Johns Hopkins University

625.725

Theory of Statistics I

Hypothesis Testing Final Report

Author:
Jack Moody

Professors:
Dr. Thomas Woolf
Dr. Burhan Sadiq

August 22, 2022

Contents

List of Figures

Executive Summary

The purpose of this paper is 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. The dataset utilized is a refined dataframe comprised of data from the 2015-2016 NHANES dataset. All of the data engineering and analysis was written in Python and the GitHub repository is publicly available for review. The main findings involve p-values from the various hypothesis tests, including a One Population Proportion Test, Two Population Proportion Test, Comparing Means (and their standard error), Paired Tests, Goodness-of-Fit Test, and a Log-Likelihood Test. Beyond that, methods and results are discussed for implementations of Partial Least Squares and their ability to accurately or inaccurately predict the RMSE (the average deviation between the predicted value of a predictor and the observed value) of a response variable given a set of predictor variables. Finally this paper discusses K-fold cross validation techniques and their ability to predict the RMSE of a response variable given a set of predictor variables similarly to PLS, however a linear regression model is used instead of PLS regression.
Keywords: *Hypothesis testing, PLS, NHANES, K-fold cross-validation, Python*

1 Project Description

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 [2].

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-PM and K-fold cross-validation.

Hypothesis testing was chosen because it is a fundamental part of statistical analysis. Specifically, a One Population Proportion Test, Two Population Proportion Test, Comparing Means(and their standard error) Test, Paired Tests, Goodness-of-Fit Test, and a Log-Likelihood Test were conducted.

Partial Least Squares Path Modeling (PLS-PM) was chosen due to it being the subject of one of the selected papers associated with hypothesis testing [5]. 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. I attempted to use a python library called PLSPM, which was adapted from an R library of the same name. However, after many days of struggling to implement it effectively, I resorted to regular PLS implementation for this paper. This was done using Scikit-learn's model_selection functions called RepeatedKFold, train_test_split,cross_decomposition, PLSRegression and my implementation can be seen on my GitHub for this project.

K-fold cross-validation (CV) has also been chosen because it is a common method for model selection and the main focus of one of the provided papers [4]. This K -fold cross-validation method included fitting a multiple linear regression model to the NHANES dataset and then performing a LOOCV to evaluate the model prediction. This was done using the sklearn.model_selection library in python via the sklearn module.

2 Methodology

2.1 Exploratory Data Analysis

For this project, I have decided to focus solely on the ('SMQ020', 'RIAGENDR', 'RIDAGEYR', 'DMDEDUC2','BMXWT', 'BMXHT', 'BMXBMI', 'BPXSY1', 'BPXSY2') columns of the data set, which I have renamed to ('smoking','gender', 'age','education','weight', 'height','bmi', bp1, bp2) because that is what the column names correlate to [2], in order to make it more human-readable (bp in this context means blood pressure).

Further exploratory data analysis was done to get rid of outliers and ensure the normal distribution of the data set. Figure 1 shows the raw histograms prior to data adjusts and so does Figure 2 with the box plots.

After the data cleaning of the data set, every column of data resembles a normal distribution. It is important to note that age is not in these figures because it was already a normal distribution, so did not need any adjusting. However, the other columns were either left or right-skewed, so they needed to be adjusted, as seen in Figure 3 and Figure 4. If there is more curiosity about how this was done, please see the GitHub page associated with this project.

2.2 Hypothesis Testing

Using [7] Chapter 10, we can define hypothesis testing. Let's consider the following two hypotheses for an example where we have two groups of rats randomly divided amongst a group that was exposed to asbestos, and the other group is not:

The Null Hypothesis: The disease rate is the same in the two groups.
The Alternative Hypothesis: The disease rate is not the same in the two groups.

If the exposed group has a much higher rate of disease than the unexposed group then we will reject the null hypothesis and conclude that the evidence favors the alternative hypothesis. This is an example of hypothesis testing.
More formally, suppose that we partition the parameter space Θ into two disjoint sets Θ_0 and Θ_1 and that we wish to test

$$H_0 : \theta \in \Theta_0 \text{ versus } H_1 : \theta \in \Theta_1.$$

We call H_0 the null hypothesis and H_1 the alternative hypothesis. The basic formula for this (also known as the Wald test) can be shown as:

$$W = \frac{Best\ Estimate - Hypothesized\ Estimate}{Standard\ Error\ of\ Estimate} = \frac{\widehat{\theta} - \theta_0}{\widehat{\text{se}}}$$

Where $H_0 : \theta = \theta_0 \quad$ and $\quad H_1 : \theta \neq \theta_0$.

