Data Analysis Demystified: A Beginner's Step-by-Step Guide

Ketana Kakani[1], Praveen Kumar P T V [2], Santanu Koley[3], Radhika T S L [4]*

Department of Mathematics, BITS Pilani, Hyderabad Campus,

Hyderabad, India- 500078.

***Corresponding author: radhikatsl@hyderabad.bits-pilani.ac.in**

Abstract

This paper is a comprehensive resource for individuals seeking to analyze medical data using statistical methods. The dataset utilized in this analysis is from the repository generated in our prior research projects. Beginning with an elucidating flowchart, the paper outlines a systematic procedure to facilitate a thorough comprehension of statistical techniques and the underlying data. This enables readers to navigate the analytical process clearly and confidently, ensuring a complete understanding of each step involved.

Keywords

Statistical analysis, medical dataset, Linear regression, Regularized regression models.

1. Introduction

Statistical analysis is fundamental for evidence-based decision-making and problem-solving across various disciplines and applications, including business, healthcare, social sciences, engineering, and environmental science. It empowers professionals and researchers to harness data effectively, extract actionable insights, and optimize outcomes in diverse contexts. A thorough examination of existing literature demonstrates the concerted efforts of researchers to guide the conducting of various analyses. For instance, Simpson [1] offered valuable assistance to an emerging researcher by crafting a tailored data analysis plan for a quantitative study. It focused on condensing study data and identifying relevant statistical tests. It organized variables by characteristics, used descriptive statistics for summarization, defined them as dependent or independent, and employed inferential statistics to select tests based on the association among variables. Such methodological clarity underscores the importance of employing scientific rigor in data analysis across all sectors, stressed Tao, Luo, and Yan [2]. They emphasized that data analysis underpins informed decision-making, evidence-based policies, innovation, and societal advancement, cementing its indispensable role in navigating the complexities of the modern landscape. Furthermore, the review paper authored by Yan, Robert, and Li [3] delved into critical statistical design considerations within biomedical studies, aiming to enhance scientists' proficiency in statistical methodologies. It addressed key concepts such as sample size determination, data summarization techniques, test methodologies, and common pitfalls, aiming to

foster robust statistical reasoning and minimize study design and analysis errors. Such endeavors underscore the ongoing commitment within the scientific community to promote methodological rigor and enhance the quality of research across various domains. RW Cooksey [4] simplified descriptive statistics, aiming to summarize data efficiently. The work employed graphical representations or numerical indices to depict characteristics while exploring probability and normal distribution, prioritizing general trends over individual data interpretation. Schneider Hommel and Blettner [5] introduced uni- and multivariable regression models followed by illustrative examples on pre-analysis considerations and result interpretation. Pitfalls in linear regression analysis were thoroughly examined for performance and interpretation. Ali and Younas [6] discussed the purpose of regression analysis as description, estimation, prediction, and control. Different types of regression analysis were used, namely linear, logistic, and multiple regression, and factors affecting sample size, missing data, and the nature of the sample were discussed. Bzovsky, Phillips, and Guymer [7] discussed using linear and logistic regression to study the relationship between predictor and response variables for continuous and dichotomous outcomes in a clinical study on patients receiving anti-vascular endothelial growth factor therapy. It stated that these models helped to understand risk factors associated with the disease. Alexopoulos [8] focused on linear regression involving one or more independent variables that predict the quantitative dependent variable. Later, the models were tested for accuracy using ANOVA, and the violations of assumptions in the model were checked. Bryan and Stanton [9] focused on using multiple linear regression instead of conducting a series of simple regressions and tested for the underlying assumptions, correlations among predictors, influential observations, and exploration of model structures. Hao and Hailong [10] focused on establishing a mathematical model and predicting the results based on the existing factor data. It explained the factor analysis method in prediction, where the correlation is calculated for relevant variables. Then, they studied the causation to see if there was any other casualty and calculated the regression equations. The paper by Huei and Liang [11] discussed the average velocity of blood in each arterial segment for healthy and diseased conditions using statistical analysis, namely Karl Pearson's correlation, linear regression, and Wilcoxon signed-rank test. Noora [12] focused on detecting multicollinearities, such as correlation coefficients, Variance Inflation Factor, and eigenvalue methods, and advanced regression techniques, such as principle component regression, weighted regression, and ridge regression, were adopted. Jamal [13] discussed the impact of multicollinearity and its existence on predictor variables in hypothesis testing. The paper recommended ignoring and dismissing the models with high correlation as it is difficult to interpret the results. The study by Vatcheva et al. [14] demonstrated the effect of different degrees of multicollinearity among predictors using pairwise Pearson product-moment correlation coefficients on two outcome variables, namely systolic and diastolic blood pressure, in linear regression analysis for generated simulated datasets. In the work by Breusch and Pagan they developed a model for heteroscedasticity disturbances for linear regression model using Lagrangian multiplier test[15]. The paper by Osborne et al. [16] discusses the assumptions of multiple regression that

