Social Salary Study

S. Arora, J. Harmse, V. Mulholland

April 22, 2018

Study Overview

- Clearly state the null hypothesis
- Include study overview with specific methods used
- Support terminology used with other papers
- Define social standards
- Define the relationship behind

Is a person's social standards correlated with a person's drive for financial wealth?

The null hypothesis is that the people who are driven by financial wealth in their careers and those that are driven by job satisfaction share the same social standards in terms of spending patterns.

> We have subjects from all walks of life and are particularly interested in their attitude towards finances and careers.
>
> There could be many factors influencing a participants interest in pursuing a salary increase that could be correlated with the time since the last increase in salary for instance. The question was designed to ask whether a user
>
> Self-reporting is innacurate. The study was designed acknowledging that there could be a certain amount of bias in each person's answers knowing that they may not not have had access to their financial information on hand when responding. Since we are mainly interested in one's relative relatioship between the different monetary values (hence the ratio and the percentages), we are making the assumption that one's biased impression of their finances can still be representive of their beliefs when indicating whether money is the main driver in their career.
>
> That being said, this study has its limitations and could have been improved namely in the categorization of expense categories. It can be difficult to differentiate between the different types of conspicuous spending, and one could consider that living expenses includes the over-priced rent for instance. We would have had to use a benchmark of the average costs which would have made us much more reliant and the actual monetary values that were input, or if we maintain the path we chose of only being reliant on the perceived attitude toward one's finances, we would have had to be more explicit in how we differentiate frivolous spending over basic needs.
>
> The goal of this study is to determine whether a person's social standards are correlated with a person's expected salary. The idea behind the hypothesis is that people who have higher social standards expect a higher salary. The opposite can also be argued - does a person's expected salary determine a person's social standards? This study does not aim to determine which variables are explanatory or a response, but rather to determine whether a strong correlation exists between social standards and expected salary. We found it reasonable to assume that a person who is driven by money would expect to earn more than the average person who has the same skillset and experience.

Considering that participants identified several different country job locations, it was determined that a possible confounder could be the cultural attitude towards spending. Since our respondents were mostly from North America, we categorized the country as either North American or not to get somewhat of an even split.

Methodology

The test statistic will be attempting to identify if there is a strong correlation between social standards and a person's salary expectations. A positive correlation would be expected between the continuous numerical measurement of social standards and the normalized continuous expected salary range. Social standards and expected salaries are expected to both form t-distributions given the survey responses. A linear regression model seems to be an appropriate choice for the study, since our response variable (expected salary) is a continuous range and the explanatory variables related to social standards are expected to have a linear relationship with expected salary.

survey study design

The questions in the Appendix under Survey Questions and are conceptualized from two topics, one pertaining to the salary motivations and the latter is a measure of a participants social standards. This is admittedly a difficult concept to measure, thus our focus is mainly to delineated the difference between essential expenditure versus lifestyle enhacement spendings.

Certain safety checks are put in place to prevent users from entering invalid data. For instance, the spending categories have to add up to 100% of their expenditure

data collection methods

The following table summarises the key fields populated by the survey data and the calculated value, `ratio`, namely the response variable as a ratio of the two salary values.

Features	Description
`salary_base`	An indicator meant to be a subjective baseline of what salary a person of their expertise would earn.
`salary_expect`	The expected salary combined with the base salary provides a relative indicator to the respondents pursuit of monetary gains.
`no_increase_acceptance`	A binary metric used to flag those that are more driven by money than others.
`ratio`	$\frac{Salary_{expect}-Salary_{base}}{Salary_{base}}$ allows us to differentiate those who have a desire for a high increase in salary vs those that are satisfied with a modest amount.
`living_expenses`	Living Expenses (utilities, rent, mortgage, transportation, property taxes if owner, etc.)
`savings`	Savings (retirement, investments, emergency funds, etc.)
`vacation`	Vacation (lodging, transportation, day trips, etc.)
`daily_leisure`	Daily Leisure (eating out, books, movies, self-care, etc.)
`consumption_goods`	Consumption Goods (clothing, electronics, other luxury items, etc.)
`sports_hobbies`	Personal Sports and Hobbies (sporting goods and services, gym, arts and crafts, etc.)
`other`	Other (health care, taxes, dependent expenses, etc.)

The data is anonymized before being made visible and then uploaded as a processed dataset. Several wrangling steps are performed to process and wrangle the data into a usable format.

Analysis

methodology

Our analysis wants to compare the two defined groups (participants who are driven by wealth and those who are not) in terms of spending habits.

