Multivariate Analysis on Heart Disease

Joshua Fontes, Mengqi Yin, Sisang Cho

Math 257 - Dr. Gottlieb

San Jose State University

May 20, 2019

Introduction

Heart disease is the leading cause of death in the United States, accounting for approximately 1 in every 4 deaths annually. The necessity for improving understanding of heart disease in order to prevent it is apparent and it is the motivation for conducting this analysis. The dataset used in this report was retrieved from the UCI Machine Learning Repository. It was collected in from heart disease patients in Cleveland and contains 14 attributes concerning heart disease related measures. There are five quantitative variables and nine categorical variables. The patient attributes used for statistical analysis in this report are age, sex, cholesterol (chol), resting blood pressure (trestbps), maximum heart rate achieved (thalach), ST depression induced by exercise relative to rest (oldpeak), and the number of major blood vessels colored by fluoroscopy (ca).

Originally a classification problem, the purpose of the dataset was to help detect the presence of heart disease in a patient. In contrast, this report investigates how relationships between different health measures relate to heart disease. First, hypothesis testing and simultaneous confidence intervals are used to explore population mean differences between males and females in different health related measures, such as cholesterol and resting blood pressure. Then, a two-way MANOVA model was constructed to determine if sex and the number of major blood vessels colored by fluoroscopy were significant factors affecting cholesterol, resting blood pressure, and maximum heart rate achieved. Finally, a factor analysis and principal components analysis was conducted to better understand the variability in the data and how certain variables correlate with each other.

One Sample Inferences

Validation of Model Assumptions:

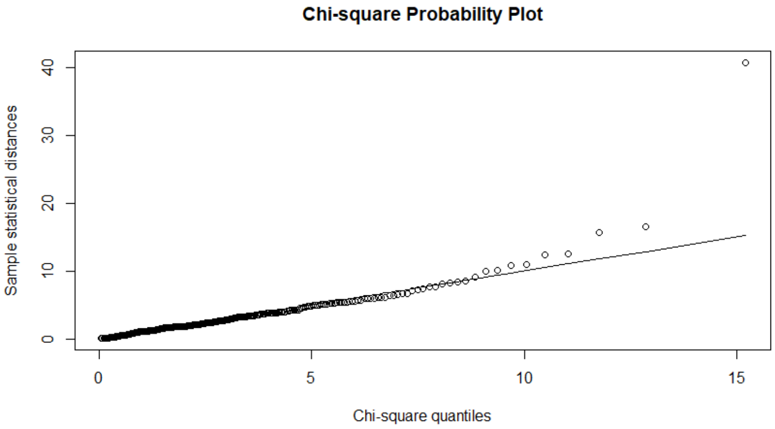

Figure 1: The figure above shows chi-square plot of the ordered statistical distance

Most of the statistical techniques conducted and discussed in the following sections assume that each vector observation Y_j comes from a multivariate normal distribution. Figure 1 shows a plot of the Mahalanobis distances, a statistical calculation used to check for multivariate normality. With the exception of one outlier, the statistical distances approximately follow the distribution of theoretical chi-square quantiles shown by the straight black line. This is strong evidence that our assumption for multivariate normality between our response variables is satisfied. Another assumption is that the populations of each group of response variables that we are testing share equal covariances matrices. The respective Bartlett's test for equality of covariance matrices was conducted with each of the following statistical procedures.

Hypothesis Testing and Simultaneous Confidence Intervals:

Our first question was whether men and women had significant population differences in average resting blood pressure, serum cholesterol, and maximum heart rate achieved. The approximate Hotelling's T^2 statistic was used for comparing responses from one population with independent responses from another population. We performed Bartlett's test to check for homogeneity of covariance matrices as the Hotelling's T^2 Test procedures assume equal covariance matrices between groups. The null hypothesis is that the covariance matrices for female and male were equal. The Bartlett's test statistic, 32.3995, is larger than the critical value, 12.59159. Thus, we rejected the null hypothesis at the 0.05 significance level and concluded that the female covariance was different from the male covariance. Consequently, it was reasonable for us to approximate distributions of T^2 when the population covariance matrices were not equal. Using the approximate Hotelling's T^2 Test for comparing mean vectors from two populations, the null hypothesis was that mean vectors were the same between female and male. The test statistic for this hypothesis testing was 10.11927, which is larger than the critical value, 5.991465. Thus, we concluded that there is a difference in average resting blood pressure, serum cholesterol, and maximum heart rate achieved between female and male.