There were 6 types of hypothesis tests conducted on this data set:

1. **One Population Proportion**:
   This is the most basic hypothesis test and is rarely used in practice because it is rare that there is a specific fixed value to use for comparison.

   The test involves having a specified particular value as the null H_0 for the proportion, and we wish to asses if the data are compatible with this true parameter being equal to the specified H_0

2. **Two Population Proportion**:
   Comparative tests are used more frequently than tests comparing one population to a fixed value. A two-sample test of proportions is used to assess whether the proportion of individuals with some trait differs between two sub-populations.

3. **Comparing Means (and their standard error)**:
   Example 10.8 of [7] has a great theoretical intro to this test. But, allow me to attempt to describe it. Tests of means have many similarities to tests of proportions. Just as with proportions, for comparing means there are one and two-sample tests, z-tests and t-tests, and one-sided and two-sided tests

4. **Paired Tests**:
   A paired test is a modified form of a mean test that can be used when we are comparing two repeated measurements on the same unit.

5. **Goodness-of-Fit Tests**:
   A Goodness-of-Fit test gives a solution to validate our theoretical assumptions about data distributions.

6. **Log-Likelihood Test**:
   A likelihood ratio test compares the goodness of fit of two nested regression models. A nested model is simply one that contains a subset of the predictor variables in the overall regression model. This has a nice definition in [7] Chapter 10 Definition 10.21: Consider testing

   $$H_0 : \theta \in \Theta_0 \text{ versus } H_1 : \theta \notin \Theta_0.$$

   The likelihood ratio statistic is

   $$\lambda = 2 \log \left(\frac{\sup_{\theta \in \Theta} \mathcal{L}(\theta)}{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta)} \right) = 2 \log \left(\frac{\mathcal{L}(\widehat{\theta})}{\mathcal{L}\left(\widehat{\theta}_0\right)} \right)$$

   where $\widehat{\theta}$ is the MLE and $\widehat{\theta}_0$ is the MLE when θ is restricted to lie in Θ_0.

2.3 Partial Least Squares Based Model Selection

The partial least squares path modeling or partial least squares structural equation modeling (PLS-PM, PLS-SEM) is a method for structural equation modeling that allows estimation of complex cause-effect relationships in path models with latent variables as described in [4]. However, I was unable to actually get this to work in my analysis, so I just used basic PLS in order to still work towards the goal of PLS-PM. I will now describe how PLS works:

1. Standardize both the predictor and response variables.

2. Calculate the M number of linear combinations from the original predictor variables that explain a significant amount of variation in both the response variable and the predictor variables.

3. Using the method of least squares, fit this to a linear regression model using the PLS components as the predictors.

4. Use k-fold cross validation to find the optimal number of PLS component to keep in the model.

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

K-fold cross-validation uses the following approach to evaluate a model:

1. Randomly divide the dataset into k groups, or "folds", of roughly equal size.

2. Choose one of the folds to be the holdout or "truth" set. Fit the model on the remaining k-1 folds. Calculate the test MSE (Mean Square Error) on the observations in the fold that was held out.

3. Calculate the overall test MSE to be the average of the k test MSE's.

A more mathematical definition of MSE can be found on page 91 of [7]:
The MSE can be written as

$$\text{MSE} = \text{bias}^2\left(\widehat{\theta}_n\right) + \mathbb{V}_\theta\left(\widehat{\theta}_n\right)$$

However, for our purposes of using it computationally, a more gentle formula would be:

$$\text{MSE } = \frac{1}{n}\sum(y_i - f(x_i))^2$$

where n is the total number of observations, y_i is the response value for the ith observation, and $f(x_i)$ is the predicted response value of the ith observation.

It is important to note, that the closer the predictions are to the observations, the smaller the MSE will be. This analysis was based in the findings of [5].