Variable Type	Variable Name
`response`	`Participant Group`
`explanatory`	`living_expenses`
`explanatory`	`savings`
`explanatory`	`vacation`
`explanatory`	`daily_leisure`
`explanatory`	`consumption_goods`
`explanatory`	`sports_hobbies`
`explanatory`	`other`

Given the nature of our model variables, a logistic regression model would be the the most appropriate model choice - all the explanatory variables are continuous, whereas the response variable is a binary outcome.

Before excepting that this model is sufficient, it should be considered whether we are dealing with any confounding variables.

Each person who completed the survey had to report their country of employment. Data was collected from a number of different countries. The country a person works in has the potential to influence a person's spending behaviours regardless of their response group. For example, people from different countries may not spend the same amount on vacation, as found by MoveHub.

We need to consider whether country has any significant effect our model. The Appendix visualization shows a number of single observations for different countries, and a larger number of observations from Canada, South Africa and the US. A arguably logical solution to the handling of single country observations would be to group the countries by similarities. Seeing that Canada and the United States of America are neighbouring countries it we can group these two countries as `North America`.

South Africa, Nigeria and Botswana have a lot in common in terms of lifestyle, which means that we could group these observations as `Africa`. The other single observations should be omitted for this comparison, because it would require arbitrary assumptions for classifying these observations.

results

In the Appendix Continent section two logistic regression models are compared to determine the effect of the potential confounding variable.

The Anova test results indicates a p-value of 0.05046. Since this value is on the verge of being below a significance level of 0.05 the `continent` variable (aggregation of the `country` variable) should be included in the model as a variable that potentially has a significant influence.

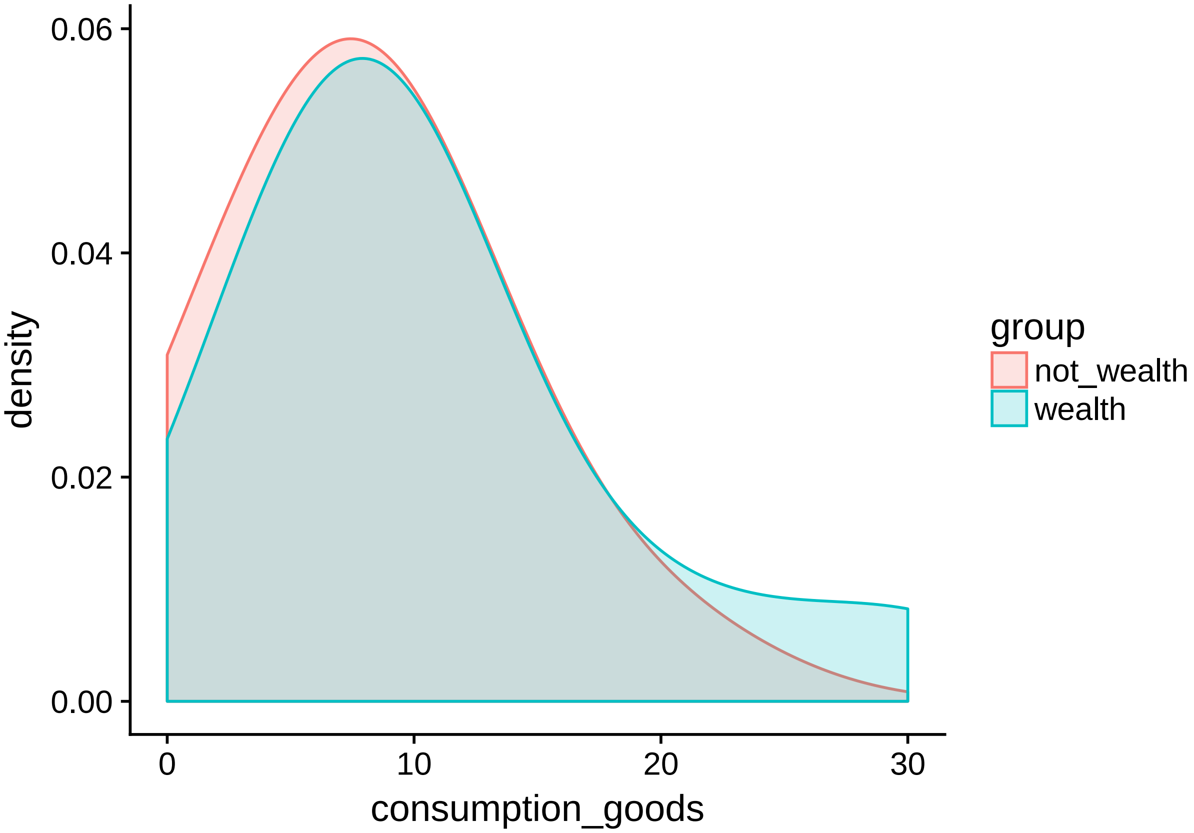
Below follows the results of the logistic regression model which includes the confounding variable as an explanatory variable.

