Can We Predict Big 5 Personality Traits from Demographic Characteristics?

Ethan Woods & Dr. David Han
The University of Texas at San Antonio, TX 78249

Abstract

Here we aim to predict the Big Five personality traits based on the demographic information using a generalized linear model. Data was obtained from openpsychometrics.org, pre-processed in MS Excel, and imported to R for statistical analysis. First, it was attempted to predict each individual response item using an ordinal regression model. It was however found to be not viable, even after various weightings were then applied to the demographic data. The response variables were then aggregated to form five categories, one for each personality trait: **conscientiousness, agreeableness, neuroticism, openness to experience, and extraversion**. We then applied a dimension reduction technique to the country variable as well as the race variable in order to achieve an adequate model fit. It was determined that although the demographic information could be useful, precise prediction of the Big Five traits require other information that was not captured in the dataset.

Introduction & Background

Factor analysis performed in the 1980's and onwards has indicated that a five-factor model for personality is most appropriate. These factors are contained within the OCEAN acronym: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism.

Openness to experience is the extent to which people are receptive to new ideas. Conscientiousness describes the degree of thoroughness applied to one's work. Extraversion is a measure of how outgoing or sociable a person is. Agreeableness describes the degree to which an individual is cooperative and considerate. Neuroticism is the tendency to harbor negative emotions.

It may be possible to predict the personality traits from demographic variables such as age, handedness, country of residence, gender, and so on. In this work, we attempted to construct a generalized linear model to predict the Big Five personality traits from demographic characteristics. Through this research, we aimed to identify statistically significant factors that may contribute to the development of certain personality characteristics.

Research Objective

- to establish a predictive model of the Big Five personality traits from the demographic characteristics of an individual

Methods

Dataset was acquired from the Open-Source Psychometrics Project (https://openpsychometrics.org/_rawdata/). This large dataset (*n*=19,719) was exported to MS Excel for pre-processing. Countries with an insufficient amount of data were removed to correct the rank deficiencies in the dataset. Data was then imported to R for statistical analysis. The programming language R was then used for data visualization and to perform all the predictive analyses.

	A	B	C	D	E
1	race	age	engnat	gender	hand
2	1	13	1	1	1
3	1	36	1	1	1
4	1	16	1	1	1
5	1	22	1	1	1
6	1	54	1	1	1
7	1	14	1	1	1
8	1	30	1	1	1
9	1	40	1	1	1

Figure 1. A screenshot of the psychometric data after exporting to MS Excel

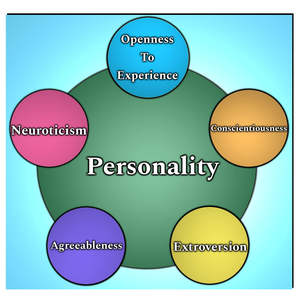
```
B5Data<-read.csv("C:/Users/Ethan/Desktop/Big_5_Data2
.csv", header=T)

summary(B5Data)
```

Figure 2. A screenshot of the R command used to import the psychometric data to R

Results

Figure 3. A table of residual deviance, AIC, and accuracy values for each item; note low accuracy achieved when predicting individual response items.



Test Item	Residual Deviance	AIC	Accuracy
E1	70.2300	254.2300	0.2736
E2	72.7900	256.7900	0.2503
E3	69.6600	253.6600	0.2967
E4	71.0500	255.0500	0.2721
E5	69.4200	253.4200	0.2859
E6	69.1400	253.1400	0.3158
E7	73.3000	257.3000	0.2458
E8	71.2700	255.2700	0.2628
E9	73.8700	257.8700	0.2401
E10	70.1600	254.1600	0.3311
N1	72.3500	256.3500	0.2646
N2	70.0300	254.0300	0.3096
N3	65.3400	249.3400	0.3694
N4	71.8900	255.8900	0.3028
N5	72.5500	256.5500	0.2712
N6	73.0600	257.0600	0.2630
N7	72.2300	256.2300	0.2838
N8	72.1100	256.1100	0.2682
N9	72.6000	256.6100	0.2823
N10	72.3900	256.3900	0.2598
A1	65.3300	249.3300	0.3997
A2	62.6600	246.6600	0.3882
A3	62.8400	246.8400	0.4002
A4	57.2500	241.2500	0.4274
A5	63.8400	247.8400	0.3834
A6	61.8400	245.8400	0.4034
A7	63.9500	247.9500	0.3877
A8	62.4700	246.4700	0.3979
A9	60.1300	244.1300	0.4217
A10	63.4600	247.4600	0.3551
C1	67.0600	251.0600	0.3407
C2	73.8800	257.8800	0.2542
C3	60.4800	244.4800	0.3886
C4	69.8700	253.8700	0.2902
C5	70.1300	254.1300	0.2954
C6	73.6100	257.6100	0.2510
C7	65.7800	249.7800	0.3354
C8	66.5600	250.5600	0.3495
C9	71.3500	255.3500	0.2823
C10	62.8200	246.8200	0.3592
O1	62.3400	246.3400	0.3603
O2	64.2100	248.2100	0.3501
O3	57.2100	241.2100	0.4642
O4	63.4800	247.4800	0.3812
O5	58.6600	242.6600	0.3905
O6	55.5900	289.5900	0.5305
O7	56.0800	240.0800	0.4201
O8	69.8900	253.8900	0.3053
O9	56.4200	240.4200	0.4364
O10	58.8100	242.8100	0.3973