Table 1 shows the 95% Scheffe and Bonferroni simultaneous confidence intervals for the differences in mean components. Because the confidence intervals for the difference in mean does not cover $\vec{0}' = [0\ 0\ 0]$, it appeared the female average mean vectors are different from those of male, which is consistent with the Hotelling's T^2 test.

		Lower Bound	Upper Bound
Bonferroni	trestbps (resting blood pressure)	1.457	2.816
	chol (serum cholesterol)	19.849	24.176
	thalach (maximum heart rate achieved)	1.377	2.950
Scheffe	trestbps (resting blood pressure)	0.768	3.505
	chol (serum cholesterol)	17.653	26.371
	thalach (maximum heart rate achieved)	0.579	3.748

Table 1: The table above shows 95% simultaneous confidence intervals on *trestbps*, *chol*, and *thalach* to compare differences between sex.

Furthermore, we compared the mean vectors from five levels of *ca* (number of major vessels colored by fluoroscopy). First, due to unbalanced sample size in each level of ca, we reduced the five different levels into two groups. Group 1 refers to patients with less than or equal to 1 major blood vessel and group 2 includes patients with greater than 1 major blood vessel colored by fluoroscopy. We first noticed that the sample covariance for group 1 were equal to the sample covariance for group 2 with Bartlett's test statistics 2.8395, which is smaller than the critical value 12.5916. As a result, it was reasonable to use the pooled covariance matrix for computing the Hotelling's T^2 test to investigate the null hypothesis that mean responses for group 1 are the same as those for group 2. The test statistic of 11.49878 was larger than the critical value 7.9573. Thus, we concluded that there were unequal population average responses between patients with less than or equal to 1 major blood vessel and patients with greater than 1 major blood vessel colored by fluoroscopy.

Table 2 shows the 95% Scheffe and Bonferroni simultaneous confidence intervals for the differences in mean components. It was obvious that $\vec{0}' = [0\ 0\ 0]$ was included in all confidence intervals for the difference in mean, hence, there were no differences in mean vectors between group 1 and group 2, which contradicted with the Hotelling's T^2 test result. The contradiction between the result of simultaneous confidence intervals and Hotelling's T^2 test can occur because the simultaneous confidence intervals produce a less compact region. Computing a confidence region produces a more compact estimate and possibly would produce the same result as our Hotelling's T^2 test. However, we computed simultaneous confidence intervals instead because a confidence region is much harder to meaningfully interpret.

		Lower Bound	Upper Bound
Bonferroni	trestbps (resting blood pressure)	-11.423	0.453
	chol (serum cholesterol)	-27.725	7.550
	thalach (maximum heart rate achieved)	-0.158	15.338
Scheffe	trestbps (resting blood pressure)	-12.443	1.473
	chol (serum cholesterol)	-30.753	10.579
	thalach (maximum heart rate achieved)	-1.488	16.668

Table 2: The table above shows the 95% simultaneous confidence intervals on *trestbps*, *chol*, and *thalach* to compare differences between the number of major blood vessels colored by fluoroscopy (*ca*).

Two-way MANOVA

We used a two-way MANOVA model, which is regarded as an extension of the two-way ANOVA, to analyze the impact of two factors, *sex* and *ca*, on the response variables *trestbps*, *cholesterol*, and *thalach*.

The model can be expressed as $Y_{lkr} = \mu + \tau_l + \beta_k + \gamma_{lk} + \varepsilon_{lkr}$ where Y_{lkr}represents a matrix of our response vectors, μ represents the overall population average response vector, τ_l represents the difference in average response based on sex, β_k represents the difference in average response based on ca, γ_{lk} represents the difference in average response based on an interaction between our two factors, and ε_{lkr} represents the error in our model.

Source	Df	Wilk's lambda	approx F	num DF	Pr(>F)
sex	1	0.95671	4.4038	3	0.004761
ca	4	0.86521	3.6225	12	2.676e-05
sex:ca	3	0.95243	1.5975	9	0.11

Table 3: The table above shows the summary results of the Two-Way MANOVA analysis.