.rownames	Estimate	Std..Error	z.value	Pr...z..
(Intercept)	-15.2712437	8.4172634	-1.8142766	0.0696352
living_expenses	0.1244577	0.0833833	1.4925982	0.1355424

.rownames	Estimate	Std..Error	z.value	Pr...z..
savings	0.0495793	0.0871461	0.5689214	0.5694095
vacation	0.1028606	0.1009215	1.0192139	0.3081014
daily_leisure	0.0480662	0.1035532	0.4641690	0.6425267
consumption_goods	0.2694227	0.1345632	2.0022028	0.0452629
sports_hobbies	0.3313781	0.1709313	1.9386624	0.0525425
continentNorth America	2.7595625	1.7060930	1.6174748	0.1057758

Consumption goods shows some significance. This indicates that people who are classified as driven by wealth may tend to spend more on consumption goods than people who are not driven by wealth.

```r
ggplot(data = survey_results_continent, aes(x = consumption_goods, group = group, colour = group)) +
  geom_density(alpha = 0.2, aes(fill = group), bw = 5)
```



However, the model includes multiple comparisons. The p-values need to be adjusted in order to account for random significance.

P_original	P_adjust(FDR)	P_adjust(Bonf)
0.1355424	0.2371992	0.9487968
0.5694095	0.6425267	1.0000000
0.3081014	0.4313420	1.0000000
0.6425267	0.6425267	1.0000000
0.0452629	0.1838986	0.3168405
0.0525425	0.1838986	0.3677972
0.1057758	0.2371992	0.7404308

Adjusting the p-value removes all significance from the model. The lack of significance can be attributed to the lack of more data. If more data was collected, the study would have had the potential to gain more power.

Discussion

Study Design

> what did you do well to make this study as causal as possible?

The questions were constructed to account for all types of spending so that the respondent could better consider their periodic spending distribution. There are many subjective and psychological features that would contribute to someone's self-assessment of expected and base salary estimates which was accounted for when stating that we are looking at a person's drive for money. It could be

> what was not done well and how did that effect your studies conclusions?

Clarifying the spending categories is a shortcoming of our study. There is a tradeoff between making our survey straightforward and being too transparent about the agenda behind the analysis with very specific questions. There is some ambiguity behind the concept of social standards which we tried to account for in the vacation, hobbies, and daily leisure categories, but acknowledge that one could be partaking in conspicuous consumption while categorizing it as a living expense, such as paying a very high rent to live in the nicest neighbourhood. The "Other" category could also be misleading because there could be some frivolous expenses that are not accounted for.

Self-assessments aren't ideal since the participant is require to think objectively on the spot about their finances. This could inject a considerable source of bias, and would have required a more thorough assessment method than a survey. Providing and export of a bank categorization of one's spendings would be a better method for a true representation.

On the other hand, there are also a lot of psychological fe

> what would you do differently next time to improve your survey/study design and why?

It was good to use the point system to divide the spending because it forced the participant to consider each category of interest and then associate the rest with "other". A limitation of our design was that the

Results Discussion

Conclusion

Appendix

Survey Questions

1. What is your country of employment/future employment? (used for determining currency for following questions)
2. Assuming the country's currency specified above, what should someone with your qualifications and experience expect to receive as an annual salary?
3. Assuming the country's currency specified above, what is the annual salary that you aim to receive 1 year from now?
4. Assuming high job satisfaction, would you keep a job that does not give you a salary increase over the next two years?
5. Please assign an approximate percentage of your current yearly expenses to the following categories (must sum up to 100).
   - Living Expenses (utilities, rent, mortgage, transportation, property taxes if owner, etc.)
   - Savings (retirement, investments, emergency funds, etc.)
   - Vacation (lodging, transportation, day trips, etc.)
   - Daily Leisure (eating out, books, movies, self-care, etc.)
   - Consumption Goods (clothing, electronics, other luxury items, etc.)
   - Personal Sports and Hobbies (sporting goods and services, gym, arts and crafts, etc.)
   - Other (health care, taxes, dependent expenses, etc.)

