Statistical Study of Backpain

s1899215

March 6, 2019

1 Introduction

The purpose of this study is to understand the medical conditions causing the pain. Two seperate datasets were gathered for this study. One dataset contains the general details of patients and a survey rating their pain at inclusion and at 12 month follow-up. This dataset is called "Covariates". The other dataset contains the levels of certain proteins taken from the blood sample of the patients. This dataset is called "Biomarkers". These two datasets are analyzed in order to further understand the relationship between biomarkers and pain.

2 Exploring Raw Data

2.1 Biomarkers Data

Table 1: Biomarkers raw data

Biomarker	IL-8	VEGF-A	OPG	TGF-beta-1	IL-6	CXCL9	CXCL1	IL-18	CSF-1
126-0weeks	7.63	11.51	10.20	8.83	3.52	6.16	9.45	7.91	8.41
126-6weeks	7.12	11.59	10.41	8.87	3.89	6.12	9.06	7.92	8.39
127-0weeks	6.93	10.92	10.30	6.59	2.73	6.14	7.31	7.95	8.40
127-6weeks	7.16	11.58	10.39	8.61	2.60	6.35	8.61	7.94	8.51
127-12months	6.87	11.13	10.25	7.44	3.92	6.15	8.79	7.94	8.46

Table 1 displays the first 5 rows of the Biomarkers data set. There are a total of 351 observations and 10 columns. Column 1 contains both the patient ID and the time(week) when the blood sample was tested. The other columns contain the protein levels from the blood sample.

Table 2: Missing data in Biomarker dataset

Biomarker	IL-8	VEGF-A	OPG	TGF-beta-1	IL-6	CXCL9	CXCL1	IL-18	CSF-1
4	3	4	4	4	4	4	4	4	4

Table 2 summarizes the missing data in each column. Since missing data can prevent us from carrying on to the next step of analysis, we have removed any rows that are missing data. Total of 4 rows were removed from Biomarkers dataset.

2.2 Covariates Data

Table 3: Covariates raw data

PatientID	Age	Sex (1=male, 2=female)	Smoker (1=yes, 2=no)	VAS-at-inclusion	Vas-12months
1	56	1	2	3.0	4.0
3	32	1	2	7.2	0.5
4	43	2	2	2.7	0.5
5	25	2	2	3.0	3.9
6	39	1	2	3.5	5.0

Table 3 displays the first 5 rows of the Covariate data set. There are a total of 118 observations and 6 columns. Each column is self-evident. The last two columns contain ratings of patient pain (VAS) at inclusion, and at 12 months. VAS 0 means no pain and VAS 10 is intolerable pain. This data set was missing two rows of data, thus removed

Table 4: Missing data in Covariate dataset

PatientID	Age	Sex (1=male, 2=female)	Smoker (1=yes, 2=no)	VAS-at-inclusion	Vas-12months
0	0	0	0	0	2

2.2.1 Covariate Data - Simple Analysis

First, let us examine the patients based upon: age, sex, and smokers. This will give us an insight on whether any of these variables are unequally represented in the data set. Unequal representation shapes the question we will formulate for the hypothesis testing. Based on figures 1 and 2, all these variables are almost equally represented.

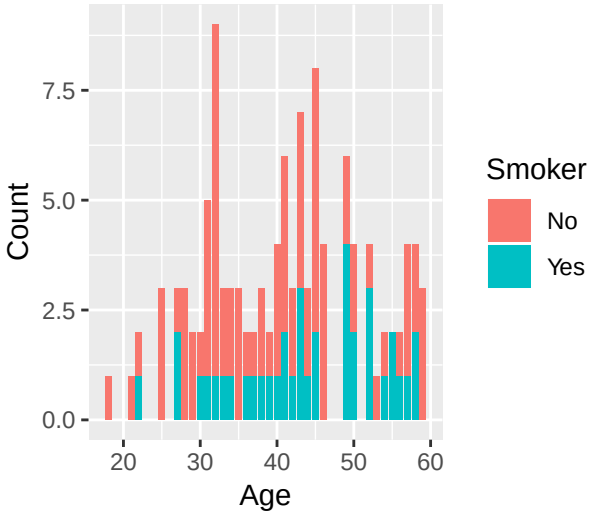

Fig 1: number of smokers in each Age group

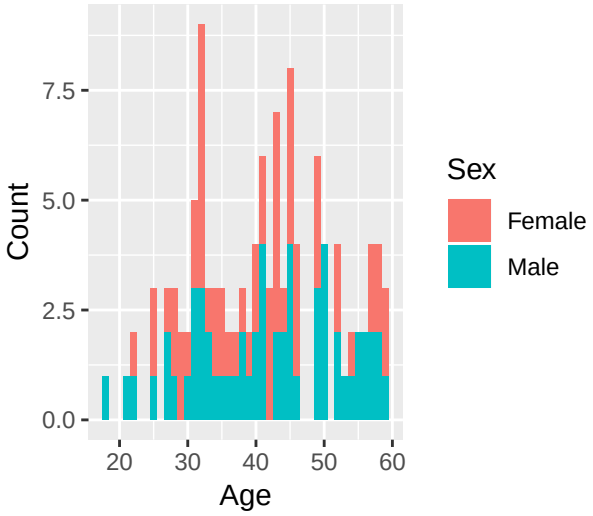

Fig 2: Different sex in each age group

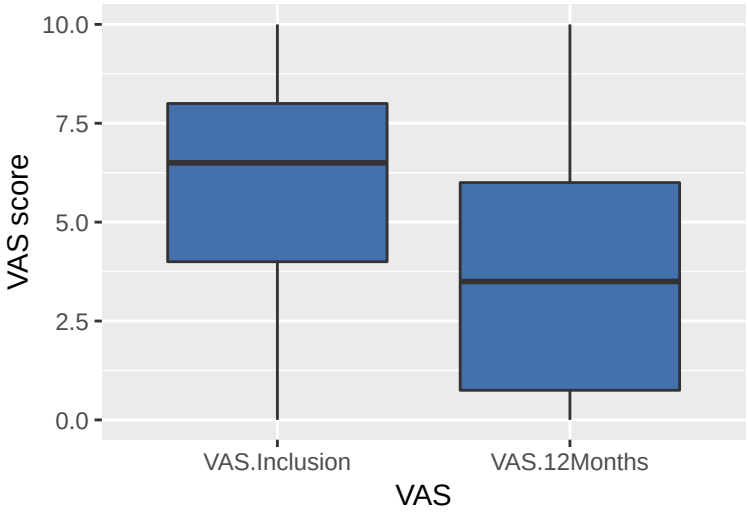

Fig 3: Comparing VAS at inclusion and at 12 months

Second, let us examine whether there was any changes in VAS. Based on the histograms provided in figure 3, the VAS at 12 months appears to be significantly lower than VAS at inclusion. We need to explore further to under how VAS changed based on sex and smokers. Both figure 4 and 5 clearly shows that there is a significant change in the VAS for each category.