3 Results

3.1 Hypothesis Testing

1. **One Population Proportion**:
   The test case chosen for this test is a situation where the rate of lifetime smoking in another country was known to be 40%, and we wished to assess whether the rate of lifetime smoking in the US is different from 40%. We carry out the (two-sided) one-sample test that the population proportion of smokers is 0.4, and obtain a test statistic of 2.86 and a p-value of 0.004. This indicates that the NHANES data are compatible with the proportion of (ever) smokers in the US being 40%.

   This test was carried out in two ways, the first was finding the test statistic by hand, and the second was using the python library stasmodels. Both methods yielded the same result, which helps show that the method conducted by hand was correct.

2. **Two Population Proportion**:
   Comparative tests are used more frequently than tests comparing one population to a fixed value. A two-sample test of proportions is used to assess whether the proportion of individuals with some trait differs between two sub-populations. For example, I chose to compare the smoking rates between females and males. Since smoking rates vary strongly with age, we do this in the subpopulation of people between 18 and 25 years of age.

   This was calculated in two ways, one by hand and one using statsmodels. We find that men are about 8% more likely to smoke than women between the ages of 18 to 25 (given that 26% of women between 18-25 smoke and 33% of men between 18-25 smoke according to the sample population in the NHANES dataset), and it's statistically significant because the p-value is around 0.005.

3. **Comparing Means (and their standard error)**:
   Just as with proportions, for comparing means there are one and two-sample tests, z-tests and t-tests, and one-sided and two-sided tests.

   Similar to the tests of proportions, one-sample tests of means are not very common, but we will do an example of it anyway for a better understanding of our data. We compare systolic blood pressure to the fixed value of 120 (which is the lower threshold for "pre-hypertension") and find that the mean is significantly different from 120 (the point estimate of the mean is 124.6 so we will say 125).

   A formal test of the H_0 was also run to see if the mean blood pressure of women between 60 - 70 is equal to the mean blood pressure of men of the same age range. The results show that the mean systolic blood pressure is slightly higher in men than in women (131 mm/Hg vs 130 mm/Hg), but that the difference is not statistically significant (the p-value is 0.16).

   While in this part of hypothesis testing, it is important to discuss the standard error. How can we estimate the standard error of the mean difference? Well, it can be done in a few ways. One way is by using the statsmodels library, which uses the "pooled" and "unequal" approaches of estimating the variance. If the variances are equal, then there

should be little difference between the two approaches. However, even with a moderate difference, the results for the two methods are typically similar.

I tested the 10-year age band between the BMI data and allow the statsmodel library to assess the evidence for any difference in BMI for men and women. The results are then printed per 10-year bands as shown in Figure 5

4. **Paired Tests**:
   This test took advantage of how our dataset has two measurements of a subject's blood pressure. Even though the measurements are repeated, there is no guarantee that the mean is the same each time, i.e. the mean blood pressure may be slightly lower on the second measurement compared to the first. A paired test is a modified form of a mean test that can be used when we are comparing two repeated measurements on the same unit.

   A paired t-test for means is equivalent to taking the difference between the first and second measurements and using a one-sample test to compare the mean of these differences to zero. I did a paired test of the entire NHANES dataset, the first measurement of systolic blood pressure is on average 0.68 mm/Hg greater than the second measurement. While this difference is not large, it is strongly statistically significant. That is, there is strong evidence that the mean values for the first and second blood pressure measurements differ.

5. **Goodness-of-Fit Tests**:
   One of the traditional statistical approaches, the Goodness-of-Fit test gives a solution to validate our theoretical assumptions about data distributions.

   For this test, I wanted to verify that our height variable follows a normal distribution. We find that it is actually not a normal distribution! Despite our best efforts to make it one earlier. This is shown because our chi-square value is 11.1, which is far too high to allow us to accept the normal distribution as this variable's actual distribution.

6. **Log-Likelihood Test**:
   A likelihood ratio test compares the goodness of fit of two nested regression models. A nested model is simply one that contains a subset of the predictor variables in the overall regression model

   Here we try to see which of the two models defined below could better predict a person's weight. The first model includes the age, BMI, bp1, and bp2 of the subject, while the nested model only includes the age and BMI. After conducting the analysis, the Chi-Squared test statistic is 39.3 and our p-value is $2.9 * 10^{-}9$, so we will reject the null hypothesis and this shows that the model includes age, BMI, bp1, and bp2 offered a significant fit in predicting a subject's weight.

