

# Data 621 Blog #1

Jean Jimenez

2024-04-14

## Simple Linear Regression of Stroke Dataset

```
library(tidyverse)
```

```
## Warning: package 'dplyr' was built under R version 4.3.2
```

```
## -- 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    1.5.1
```

```
## v ggplot2    3.4.4      v tibble     3.2.1
```

```
## v lubridate  1.9.3      v tidyr      1.3.0
```

```
## v purrr      1.0.2
```

```
## -- 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 errors
```

### Introduction

Simple linear regression is a tool that can be used to predict outcomes based on data. Simple Linear Regression models can be trained on real life data to help predict events before they happen.

In the case of healthcare, it is important to know when a stroke is about to occur so that a team can intervene. Fortunately, there is a lot of data that exists on stroke patients. We can use some of this data and simple linear regression to help determine factors that lead to the prediction of stroke events and certain outcomes.

In this blog post, I aim to explore simple linear regression on a stroke dataset. Specifically, I will use it to determine the relationships between certain blood biomarkers and poor stroke outcomes.

I first begin by defining the variables in the dataset:

```
import pandas as pd
```

```
data = {
```

```
    "VARIABLE NAME": [
```

```
        "Auto Lymphocyte #", "Auto Lymphocyte %", "Auto Neutrophil #",
```

```
        "Auto Neutrophil # last", "Auto Neutrophil %", "Auto Neutrophil % last",
```

```

    "Platelet Count - Automated last", "ratio_of_Lymphocyte_Neutrophil_pct",
    "ratio_of_Platelet_Neutrophil_cnt", "Bilirubin Total, Serum",
    "C-Reactive Protein, Serum", "Creatinine, Serum",
    "D-Dimer Assay, Quantitative", "Albumin, Serum",
    "MRS_discharge_score_cleaned"
],
"DEFINITION": [
    "The absolute number of lymphocytes, a type of white blood cell, as measured automatically.",
    "The percentage of lymphocytes out of total white blood cells, as measured automatically.",
    "The absolute number of neutrophils, a type of white blood cell, as measured automatically.",
    "The absolute number of neutrophils from the last measurement, automatically determined.",
    "The percentage of neutrophils out of total white blood cells, as measured automatically.",
    "The percentage of neutrophils from the last measurement, automatically determined.",
    "The most recent automated count of platelets, which are cell fragments important for clotting.",
    "The ratio of the percentage of lymphocytes to neutrophils.",
    "The ratio of the absolute count of platelets to neutrophils.",
    "The total amount of bilirubin in the serum, indicating liver function.",
    "The level of C-reactive protein in the serum, a marker of inflammation.",
    "The level of creatinine in the serum, indicating kidney function.",
    "The quantitative result of a D-Dimer assay, used to help rule out the presence of an inappropriate blood clot.",
    "The level of albumin in the serum, a protein that can indicate nutritional status and liver function.",
    "A cleaned score based on the Modified Rankin Scale at discharge, assessing the degree of disability."
],
"THEORETICAL EFFECT": [
    "Could indicate an immune response or a decrease in immunity depending on the level.",
    "Might reflect changes in the immune system or be indicative of specific health conditions.",
    "High levels could suggest infection or inflammation, while low levels could indicate a compromised immune system.",
    "Past neutrophil levels could help in observing trends in a patient's immune response over time.",
    "A higher percentage could indicate an acute infection or chronic inflammation.",
    "Past percentages could provide context to current immune function and response to treatment.",
    "Low levels can indicate thrombocytopenia and a risk for bleeding; high levels could suggest clotting disorders.",
    "A higher ratio may suggest a viral infection, while a lower ratio could indicate bacterial infection.",
    "This ratio can help in diagnosing and monitoring the severity of infections or inflammatory conditions.",
    "Increased levels may indicate liver damage or disease; lower levels might be seen in certain liver conditions.",
    "Elevated levels suggest inflammation or infection; it is a broad marker and not specific to any one condition.",
    "Elevated levels can indicate renal dysfunction or failure, while low levels can occur with renal disease.",
    "Elevated results may suggest the presence of thrombosis or an increased risk for clotting disorders.",
    "Low levels can indicate malnutrition, liver disease, or chronic illnesses.",
    "This score can help in predicting patient outcomes and the need for post-discharge care."
]
}

variables_table = pd.DataFrame(data)
variables_table

