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 4, 2015

1 Introduction

In this paper we seek to determine whether there are significant differences in the concentrations of 3 substances in the blood of elderly patients, based on gender and age groups. The dataset used is calcium.dat, which can be found at `http://www.amstat.org/publications/jse/jse_data_archive.htm`. This dataset was chosen for this project because of its reasonable size (1424 data points), and because it contained real clinical data that would give us a chance to exercise our ability to use data that likely contains errors. All tests in this paper are done with a significance value of 0.05, and the R code used is shown in the appendix.

2 The Experiment

The experiment was set up with the 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/Litre (Alksphos)
- Calcium mmol/L (Cammol)
- Inorganic Phosphorus mmol/L (Phosmmol)

There are 6 different labs from which the data is extracted. Next to these features, the sex, age, age-group 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, but for this paper only the calcium.dat data is considered. The research question we want to answer is: What influence does age and sex have on the given concentrations in the body for patients over the age of 65, and can a predictive model be created to estimate these values?

3 Data Analysis

3.1 Preparation

Before loading the data into R, there were a few steps that needed to be taken. First, the table reading method in R will not work correctly if there are missing data points with no representations, so these points were manually replaced with an underscore to allow us to tell R to replace these underscores with *NA*. Next headers were used to simplify our code, and let us know when we

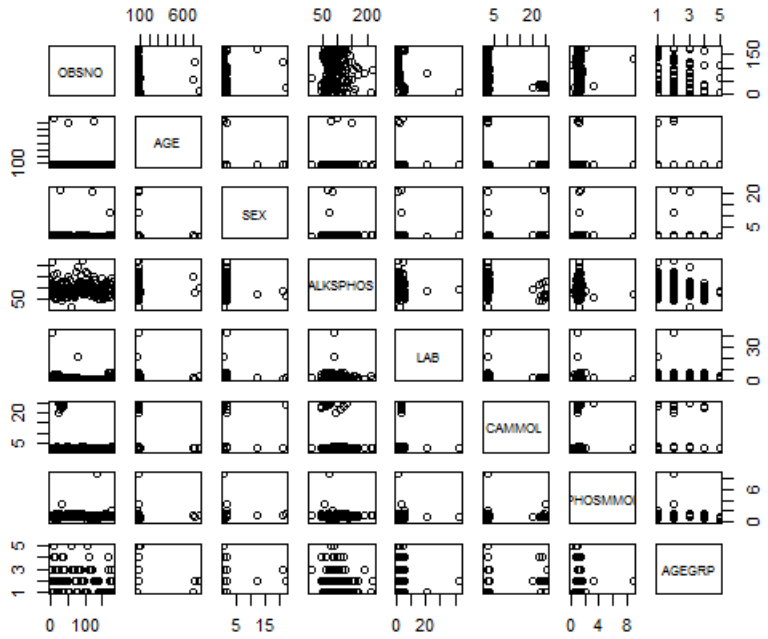


Figure 1: Pairsplot of original data

are referring to each variable by name, these were OBSNO, AGE, SEX, ALKSPHOS, LAB, CAMMOL, PHOSMMOL, AGEGRP, which represent patient number, age, sex, alkaline phosphatase in international units/litre, calcium in mmol/litre, inorganic phosphorus in mmol/litre and age group of patient respectively. This allowed us to create a pairs plot to get a general overview of the data, and in Fig: **??**, just from observing how many points lie a great distance from the main clusters in each pair, it is clear there are a large number of outliers in our original data.

For example, lab 3 has strange measurements for cammol, which appears to be a misplaced decimal, causing all measurements to be an order of magnitude larger than all other measurements for cammol. This, as well as the other behaviour of the data for cammol, alksphos and phosmmol can be seen in QQ-plots of all data for each substance.(Fig: 2)

It is clear from Fig: 2 that our populations are certainly not normal, and this is perhaps due to the outliers. By experimentation, the values were changed in the following manner:

- 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 or under 0.2
- Divided Cammol of Lab 3 by 10 (this is visually tested by using a pairs plot)
- Removed Cammol over 3 or under 1.9
- Removed Alksphos over 150 or under 20