are robust to the violation, such as linearity, measurement reliability, heteroscedasticity, and normality. Hickey et al. [17] outlined some multivariate linear regression model-checking diagnostic techniques, such as heteroscedasticity and autocorrelation. Nayak and Tantravahi [18] explored key feature variables employing regularized linear regression, ensemble methods, and boosting algorithms tailored for correlated features in medical datasets. Subsequently, these methods underwent testing utilizing the quantile loss function to ascertain the optimal regression model for prediction, followed by the computation of quantile intervals. Ogutu, Schulz, and Piepho [19] focused on identifying approaches, namely ridge, lasso, and elastic net regression, to efficiently and accurately predict breeding values in genomic selection and then evaluate them for prediction. The model introduced by Tibshirani [20], the 'lasso,' minimizes the residual sum of squares with coefficients constrained by a constant. Their results showed that it yields exact 0 coefficients, creating interpretable models with favorable properties of subset selection and ridge regression. The widely used Lasso regularization is showcased on a real-life dataset of adults with autism in R software by Finch and Finch [21]. Findings on executive functioning and intelligence test correlations were discussed. The work by Hui and Trevor [22] explained the elastic net for the sparse model while accommodating the correlation effect. It explained the importance of penalties for increasing the accuracy of the model.

Continuing efforts to elucidate the process of conducting statistical analysis address a critical necessity within the research community. Such endeavors are pivotal in advancing scientific inquiry by offering guidance, enhancing methodological clarity, promoting reproducibility, and facilitating knowledge dissemination and skill development. This paper is a comprehensive guide for individuals interested in analyzing medical data using statistical methods. Leveraging a dataset from prior research, we presented a structured approach via a detailed flowchart accompanied by a systematic procedure that elucidates statistical techniques and data comprehension. This framework empowers readers to navigate the analytical process with confidence and thoroughness. The paper is organized as follows: Section 2 delves into the data utilized in the study, followed by a detailed exposition of the methodology in Section 3, followed by results and discussions in Section 4, and concluding remarks in Section 5.

2. Data Description

The dataset was generated from our earlier studies, where we utilized a Computational Fluid Dynamics (CFD) prototype based on sonography images of the human femoral artery. This model was simulated in the COMSOL multi-physics software under various anatomical and physiological blood flow conditions within the artery to generate a diverse dataset representing different blood flow scenarios. We incorporated the time-independent continuity and momentum equations for incompressible fluid flows, employing the laminar flow interface. For the numerical solution of the velocity and pressure fields, we used P1-P1 linear finite element discretization, dividing the computational domain into small finite elements and

approximating velocity and pressure within each element using linear functions. Given the importance of tolerance in determining the efficiency of a numerical method, we set the relative tolerance to 10^{-3} to define the acceptable error relative to the magnitude of the solution. In the next step, we identified appropriate fluid (blood) models for femoral arterial components, namely the CFA (Common Femoral Artery), SFA (Superficial Femoral Artery), and the DFA (Deep Femoral Artery). [18].

The dataset comprised the following predictor and response variables.