```

	VARIABLE NAME	THEORETICAL EFFECT
## 0	Auto Lymphocyte #	Could indicate an immune response or a decrease in immunity depending on the level.
## 1	Auto Lymphocyte %	Might reflect changes in the immune system or be indicative of specific health conditions.
## 2	Auto Neutrophil #	High levels could suggest infection or inflammation, while low levels could indicate a compromised immune system.
## 3	Auto Neutrophil # last	Past neutrophil levels could help in observing trends in a patient's immune response over time.
## 4	Auto Neutrophil %	A higher percentage could indicate an acute infection or chronic inflammation.
## 5	Auto Neutrophil % last	Past percentages could provide context to current immune function and response to treatment.

```
## 6      Platelet Count - Automated last ... Low levels can indicate thrombocytopenia and a...
## 7 ratio_of_Lymphocyte_Neutrophil_pct ... A higher ratio may suggest a viral infection, ...
## 8      ratio_of_Platelet_Neutrophil_cnt ... This ratio can help in diagnosing and monitori...
## 9      Bilirubin Total, Serum ... Increased levels may indicate liver damage or ...
## 10     C-Reactive Protein, Serum ... Elevated levels suggest inflammation or infect...
## 11     Creatinine, Serum ... Elevated levels can indicate renal dysfunction...
## 12     D-Dimer Assay, Quantitative ... Elevated results may suggest the presence of t...
## 13     Albumin, Serum ... Low levels can indicate malnutrition, liver di...
## 14     MRS_discharge_score_cleaned ... This score can help in predicting patient outc...
##
## [15 rows x 3 columns]
```

## Data Exploration

### Importing, Processing

I begin by importing and processing the data set to have a target variable.

I am using the MRS\_discharge\_score\_cleaned to create the target variable

A MRS score of 6 means dead. A MRS score of 5 means severely disable. 4 Means greatly disable

We will see if these blood bio markers have any impact on MRS score of 4 or 5 or 6.

```
library(tidyverse)

shout2023_biomarkers = read_csv("~/Masters/Data621/blogs/shout2023_biomarkers.csv")

## Rows: 29662 Columns: 10
## -- Column specification -----
## Delimiter: ","
## dbl (10): auto_lymphocyte_rat, auto_neutrophil_rat, auto_neutrophil_last_rat...
##
## i Use 'spec()' to retrieve the full column specification for this data.
## i Specify the column types or set 'show_col_types = FALSE' to quiet this message.

blog1_dat= shout2023_biomarkers %>%
  mutate(TARGET= if_else(MRS_discharge_score_cleaned %in% c(4,5,6),1,0, missing=NULL))

blog1_dat = blog1_dat %>%
  select(-MRS_discharge_score_cleaned)
```

### Exploration

Now let us explore the distribution of predictor variables and look at the summary statistics.

```
library(summarytools)

## Warning: package 'summarytools' was built under R version 4.3.3

##
## Attaching package: 'summarytools'
```

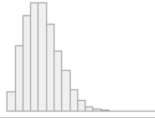
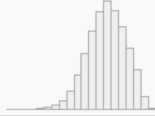
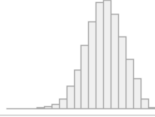


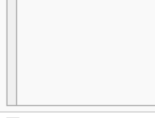

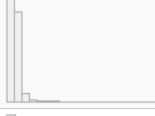

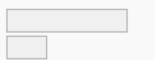
```
## The following object is masked from 'package:tibble':  
##  
##      view  
  
blog1_stats = dfSummary(blog1_dat, stats = c("mean", "sd", "med", "IQR", "min", "max", "valid", "n.miss"  
#view(blog1_stats)
```

