STATISTICAL ANALYSIS OF URINARY BIOMARKERS FOR PANCREATIC CANCER

CS555 FINAL PROJECT SP'23

Kadam, Pooja Suresh
BUID - U31752372

CONTENTS

1. Introduction

Pancreatic cancer is an extremely deadly type of cancer. Once diagnosed, the five-year survival rate is less than 10%. However, if pancreatic cancer is caught early, the odds of surviving are much better. Unfortunately, many cases of pancreatic cancer show no symptoms until the cancer has spread throughout the body. A diagnostic test to identify people with pancreatic cancer could be enormously helpful.

In a paper by Silvana Debernardi and colleagues, published this year in the journal PLOS Medicine, a multi-national team of researchers sought to develop an accurate diagnostic test for the most common type of pancreatic cancer, called pancreatic ductal adenocarcinoma or PDAC. They gathered a series of biomarkers from the urine of three groups of patients:

- Healthy controls
- Patients with non-cancerous pancreatic conditions, like chronic pancreatitis
- Patients with pancreatic ductal adenocarcinoma

When possible, these patients were age- and sex-matched. The goal was to develop an accurate way to identify patients with pancreatic cancer.

2. Research Scenario and Research Questions

Research Scenario:

In this research paper, the Silvana and team had sought to develop a diagnostic test to detect the most common type of pancreatic cancer based on the biomarkers from different group of individuals.
The key features are four urinary biomarkers: creatinine, LYVE1, REG1B, and TFF1.

- Creatinine is a protein that is often used as an indicator of kidney function.
- LYVE1 is lymphatic vessel endothelial hyaluronan receptor 1, a protein that may play a role in tumor metastasis
- REG1B is a protein that may be associated with pancreas regeneration.
- TFF1 is trefoil factor 1, which may be related to regeneration and repair of the urinary tract.

The goal of this experiment is to predict the presence of pancreatic cancer before it is clinically diagnosed. It is an attempt to be able to detect the pancreatic cancer based on the presence and value of biomarkers in urine samples and then follow the clinical diagnostic methods to confirm the onset of disease at much earlier stages of its development.

Research Questions:

1. To Examine whether there is linear relationship between age of the patients and the biomarkers – REG1B, LYVE1 and TFF1.

2. To investigate whether the different categories of diagnosis, i.e., 1 = control (no pancreatic disease), 2 = benign hepatobiliary disease (119 of which are chronic pancreatitis); 3 = pancreatic cancer, are not due to age differences.