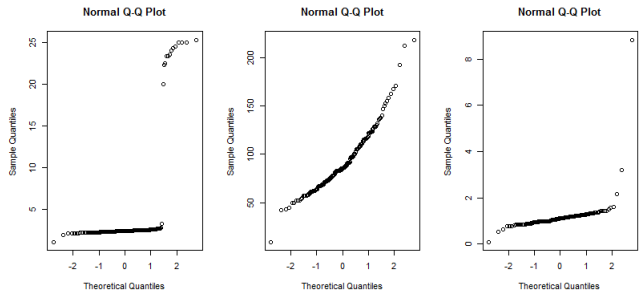


Figure 2: QQ plot of raw data, from left to right, Cammol, alksphos, phosmmol

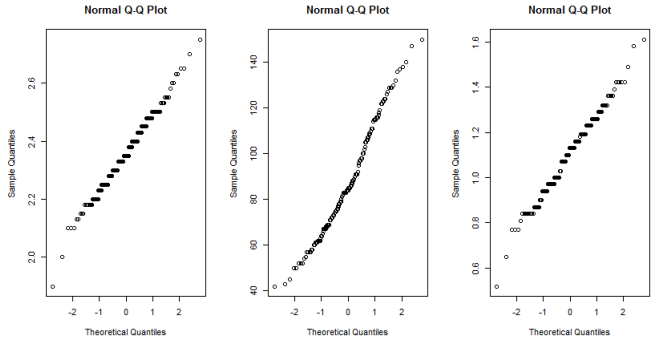


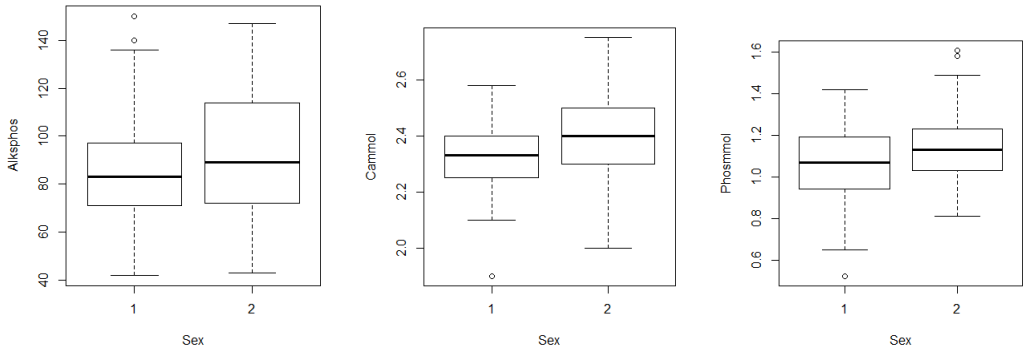
Figure 3: QQ plot after removing outliers, from left to right, Cammol, alksphos, phosmmol

Now from the Normal QQ-Plots of the transformed data Fig: 3, we can see that we indeed have normal, or at least reasonably normal distributions. It should be noted that the stair like pattern for cammol and phosmmol concentrations are likely due to rounding.

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 Fig4 the relation between the different concentrations and the sex groups are visually represented.

There seems to be a subtle difference between the sexes for all concentrations, with females having slightly higher concentrations across all substances. To get a better understanding of these relationships we perform the Kruskal-Wallis rank sum test, to determine whether our populations are identical. The results can be seen in table: 1



(a) Alksphos (b) Cammol (c) Phosmmol

Figure 4: Boxplots of the concentrations between sexes

Table 1: Kruskal-Wallis rank sum test

Variable	p
CAMMOL~SEX	0.001539
CAMMOL~LAB	1.433e-05
CAMMOL~AGEGRP	0.5451
ALKSPHOS~SEX	0.007631
ALKSPHOS~LAB	0.001325
ALKSPHOS~AGEGRP	0.6927
PHOSMMOL~SEX	0.007711
PHOSMMOL~LAB	0.159
PHOSMMOL~AGEGRP	0.006144

Table 2: The number of patients per age group

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

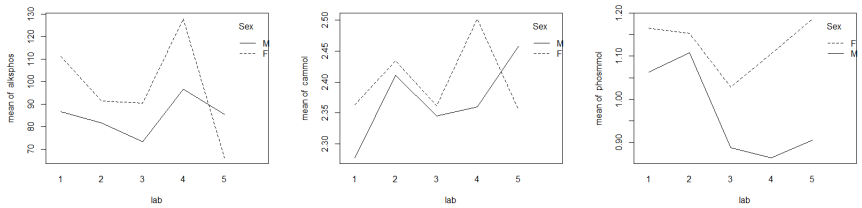
To see if sex, lab and age have an influence on the concentrations, the interaction plots in Fig5 and Fig6 are made. This shows that all three features have an influence on the measured concentrations. The age has an influence over the concentrations, which is good for the research question. But the lab and sex also influence the outcome, which means that these need to be taken into account with modeling the data.