Results - con't

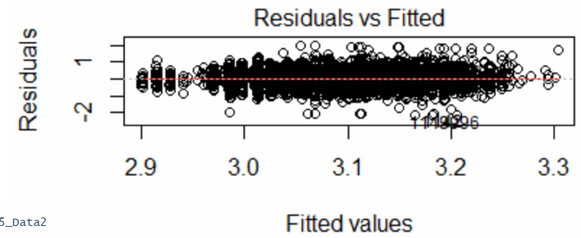


Figure 4. A residual plot when the Extraversion domain is predicted from all demographic variables

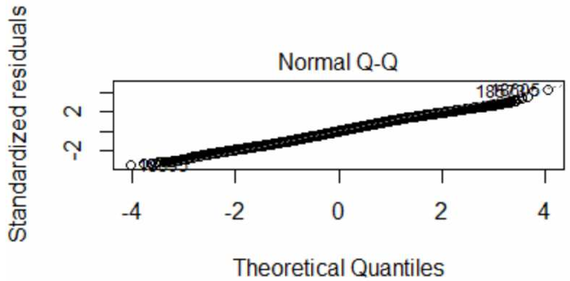


Figure 5. A Q-Q plot constructed for the Neuroticism domain when the trait was predicted from all demographic variables.

There appears to be a linear fit, indicating the data is roughly normally distributed.

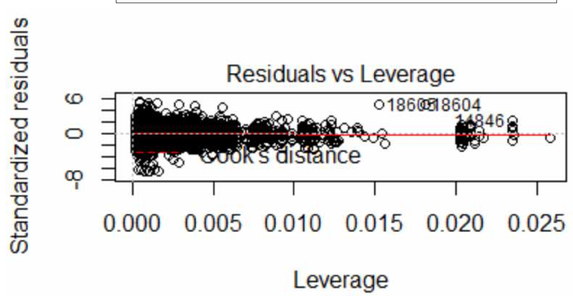


Figure 6. A residuals vs. leverage plot constructed when the Agreeableness domain was predicted from demographics

Conclusions

- Using a large open psychometric dataset, it was attempted to statistically predict the Big Five personality traits based on the demographic information of an individual.
- Based on the ordinal regression models, it was concluded that the demographic variables are poor indicators for the individual response items on the Big Five personality assessment.
- Based on the generalized linear models on the aggregated scores over each domain, it was found that the demographic variables are statistically significant but they alone are insufficient to predict the five personality domains with a strong accuracy. Other pertinent information of an individual that was not captured by the dataset is required to improve the predictive power of the personality analysis.

References

- _____ (2020). *Open-Source Psychometrics Project*. retrieved from https://openpsychometrics.org/
- Goldberg, L.R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, **4:** 26-42.
- Hirschfeld, G., von Brachel, R., and Thielsch, M. (2014). Selecting items for Big Five questionnaires: at what sample size do factor loadings stabilize? *Journal of Research in Personality*, **53:** 54–63.

Acknowledgements

I would like to thank Dr. Han for all of his guidance. He was immensely patient and helpful throughout the entire research process. I was exposed to many new statistical techniques that I had never seen before, and I am thankful that I was provided with this opportunity to expand my knowledge.

UTSA