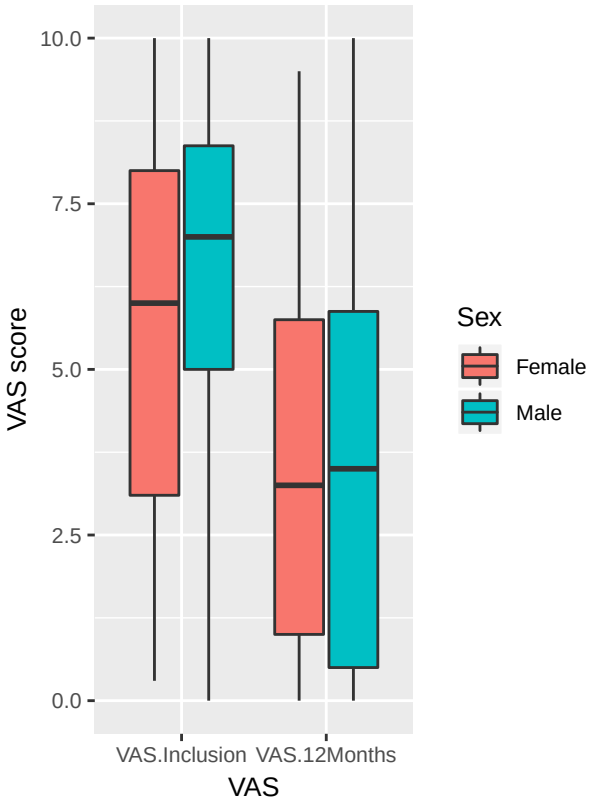

Fig 4: Comparing VAS on different sex

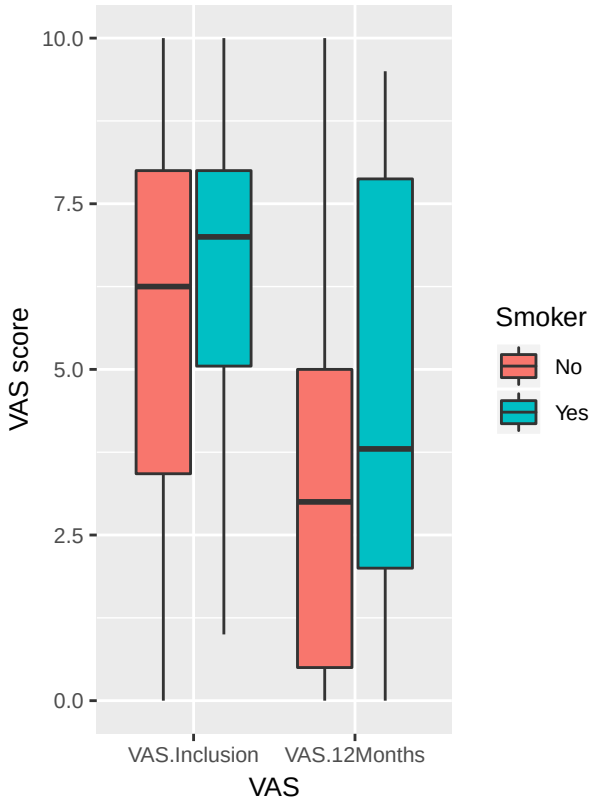

Fig 5: Comparing VAS on smokers

3 Cleaning/Manipulating Data

As per Table 1, we see that the time of observation and patient ID needs to be separated from column 1. Also, the 12 months needs to be changed to 52 weeks in order to maintain the same units. Table 5 displays the first 10 observations of the Biomarkers data set with the modification.

Table 5: Biomarkers data after seperating the patient ID and week of observation

PatientID	WeeksObs	IL-8	VEGF-A	OPG	TGF-beta-1	IL-6	CXCL9	CXCL1	IL-18	CSF-1
126	0	7.63	11.51	10.20	8.83	3.52	6.16	9.45	7.91	8.41
126	6	7.12	11.59	10.41	8.87	3.89	6.12	9.06	7.92	8.39
127	0	6.93	10.92	10.30	6.59	2.73	6.14	7.31	7.95	8.40
127	6	7.16	11.58	10.39	8.61	2.60	6.35	8.61	7.94	8.51
127	52	6.87	11.13	10.25	7.44	3.92	6.15	8.79	7.94	8.46
128	0	8.62	12.51	10.56	8.51	3.71	7.34	9.90	8.72	8.72
128	6	6.94	11.50	10.51	7.46	3.84	7.14	8.57	8.62	8.51
128	52	6.47	11.05	10.14	6.45	4.65	8.00	8.18	8.71	8.56
129	0	8.16	11.16	10.61	8.76	3.85	5.81	9.18	7.49	8.39
129	6	6.57	10.72	10.23	6.82	2.98	6.11	6.69	7.23	8.16

Box plots of each biomarker compared with each week of observation revealed that there are some changes to biomarker over the period of time. Figure 6 shows that almost all except “IL-6” have had changes between the weeks of observation.