Under the null hypothesis that all interactions between sex and ca are equal, $H_0: \gamma_{female,0} = \gamma_{female,1} = \ldots = \gamma_{male,3} = \gamma_{male,4}$, the Wilk's lambda statistic is 0.9524 and p-value is 0.11 which is bigger than the significance level of 0.05. Thus, we must fail to reject our null hypothesis that *sex* and the number of major blood vessels colored by fluoroscopy do not have a significant interaction. The null hypotheses of main effects of factor 1 and factor 2 are$H_0: \tau_{female} = \tau_{male}$ and $H_0: \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4$ respectively and the alternative hypotheses are at least one coefficient is not equal. The Wilk's lambda statistics are 0.95671 and 0.8652 for factor 1 and 2 and both p-values are less than significant level 0.05. Therefore, there is strong evidence that sex and ca have significant effect on cholesterol, resting blood pressure, and maximum heart rate achieved according to our test results with MANOVA. Consequently, our final model does not include an interaction term and can be expressed as $Y_{lkr} = \mu + \tau_l + \beta_k + \varepsilon_{lkr}$.

Validation of Model Assumptions:

The MANOVA procedure requires that the response variables in or model share multivariate normality, which was validated in the previous section under hypothesis testing. Furthermore, MANOVA also requires that the response variables in our model also have univariate normality. This assumption is checked in figure 2 below. Figure 2 shows a normal quantile plot of studentized residuals from our model. Figure 2 shows that the marginal distributions in our model appear a little skewed and not symmetric. However, they do approximately follow the distribution of theoretical quantiles for the normal distribution shown by the black lines. Thus, it is reasonable for our assumption of univariate normality in our responses to be satisfied.

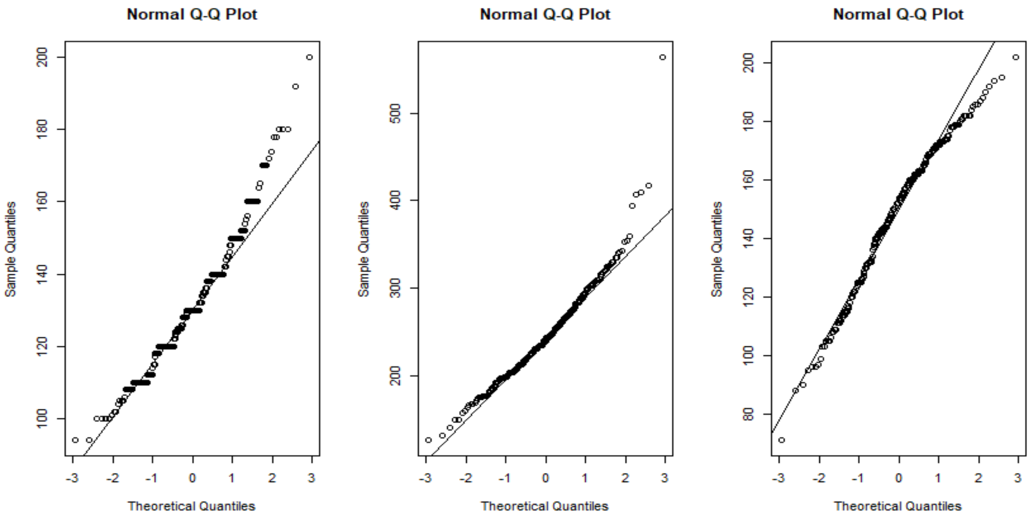

Figure 2: Marginal distributions for univariate normality of trestbps, chol, and thalach

Another assumption in the MANOVA procedure is that the covariance matrices of the potentially different populations are the equal. Thus the test for equality in covariance matrices is required and common method is Bartlett's test. In Two-way MANOVA, we needed to check whether the covariance matrices for each combination of levels from factor 1 and 2 are equal or not. For instance, first covariance matrix S1 was the combination of male and no major blood vessels. *Sex* had 2 levels, *Ca* had 4 levels and the number of the combinations were 4*2=8.

$$H_0 : {\sigma_1}^2 = \sigma_2 \quad {}^2 = \ldots = {\sigma_8}^2$$

$$H_1 : {\sigma_i}^2 \neq {\sigma_j}^2 \; for \; at \; least \; one \; pair \; of \; i, j$$

The test was based on chi-square approximation and Bartlett's statistics gives,

$$M = [\sum_t \quad (n_t - 1)] ln|S_{pooled}| - \sum_t \quad [(n_t - 1) ln|S_l|]$$

$\boldsymbol{s_l}$is the lth group sample covariance matrix and $\boldsymbol{s}_{pooled}$ is the pooled sample covariance matrix. We used R programming for this test and obtained 68.149 as test statistics. The critical value was $\chi^2_{p(p+1)(g-1)/2}(\alpha)$ and gained 58.124 when significance level was 0.05. Since the statistics were bigger than critical value, we could reject the null hypothesis and conclude that the covariance matrices of the sex and ca with different levels were not same. However, we had large samples which could robust nonnormality and have little impact on the MANOVA test. Therefore, we determined to continue to use MANOVA test even though the test for equality of covariance matrices results to reject of the null hypothesis.

