data, with appropriate VIFs and significant estimates for each predictor. Altogether, the predictors included in the model explain 23.5% of the variance in household income. This is a good effect size, considering the nature of the phenomenon under investigation, which is expected to be sensitive to a number of personal and environmental factors.

However, the sample characteristics indicate that we should not generalize the results to different populations.

One important aspect to be discussed is age. Participants' age ranged between 50 and 100 years (M = 70.38, SD = 8.48), with 73.57% of the sample being retired. The mean age of participants and high rate of retirement might be a reason why retirement wasn't a strong predictor and why age wasn't even included in the final model. In a different age range, including younger subjects, the relationship between retirement and income, as well as between age and income, might have been different. Including a more diverse age range could also increase the variance in income and cause other variables, like years of education, to become significant predictors.

Additionally, the plot of the linear relationship between household income and age, grouped by retirement (Figure 3) suggests that, in this sample, age of retirement might be relatively stable, that is, people retire at about the same age. This could mean that age and retirement account for the same variance in the data, which explains why the final model didn't include both variables simultaneously. Thus, in a context where people don't generally retire in their mid-sixties, these variables could play a different role.

Data was collected in 2008, 2009, and 2011. Since increased life expectancy tends to affect retirement age and, consequently, the household's financial dynamics, the model shouldn't be expected to generalize to different generations either.

Overall, the results are better applicable to a population of older adults in a country with similar characteristics regarding retirement policies and income distribution, at a similar time point.

Supplementary Material

Table 1. Correlation Matrix Including All Numerical Variables in the Dataset

Variable	*M*	*SD*	1	2	3	4
1. n_household	1.79	0.55				
2. age	70.38	8.48	-.32			
3. age_partner	68.49	6.77	-.32	.76		
4. years_of_education	11.21	3.95	.22	-.27	-.18	
5. household_income	29122.56	20100.05	.44	-.23	-.17	.23

Table 2. Estimates of the Multiple Linear Regression of Household Income as a Function of Number of People in the Household, Retirement, Number of Books at Age 10, and Language Abilities at Age 10

Predictor	Estimated coefficient	Standard Error	*t*	*p*-value
(Intercept)				
N Household	14462	1579	9.101	0.00000
Retired (yes)	1997	1997	-2.283	0.02296
Books at Age 10 (-25)	-5413	1804	-3.000	0.00287
LanguageAbilities (same/worse)	-5047	1737	-9.906	0.00386

