Biomarkers and Pain - Analysis

Project by
Student number - s2511090

1 Introduction

This project revolves around the statistical analysis of the data collected by two prominent Scandinavian university hospitals. The obtained data consists of medical history of individuals who are afflicted by a medical condition that induces pain and their respective responses. The dataset specifically includes particular details on the levels of specific biomarkers that influences patients' pathophysiology.

2 Objective

The principle objective of this study is to investigate and elucidate a comprehensive understanding of the relationship between the given biomarker levels and the VAS(Visual Analog Scale) levels. Furthermore, we intend to scrutinize the impact of other contributing factors such as gender, smoking status, biomarker concentrations and time-frame. We want to study how these variables affect the reported pain levels over a span of 12-month period.

3 Data collection

The data used to perform the required statistical tests and modelling is obtained from two Scandinavian university hospitals. The study acquired the data from patients suffering from a painful medical condition. The patients were requested to rate the pain levels ranging from 0-10 using a Visual analog scale which were recorded at 2 distinct time points: at 0 week(during the acute phase of the condition) and a 12 month follow up. Blood samples were drawn from patients and recorded at the time of inclusion, 6 weeks later and 12 months later. The blood samples were analysed for the concentration of specific biomarkers. The list of biomarkers involve Interleukin-6 (IL-6), Vascular endothelial growth factor A (VEGF-A), Osteoprotegerin (OPG), Latency-associated peptide transforming growth factor beta 1 (TGF-beta-1), Interleukin-8 (IL-8), C-X-C motif chemokine 9 (CXCL9), C-X-C motif chemokine 1 (CXCL1), Interleukin-18 (IL-18) and Macrophage colony-stimulating factor 1 (CSF-1). Alongside these biomarkers, details of various covariates were also recorded for all patients which will facilitate robust statistical analyses to explore the interplay between these biomarkers and pain levels.

3.1 Data Pre-processing

The original data cannot be considered for analysis without preprocessing as it may contain errors and blank entries. For data preprocessing we apply a two-pronged approach. To enhance data reliability, first we conducted thorough data cleaning, which involved identifying and rectifying inconsistencies within the dataset. Subsequently, to fill the gaps caused by missing values, we employed mean value imputation, ensuring the completeness of data. Majority of the pre-processing of the data is done in MS Excel. As the data used to perform the statistical analysis involves data for the time of inclusion and 12 month follow-up, the data cleaning is done for Patient-ID 40, 49, 117, 122 and 126 as these do not contain either the inclusion entries or 12 month follow up entries. Hence it is cleaned from the dataset to produce more accurate results. This makes the total number of patients included in the study to be 113 instead of 118. For the missing values in covariates data, mean imputation method is used to provide data for blank cells. This helps to generate more accurate and reliable results in our analysis.

4 Methodology

Two methods are utilized to analyse the data. Both of the methods have different approach. The methodology is explained in brief as follows:

4.1 Hypothesis Testing

A statistical hypothesis test is a method of statistical inference used to decide whether the data at hand sufficiently support a particular hypothesis[4]. Hypothesis testing allows us to make probabilistic statements about population parameters. There are several types of questions whose answers the research team wants to discover. However the scope of this report is restricted to determine whether the levels of biomarkers during the acute phase of the condition differ between male and female patients. On that premise, to determine whether there is a statistical relationship between the biomarker levels at the inclusion among male and female patients, we perform a hypothesis test. The first step of any hypothesis testing is to formulate appropriate null and alternate hypothesis. This type of hypothesis typically involves comparing two independent groups to determine if there is a significant difference between their biomarker levels at inclusion. Hence we consider following null hypothesis $H0$ and alternate hypothesis $H1$ to check whether there exists a statistical relationship between biomarker levels for male and female patients at the time of inclusion.

- H_0: The levels of a biomarker at inclusion do not vary between male and female patients. ($\mu_1 = \mu_2$)
- H_1: The levels of a biomarker at inclusion vary between male and female patients. ($\mu_1 \neq \mu_2$)

Here we consider the random variables X and Y as the biomarker levels at inclusion for males and females respectively. Along with that the parameters of distribution are taken as the population mean of the biomarker levels at inclusion μ and the population standard deviation of the biomarker levels at inclusion σ. The significance level for the above stated hypothesis is taken as 0.05.