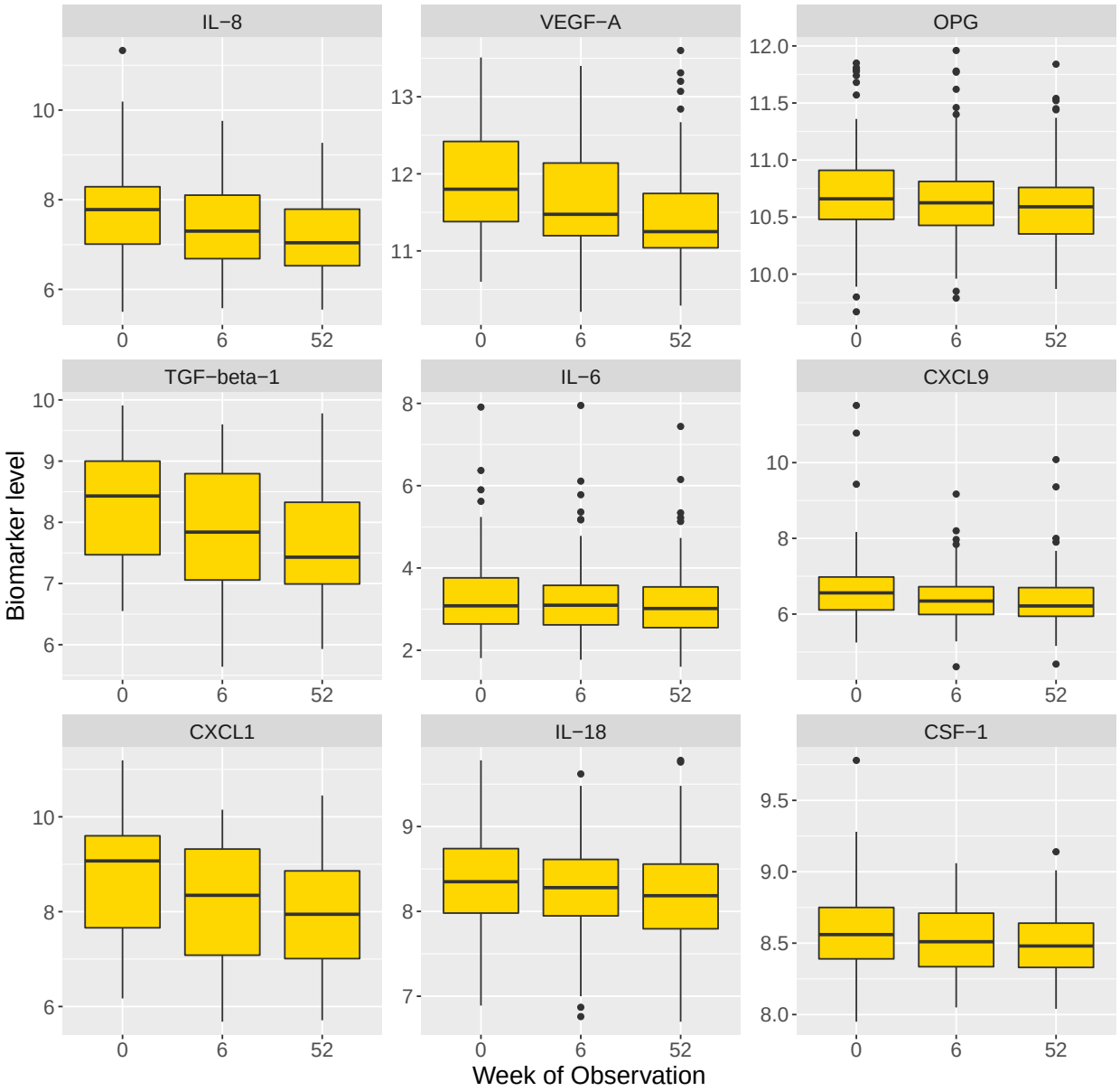

Fig 6: Biomarker value over the course of week of observation

4 Question 1: Statistical Hypothesis Testing

4.1 Choosing a Question

Simple analysis on the data sets revealed that there was a change in most biomarker levels and improvement of VAS with some patients. During this research we are interested in finding changes of biomarker levels among the patients whose VAS levels have been reduced. A total of 116 patients reported that their VAS at 12 months was lower than at inclusion. In order to conduct this research, both data sets were merged together. Any patient ID that did not contain all three different observations (Week 0, 6 and 52) were removed and only 79 patient's data remained for testing. Studying these patient's biomarker levels may reveal a better understanding of the relationship between biomarker levels and pain.

4.2 Formulate the Question as Hypotheses

- Hypothesis: Did the μ of each biomarker remain constant between week 0 and week 52.
- Random variables: Biomarker levels at each observation week.
- Distribution: Since these are continious random variables and the population μ is unknown, the distribution is assumed to be t-distribution. Furthermore, the biomarkers are measured twice during this test. Therefore these samples are paired.

4.3 Perform Suitable Hypothesis Test

Hypothesis:

- H0: $\mu_{week0} = \mu_{week52}$
- Ha: $\mu_{week0} \neq \mu_{week52}$ (two-tailed)

The p-values of paired sample t.test for biomarkers for week 0 and week 52 with $\alpha = 0.05$ are posted in table 6. Based on the results, p.values of IL-8, VEGF-A, OPG, TGF-beta-1, CXCL9, CXCL1, IL-18 and CSF-1 are below 0.05. Therefore these null hypotheses will be rejected. In other words, μ of these biomarkers have changed between week 0 and week 52. Therefore we can make an assumption that pain can be explained through these biomarkers.

Table 6: p.values for each hypothesis test

Biomarker	Week 0 and Week 52
IL-8	0.0000110
VEGF-A	0.0000034
OPG	0.0234495
TGF-beta-1	0.0000483
IL-6	0.3552379
CXCL9	0.0149160
CXCL1	0.0000329
IL-18	0.0209696
CSF-1	0.0021744

4.4.1 Potential Problem of Multiple Testing

When many hypotheses are tested, and each test has a specified Type I error probability, the probability that at least some Type I errors are committed increases, often sharply, with the number of hypotheses. This may have serious consequences if the set of conclusions must be evaluated as a whole [1]. The probability of making at least one type I error assuming that our tests are independent and that null hypotheses are true is given by formula[2] below:

Pr(at least one significant result | all k H0's are true)

= 1 - Pr(no significant results | all k H0's are true)

$= 1 - (1 - \alpha)^k$

Where $\alpha = 0.05$ and k = number of tests = 9

Therefore:

Pr(at least one significant result | all k H0's are true) = 0.6302494

The probability of making at least one type I error is 0.63

4.4.2 Bonferroni Correction

Bonferroni correction is a classical approach to limit the possibility of getting a Type I error when testing multiple hypotheses[3]. This is achieved by adjusting the α level by dividing the α by the number of tests(n_{tests}). In our test, the $\alpha_{adjusted}$ below:

$\alpha_{adjusted} = \frac{\alpha}{n_{tests}}$

Where $\alpha = 0.05$ and $n_{tests} = 9$

Therefore $\alpha_{adjusted} = 0.006$

Table 7: p.values for each hypothesis test

Biomarker	p.values of alpha	p.values of adjusted alpha
IL-8	0.0145484	0.0000110
VEGF-A	0.0847731	0.0000034
OPG	0.3927919	0.0234495
TGF-beta-1	0.0328849	0.0000483
IL-6	0.9194881	0.3552379
CXCL9	0.0731432	0.0149160
CXCL1	0.0161861	0.0000329
IL-18	0.2004710	0.0209696
CSF-1	0.0742271	0.0021744

We will now run the test again with the $\alpha_{adjusted}$ and reject null hypotheses when the p.value is below the $\alpha_{adjusted}$ value. The table 7 compares p.values from both hypotheses tests($\alpha = 0.5$ and $\alpha_{adjusted} = 0.006$). The results shows that with the Bonferroni Correction, L-8, VEGF-A, TGF-beta-1, CXCL1 and CSF-1 reject the null hypotheses. Using this results we can conclude that these biomarkers can be used to understand the patient's pain level.