## Data Frame Summary

blog1\_dat

Dimensions: 29662 x 10

Duplicates: 119

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	auto_lymphocyte_rat [numeric]	Mean (sd) : 22.2 (11) min ≤ med ≤ max: 0 ≤ 21.2 ≤ 94.3 IQR (CV) : 15.1 (0.5)	643 distinct values		27817 (93.8%)	1845 (6.2%)
2	auto_neutrophil_rat [numeric]	Mean (sd) : 66.9 (12.8) min ≤ med ≤ max: 0 ≤ 67.4 ≤ 99.1 IQR (CV) : 17.6 (0.2)	741 distinct values		27818 (93.8%)	1844 (6.2%)
3	auto_neutrophil_last_rat [numeric]	Mean (sd) : 65.3 (12.7) min ≤ med ≤ max: 0 ≤ 65.7 ≤ 99 IQR (CV) : 17.3 (0.2)	749 distinct values		27818 (93.8%)	1844 (6.2%)
4	platelet_count [numeric]	Mean (sd) : 242.2 (91.5) min ≤ med ≤ max: 6 ≤ 228 ≤ 2764 IQR (CV) : 98 (0.4)	705 distinct values		29535 (99.6%)	127 (0.4%)
5	ratio_of_Lymphocyte_Neutrophil_pct [numeric]	Mean (sd) : 0.4 (0.7) min ≤ med ≤ max: 0 ≤ 0.3 ≤ 63 IQR (CV) : 0.3 (1.9)	282 distinct values		27816 (93.8%)	1846 (6.2%)
6	ratio_of_Platelet_Neutrophil_cnt [numeric]	Mean (sd) : 48.4 (100.3) min ≤ med ≤ max: 0.1 ≤ 42.5 ≤ 15500 IQR (CV) : 28.6 (2.1)	8275 distinct values		26614 (89.7%)	3048 (10.3%)
7	bilirubin_total [numeric]	Mean (sd) : 0.6 (0.4) min ≤ med ≤ max: 0.1 ≤ 0.5 ≤ 24.4 IQR (CV) : 0.4 (0.8)	60 distinct values		26500 (89.3%)	3162 (10.7%)
8	creatine_serum [numeric]	Mean (sd) : 1.2 (1.1) min ≤ med ≤ max: 0.2 ≤ 1 ≤ 19.6 IQR (CV) : 0.5 (0.8)	762 distinct values		27753 (93.6%)	1909 (6.4%)
9	albumin [numeric]	Mean (sd) : 3.9 (22.6) min ≤ med ≤ max: 0.8 ≤ 3.9 ≤ 3701 IQR (CV) : 0.7 (5.7)	54 distinct values		26768 (90.2%)	2894 (9.8%)
10	TARGET [numeric]	Min : 0 Mean : 0.2 Max : 1	0 : 22267 (75.1%) 1 : 7395 (24.9%)		29662 (100.0%)	0 (0.0%)

Generated by [summarytools](#) 1.0.1 (R version 4.3.1)

2024-04-14

The `dfSummary` function of `summarytools` package displays summary statistics of all variables in the dataframe, as well as showing the distribution and missing data.

The distribution of most data seem normally distributed.

Since each predictor is its own biomarker, I am leaving them all in (ignoring multicollinearity).

There are about a bit under 30k records total. In terms of missing data, some rows are missing as much as

3000 records (10%).

## Data Preparation

Since the columns that have missing data seem to have a large range and not distributed following a pattern, I will not use mean or median imputation to fill in for missing data. Instead, I will just exclude all missing values.

```
clean_blog1_dat = na.omit(blog1_dat)

blog1_stats_clean = dfSummary(clean_blog1_dat, stats = c("mean", "sd", "med", "IQR", "min", "max", "val.

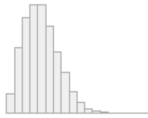
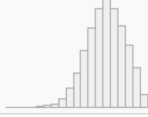
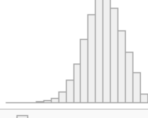







#view(blog1_stats_clean)
```