3.2 Partial Least Squares Based Model Selection

Since I was unable to produce results for PLS-PM, I will go over my results for PLS. I conducted three total tests:

1. For the first test, I set the predictor variables to include ["bmi", "weight", "bp1", "bp2"], and the response variable to be age. I then did a 10 split repeated kfold for the cross validation. From there I calculated the MSE using that kfold algorithm and plotted that against the number of PLS components, resulting in Figure 6. Finally, I calculated the RMSE, which came out to 15.5, which is the average deviation between the predicted value of age and the observed value for age for the observations made in the NHANES data set. This is fairly large, and perhaps a different regression model would have fit better.

2. For the second test, I set the predictor variables to include ["age", "bmi", "bp1", "bp2"], and the response variable to be weight. I then did a 10 split repeated kfold for the cross validation. From there I calculated the MSE using that kfold algorithm and plotted that against the number of PLS components, resulting in Figure 7. Finally, I calculated the RMSE, which came out to 10.01, which is the average deviation between the predicted value of weight and the observed value for weight for the observations made in the NHANES data set. This is fairly large, and perhaps a different regression model would have fit better.

3. For the third test, I set the predictor variables to include ["age", "weight", "bp1", "bp2"], and the response variable to be BMI. I then did a 10 split repeated kfold for the cross validation. From there I calculated the MSE using that kfold algorithm and plotted that against the number of PLS components, resulting in Figure 8. Finally, I calculated the RMSE, which came out to 3.06, which is the average deviation between the predicted value of weight and the observed value for weight for the observations made in the NHANES data set. This is a pretty accurate finding, which means this is a good fit for the model.

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

As a review from the project description section, this K -fold cross-validation method included fitting a multiple linear regression model to the NHANES dataset and then performing a LOOCV to evaluate the model prediction.

Overall, three different tests were conducted:

1. I first wanted to see if age and weight were a good predictor of bmi. I took the predictor variables as ['age', 'weight'] and the response variable as BMI. I then used a 10 split k-fold and used a linear regression model to be used as the evaluation for the k-fold cv. After the analysis, age and weight are good predictors of bmi because the RMSE (root mean square error) is 1.57.

2. I then wanted to see if age and BMI would be a good predictor of blood pressure. I took the predictor variables as ['age', 'bmi'] and the response variable as blood pressure. I then used a 10 split k-fold and used a linear regression model to be used as the evaluation for the k-fold cv. After the analysis, age and bmi are not good predictors of blood pressure because the RMSE is 3.34.

3. Finally, I wanted to see if either age or BMI along with blood pressure 1 would be a better predictor of blood pressure 2. I took the predictor variables as ['bp1', 'bmi'] and

['age','bp1'] then and the response variable as blood pressure 2. I then used a 10 split k-fold and used a linear regression model to be used as the evaluation for the k-fold cv. After the analysis,both age and bmi are solid predictors along with bp1 to predict bp2 because they both come out at the same RMSE of 1.94, which is still fairly good.

4 Conclusion

Overall, I learned a lot during this project and I'm happy I did it. Working with NHANES was a fun and interesting challenge, given it's high dimensionality and variety of both numerical and categorical data types. I was able to properly implement all my desired tasks except for PLS-PM, which was more due to a technical error than anything else. Given more time, I would have tried to implement this by hand.

I believe that all of the analysis methods I chose are fairly suitable given this type of data. However, for the PLS, I would not use PLSRegression again because it did not fit the data well, I would try PLSCanonical because it also includes a transformer which might help reduce the dimensionality that we see in this dataset, allowing for more accurate results.

There were not many difficulties encountered with this project or the dataset outside of what was mentioned previously. The NHANES dataset, despite being very large, it was relatively easy to perform the basic data engineering needed to get the data set into a usable data frame. Given more time, I would try to reproduce this work in R, since it is a language I am less familiar with.

5 Appendix

Please go to my GitHub page (`https://github.com/jackdmoody0427/625.725_Project.git`) for the code and dataset used for this analysis.

References

[1] Hervé Abdi. "Partial least square regression (PLS regression)". In: *Encyclopedia for research methods for the social sciences* 6.4 (2003), pp. 792–795.