5 Question 2: Regression modelling

In section 2.2.1 we were able to find that both data sets displayed some sort of correlation. We will now build a model to predict VAS at month 12 using biomarkers at inclusion and covariates as explanatory variables. The model will be trained on 80% of data and it will be tested on the 20% of the data(out of sample).

5.1 Model Description

- Explnatory variables : There are 13 explanatory variables in the data set. They are Age, Male, Smoker, VAS.Inclusion, IL-8, VEGF-A, OPG, TGF-beta-1, IL-6, CXCL9, CXCL1, IL-18 and CSF-1. Both "Sex" and "Smoker" variables are categorical and the rest are numerical variables. These variables are represented by $X_1, X_2, ...X_{13}$
- Response variable: VAS.12Months is a numerical variable which we will try to predict using the explanatory variables.

- General formula:

$$y = \beta_0 + \beta_{Age}X_1 + \beta_{Male}X_2 + \beta_{Smoker}X_3 + \beta_{VAS.inclusion}X_4 + \beta_{IL-8}X_5 + \beta_{VEGF-A}X_6 + \beta_{OPG}X_7 + \beta_{TGF-beta-1}X_8 + \beta_{IL-6}X_9 + \beta_{CXCL9}X_{10} + \beta_{CXCL1}X_{11} + \beta_{IL-18}X_{12} + \beta_{CSF-1}X_{13}$$

The β values from training data are presented in table 8. These will be substituted in the general formula above to predict the y value (VAS.12Months).

Table 8: Model description

	beta	Std. Error	t value	Pr(>|t|)
(Intercept)	11.9456227	11.3320639	1.0541436	0.2950716
Age	0.0085684	0.0340677	0.2515104	0.8020810
Male	0.0178643	0.6477933	0.0275771	0.9780699
Smoker	0.3628821	0.6974263	0.5203161	0.6043171
VAS.Inclusion	0.2548550	0.1234032	2.0652231	0.0422240
IL-8	0.7562968	0.6746334	1.1210486	0.2657068
VEGF-A	0.6833993	0.7688247	0.8888883	0.3767963
OPG	-2.5147641	0.8179296	-3.0745485	0.0029053
TGF-beta-1	-1.1061791	0.6985479	-1.5835409	0.1173447
IL-6	1.3184344	0.3333348	3.9552857	0.0001672
CXCL9	-0.0490029	0.3695518	-0.1326008	0.8948506
CXCL1	0.0947077	0.6113298	0.1549208	0.8772839
IL-18	-0.7844854	0.5622477	-1.3952666	0.1668955
CSF-1	1.5412070	1.4573327	1.0575533	0.2935235

5.2 Model Fit

- Residual Standard Error: 2.71
- R-Squared: 0.34
- Adjusted R-Squared: 0.23

According to the table 9, VAS.Inclusion, OPG and IL-6 displaying a statistically significant relationship at the $p < 0.05$ cut-off level. According to the R-squared value, the regression explains 34% of the variablity of the y[3]. Based on figure 7, the residuals do not display any sign of Normal distribution. Also according figure 8 the residuals seems to be independent. Figure 9 assures that there are no patterns between residuals and fitted values.

