Probability and Statistics Final Report: Difference in Means and Regression Modelling

s1861053

The code for this report is hosted at: https://github.com/sharkie58/probability-and-statistics-report.git.

Introduction

This report addresses the statistical analysis and regression modelling of 9 randomly selected biomarkers related to inflammation in patients with disc herniation. 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 at inclusion and VAS at 12 months).

Among other covariates, understanding the effect of sex and gender on health is instrumental in delivering appropriate treatment and tailoring dosages of drugs. The possibility of sex-specific treatment response is still understudied (Mazure and Jones, 2015), and studying the biomarker levels by sex could help develop new approaches to prevention, diagnosis and treatment specific to patients' sex. Hence, the first part of this report focused on the question: Do the levels of each biomarker vary between males and females at inclusion?

In the second part of the report, a regression model was constructed to make predictions on the patients' recovery based on the available covariates, including patients' sex, age, smoking, and pain levels at inclusion.

Methods

Data manipulation, statistical hypothesis testing and regression modelling were performed in R. The software specifications are included in Appendix 1, and full scripts are available in Appendix 2, 3, and 4.

Data Manipulation

The original datasets biomarkers.xlsx and covariates.xlsx were collated into a single dataframe using left join to only keep patients with complete biomarker values 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), these observations were deleted for the regression modelling. Normality was checked visually with histograms.

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, normal distributions were assumed using the Central Limit Theorem, as sample size of each group is large (>30).

Test:

A Welch two-sample t-test was used to compare the means of two groups (male, female). Welch's t-test performs better than Student's t-test for unequal sample sizes and variances between groups, and gives the same result when sample sizes and variances are equal (Delacre et al., 2022). Furthermore, variances do not need to be pre-tested (Rasch et al., 2011). The Welch t-test was chosen as it is a standard and 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 $\alpha = 0.05$, the probability of making at least one Type I error is 0.37.

Bonferroni Correction

Bonferroni's correction was used to deal with the increasing Type I error with multiple hypothesis testing. Instead of having the probability of Type I error, $\alpha = 0.05$, for all independent tests, this was set as the probability of Type I error across all 9 tests, lowering the p value for each individual test to 0.006 to reach significance. The adjusted p-value accounted for this.

Regression Modelling

An 80-20 split was performed on the data to obtain a training and a testing dataset to adequately estimate true model performance on test data. The following assumptions of linear regression (Thulin, 2025) were fulfilled: the model is linear in the parameters and the random errors are homoscedastic and normally distributed (Appendix 5). It was assumed the observations were independent. A multiple regression model was created with VAS at 12 months as the response variable and all biomarkers and all other covariates as explanatory variables. Predictions were made and were plotted against previously split test data.

As multiple explanatory variables were not significantly related to the response variable (Appendix 6), models with only significant variables were produced in iteration, resulting in another model with only IL-6 and VAS at inclusion as explanatory variables (Appendix 7). The predictions of this smaller model were fitted to the actual test values.

Results

Statistical Hypothesis Testing

Welch Two-Sample t-test

Significant difference was found using a Welch t-test without a correction for multiple hypothesis testing, between males and females for four biomarker levels: CSF_1, CXCL1, TGF_beta_1, and VEGF_A. The corresponding difference between means, means of the two groups, p-values and 95% confidence intervals are given in Table 1 below.

Table 1: Welch t-test results comparing levels of 9 biomarkers between males and females, including mean values, difference between means, p-values, and 95% confidence intervals.

X	biomarker	difference	male	female	p	conf.low	conf.high
1	csf_1	-0.145	8.53	8.67	0.006	-0.25	-0.04

X	biomarker	difference	male	female	p	conf.low	conf.high
2	cxcl1	-0.619	8.36	8.98	0.006	-1.06	-0.18
3	cxcl9	0.003	6.63	6.62	0.985	-0.33	0.33
4	il_18	0.128	8.43	8.31	0.249	-0.09	0.35
5	il_6	-0.211	3.23	3.44	0.253	-0.58	0.15
6	il_8	-0.161	7.64	7.80	0.329	-0.49	0.16
7	opg	-0.118	10.67	10.79	0.132	-0.27	0.04
8	tgf_beta_1	-0.345	8.07	8.42	0.046	-0.68	-0.01
9	vegf_a	-0.252	11.75	12.00	0.042	-0.49	-0.01

Welch Two-Sample t-test with Bonferroni Correction

There was no significant difference between male and female levels of the 9 biomarkers when the p-values were adjusted with Bonferroni's correction. Results are given in Table 2 below.

