MSDS 6372 Project 2

Principal component Analysis

on Heart Disease Factors

**Team members:**

Kottegoda Manjula

Messa Araya

Kumar Rajeev

Date: 2016-11-06

**Introduction**

Heart Disease is the leading cause of death in the United States and is primarily caused by the coronary artery disease, or plaque build-up on the inner walls of arteries resulting in restricted blood flow. In most cases the plaque build-up happens from a young age and often goes unnoticed until a catastrophic event.

Heart Disease has been the subject of countless research studies. As a result reliable data is now available on the subject where modeling can be performed with moderate to high degrees of accuracy. Our team decided to analyze such a dataset with the hope that our study will, in some way, provide meaningful insight into early detection which could help manage risk and avoid disastrous consequences if left untreated.

The Heart Disease Dataset from the UCI Machine Learning Repository [1] was used in this study. We performed analysis on 14 explanatory variables (Fig 1). Since there was a high likelihood of variables being correlated we chose Principal Component Analysis (PCA) as the preferred method to examine the data. This paper will examine which variables are correlated and which variables directly influence in the prediction of the existence of heart disease in patients.

**Descriptive Statistics**

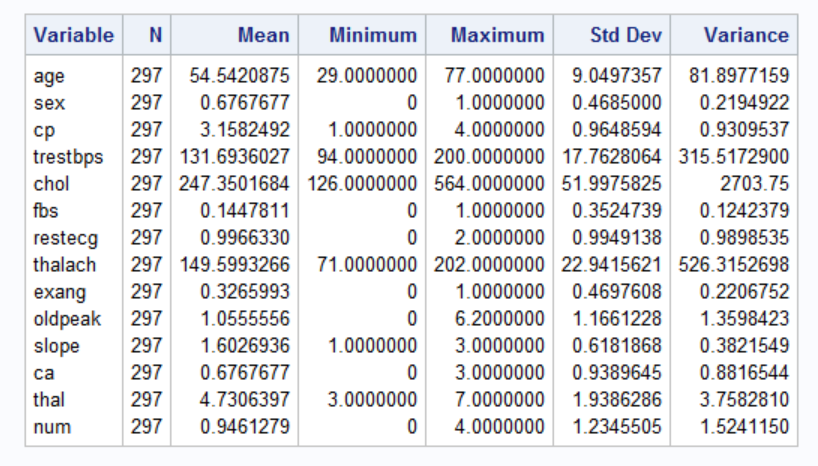
|  |  |  |  |
| --- | --- | --- | --- |
| **Variable name** | **Short desciption** | **Variable name** | **Short description** |
| age | Age of patient | thalach | maximum heart rate achieved |
| sex | Sex, 1 for male | exang | exercise induced angina (1 yes) |
| cp | chest pain | oldpeak | ST depression induc. ex. |
| trestbps | resting blood pressure | slope | slope of peak exercise ST |
| chol | serum cholesterol | ca | number of major vessel |
| fbs | fasting blood sugar larger 120mg/dl (1 true) | thal | no explanation provided, but probably thalassemia (3 normal; 6 fixed defect; 7 reversable defect) |
| restecg | resting electroc. result (1 anomality) | num | diagnosis of heart disease (angiographic disease status) |

*Fig 1. The variables used for analysis*

The variable we want to predict is **num** with Value 0: < 50% diameter narrowing and Value 1: > 50% diameter narrowing. We assume that every value with 0 means heart is okay, and 1,2,3,4 means heart disease

**Basic Statistics and Assumptions of PCA**

***Descriptive Statistics***

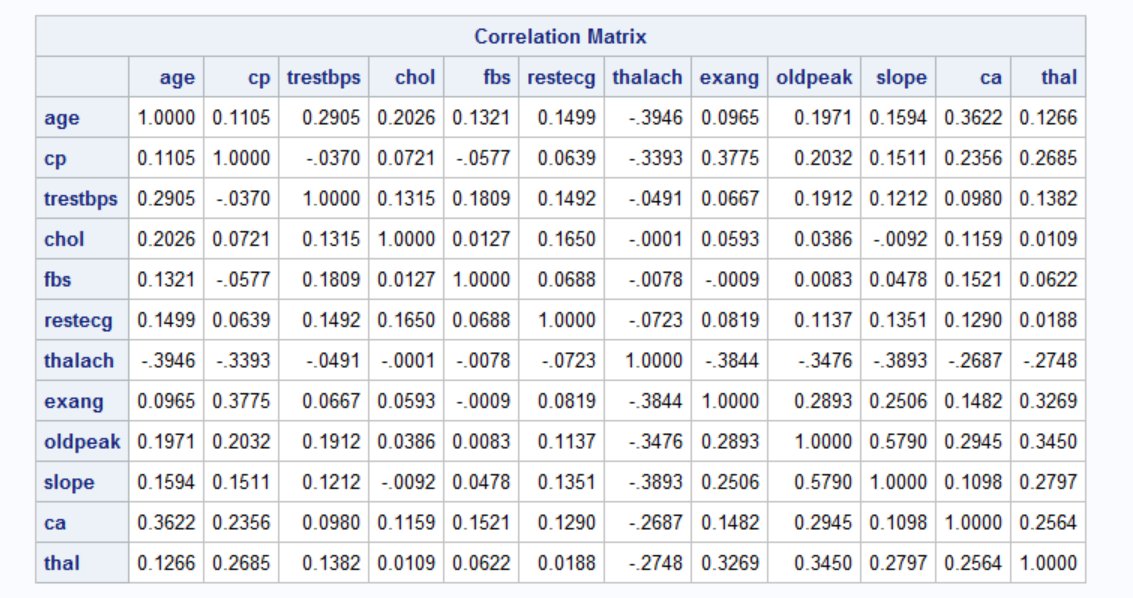


*Table1. Descriptive statistics of the variables*



*Fig 2. Scatterplot and histograms of the variables*

There are orders of magnitude differences in the standard deviations and means of the different variables (Table 1). This is expected because different variables will have varying ranges of possible values. The scatter plot shows most of the variables seem to have no real relationship with each other although there are some that do have collinearity (fig 2). Also, the histograms do not show any normality.



*Table 2 Correlation of Variables*

The variables that correlate with each other make sense (Table 2) as there is no strong linear relationship between variables. Given the number of variables and correlation between them (its existence and that it’s linear) these set of data is a good candidate for PCA using the correlation matrix and multiple linear regression with the selected principal components.

***Assumptions[2]***

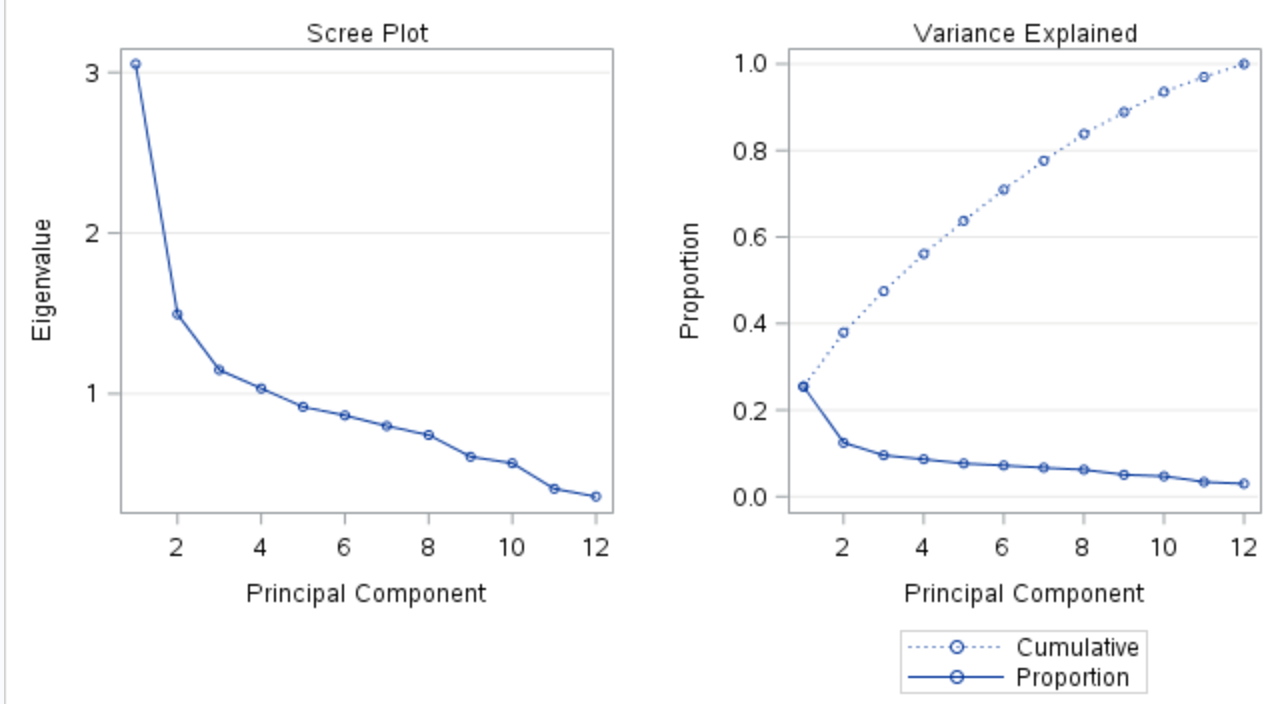
The following four basic assumptions of PCA are satisfied:

* **Data Continuity:**  Data or measurements need to be continuous. Our dataset does not have any discrete value hence continuity is satisfied.
* **Linearity:** There needs to be a **linear relationship between all variables** since PCA is based on Pearson correlation coefficients. The scatterplot and correlation matrix show there is colinearly between variables. Hence, there is linearity behavior between the variables.
* **Sample adequacy:** For PCA to produce a reliable result, large enough sample sizes are required. Our dataset has 297 observations and is adequately enough.
* **No significant outliers:** There should be **no significant outliers**. There are not substantial outliers in our data set.

Hence, PCA is appropriate for our chosen dataset to prepare it for the regression.

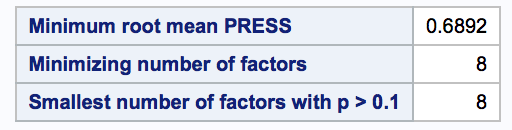
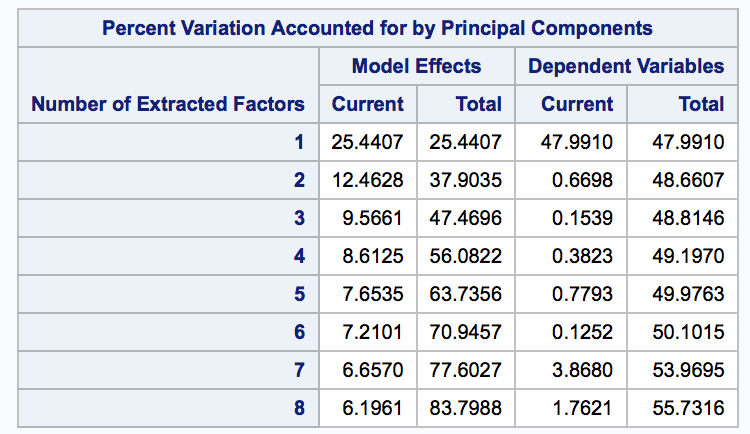
***Principal Component Analysis***

First the full PCA was run using SAS’s PROC PRINCOMP to generate the eigenvectors for the components and a Scree plot to begin to determine which components should be selected for the regression model.



*Fig 3. Scree Plot and Variance Explained*

The above Scree plot (fig 3) has no extremely obvious ‘elbow’, there seems to be an inflection point at component 3. The proportion of variance explained also has no obvious leveling off point. The increase slows around points 4-8 and slows even more after component 8. In order to more clearly determine which components to select the PROC PLS with cross validation and the PCR method was run.



*Table 3. Accounted Variation Table 4. Cross Validation*

From Table 3 and 4, it is evident that selection of 8 components makes the press statistics to be the smallest. Moreover, the 8 components explains 83% of the total variation. Therefore, the final model will have 8 factors.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Eigenvectors** | | | | | | | | |
|  | **Prin1** | **Prin2** | **Prin3** | **Prin4** | **Prin5** | **Prin6** | **Prin7** | **Prin8** |
| **age** | 0.298686 | 0.396460 | 0.151479 | -.165697 | -.474436 | 0.091428 | -.243692 | 0.036575 |
| **cp** | 0.289789 | -.278238 | 0.450798 | -.075745 | 0.213203 | -.073314 | -.085248 | 0.066713 |
| **trestbps** | 0.176362 | 0.453249 | -.276838 | 0.074592 | 0.199307 | 0.456010 | -.320046 | 0.389511 |
| **chol** | 0.098083 | 0.361288 | 0.447882 | 0.380100 | 0.120792 | 0.360021 | 0.256990 | -.522654 |
| **fbs** | 0.076418 | 0.379392 | -.260146 | -.493554 | 0.422075 | -.322393 | -.082531 | -.473786 |
| **restecg** | 0.149209 | 0.312123 | 0.097017 | 0.515948 | 0.201299 | -.646523 | -.045200 | 0.318826 |
| **thalach** | -.396860 | 0.147668 | -.085014 | 0.105417 | 0.350734 | 0.103616 | 0.382129 | 0.143678 |
| **exang** | 0.330898 | -.268734 | 0.174041 | 0.044359 | 0.380454 | 0.060899 | -.366595 | -.095470 |
| **oldpeak** | 0.398513 | -.096147 | -.343025 | 0.203023 | -.104569 | 0.053773 | 0.327855 | -.052070 |
| **slope** | 0.355898 | -.142267 | -.451639 | 0.288387 | -.139115 | -.096526 | 0.067009 | -.300380 |
| **ca** | 0.308478 | 0.197425 | 0.228503 | -.368554 | -.152437 | -.175741 | 0.525935 | 0.224522 |
| **thal** | 0.332199 | -.159988 | -.090193 | -.183498 | 0.367680 | 0.263131 | 0.294988 | 0.264756 |

*Table 5. Eigenvector for Principal Components*

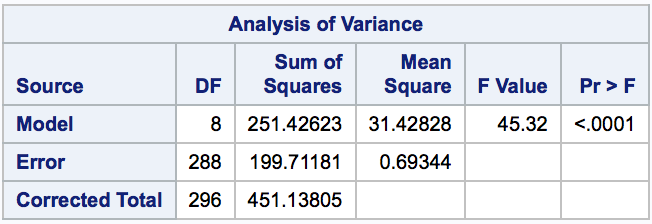
Based on the significant factors in each component (Table 5) I would base these as being representative of the following: Prin1 is related to heart stress testing parameters such as maximum heart rate achieved (thalach), oldpeak and slope. The oldpeak and slope has a positive relationship while thalach has a negative relationship. Prin2 is the combination of the two measures age and resting blood pressure(trestbps). The seniors have the higher chance of heart disease. Also, the person with high blood pressure is more susceptible to heart disease. Prin3 is a combination of cp and chol. Higher the cholesterol level, the higher is the rate of diagnosis of heart disease. Also, chest pain has been linked to higher occurrences of coronary disease. Prin4 is a combination of blood sugar and resting ecg. Both variables have a positive relationship with heart disease.

**Multiple Regression**

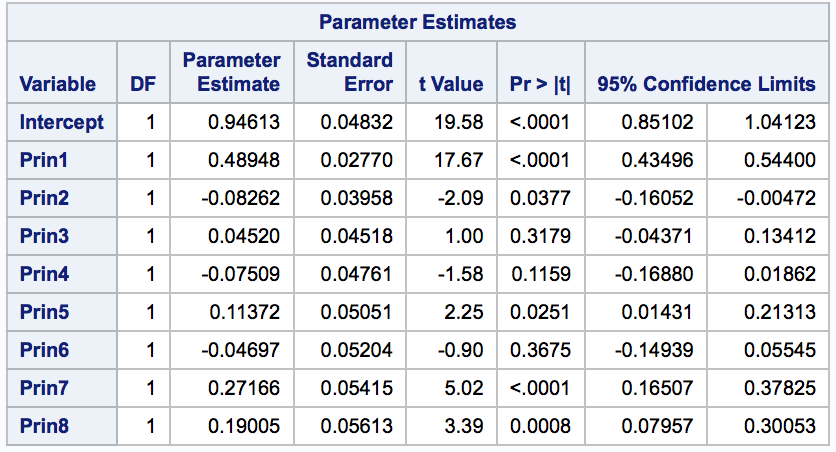
The model being tested is:

**Diagnosis of heart disease(Num) = β0 + β1 Prin1 + β2 Prin2 + β3 Prin3+ β4 Prin4+ β5 Prin5+ β6 Prin6+ β7 Prin7+ β8 Prin8.**

The coefficient β0 represents the mean diagonsis of heart disease while the other eight represent the magnitude and direction (negative or positive) of each principal component used in the model. The F-test resulting from this model is significant (p-value of <0.0001) showing that our model is apporpriate for explaining at least some of the variation (Table 6 ).

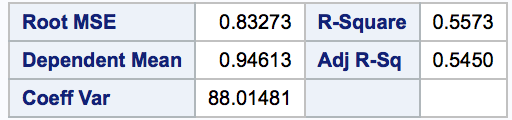


*Table 6. Overall F-test*



*Table 7. Parameter Estimates for regression model*

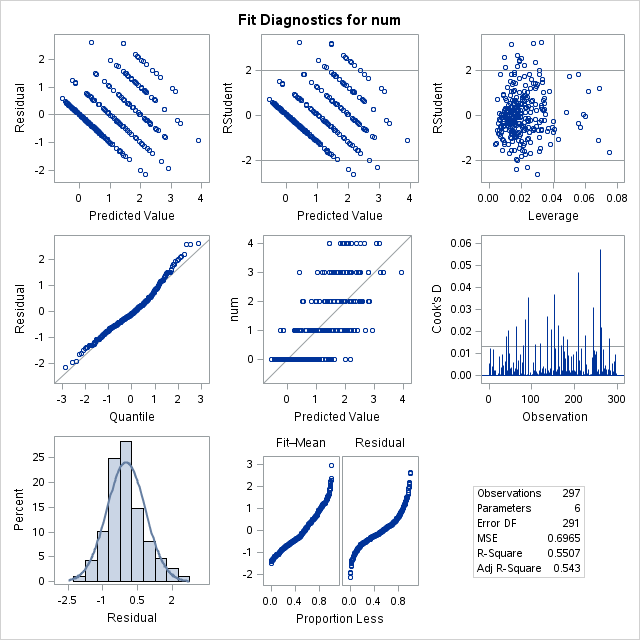
Surprisingly only five of our parameter estimates are significant (p-values <0.05) while the other three parameter estimates (Prin3, Prin4, and Prin6) are not significant (p-values of 0.3179, 0.1159, and 0.3675 respectively). Even Prin1 does not have a very large absolute value and so can only have a moderate influence on the heart disease given the real-world constraints (Table 6). Therefore, only Prin1 (0.489) has moderate impact on heart disease.

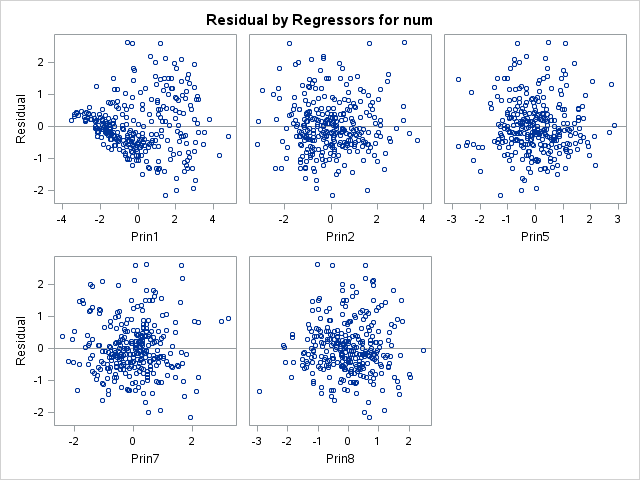


*Table 8. R2 table*

The AdjR2 value (Table 8) for the model is moderate (0.545), which shows our explanatory variables explain the response variable well.

*Fig 4. Model residuals and Cook’s D*





Observing at the residual plot for the model (fig 5) the residuals are arranged in to bands (strips). The plot has a series of lines because these correspond to the distinct values of Y (response) with the lowest line corresponding to category with Y = 0, the next line corresponds to Y = 1, then Y = 2 and so on up to Y=4. Otherwise the residual plots are symmetric with respect to the zero line. The histogram and QQ plot both indicate normality. The residual plots for each of the four principal components look good and show no strong evidence of problems with non-constant *Fig 5. Residuals for components*

evidence (fig 4). There is one high value of Cook’s D in fig 3 however the actual value of it is about 0.055 which isn’t high enough to cause much concern.

**Conclusion**

Heart disease is the leading killer both in the United States and increasing throughout the world. Some of the more recognizable indicators of the disease are physiological in nature such as chest pain, shoulder discomfort and exhaustion. It is the unnoticed indicators lurking in the background that are the most dangerous and this study examined some of these no-so-obvious indicators and their influence towards developing heart disease.

From our Principle Component Analysis results, we can see that the first four principle components explain close to 60% of the variability while the first eight account for 83% of the variability. Furthermore from the first four principle components alone some strong influences can be established.

The ST segment depression (oldpeak), a measurement in electrocardiogram, and a lower heart rate (thalach) both measured during stress testing each account for almost 10% of the variability in predicting the existence of heart disease. A lower heart rate during exercise generally speaks to a strong heart that efficiently pumps blood without having to overexert itself. The peak slope maintained in the treadmill during stress test also has a strong influence in the prediction. Overall the first four principle components seems to suggest that measurements of exercise tolerance in a patient is the most reliable predictor for heart disease.

Continuing the analysis into the other influential variables show that the resting blood pressure (trestbps) and chest pain type (cp) also has significant influence. Age being included confirms the fact that plaque build-up happens over time as age can explain close to 7% of the variability. Lifestyle habits are also reflected in the measurements in the form of cholesterol (chol) and blood sugar (Fbs) both of which play a key role influencing the likelihood of heart disease. Cholesterol (chol) alone can account for 14% of the variation thus managing cholesterol trough lifestyle choices can be a key preventative measure.

Understanding the key indicators, getting routine measurements and taking preventative actions based on all indicators seems to be the most effective way to fight this hidden epidemic that is prevalent across the nation.

**References**

1. *Heart disease Dataset:* <http://archive.ics.uci.edu/ml/datasets/Heart+Disease>
2. *PCA Assumptions:* [*https://statistics.laerd.com/spss-tutorials/principal-components-analysis-pca-using-spss-statistics.php*](https://statistics.laerd.com/spss-tutorials/principal-components-analysis-pca-using-spss-statistics.php)

**Appendix**

**PROC** **IMPORT** DATAFILE= "\\client\C$\Users\ARAYA ABADI\Desktop\EXPERMENTALSTAT2\_6372\Cleveland\_heartdisease\_data.csv"

OUT= heartdisease

DBMS=csv

REPLACE;

GETNAMES=YES;

**RUN**;

**proc** **print** data=heartdisease;

**run**;

**proc** **means** data=heartdisease n mean min max std var;

**run**;

title 'histogram and scatter';

**proc** **sgscatter** data=heartdisease;

matrix age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal num / diagonal=(histogram) group=num;

**run**;

title 'full PCA with num';

**proc** **princomp** data=heartdisease out=heartdiseaseP2;

var age /\*sex\*/ cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal;

**run**;

title 'PCR with cross num';

**proc** **pls** data=heartdisease method=PCR cv=one cvtest (stat=PRESS);

model num = age /\*sex\*/ cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal;

**run**;

title 'PCR with 8';

**proc** **pls** data=heartdisease method=PCR nfac=**8**;

model num = age /\*sex\*/ cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal;

**run**;

title "Reg 8 Prin CI";

**proc** **reg** data=heartdiseaseP2;

model num = Prin1 Prin2 Prin3 Prin4 Prin5 Prin6 Prin7 Prin8 / CLB;

**run**;

|  |  |
| --- | --- |
| Grade: 96 | |
| Mechanics | |
| Spelling & Grammar | good |
| Flow of thought | good |
| Sections | Well defined |
| Figures & tables | Well numbered but may need better organization. |
| Code | Included |
| Title | Good |
| Analysis | |
| Data Description | An interesting problem with clear introduction of dataset. |
| EDA | Very detailed. |
| Method & interpretation | Well described. The PCA is conducted with careful assumption checking. But the assumptions for MLR need further consideration and validation. |
| Diagnostics | Diagnostic plot is included. |
| Conclusion | Reasonable conclusion is included. |