Johns Hopkins University

625.725

Theory of Statistics I

Full Methods for Hypothesis Testing Project

Author:
Jack Moody

Professors:
Dr. Thomas Woolf
Dr. Burhan Sadiq

August 9, 2022

Contents

List of Tables

List of Figures

Executive Summary

This project utilizes the 2015-2016 NHANES data set to conduct a series of experiments in hypothesis testing, partial least squares (PLS) based model selection, and multiple K-fold cross-validations for model selection. **Keywords:** *Hypothesis testing, PLS, NHANES*

1 Project Description

Present the dataset used and the associated analyses performed. Do so in such a way that your analyses can be replicated. Highlight the connection to your entry paper from the topic choices

The data set chosen was the 2015-2016 National Health and Nutrition Examination Survey (NHANES) survey. It is a research program conducted by the National Center for Health Statistics (NCHS) to assess the health and nutritional status of adults and children in the United States, and to track changes over time.

I chose this particular data set because I have never worked with health-related data before and felt it would be a good way to try something new. I decided to use this data set to perform a variety of hypothesis tests and method selection algorithms, specifically PLS and K-fold cross-validation.

Normal hypothesis testing was chosen because it is a fundamental part of statistical analysis. Since this is course spends a decent amount of time on it, I felt it was necessary to include it.

PLS was chosen due to it being the subject of one of the selected papers associated with hypothesis testing. This paper and the PLS criteria involved are important to understand because the authors try to address how real-world data sets are often more complicated than the baseline models created in more abstract settings.

K-fold cross-validation (CV) has also been chosen because it is a common method for model selection. However, the specific paper included introduces a slightly more nuanced perspective on K-fold CV by introducing K - 1 folds of the data for model validation, while the other fold is for model construction. This provides K - 1 predicted values for each observation. These values are then averaged to produce a final predicted value. Then, the model selection based on the averaged predicted values can reduce variation in the assessment due to the averaging.

2 Methodology

Describe how your results were determined and the mathematical and statistical methods that you used.

2.1 Exploratory Data Analysis

I have completed all of the data engineering for the dataset. I have decided to focus solely on the ('SMQ020', 'RIAGENDR', 'RIDAGEYR','DMDEDUC2','BMXWT', 'BMXHT', 'BMXBMI') columns of the data set, which I have renamed to ('smoking','gender', 'age','education','weight', 'height','bmi') because that is what the column names correlate to, in order to make it more human readable.

Further exploratory data analysis was done to get rid of outliers and ensure normal distribution of the data set. Below shows the results post outlier removal:

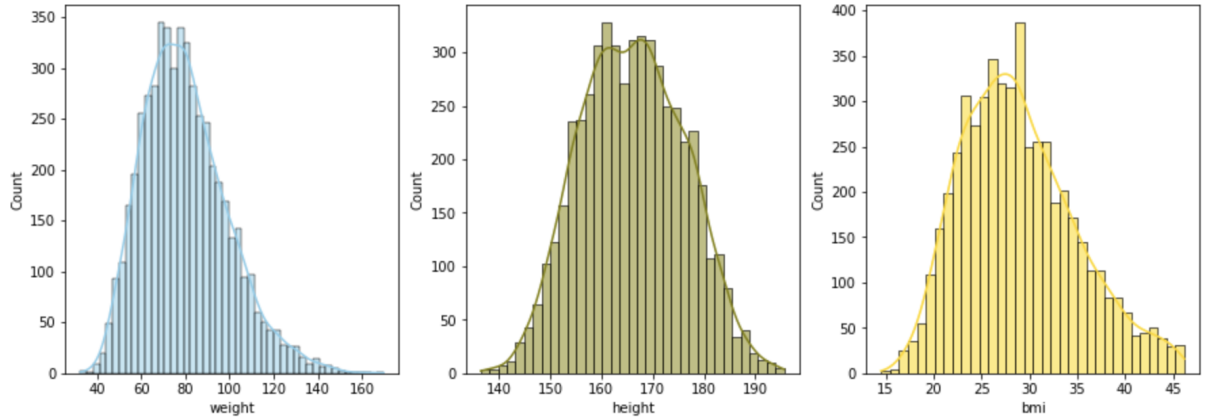

Figure 1: Histograms of Weight, Height, and BMI post outlier removal

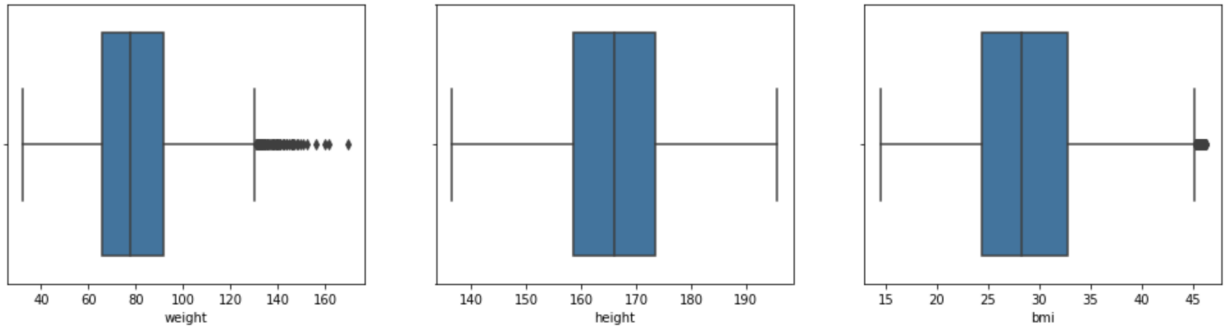

Figure 2: Box and Whisker Plots of Weight, Height, and BMI post outlier removal

2.2 Hypothesis Testing

Will be using chapter 8 of Casella and Berger to go over a proof of this then the code to test it out

2.3 Partial Least Squares Based Model Selection

Will be using the provided paper to go over the proofs and then demonstrating how I used the NHANES data set to test this method

2.4 Multiple predicting K -fold cross-validation for model selection

Will be using the provided paper to go over the proofs and then demonstrating how I used the NHANES data set to test this method

3 Results and Discussion

Provide textural and graphical results of your analyses along with interpretation of those results.

3.1 Summary Statistics

The mean and standard deviation of the weight, height, and BMI between men and women in age demographics of (18,30], (30,40], (40,50], (50,60], (60,70], and (70,80] is shown in the table below. These summary statistics will provide us with a baseline for the work discussed later on in this paper.

		weight		height		bmi	
		mean	std	mean	std	mean	std
age	gender						
(18, 30]	female	71.971295	18.605828	161.450094	6.867726	27.558161	6.624561
	male	83.320879	19.261460	174.909890	7.701208	27.183297	5.771789
(30, 40]	female	76.091111	19.726570	160.744444	7.207113	29.344889	6.871599
	male	88.335147	19.706370	174.084354	7.755265	29.067347	5.684898
(40, 50]	female	77.595168	18.692238	160.225630	7.163041	30.174370	6.717175
	male	88.724548	19.747783	173.806202	7.563134	29.247028	5.654652
(50, 60]	female	75.917007	16.789526	159.985261	6.917037	29.602041	5.933152
	male	86.674266	18.809328	172.935892	8.383587	28.866591	5.282098
(60, 70]	female	75.776485	17.716172	158.029929	6.982854	30.254157	6.356145
	male	86.022196	18.078713	171.831742	7.290485	29.017422	5.208027
(70, 80]	female	70.566247	15.371260	156.253401	6.688228	28.836272	5.698754
	male	81.334704	16.850614	170.379177	7.485395	27.899743	4.847672

Figure 3: Mean and Standard Deviation between men and women of various ages

Figure 3 shows summary statistics between age and gender of our data set.

3.2 Partial Least Squares Based Model Selection

I have not done the code for this yet. But I will be using python libraries to assist in this.

3.3 Multiple predicting K -fold cross-validation for model selection

I have not done the code for this yet. But I will be using python libraries to assist in this.

4 Conclusion

Discuss the results in terms of suitability of the chosen analysis methods for the data, difficulties encountered, and suggestions for alternative types of analyses.

None yet

References

1. Cassella, G. and Berger, R. (2002). Statistical Inference. (2nd edition) Wadsworth, Inc., Belmont, CA. Chapter 8
2. Wasserman, L. (2010). All of Statistics. Springer Nature Switzerland., Chapter 10
3. Sharma, Pratyush, Marko Sarstedt, Galit Shmueli, Kevin H. Kim, and Kai Oliver Thiele. "PLS-based model selection: The role of alternative explanations in information systems research." Journal of the Association for Information Systems20, no. 4 (2019): 4.
4. Jung, Yoonsuh. "Multiple predicting K-fold cross-validation for model selection." Journal of Nonparametric Statistics 30, no. 1 (2018): 197-215.
5. Sharma, Pratyush, and Kevin H. Kim, "Model Selection in Information Systems Research Using Partial Least Squares Based Structural Equation Modeling." Thirty Third International Conference on Information Systems, Orlando 2012
6. Partial Least Squares (PLS) Regression by Herv´e Abdi
7. Demystifying hypothesis testing with simple Python examples by Tirthajyoti Sarkar
8. 17 Statistical Hypothesis Tests in Python (Cheat Sheet)
9. A short introduction to model selection by David Schönleber
10. The Complete Guide to R-squared, Adjusted R-squared and Pseudo-R-squared by Sachin Date
11. k fold cross validation in python

APPENDIX This will likely be a link to my github repo for this project