Table 2: Welch t-test results with Bonferroni correction comparing levels of 9 biomarkers between males and females, including mean values, difference between means, adjusted p-values, and 95% confidence intervals.

X	biomarker	difference	male	female	p_adj	conf.low	conf.high
1	csf_1	-0.145	8.53	8.67	0.055	-0.25	-0.04
2	cxcl1	-0.619	8.36	8.98	0.056	-1.06	-0.18
3	cxcl9	0.003	6.63	6.62	1.000	-0.33	0.33
4	il_18	0.128	8.43	8.31	1.000	-0.09	0.35
5	il_6	-0.211	3.23	3.44	1.000	-0.58	0.15
6	il_8	-0.161	7.64	7.80	1.000	-0.49	0.16
7	opg	-0.118	10.67	10.79	1.000	-0.27	0.04
8	tgf_beta_1	-0.345	8.07	8.42	0.411	-0.68	-0.01
9	vegf_a	-0.252	11.75	12.00	0.374	-0.49	-0.01

Regression Modelling

The model showed a low adjusted R-squared (Adj. R-squared = 0.26), and only 5 explanatory variables having a significant relationship with VAS at 12 months (out of 13 explanatory variables).

The predicted values underestimated the actual values of VAS at 12 months (Fig. 1), with heteroscedastic residuals and RMSE 3.20.The smaller model showed a lower adjusted R-squared (Adj. R-squared = 0.23), a higher RMSE (RMSE = 3.26), and similarly underestimating predictions for VAS at 12 months (Appendix 8).

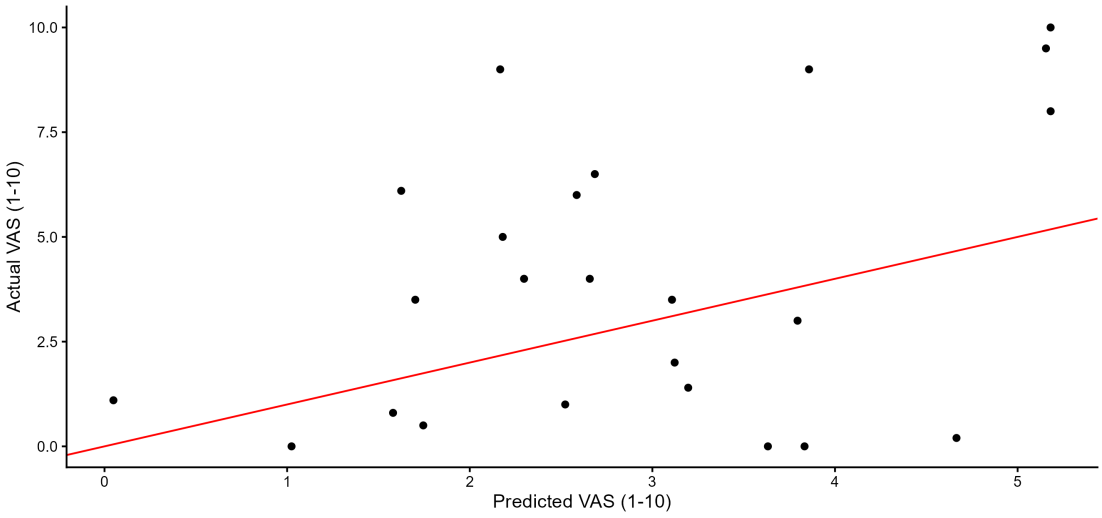


Figure 1: Predicted values of VAS at 12 months from multiple regression model using train data, plotted against actual VAS at 12 months from test data. RMSE = 3.20.