[2] *About the National Health and Nutrition Examination Survey*. 2017. URL: `https://www.cdc.gov/nchs/nhanes/about_nhanes.htm` (visited on 08/22/2022).

[3] George Casella and Roger L Berger. *Statistical inference*. Cengage Learning, 2021.

[4] Yoonsuh Jung. "Multiple predicting K-fold cross-validation for model selection". In: *Journal of Nonparametric Statistics* 30.1 (2018), pp. 197–215.

[5] Pratyush Sharma et al. "PLS-based model selection: The role of alternative explanations in information systems research". In: *Journal of the Association for Information Systems* 20.4 (2019), p. 4.

[6] Pratyush Nidhi Sharma and Kevin Hyunkyung Kim. "Model selection in information systems research using partial least squares based structural equation modeling". In: (2012).

[7] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Vol. 26. Springer, 2004.

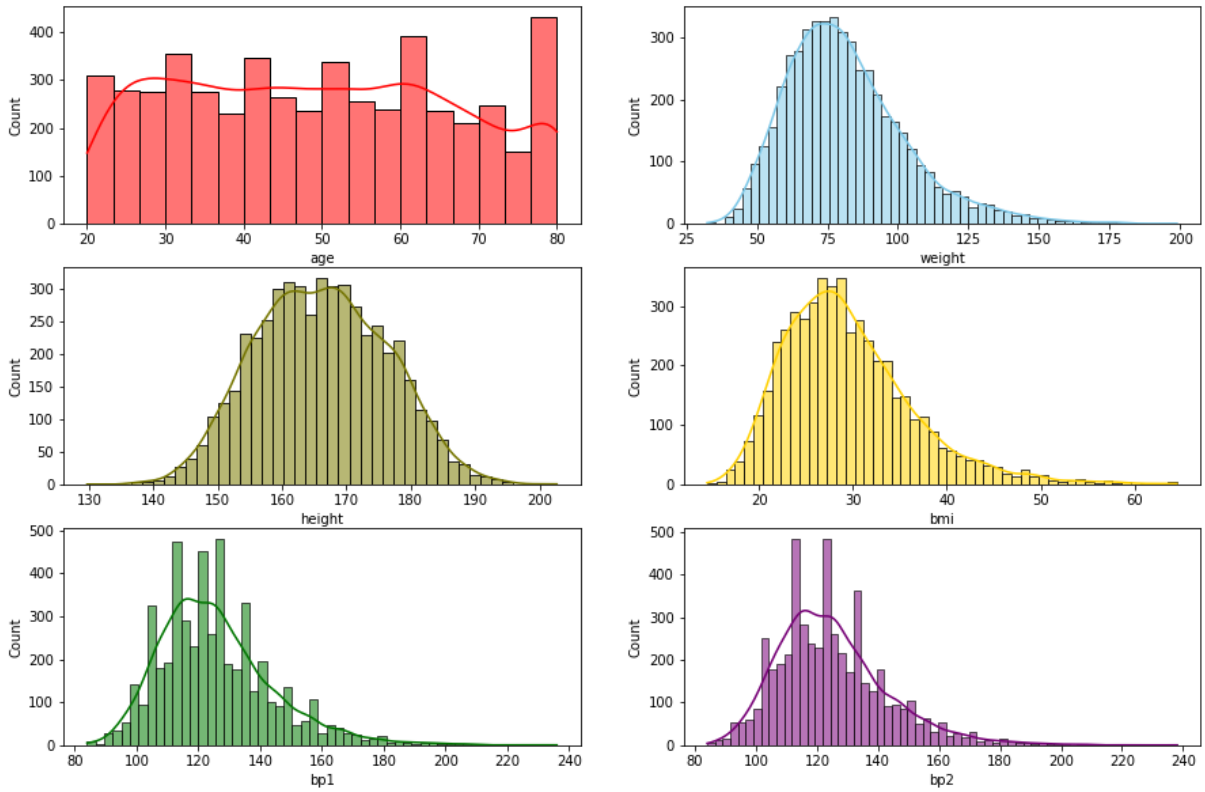

Figure 1: Histograms of Age, Weight, Height, BMI, BP1, and BP2 prior to data cleaning

