Formative Assignment

B149008

January 2024

1 Introduction

The theory of basic human values was first developed by Shalom H. Schwartz [2]. It aims to describe human behaviour and thought processes by focusing on understanding what may drive mental processes and behaviours present in all humans, and especially understanding where and how differences among individuals may arise. It also outlines the features that all of these values might share with each other, along with how they differ as well. He outlined ten personal values with a motivation or goal associated with each of them, and he grouped and described them in the following manner:

Openness to change

- Self-Direction – the ability to think and act independently, typically expressed through choices, creations or exploration.
- Stimulation – pursuing novelty, excitement and challeneges in one's life.

Self-Enhancement

- Hedonism – the prioritization of pleasure in one's life.
- Achievement – demonstrating competence in accordance with societal standards to obtain personal success.
- Power – Prestige, control or dominance over other people or resources.

Conservation

- Security – Stability, safety and harmony in relationships, society, or self.
- Conformity – restricting oneself to behave in accordance with societal standards and ideals.
- Tradition – dedication and commitment to fulfilling and participating in the customs and following ideals from one's religion or culture.

Self-Transcendence

- Benevolence – the desire to behave and act in a way which improves and enhances the welfare of those who are around oneself (described by Schwartz as the "in-group").
- Universalism – the desire to be broadminded, promote welfare, harmony, and justice for all people.

The aim of this report is to use longitudinal survey data to fit and evaluate a linear regression model predicting how an individual's prioritisation of the value of universalism may vary with age.

2 The Data

The European Social Survey (ESS) evaluates human values as part of a biannual survey [3]. It asks the participants questions, nested within countries, which provide a score that each individual may assign to each of the ten values described in §1. All analysis of the data was carried out using `R` Statistical Software [1].

The data is made up of 16 variables in total, with 10 of them being the respective scores associated with each of the values. The remaining variables are given by the following:

- `indiv`: the individual's ID in the data.
- `Country`: the individual's country of origin.
- `gender`: the individual's sex.
- `age`: the individual's age (in years).
- `eduyrs`: the number of years of education completed by the individual.
- `income`: individual's house income (split into 12 categories).

When performing statistical analysis with this data, we have four possible explanatory variables, namely `gender`, `age`, `eduyrs`, and `income`. The scores for each of the ten values are our possible response variables.

The dataset contains 36,537 entries and has 7,499 missing values in total. In order to decide upon an explanatory and response variable, we may look more closely to see which variables in particular have the largest number of missing values. We have the number of missing values per each variable given in tables 1 and 2.

Variable	**No. of missing values**
`gender`	31
`age`	164
`eduyrs`	445
`income`	6787

Table 1: No. of missing values for each of the possible explanatory variables.

Variable	**No. of missing values**
Self-Direction	9
Stimulation	10
Hedonism	2
Achievement	6
Power	2
Security	7
Conformity	26
Tradition	6
Benevolence	4
Universalism	0

Table 2: No. of missing values for each of the possible response variables.

We see that among the possible response variables, universalism is the only variable with no missing values. Hence, to ensure we have as much data for analysis as possible, we may choose to predict universalism. We also see that among the possible explanatory variables, `gender` has the fewest

missing values. However, it is a binary variable, whereas for the purpose of this report it would be beneficial to have a continuous or categorical explanatory variable. Hence, we may choose `age` to be our primary explanatory variable in this case. In order to ensure there are no issues in modelling the data, we may remove all entries containing a missing value. In doing this, we are then left with a total of 36,362 entries. We can visualise the relationship between our variables with the plot shown in figure 1.

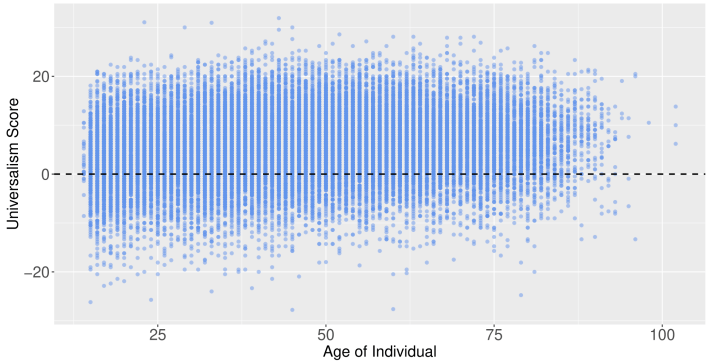

Figure 1: Plot showing how the universalism score of an individual may vary by age.

