Statistics

Elena Williams

Short intro

Dataframe of biomarkers in each paper. A dataset containing the information about papers mentioning pain biomarkers.

In the presented statistical report, I am going to analyze a dataset from an ongoing study held at two Scandinavian university hospitals. The research was concerned with patients that had a medical condition causing pain. Blood samples were taken at inclusion and analyzed. Some of the proteins are believed to influence the condition.

More and more clinical research findings indicate that men and women have a different biomarkers levels [1, 2, 3] and even some of the evidence shows that there are sex differences in pain responses [4]. We endeavored to identify objective blood biomarkers for pain, a subjective sensation with a biological basis, using a stepwise discovery, prioritization, validation, and testing in independent cohorts design. We studied psychiatric patients, a high risk group for co-morbid pain disorders and increased perception of pain.

A data frame with 794 rows and 6 variables: * Participant ID - ID attached to each of the participant of the study * Diagnosis - sample included patients with psychiatric conditions like Bipolar disorder (BP), Schizoaffective disorder (SZA), Schizophrenia (SZ), Major depressive disorder (MDD), Post-traumatic stress disorder (PTSD), Mood disorder (MOOD) and others. * Gender - male or female * Age - age of the participants * Ethnicity Caucasian, African American, Hispanic, Asian American, Mixed, Asian * Pain Scale - reported pain level on a scale from 1 to 10

Content structure:

Exploratory data analysis – Cleaning Data – Setting the Hypothesis – Visualising the Distribution – Comparing skewness and kurtosis of the factor variables – Testing variables on homogenity of variances using Bartlett's test

Testing the hypothesis the difference in means – Two-tailed T-test – One-way ANOVA – Computing Tukey Honest Significant Differences

Regression Analysis – Building a Model – Exploring Diagnostic Plots – Checking Multicollinearity – Evaluating the performance of the model on the test set calculating Root Mean Squared Error

1. Exploratory data analysis

a Cleaning data

I note that there are blank cells in the data set. The pain variable has 181 missings. These cells will be converted in NAs (namely missing data points in R) which allow us to progress further with the robust data analysis.

```
## Warning: NAs introduced by coercion
```

(b) Setting hypothesis

In a given sample we have 59 men and 58 women. I would like to examine whether the levels of 9 different biomarkers taken at the beginning of the study vary between males and females.

The following hypothesis were set:

H0.1 : There is no difference in pain perception between men and women

Ha.1: There is a difference in pain perception between men and women

H0.2 : There is no difference in pain perception between the given diagnosis

Ha.2: There is a difference in pain perception between the given diagnosis

Based on the null hypothesis we assume that there is no difference in biomarker levels between men and women. Otherwise, the alternative hypothesis is that the true difference in means is not equal to 0.

The pain levels is a continuous random variable and the gender, diagnosis and ethnicity characteristics are discrete random variables.

Before performing a hypothesis tests I will look at the distributions of the variables of interest using histogram plot and grouped bar charts.