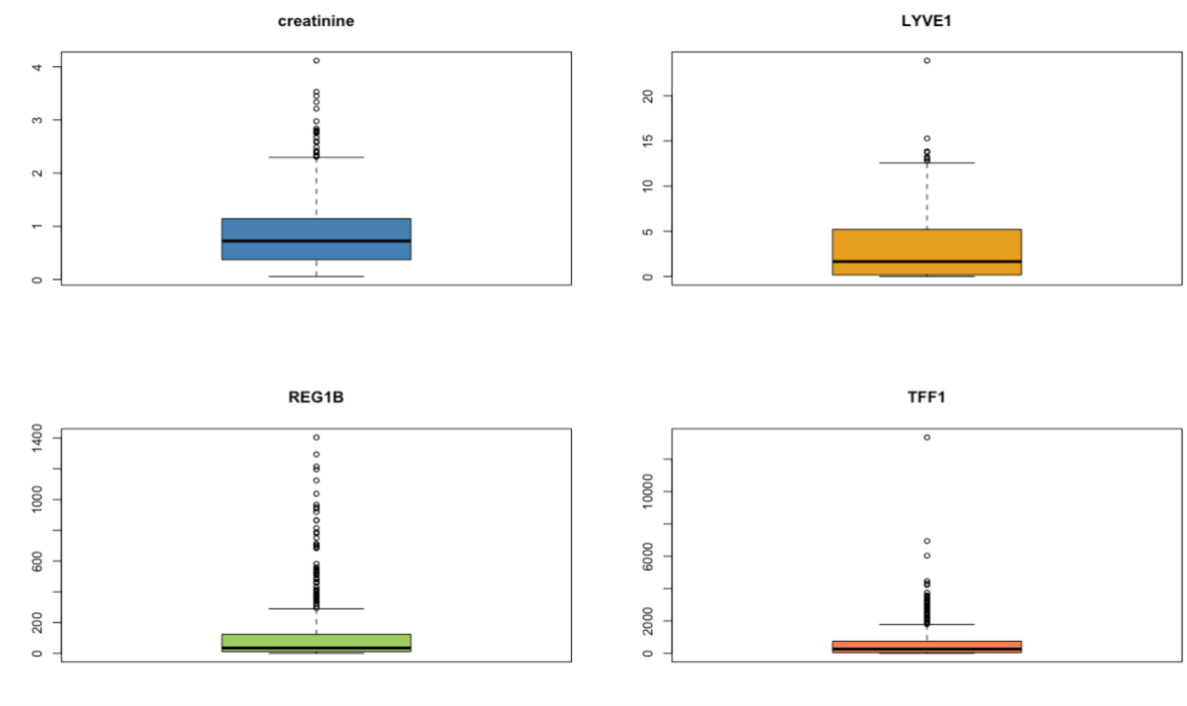
3. Dataset Overview

This dataset consists of 590 observations and 14 attributes. Below are the variables that will be used for investigating the research questions are as below:

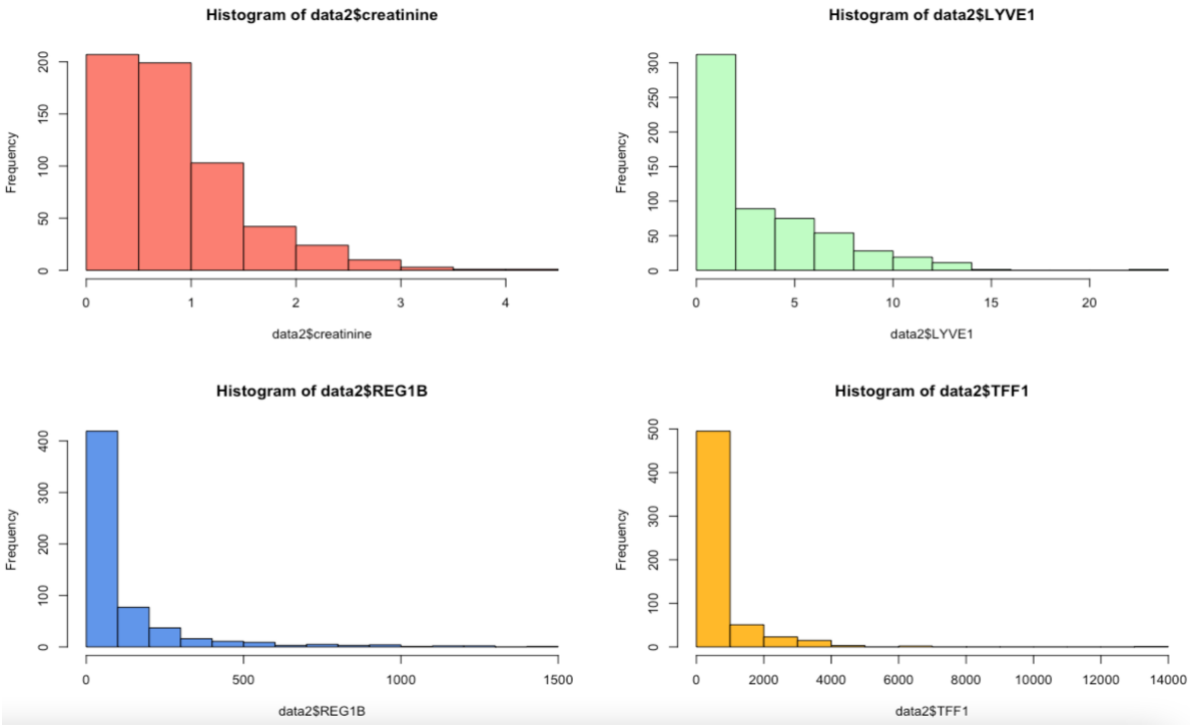
Column name	Details
sample_id	Unique string identifying each subject
patient_cohort	Cohort 1, previously used samples; Cohort 2, newly added samples
age	Age in years
sex	M = male, F = female
diagnosis	1 = control (no pancreatic disease), 2 = benign hepatobiliary disease (119 of which are chronic pancreatitis); 3 = Pancreatic ductal adenocarcinoma, i.e. pancreatic cancer
creatinine	Urinary biomarker of kidney function
LYVE1	Urinary levels of Lymphatic vessel endothelial hyaluronan receptor 1, a protein that may play a role in tumor metastasis
REG1B	Urinary levels of a protein that may be associated with pancreas regeneration.
TFF1	Urinary levels of Trefoil Factor 1, which may be related to regeneration and repair of the urinary tract