Conclusions

Four biomarkers (CSF-1, CXCL1, TDF-beta-1, and VEGF_A, Table 1) showed a significant difference between males and females. However, as there was no correction for multiple testing, the probability that one of these is a Type I error is 0.37. The evidence in Table 2 shows no significant difference in means between males and females when testing hypotheses with Bonferroni's correction, hence the null hypothesis was accepted.

Bonferroni's correction is the most commonly used correction for multiple testing, however it has its own limitations, such as the assumption of all null hypotheses to be true (Herzog, 2019). This correction lowers the alpha to ensure the probability of committing one or more Type I errors is less than 0.05, however is more relevant for cases with joint intersection of hypotheses, as Type I errors are independent of each other and can coexist (Garcia Perez, 2023). As such, other corrections can be tested in the future, such as Holms correction, to compare results and determine whether any of the selected biomarkers have different levels for males and females.

The multiple regression model did not perform well in predicting the pain outcomes for patients after 1 year from all biomarkers at inclusion and all covariates. It underestimated high VAS values by almost a half (Fig.1), and most of the explanatory variables were not significant in the model. The second model presented in Appendix 6 and 7 only considered explanatory variables that were significant, however its adjusted R-square value was lower and the predictions were underestimating actual values in a similar trend.

The model could potentially be improved by testing for interactions in the model, including more covariates, or more closely assessing whether linear regression is appropriate. The large underestimation of the predictions could suggest another probability distribution may be needed.

References

Datanovia. (2020). How to Perform Multiple T-test in R for Different Variables - Datanovia. https://www.datanovia.com/en/blog/how-to-perform-multiple-t-test-in-r-for-different-variables/

Delacre, M., Lakens, D., & Leys, C. (2022). Why Psychologists Should by Default Use Welch's t-test Instead of Student's t-test. International Review of Social Psychology, 35(1). https://doi.org/10.5334/IRSP.661

García-Pérez, M. A. (2023). Use and misuse of corrections for multiple testing. Methods in Psychology, 8, 100120. https://doi.org/10.1016/J.METIP.2023.100120

Herzog, M. H., Francis, G., & Clarke, A. (2019). The Multiple Testing Problem. 63–66. https://doi.org/10.1007/978-3-030-03499-3_5

Mazure, C. M., & Jones, D. P. (2015). Twenty years and still counting: including women as participants and studying sex and gender in biomedical research. BMC Women's Health 2015 15:1, 15(1), 1–16. https://doi.org/10.1186/S12905-015-0251-9

Rasch, D., Kubinger, K. D., & Moder, K. (2009). The two-sample t test: pre-testing its assumptions does not pay off. Statistical Papers 2009 52:1, 52(1), 219–231. https://doi.org/10.1007/S00362-009-0224-X

Thulin, M. (2025). 8. Regression Models. In Modern statistics with R : from wrangling and exploring data to inference and predictive modelling. CRC Press.

Appendices

Appendix 1: Software Specifications

Package	Version
knitr	1.50
broom.helpers	1.22.0
ggfortify	0.4.19
rsample	1.3.1
rstatix	0.7.3
gtsummary	2.4.0
broom	1.0.10
ggplot2	4.0.0
janitor	2.2.1
tidyr	1.3.1
dplyr	1.1.4
readxl	1.4.5
gt	1.1.0