## Data Frame Summary

clean\_blog1\_dat

Dimensions: 24182 x 10

Duplicates: 0

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	auto_lymphocyte_rat [numeric]	Mean (sd) : 22.1 (11) min ≤ med ≤ max: 0 ≤ 21.2 ≤ 94.3 IQR (CV) : 15.1 (0.5)	629 distinct values		24182 (100.0%)	0 (0.0%)
2	auto_neutrophil_rat [numeric]	Mean (sd) : 67 (12.8) min ≤ med ≤ max: 1 ≤ 67.4 ≤ 99.1 IQR (CV) : 17.6 (0.2)	730 distinct values		24182 (100.0%)	0 (0.0%)
3	auto_neutrophil_last_rat [numeric]	Mean (sd) : 65.3 (12.7) min ≤ med ≤ max: 0 ≤ 65.7 ≤ 98.3 IQR (CV) : 17.3 (0.2)	738 distinct values		24182 (100.0%)	0 (0.0%)
4	platelet_count [numeric]	Mean (sd) : 242 (91.4) min ≤ med ≤ max: 6 ≤ 227 ≤ 2764 IQR (CV) : 97 (0.4)	677 distinct values		24182 (100.0%)	0 (0.0%)
5	ratio_of_Lymphocyte_Neutrophil_pct [numeric]	Mean (sd) : 0.4 (0.7) min ≤ med ≤ max: 0 ≤ 0.3 ≤ 63 IQR (CV) : 0.3 (1.9)	262 distinct values		24182 (100.0%)	0 (0.0%)
6	ratio_of_Platelet_Neutrophil_cnt [numeric]	Mean (sd) : 47.7 (32.5) min ≤ med ≤ max: 0.1 ≤ 42.4 ≤ 1825 IQR (CV) : 28.6 (0.7)	8008 distinct values		24182 (100.0%)	0 (0.0%)
7	bilirubin_total [numeric]	Mean (sd) : 0.6 (0.4) min ≤ med ≤ max: 0.1 ≤ 0.5 ≤ 24.4 IQR (CV) : 0.4 (0.8)	58 distinct values		24182 (100.0%)	0 (0.0%)
8	creatine_serum [numeric]	Mean (sd) : 1.3 (1.1) min ≤ med ≤ max: 0.2 ≤ 1 ≤ 19.6 IQR (CV) : 0.5 (0.8)	733 distinct values		24182 (100.0%)	0 (0.0%)
9	albumin [numeric]	Mean (sd) : 3.8 (0.6) min ≤ med ≤ max: 0.8 ≤ 3.9 ≤ 5.9 IQR (CV) : 0.7 (0.1)	52 distinct values		24182 (100.0%)	0 (0.0%)
10	TARGET [numeric]	Min : 0 Mean : 0.2 Max : 1	0 : 18142 (75.0%) 1 : 6040 (25.0%)		24182 (100.0%)	0 (0.0%)

Generated by [summarytools](#) 1.0.1 (R version 4.3.1)

2024-04-14

Now there is no missing data.

We will proceed with building the linear model.

## Building Linear Regression Model

I will use the `lm()` function to generate the linear model from the `stats` package.

I will use all predictor variables.

Use the `summary()` function to display the results of the linear regression model.

```
library(stats)

model1= lm(TARGET ~ ., data = clean_blog1_dat)

summary(model1)

##
## Call:
## lm(formula = TARGET ~ ., data = clean_blog1_dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.22203 -0.26789 -0.19532  0.04498  1.46910
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.457e-01  6.172e-02   7.222 5.26e-13 ***
## auto_lymphocyte_rat -2.163e-03  7.435e-04  -2.910  0.00362 **
## auto_neutrophil_rat -4.817e-04  6.655e-04  -0.724  0.46920
## auto_neutrophil_last_rat  2.548e-03  3.248e-04   7.844 4.55e-15 ***
## platelet_count  3.908e-04  3.156e-05  12.385 < 2e-16 ***
## ratio_of_Lymphocyte_Neutrophil_pct  1.923e-02  4.173e-03   4.608 4.09e-06 ***
## ratio_of_Platelet_Neutrophil_cnt -4.304e-04  1.074e-04  -4.006 6.20e-05 ***
## bilirubin_total  4.053e-02  6.462e-03   6.273 3.61e-10 ***
## creatine_serum  2.535e-03  2.592e-03   0.978  0.32801
## albumin       -1.020e-01  5.040e-03 -20.248 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4226 on 24172 degrees of freedom
## Multiple R-squared:  0.04738,    Adjusted R-squared:  0.04703
## F-statistic: 133.6 on 9 and 24172 DF,  p-value: < 2.2e-16
```

## Linear Regression Model Interpretation

Coefficients Interpretation:

The model shows that platelet count, total bilirubin, and the ratio of lymphocyte to neutrophil percentage all have positive coefficients. This suggests that increases in these predictors are associated with a higher likelihood of severe disability or death.

Conversely, albumin demonstrates a significant negative relationship with “TARGET”. Specifically, its coefficient of -0.102 indicates that higher levels of albumin are associated with lower probabilities of severe disability or death, which might reflect better overall health status.

Statistical Significance:



Variables such as `auto_lymphocyte_ratio`, `auto_neutrophil_last_ratio`, `platelet_count`, ratio of lymphocyte to neutrophil percentage, ratio of platelet to neutrophil count, total bilirubin, and albumin all show statistically significant p-values. This strong evidence against the null hypothesis suggests these factors significantly influence the severity of disability or risk of death.