Factor Analysis

We conducted a factor analysis to explain the covariance relationships among different variables by factors which are unobservable random quantities. At least five continuous variables were required to predict two factors but the heart disease dataset only had four related continuous variables. Thus, we decided to use the quantitative variables *trestbps*, *chol*, *thalach*, and *oldpeak*. Table 4 through table 7 below show the results of our factor analysis.

trestbps	chol	thalach	oldpeak
0.962	0.997	0.881	0.005

Table 4: Uniqueness

	Factor1
trestbps	0.194
chol	
thalach	-0.345
oldpeak	0.997

Table 5: Loadings

	Factor1
SS loadings	1.154
Proportion Var	0.289

Table 6: Variabilities

According to results of the factor analysis, test of the hypothesis that 1 factor was sufficient. The chi square statistic was 4.14 on 2 degree of freedom and p-value was 0.126. The factor was dominated by *oldpeak* and small of proportion of resting heart rate. We could say that high depression level and resting heart level will increase the possibility of having heart disease. On the contrary, maximum heart rate reduced the overall score.

The uniquenesses of *trestbps*, *chol*, and *thalach* were too high to be explained by this model but the variable *oldpeak* is well described by the model. The proportion of variability of factor 1 was 0.289, therefore it was hard explain total variability with factor1. We needed more continuous variables to estimate more factors and increase the percentage of explanation of total variability.

	trestbps	**chol**	**thalach**	**oldpeak**
trestpbs	-6.96951e-09	1.12679e-01	2.013158e-02	8.828009e-06
chol	1.126792e-01	3.721106e-08	8.755162e-03	-9.647578e-05
thalach	2.013158e-02	8.755162e-03	-4.96739e-08	-2.269248e-05
oldpeak	8.828009e-06	-9.647578e-05	-2.26924e-05	-2.717442e-08

Table 7: Residual matrix

This is the residual matrix of this two-factor model. The residual correlation of all variables have small variances which means that the model can be well explained by 4 different variables.

Principal Components Analysis

A principal components analysis was conducted to better understand variability in the data and to try to reduce the dimensionality of the data. The response variables cholesterol (chol), resting blood pressure (trestbps), maximum heart rate achieved (thalach), and ST depression induced by exercise relative to rest (oldpeak). To reduce all variables to similar units, the analysis was computed using the correlation matrix of the variables, rather than the covariance matrix.

PC1	**PC2**	**PC3**	**PC4**
1.1983	1.0334	0.934	0.7898

Table 8: Standard Deviations

	PC1	PC2	PC3	PC4
trestbps	0.4184	0.5099	0.6904	-0.2969
chol	0.2141	0.7215	-0.6584	0.0101
thalach	-0.5771	0.4341	0.2976	0.6244
oldpeak	0.6678	-0.1757	0.0357	0.7224