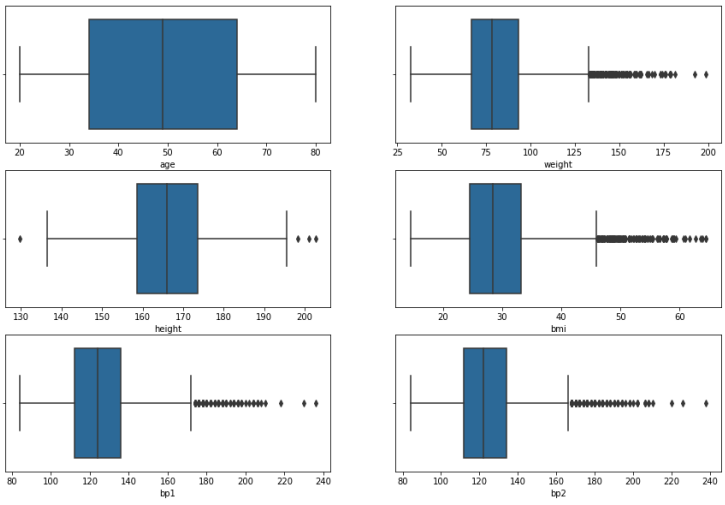

Figure 2: Box and Whisker Plots of Age, Weight, Height, BMI, BP1, and BP2 prior to data cleaning