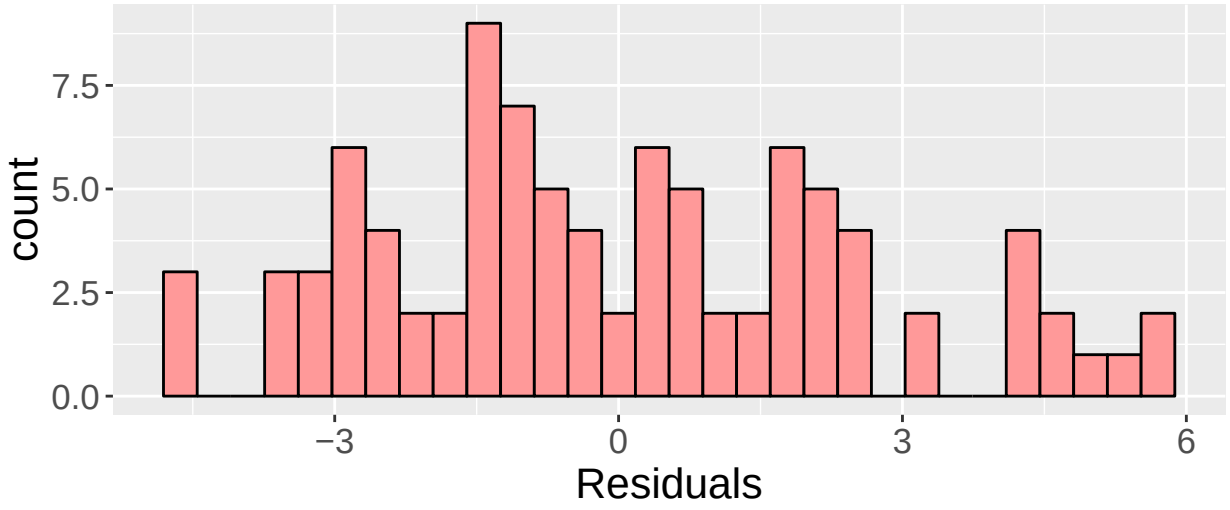

Fig 7: Residual histogram

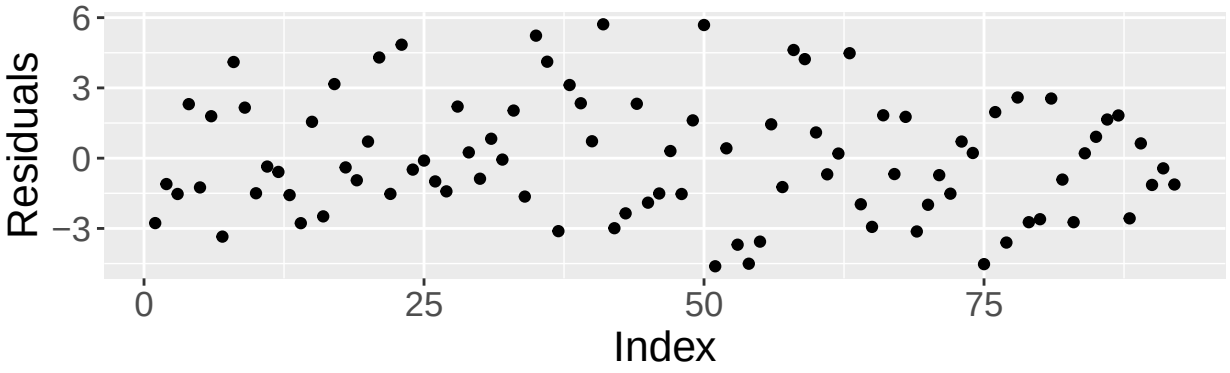

Fig 8: Residual scatter plot

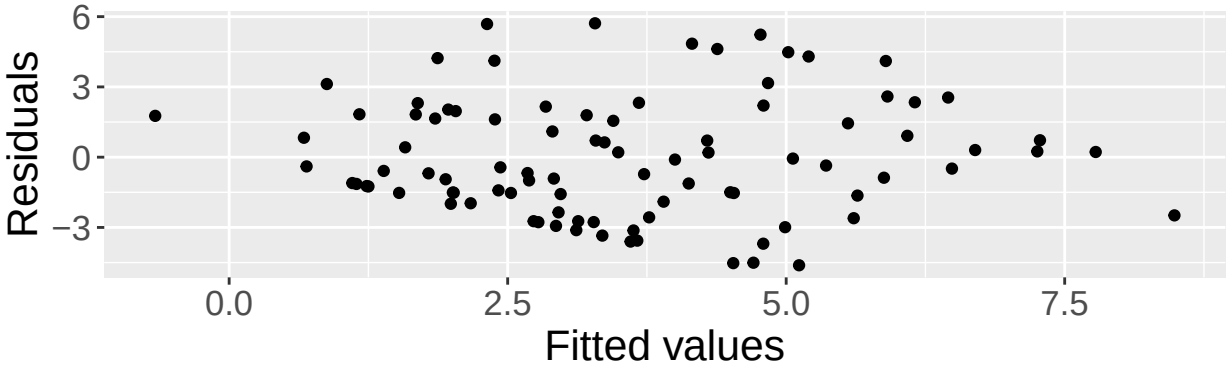

Fig 9: Residuals vs fitted values

We built the prediction model again using only VAS.Inclusion, OPG and IL-6 as the explanatory variables. Therefore our new equation would be:

$y = \beta_0 + \beta_{VAS.Inclusion} X_1 + \beta_{OPG} X_2 + \beta_{IL-6} X_3$

Model fit parameters are calculated again using the latest formula and results are displayed in table 9.

Table 9: Model description

	beta	Std. Error	t value	Pr(>|t|)
(Intercept)	18.2698149	7.2391916	2.523737	0.0134065
VAS.Inclusion	0.2674902	0.1069125	2.501954	0.0142001
OPG	-1.9905502	0.7016747	-2.836856	0.0056538
`IL-6`	1.4723280	0.2901399	5.074545	0.0000021

Summary of the model fit:

- Residual Standard Error: 2.69
- R-Squared: 0.27
- Adjusted R-Squared: 0.24

Based on the results, the new Residual Standard Error and Adjusted R-Squared displayed some improvement, but not significant. The latest R-Squared is lower than the previously calculated R-Squared value. R-Squared value increases when we add more explanatory variables. Since the new equation contains less explanatory variables compared to the previous formula, the R-Squared value is reduced.

5.3 Testing the Model (out of sample)

VAS. Inclusion, OPG and IL-6 will be used as explanatory variables to test the model as these variables displayed statistical significance and improvement in model fit.