Figure 1. Graphic Representation of Correlations

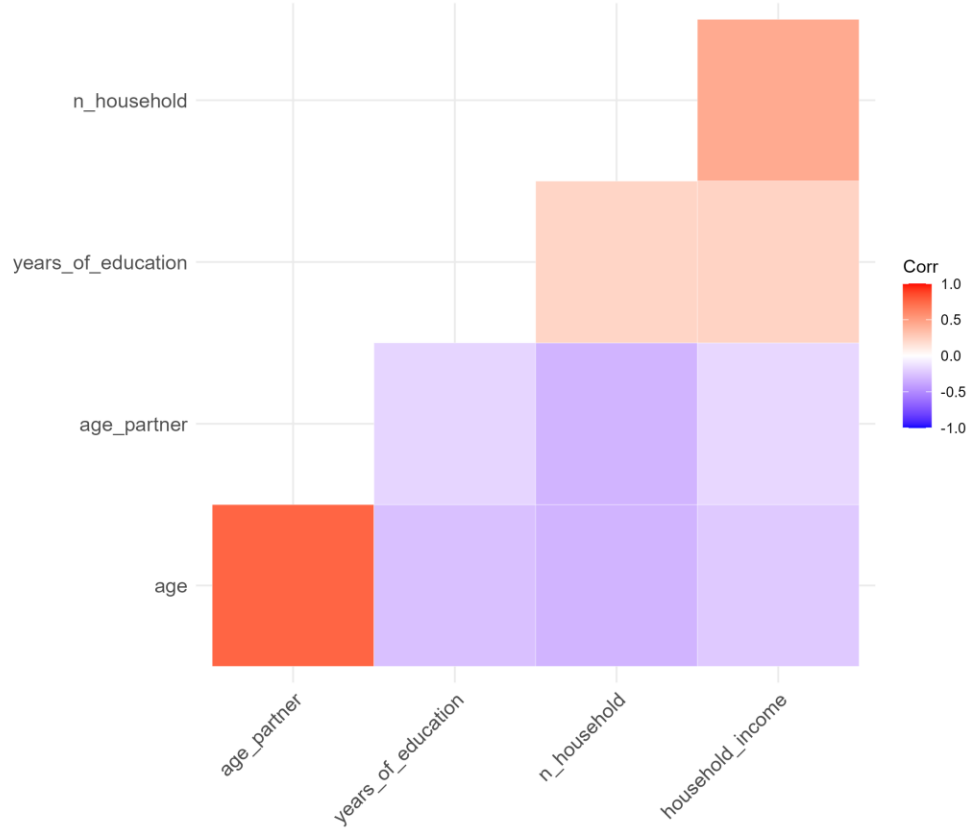

Figure 2. Relationship Between Household Income and Retirement

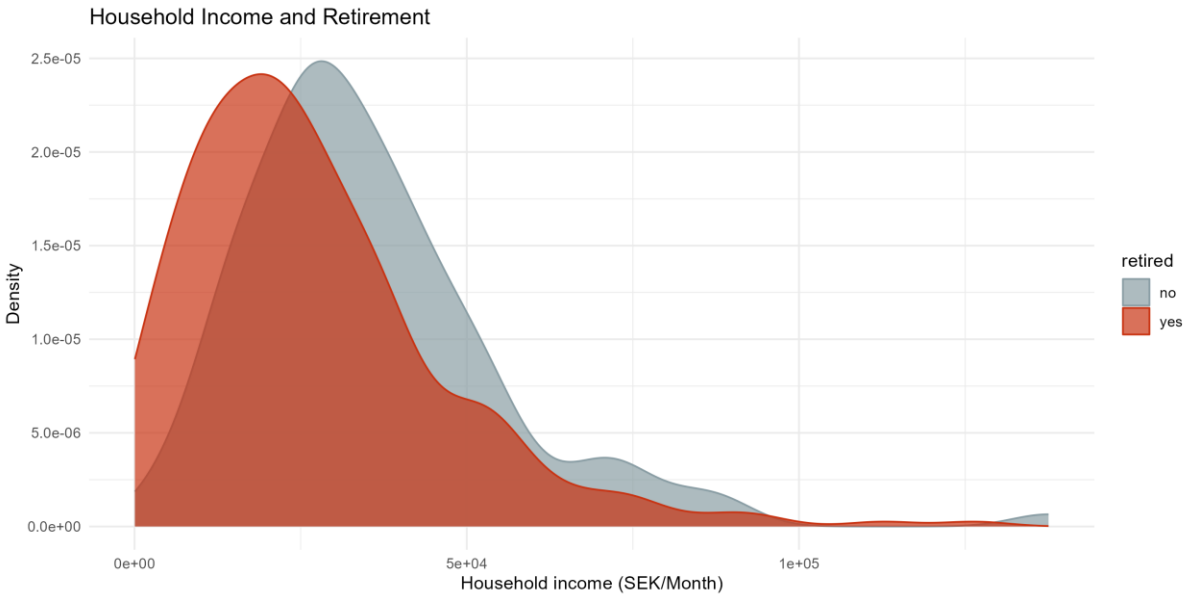

Figure 3. Linear Relationship Between Household Income and Age, Grouped by Retirement