R version: 4.3.1
Operating system: Windows 11

Appendix 2: Initial Data Manipulation Script

```r
# Probability and Statistics
# Final Assessment: Data Manipulation Script

# Load packages ------------------------------------------------------------
library(readxl) # for reading excel files
library(dplyr) # for pipe operators and manipulating data frames
library(tidyr) # for manipulating data frames, separate_wider_delim()
library(janitor) # for cleaning column names


# Load data ----------------------------------------------------------------
biomarkers <- read_excel("data/biomarkers.xlsx")
covariates <- read_excel("data/covariates.xlsx")

# Explore data -------------------------------------------------------------
biomarkers
covariates


# Connect biomarkers with covariates datasets ------------------------------

# Separate the column "Biomarkers" to two columns "PatientID" and "Time"
biomarkers_id <- separate_wider_delim(biomarkers,
                                      Biomarker,
                                      delim = "-",
                                      names = c("PatientID", "Time"))
biomarkers_id

# Convert Patient ID in both datasets to numeric
covariates$PatientID <- as.numeric(covariates$PatientID)
biomarkers_id$PatientID <- as.numeric(biomarkers_id$PatientID)

# Calculate number of patients in each dataset
n_patients <- length(unique(biomarkers_id$PatientID))
sprintf("The number of patients in the biomarkers dataset is %s.", n_patients)

# Check that the number of patients in the covariates dataset is the same
n_patients2 <- length(unique(covariates$PatientID))
sprintf("The number of patients in the covariates dataset is %s.", n_patients2)

# Left join datasets to only keep data for patients with biomarker measurements
biomarkers_covariates <- left_join(biomarkers_id, covariates, by = "PatientID")
summary(biomarkers_covariates)
```

```r
# Prepare data for Statistical Analysis and Regression Modelling ----------

# Create a clean dataset of biomarkers at inclusion and all covariates
biomarkers_covariates_clean <- biomarkers_covariates %>%

  # Limit Time to only at inclusion
  filter(Time == "0weeks") %>%

  # Remove redundant Time column
  select(-Time) %>%

  # Reorder the dataset by PatientID
  arrange(PatientID) %>%

  # Rename columns with long names
  rename("Sex" = "Sex (1=male, 2=female)",
         "Smoker" = "Smoker (1=yes, 2=no)",
         "VAS-0" = "VAS-at-inclusion",
         "VAS-12" = "Vas-12months") %>%

  # Clean all column names with the janitor package
  clean_names()


# Check that the number of patients in the clean dataset is the same
n_patients3 <- length(unique(biomarkers_covariates_clean$patient_id))
sprintf("The number of patients in the clean dataset is %s.", n_patients3)

# 1 patient is missing data for inclusion.

# Check for NAs
nas <- sum(is.na(biomarkers_covariates_clean))
sprintf("There are %s NAs in biomarkers_covariates_clean.", nas)

# Find the NAs
colSums(is.na(biomarkers_covariates_clean)) # 2 missing values in VAS after 12 months.

# Check distributions ----------------------------------------------------

female <- biomarkers_covariates_clean %>%
  filter(sex == 2)

hist(female$il_8) # normal
```

```r
hist(female$vegf_a) # normal
hist(female$opg) # normal
hist(female$tgf_beta_1) # not normal
hist(female$il_6) # right skew
hist(female$cxcl9) # right skew
hist(female$cxcl1) # left skew
hist(female$il_18) # normal
hist(female$csf_1) # normal

male <- biomarkers_covariates_clean %>%
  filter(sex == 1)

hist(male$il_8) # normal
hist(male$vegf_a) # normal
hist(male$opg) # normal
hist(male$tgf_beta_1) # not normal
hist(male$il_6) # right skew
hist(male$cxcl9) # right skew
hist(male$cxcl1) # left skew
hist(male$il_18) # not normal
hist(male$csf_1) # normal

# Save data ------------------------------------------------------------------
write.csv(biomarkers_covariates_clean, "data/biomarkers_covariates_clean.csv")
```