Here, we see that while there is no obvious or strong relationship between universalism score and age, there does seem to be a slight increase in the minimum values of the universalism score with increasing age. In addition to this, table 3 also outlines some of the descriptive statistics for our explanatory and response variable. We see that `age` ranges from 14 to 102. This means that when

Statistic \ Variable	Age	Universalism
Minimum value	14	-27.778
1st Quartile	32	1.905
Median	45	6
Mean	46.16	5.811
3rd Quartile	60	10
Maximum value	102	31.9057

Table 3: Descriptive statistics for the explanatory and response variable.

modelling, we'd like to ensure that we can get a meaningful intercept that is realistic for the data. Hence, we may center the `age` variable around its mean. In doing this, the new descriptive statistics are given in table 4.

Statistic	Value for `age` (centred)
Minimum value	-32.16083
1st Quartile	-14.16083
Median	-1.16083
Mean	0
3rd Quartile	13.83917
Maximum value	55.83197

Table 4: Descriptive statistics for centred `age` variable.

3 Modelling the Data

3.1 Simple linear regression

In order to understand the impact that age may have on the universalism score of an individual, we make use of linear regression. Hence, we fit a model with the following equation, representing the relationship between the universalism score of the i-th individual and their age:

$$\text{Universalism}_i = \beta_0 + \beta_1 \cdot \text{age}_i + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

where β_0 is the model's intercept, β_1 is the model's slope and ϵ is an error with mean 0 and variance σ^2. In practical terms, the model's intercept represents the value of universalism score we expect the average individual from our sample to have. The slope represents the impact one's age may have on their universalism score.

Upon implementing this model in `R`, we obtain a model with an intercept of 5.811 (3 d.p.) and slope of 0.053 (3 d.p.). Both of these coefficients have a p-value that is significant. Hence, we have the equation given by

$$\text{Universalism}_i = 5.811 + 0.053 \cdot \text{age}_i + \epsilon_i, \quad i = 1, 2, \ldots, n. \tag{2}$$

So, we have that for the average individual from our sample, we expect them to have a universalism score of 5.811, represented by the intercept, and that for each unit increase in age, we expect to see the universalism score to increase by 0.053. While this is a positive association between age and the universalism score, it is very small. In order to further test the validity of this model, we may also inspect the model diagnostic plots, shown in figure 2. We see from figure 2 that our model satisfies the general assumptions of linearity, normality of the data, and homoskedasticity.

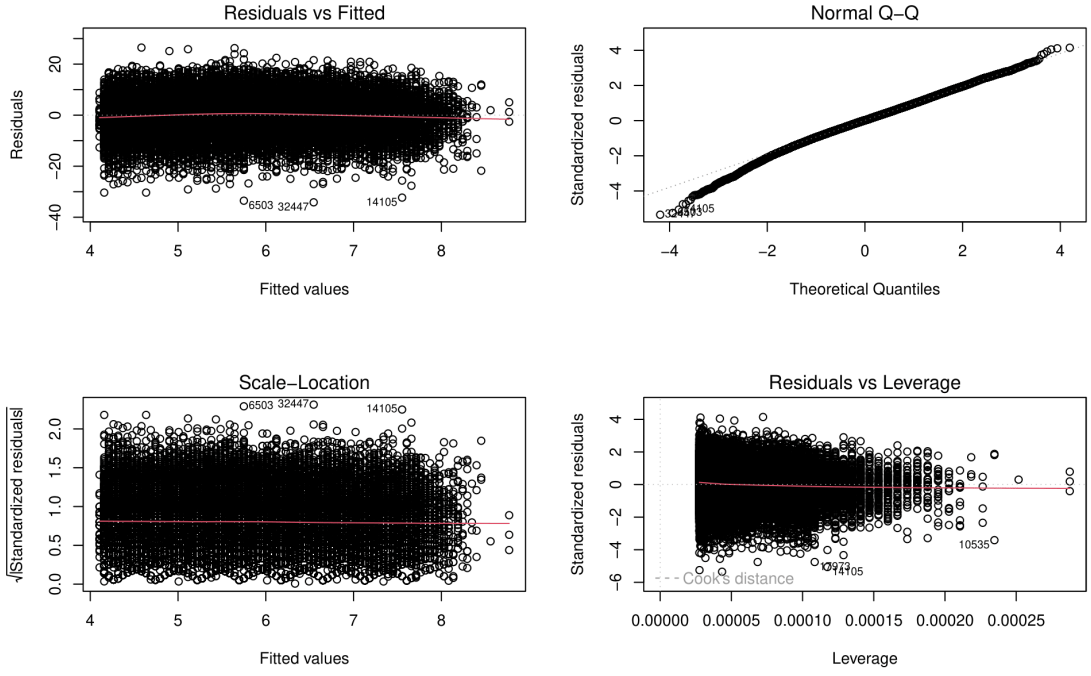

Figure 2: Model diagnostics plots for the simple linear regression model, modelling universalism score with age.

