```
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
## Warning: package 'tidyr' was built under R version 4.2.3
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.1
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(ggplot2)
library(kableExtra)
```

```
## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 2 > 1' in coercion to 'logical(1)'
```

```
##
## Attaching package: 'kableExtra'
##
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(readxl)
```

```
# Loading the Excel files
biomarkers <- read_excel("biomarkers.xlsx")
biomarkers
```

```
## # A tibble: 347 x 10
##    Biomarker    `IL-8` `VEGF-A`   OPG `TGF-beta-1` `IL-6` CXCL9 CXCL1 `IL-18`
##    <chr>         <dbl>    <dbl> <dbl>        <dbl>  <dbl> <dbl> <dbl>   <dbl>
##  1 126-0weeks     7.63     11.5  10.2         8.83   3.52  6.16  9.45    7.91
##  2 126-6weeks     7.12     11.6  10.4         8.87   3.89  6.12  9.06    7.92
##  3 127-0weeks     6.93     10.9  10.3         6.59   2.73  6.14  7.31    7.95
##  4 127-6weeks     7.16     11.6  10.4         8.61   2.6   6.35  8.61    7.94
##  5 127-12months   6.87     11.1  10.2         7.44   3.92  6.15  8.79    7.94
##  6 128-0weeks     8.62     12.5  10.6         8.51   3.71  7.34  9.9     8.72
##  7 128-6weeks     6.94     11.5  10.5         7.46   3.84  7.14  8.57    8.62
##  8 128-12months   6.47     11.0  10.1         6.45   4.65  8     8.18    8.71
##  9 129-0weeks     8.16     11.2  10.6         8.76   3.85  5.81  9.18    7.49
## 10 129-6weeks     6.57     10.7  10.2         6.82   2.98  6.11  6.69    7.23
## # i 337 more rows
## # i 1 more variable: ‘CSF-1‘ <dbl>
```

```
covariates <- read_excel("covariates.xlsx")
covariates
```

```
## # A tibble: 118 x 6
##    PatientID   Age ‘Sex (1=male, 2=female)‘ ‘Smoker (1=yes, 2=no)‘
##        <dbl> <dbl>                    <dbl>                  <dbl>
##  1         1    56                        1                      2
##  2         3    32                        1                      2
##  3         4    43                        2                      2
##  4         5    25                        2                      2
##  5         6    39                        1                      2
##  6         7    38                        2                      2
##  7         8    49                        1                      1
##  8         9    43                        2                      1
##  9        13    54                        2                      1
## 10        14    41                        1                      2
## # i 108 more rows
## # i 2 more variables: ‘VAS-at-inclusion‘ <dbl>, ‘Vas-12months‘ <dbl>
```

Data Cleaning and Consistency Checks

```
#Assignment instructions explain that Biomarker column (biomarkers) indicates patient number and time p
#So 'Patient ID' set as Primary key for join.
#Split Biomarker column in biomarkers
split_data <- strsplit(biomarkers$Biomarker, "-")
biomarkers$PatientID <- sapply(split_data, `[`, 1)
biomarkers$Biomarker <- sapply(split_data, `[`, 2)

biomarkers$Biomarker <- sub("weeks", "", biomarkers$Biomarker)
biomarkers$Biomarker <- sub("months", "", biomarkers$Biomarker)
```

```
# change datatpe for 'Patient ID' in covariates to correct datatype for match
covariates$PatientID <- as.character(covariates$PatientID)
```

```
#Left join on biomarkers and covariates datasets and new df labelled **combined_data_set
combined_data_set <- left_join(biomarkers, covariates, by = "PatientID")
```

```r
#Summary statistics for new combined_df - Followed by handling of 'null' values
summary_result <- summary(combined_data_set)
summary_result
```