Appendix 3: Hypothesis Testing Script

```r
# Probability and Statistics
# Final Assessment: Statistical Hypothesis Testing


# Load packages ------------------------------------------------------------
library(dplyr) # for pipes and manipulation
library(rstatix) # for a t_test (wrapper for t.test)

# Load data ----------------------------------------------------------------

# Load merged dataset and set the first column as index
biomarker <- read.csv("data/biomarkers_covariates_clean.csv", row.names = 1)

# Hypothesis ---------------------------------------------------------------

# Question:
# Do the levels of each biomarker vary between males and females at inclusion?

# 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.

# Welch two sample t-tests -------------------------------------------------

# Use the Welch test to compare the means of each of 9 biomarkers

# To perform t-tests in a single loop, convert data from wide to long format
# The script to perform all 9 tests together was taken from Datanovia (2020).
biomarker_long <- biomarker %>%
  select(il_8:csf_1,sex) %>% # remove columns that won't be used in testing
  pivot_longer(-sex, names_to = "biomarker", values_to = "value")

# See the long dataset
biomarker_long

# Construct Welch t-tests for all 9 biomarkers
h_test <- biomarker_long %>%
  group_by(biomarker) %>%
  t_test(value ~ sex, detailed = TRUE)

# See test results
h_test
```

```r
# Polish the table
h_test_tbl <- h_test %>%
  # select columns to report
  select(biomarker, estimate, estimate1, estimate2, p, conf.low, conf.high) %>%
  # adjust number of significant figures
  mutate(
    estimate = round(estimate, digits = 2),
    estimate1 = round(estimate1, digits = 2),
    estimate2 = round(estimate2, digits = 2),
    p = round(p, digits = 2),
    conf.low = round(conf.low, digits = 2),
    conf.high = round(conf.high, digits = 2)
    ) %>%
  # rename columns to meaningful names
  rename(
    c(difference = estimate,
      male = estimate1,
      female = estimate2)
    )

h_test_tbl

# Bonferroni Correction ----------------------------------------------------

# The probability of making a Type I error is 0.05 across all 9 t-tests, lowering
# the p-value necessary to reach significance for each individual test to 0.006.

# Adjust the hypothesis test with Bonferroni correction
h_test_bonferroni <- biomarker_long %>%
  group_by(biomarker) %>%
  t_test(value ~ sex, detailed = TRUE) %>%
  adjust_pvalue(method = "bonferroni")

# See test results
h_test_bonferroni

# Create a table for test results with Bonferroni correction
bonferroni_tbl <- h_test_bonferroni %>%
  # select columns to report
  select(biomarker, estimate, estimate1, estimate2, p.adj, conf.low, conf.high) %>%
  # adjust number of significant figures
  mutate(
    estimate = round(estimate, digits = 2),
    estimate1 = round(estimate1, digits = 2),
```

```
    estimate2 = round(estimate2, digits = 2),
    p.adj = round(p.adj, digits = 2),
    conf.low = round(conf.low, digits = 2),
    conf.high = round(conf.high, digits = 2)
  ) %>%
  # rename columns to meaningful names
  rename(
    c(difference = estimate,
      male = estimate1,
      female = estimate2,
      p_adj = p.adj)
  )


# See table
bonferroni_tbl

# None of the tests fulfill significance for the alternative hypothesis with
# Bonferroni correction.

# Save tables for report ---------------------------------------------------
write.csv(h_test_tbl, "data/hypotheses_tests_table.csv")
write.csv(bonferroni_tbl, "data/hypotheses_tests_table_bonferroni.csv")
```

Appendix 4: Regression Modelling Script

```
# Probability and Statistics
# Final Assessment: Regression Modelling

# Load packages ------------------------------------------------------------
library(rsample) # for splitting data into test and training datasets
library(ggplot2) # for plotting data
library(ggfortify) # for checking assumptions
library(gtsummary) # for regression summary table
library(broom.helpers) # for regression summary table
library(gt) # to save gtsummary tables


# Load data ----------------------------------------------------------------

# Load merged dataset and set the first column as index
biomarker <- read.csv("data/biomarkers_covariates_clean.csv", row.names = 1)

# Split data ---------------------------------------------------------------

# Set seed for reproduction of random sampling
set.seed(1212)

# Perform a random 80/20 split
split <- initial_split(biomarker, prop = 0.8)

# 80% of values in training dataset
train <- training(split)

# 20% of calues in testing dataset
test <- testing(split)


# Construct a multiple regression model ------------------------------------

# Response variable: 12 month VAS
# Explanatory variables: all 9 biomarker levels at inclusion, age, sex, smoker
# and VAS at inclusion

# 2 observations with missing VAS at 12 months deleted.

# Fit the model
model <- lm(
  vas_12 ~ il_8 + vegf_a + opg + tgf_beta_1 + il_6 + cxcl9 + cxcl1 + il_18 +
```