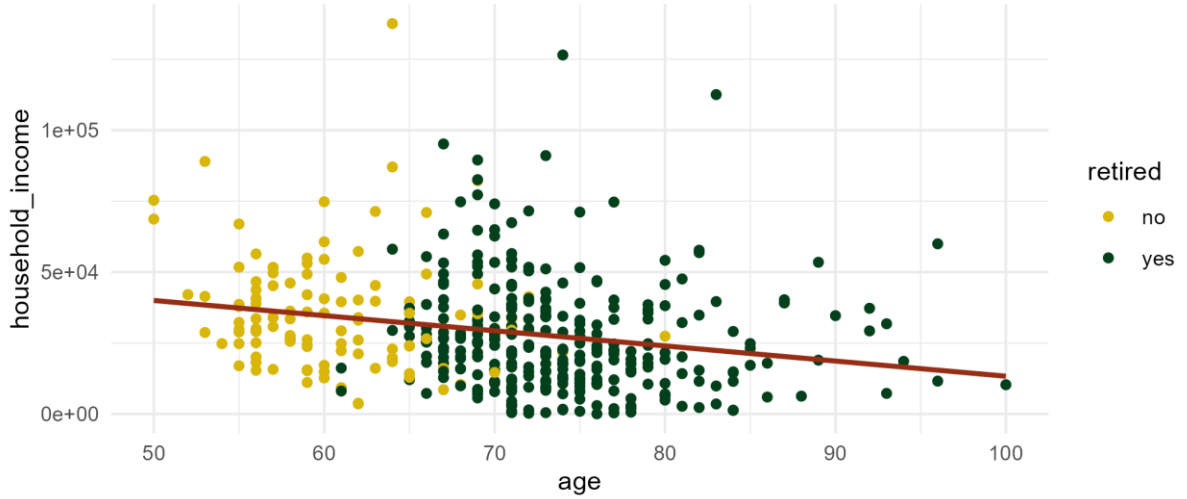

Figure 4. Comparison of Density of Household Income by Number of Books at Age 10

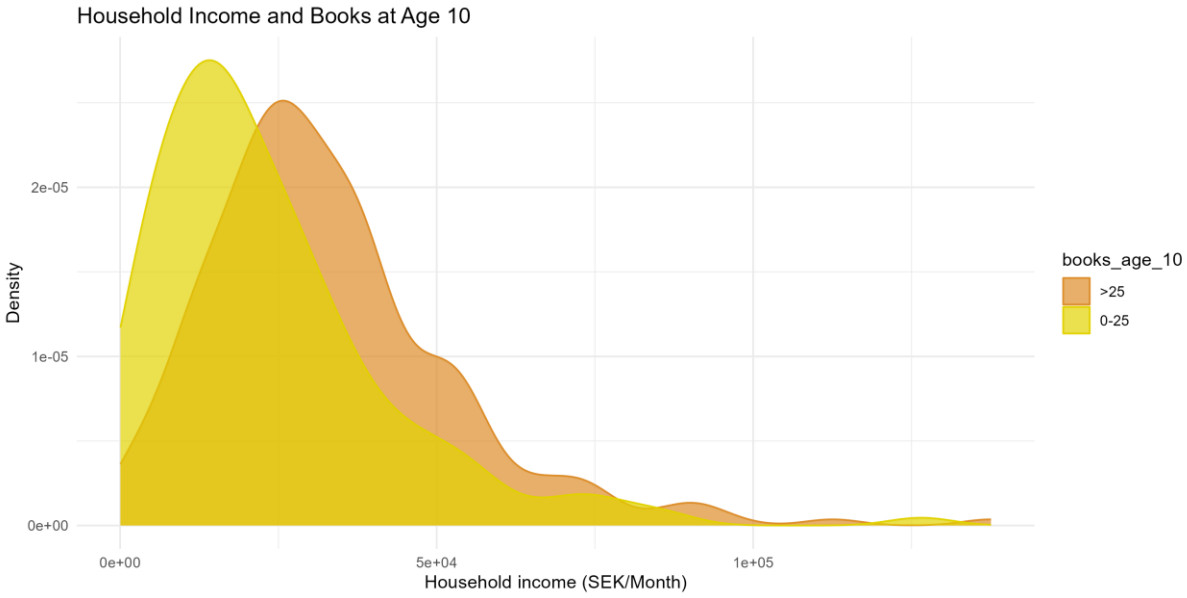

Figure 5. Comparison of Density of Household Income by Language Ability at Age 10