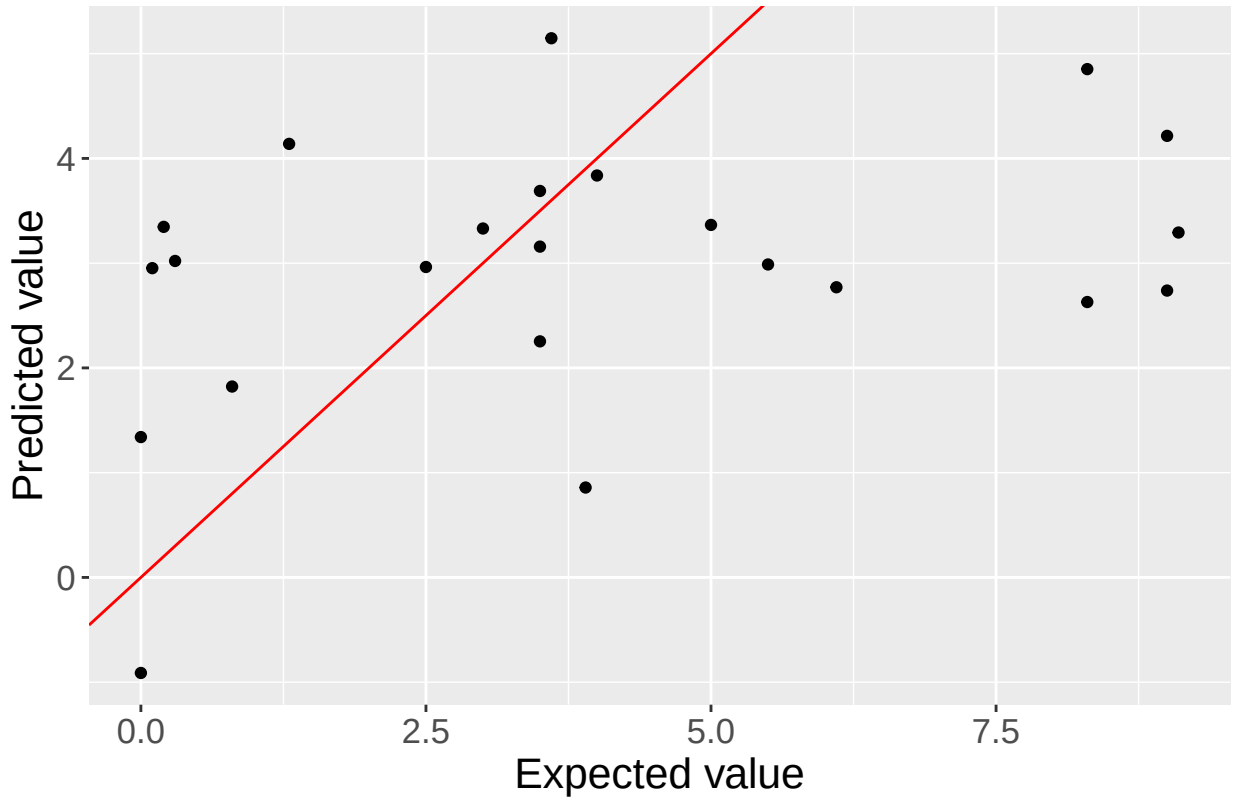

Fig 10: Predicted value vs expected value

The regression model appears to be a weak predictor. According to figure 10, the points are dispersed and does not follow a pattern. Ideally the scatter plot should be closely following the red line. Residual Sum of Squares (RSS) is 210.83. If we were to run this test with all the variables, the RSS would result in 197.25.

5.4 Conclusion

In conclusion, the linear model we derived in section 5.2 is not useful to predict VAS at 12 months. Linear model with all 13 variables had a lower RSS compared to the linear model with only statistically significant variables. The statistical significance of each explanatory variable changes significantly depending on the selected training data. Perhaps more training data might give us a model that can predict with better accuracy. It is also possible that these variables could have no linear relationship.

6 Final Remarks

In this paper, we were able to identify the biomarkers that are associated with patient pain levels. Since the probability of committing type I error increases with multiple testing, the biomarkers were identified using Bonferroni Correction method. Biomarkers IL-8, VEGF-A, TGF-beta-1, CXCL1 and CSF-1 displayed statistical significance in the test. Therefore, it can be concluded that these biomarkers can be tested to understand the pain level of a patient.

The linear regression prediction model was weaker in prediction. VAS. Inclusion, OPG, and IL-6 were used as the explanatory variables to build the linear regression model. Even though these variables displayed statistical significance, their RSS and Adjusted R-Squared values were not far off from the model with all 13 variables. It maybe possible that these explanatory variables might have non-linear relationship, or the sample size of training data was not adequate to build a better model. The current results may not be favourable, but future research may include non-linear prediction models to provide more accuracy.

References

[1] Shaffer J. “Multiple Hypothesis Testing” Annual Rev. Physchol, 46 (1995), 561-584.

[2] Bangdiwala S. “Multiple Hypothesis Testing” International Journal of Injury Control and Safety Promotion, 24:1 (2017), 140-142.

[3] Devore J. and Berk K. “Modern Mathematical Statistics With Applications” Second edition, 2012, 707-710.

[4] Folsentein J. “Lecture 9: Multiple tests, Bonferroni correction, FDR” lecture notes, Washington University, delivered 2011.

[5] Goldman M. “Statistics for Bioinformatic” lecture notes, STATC141, UC Berkeley, delieverd 2008 Spring.

[6] Sedgwick P. “Multiple hypothesis testing and Bonferroni’s correction” BMJ, 349 (2014), g6284.

[7] Akey J. “Lecture 10: Multiple Testing” lecture notes, Washington University, delievered 2008 April.

[8] Fan Z. “Lecture 11: Testing Multiple Hypotheses” lecture notes, STATS 200, Stanford University, delievered 2016 Fall.

[9] Austin S, Dialsingh I and Altman N. “Mutiple Hypothesis Testing: A Review” Journal of the Indian Society of Agricultural Statistics, 68 (2014), 303-314.