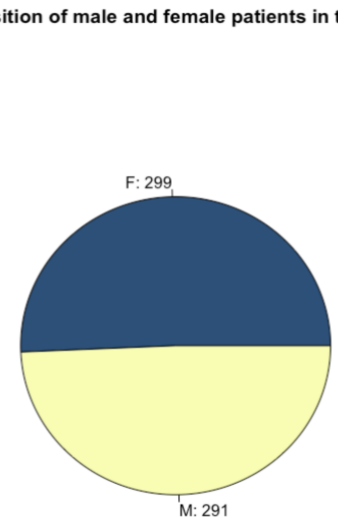
The main feature variables – creatinine, LYVE1, REG1B, TFF1 were visualized to check its distribution. Hence, boxplot and histograms were plotted of these variables.

The below histogram shows the presence of outliers in the dataset. Majority of outliers are found in the variable – creatinine, REG1B and TFF1. The mean and minimum value is also the same for the variables – TFF1 and REG1B.



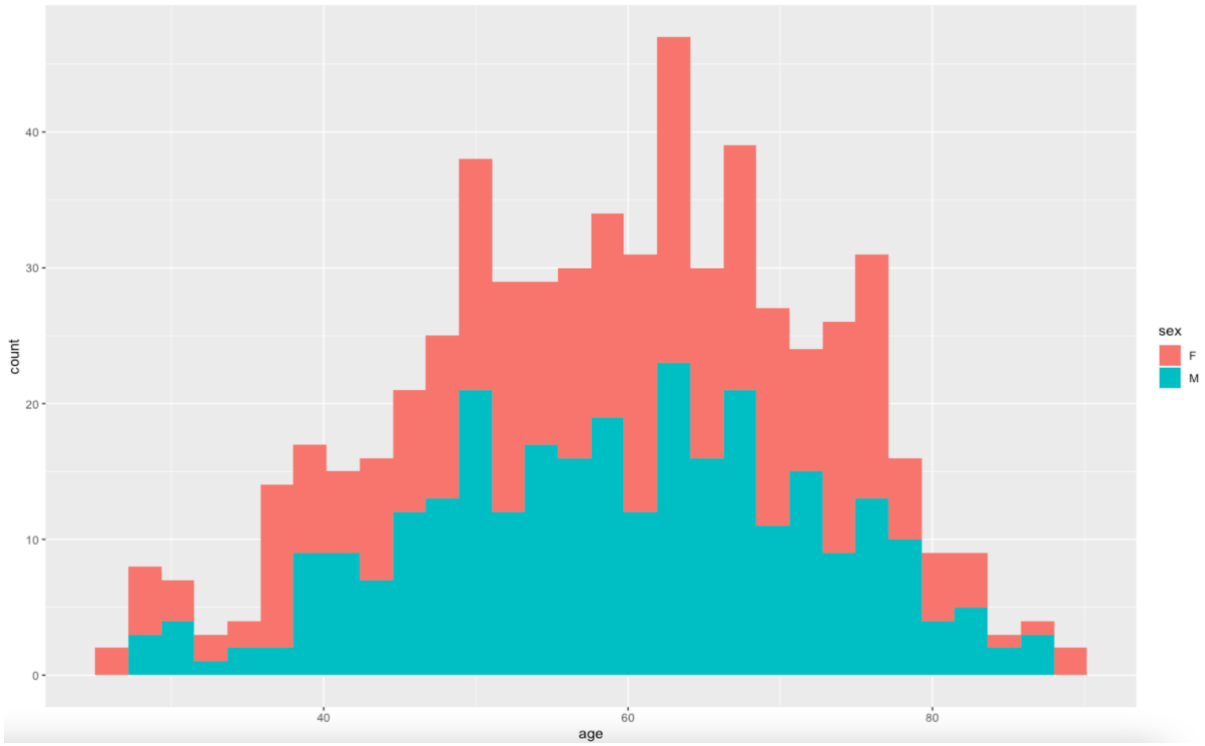
The below histogram shows the distribution of the 4 feature variables – creatinine, LYVE1, REG1B, TFF1. It is observed that the distribution is skewed towards right.