In contrast, predictors like `auto_neutrophil_ratio` and creatinine serum have non-significant p-values, indicating their effects on “TARGET” may be negligible in this model.

Model Fit and Efficacy:

The Residual Standard Error of 0.4226 on 24,172 degrees of freedom reflects the average deviation of data points from the fitted line, measured in the scale of “TARGET”.

The Multiple R-squared value of 0.04738, though relatively low, indicates that about 4.738% of the variability in “TARGET” can be accounted for by the predictors included in this model. This suggests that while the model captures a portion of the factors influencing outcomes, much of the variability remains unexplained, highlighting the complex nature of disability and mortality.

The F-statistic of 133.6 and its associated very small p-value ( $< 2.2e-16$ ) confirm the overall statistical significance of the model, indicating that it successfully identifies at least some key factors affecting severe outcomes in patients.

## Creating an Improved Linear Regression Model.

To make this model better, I've included interaction terms such as `auto_lymphocyte_rat`, `platelet_count`, `auto_neutrophil_last_rat`, `ratio_of_Lymphocyte_Neutrophil_pct`, and `platelet_count`, `bilirubin_total`. These interactions might reveal combined effects that are not apparent when considering the variables independently.

Interaction terms are chosen based on logical associations that might exist between the variables. For example, how the lymphocyte ratio interacts with platelet count could be relevant in the context of immune responses or inflammation that both could relate to the severity of the medical condition leading to disability or death.

```
enhanced_model = lm(TARGET ~ auto_lymphocyte_rat + auto_neutrophil_last_rat +
                    platelet_count + ratio_of_Lymphocyte_Neutrophil_pct +
                    ratio_of_Platelet_Neutrophil_cnt + bilirubin_total +
                    albumin +
                    auto_lymphocyte_rat:platelet_count +
                    auto_neutrophil_last_rat:ratio_of_Lymphocyte_Neutrophil_pct +
                    platelet_count:bilirubin_total,
                    data = clean_blog1_dat)

summary(enhanced_model)

##
## Call:
## lm(formula = TARGET ~ auto_lymphocyte_rat + auto_neutrophil_last_rat +
##     platelet_count + ratio_of_Lymphocyte_Neutrophil_pct + ratio_of_Platelet_Neutrophil_cnt +
##     bilirubin_total + albumin + auto_lymphocyte_rat:platelet_count +
##     auto_neutrophil_last_rat:ratio_of_Lymphocyte_Neutrophil_pct +
##     platelet_count:bilirubin_total, data = clean_blog1_dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.40082 -0.26799 -0.19535  0.08848  1.65038
```

```

##
## Coefficients:
##
## (Intercept) 5.591e-01
## auto_lymphocyte_rat -7.533e-03
## auto_neutrophil_last_rat 9.794e-04
## platelet_count 3.400e-04
## ratio_of_Lymphocyte_Neutrophil_pct 1.448e-02
## ratio_of_Platelet_Neutrophil_cnt -4.668e-04
## bilirubin_total 1.864e-02
## albumin -1.018e-01
## auto_lymphocyte_rat:platelet_count 2.301e-07
## auto_neutrophil_last_rat:ratio_of_Lymphocyte_Neutrophil_pct 4.200e-03
## platelet_count:bilirubin_total 8.342e-05
##
## Std. Error t value
## (Intercept) 4.288e-02 13.041
## auto_lymphocyte_rat 1.168e-03 -6.452
## auto_neutrophil_last_rat 3.822e-04 2.562
## platelet_count 7.067e-05 4.812
## ratio_of_Lymphocyte_Neutrophil_pct 4.270e-03 3.392
## ratio_of_Platelet_Neutrophil_cnt 1.070e-04 -4.361
## bilirubin_total 1.559e-02 1.196
## albumin 4.971e-03 -20.474
## auto_lymphocyte_rat:platelet_count 2.741e-06 0.084
## auto_neutrophil_last_rat:ratio_of_Lymphocyte_Neutrophil_pct 6.124e-04 6.859
## platelet_count:bilirubin_total 5.883e-05 1.418
##
## Pr(>|t|)
## (Intercept) < 2e-16 ***
## auto_lymphocyte_rat 1.13e-10 ***
## auto_neutrophil_last_rat 0.010406 *
## platelet_count 1.50e-06 ***
## ratio_of_Lymphocyte_Neutrophil_pct 0.000694 ***
## ratio_of_Platelet_Neutrophil_cnt 1.30e-05 ***
## bilirubin_total 0.231782
## albumin < 2e-16 ***
## auto_lymphocyte_rat:platelet_count 0.933106
## auto_neutrophil_last_rat:ratio_of_Lymphocyte_Neutrophil_pct 7.12e-12 ***
## platelet_count:bilirubin_total 0.156231
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4222 on 24171 degrees of freedom
## Multiple R-squared: 0.04925, Adjusted R-squared: 0.04885
## F-statistic: 125.2 on 10 and 24171 DF, p-value: < 2.2e-16