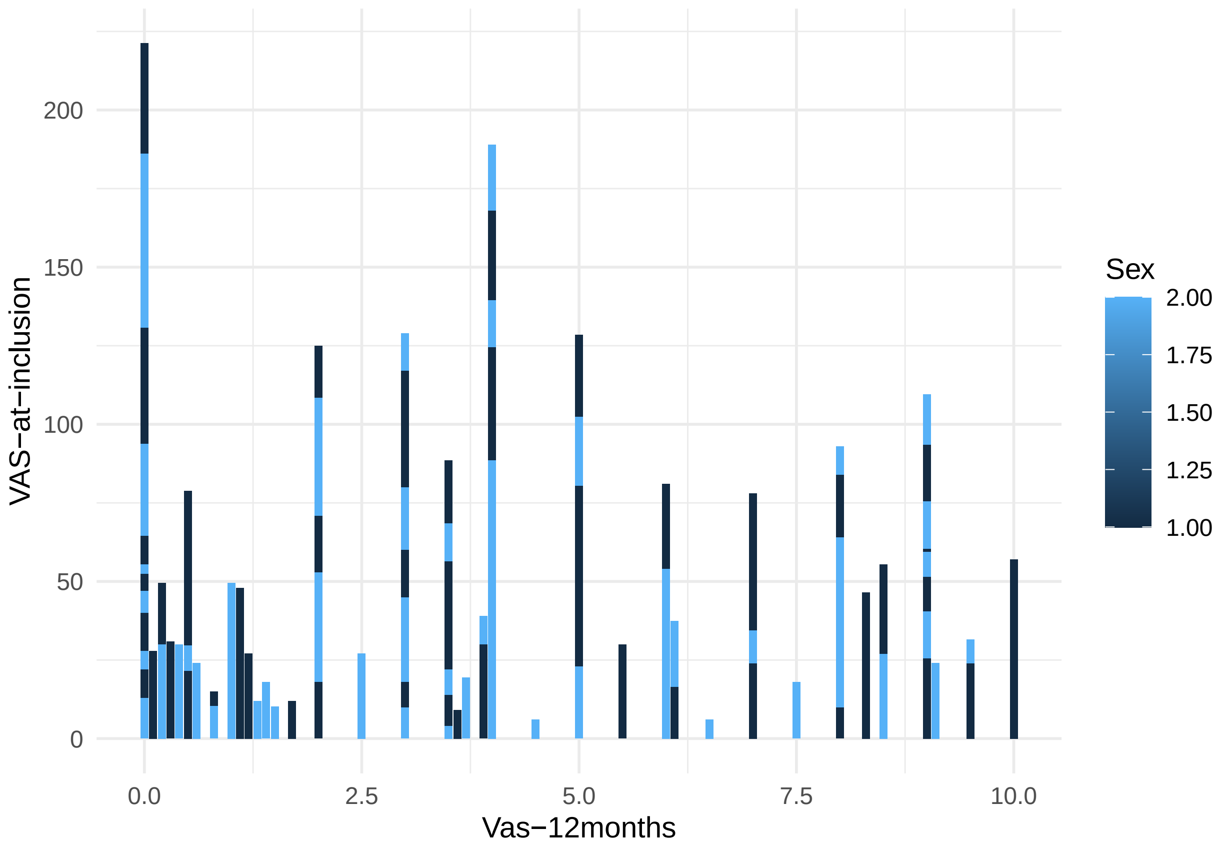
```
##   Biomarker              IL-8            VEGF-A           OPG
##  Length:347         Min.   : 5.500   Min.   :10.21   Min.   : 9.67
##  Class :character   1st Qu.: 6.690   1st Qu.:11.14   1st Qu.:10.41
##  Mode  :character   Median : 7.360   Median :11.51   Median :10.62
##                     Mean   : 7.424   Mean   :11.66   Mean   :10.67
##                     3rd Qu.: 8.130   3rd Qu.:12.15   3rd Qu.:10.84
##                     Max.   :11.330   Max.   :13.60   Max.   :11.96
##
##    TGF-beta-1         IL-6           CXCL9           CXCL1
##  Min.   :5.640   Min.   :1.600   Min.   : 4.61   Min.   : 5.680
##  1st Qu.:7.065   1st Qu.:2.580   1st Qu.: 5.99   1st Qu.: 7.165
##  Median :7.900   Median :3.070   Median : 6.35   Median : 8.370
##  Mean   :7.947   Mean   :3.249   Mean   : 6.47   Mean   : 8.292
##  3rd Qu.:8.815   3rd Qu.:3.665   3rd Qu.: 6.85   3rd Qu.: 9.340
##  Max.   :9.910   Max.   :7.950   Max.   :11.51   Max.   :11.190
##
##      IL-18           CSF-1        PatientID              Age
##  Min.   :6.700   Min.   :7.950   Length:347         Min.   :18.00
##  1st Qu.:7.930   1st Qu.:8.350   Class :character   1st Qu.:32.00
##  Median :8.250   Median :8.530   Mode  :character   Median :41.00
##  Mean   :8.293   Mean   :8.537                       Mean   :40.77
##  3rd Qu.:8.675   3rd Qu.:8.700                       3rd Qu.:49.00
##  Max.   :9.780   Max.   :9.780                       Max.   :59.00
##
##  Sex (1=male, 2=female) Smoker (1=yes, 2=no) VAS-at-inclusion  Vas-12months
##  Min.   :1.000          Min.   :1.000        Min.   : 0.0     Min.   : 0.000
##  1st Qu.:1.000          1st Qu.:1.000        1st Qu.: 4.0     1st Qu.: 0.800
##  Median :2.000          Median :2.000        Median : 6.5     Median : 3.500
##  Mean   :1.504          Mean   :1.674        Mean   : 6.0     Mean   : 3.655
##  3rd Qu.:2.000          3rd Qu.:2.000        3rd Qu.: 8.0     3rd Qu.: 6.000
##  Max.   :2.000          Max.   :2.000        Max.   :10.0     Max.   :10.000
##                                                               NA’s   :6
```

```r
combined_data_set <- combined_data_set %>%
  replace_na(list("Vas-12months" = 0.0))
view(combined_data_set)
```

```r
# renaming relevant columns for easier readibility.
combined_data_set <- combined_data_set %>%
  rename(Sex = `Sex (1=male, 2=female)`)
```