The above pie chart shows the composition of male and female in the study. The percent of male and female is same for the study.

Below, the distribution of age for male and female. It is observed that the age distribution of both male and female is normally distributed.



4. Statistical methods used

The statistical methods that were used for investigating the research questions are:

1. Multiple Linear Regression
    - To examine whether there is linear relationship between age of the patients and the biomarkers - REG1B, LYVE1 and TFF1, multiple regression analysis was performed. Initially, linearity was checked by plotting the scatterplots and finding the correlation coefficients.
    - Then, multiple linear regression model was fit for each biomarker with age.
    - Then, the model assumptions were checked including – normality, constant variance and linearity. This was done by examining the residuals of the model.
    - To check the linearity, r-squared value and F-statistic was calculated and p-value is determined with alpha as 0.05
    - Hypothesis for F-test –
      $\boldsymbol{H0}$: $\boldsymbol{\beta 1} = \boldsymbol{\beta 2} = \cdots = \boldsymbol{\beta k} = 0$
      $\boldsymbol{H1}$: $\boldsymbol{\beta i} \neq \boldsymbol{0}$ for at least one *i*

2. ANOVA
    - ANOVA (analysis of variance) was used to investigate whether the different categories of diagnosis (control, benign hepatobiliary disease, pancreatic cancer) are not due to age differences.
    - Checked that the age data for each diagnosis category are normally distributed and have equal variances by plotting histogram and boxplots.
    - Conducted an ANOVA with diagnosis as the independent variable and age as the dependent variable.
    - The null hypothesis is that there is no difference in age across diagnosis categories. The alternative hypothesis is that there is a difference.
      $H0$: $\mu 1 = \mu 2 = \mu 3$. $H1$: $\mu i \neq \mu j$ for some i and j. $\alpha = 0.05$
      $F = MSB / MSW$ with $k - 1 = 2$ and $n - k = 590 - 3 = 587$ degrees of freedom
    - F-distribution with 2, 587 degrees of freedom and associated with $\alpha = 0.05$. Decision Rule: Reject $H0$ if $F \geq 3.01$
    - If the p-value is less than the chosen significance level (usually 0.05), then you reject the null hypothesis and conclude that there is a difference in age across diagnosis categories.