```

## Enhanced Model Interpretation

### Model Coefficients and Interpretation:

Main effects: Most of the predictors retained their significance and direction of influence from the previous model. For instance, albumin continues to show a strong negative relationship with “TARGET,” suggesting higher albumin levels are associated with a decreased risk of severe outcomes. Platelet count and ratio of Lymphocyte Neutrophil pct remain positive, indicating an increase in these variables correlates with an increased risk.

Interaction terms: The interaction between `auto_neutrophil_last_rat` and `ratio_of_Lymphocyte_Neutrophil_pct` is particularly notable with a coefficient of 0.0042 and is highly significant ( $p < 2.2e-12$ ), suggesting that the combined effect of these predictors on “TARGET” is greater than their individual effects. However, other interaction terms, such as `auto_lymphocyte_rat:platelet_count`, did not demonstrate a significant effect, suggesting that the simple interaction of these two factors does not notably influence the outcome within the model’s context.

Statistical Significance:

Predictors like `auto_lymphocyte_rat`, `ratio_of_Platelet_Neutrophil_cnt`, and `auto_neutrophil_last_rat:ratio_of_L` are highly significant, reflecting their substantial impact on “TARGET”. Conversely, the interaction term `platelet_count:bilirubin_total` and `bilirubin_total` itself did not achieve statistical significance, indicating that their contributions might be more complex or require different modeling approaches to be adequately captured.

Model Fit and Evaluation:

The Residual Standard Error (RSE) slightly decreased to 0.4222 from 0.4226, indicating a minor improvement in the model’s accuracy in predicting the data points.

The Multiple R-squared increased marginally to 0.04925 from 0.04738, and the Adjusted R-squared followed suit to 0.04885 from 0.04703, suggesting a slight enhancement in the model’s explanatory power due to the addition of interaction terms.

The F-statistic remains highly significant ( $p\text{-value} < 2.2e-16$ ), confirming that the model is statistically meaningful and that the inclusion of new terms is justified.

## Conclusion

This analysis of stroke data using simple linear regression provides valuable insights into the factors contributing to severe outcomes such as severe disability or death. By examining a range of blood biomarkers, we can identify key predictors that correlate with these serious health events. Notably, variables like platelet count and total bilirubin were positively associated with higher risk scores, suggesting that these factors can serve as critical indicators in predicting severe outcomes. Conversely, albumin exhibited a negative relationship with the target variable, highlighting its potential protective role against severe disability or mortality.

The initial regression model indicated that while significant relationships exist among some biomarkers and the severity of stroke outcomes, the overall variability explained by the model was limited. This underscores the complex nature of stroke-related disabilities and fatalities, which likely involve multifaceted interactions between various biological systems.

To refine the predictive accuracy, an enhanced model incorporated interaction terms, addressing the potential combined effects of biomarkers. The inclusion of interactions such as the one between `auto_neutrophil_last_rat` and `ratio_of_Lymphocyte_Neutrophil_pct` proved to be highly significant, suggesting that the interplay between these markers significantly impacts stroke outcomes. This advanced model approach demonstrated a slight improvement in explaining the variability in patient outcomes, as evidenced by a modest increase in the R-squared value.

Despite these improvements, the persistently low R-squared value after model enhancement highlights the possibility of unaccounted factors or the need for alternative modeling techniques to fully capture the dynamics influencing severe stroke outcomes.

In conclusion, the complexity of stroke demands ongoing research efforts to refine predictive models and enhance the accuracy and reliability of these predictions. This is crucial for timely and effective interventions that can potentially reduce the incidence of severe outcomes in stroke patients.