[10] Teetor P. “R Cookbook” First edition, 2011

[11] Chang W. “R Graphics Cookbook” First edition, 2012

Appendix

```
#name the data files
biomarkerFile <- "biomarkers.xlsx"
covariateFile <- "covariates.xlsx"

#Set file paths
biomarkerFilePath <- paste0(getwd(),"/data/",biomarkerFile)
covariateFilePath <- paste0(getwd(),"/data/",covariateFile)

#Read the files
biomarkerData <- read_xlsx(biomarkerFilePath)
covariateData <- read_xlsx(covariateFilePath)

#Convert to data.table
biomarkerData <- as.data.table(biomarkerData)
covariateData <- as.data.table(covariateData)
```

```
#Number of observations in raw file
BioMarkerNumOfObservationsRaw <- nrow(biomarkerData)

#Sample display of the raw file
kable(biomarkerData[1:5,], caption = " Biomarker raw data")
```

```
#Table with NAs in the data
kable(biomarkerData[, lapply(.SD, function(x) sum(is.na(x))),
                    .SDcols = 1:ncol(biomarkerData)],
      caption = "Missing data in Biomarker dataset")

#Remove rows that contains NA
biomarkerData <- na.omit(biomarkerData)
```

```
#Number of observations from Covariate data
CovNumOfObservationsRaw <- nrow(covariateData)

#Sample display of covariate data
kable(covariateData[1:5,], caption = " Covariates raw data")
```

```
#Summary table of NAs in the data
kable(covariateData[,lapply(.SD, function(x) sum(is.na(x))),
                    .SDcols = 1:ncol(covariateData)],
      caption = "Missing data in Covariate dataset")

#Remove the rows that contains NAs
covariateData <- na.omit(covariateData)

#Convert PatientID into
covariateData[,PatientID := factor(PatientID)]
```

```
#Change the column names of the data
setnames(covariateData, c("Sex (1=male, 2=female)",
                          "Smoker (1=yes, 2=no)",
                          "VAS-at-inclusion",
                          "Vas-12months"),
         c("Sex", "Smoker", "VAS.Inclusion", "VAS.12Months"))
```

```r
#Saving the names of the columns
covariateColNames <- names(covariateData)

#Change the Sex and Smoker columns from 1 and 2 into corresponding categorical variable
covariateData[,Sex := factor(ifelse(Sex == 1, "Male", "Female"))]
covariateData[,Smoker := factor(ifelse(Smoker == 1, "Yes", "No"))]

#Reshape the data in order to plot easily
covariateDataMelt <- melt(covariateData, id.vars = c("PatientID", "Age",
                                                     "Sex", "Smoker"))
```

```r
#Plot number of smokers in each age group
g1 <- ggplot(data = covariateData, aes(x = Age)) +
  geom_bar(aes(fill = Smoker)) +
  xlab("Age") +
  ylab("Count") +
  labs(caption= "Fig 1: number of smokers in each Age group") +
  theme(plot.caption = element_text(hjust = 0))

g2 <- ggplot(data = covariateData, aes(x = Age)) +
  geom_bar(aes(fill = Sex)) +
  xlab("Age") +
  ylab("Count") +
  labs(caption= "Fig 2: Different sex in each age group") +
  theme(plot.caption = element_text(hjust = 0))

grid.arrange(g1,g2,ncol = 2)
```

```r
#Box plot of VAS at inclusion and VAS at months
ggplot(covariateDataMelt, aes(x = variable, y = value)) +
  geom_boxplot(fill = "#4271AE") +
  xlab("VAS") +
  ylab("VAS score") +
  labs(caption= "Fig 3: Comparing VAS at inclusion and at 12 months") +
  theme(plot.caption = element_text(hjust = 0))
```

```r
#VAS comparrison for sex
VAS.ComparrisonSexGraph <- ggplot(covariateDataMelt,
                                  aes(x = variable, y = value, fill = Sex)) +
  geom_boxplot() +
  xlab("VAS") +
  ylab("VAS score") +
  labs(caption= "Fig 4: Comparing VAS on different sex") +
  theme(plot.caption = element_text(hjust = 0), text = element_text(size=10))

#VAS comparrison for smokers
VAS.ComparrisonSmokersGraph <- ggplot(covariateDataMelt, aes(x = variable,
                                                             y = value,
                                                             fill = Smoker)) +
  geom_boxplot() +
  xlab("VAS") +
  ylab("VAS score") +
  labs(caption= "Fig 5: Comparing VAS on smokers") +
  theme(plot.caption = element_text(hjust = 0), text = element_text(size=10))
```

```r
#Plpot the last two graphs side by side
grid.arrange(VAS.ComparrisonSexGraph,VAS.ComparrisonSmokersGraph,ncol = 2)
```

```r
#Extract the biomarkers
biomarkerNames <- colnames(biomarkerData)[2:ncol(biomarkerData)]
#Extract the PatientID from Biomarker Column. Every character before the "-"
biomarkerData[,PatientID :=as.factor(sub("\\-.*", "", biomarkerData$Biomarker))]
#Extracting the characters right of "-" on Biomarker column
extractedTime <- sub(".*-", "", biomarkerData$Biomarker)
#Extract the number from the previous string
extractedTime <- gsub("[^0-9.]", "",  extractedTime)
#Replace 12 months with 52 weeks
extractedTime[which(extractedTime == 12)] <- 52
#Create a new column called WeeksObs which contains the extracted time
biomarkerData[,WeeksObs := factor(extractedTime, levels = c("0","6","52"))]
#Re-arrange the columns
biomarkerData <- setcolorder(biomarkerData, c("PatientID", "WeeksObs", biomarkerNames))
#Remove biomarker column
biomarkerData[,Biomarker := NULL]
#Table of the biomarkerData
kable(head(biomarkerData,10),
      caption = "Biomarker data after seperating the patient ID and week of observation")
```

```r
#Reshape the data for ease of plotting
biomarkerDataMelt <- melt(biomarkerData, id.vars = c("PatientID", "WeeksObs"))
#Biomarker box plot of each biomarker at each week
ggplot(data = biomarkerDataMelt, aes(x = WeeksObs, y = value)) +
  geom_boxplot(fill = "#FFD700") +
  facet_wrap(~variable, scales = "free") +
  xlab("Week of Observation") +
  ylab("Biomarker level") +
  labs(caption= "Fig 6: Biomarker value over the course of week of observation") +
  theme(plot.caption = element_text(hjust = 0), text = element_text(size=16))
```

```r
#Create a new column to get the difference of VAS at inclusion and 12 mos
covariateData[,PainDiff := VAS.Inclusion - VAS.12Months]

VAS.improvement <- covariateData[,Improved := ifelse(PainDiff > 0,
                                                     "Pain improved",
                                                     "Pain did not improve")]

painImproved <- length(VAS.improvement$PainDiff > 0)
```

```r
#Left Join on PatientID
mergedCovBioDT <- covariateData[biomarkerData, on = .(PatientID)]
#Count number of observation for each PatientID
mergedCovBioDT <- mergedCovBioDT[,Count := .N,by = PatientID]
#Exclude any PatientID that does not contain 3 different weeks of observation
mergedCovBioDT <- mergedCovBioDT[Count == 3 & Improved == "Pain improved",]
#Remove the column Count and Improved
mergedCovBioDT[,`:=`(Count = NULL, Improved = NULL)]