Exploratory Data Analysis(EDA)

```r
# Exploration of VAS Columns, and Sex with Bar graph
combined_data_set %>%
ggplot(aes(x = `Vas-12months`, y = `VAS-at-inclusion`, fill = Sex)) +
geom_col()+
theme_minimal()
```



```
# Central Tendancy of 'VAS_scores_dates(Inclusion and 12months), & 'Sex' Variables
summary_table <- combined_data_set %>%
  group_by(Sex) %>%
  summarise(
    Mean_Vas_12months = mean(`Vas-12months`),
    Mean_VAS_at_inclusion = mean(`VAS-at-inclusion`)
  )
summary_table
```

```
## # A tibble: 2 x 3
##     Sex Mean_Vas_12months Mean_VAS_at_inclusion
##   <dbl>             <dbl>                 <dbl>
## 1     1              3.64                  6.36
## 2     2              3.55                  5.64
```

Introduction - *Do the levels of inclusion vary between Males and Females?*

The decision to explore whether the levels of inclusion vary between Males and Females, is a result of initial data exploration. EDA indicates a fair distribution of data collected, between Males and Females (See Appendix), for the respective VAS at inclusion dates.

Summary statistics of the same variables, suggests that while the mean value went down at (VAS 12 months) of the study, it was a fairly small decline (See Appendix)

This prompted further analysis on this research topic with a focus on the random variables that are biomarker levels for Male and Female. (See Appendix)

Methodology ### Data cleaning steps and consistency checks (See Appendix)) ### Summary Statistics (See Appendix) ### Hypothesis Testing ### Regression Modelling

Research Hypothesis Null Hypothesis:

There is no significant difference in biomarker levels at inclusion between Male and Female groups.