EDA

Confounding Variables

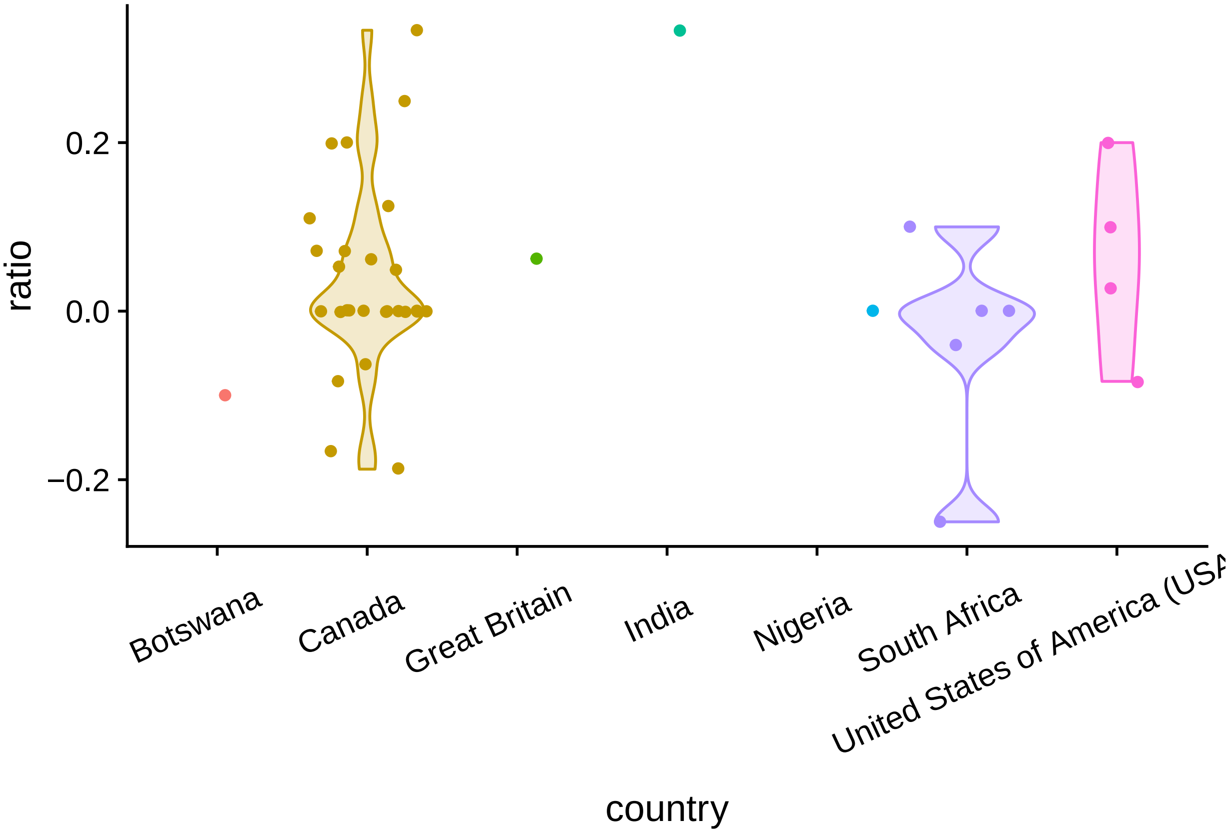
Continents

We want to determine whether people who are driven by wealth have different spending habits than people who are not driven by wealth. The nature of the hypotheiss and the data makes a logistic regression model the obvious choice for determining whehter any expense category is significantly different between the two groups.

Before we start building the model we should consider whether we are dealing with any confounding variables.

Each person who completed the survey had to report their country of residence. Data was collected from a number of different countries. The study's response variable is standardised salary ratio, which does not require taking the person's country into account. However, the country a person live has the potential to play a role in a person's spending habits regardless of their salary ratio. For example, a person from Africa would not necessarily spend a lot on vacations in comparison to a person from North America which may be a result of something like cultural differences.

We need to consider whether country has any significant effect on either out explanatory expense variables or our response variable.



After removing the observations that do not match our criteria for either a person who is driven by wealth or not driven by wealth, we are left with a relatively small number of observations. The visualization above shows a number of single observations for different countries, but quite a few observations from Canada, South Africa and the US.

The simplest way of determining whether country has an influence on our model, we should influde a person's country as an explanatory variable and determine whether this variable has any statistical significance. Country is a categorical variable which makes the single observations difficult to work with when trying to determine its effect. A arguably logical solution to the handling of the single observations would be to group the countries by similarities. Seeing that Canada and the United States of America are neighbouring countries it we can group these two countries as `North America`.

South Africa, Nigeria and Botswana have a lot in common in terms of cultural lifestyle, which means that we could group these observations as `Africa`. The other single observations should be omitted for this comparison, because it would require arbitrary assumptions where these data points would fit in.

```
## Warning in tidy.anova(.): The following column names in ANOVA output were
## not recognized or transformed: Resid..Df, Resid..Dev, Deviance
```

Resid..Df	Resid..Dev	df	Deviance	p.value
31	38.15569	NA	NA	NA
30	34.32947	1	3.82622	0.0504566