111 women and 371 men recorded their pain levels **(c)** Exploring the data set

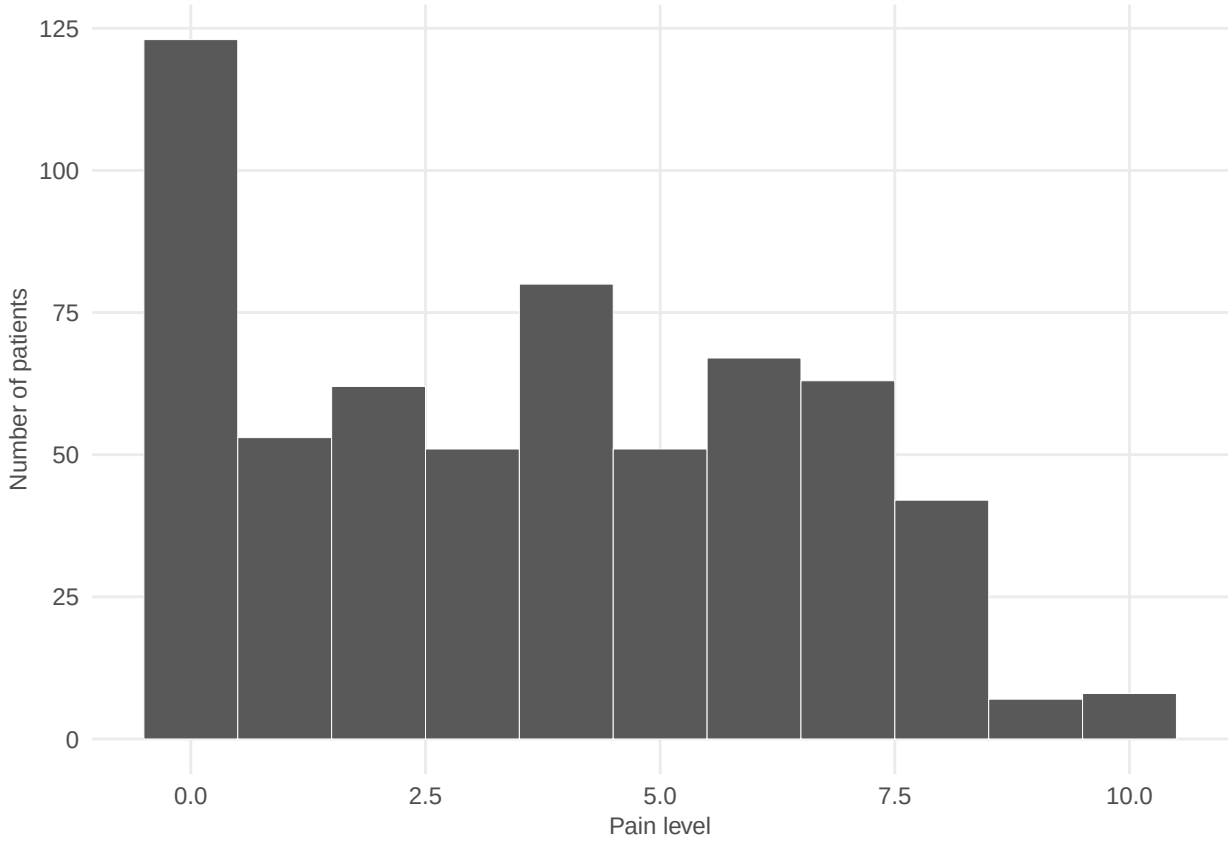