4.1.1 Potential problems with multiple testing

There are potential problems with multiple testing which includes an increased risk of making a false positive error i.e. Type I error. If we assume that all tests are independent and that all null hypotheses are true then probability of making at least one type I error can be calculated as:

$$P(\text{At least one Type I error}) = 1 \text{ - } (1 \text{ - } \alpha)^n = 1 \text{ - } (1 \text{ - } 0.05)^9 = 0.3697505903$$

The potential problems with multiple testing are as follows:

1. Type I Error Inflation: The more hypothesis tests you run, the greater the chance of encountering Type I errors, which leads to rejecting null hypotheses even if they are actually true. This amplification arises because each test is executed at a pre-defined significance level α which leads to an increased cumulative rise in getting Type I errors[5].

2. Family-wise Error: This is the probability of making at least one Type I error across all the tests. Controlling the family-wise error rate is important to maintain the overall significance level. The Bonferroni correction is used to correct the familywise error. The Bonferroni correction sets the significance cut-off at α/n [5].For the given hypothesis testing, n = 9 and α = 0.05, hence the new cut-off will be α_{new} = 0.0056 and we'd only reject a null hypothesis if the p-value is less than 0.0056.

3. False Discovery Rate: The false discovery rate (FDR) is the expected proportion of type I errors. The FDR works by estimating some rejection region so that on average, FDR $< \alpha$ [5].

4.2 Regression Modelling

In a statistical setting, regression modelling is referred as the process of estimating the relationship between a response variable (also called as dependent variable) and one or more explanatory variables (also called as independent variable)[2] [3]. A simple linear regression model involves only one independent variable and hence it can be written as,

$$y = \beta_0 + \beta x$$

where y is the response variable, β_0 is the intercept (value of y when all the independent variables are equal to zero), β is estimated regression coefficient and x is explanatory variable. A multiple linear regression model involves more than one independent variables. The formula for multiple regression is as stated below,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_i x_i$$

where y is the response variable, β_0 is the intercept, $\beta_1, \beta_2, \beta_i$ are estimated regression coefficients and x_1, x_2, x_i are explanatory variables.
The problem at hand involves multiple regression method to analyse the given data as there are more than one independent variables. For the purpose of this study a multiple linear regression model is created using the 12-month VAS as the response variable and biomarker levels (at inclusion) and covariates as explanatory variables. There are 9 individual biomarkers and 4 covariates (age, sex, smoking status of patient and VAS at inclusion) which account for a total of 13 explanatory variables whereas, the pain level observations at 12-month act as response variable.

5 Results

For hypothesis we tested whether there is significant difference in the levels of biomarkers at the acute phase of the condition between male and female patients. The results obtained by performing two sample t-test in RStudio[1] by using t.test()[6] for hypothesis testing gives the following results. The results contain value of the t-statistic, degrees of freedom, p-value and confidence interval for every biomarker.

Biomarker name	t-statistic	Degree of Freedom	p-value	Confidence Interval	
IL-6	-1.2375	104.15	0.2187	-0.6059797	0.1402654
VEGF-A	-2.4549	108.66	0.01568	-0.54025160	-0.05757421
OPG	-1.687	109.01	0.09446	-0.29016907	0.02332696
TGF-beta-1	-2.051	110.96	0.04263	-0.70116169	-0.01206513
IL-8	-0.93482	110.09	0.3519	-0.4906560	0.1761259
CXCL9	-0.11925	110.77	0.9053	-0.3613378	0.3203165
CXCL1	-2.7312	111	0.007343	-1.0616653	-0.1688673
IL-18	1.0928	110.22	0.2769	-0.1017290	0.3518293
CSF-1	-3.0024	99.837	0.003385	-0.26286853	-0.05368536

Table 1 - Welch Two-sided T-Test Results for every biomarker

Another set of results are obtained by performing a multiple regression on the 80% of the data (96 patients out of 113 patients) to investigate the relationship between pain level at 12 months and a set of predictor variables, including biological markers (IL-6, VEGF-A, OPG, TGF-beta-1, IL-8, CXCL9, CXCL1, IL-18, CSF-1), demographic variables (age, sex), smoking condition, and the initial pain level at inclusion. The goal of this study is to identify the factors among the independent variables that influence the dependent variable. The table below provides the details of the estimated coefficients for the predictor variables.