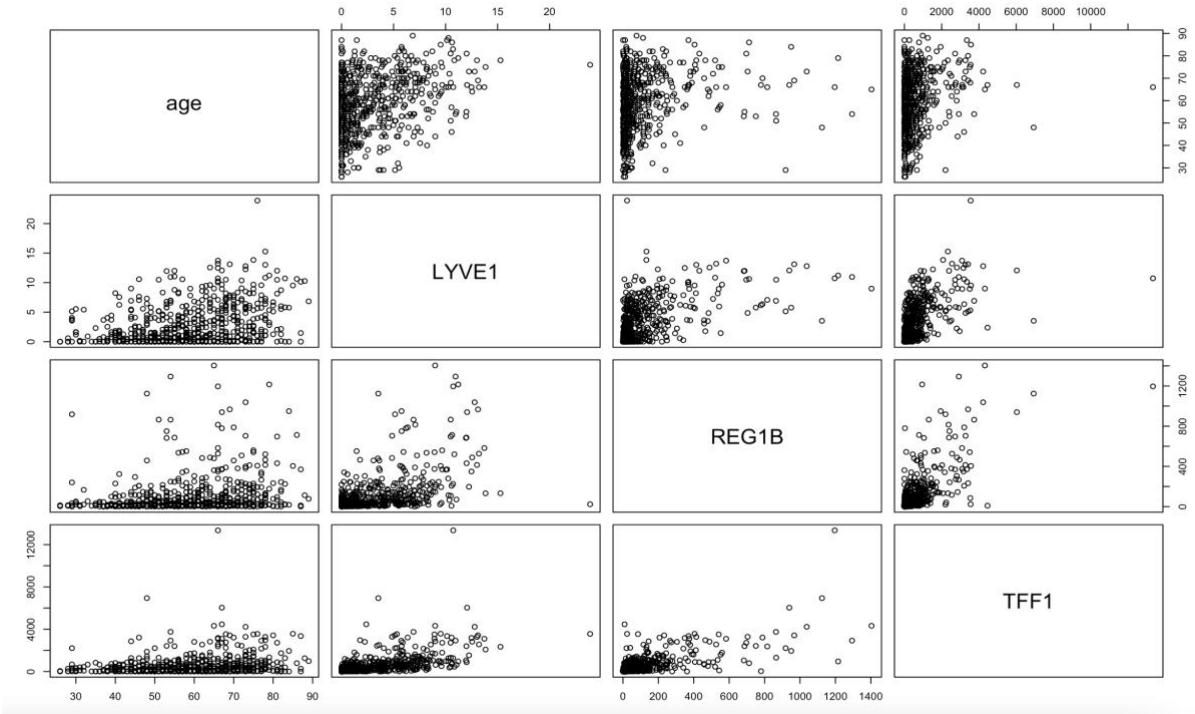
5. Results and Discussion

Multiple Linear Regression

Checking the linearity of the variables by plotting the scatterplots. Upon checking the scatterplots (please refer below image), it is observed that there is weak linear relation between the biomarker LYVE1 with age and REG1B with age. Whereas there is no linear relation between the biomarker TFF1 with age.

The relationship between variables is mostly random and is fairly positive for biomarker LYVE1 with age and REG1B with age (but mostly it shows there is no association).

Nevertheless, the relationship check was continued by applying multiple linear regression to quantify and confirm the observations done visually.



Correlation table:

	age	LYVE1	REG1B	TFF1
age	1	0.3298	0.2043	0.2079
LYVE1	0.3298	1	0.5431	0.5774
REG1B	0.2043	0.5431	1	0.6902
TFF1	0.2079	0.5774	0.6902	1