Predictors

Den_CFA: Blood Density in the common femoral artery (CFA) (Kg/m^3)

Vis_CFA: Blood viscosity in the CFA (Pa.s)

Den_SFA: Blood Density in the superficial femoral artery (SFA) (Kg/m^3)

Vis_SFA: Blood viscosity in the SFA (Pa.s)

YS1: Blood Yield stress in the Deep femoral artery (DFA) (N/m^2)

Mp1: Blood model parameter in DFA (s)

INVEL: Blood velocity at the CFA entrance (m/s)

Response Variables

Vel_SFA: Average blood velocity in SFA (m/s)

Vel_DFA: Average blood velocity in DFA (m/s)

Pre_CFA: Average blood pressure in CFA (Pa)

Pre_SFA: Average blood pressure in SFA (Pa)

Pre_DFA: Average blood pressure in DFA (Pa)

WSS_CFA: Wall shear stress in CFA (Pa)

WSS_SFA: Wall shear stress in SFA (Pa)

WSS_DFA: Wall shear stress in DFA (Pa)

This study analyzes the dataset to derive insights pursued through the following objectives.

Objectives

- Utilize linear models for evaluating significant model parameters.
- Investigate the predictive accuracy of linear models.
- Assess the coherence between predictions and clinical data.

3. Methodology

The data analysis procedure follows a sequential approach outlined in Figure (1). Initially, the dataset was imported into the Python workspace, and univariate analysis was conducted to examine the distribution of the response variable, utilizing descriptive statistics such as mean, standard deviation, skewness, and kurtosis. Subsequently, scatter plots and Karl Pearson's correlation coefficient were employed to identify significant predictor variables for each response variable. These findings informed the appropriateness of a multiple linear regression model, with subsequent confirmation of multicollinearity among predictor variables. It is to be noted that the model-building process is complete if significant (predictor) variables are uncorrelated; otherwise, regularized linear models are to be developed to address collinearity issues.

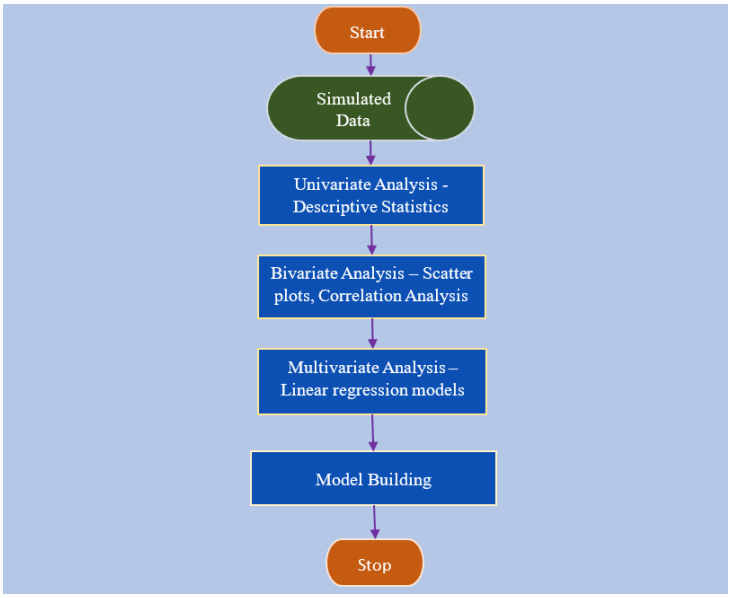

Figure 1. Work Flow Diagram

Figure (2) displays the comprehensive flowchart detailing the workflow implementation. The process begins with importing the dataset into the Python workspace and generating descriptive statistics for the response variables to assess their distribution. Subsequently, simple linear regression models are constructed to identify and eliminate insignificant predictor variables for each response variable. Upon discovering multiple significant predictor variables affecting the response variable, the analysis progresses to building multiple regression models. While individual variables were found to affect the response variable in the previous step, evaluating a group of predictor variables' significance involves computing R^2, t, and F statistics. This step helps eliminate insignificant predictors, selecting the best model with the most appropriate predictor variables.

Following this, the validity of the multilinear models is to be assessed in terms of the factors, residual randomness, autocorrelation, and heteroscedasticity. The results indicated the necessity of constructing regularized linear regression models, which were addressed in the subsequent stage. Finally, the best model is identified

based on performance metrics, and the statistical model for each response variable is built.