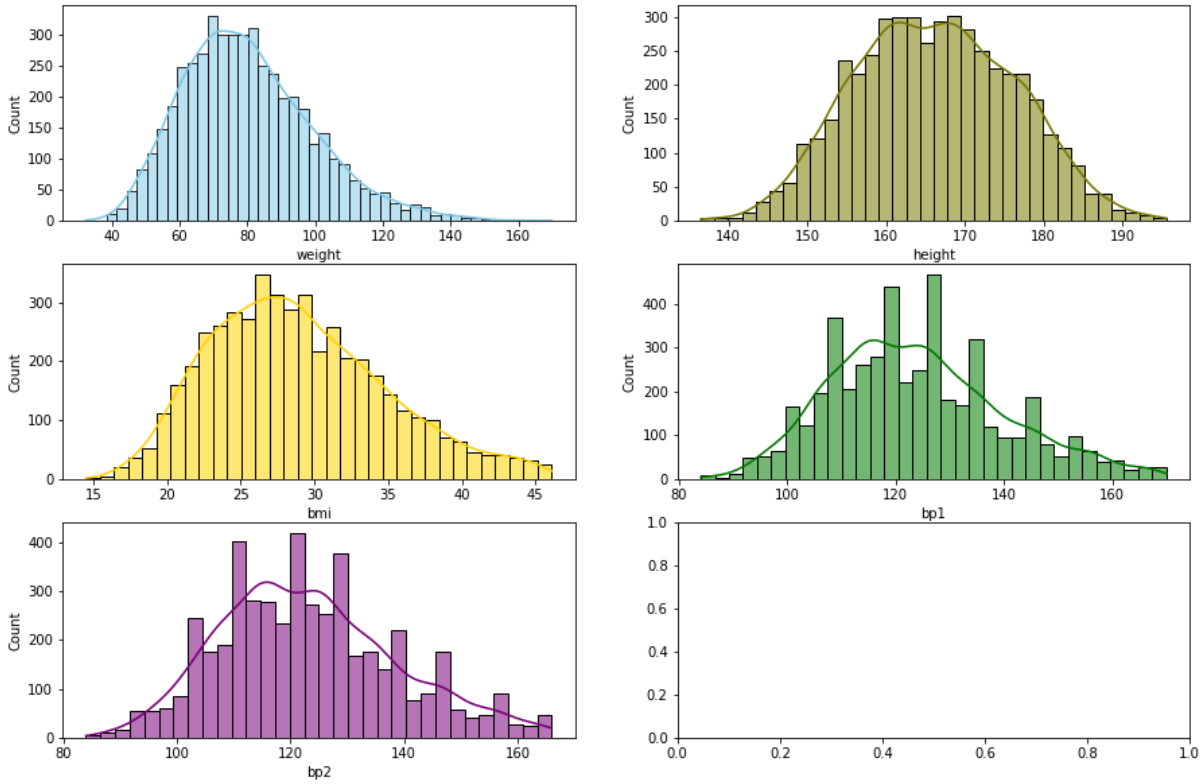

Figure 3: Histograms of Weight, Height, BMI, BP1, and BP2 post data cleaning

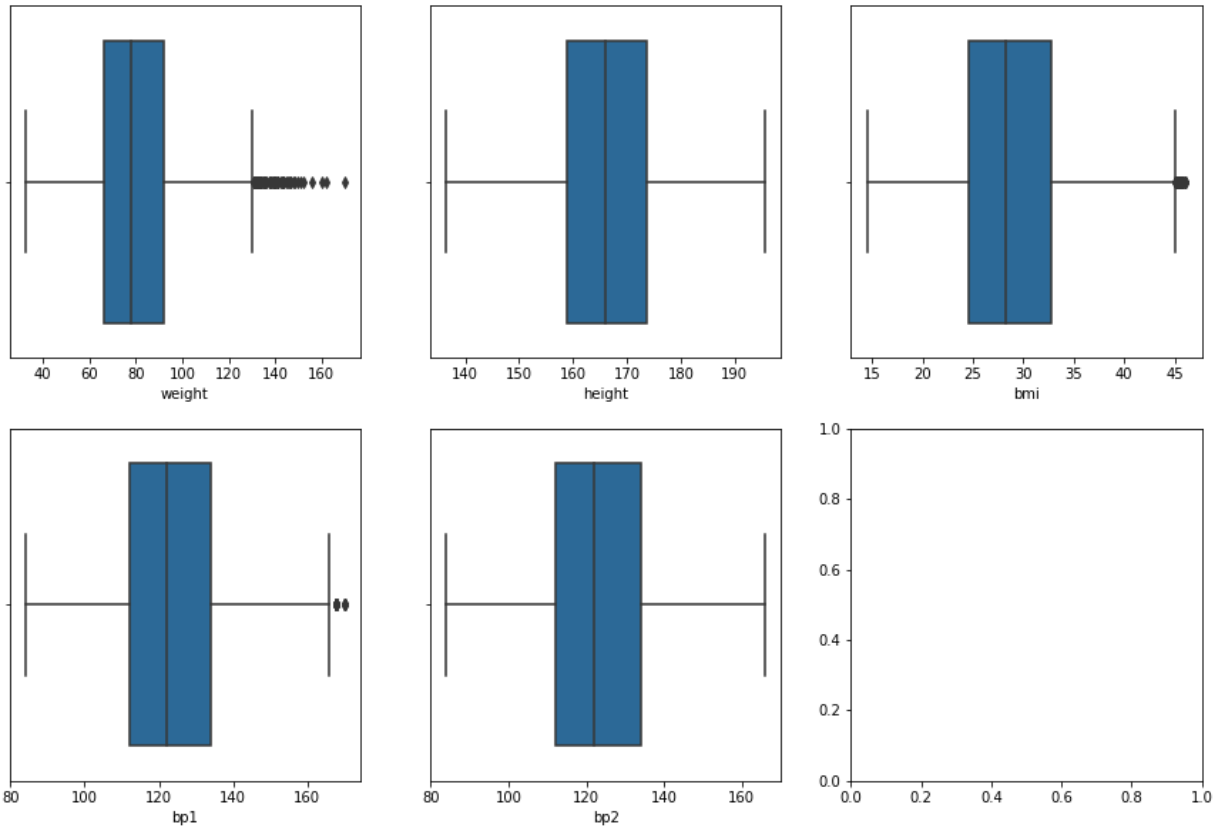

Figure 4: Box and Whisker Plots of Weight, Height, BMI, BP1, and BP2 post data cleaning

```
(18, 30]
pooled:  (1.0017886221345094, 0.3164456935802199)
unequal: (1.0122353553681755, 0.31142553532117845)

(30, 40]
pooled:  (0.9187850562544776, 0.35820801034676686)
unequal: (0.9162500283509661, 0.35953577129909386)

(40, 50]
pooled:  (2.1982214733482652, 0.02793332716600004)
unequal: (2.2365741257435063, 0.025314187579225614)

(50, 60]
pooled:  (1.5520439922611124, 0.12065169773596994)
unequal: (1.5534374923624668, 0.12031865168395503)

(60, 70]
pooled:  (2.7610542104653644, 0.005761511215181608)
unequal: (2.751825627904846, 0.005926406773513805)

(70, 80]
pooled:  (2.480691869267047, 0.01311276708669278)
unequal: (2.4788573018563804, 0.013180402543941755)
```

