Multivariate Regression with Categorical Responses

Logistic Regressions for Predicting Personalities with Python and Epidemic Disease Outbreak using SAS

*May 7, 2018*

## Authors

Robert Nakano [robertnakano@gmail.com](mailto:robertnakano@gmail.com)

Sunny Jelokhani [afsanehkhani35@gmail.com](mailto:afsanehkhani35@gmail.com)

Ziyi Wang [joannewangziyi@gmail.com](mailto:joannewangziyi@gmail.com)

## Abstract

This report examines Categorical Responses from the perspective of a student studying linear multivariate regression analysis. The report begins with an overview of categorical responses, including definitions and different types. It proceeds to examine two case studies: predicting introvert/extrovert personalities based on forum posts of a user, and the epidemic outbreak of a disease. The former uses code in Python to achieve the analysis while the later is written in SAS. Both possibilities explore the viability of a logistic regression for solving their respective problem.

## 

## Table of Contents

[**Authors**](#_1c9eepwggbk4) **1**

[**Abstract**](#_x5p3sg3ew2rx) **1**

[**Introduction**](#_ng65fflrqy1m) **2**

[**Background**](#_1opqy99dludp) **2**

[What is a categorical response?](#_wzqadd66y6qt) 2

[Types of Categorical Responses](#_frcs799skena) 2

[Type of Polytomous Categorical Responses:](#_oioudz1wvkx6) 2

[How Do You Predict a Categorical Response?](#_sxofzkubksy1) 3

[Why is a Logistic Regression Needed?](#_lqdf9kbyhyjg) 3

[Assumptions for Logistic Regressions](#_jaiy5bvv5f7f) 3

[**Case Study 1: Using Blog Posts to Predict Personality Type**](#_k0zfvl62wk0r) **4**

[What is MBTI?](#_8qbizhvl1b3f) 4

[Research Question](#_k8z5z1hxy8bq) 5

[Procedure](#_hri4i2gt7kft) 6

[**Results and discussion**](#_i0t236ed5mtw) **7**

[**Summary/conclusions**](#_mc9apz9wbrxs) **8**

[**References**](#_l1ulvbjvr5d) **9**

[MBTI References](#_auikjfuo69a9) 9

[References for R, SAS, & Python](#_gaqf9l15b91) 9

[Background References](#_ubqhlp9x52zp) 10

[Image References](#_zdjecc7qxhv0) 10

## 

## Introduction

In Regression Analysis, (STAT 510), we’ve learned how to model a linear regression with a numeric response variable, but what if the response variable (Y) is not numeric? In this paper, we discuss the different types of non-numeric response variables (categorical responses), a potential solution for analysis, and two case studies dealing with a binary response variable and ordinal response variable.

## Background

#### What is a categorical response?

Response data that is measured by categories instead of continuously.

Also called: Qualitative (vs. Quantitative),

Examples of Qualitative data include color of a sample, texture of a surface, aroma of a reaction

Quantitative examples include mass of a sample, length of a piece of wire, molecules in a mole.

### Types of Categorical Responses

**Binary**: 1 or 0. The response variable is one of two things. Examples include Smoker or Non-smoker, Netflix Thumbs up or Thumbs down, and People who dance in front of the mirror or people who don’t dance in front of the mirror

**Polytomous** (more than 1 outcome)

#### Type of Polytomous Categorical Responses:

**Ordinal**: On an ordered spectrum, Multiple Categories that can be ordered. E

**Interval**: numerical distance between data points

**Nominal categorical responses**: unordered categorical responses (unlike ordinal). Examples include:

Worst Television Show?

iZombie, Santa Clarita Diet, Ironfist

Who is your favorite singer?

Taylor Swift, The Biebs, Beyonce, Shakira, Robert Nakano

Favorite Dessert?  
Tiramisu, Chocolate Souffle, Beef jerky, Froyo!

#### How Do You Predict a Categorical Response?

Different methods are used to predict different types of categorical responses.