3.2 Including an interaction term

Upon performing and testing the simple linear regression model, we may now consider a model where we also have an interaction term between the age and sex of an individual. However, before modelling, we may also check whether there is any obvious relationship between an individual's sex and their universalism score. This is shown in the boxplot in figure 3. Here, we see that there is a very slight increase in universalism score for women compared to men. We may further check this with our model. In order to include gender in our model, we must first encode the gender variable to be a

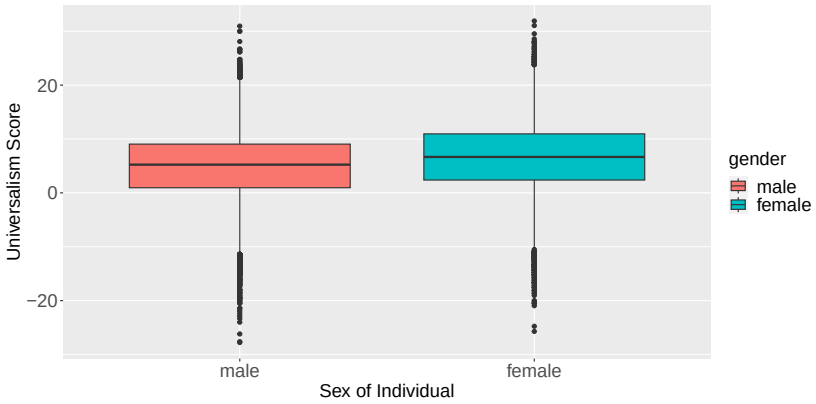

Figure 3: Boxplot showing the difference in universalism scores between males and females.

dummy variable, which is either 0 or 1 if the individual is a male or female, respectively. Hence, upon including our interaction term, we have the following model:

$$\text{Universalism}_i = \beta_0 + \beta_1 \cdot (\text{age} \times \text{gender})_i, \quad i = 1, 2, \ldots, n, \tag{3}$$

where β_1 is the coefficient associated with the impact that age might have on an individual given that they are female as compared to male. Upon implementing this in `R`, we get the following model:

$$\text{Universalism}_i = 5.799 + 0.046 \cdot (\text{age}_i \times \text{gender}_i), \quad i = 1, 2, \ldots, n. \tag{4}$$

All of these coefficients are significant, and we notice from this that being a female compared to being a male has a positive coefficient, suggesting that when compared to males, females are likely to have a higher universalism score, by factor of 0.046. This result agrees with what we saw in figure 3, but it is worth noting that the coefficients for the simple linear regression model shown in equation (2) are not too different from the coefficients of this model. Hence, we may conclude that while the sex of an individual may have an impact on their universalism score, it is not particularly large. As with the simple linear regression model in §3.1, the model coefficients for this interaction model are all significant and the model satisfies our modelling assumptions of linearity, normality and homoskedasticity.

3.3 Including a squared term

Finally, we may now include a term in our simple linear regression model from §3.1 which represents age^2. Then, our new model is given by

$$\text{Universalism}_i = \beta_0 + \beta_1 \cdot \text{age}_i + \beta_2 \cdot (\text{age}_i)^2, \quad i = 1, 2, \ldots, n, \tag{5}$$

where β_2 is the coefficient associated with the age^2 variable in our model. Upon implementing this in `R`, we obtain the following model:

$$\text{Universalism}_i = 6.3270 + 0.0589 \cdot \text{age}_i - 0.0016 \cdot (\text{age}_i)^2, \quad i = 1, 2, \ldots, n. \tag{6}$$

Here, since the coefficient associated with age^2 is negative, we have that per unit increase in age, the *increase in universalism score* decreases by 0.0016. As before, all coefficients in this model are significant and the diagnostic plots validate our general modelling assumptions of linearity, normality and homoskedasticity.

4 Conclusions

In the report we have used linear regression to try and understand the association age may have with the importance that an individual places on the value of universalism, described by Schwartz as the desire for welfare for all people [2]. In implementing a simple linear regression model we found that in general, an increase in age was associated with a higher universalism score. While age is an important factor, we also considered the sex of the individuals, and how this may interact with their age. We found that when compared to males, being a female was associated with a slightly higher universalism score, suggesting that women may value universalism more than men on average. Finally, we also included a squared term to describe how the relationship between age and universalism score varies with age, and we found that as age increases, the positive association between age and universalism score generally weakens.

References

[1] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: https://www.R-project.org/.

[2] Shalom Schwartz. "An overview of the Schwartz theory of basic values". In: *Psychology and Culture Article* 2.1 (Dec. 2012), pp. 1–20. DOI: https://doi.org/10.9707/2307-0919.1116. URL: https://scholarworks.gvsu.edu/cgi/viewcontent.cgi?article=1116&context=orpc.

[3] Norwegian Science Data Services. *ESS Round 1: European Social Survey Round 1 Data*. 2002. URL: http://www.europeansocialsurvey.org/data.