Figure 5: Table showing any difference between men's and women's BMI in 10 year bands

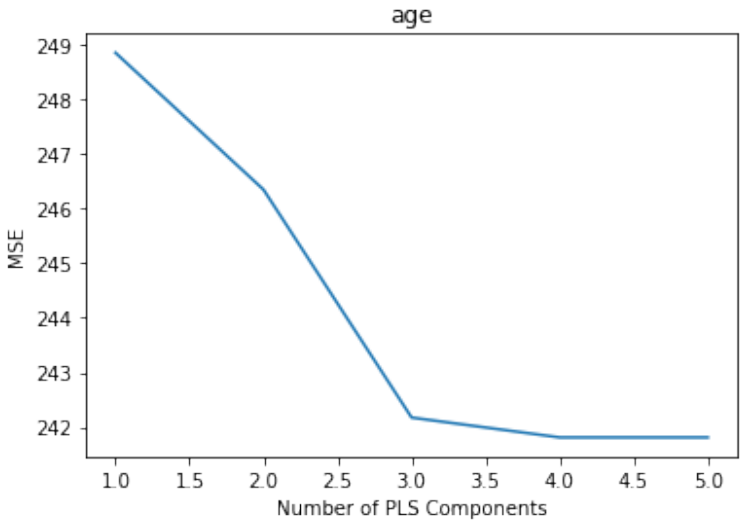

Figure 6: Test MSE Vs. number of components for age

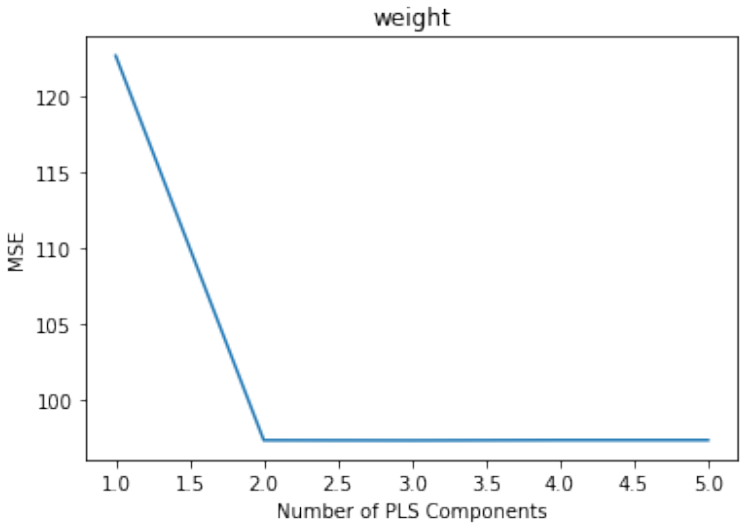

Figure 7: Test MSE Vs. number of components for weight

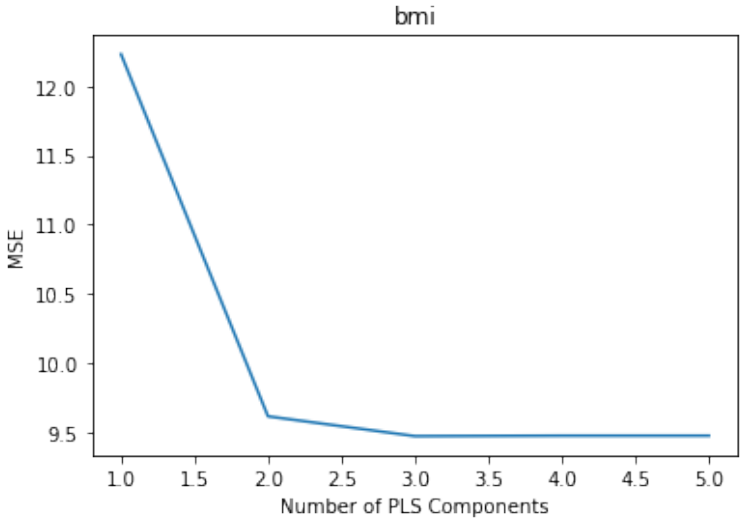

Figure 8: Test MSE Vs. number of components for bmi