Explanatory Variable	Coefficients	Standard Error	t-value	p-value
Intercept	0.004881	10.340823	0.00	0.99962
IL-6	0.711476	0.371381	1.916	0.05888 .
VEGF-A	1.348214	0.700076	1.926	0.05759 .
OPG	-2.144222	0.776915	-2.760	0.00713 **
TGF-beta-1	-0.995913	0.642972	-1.549	0.12525
IL-8	0.634761	0.599247	1.059	0.29259
CXCL9	-0.193747	0.333429	-0.581	0.56278
CXCL1	-0.128593	0.483203	-0.266	0.79081
IL-18	-0.462399	0.534127	-0.866	0.38917
CSF-1	1.885928	1.353646	1.393	0.16732
age	-0.004782	0.031255	-0.153	0.87878
Sex	0.276725	0.597500	0.463	0.64449
Smoking status	-0.272159	0.651274	-0.418	0.67712
VAS inclusion	0.289959	0.111364	2.604	0.01094 *

Table 2 - Estimated coefficients for predictor variables
Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
The significance is denoted using asterisks (*), more the asterisks higher the significance

The above model is used to make predictions for the remaining 20 % of the patients to perform out-of-sample evaluation of the model, where we compare the predicted 12-month VAS to actual 12-month VAS. The following table summarizes the predicted values and the actual values of pain levels at 12 months for the patients included in the remaining 20%:

PatientID	Actual VAS	Predicted VAS
23	8.0	5.6417938
32	9.5	4.5053537
56	0.0	1.8484814
61	0.0	1.5112938
65	3.5	4.0770432
66	0.5	2.6665980
90	7.5	4.8855477
92	8.3	5.1629244
96	0.0	3.0942348
100	1.0	2.9685095
102	8.5	4.9123494
127	10.0	4.6308236
132	4.0	3.3689776
136	0.0	0.8543439
140	9.0	3.8675909
143	2.0	4.0792430
150	9.0	2.8383499

Table 3 - Actual and predicted VAS at 12 months for 20% of data

6 Statistical Analysis

6.1 Analysis of Hypothesis test

In this test we have stated the null hypothesis (H0) as there is no significant difference in the mean test scores of biomarker levels at inclusion between males and females, while the alternative hypothesis (H1) suggests that there is a significant difference. The T-Statistic value represents how many standard errors the sample means are away from each other. The p-value indicates that the probability of observing the data if the null hypothesis is true. A p-value less than or equal to a predetermined significance level ($\alpha = 0.05$) indicates a statistically significant result, meaning the observed data provide strong evidence against the null hypothesis. The confidence interval attribute of the t-test suggests that for 95% confidence interval, the true difference in mean test scores between males and females could range from (lower limit, upper limit)[3]. We perform these analysis steps to check whether to accept or reject the null hypothesis.

Biomarker name	p-value	Hypothesis Decision for ($\alpha = 0.05$)	Hypothesis Decision by Bonferroni Correction for ($\alpha_{new} = 0.0056$)
IL-6	0.2187	Accept H0	Accept H0
VEGF-A	0.01568	Reject H0	Reject H0
OPG	0.09446	Accept H0	Accept H0
TGF-beta-1	0.04263	Reject H0	Reject H0
IL-8	0.3519	Accept H0	Accept H0
CXCL9	0.9053	Accept H0	Accept H0
CXCL1	0.007343	Reject H0	Accept H0
IL-18	0.2769	Accept H0	Accept H0
CSF-1	0.003385	Reject H0	Reject H0

Table 4 - Hypothesis decision results for all biomarkers

For biomarker CXCL1, the t-statistics -2.7312 is negative, which indicates that the mean test score for males is lower than that for females. The corresponding p-value 0.007343 is less than the significance level $\alpha = 0.05$. This helps in the decision making to reject the null hypothesis, suggesting that there is a significant difference. The confidence interval for this biomarker is (-0.54025160, -0.05757421), which states that the true difference in mean test scores between males and females lies between -0.54025160 and -0.05757421. However, due to Bonferroni correction we compare the p-value against the new significance level $\alpha_{new} = 0.0056$. Hence we detect the type I error in this case by performing Bonferroni correction and rule out the possibility of getting such Type I errors.

6.2 Analysis of Multiple Regression Model

For the regression modelling, we performed multiple linear regression analysis considering VAS at 12 month as the response variable and explanatory variables such as sex, age and smoking status of the patients along with the biomarker levels at inclusion. To analyse the multiple regression we discuss each parameter of the results such as coefficients, t-value, p-value, adjusted R-squared value, F-statistic and residual standard error.