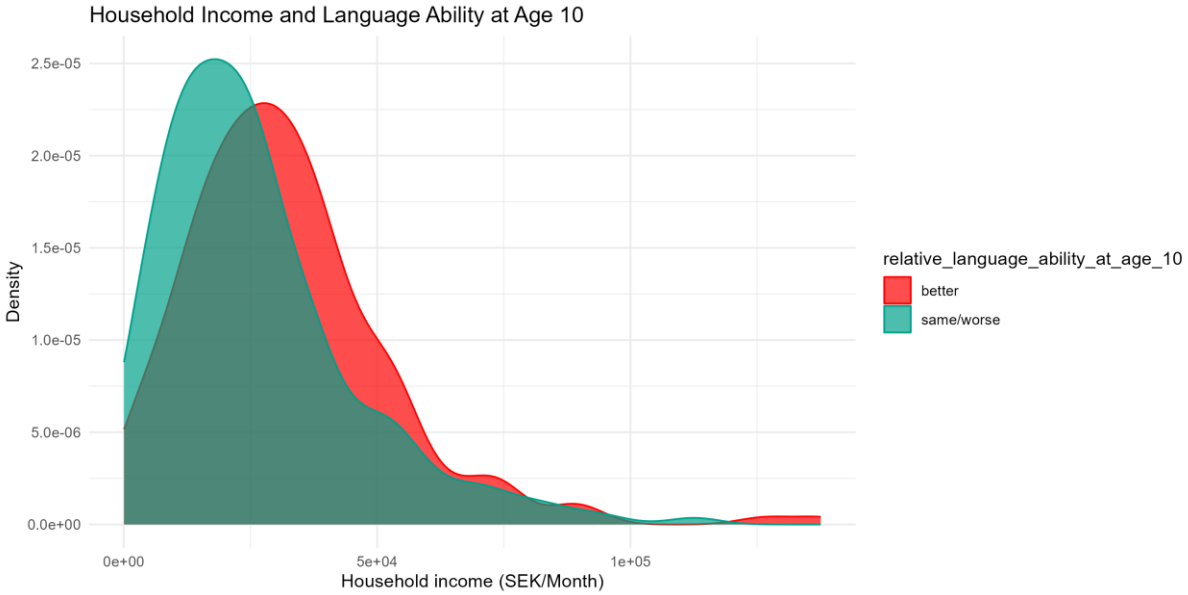

Figure 6. Comparison of Density of Household Income by Math Ability at Age 10

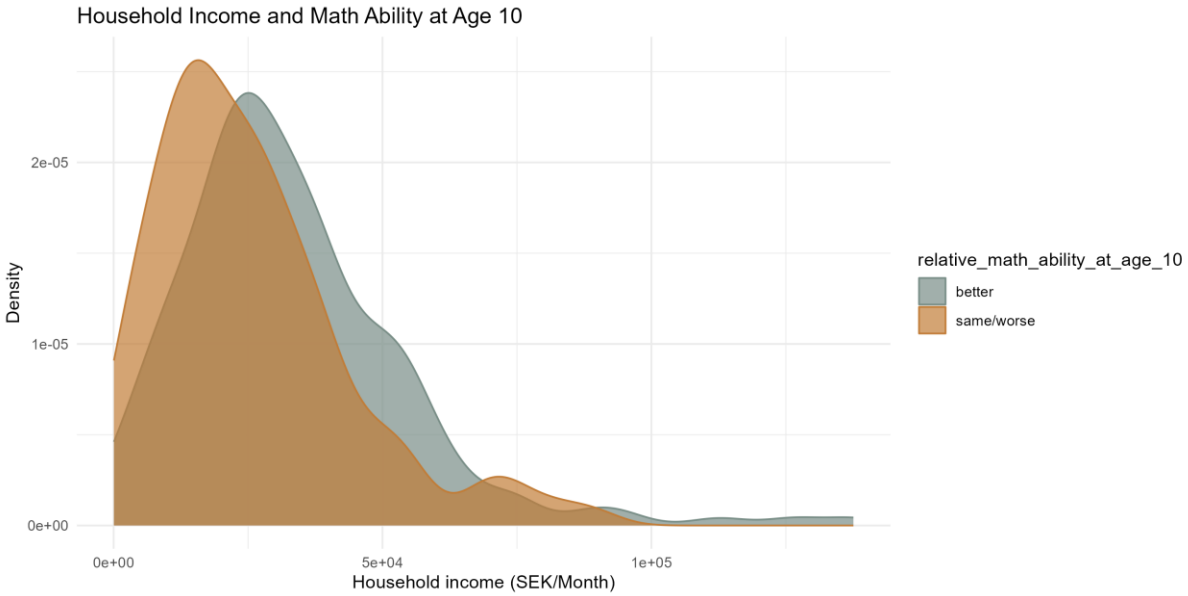

Figure 7. Relationship Between Household Income and Living with Partner