A Logistic Regression can be used to model Binary and ordinal categorical responses. If a ordinal response has even intervals, a linear regression can be used. Nominal categorical responses can be modeled using a multinomial logit model

#### Why is a Logistic Regression Needed?

A Logistic Regression is useful for the following reasons:

1. The residuals cannot be normally distributed (as the OLS model assumes), since they can only take on one of several values for each combination of level of the Independent Variables.
2. The OLS model makes nonsensical predictions, since the DV is not continuous - e.g., it may predict that someone does something more than ‘all the time’. Something like 1.3 doesn’t make sense in a binary response.
3. For nominal DVs, the coding is completely arbitrary.

#### Assumptions for Logistic Regressions

**Don’t Need**

Doesn’t need to be linear

Distribution doesn’t need to be normal

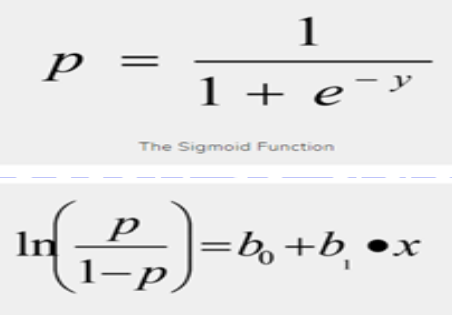
**Do Need:**

Binary or ordinal data

Independent observations (no matched data)

Independent variables need to be linearly related to the log odds

Logistic Regression Formula:



A binary logistic model is used to estimate the probability of a binary response based on one or more predictor (or independent) variables (features). It allows one to say that the presence of a risk factor increases the odds of a given outcome by a specific factor. The model is a direct probability model and not a classifier.

## Case Study 1: Using Blog Posts to Predict Personality Type

#### What is MBTI?

The Myers Briggs Type Indicator (MBTI) is a personality type system that divides everyone into

16 distinct personality types across 4 axis:

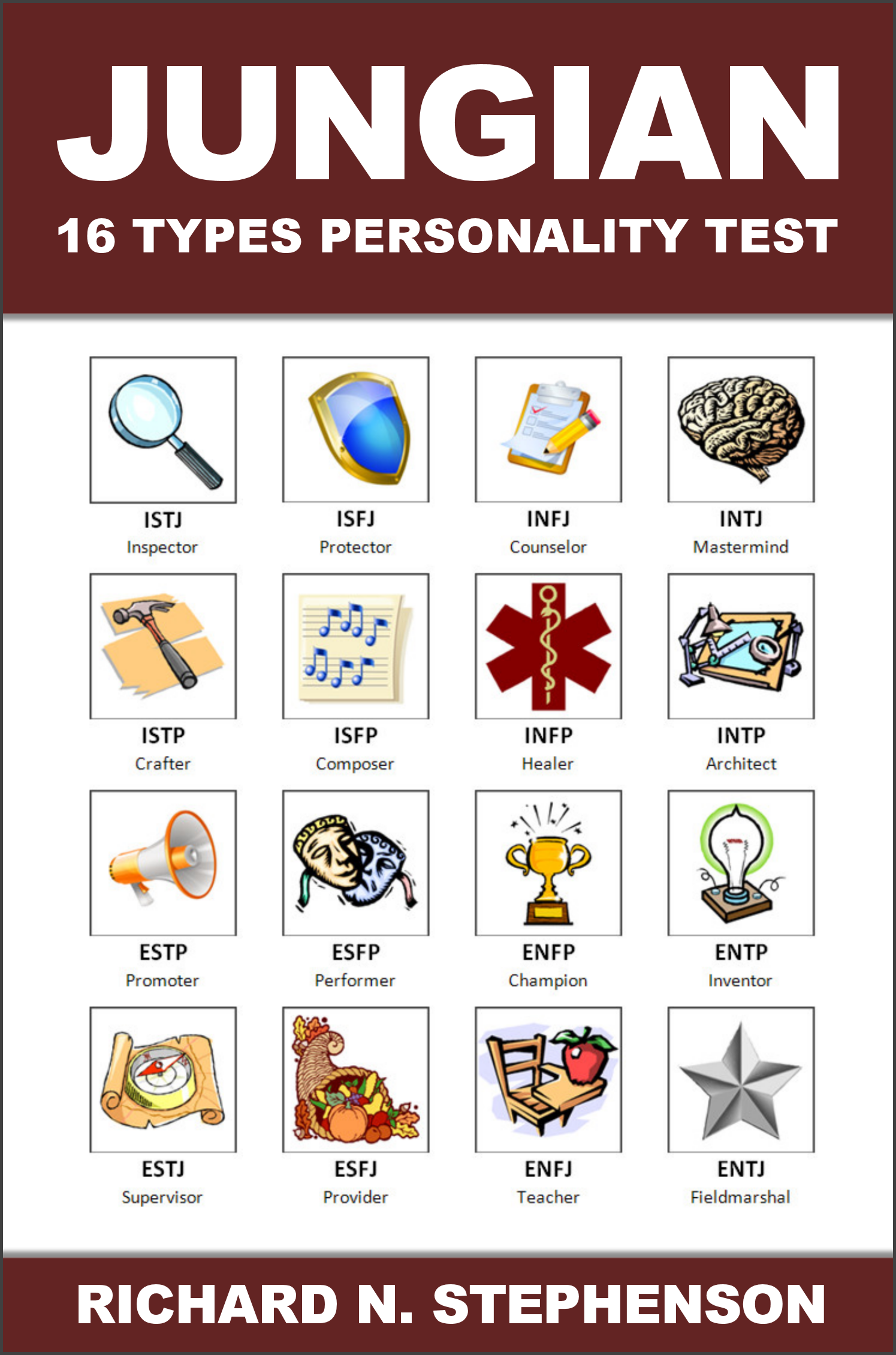
· Introversion (I) – Extroversion (E)