The coefficients represent the estimated effect of each explanatory variable on the response variable. the coefficient associated with every predictor variable represents the change in the dependent variable for a one-unit change in the corresponding predictor variable while holding all other variables constant. In case of t-values, higher the associated t-value greater is the significance.

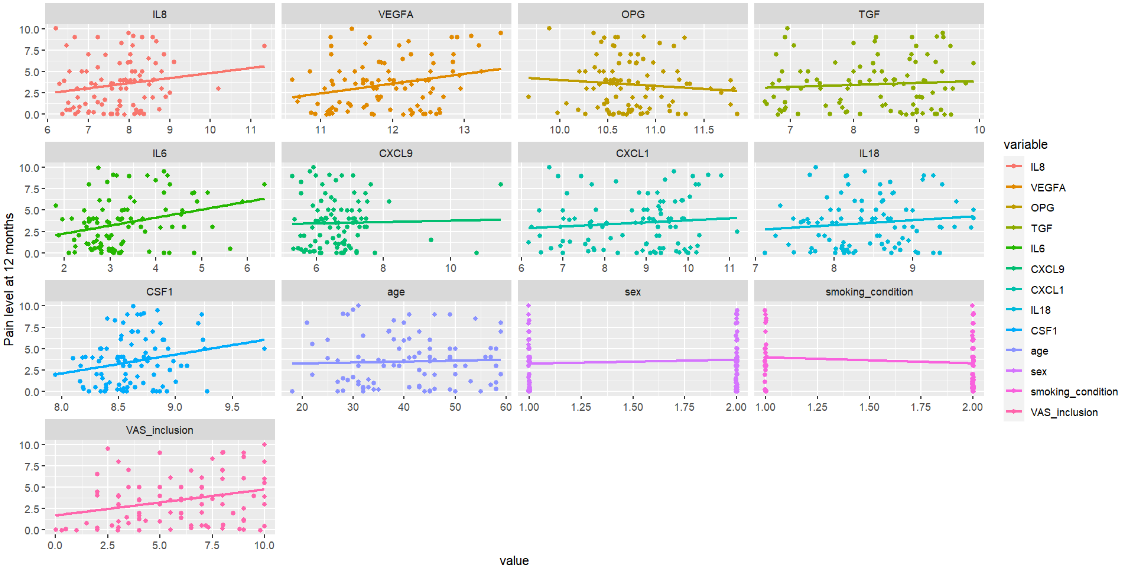

Figure 1: Multiple Regression for VAS at 12 month as response variable

P-value is the most crucial factor for determining the statistical significance of the model. The biomarker OPG has a statistically significant negative impact on the response variable. VAS at inclusion also has a p-value 0.01094 indicating statistical significance.

The adjusted R-squared value (0.1938) accounts for all the predictors in the model and indicates how well the independent variables collectively explain the variance in response variable. F-statistic tests whether the model as a whole is statistically significant. In this case the model is statistically significant as the overall p-value of the model is 0.002719 which is less than the significance level. The residual standard error (2.573) represents the average variation in the observed values of pain level at 12 month from the predicted values by the model.

7 Conclusion

From the hypothesis testing we can conclude that for biomarkers Interleukin-6 (IL-6), Osteoprotegerin (OPG),Interleukin-8 (IL-8), C-X-C motif chemokine 9 (CXCL9), C-X-C motif chemokine 1 (CXCL1) (* because of Bonferroni correction) and Interleukin-18 (IL-18) there is no significant difference in the mean of the biomarker levels at inclusion between males and females. The biomarkers Vascular endothelial growth factor A (VEGF-A), Latency-associated peptide transforming growth factor beta 1 (TGF-beta-1) and Macrophage colony-stimulating factor 1 (CSF-1) show significant difference in the mean of the biomarker levels at inclusion between males and females.
From the multiple regression test we can state that the model having 12-month VAS as the response variable and covariates & biomarker levels at inclusion as independent variables is statistically significant.

References

[1] R documentation. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/t.test.

[2] Regression analysis wikipedia. https://en.wikipedia.org/wiki/Regression_analysis.

[3] Sheldon M. Ross. *Introduction To Probability And Statistics For Engineers And Scientists*.