Table 9: Principal Components

PC1	PC2	PC3	PC4
35.89%	26.69%	21.81%	15.59%

Table 10: Proportion of Variance

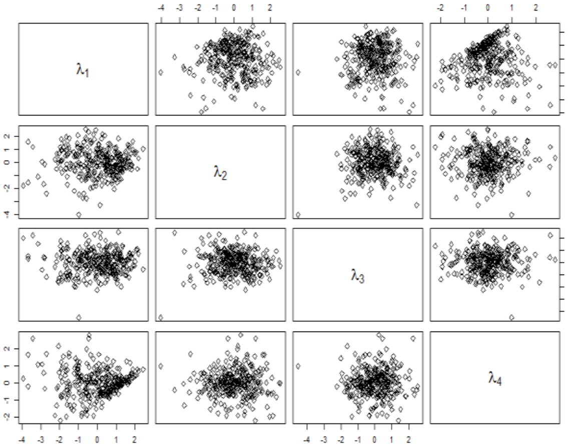

Figure 3: Scatterplots for PCA

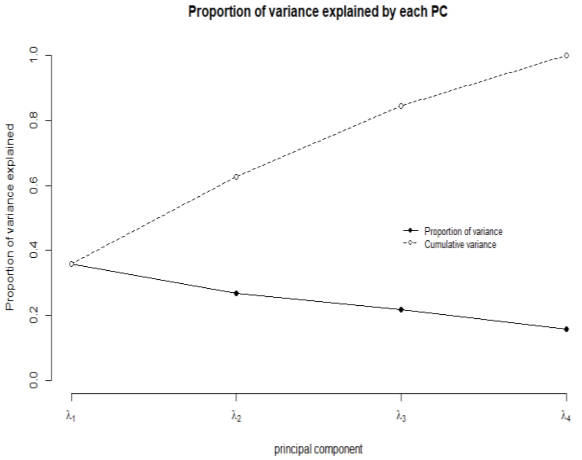

Figure 4: Scree Plot for PCA

The first 3 principal components capture about 84.4% of variability in the correlation matrix. It may be possible to adequately summarize the data with only 3 variables, rather than 4. However, the last principal component captures about 15.6% of the variability in the correlation matrix, which is arguably large enough that it needs to be included.

The first principal component can be described as an overall health index or an aggregate "heart disease" measure. Cholesterol, resting blood pressure, and old peak are contrasted against maximum heart rate achieved. This makes practical sense because higher values of cholesterol, resting blood pressure, and ST depression are generally associated with an unhealthy human body. In contrast, higher values of maximum heart rate are generally good because this indicates a strong and healthy heart.

Principal component 2 is a contrast between old peak and all other response variables, although old peak has a rather small coefficient. PC2 is mainly dominated by the grouping of resting blood pressure, cholesterol, and maximum heart rate. Principal component 3 is a contrast between cholesterol and all other response variables. Although, PC3 is dominated by cholesterol and resting blood pressure. Principal component 4 is a contrast between resting blood pressure and all other response variables. Although, PC4 is dominated by maximum heart rate achieved and old peak.

In summary, the variability in the dataset seems to be relatively evenly dispersed throughout all four principal components. I would argue that all four variables included in the analysis provide essential information for describing heart disease in heart disease patients.

Conclusion

We conducted various multivariate analyses to find evidence of any significant relationship between the two factors sex and ca, and the three response variables which are average resting blood pressure, serum cholesterol, and maximum heart rate.

First, we tested the significance of *sex* and *ca* on the response variables using Hotelling's T^2 test. Depending on the equality of covariance matrices in each population, we used approximate T^2 test or the pooled covariance matrix. Through the tests, we concluded that *sex* and *ca* both had a significant impact on the response variables. We also calculated 95% Scheffe and Bonferroni simultaneous confidence intervals for the difference in mean components. All confidence intervals for *sex* excluded 0 but the intervals for *ca* all included 0 which contradicted the result of the T^2 test. One possible explanation for this is that the Hotelling T^2 test produces a larger region than the simultaneous confidence interval. Furthermore, we examined the impact of the two factors on the response variables with Two-way MANOVA. The result of the MANOVA procedure showed strong evidence of significance for the two factors which agreed with the Hotelling T^2 result. It also showed that the interaction between sex and ca was not significant. The assumption of univariate normality in our responses was satisfied, judging by the normal quantile plot of studentized residuals.

Factor analysis was utilized for explaining the covariance relationships among different response variables by factors which are unobservable random quantities. We added one more continuous variable, *oldpeak*, to increase the explanation of total variability. We obtained one factor which was dominated by *oldpeak* and could explain 28% of the total variability. The heart disease dataset included age and four continuous variables, but future studies should collect data on more continuous variables to increase the proportion of variability for explaining total variability. Lastly, we used principal components analysis to reduce the dimension of the data and further understand variability in the data. According to the first principal component which explains 35% of the variability in the total population, high level of cholesterol, blood pressure and depression level increased the risk of heart disease and maximum heart level decreased the risk. The cumulative proportion of variance of first 3 principal components was over 80 percent but we also recommend considering all components since principal 4 had 15.6% variability.

References

Heart Disease Facts & Statistics. (n.d.). Retrieved from https://www.cdc.gov/heartdisease/facts.htm

Johnson, R. A., & Wichern, D. W. (1992). Applied multivariate statistical analysis. Englewood Cliffs, N.J: Prentice Hall.

Ronit. (2018, June 25). Heart Disease UCI. Retrieved from https://www.kaggle.com/ronitf/heart-disease-uci

UCI Machine Learning Repository: Heart Disease Data. (n.d.). Retrieved from https://archive.ics.uci.edu/ml/datasets/heart disease