Experimental Design and Data Analysis: Calcium, inorganic phosphorus and alkaline phosphatase levels in elderly patients

Andrew Bedard(2566978) & Simone van Gompel(2567525)
Group 19

June 3, 2015

1 Introduction

In this paper the process of analysing a certain dataset is laid out. The dataset used is calcium.dat, which can be found at `http://www.amstat.org/publications/jse/jse_data_archive.htm`. This dataset is used for this project is because it has a good number of entries and enough features to be able to do data analysis on. In the rest of this paper the experiment is explained, the data is analysed, the found results are shown and all of this will be discussed at the end. In the appendix the used R code is shown.

2 The Experiment

The experiment that was set up had as goal to see if age and sex has an influence on certain concentrations in the body. The concentrations that were measured are:

- Alkaline Phosphatase International Units/Liter

- Calcium mmol/L

- Inorganic Phosphorus mmol/L

There are 6 different labs from which the data is extracted. Next to these features, the sex, age, agegroup and patient observation number are recorded. In the calcium.dat the original data is stored with errors, in calciumgood.dat the data is already cleaned up. In this project only the calcium.dat data is used to explain how the cleaning up of the data is done. The research question we want to answer is: What influence does age have on the given concentrations in the body?

3 Data Analysis

3.1 Preparation

The data needed to be prepared to be able to read it in R. This preparation exists of replacing the empty fields by an underscore, with this the data can be read in R. First off a pairs plot was made of the data to be able to see how the data relates to each other. In Fig1 the foremost problem is clear, the outliers in the data are big and unlogical. Furthermore has Lab more than 6 categories, these problems can be explained by human error. And lastly the lab 3 has strange measurements with the cammol, this might be because of confusion of the measurement unit. The following values were changed:

- Removed Ages over 110

- Removed Sex which is not in the category 1 or 2

- Removed Lab categories which are over 6

- Removed Phosmmol over 2

- Divided Cammol of Lab 3 by 10 (this is visually tested by using a pairs plot)

After removal the same plot is made, see: Fig2. Here you can see the correlations between the features better than in Fig1.
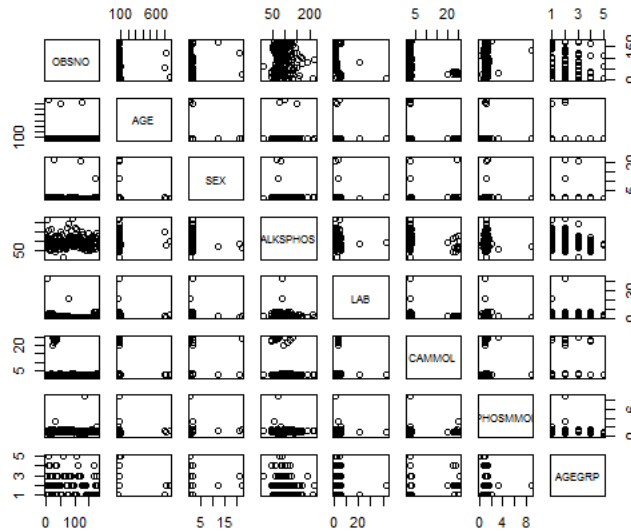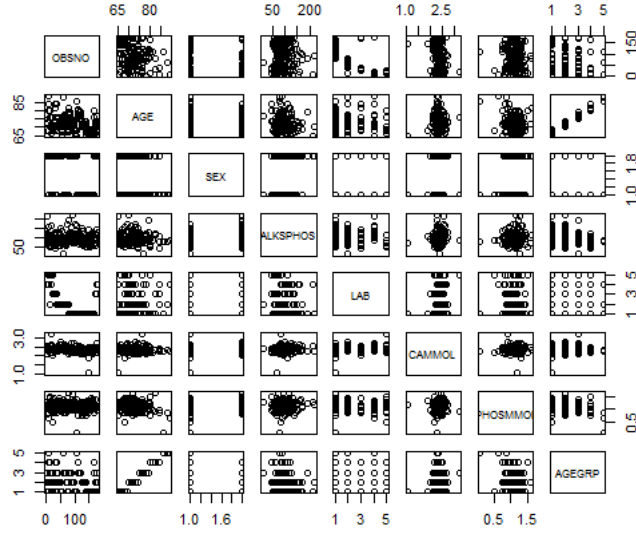

Figure 1: Pairsplot of the original data

Figure 2: Pairsplot of the updated data

3.2 Analysis

The feature that needs to be analyzed is age, this can be done by either using the feature age or the feature age group. The age group exists of 5 levels and is thus categorical. The different age groups are: 1=65-69, 2=70-74, 3=75-79, 4=80-84, 5=85-89 Years. In Fig3 the relation between the different concentrations and the age groups are visually represented.
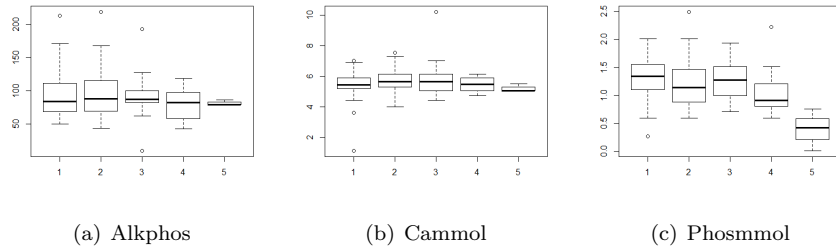

(a) Alkphos

(b) Cammol

(c) Phosmmol

Figure 3: Boxplots of the concentrations with the age groups

There is some difference between the different age groups with all of the concentrations. But with Alksphos the difference only seems to be in the variety of the values, the same holds for Cammol. For Phosmmol there does seems to

Table 1: The number of patients per age group

Groups	1	2	3	4	5
Patients	56	70	38	10	3

be a difference in the age groups, the median drops with the older ages. The difference in the first two concentrations is because of the number of people in the different age groups, see Table1.

4 Results

5 Discussion

6 R-Code