[4] Statistical hypothesis testing. https://en.wikipedia.org/wiki/Statistical_hypothesis_testing.

[5] Statistics for bioinformatics. https://www.stat.berkeley.edu/~mgoldman/Section0402.pdf.

[6] Two sample t-test. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/t.test.

8 Appendix

8.1 Github repository

https://github.com/rasikadalvi/s2511090/blob/main/s2511090.r

8.2 R Code for hypothesis test

```
library(readxl)
library(dplyr)
library(broom)
library(ggplot2)
library(reshape2)
library(caret)
## Introductory Probability and Statistics
## Project - Biomarkers and Pain level
## Student number- s2511090

## Hypothesis testing question - Do the levels at inclusion vary between males and females?

# Pre-Processed Biomarker file
biomarkers_data <- read_excel("biomarkers.xlsx")
# Pre-Processed Covariate file
covariates_data <- read_excel("covariates.xlsx")

# Extract covariates data for females
female_data <- covariates_data %>%
  filter(covariates_data[,3] == "2")

# Extract covariates data for males
male_data <- covariates_data %>%
  filter(covariates_data[,3] == "1")

# Combine biomarker data for female patients
female_inclusion<- biomarkers_data %>%
  filter(PatientID %in% female_data$PatientID, Timepoint == "0weeks")

# Combine biomarker data for male patients
male_inclusion<- biomarkers_data %>%
  filter(PatientID %in% male_data$PatientID, Timepoint == "0weeks")

### Hypothesis testing - Do the levels at inclusion vary between males and females?
  ## H0 (NULL Hypothesis): The levels of biomarker "X" at inclusion do not differ between females
  ## H1 (Alternate Hypothesis): The levels of biomarker "X" at inclusion differ between females ar

### Hypothesis testing done by Welch Two Sample t-test for every biomarker
## Biomarker IL-6
IL6_0W <- t.test(male_inclusion$‘IL-6‘, female_inclusion$‘IL-6‘)
IL6_0W
## Biomarker VEGF-A
VEGFA_0W <- t.test(male_inclusion$‘VEGF-A‘, female_inclusion$‘VEGF-A‘)
VEGFA_0W
## Biomarker OPG
OPG_0W <- t.test(male_inclusion$OPG, female_inclusion$OPG)
OPG_0W
## Biomarker TGF-beta-1
TGFbeta1_0W <- t.test(male_inclusion$‘TGF-beta-1‘, female_inclusion$‘TGF-beta-1‘)
TGFbeta1_0W
## Biomarker IL-8
IL8_0W <- t.test(male_inclusion$‘IL-8‘, female_inclusion$‘IL-8‘)
IL8_0W
## Biomarker CXCL9
```

```
CXCL9_OW <- t.test(male_inclusion$CXCL9, female_inclusion$CXCL9)
CXCL9_OW
## Biomarker CXCL1
CXCL1_OW <- t.test(male_inclusion$CXCL1, female_inclusion$CXCL1)
CXCL1_OW
## Biomarker IL-18
IL18_OW <- t.test(male_inclusion$‘IL-18‘, female_inclusion$‘IL-18‘)
IL18_OW
## Biomarker CSF-1
CSF1_OW <- t.test(male_inclusion$‘CSF-1‘, female_inclusion$‘CSF-1‘)
CSF1_OW
```