· Intuition (N) – Sensing (S)

· Thinking (T) – Feeling (F)

· Judging (J) – Perceiving (P)

A diagram of the 16 types is shown below.

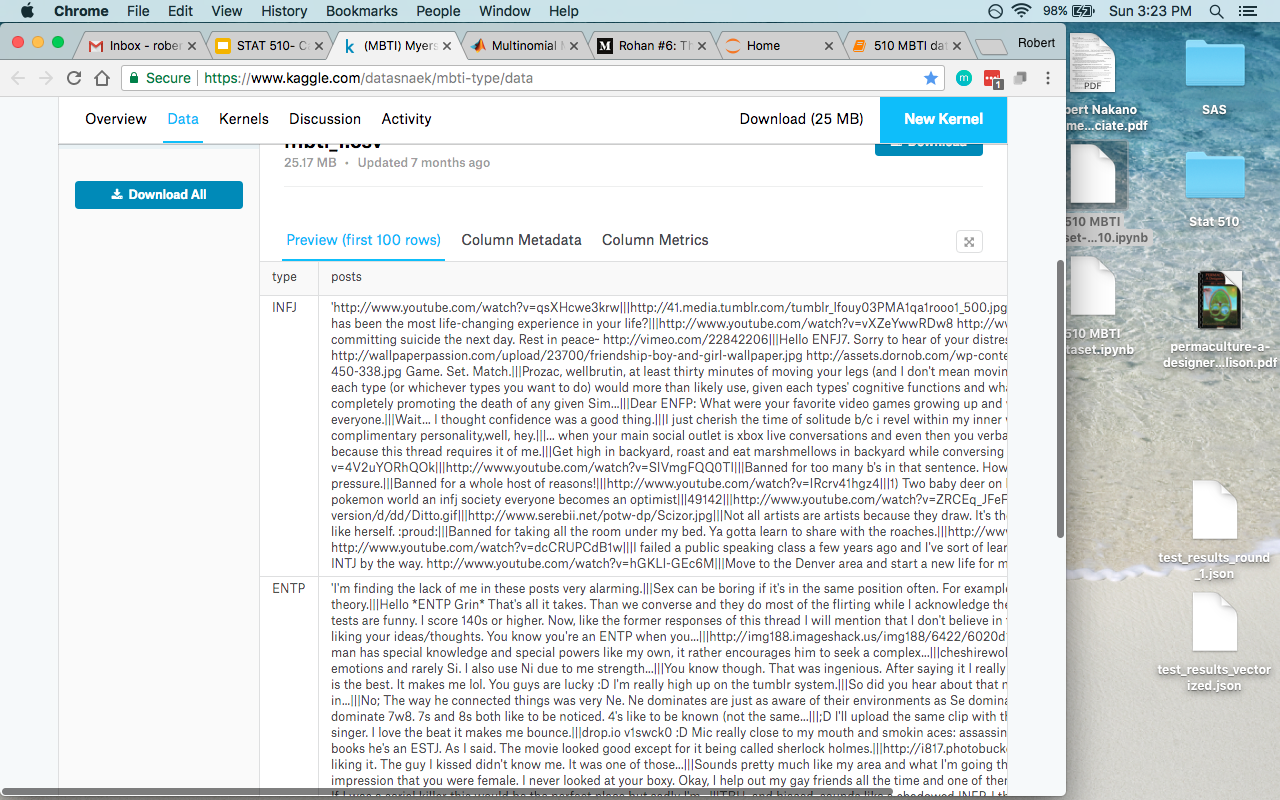


A free self-test to determine your personality type and further information can be obtained from <https://youtopiaproject.com/youtopia-16-assessment/>.

#### Research Question

In this case study, we explore the first axis, introversion vs. extroversion, a binary variable. Do using certain words indicate Introverted/Extrovertedness?

In this case study, we use a data set from kaggle with 8675 blog posts with a matched personality type (<https://www.kaggle.com/datasnaek/mbti-type/data>).



#### Procedure

1. Clean data
2. Create a column for I vs. E (binary)
3. Tokenize/Vectorize dataset (bag of words)
4. Count words & create giant matrix
5. Run a logistic regression
6. Summarize data

##### Python code

Here is the link to the entire python code for the project with output

<https://drive.google.com/open?id=1qSQU-UmlPtrouVIiaAt78ZFeqr-1ExJ6>

A binary column was created to run a binary logistic regression over. Stopwords, words to be removed from the model were removed. The remaining words were sent through a Countvectorizer to create a matrix for the logistic regression to process. Multiple iterations of the Logistic Regression were run, with disappointing univariate statistics. Thus, Recursive Feature Selection with Cross Validation was selected as a way to select a model.

## Case Study 2: Epidemic Disease Outbreak using SAS

#### Research Question

How to predict a categorical response?

Ordered (Ordinal)

To predict a categorical response base on explanatory variables we use Logistic Regression in SAS.

The following example illustrates the fitting and interpretation of a multiple logistic regression model that involves both quantitative and categorical predictor variables.

In a health study to investigate an epidemic outbreak of a disease that is spread by mosquitoes, individuals were randomly sampled within two sectors in a city to determine if the person had certain specific symptoms associated with the disease. If the person was determined to have contracted the disease, the response variable Y = 1, and 0 if not. In this study, the following three predictor variables were included:

· Age (X1): a quantitative variable;

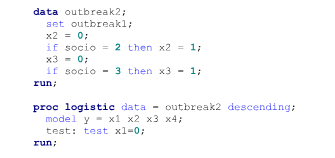
· Socioeconomic status of household. This is a categorical variable with three levels. It is coded by two indicator variables (X2 and X3) as follows:

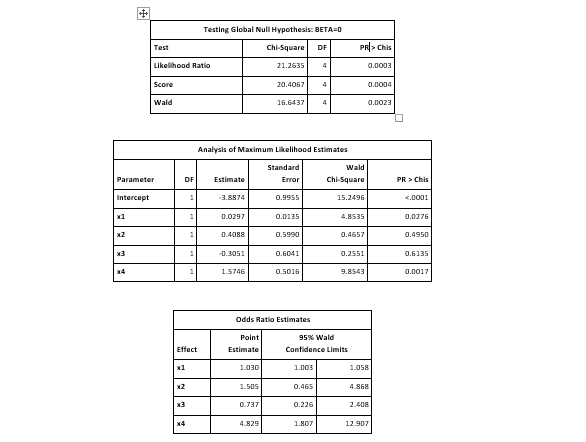
|  |  |  |
| --- | --- | --- |
| Class | X2 | X3 |
| Upper | 0 | 0 |
| Middle | 1 | 0 |
| Lower | 0 | 1 |

The reason why the upper socioeconomic class was chosen as the reference class (X2 = 0 and X3 = 0) is that it was expected that this class would have the lowest disease rate among the socioeconomic classes.

· City sector (X4): Since there were only two sectors in the study, one indicator variable was used, defined as X4 = 0 for sector 1 and X4 = 1 for sector 2.

The primary purpose of the study was to assess the strength of the association between each of the predictor variables and the probability of a person having contracted the disease.





#### Procedure

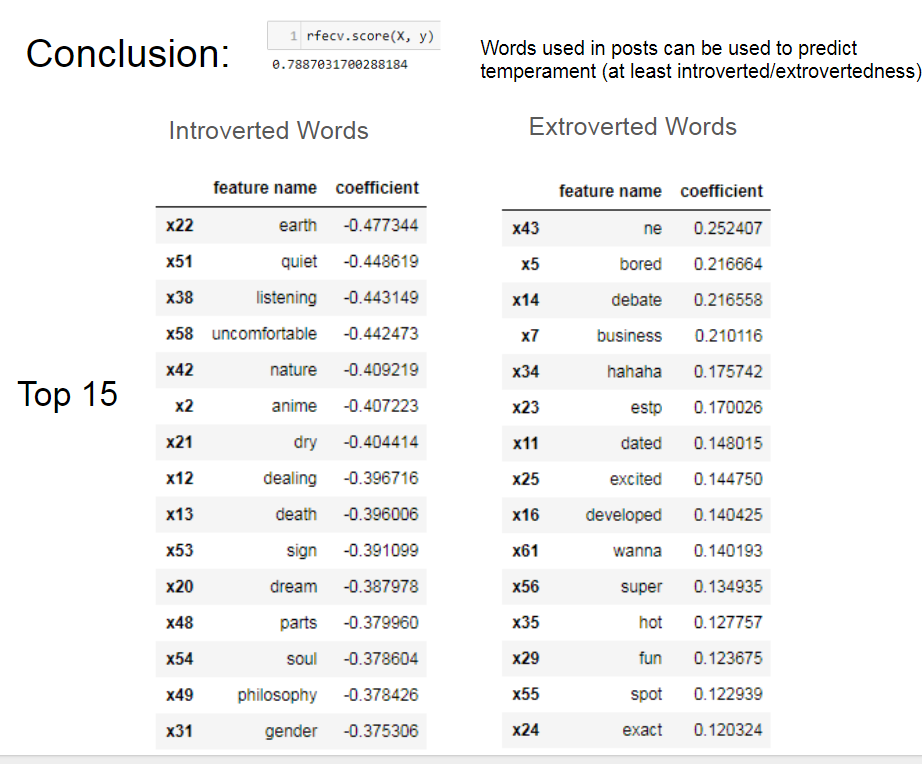
In this case we used Logistic Regression. Base on SAS outcome we can interpret the result in terms of odds ratios. Point Estimated in logistic regression is log(odds ratios). The explanatory variables with larger odds ratio increase the probability of having the disease more than other variables. In “Linear Hypothesis Testing Results “ table Chi-sq Proc Logistic was performed with the “test” option to test influence of X1(age) on having contracted the disease.

## Results and discussion

By looking at “Odds Ratio Estimates” table, in column Point Estimate Odds Ratios of X1(age) and X3( household status ) with significant P-Value the probability of having the disease is increasing, since their Point Estimates were getting closer to or greater than 1. Therefore with the result we expect more probability of having contracted to the disease among people who are classified as “lower” in socioeconomic class.