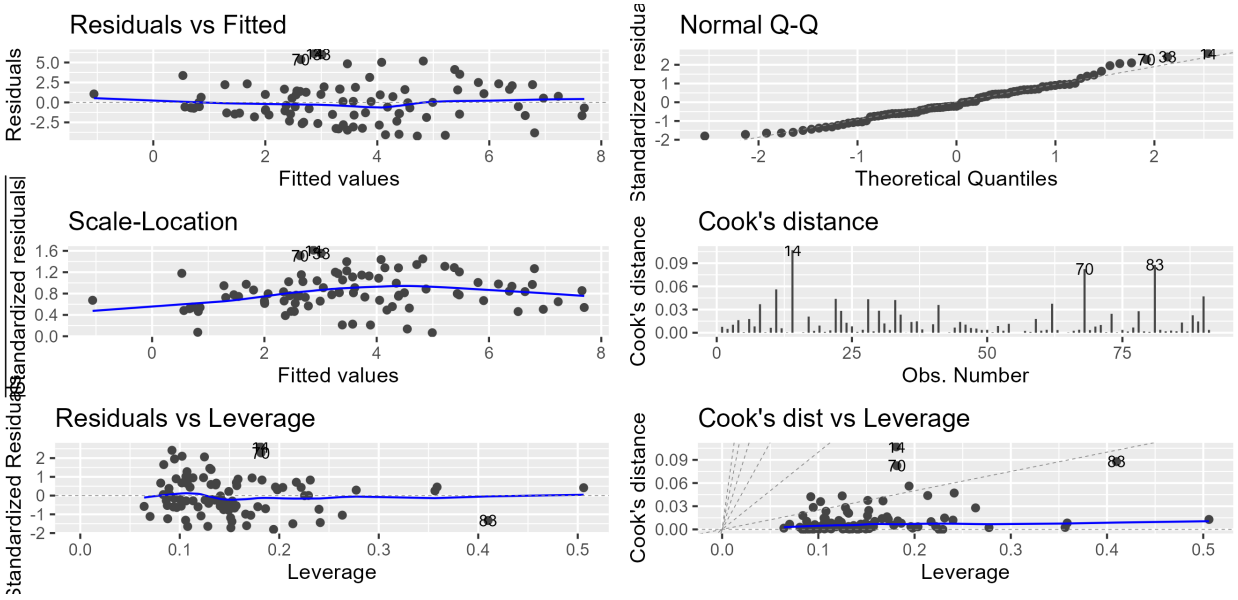
```
    csf_1 + age + sex + smoker + vas_0,
  data = train
  )

# See model summary
summary(model)

# Evaluate the model by checking assumptions  ------------------------------

# Check assumptions (taken from Thulin (2025)):

# Plot residuals using ggfortify package:
residual_plots <- autoplot(model, which = 1:6, ncol = 2, label.size = 3)
residual_plots

# 1. The model is linear in the parameters
# Assumption fulfilled: Residuals vs fitted show a straight line.

# 2. The observations are independent
# We assume they are independent as this is harder to assess visually.

# 3. Homoscedasticity (Random errors have the same variance)
# Assumption fulfilled: Scale-Location plot shows approximately even spread of residua

# 4. Normally distributed random errors
# Assumption fulfilled: Residuals (estimates of random errors) follow a normal
# distribution shown in the Normal QQ plot.


# Create a table for regression results for training data--------------------

# Create table
regression_tbl <- tbl_regression(model,
                                 intercept = TRUE,
                                 conf.level = 0.95,
                                 tidy_fun = broom.helpers::tidy_with_broom_or_parameters
# See table
regression_tbl

# Generate predictions and compare to actual test data ----------------------
predictions <- predict(model, test)

pred_plot <- ggplot(test, aes(x=predictions, y=vas_12)) +
  geom_point() +
```

```
  theme_classic() +
  geom_abline(intercept=0,
              slope=1,
              colour = "red") +
  labs(x='Predicted VAS (1-10)', y='Actual VAS (1-10)')

pred_plot

# Check RMSE
sqrt(mean((test$vas_12 - predictions)^2))


# Create other models with less explanatory variables (not included in report) ----

# Create a model with only variables that have a significant relationship with vas_12
model_narrow1 <- lm(
  vas_12 ~ il_8 + opg + tgf_beta_1 + il_6 + vas_0,
  data = train
)

# See model 1 summary
summary(model_narrow1)

# The p-values of il_8 and tgf_beta_1 having a relationship with vas_12 are not
# significant in the second model.

model_narrow2 <- lm(
  vas_12 ~ opg + il_6 + vas_0,
  data = train
)

# See model 2 summary
summary(model_narrow2)

# Discard opg from the model as it is not significant in the last model.
model_narrow3 <- lm(
  vas_12 ~ il_6 + vas_0,
  data = train
)

# See model 3 summary
summary(model_narrow3)

# The highest Adjusted R-squared was reported in the first model including all
```

```r
# biomarker levels and all covariates.

# Create results table with model_narrow3 ----------------------------------
regression3_tbl <- tbl_regression(model_narrow3,
                                  intercept = TRUE,
                                  conf.level = 0.95,
                                  tidy_fun = broom.helpers::tidy_with_broom_or_parameters
# See table
regression3_tbl

# Fit model 3 to test data -------------------------------------------------
predictions_model3 <- predict(model_narrow3, test)

pred3_plot <- ggplot(test, aes(x=predictions_model3, y=vas_12)) +
  geom_point() +
  theme_classic() +
  geom_abline(intercept=0,
              slope=1,
              colour = "red") +
  labs(x='Predicted VAS (1-10)', y='Actual VAS (1-10)')

pred3_plot

# Check RMSE
sqrt(mean((test$vas_12 - predictions_model3)^2))

# Save results -------------------------------------------------------------

regression_tbl %>%
  as_gt() %>%
  gt::gtsave('figures/regression_table.png')

regression3_tbl %>%
  as_gt() %>%
  gt::gtsave('figures/regression_table_small_model.png')

ggsave(residual_plots,
       filename = "figures/residual_plots.png",
       device = "png")

ggsave(pred_plot,
       filename = "figures/prediction_vs_actual.png",
       device = "png")
```

```r
ggsave(pred3_plot,
       filename = "figures/prediction_vs_actual_smaller_model.png",
       device = "png")
```

