STA 160 Midterm Project: Narcissistic Personality Inventory Analysis

Chapman, Lamba, Spiga, Vien

May 7, 2022

Contents

1 Introduction

For this project, we chose to explore a dataset from the Open-Source Psychometrics Project, a collection of personality tests for psychological research. This dataset consists of metadata collected from 11,243 Narcissistic Personality Inventories (NPI), containing survey questions, answers, and a final "narcissism" score from a sample of respondents. It also contains demographic information on age and gender, resulting in about 11,000 rows in the dataset.

2 Data Cleaning

Before approaching any form of data exploration or analysis, we had to ensure that our data was illustratable and interpretable. The fourty questions (Q1 through Q40) are trinary put in a format a 1 for the first responsem, 2 for the second response, or 0 if the value no answer selected. The provided codebook defines which choice was the corresponding narcissistic choice to be logged in the summation in the end for a score out of 40. In order to format this scoring system into our data, we needed to transform each corresponding question to reflect where a 1 would be a narcisstic answer and 0 would be a non-narcissitic response. However, before this was done, we ensured that we removed those who did not complete the survey, so that incomplete data would not harm our later analysis. By the end of this manual encoding, we had binary columns of 1 and 0 and a reduction of 825 surveyees.

Having each column corresponding to a survey question as the letter Q with a ordinal value attached leads to a lot of ambiguity and ineffiencey in analysis. To remedy this, the next step in our data cleaning was going through each question to summarize each question with the chracteristic addressed, and renaming the column to that summary word or phrase. For example, the first question Q1 aks to pick one of the following: "I have a natural talent for influencing people," or "I am not good at influencing people." This column was then renamed "Influence." We must note that this naming system is in no way a perfect nor transparent way of paraphrasing the question, the purpose is simply for efficiency of analysis.

While the above paragraphs decribe the methodology of cleaning the columns in our dataset pertaining to questions, we did not neglect our variables describiing the surveyees. We are given 3 gender categories, 1 for Male, 2 for Female, and 3 for Other. However, we removed those with gender equal to 0 as these represented missing values. Secondly, we came across individuals with inappropriate ages. The codebook provided by the data said that anyone under the age of 14 was omitted, however we had values less than 14. Oddly, we had age values of ages greater than 100. While this is not impossible, we were very skeptical of those whose age was logged as 366 and 509 years old. We removed ages above 117, assuming this was due to input error. After these proccesses, our data was no suitable for exploration and analysis.

3 Method of Approach

4 Exploratory Data Analysis

4.1 Methodology

4.1.1 Response Distribution of Survey Questions

4.1.2 Score Against Demographics

4.1.3 Word Cloud

4.1.4 Contingency Table

4.1.5 Correlation

5 Two-Factor ANOVA Test

5.1 Methodology

5.2 Results

6 Multinomial Logistic Regression

We wanted to explore the ability to predict gender and age group given the responses a respondant inputted. One of the methods we perfomed in order to do this was a Multinomial Logistic Regression. Taking advantage of the sklearn's functions we were able to build a Multinomial Logistic model with a L_2 norm regularization. Before building the model, we did a 70/30 train-test split on the dataset, so we could measure how the model predicted unseen surveyees. In order to verify the accurarcy, we performed a Stratified 10-Fold Cross Validation and found the mean accuracy to be 0.635 with a standard deviation of 0.019.
We look at which questions have the greatest influence on deciding which gender. Given that this model is a multinomial logistic regression, we have three coefficients vectors, one for each gender category.

6.1 Methodology

6.2 Results

7 LASSO Regression

7.1 Methodology

7.2 Results

8 Member Contributions

In this section, we explain the work each member contributed to the finished report:

A Code

```
# insert code if needed in the report
```