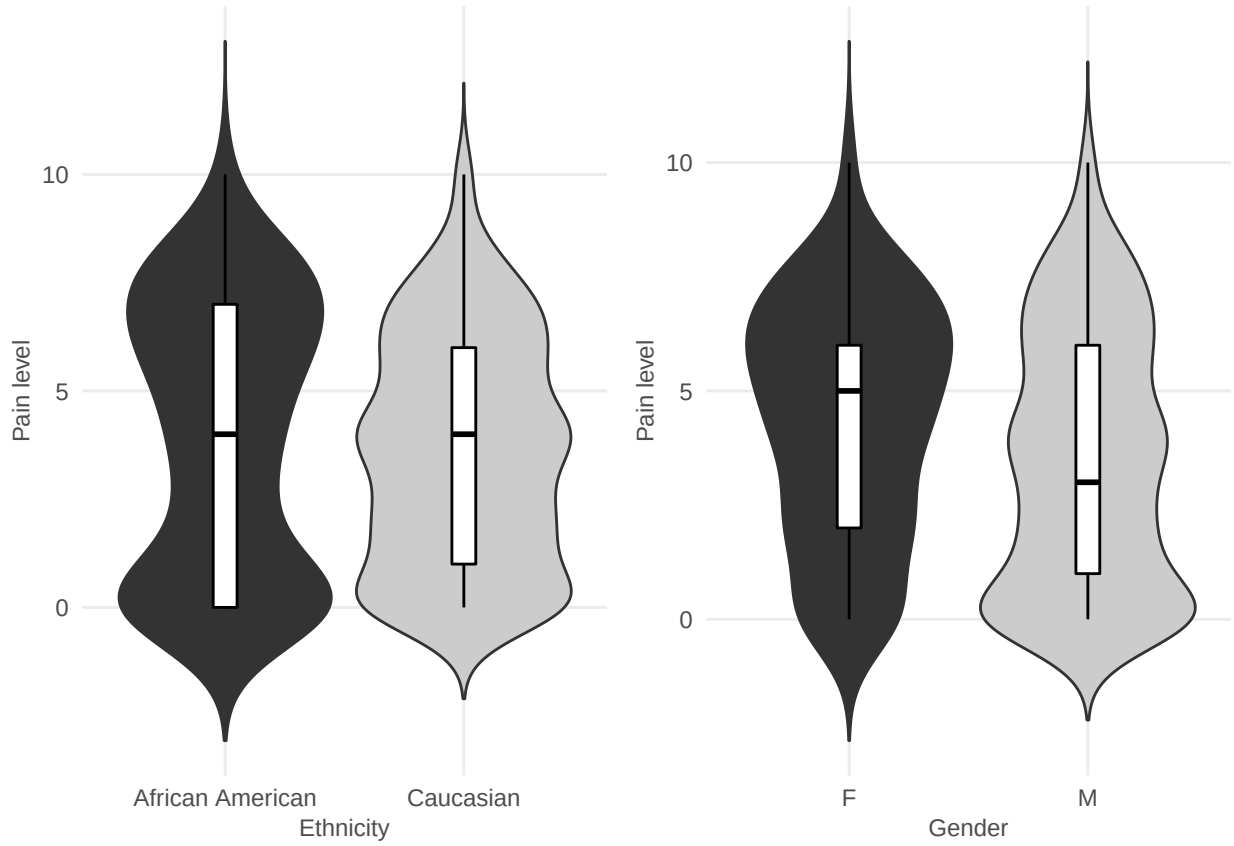
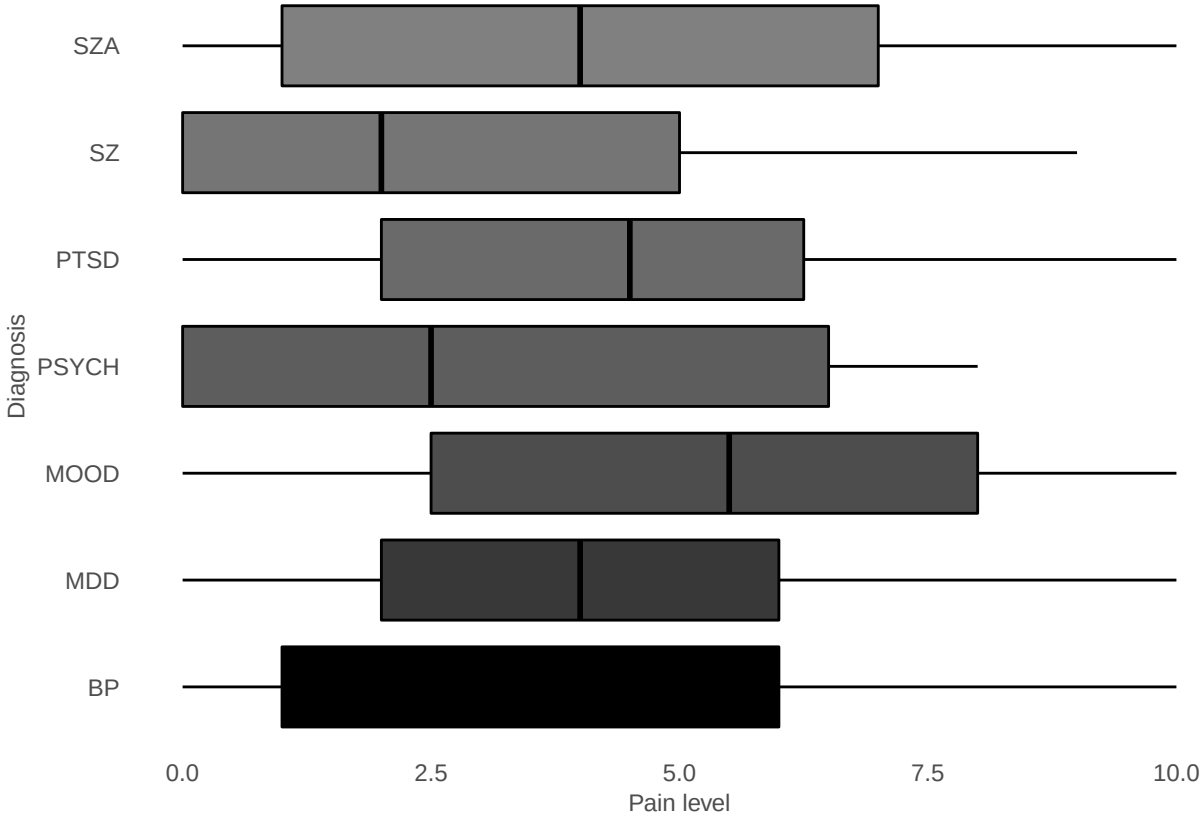


