Probability and Statistics Final Report

2025-11-11

Introduction

This report addresses the statistical analysis and regression modelling of 9 randomly selected biomarkers related to inflammation in patients with disc herniation (Moen, 2016). The original data contains levels of 9 biomarkers at inclusion, at 6 weeks and at 12 months along with covariates such as Patient age, sex, smoking status and their reported pain on a scale form 1 to 10 at inclusion and after 12 months (VAS).

Among other covariates, understanding the effect of sex and gender on health is instrumental in delivering appropriate treatment and tailoring dosages of drugs. The first requirement to include women as well as men in medical trials and result analysis was 30 years ago, and progress has been slow and the possibility of sex-specific treatment response is still understudied (Mazure and Jones, 2015). Studying the biomarker levels by sex can help develop new approaches to prevention, diagnosis and treatment specific to patients' sex. If biomarker levels differ between males and females, it suggests a different patophysiology of the condition for men and women. This can help in future applications to advise whether the treatment needs to be tailored based on sex. To narrow down the research question, only biomarker levels at inclusion will be analysed. Hence, the first part of this report will focus on the question: Do the levels of each biomarker vary between males and females at inclusion? The following hypotheses were constructed to answer this question:

- H0: Mean level of biomarker X at inclusion is the same for males and females.
- H1: Mean level of biomarker X at inclusion differs between males and females.

In the second part of the report, a regression model will be constructed to make predictions on the patients' recovery (pain after 12 months) based on the available covariates, including patients' sex, age, smoking, and pain levels.

Methods

Data manipulation, statistical hypothesis testing and regression modelling were performed in R. The full scripts are available in the Appendix.

Data Manipulation

The original datasets biomarkers.xlsx and covariates.xlsx were collated into a single dataframe and saved as a csv file "biomarkers_covariates_clean.csv".

The column names were simplified for easier manipulation, and only values of biomarkers at inclusion were selected for the purposes of the statistical hypothesis testing and regression modelling described below.

2 NAs were found in VAS at 12 months (pain levels after one year). Normality was checked visually with histograms to make assumptions for statistical tests used below.

The R script used for the data manipulation is available in Appendix 1.

Statistical Hypothesis testing

Hypotheses:

- H0: Mean level of biomarker X at inclusion is the same for males and females.
- H1: Mean level of biomarker X at inclusion differs between males and females.

$$H_0 : \mu_{xi} = \mu_{yi}$$
$$H_1 : \mu_{xi} \neq \mu_{yi}$$

where μ_{xi} is mean level of biomarker i for males, and μ_{yi} is mean level of biomarker i for females

Random Variables:

X_i : Level of biomarker i for a randomly selected male patient.

Y_i : Level of biomarker i for a randomly selected female patient.

Distribution:

While most levels of biomarkers for males and females appear to be approximately normally distributed, levels of TGF_beta_1, IL-6, CXCL9 are not normally distributed for either sexes, and level of IL-18 for male patients is not normally distributed. However, we will assume the random variables are independent with normal distributions using the Central Limit Theorem, as sample size of each group is large (>30), and the underlying sampling distribution can be approximately normal.

Test:

A Welch two-sample t-test was used to compare the means of two groups (male, female). Welch's t-test was chosen as it performs better than Student's t-test sample sizes and variances between groups, and gives the same result when sample sizes and variances are equal (Delacre et al., 2022). The Welch t-test is a robust method that can deal with unequal variances and tolerate slight departure from normality.

Multiple Hypothesis Testing:

When doing multiple independent test, the probability of producing Type I error (wrongly accepting the alternative hypothesis) increases (Herzog et al., 2019). In our case, if the null hypotheses are true and the t-tests are all independent, the probability of making at least one Type I error within the 9 t-tests is:

$$1 - (1 - \alpha)^9$$

For α 0.05, the probability of making at least one Type I error is 0.37.

Bonferroni Correction

Regression Modelling

Results

Discussion

Conclusion

References

herzog moen mazure and jones delacre

Appendices

Appendix 1: Initial Data Manipulation Script