Appendix 5: Residual Plots for Checking Assumptions



Appendix 6: Regression Model Parameters with all biomarkers and covariates as explanatory variables

Characteristic	Beta	95% CI	p-value
(Intercept)	6.4	-15, 28	0.6
il_8	1.3	0.06, 2.5	0.039
vegf_a	1.3	-0.21, 2.8	0.090
opg	-1.6	-3.3, -0.01	0.049
tgf_beta_1	-1.5	-3.0, -0.05	0.043
il_6	1.1	0.51, 1.8	<0.001
cxcl9	-0.31	-1.0, 0.39	0.4
cxcl1	0.03	-1.0, 1.1	>0.9
il_18	-0.35	-1.4, 0.71	0.5
csf_1	0.32	-2.5, 3.1	0.8
age	-0.01	-0.08, 0.05	0.7
sex	-0.39	-1.6, 0.85	0.5
smoker	-0.18	-1.5, 1.1	0.8
vas_0	0.25	0.01, 0.49	0.045

Abbreviation: CI = Confidence Interval

Appendix 7: Regression Model Parameters with IL_6 and VAS_0 as Explanatory Variables

Characteristic	Beta	95% CI	p-value
(Intercept)	-2.2	-4.4, -0.03	0.047
il_6	1.1	0.58, 1.6	<0.001
vas_0	0.33	0.13, 0.54	0.002

Abbreviation: CI = Confidence Interval

Appendix 8: Predicted vs. Actual Values for Model with IL_6 and VAS_0 as Explanatory Variables, RMSE = 3.26