Figure 1: The distribution of pain scores among the patients

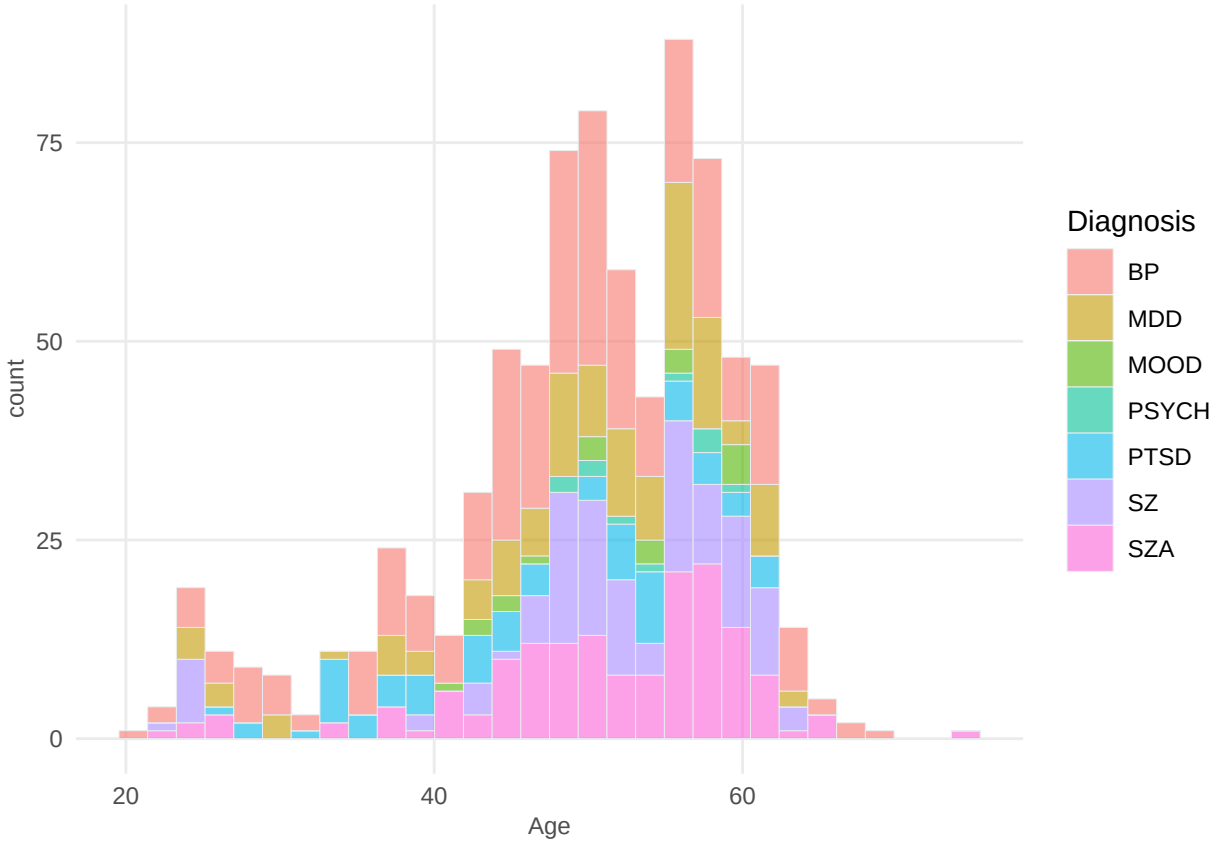
In figure 1 we note that the pain scores have a slightly right-skewed distribution.



blahblah blah



```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Then I compared the coefficients of skewness and kurtosis for both genders. These measures represent the asymmetry and the *"tailedness"* of biomarker distribution. Overall the skewness and kurtosis coefficients look all right for most of the variables except CXCL9. The estimated skewness is 2.73 and 1.8, the kurtosis is 11.5 and 6.87 for women and men respectively. High kurtosis indicates that we have rare patients in our sample with extreme protein levels.

Table 1: Skewness and Kurtosis coefficients

	N	Mean	SD	Median	Skew	Kurtosis
BP	219	3.35	3.0	3.0	0.32	-0.99
MDD	124	4.31	4.0	4.0	0.00	-1.04
MOOD	20	4.85	5.5	5.5	-0.25	-1.39
PSYCH	8	3.38	2.5	2.5	0.24	-1.95
PTSD	72	4.35	4.5	4.5	-0.21	-0.97
SZ	92	2.71	2.0	2.0	0.47	-1.18
SZA	71	3.82	4.0	4.0	0.22	-1.30

Then I compared the coefficients of skewness and kurtosis for both genders. These measures represent the asymmetry and the *"tailedness"* of biomarker distribution. Overall the skewness and kurtosis coefficients look all right for most of the variables except CXCL9. The estimated skewness is 2.73 and 1.8, the kurtosis is 11.5 and 6.87 for women and men respectively. High kurtosis indicates that we have rare patients in our sample with extreme protein levels.

Table 2: Skewness and Kurtosis coefficients

	N	Mean	SD	Median	Skew	Kurtosis
F	126	4.36	5	5	-0.19	-0.97
M	481	3.49	3	3	0.29	-1.08

Finally, I have checked the data on homogeneity of variances using Bartlett's test. Proteins IL.6 and CSF.1 have shown a significant P-value at 0.05 level. For these variables the variance is not homogeneous and correction is needed.

Table 3: Results of Bartlett's tests for homogeneity of variances

Varibale	K.squared	p.value
Gender	1.286	0.2568
Diagnosis	8.129	0.2288

(c) T-test. Comparing the means

I have chosen a two-sample T-test in which the test statistic follows a Student's t-distribution under the null hypothesis [5].

I have used this test to indentify whether the mean difference between two sexes is significant. Taking into consideration the results of Bartlett's test Welch Two-Sample T-test was applied for proteins IL.6 and CSF.1 .

The results have shown that biomarkers VEGF.A, TGF.beta.1, CXCL1 and CSF.1 have a difference in means between men and women at 0.05 significance level. We reject the null hypothesis that the difference in means is equal to 0 and accept an alterantive hypothesis.

Tests with other proteins have shown no significant difference in means.

To conclude, women and men with medical conditions causing pain have a mean difference in VEGF.A, TGF.beta.1, CXCL1 and CSF.1 proteins levels at inclusion. No significant differences were found in IL.8, OPG, IL.6, CXCL9, IL.18 protein levels.

Table 4: Results of Two Sample t-test

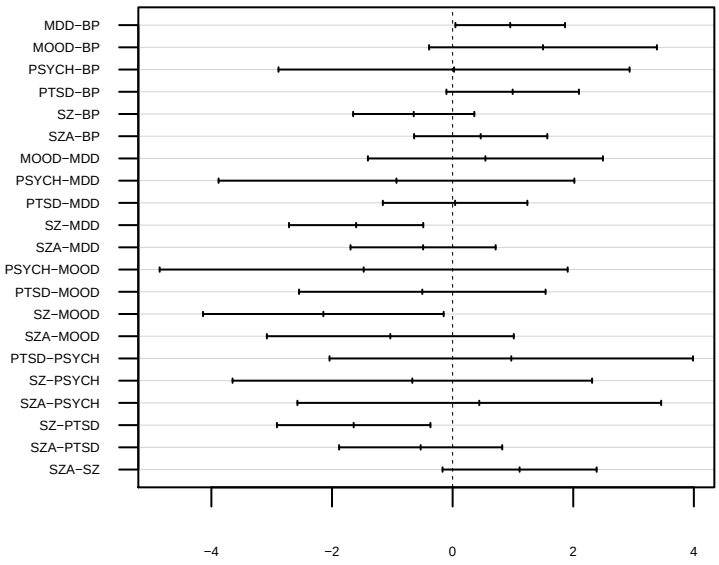
Mean-1	Mean-2	T-statistic	P-value	DF
4.357	3.489	3.14	0.001774	605

(d) ANOVA

The one-way analysis of variance (ANOVA), also known as one-factor ANOVA, is an extension of independent two-samples t-test for comparing means in a situation where there are more than two groups. In one-way ANOVA, the data is organized into several groups base on one single grouping variable (also called factor variable). This tutorial describes the basic principle of the one-way ANOVA test and provides practical anova test examples in R software.

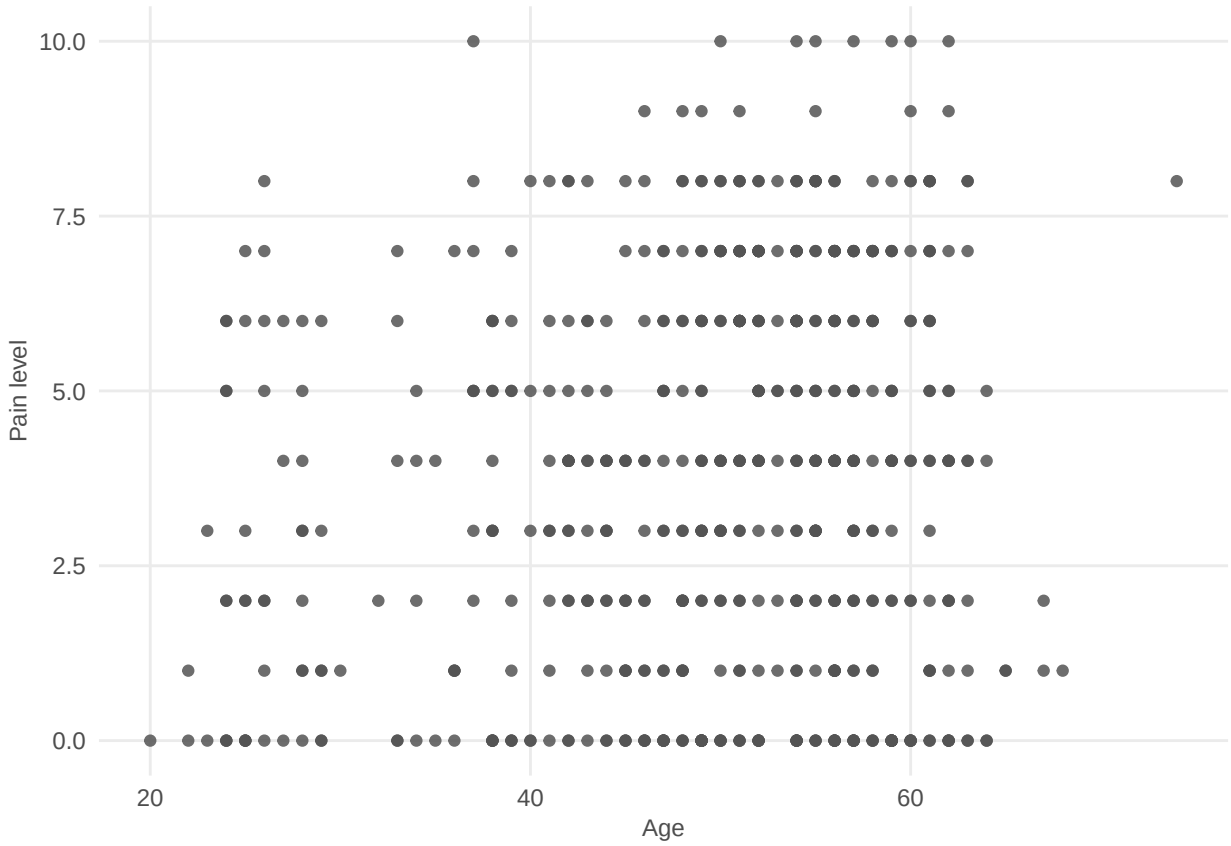
ANOVA test hypotheses:

Null hypothesis: the means of the different groups are the same Alternative hypothesis: At least one sample mean is not equal to the others. The observations are obtained independently and randomly from the population defined by the factor levels The data of each factor level are normally distributed. These normal populations have a common variance. (Levene's test can be used to check this.)



Regression analysis

(a) Scatter plot

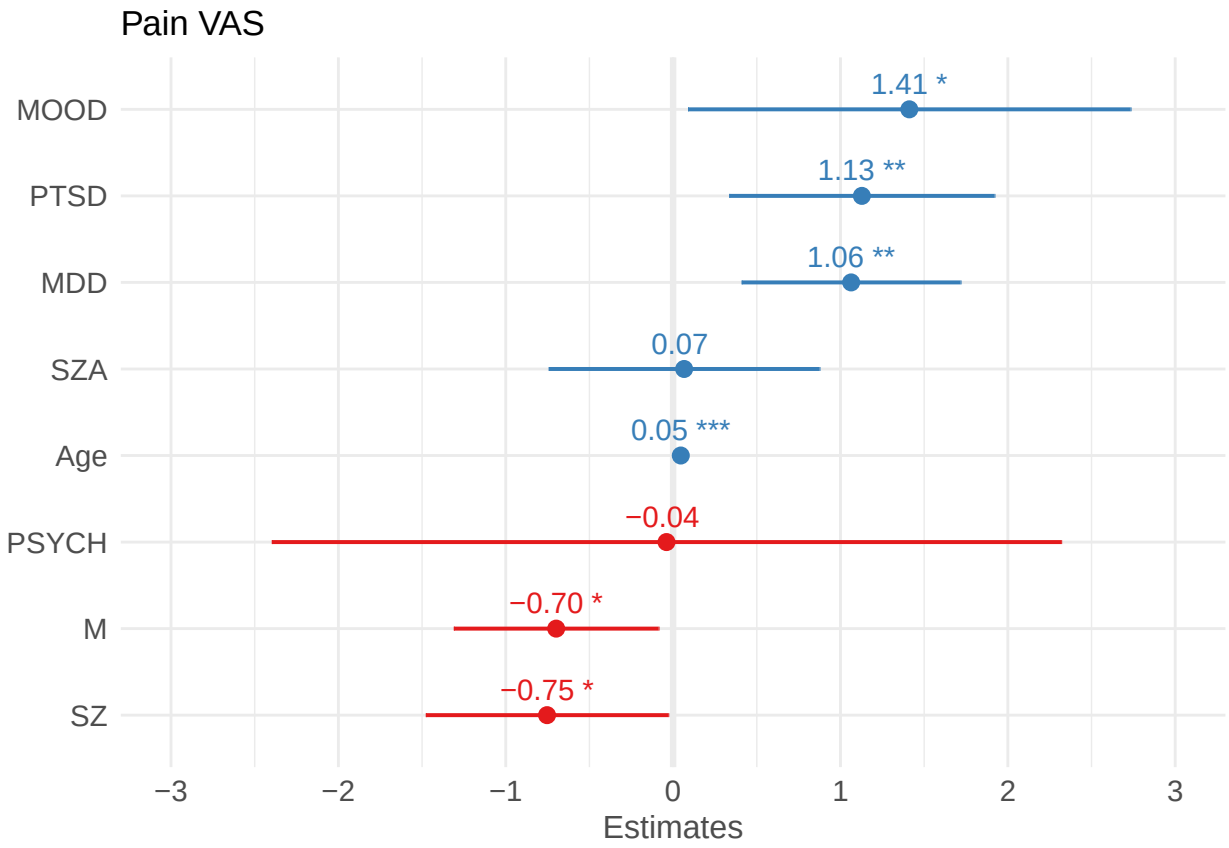


I have constructed a regression model to make predictions on how well patients with medical conditions will recover.

The dependent variable is a pain level measured one year after onset and independent variables are biomarker levels at inclusion. Covariates such as age, sex and smoke status were also included in the model.

The results have shown that proteins OPG, TGF.beta.1 and IL.6 are strongly related to pain at 0.05 significance level. Both estimates have negative sign meaning that the higher the pain the lower the protein level or vice versa.

Table 6: Results of Regression model


Pain VAS
MOOD
PTSD
MDD
SZA
Age
PSYCH
M
SZ
1.41 *
1.13 **
1.06 **
0.07
0.05 ***
−0.04
−0.70 *
−0.75 *
−3
−2
−1
0
1
2
3
Estimates