Alternative Hypothesis: There is a significant difference in biomarker levels at inclusion between Male and Female groups.

This report will employ a t-test on the cleaned data sets that has been merged into one (combined_data_set) that includes biomakers and Sex. The t-test is an ideal choice for assessing differences between groups, and observing mean and variance.

```r
#Select all biomarkers for analysis
biomarkers_to_compare <- c("IL-8", "VEGF-A", "OPG", "TGF-beta-1", "IL-6", "CXCL9", "CXCL1", "IL-18", "C
```

```r
# Loop through each variable (biomarker) and perform multiple t-tests
for (biomarker in biomarkers_to_compare) {
  t_test_outcome <- t.test(combined_data_set[[biomarker]] ~ Sex, data = combined_data_set)

  cat("T-test for", biomarker, ":\n")
  print(t_test_outcome)
  cat("\n")
}
```

```
## T-test for IL-8 :
##
##  Welch Two Sample t-test
##
## data:  combined_data_set[[biomarker]] by Sex
## t = 0.71308, df = 344.69, p-value = 0.4763
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  -0.1213355  0.2593527
## sample estimates:
## mean in group 1 mean in group 2
##        7.458837        7.389829
##
##
## T-test for VEGF-A :
##
##  Welch Two Sample t-test
##
## data:  combined_data_set[[biomarker]] by Sex
## t = -1.6427, df = 328.85, p-value = 0.1014
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  -0.25965580  0.02334351
## sample estimates:
## mean in group 1 mean in group 2
##        11.59756        11.71571
##
##
## T-test for OPG :
```

```
##
##  Welch Two Sample t-test
##
## data:  combined_data_set[[biomarker]] by Sex
## t = -2.8495, df = 343.09, p-value = 0.004643
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  -0.20430298 -0.03743921
## sample estimates:
## mean in group 1 mean in group 2
##        10.60919        10.73006
##
##
## T-test for TGF-beta-1 :
##
##  Welch Two Sample t-test
##
## data:  combined_data_set[[biomarker]] by Sex
## t = -1.0036, df = 342.37, p-value = 0.3163
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  -0.30199489  0.09793974
## sample estimates:
## mean in group 1 mean in group 2
##        7.895058        7.997086
##
##
## T-test for IL-6 :
##
##  Welch Two Sample t-test
##
## data:  combined_data_set[[biomarker]] by Sex
## t = -1.6045, df = 325.33, p-value = 0.1096
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  -0.36213126  0.03677644
## sample estimates:
## mean in group 1 mean in group 2
##        3.167151        3.329829
##
##
## T-test for CXCL9 :
##
##  Welch Two Sample t-test
##
## data:  combined_data_set[[biomarker]] by Sex
## t = 2.5415, df = 335.86, p-value = 0.01149
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  0.04825458 0.37873944
## sample estimates:
## mean in group 1 mean in group 2
##        6.577326        6.363829
##
```

```
##
## T-test for CXCL1 :
##
##  Welch Two Sample t-test
##
## data:  combined_data_set[[biomarker]] by Sex
## t = -1.722, df = 338.55, p-value = 0.08599
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  -0.49238946  0.03270308
## sample estimates:
## mean in group 1 mean in group 2
##        8.175814        8.405657
##
##
## T-test for IL-18 :
##
##  Welch Two Sample t-test
##
## data:  combined_data_set[[biomarker]] by Sex
## t = 3.5687, df = 342.67, p-value = 0.0004099
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  0.09743075 0.33671410
## sample estimates:
## mean in group 1 mean in group 2
##        8.402558        8.185486
##
##
## T-test for CSF-1 :
##
##  Welch Two Sample t-test
##
## data:  combined_data_set[[biomarker]] by Sex
## t = -3.383, df = 330.87, p-value = 0.000803
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  -0.14191374 -0.03755537
## sample estimates:
## mean in group 1 mean in group 2
##        8.492151        8.581886
```

Findings for t-test of multiple tests

From the t-tests above, the data looks at the mean levels of biomarkers between Male and Female, alongside its Probability (p-value) for multiple tests.