#### Do using certain words indicate a certain temperament?

Yes. We were able to predict, with 78.9% accuracy a person’s introvertedness or extrovertedness by 64 words used. It is was surprising to see words such as earth, anime, and parts with strong introverted coefficients, while words such as wanna and spot were surprising to see to indicate extrovertedness. Ultimately, a binary logistic regression seems to yield meaningful results.



Cross validation seemed to be an appropriate method to eliminate features, as words such as “quiet” and “listening” seem like very likely words to be included as an introverted words. Additionally, some words should have been removed from this list (ne, estp), as they are metacognitive of personality theory and lack interesting implications.

## Summary/conclusions

A summary and explanation of the significance of your findings and a set of recommendations for future work

Thus, logistic regressions are one tool that can sometimes provide meaningful results in analyzing categorical responses. For our case study one, next steps would include:

1. More data: the dataset was skewed, largely towards Ns and introverts.
2. Better computer: Only 1000 of the most frequent words were used in the analysis do to limited processing power. Running this model with all 177,000 words could be quite interesting.
3. Compare performance of other models/classifiers: It would be interesting to compare the performance of a logistic regression with other models and classifiers.
4. Try using ngrams: ngrams could provide a way to add more meaning significance.
5. Look at each other axis N/S, F/T, J/P: this could expand the interpretation of the model.
6. Review stop words list: carefully reviewing which words are MBTI related and removing them from the list could provide more meaningful predictive word results.

## References

Link to slides of presentation

<https://tinyurl.com/y93zm5bv>

Python Code used for MBTI analysis

<https://drive.google.com/open?id=1qSQU-UmlPtrouVIiaAt78ZFeqr-1ExJ6>

### MBTI References

MBTI Dataset

<https://www.kaggle.com/datasnaek/mbti-type/data>

A MBTI test and information on different personality types

<https://youtopiaproject.com/youtopia-16-assessment/>

CASE STUDY 2 References

Lillis, D. n.d.Generalized Linear Models in R, Part 1: Calculating Predicted

### References for R, SAS, & Python

<https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/>

R tutorial on running a logistic regression:

<https://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/>

R tutorial on building a logistic regression from scratch:

<https://www.analyticsvidhya.com/blog/2015/10/basics-logistic-regression/>

Python walkthrough: <https://towardsdatascience.com/logistic-regression-using-python-sklearn-numpy-mnist-handwriting-recognition-matplotlib-a6b31e2b166a>

<https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8>

SAS example:

<http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_logistic_sect060.htm>

Recursive Feature Selection with Cross Validation example in Python

<http://scikit-learn.org/stable/auto_examples/feature_selection/plot_rfe_with_cross_validation.html#sphx-glr-auto-examples-feature-selection-plot-rfe-with-cross-validation-py>

Changing order of dataframes

<https://stackoverflow.com/questions/13148429/how-to-change-the-order-of-dataframe-columns>

<http://thestatsgeek.com/2014/02/08/r-squared-in-logistic-regression/>

For multinomial models (more than 1 possible response)

<https://www.mathworks.com/help/stats/multinomial-models-for-nominal-responses.html>

<http://amunategui.github.io/multinomial-neuralnetworks-walkthrough/>

<https://www.theanalysisfactor.com/logistic-regression-models-for-multinomial-and-ordinal-variables/>

R walkthrough

### Background References

General information on categorical data

<http://www.ucd.ie/statdept/classpages/categorical_data_analysis/cda1.pdf>

Ordinal Responses

<https://en.wikipedia.org/wiki/Ordinal_regression>

<https://en.wikipedia.org/wiki/Ordered_logit>

Example of ordered regression problems:

<https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/>

In SPSS

<https://www.ibm.com/support/knowledgecenter/en/SSLVMB_22.0.0/com.ibm.spss.statistics.help/spss/categories/idh_catr.htm>

### Image References

<https://www.twenty20.com/photos/1825e01e-3a43-43a8-929e-e0daa7d09cd1>