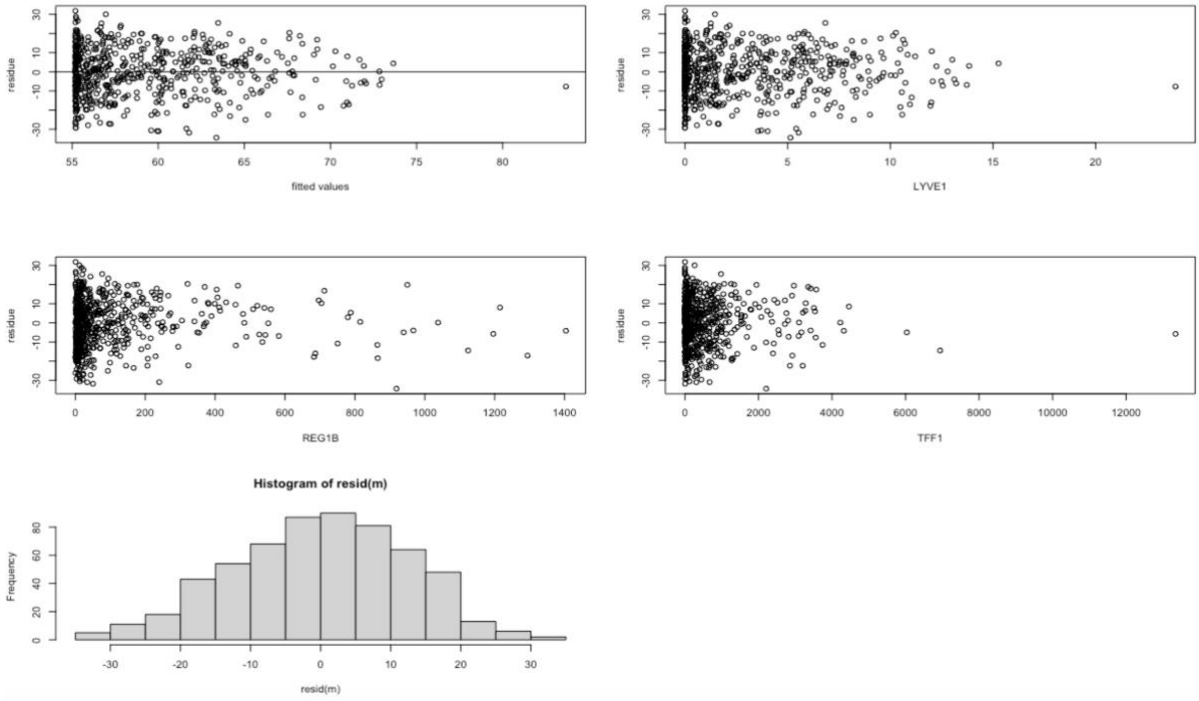
As per above correlation coefficients, there is no significant correlation between age and the biomarkers.

Fitting the model:

Upon fitting the model, below is the equation for MLR:

Y = 55.184 + 1.174 * LYVE1 + 0.002 * REG1B + 0.0001 * TFF1

Assessing model assumption by plotting residuals:



The above residual plot shows random scattering of points indicating that there is linear relation between the biomarkers. Although, there is no same amount of scattering in the residual plots, hence, there is no constant variance. The histogram of residuals is normally distributed as observed from the above histogram.

Calculating multiple R-squared:
Multiple R-squared: 0.1097

Decision rule for F-test: F-distribution with 3, 586 degrees of freedom and associated with α = 0.05
Decision Rule: Reject H_0 if F ≥ 2.62; Otherwise, do not reject H_0

Result:

F-statistic (on 3 and 586 DF)	24.07
p-value	1.061e-14

F-statistic is 24.07. Hence, reject the null hypothesis. We have significant evidence at the $\alpha = 0.05$ level that βLYVE1 $\neq 0$ *and/or* βTFF1 $\neq 0$ *and/or* βREG1B $\neq 0$

Here, $p < 0.05$

There is evidence of a linear association between annual age and biomarkers – LYVE1 and/or REG1B and/or TFF1

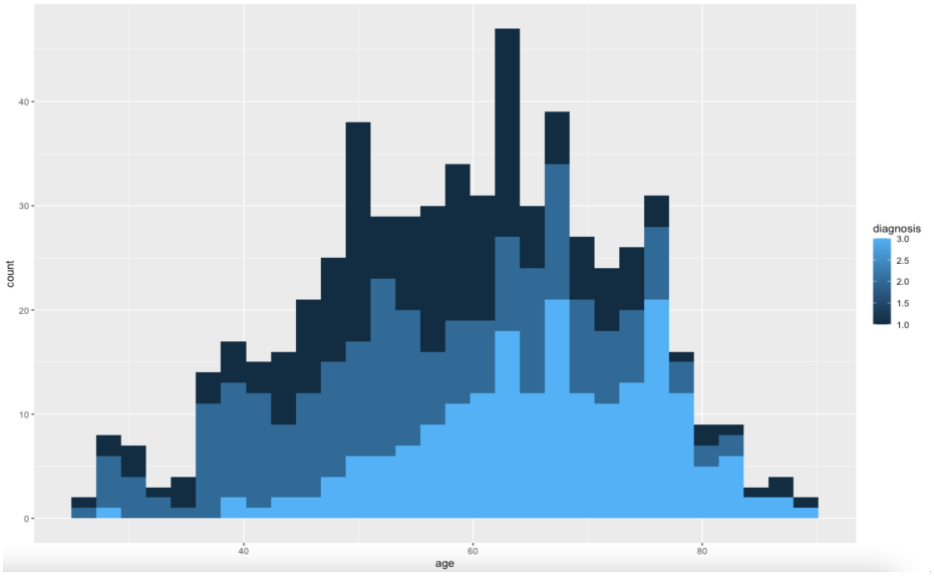
Confidence interval at alpha=0.05 is calculated and it is as below.
Confidence interval at alpha =0.05:

	2.50%	97.50%
(Intercept)	53.8361	56.5323
data3$LYVE1	0.8053	1.5425
data3$REG1B	-0.0052	0.0094
data3$TFF1	-0.0013	0.0016

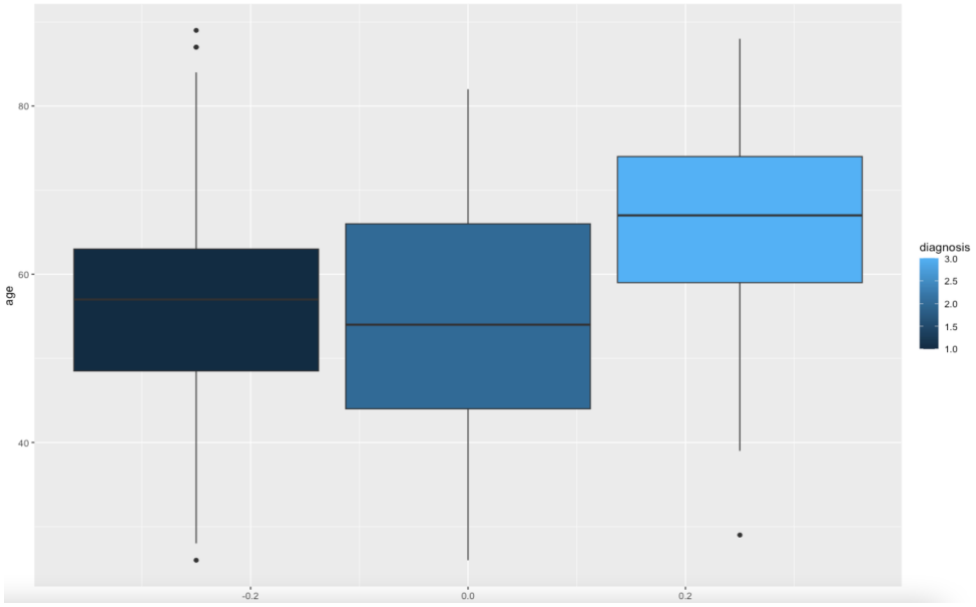
Only the biomarker LYVE1 has confidence interval that cannot be rejected, the remaining value for other 2 biomarkers do not show any association with the variable age.

2. ANOVA

The age data was checked for normality and equal variance by plotting histogram and boxplots. As per below histogram, the age across categories 1 and 2 is normally distributed whereas the age across category 3 is slightly skewed towards left.



The below boxplot shows the distribution of 3 diagnosis categories across age. There are few outliers for category 1 and 3.



Result:
The F-statistic was calculated to be 61.74 which is greater than 3.01.

We have significant evidence at the α = 0.05 that there is a difference in age across diagnosis categories - control, benign hepatobiliary disease, pancreatic cancer.

6. Conclusion

1. There is linear relation between the biomarker LYVE1 with age, i.e, with increase in age there increase in the presence of biomarker by 1.174 time per 1 year increase in age. The remaining 2 biomarkers are not linearly correlated with age and show no association.

2. Upon performing ANOVA it can be concluded that among the 3 categories mentioned in the column diagnosis – 1=control, 2=benign hepatobiliary disease, 3=pancreatic cancer, there is proof of the presence of different age groups across these categories.

7. Limitations

1. Lot of data was missing in the original dataset which may have had influence in the investigation.

2. The presence of outliers in the dataset may have also influenced the result of the investigation.

3. The assumptions made in the investigation may not be satisfactory and additional tests like pair wise comparisons might be needed to further investigate the relevance of the research questions.