It's evident to see a significant difference in the mean levels of bio markers for Male and Female, by observing the various score of p-values that are summarized below.

```
#Summary results:
findings <- sapply(biomarkers_to_compare, function(b) t.test(combined_data_set[[b]] ~ Sex, data = combi
print(findings)
```

```
##          IL-8       VEGF-A          OPG   TGF-beta-1          IL-6        CXCL9
## 0.4762772388 0.1014070294 0.0046428322 0.3162980851 0.1095645183 0.0114876457
##        CXCL1        IL-18        CSF-1
## 0.0859857030 0.0004099459 0.0008030102
```

With significance level alpha set to 0.05, the data indicates that biomarkers IL-8, VEGA-A,TGF-beta-1, IL-6, CXCL1 are above the set threshold, and therefore not statistically significant.

On the contrary, the p-values for biomarkers OPG, CXCL9, IL-18, CSF-1 point to a significant result, as these values are P<0.05.

Overall, the multiple tests carried out in the analysis show both non-significant and significant results.

Potential problems of multiple testing is the increased likely hood of getting a ‘Type 1 error’ that affects statistical analysis. This can be calculated as:

```r
#Set P-Values
pvalues <- c(0.4763, 0.1014, 0.004643, 0.3163, 0.1096, 0.01149, 0.08599, 0.0004099, 0.000803)
m <- length(pvalues)
alpha <- 0.05
p_fwer <- 1 - (1 - alpha)^m
cat("m:", m, "\n", "alpha:", alpha, "\n", "p_fwer:", p_fwer, "\n")
```

```
## m: 9
##  alpha: 0.05
##  p_fwer: 0.3697506
```

The Bonferroni correction mitigates this problem by adjusting the significance level over a number of tests to deal with probability of getting a ‘type 1’ error. Andrade, explains that the Bonferroni correction is lowered from the conventional P<0.05, after dividing the p-value from tests performed to dealth with the multiple tests problem (Andrade, 2019).

```r
#Use the p_fwer func to apply the Bonferroni correction
p_fwer <- 1 - (1 - alpha/length(pvalues))^length(pvalues)
cat("m:", length(pvalues), ", alpha:", alpha, ", p_fwer:", p_fwer, "\n")
```

```
## m: 9 , alpha: 0.05 , p_fwer: 0.04890317
```

Regression Modelling This section will use a linear regression model on the combined_data_set with 80:20 random split for training and testing.

```r
set.seed(32)
train_index <- sample(1:nrow(combined_data_set), 0.8 * nrow(combined_data_set))
#split datasets - Train_Test
train_data <- combined_data_set[train_index, ]
test_data <- combined_data_set[-train_index, ]
#Extract response and predictor variables
response_variable <- combined_data_set$`Vas-12months`
biomarkers_data <- combined_data_set[, c("IL-8", "VEGF-A", "OPG", "TGF-beta-1", "IL-6", "CXCL9", "CXCL1
covariates_data <- combined_data_set[, c("Age", "Sex")]
#Fit linear model
linear_model <- lm(response_variable ~ ., data = biomarkers_data)
summary(linear_model)
```

```
##
## Call:
## lm(formula = response_variable ~ ., data = biomarkers_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1351 -2.4079 -0.3098  1.6741  7.3797
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.37829    6.65475   0.508   0.6120
## ‘IL-8‘        0.71258    0.32793   2.173   0.0305 *
## ‘VEGF-A‘      0.59187    0.43109   1.373   0.1707
## OPG          -2.15643    0.45792  -4.709 3.64e-06 ***
## ‘TGF-beta-1‘ -0.32677    0.36913  -0.885   0.3767
## ‘IL-6‘        0.92722    0.18430   5.031 7.95e-07 ***
## CXCL9        -0.19326    0.22092  -0.875   0.3823
## CXCL1        -0.31498    0.23592  -1.335   0.1827
## ‘IL-18‘      -0.03502    0.29951  -0.117   0.9070
## ‘CSF-1‘       1.73002    0.82513   2.097   0.0368 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.866 on 337 degrees of freedom
## Multiple R-squared:  0.1637, Adjusted R-squared:  0.1414
## F-statistic: 7.329 on 9 and 337 DF,  p-value: 9.174e-10
```

The model is then applied to unseen test data (the remaining 20%) to make predictions.

```
#predictions on test data (remaining 0.2)
predicted_vas_12months <- predict(linear_model, newdata = test_data)
#new data frame for comparisons
prediction_comparison <- data.frame(Actual_VAS_12months = test_data$`Vas-12months`, Predicted_VAS_12mont
#output
print(prediction_comparison)
```

```
##    Actual_VAS_12months Predicted_VAS_12months
## 1                  0.0              3.7725547
## 2                  4.0              5.2477740
## 3                  5.0              6.1286773
## 4                  5.0              3.1906022
## 5                  4.0              2.9311603
## 6                  4.0              2.6412614
## 7                  4.0              2.1272368
## 8                  4.0              4.0042218
## 9                  4.5              5.3460973
## 10                 0.0              5.1350530
## 11                 6.0              2.3278393
## 12                 3.5              2.7468612
## 13                 1.5             -0.6929571
## 14                 8.0              3.8510314
## 15                 0.0              2.4662929
## 16                 3.0              5.9142237
## 17                 1.3              5.2907074
## 18                 9.0              4.2186204
## 19                 0.5              3.7269420
## 20                 2.5              2.3805037
## 21                 8.5              4.8208328
## 22                 4.0              3.0832672
## 23                 1.4              1.6691920
## 24                 8.0              4.6813964
## 25                 2.0              2.3323176
## 26                 1.5              1.0048073
## 27                 8.0              5.8550899
## 28                 8.0              3.8398550
## 29                 3.0              4.4236794
## 30                 3.0              2.8636769
## 31                 9.5              4.1841050
## 32                 0.0              3.6440679
## 33                 0.5              3.9704497
## 34                 2.0              2.2645869
## 35                 0.0              3.1798488
## 36                 9.1              3.5487578
## 37                 1.1              3.3734036
## 38                 0.1              3.8600342
## 39                 0.1              4.3733632
## 40                 0.0              3.0603311
## 41                 0.0              2.2565745
## 42                 0.6              2.7441080
## 43                 9.0              4.4822417
## 44                 2.0              4.4300469
## 45                 9.0              3.9277117
## 46                 3.5              3.2574323
## 47                 0.5              3.1976666
## 48                 2.0              4.8598914
## 49                 3.0              5.4708049
## 50                 5.0              4.3630828
## 51                 5.0              4.0040216
## 52                 6.0              3.7546889
## 53                 6.0              3.8226129
## 54                 8.3              3.3464665
## 55                 8.3              3.0553244
## 56                10.0              4.7038428
## 57                 2.0              4.3061014
## 58                 8.3              3.6028038
## 59                 3.0              5.6687441
## 60                 3.0              3.7517092
## 61                 7.0              5.3151001
## 62                 6.0              5.5304251
## 63                 8.5              4.6743059
## 64                 1.0              3.2266567
## 65                 1.0              3.9258429
## 66                 6.1              2.1170520
## 67                 0.0              3.6170110
## 68                 0.0              3.2187782
## 69                 0.5              3.2171987
## 70                 0.2              3.7031194
```