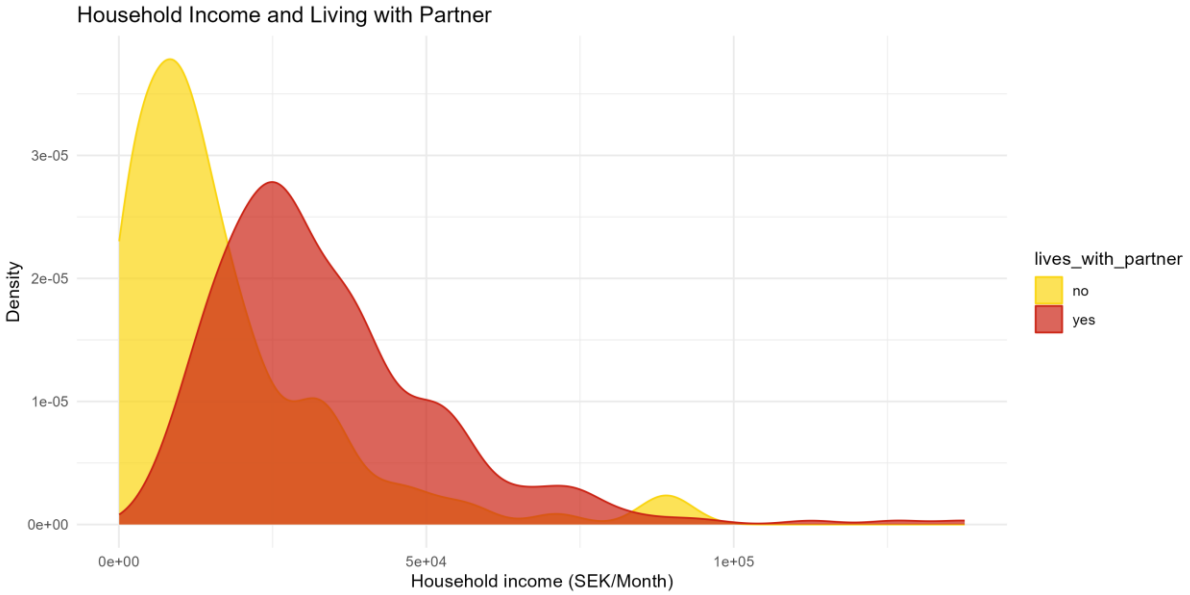

Figure 8. Relationship Between Household Income and Number of People in the Household, Grouped by Living with Partner

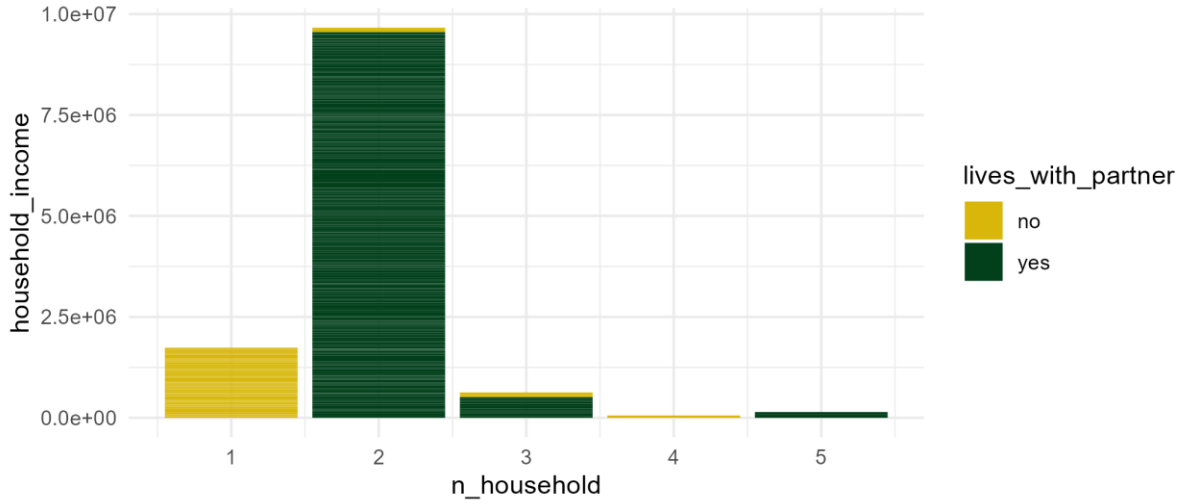

Figure 9. Variance Inflation Factor