#Number of patients to conduct hypothesis test
numOfPatientsForTest <- length(unique(mergedCovBioDT$PatientID))
```

```r
#Seperate biomarkers according to week of observation

#Seperate week 0
Week0 <- mergedCovBioDT[WeeksObs == 0]
#Remove unwanted columns
Week0 <- Week0[,`:=`(PatientID = NULL, Age = NULL, Sex = NULL, Smoker = NULL,
                     VAS.Inclusion = NULL, VAS.12Months = NULL,
                     PainDiff = NULL, WeeksObs = NULL)]

#Seperate week 6
Week6 <- mergedCovBioDT[WeeksObs == 6]
#Remove unwanted columns
Week6 <- Week6[,`:=`(PatientID = NULL, Age = NULL, Sex = NULL, Smoker = NULL,
                     VAS.Inclusion = NULL, VAS.12Months = NULL,
                     PainDiff = NULL, WeeksObs = NULL)]

#Seperate week 52
Week52 <- mergedCovBioDT[WeeksObs == 52]
#Remove unwanted columns
Week52 <- Week52[,`:=`(PatientID = NULL, Age = NULL, Sex = NULL, Smoker = NULL,
                       VAS.Inclusion = NULL, VAS.12Months = NULL,
                       PainDiff = NULL, WeeksObs = NULL)]
```

```r
t.test.Biomarkers <- function(data1,data2, biomarker,alpha = 0.05,...) {
  #Extract the data of the biomarker from both data sets as a vector
  data1 <- data1[, get(biomarker)]
  data2 <- data2[, get(biomarker)]
  #Calculate the confidence interval
  confInterval <- 1-alpha
  #Run the t-test
  res <- t.test(data1, data2, alternative = "two.sided",
                paired = TRUE, conf.level = confInterval)
  #Extract the p.value
  pValue <- res$p.value
  #return p.value
  return(pValue)
}

alpha <- 0.05
#Running t-test for different combination of weeks for each biomarker
list1 <- lapply(biomarkerNames, FUN = t.test.Biomarkers,
                data1 = Week0, data2 = Week52, alpha = alpha)

#Amulgamate p.values with the biomarker
p.values <- data.table(Biomarker = biomarkerNames, C1 = do.call(rbind,list1))

#Remaning the columns
setnames(p.values, c("C1.V1"), c("Week 0 and Week 52"))

#Biomarker that have p.value lower than alpha
rejectedBiomarkersTest1 <- p.values[`Week 0 and Week 52` < alpha,Biomarker]
```

```r
#Display the table
kable(p.values, caption = "p.values for each hypothesis test")
```

```r
#Calculating the pr of at least 1 type I error
Pr <- (1- alpha)^nrow(p.values)
```

```r
alphaAdjusted = alpha/nrow(p.values)
```

```r
#Saving the p.values from test into a new data.frame
p.valuesTest1 <- as.data.frame(p.values)

#Running t-test for different combination of weeks for each biomarker
list1 <- lapply(biomarkerNames, FUN = t.test.Biomarkers,
                data1 = Week0, data2 = Week6, alpha = alphaAdjusted)

#Amulgamate p.values with the biomarker
p.values <- data.table(Biomarker = biomarkerNames,
                       C1 = do.call(rbind,list1))


#Join p.values from alpha and adjusted alpha
p.values <- p.values[p.valuesTest1, on = .(Biomarker)]

#Remaning the columns
setnames(p.values, 2:3, c("p.values of alpha", "p.values of adjusted alpha"))

#Display the table
kable(p.values, caption = "p.values for each hypothesis test")

#Extract the biomarkers that were rejected
rejectedBiomarkersTest2 <- p.values[`p.values of adjusted alpha` < alphaAdjusted,Biomarker,]
```

```r
#Remove columns that are not required for regression modeling
covariateData[,`:=` (PainDiff = NULL, Improved = NULL,
                     Male = factor(ifelse(Sex == "Male", 1,0)),
                     Smoker = factor(ifelse(Smoker == "Yes",1,0)), Sex = NULL)]


#Rearrange the columns
setcolorder(covariateData, c("PatientID", "Age", "Male",
                             "Smoker","VAS.Inclusion", "VAS.12Months"))

#Left join data
mergedData <- covariateData[biomarkerData, on = .(PatientID)]

#Extract only week 0 of the data
mergedData <- mergedData[WeeksObs == 0]

#Remove unwanted columns
mergedData[,`:=` (WeeksObs = NULL, PatientID = NULL)]

#Remove NAs
mergedData <- mergedData[complete.cases(mergedData),]
#Reaggraning the columns
colNames <- colnames(mergedData)
colNames <- colNames[-which(colNames == "VAS.12Months")]
colNames <- c(colNames, "VAS.12Months")
```

```r
#Reaggrange the data.table
setcolorder(mergedData, colNames)
```

```r
#Splitting the data into train and test data set
set.seed(5)
train_i <- sample(1:NROW(mergedData), size = round(0.8*NROW(mergedData)),
                  replace = FALSE)

y_trn <- mergedData[train_i, ]

y_test <- mergedData[-train_i, ]

#Linear modeling
mod1 <- lm(`VAS.12Months`~., y_trn)

#Covering the parameters into a table
fm1.table <- xtable(mod1)
#Renaming some rows - happens to be a bug which is adding 1 infront of categorical variables
rownames(fm1.table)[3:4] <- c("Male", "Smoker")
#Rename the first column
colnames(fm1.table)[1] <- "beta"
#Print the table
kable(fm1.table, caption = "Model description")
```

```r
#Extract residuals
res <- as.data.frame(mod1$residuals)
colnames(res) <- "Residuals"

#Plot histogram of the residuals
ggplot(res, aes(x = Residuals)) +
geom_histogram(fill = "#FF9999", colour="black") +
labs(caption= "Fig 7: Residual histogram") +
theme(plot.caption = element_text(hjust = 0), text = element_text(size=16))

#Scatter plot of residuals
g1 <- ggplot(res, aes(x = 1:nrow(res),y = Residuals)) +
geom_point() +
xlab("Index") +
ylab("Residuals")
labs(caption= "Fig 8: Residual scatter plot") +
theme(plot.caption = element_text(hjust = 0), text = element_text(size=16))

#Extract residuals and fitted values
df <- augment(mod1)

#Plot residuals vs fitted values
g2 <- ggplot(df, aes(x = .fitted, y = .resid)) +
geom_point() +
xlab("Fitted values") +
ylab("Residuals") +
labs(caption= "Fig 9: Residuals vs fitted values") +
theme(plot.caption = element_text(hjust = 0), text = element_text(size=16))

#Plot two graphs together
```

```r
grid.arrange(g1,g2, nrow = 2)
```

```r
#Extracting the columns that have statistical significance
y_trainData <- y_trn[,c("VAS.Inclusion", "OPG","IL-6", "VAS.12Months")]

#Linear modeling
mod2 <- lm(`VAS.12Months`~., y_trainData)

#Covert mod2 to a table
fm2.table <- xtable(mod2)

#Rename the first column
colnames(fm2.table)[1] <- "beta"
#Print the table
kable(fm2.table, caption = "Model description")
```

```r
y_hat1 <- predict(mod1, y_test)
y_hat2 <- predict(mod2, y_test)


y <- data.table(y.hat.1 = y_hat1, y.hat.2 = y_hat2, y = y_test[,VAS.12Months])

RSS1 <- sum((y$y - y$y.hat.1)^2)
RSS2 <- sum((y$y - y$y.hat.2)^2)

ggplot(y, aes(x = y, y=y.hat.2)) +
geom_point() +
geom_abline(slope = 1, intercept = 0, color = "red") +
xlab("Expected value") +
ylab("Predicted value") +
labs(caption= "Fig 10: Predicted value vs expected value") +
theme(plot.caption = element_text(hjust = 0), text = element_text(size=16))
```