Findings for Regression Modelling

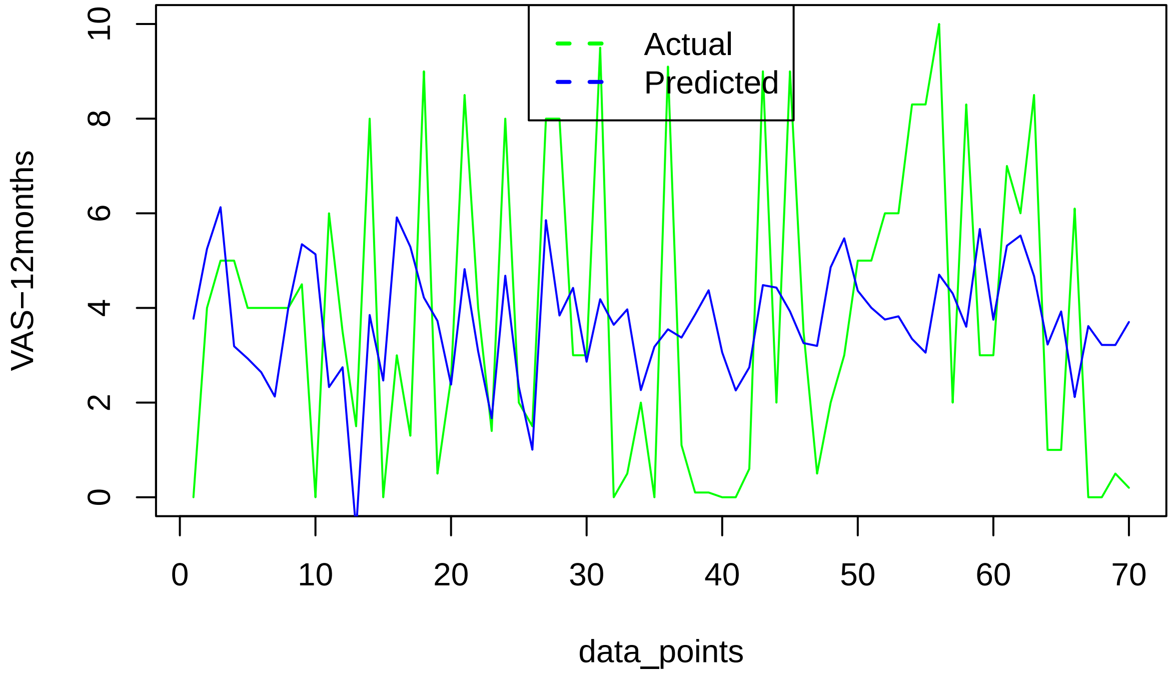
The model is designed to fit the data by using a response and predictor variables, VAS-12months and bio marker levels respectively.

The Residual standard error (RSE) offers some insights into the model. From the summary, the value 2.866 for RSE is considered fairly small, indicating this model fits the data set well.

The significance codes highlighted in the regression model is vital in focusing on the bio markers IL-8, OPG, IL-6, CSF-1 and its relationship to response variable, and point to a strong link to VAS-12months.

The model is useful in predicting the 12-month VAS of patients, as seen below. However, it's also evident where the predicted values have differed from the Actual values.

```r
#line plot for visualization
plot(prediction_comparison$Actual_VAS_12months, type = "l", col = "green", ylab = "VAS-12months", xlab =
lines(prediction_comparison$Predicted_VAS_12months, col = "blue")
legend("top", legend = c("Actual", "Predicted"), col = c("green", "blue"), lty = 2, lwd = 2)
```



Further Analysis of both data sets (Data limitations)

Time constraints: The covariates data includes data over a 12-month VAS period. However, a greater sample over a larger time period would supported further analysis into this study.

Limited Scope: There is limited scope regarding 'Health Factors' i.e 'smoking' was only considered in the study.

Conclusion

The null hypothesis is rejected based on the statistical methods conducted in this analysis. This paper has displayed through hypothesis testing, significant differences in the mean levels of bio markers between

Male and Female. This study has also found useful insights through regression modelling to explore the associations of biomarkers at 12 months.