8.3 R code for Multiple Regression

```
### Multiple Regression
set.seed(50)
#use 80% of dataset as training set and 20% as test set
sample <- sample(c(TRUE, FALSE), nrow(covariates_data), replace=TRUE, prob=c(0.8,0.2))
train  <- covariates_data[sample, ]
test   <- covariates_data[!sample, ]

VAS_12months <- train$"Vas-12months"
vector_ID <- train$PatientID
filtered_data <- biomarkers_data %>%
  filter(PatientID %in% vector_ID, Timepoint == "0weeks")

age <- train$"Age"
smoking_status <- train$"Smoker (1=yes, 2=no)"
sex <- train$"Sex (1=male, 2=female)"
VAS_inclusion <- train$"VAS-at-inclusion"

IL8 <- filtered_data$"IL-8"
VEGFA <- filtered_data$"VEGF-A"
OPG <- filtered_data$"OPG"
TGF <- filtered_data$"TGF-beta-1"
IL6 <- filtered_data$"IL-6"
CXCL9 <- filtered_data$"CXCL9"
CXCL1 <- filtered_data$"CXCL1"
IL18 <- filtered_data$"IL-18"
CSF1 <- filtered_data$"CSF-1"

# Fit the multiple regression model
model1 <- lm(VAS_12months ˜ IL8 + VEGFA + OPG + TGF + IL6 + CXCL9 + CXCL1 + IL18 + CSF1 + age + se

# Print the model summary
summary(model1)
# plot(model1)

new_train_comb <- data.frame(VAS_12months , IL8 , VEGFA , OPG , TGF , IL6 , CXCL9 , CXCL1 , IL18 ,
new_train_comb = melt(new_train_comb, id.vars='VAS_12months')
# Plot the regression plot against each explanatory variable
ggplot(new_train_comb) +
  geom_jitter(aes(value,VAS_12months, colour=variable),) + geom_smooth(aes(value,VAS_12months, col
  facet_wrap(˜variable, scales="free_x") +
  labs(y = "Pain level at 12 months")

vector_ID_test <- test$PatientID
filtered_data_test <- biomarkers_data %>%
  filter(PatientID %in% vector_ID_test, Timepoint == "0weeks")
```

```
age <- test$"Age"
smoking_status <- test$"Smoker (1=yes, 2=no)"
sex <- test$"Sex (1=male, 2=female)"
VAS_inclusion <- test$"VAS-at-inclusion"

IL8 <- filtered_data_test$"IL-8"
VEGFA <- filtered_data_test$"VEGF-A"
OPG <- filtered_data_test$"OPG"
TGF <- filtered_data_test$"TGF-beta-1"
IL6 <- filtered_data_test$"IL-6"
CXCL9 <- filtered_data_test$"CXCL9"
CXCL1 <- filtered_data_test$"CXCL1"
IL18 <- filtered_data_test$"IL-18"
CSF1 <- filtered_data_test$"CSF-1"

new <- data.frame(IL8 , VEGFA , OPG , TGF , IL6 , CXCL9 , CXCL1 , IL18 , CSF1 , age , sex , smokin

# Make predictions for the test data
predictions <- predict(model1, newdata = new)

# Create a table of fitted parameter values
fitted_parameters <- data.frame(
  Parameter = names(coef(model1)),
  Coefficient = coef(model1)
)
print(fitted_parameters)

new_prediction_data <- data.frame(predictions , IL8 , VEGFA , OPG , TGF , IL6 , CXCL9 , CXCL1 , IL

plot(model1)
```

8.4 Plots

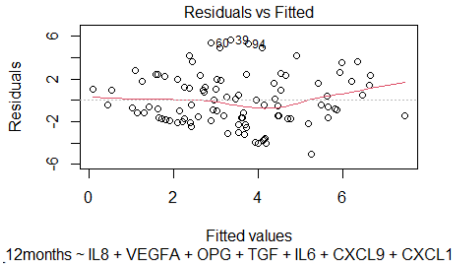

Figure 2: Residual vs Fitted plot

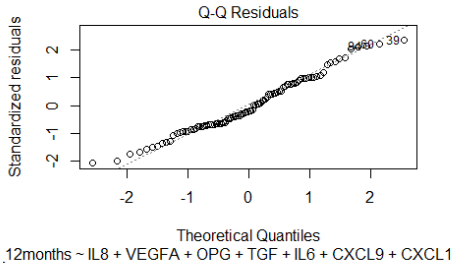

Figure 3: Q-Q plot

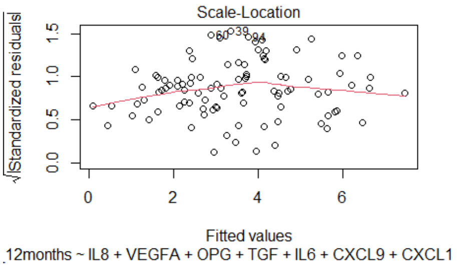

Figure 4: Scale-location

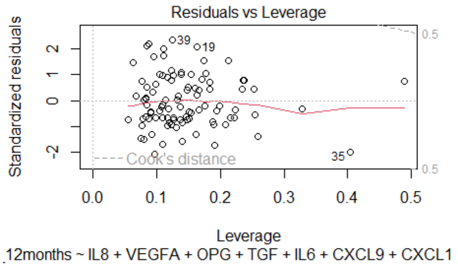

Figure 5: Residual vs Leverage

Figure 6: