Probability and Statistics

End-of-course summative assessment

Table of Contents

[Introduction 1](#_Toc33883261)

[Statistical hypothesis testing 1](#_Toc33883262)

[Regression modelling 4](#_Toc33883263)

[Annexe 6](#_Toc33883264)

[1. Distribution of biomarkers’ levels at inclusion disaggregated by sex 6](#_Toc33883265)

[2. Boxplot of biomarkers’ levels at inclusion disaggregated by sex 7](#_Toc33883266)

[3. The normality test result of biomarkers after removing the outliers 7](#_Toc33883267)

[4. Distribution of biomarkers’ levels at inclusion disaggregated by sex after removing the outliers 8](#_Toc33883268)

[5. Boxplot of biomarkers’ levels at inclusion disaggregated by sex after removing the outliers 9](#_Toc33883269)

[The Source Code: 11](#_Toc33883270)

Exam Number: B155243

*University of Edinburgh*

Introduction

In this assignment, I will be using two datasets of an ongoing research project from two Scandinavian university hospitals.

The two datasets are as follows:

* **Covariates** (covariates.xlsx): This dataset contains basic biodata about each patient and their VAS (scale of pain) at inclusion and 12 months later.
* **Biomarkers** (biomarkers.xlsx): This dataset contains individual measurements of the level of certain proteins (Biomarkers) for each patient and three different timepoints.

To use the two datasets, some basic manipulation had to be done such as renaming columns for better visibility and easier access from the code, validating the datatypes of each column and changing it where necessary and looking for any anomalies in the data.

The table for covariates summarizes as follows:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | PatientID | Age | Sex (1=male, 2=female) | Smoker (1=yes, 2=no) | VAS-at-inclusion | Vas-12months |
| count | 118.00 | 118.00 | 118.0 | 118.00 | 118.00 | 116.00 |
| mean | 75.45 | 40.86 | 1.5 | 1.67 | 6.02 | 3.62 |
| std | 43.33 | 10.14 | 0.5 | 0.47 | 2.66 | 3.08 |

We can see that the column names are not very data-friendly and that there is a difference in the number of columns between Vas-12months and the other columns. This difference is due to two missing values in the Vas-12months column (NaN values) as shown below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| PatientID | Age | Sex | Smoker | VAS0 | VAS12 |
| 42 | 27 | Male | No | 6.0 | NaN |
| 51 | 35 | Female | Yes | 7.5 | NaN |

The biomarkers dataset has a stray number at the end of the table that is due to a mistake or a human error, so after excluding it, the table summarizes as follows:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | IL-8 | VEGF-A | OPG | TGF-beta-1 | IL-6 | CXCL9 | CXCL1 | IL-18 | CSF-1 |
| count | 348.00 | 347.00 | 347.00 | 347.00 | 347.00 | 347.00 | 347.00 | 347.00 | 347.00 |
| mean | 7.41 | 11.66 | 10.67 | 7.95 | 3.25 | 6.47 | 8.29 | 8.29 | 8.54 |
| std | 0.96 | 0.67 | 0.40 | 0.95 | 0.95 | 0.79 | 1.25 | 0.58 | 0.25 |

Statistical hypothesis testing

For this part, I have chosen the question: “**Do the levels of biomarkers vary at inclusion between males and females?**”

The answer to this question lies in the comparison between males and females of the mean level of each protein and to test if this difference is significant or not. I find the question interesting because, for each of the eight biomarkers, the observations will be divided into observations for males and observations for females and then compare the difference between the two means. From a biomedical perspective, it’s also interesting to explore areas of significant differences between the two sexes and the areas of non-significant or no-differences. The biomarkers dataset contains observations at different timepoints for each individual so for our test we will be using only the observations from the time of inclusion, thus having the value “0weeks” in their Biomarker column.

Our hypothesis would be:

**H0:** The level of biomarker {n} at inclusion, does not vary between males and females ⬄ μ-male = μ-female

**H1:** The level of biomarker {n} at inclusion, varies between males and females ⬄ μ-male ≠ μ-female

The distributions are thus the samples of male and female patients at inclusion for each biomarker, and the random variable is the sample mean because its value depends on the random sample of patients selected for this research.

We will be using Student’s t-test when testing for significant mean differences between the two groups (males and females) as the population’s true variance is unknown. (Agrawal, 2019)

Carrying out inference on a mean requires some pre-conditions, on which the accuracy of our methods depends, to be satisfied. The conditions are: **Randomness** of the sample; **Independence** of the individual observations; And **Normality** of the sample. (Khan Academy, n.d.)

**The initial assumption** that we must make here is that **the sample is random, i.e. the patients were randomly** **selected.** Though we don’t have any information about the data collection mechanism that was employed, it is important to make this assumption to perform the hypothesis test.

The second **condition** for hypothesis testing is **independence**. We also must assume that the individual measurements taken at inclusion for the patients are independent.

Finally, the **third condition of normality** must be checked in two ways: First, by **visually** analyzing the data and its distribution (for each biomarker and sex) and by using a **normality test**. [Annexe 1](#Annex1) shows the distribution of biomarkers’ levels at inclusion disaggregated by sex. It is clear from these histograms that some of the biomarkers’ distributions are either heavily skewed or don’t appear to be normally distributed which can be an issue if we were to conduct hypothesis testing using Student’s T-test. To numerically verify the non-normality of the distributions, we will use the omnibus K2 statistic of D’Agostino and Pearson’s test that combines skew and kurtosis to produce a test of normality. (SciPy.org, n.d.)

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | p-value | Test Result (α=0.01) |
| Biomarker | **Sex** |  |  |
| CSF-1 | **Male** | 0.923 | Cannot Reject / Is Normal |
| **Female** | 0.012 | Cannot Reject / Is Normal |
| CXCL1 | **Male** | 0.000 | Reject / Is not Normal |
| **Female** | 0.048 | Cannot Reject / Is Normal |
| CXCL9 | **Male** | 0.000 | Reject / Is not Normal |
| **Female** | 0.000 | Reject / Is not Normal |
| IL-18 | **Male** | 0.675 | Cannot Reject / Is Normal |
| **Female** | 0.900 | Cannot Reject / Is Normal |
| IL-6 | **Male** | 0.038 | Cannot Reject / Is Normal |
| **Female** | 0.000 | Reject / Is not Normal |
| IL-8 | **Male** | 0.828 | Cannot Reject / Is Normal |
| **Female** | 0.000 | Reject / Is not Normal |
| OPG | **Male** | 0.003 | Reject / Is not Normal |
| **Female** | 0.237 | Cannot Reject / Is Normal |
| TGF-beta-1 | **Male** | 0.000 | Reject / Is not Normal |
| **Female** | 0.051 | Cannot Reject / Is Normal |
| VEGF-A | **Male** | 0.092 | Cannot Reject / Is Normal |
| **Female** | 0.406 | Cannot Reject / Is Normal |

The above table shows that 7 out of the 18 samples failed the normality test. (The null hypothesis that the distribution is normal was rejected). To deal with the normality issue, a boxplot is generated for each sample. The boxplot in [Annexe 2](#Annex2) shows that some samples have outliers that affect both the sample mean and its normality.

To remove the outliers, we use the IQR score method which finds first where the data is majorly situated and then removes any value that is beyond a threshold. (Sharma, 2018)

The IQR score method works as follows:

Q1 = 25th percentile; Q3 = 75th percentile; IQR = Q3 − Q1; Threshold = 1.5;

Outlier < [Q1 - 1.5 \* IQR] OR Outlier > [Q3 + 1.5 \* IQR]

Following the removal of the outliers, performing the normality test again gave a better result. Only two samples did not pass the test. Annex 3 shows a table with the test results and Annex 4 and Annex 5 show the histograms and the boxplots of the samples after removing the outliers. Having a sample size large enough (≈ 55 observations/sample) we can now safely conduct the t-test as the three condition are fulfilled.

Now running the t-test gives the following result:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | α | β | power | t-stat | p-value | t-test result |
| IL-8 | **0.05** | 0.902156 | 0.0978439 | -0.646581 | 0.519261 | **-** |
| VEGF-A | **0.05** | 0.465889 | 0.534111 | -2.06075 | 0.0416201 | **Reject H0** |
| OPG | **0.05** | 0.223169 | 0.776831 | -2.83788 | 0.00558852 | **Reject H0** |
| TGF-beta-1 | **0.05** | 0.482739 | 0.517261 | -2.02025 | 0.0456818 | **Reject H0** |
| IL-6 | **0.05** | 0.945143 | 0.0548574 | -0.207751 | 0.835801 | **-** |
| CXCL9 | **0.05** | 0.909472 | 0.0905281 | 0.595527 | 0.552719 | **-** |
| CXCL1 | **0.05** | 0.209845 | 0.790155 | -2.79027 | 0.00616716 | **Reject H0** |
| IL-18 | **0.05** | 0.776017 | 0.223983 | 1.21358 | 0.227463 | **-** |
| CSF-1 | **0.05** | 0.267842 | 0.732158 | -2.5896 | 0.0109716 | **Reject H0** |

In **5 out of 9 tests, the null hypothesis is rejected,** which means that with 95% confidence, we infer that **the level of biomarkers at inclusion varies significantly between males and females.**

Since we are running nine consecutive tests, **the probability of making at least one type I error is 37.0%** which is a very high percentage.

When dealing with multiple t-tests α should be adjusted in a way, so that the probability of observing at least one significant result due to chance remains below our desired significance level. (Goldman, 2008)

One solution to reduce this high probability is to use the Bonferroni correction which divides the significance level α by the number of tests (). **By lowering the value of α, it becomes harder to reject the null hypothesis, thus reducing the probability of a type I error.** (Khan Academy, n.d.) Following this correction on α and as per the following table, we do not reject **any of the hypothesis**, but we notice a large increase in the values of β:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Biomarker | α | β | power | t-stat | p-value | t-test result |
| IL-8 | **0.00555556** | **0.983487** | 0.0165133 | -0.646581 | 0.519261 | - |
| VEGF-A | **0.00555556** | **0.771724** | 0.228276 | -2.06075 | 0.0416201 | - |
| OPG | **0.00555556** | **0.530471** | 0.469529 | -2.83788 | 0.00558852 | - |
| TGF-beta-1 | **0.00555556** | **0.784218** | 0.215782 | -2.02025 | 0.0456818 | - |
| IL-6 | **0.00555556** | **0.993445** | 0.0065553 | -0.207751 | 0.835801 | - |
| CXCL9 | **0.00555556** | **0.985322** | 0.0146781 | 0.595527 | 0.552719 | - |
| CXCL1 | **0.00555556** | **0.511661** | 0.488339 | -2.79027 | 0.00616716 | - |
| IL-18 | **0.00555556** | **0.943432** | 0.0565683 | 1.21358 | 0.227463 | - |
| CSF-1 | **0.00555556** | **0.585287** | 0.414713 | -2.5896 | 0.0109716 | - |

β is the probability of a Type II error, i.e. P(Not rejecting a false H0).

Conclusion

In conclusion, though we did not reject any of the hypothesis when we applied the Bonferroni correction, as a result of decreasing the value α, we have primarily increased the probability of having false-negative tests as a tradeoff. **That is the reason why the Bonferroni correction is criticized for being conservative** (Koehrsen, 2018) **and other corrections are more recommended, such as Holm’s Step-Down Procedure which is more powerful.** (Zheng, 2018)

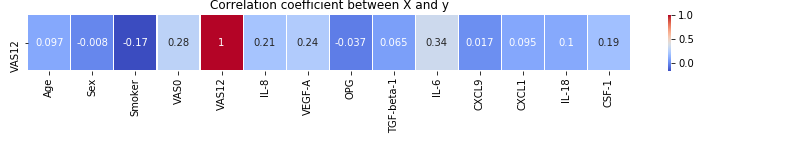
Regression modelling

For the regression modelling, we will use a **linear regression model** with **multiple explanatory variables** and try to fit a line to the data using the **least-squares** method. Our **response variable** is the 12-months VAS.

The explanatory variables used in the model are: Age, Sex, Smoker, VAS0, VAS12, IL-8, VEGF-A, OPG, TGF-beta-1, IL-6, CXCL9, CXCL1, IL-18 & CSF-1

The purpose of the regression model we are fitting to our data is to describe the relationship between the explanatory and the response variables with a linear function and use it to predict values of variables not seen previously by the model then measure how well our model performs (assessing the model fit).

The dataset is divided into two sets, a **testing set** and a **training set**. Observations will be randomly selected and placed in the two sets with a division of 80%/20% for training/test.

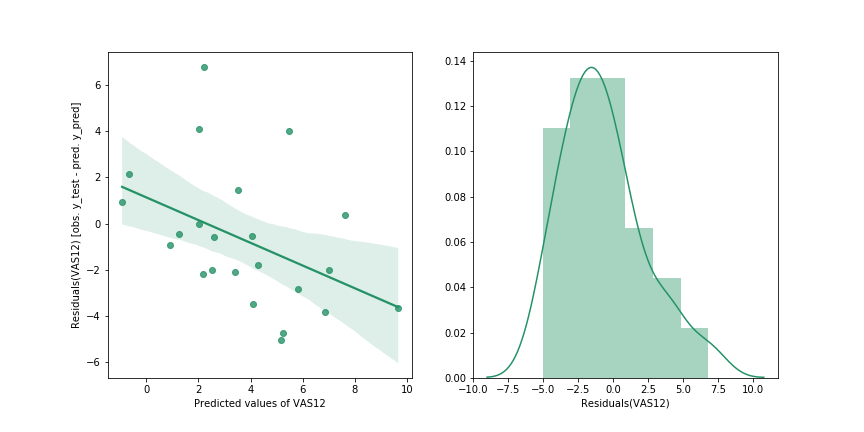
The figure on the right shows the correlation between the response variable and the explanatory variables, and it gives a clear indication of the independence of VAS12 from the other variables since the absolute value of the correlation coefficients is relatively small for all variables.

The results of the fitted parameter values are displayed in the table to the right.

Table Parameter Values of the fitted model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Coef | std err | t | P>|t| |
| Intercept | -3.2928 | 10.893 | -0.302 | 0.763 |
| Age | -0.0195 | 0.034 | -0.57 | 0.571 |
| Sex | -0.3175 | 0.618 | -0.514 | 0.609 |
| Smoker | -0.1581 | 0.66 | -0.239 | 0.811 |
| VAS0 | 0.292 | 0.114 | 2.556 | **0.013** |
| IL-8 | 1.1174 | 0.62 | 1.801 | 0.076 |
| VEGF-A | 1.4888 | 0.761 | 1.957 | 0.054 |
| OPG | -2.1262 | 0.827 | -2.572 | **0.012** |
| TGF-beta-1 | -1.7945 | 0.714 | -2.515 | **0.014** |
| IL-6 | 1.2544 | 0.379 | 3.306 | **0.001** |
| CXCL9 | -0.7432 | 0.377 | -1.972 | 0.052 |
| CXCL1 | 0.0864 | 0.539 | 0.16 | 0.873 |
| IL-18 | -0.3928 | 0.543 | -0.723 | 0.472 |
| CSF-1 | 2.4341 | 1.462 | 1.664 | 0.1 |

The coefficient of determination is R2=0.365. It implies that 36% of the variability of the response variable has been accounted for, and the remaining 64% of the variability is still unaccounted for. (Wikipedia, n.d.) The coefficient of determination is calculated here based on the result of the training data that is used to fit the model. However, if we recalculate this coefficient on predicted variables from the explanatory variables that are not used while fitting the model (the other 30% of data in the testing set), then the result of R2 turns out to be much lower, R2=-0.048. The values of both R2 are a clear indication that to model is performing very poorly on both known input variables (from the training set) and new input variables (the test set).

A scatter plot of the residuals and the predicted values of VAS12 shows that the residuals follow a descending pattern as the values of the predicted variable increases which is not a good indicator, and which indicates that there is an existing relationship between them that should not be there. The histogram also indicates that the shape of the distribution of residuals is not normal, and thus, we cannot make an assumption that the errors in the model are normally distributed.

Conclusion

In conclusion, the model is not useful for predicting the 12- month VAS of patients and corrections should be applied to the data to improve the performance of the model.

**Possible remedies to improve the model:**

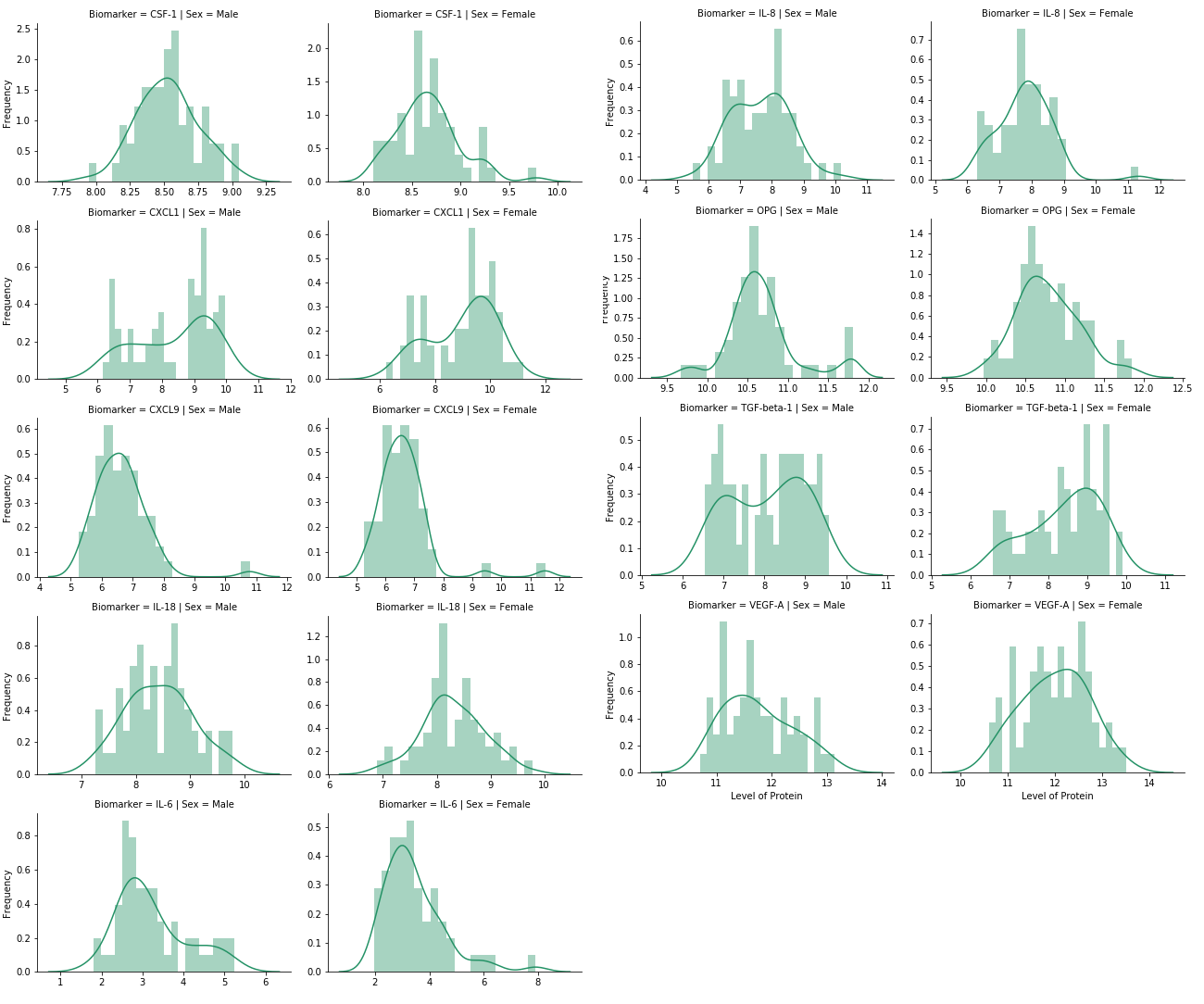
A screenshot of a computer

Description automatically generatedIf we plot a heatmap of the correlation coefficients between the explanatory variables, we notice that some variables are highly correlated such as CXCL1 with IL-8, VGF-A and TGF-beta-1. These strong correlations can cause the linear regression model’s performance to suffer. While we have 9 different predictor variables, some of them appear to be a linear combination of the others, so they don’t add any information. If the exact linear relationship holds among more than two variables, we talk about **multicollinearity**. In this case, we can safely remove one of the two variables and try to fit the model again. (Why Collinearity Is a Problem). Another area of improvement could be the standardization of data since it has varying scales, and the regressor can make assumptions about these scales. For instance, age ranges between 18 and 59 while VEGF-A ranges between 10.6 and 13.5 so the regressor can allocate more weight to the age then the biomarker VEGF-A to make a prediction. Finally, the table that describes the parameter values of the fitted model has some rows highlighted in yellow for some variables that have a p-value less than 0.05 (for a 95% confidence). These p-values mean that we can reject the null hypothesis and suggests that changes in the predictor variable are associated with changes in the response variable. This means that the other variables that are not statically significant and can be removed. Finally, I was able to get a much better R2 score by using only the predictors VAS0, OPG and IL-6.

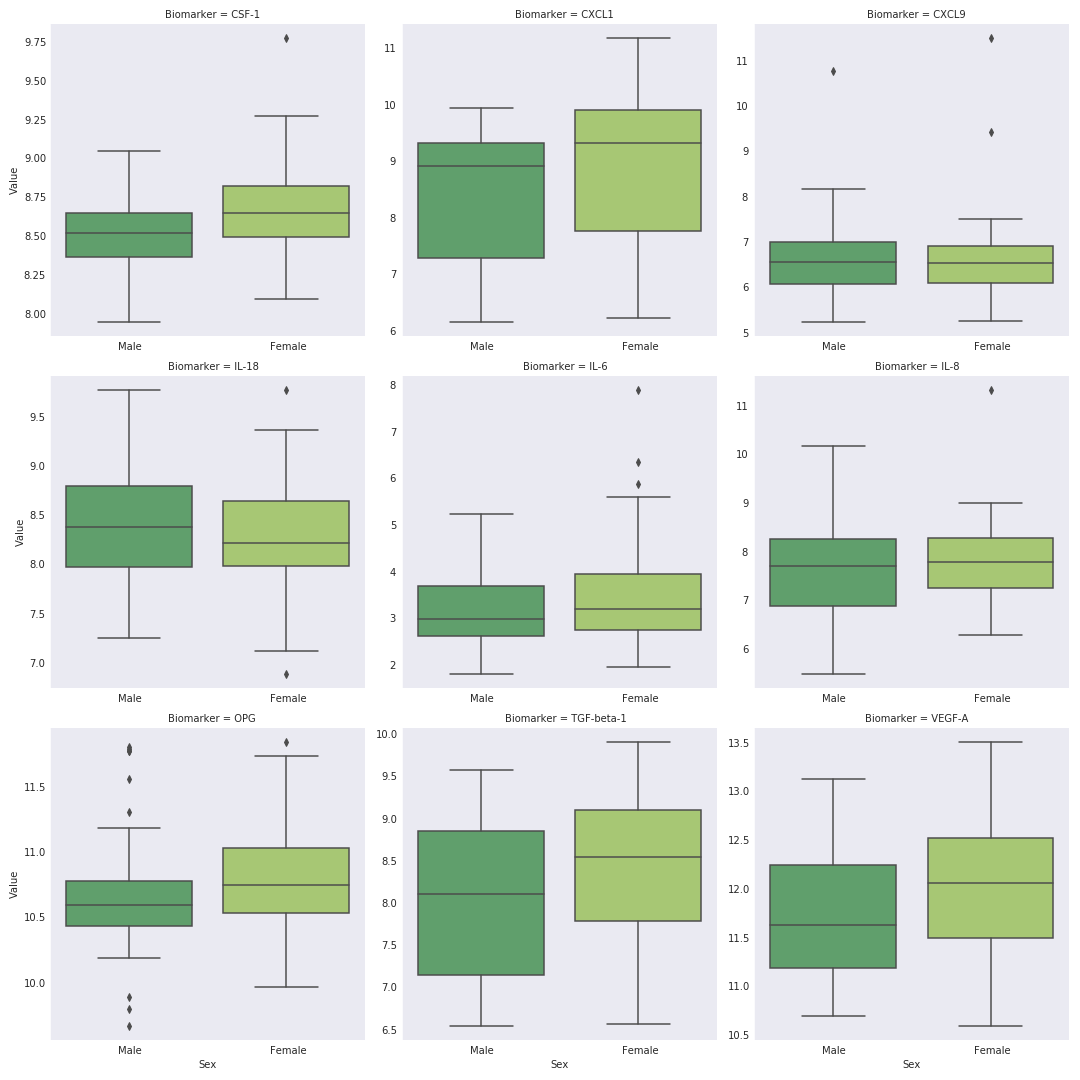
Figure The correlation coefficients between the different explanatory variables

Annexe

1. Distribution of biomarkers’ levels at inclusion disaggregated by sex



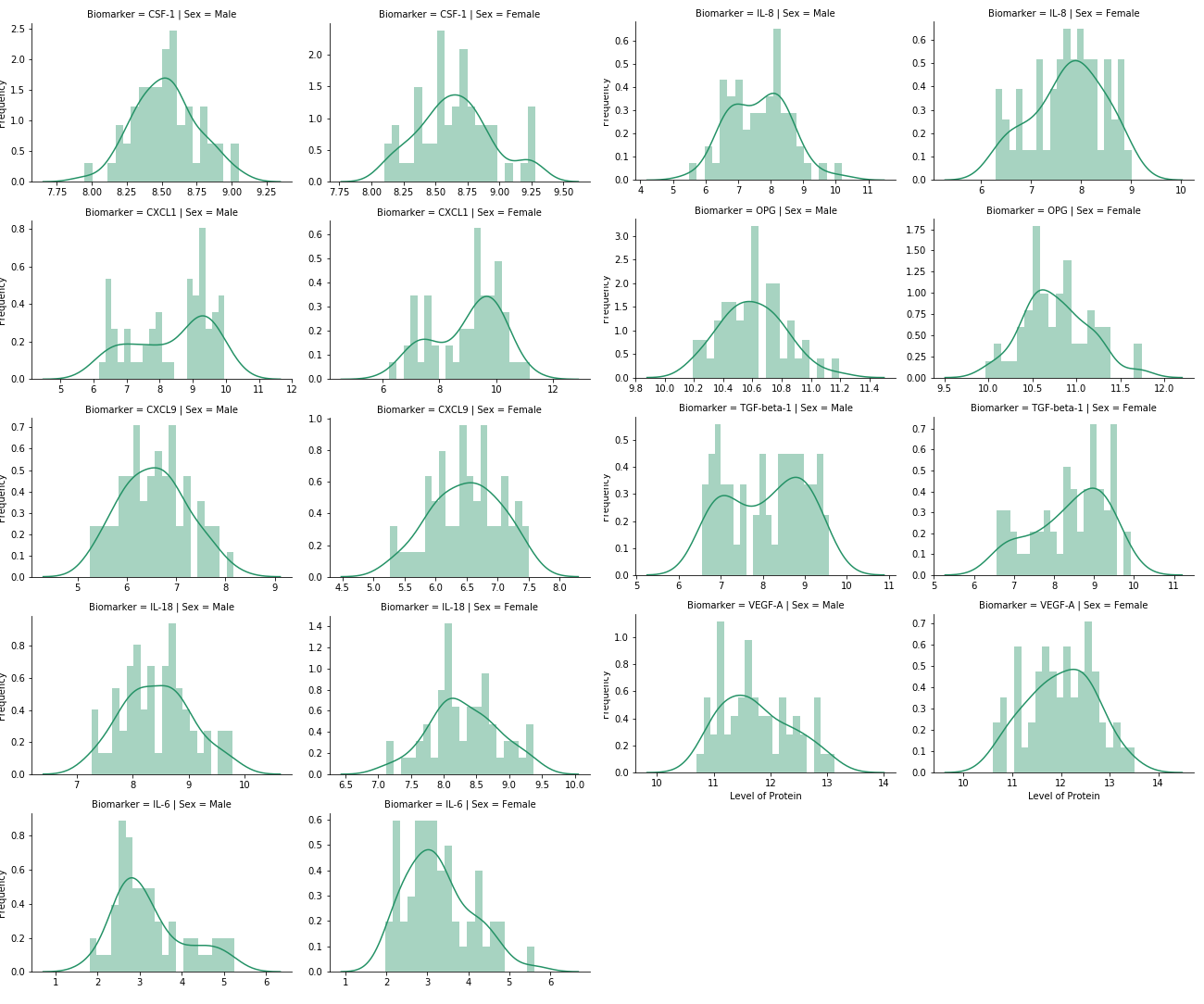
1. Boxplot of biomarkers’ levels at inclusion disaggregated by sex



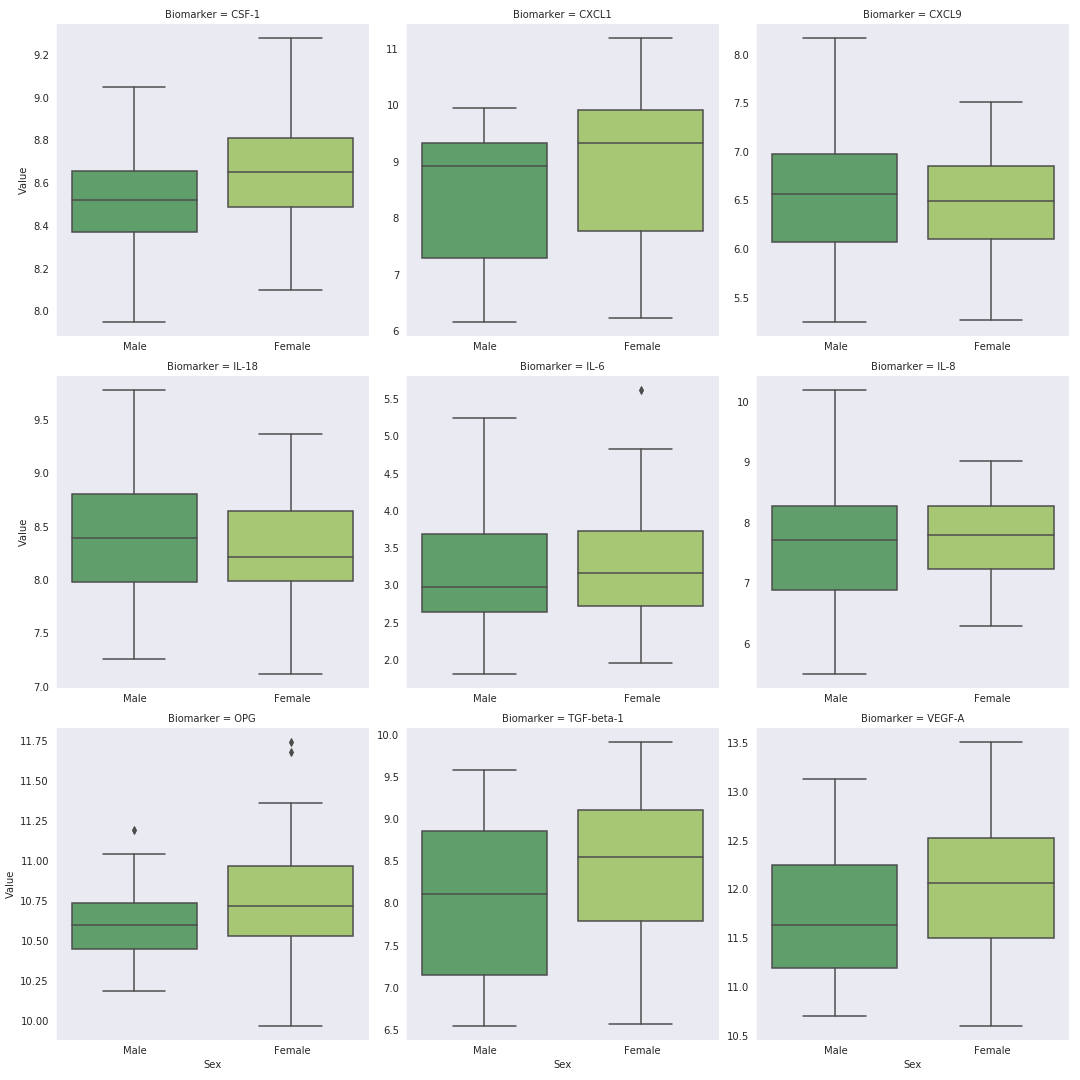
1. The normality test result of biomarkers after removing the outliers

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | p-value | Test Result (α=0.01) |
| Biomarker | **Sex** |  |  |
| CSF-1 | **Female** | 0.650 | Cannot Reject / Is Normal |
| **Male** | 0.923 | Cannot Reject / Is Normal |
| CXCL1 | **Female** | 0.048 | Cannot Reject / Is Normal |
| **Male** | 0.000 | Reject / Is not Normal |
| CXCL9 | **Female** | 0.391 | Cannot Reject / Is Normal |
| **Male** | 0.594 | Cannot Reject / Is Normal |
| IL-18 | **Female** | 0.946 | Cannot Reject / Is Normal |
| **Male** | 0.675 | Cannot Reject / Is Normal |
| IL-6 | **Female** | 0.110 | Cannot Reject / Is Normal |
| **Male** | 0.038 | Cannot Reject / Is Normal |
| IL-8 | **Female** | 0.201 | Cannot Reject / Is Normal |
| **Male** | 0.828 | Cannot Reject / Is Normal |
| OPG | **Female** | 0.535 | Cannot Reject / Is Normal |
| **Male** | 0.668 | Cannot Reject / Is Normal |
| TGF-beta-1 | **Female** | 0.051 | Cannot Reject / Is Normal |
| **Male** | 0.000 | Reject / Is not Normal |
| VEGF-A | **Female** | 0.406 | Cannot Reject / Is Normal |
| **Male** | 0.092 | Cannot Reject / Is Normal |

1. Distribution of biomarkers’ levels at inclusion disaggregated by sex after removing the outliers



1. Boxplot of biomarkers’ levels at inclusion disaggregated by sex after removing the outliers



Bibliography

Khan Academy. (n.d.). *Reference: Conditions for inference on a mean*. Retrieved from Khan Academy: https://www.khanacademy.org/math/ap-statistics/estimating-confidence-ap/one-sample-t-interval-mean/a/reference-conditions-inference-one-mean

Agrawal, Y. (2019, January 21). *Hypothesis testing in Machine learning using Python*. Retrieved from towards data science: https://towardsdatascience.com/hypothesis-testing-in-machine-learning-using-python-a0dc89e169ce https://www.xlstat.com/en/solutions/features/two-sample-t-and-z-tests

SciPy.org. (n.d.). *scipy.stats.normaltest*. Retrieved from SciPy.org: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.normaltest.html#r7bf2e556f491-1

Sharma, N. (2018, May 22). *Ways to Detect and Remove the Outliers*. Retrieved from towards data science: https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba

Goldman, M. (2008). *Spring 2008 - Stat C141/ Bioeng C141 - Statistics for Bioinformatics*. Retrieved from Berkeley: https://www.stat.berkeley.edu/~mgoldman/Section0402.pdf

Khan Academy. (n.d.). *Consequences of errors and significance*. Retrieved from Khan Academy: https://www.khanacademy.org/math/statistics-probability/significance-tests-one-sample/error-probabilities-and-power/a/consequences-errors-significance

Koehrsen, W. (2018, February 7). *The Misleading Effect of Noise: The Multiple Comparisons Problem*. Retrieved from towards data science: https://towardsdatascience.com/the-multiple-comparisons-problem-e5573e8b9578

Zheng, P. (2018, August 7). *An Overview of Methods to Address the Multiple Comparison Problem*. Retrieved from towards data science: https://towardsdatascience.com/an-overview-of-methods-to-address-the-multiple-comparison-problem-310427b3ba92

Wikipedia. (n.d.). *Coefficient of determination*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Coefficient\_of\_determination

*Why Collinearity Is a Problem.* (n.d.). Retrieved from Carnegie Mellon University: https://www.stat.cmu.edu/~larry/=stat401/lecture-17.pdf

The Source Code:

The language that was used for the coding is in Python 3.7 and it was written inside a Jupyter Notebook for clarity. The libraries used in the code are Pandas, Scipy, Numpy, Statsmodel and Seaborn.

The notebook is exported and attached in the following page and [you can use this link here to download](https://gofile.io/?c=tKz0zM) it and run it on your computer given that you have the correct python environment and Jupyter Notebooks or Jupyter Labs.