Table 3: Anova, response variable: Cammol

	$SEX * LAB$	$SEX * AGEGRP$	$LAB * AGEGRP$
Sex	0.00012	0.0003835	
Lab	0.00015		1.491e-05
$AGEGRP$		0.9361766	0.4083
$Sex : Lab$	0.51331		
$Sex : AGEGRP$		0.1214165	
$Lab : AGEGRP$			0.6595

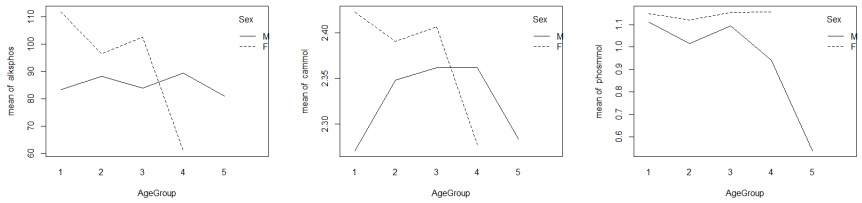
Table 4: p-values for 2-way ANOVA, response variable: Alsphos

	$SEX * LAB$	$SEX * AGEGRP$	$LAB * AGEGRP$
$Var1$	0.05781	0.07838	0.03025
$Var2$	0.01471	0.63871	0.72707
$Var1 : Var2$	0.53010	0.08933	0.58935



(a) Alksphos (b) Cammol (c) Phosmmol

Figure 5: Interaction plots of the sex and labs of the different concentrations

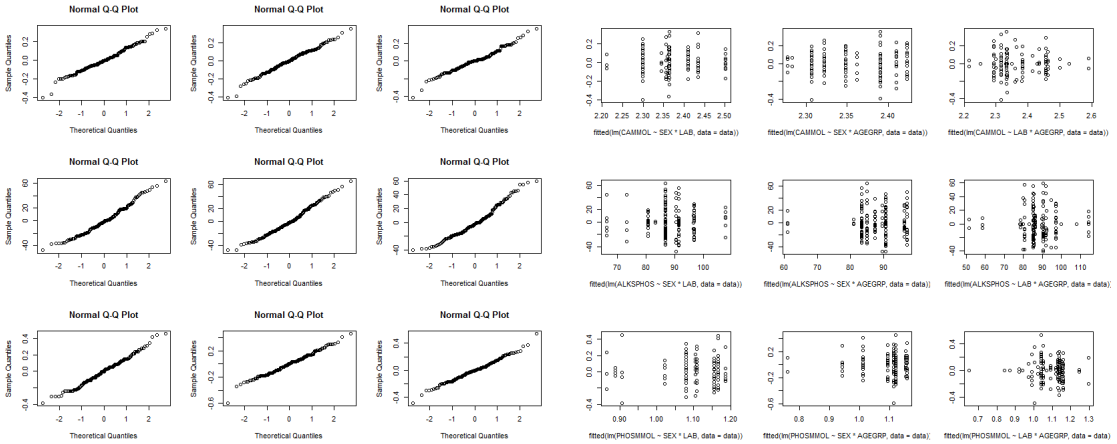


(a) Alksphos (b) Cammol (c) Phosmmol

Figure 6: Interaction plots of the sex and age groups of the different concentrations

Table 5: Anova, response variable: Phosmmol

	$SEX * LAB$	$SEX * AGEGRP$	$LAB * AGEGRP$
Sex	0.002991	0.003218	
Lab	0.006101		0.18946
$AGEGRP$		0.007732	0.02311
$Sex : Lab$	0.207392		
$Sex : AGEGRP$		0.402506	
$Lab : AGEGRP$			0.16321



(a) QQ plot of residuals for all 9 linear models (b) Plots of fitted residuals for all 9 linear models

Figure 7: Graphical tests for normality of residuals

3.3 Modeling

Modeling the data can be done in several ways. The first one tried is a step down linear regression approach for each of the concentrations. For Alksphos it follows the following steps:

- $ALKSPHOS \sim AGE + SEX + LAB + AGEGRP$
- $ALKSPHOS \sim AGE + SEX + LAB$
- $ALKSPHOS \sim SEX + LAB$
- $ALKSPHOS = 88.96 + 17.91 * SEX2 - 11.69 * LAB2 - 16.25 * LAB3 + 18.03 * LAB4 - 30.20 * LAB5$

For Cammol it follows the following steps:

- $CAMMOL \sim AGE + SEX + LAB + AGEGRP$
- $CAMMOL \sim AGE + SEX + LAB$
- $CAMMOL \sim SEX + LAB$
- $CAMMOL = 2.29 + 0.06 * SEX2 + 0.11 * LAB2 + 0.03 * LAB3 + 0.14 * LAB4 + 0.07 * LAB5$

For Phosmmol it follows the following steps:

- $PHOSMMOL \sim AGE + SEX + LAB + AGEGRP$
- $PHOSMMOL \sim AGE + SEX + LAB$
- $PHOSMMOL \sim AGE + SEX$
- $PHOSMMOL = 2.02 - 0.01 * AGE + 0.17 * SEX2$

As our research question is about what influence age has on the concentration and only Phosmmol has a model with age in it, this is not a good approach for this data.

4 Discussion

The data was realistic to work with, it contained some errors. At the start of the project, this was a challenge. To get all the data to correspond to realistic data. The good data was also supplied, but we decided against using it for the experience of working with error prone data. After the preprocessing the real data analysis could begin. The research question that was decided on was: What influence does age have on the given concentrations in the body? The reason to not only look at age and the three concentrations was to be able to see if more features had an influence on the concentrations. If this was not done the results would not be correct, because the features lab and sex also had an effect on the concentrations. In the models created it was clear that age did not have as big an influence as first thought. In two of the three models, age was not included. This means that age has a small influence on the measured concentrations, which is the answer to our research question.

5 R-Code

```
data = read.table('calcium.dat.txt', na.strings="_", header=TRUE)
# exploration
pairs(data)
# data preparation
data[!is.na(data$AGE) & !(data$AGE <= 110),]$AGE<-NA
data[!is.na(data$SEX) & !(data$SEX == 1|data$SEX ==2),]$SEX<-NA
data[!is.na(data$LAB) & !(data$LAB < 6),]$LAB<-NA
data[!is.na(data$PHOSMMOL) & !(data$PHOSMMOL < 2),]$PHOSMMOL<-NA
data$CAMMOL[!is.na(data$CAMMOL) & (data$CAMMOL > 10)] <-
          (data$CAMMOL[!is.na(data$CAMMOL) & (data$CAMMOL > 10)])/10
data$SEX <- as.factor(data$SEX)
data$LAB <- as.factor(data$LAB)
data$AGEGRP <- as.factor(data$AGEGRP)
# exploration
pairs(data)
summary(data$AGEGRP)
# data preperation, replace NA with mean
sex = data$SEX
sex[is.na(sex)]<-mean(na.omit(data$SEX))
lab = data$LAB
lab[is.na(lab)]<-mean(na.omit(data$LAB))
cammol = data$CAMMOL
cammol[is.na(cammol)]<-mean(na.omit(data$CAMMOL))
alksphos = data$ALKSPHOS
```

```
alksphos[is.na(alksphos)]<-mean(na.omit(data$ALKSPHOS))
phosmmol = data$PHOSMMOL
phosmmol[is.na(phosmmol)]<-mean(na.omit(data$PHOSMMOL))
# data preperation, make sex M and F instead of 1 and 2
Sex = as.character(sex)
Sex[sex == 1] = "M"
Sex[sex == 2] = "F"
Sex = as.factor(Sex)
AgeGroup = AGEGRP
# interaction plots
interaction.plot(lab,Sex,cammol)
interaction.plot(lab,Sex,alksphos)
interaction.plot(lab,Sex,phosmmol)
interaction.plot(AgeGroup,Sex,cammol)
interaction.plot(AgeGroup,Sex,alksphos)
interaction.plot(AgeGroup,Sex,phosmmol)
# step down linear regression
#ALKSPHOS
summary(lm(ALKSPHOS~AGE+SEX+LAB+AGEGRP,data=data))
summary(lm(ALKSPHOS~AGE+SEX+LAB,data=data))
summary(lm(ALKSPHOS~SEX+LAB,data=data))
#CAMMOL
summary(lm(CAMMOL~AGE+SEX+LAB+AGEGRP,data=data))
summary(lm(CAMMOL~AGE+SEX+LAB,data=data))
summary(lm(CAMMOL~SEX+LAB,data=data))
#PHOSMMOL
summary(lm(PHOSMMOL~AGE+SEX+LAB+AGEGRP,data=data))
summary(lm(PHOSMMOL~AGE+SEX+LAB,data=data))
summary(lm(PHOSMMOL~AGE+SEX,data=data))
```