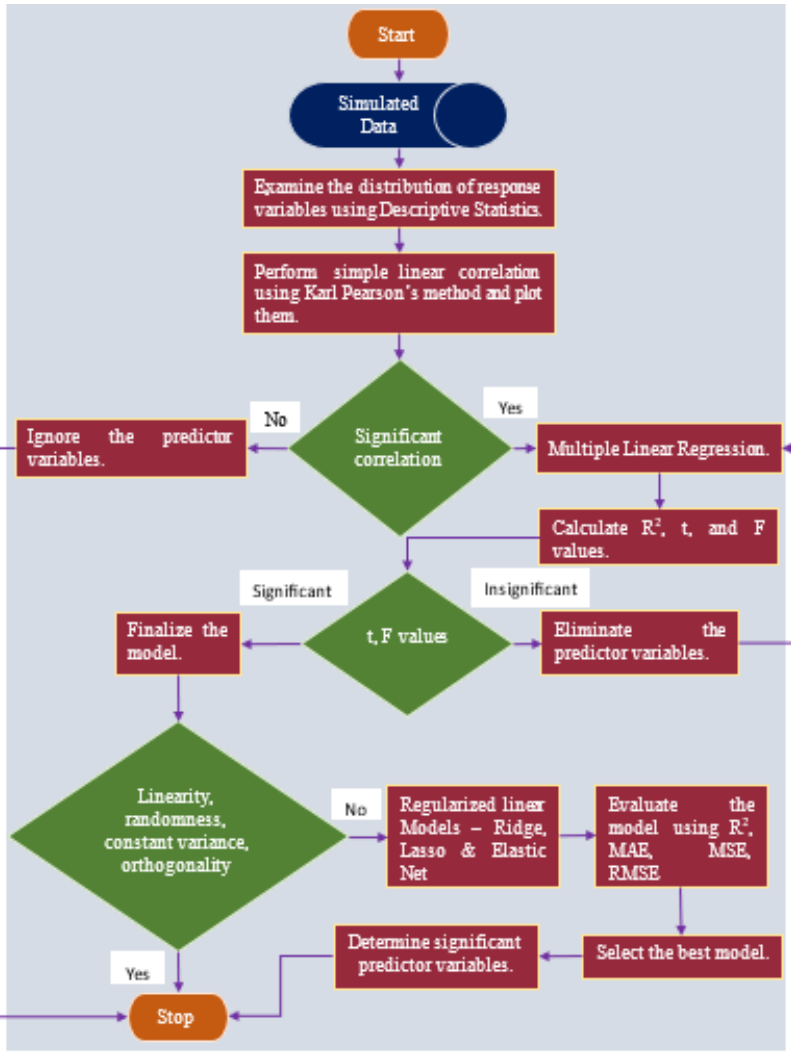

Figure 2. Flow chart of statistical analysis

4. Results and Discussions

In this section, the outcomes of the statistical examinations, as delineated in the flowchart (Figure 2), are presented alongside the resulting discussions.

Univariate Analysis

Step 1: Data summarization and presentation

Descriptive statistics serve to describe and summarize data, employing graphical representations for clarity. Mean and standard deviation values are utilized for data summarization. Table (1) presents the descriptive statistics computed for the response variables.

Table 1. Representation of descriptive statistics for response variables

Variable	Count	Mean	Std	Min	25%	50%	75%	Max	CV	Skew	Kurtosis
Vel_SFA	2187	0.13	0.01	0.12	0.12	0.13	0.15	0.15	9%	0.0231	-1.37
Vel_DFA	2187	0.11	0.01	0.09	0.10	0.11	0.12	0.12	11%	0.0089	-1.48
Pre_CFA	2187	8287.53	1.17	8285.47	8286.52	8287.54	8288.51	8289.66	0%	0.0026	-1.17
Pre_SFA	2187	8484.89	1.89	8482.09	8482.88	8484.90	8486.90	8487.76	0%	0.0006	-1.42
Pre_DFA	2187	8488.15	0.18	8487.85	8487.98	8488.16	8188.29	8488.48	0%	0.0122	-1.23
WSS_CFA	2187	0.41	0.06	0.32	0.36	0.41	0.46	0.51	14%	0.1987	-0.75
WSS_SFA	2187	0.75	0.10	0.60	0.68	0.75	0.81	0.923	13%	0.1232	-1.10
WSS_DFA	2187	0.89	0.10	0.76	0.77	0.89	0.01	0.03	11%	0.0214	-1.49

Observations

- Table (1) shows that the average blood velocity is higher in SFA than in the DFA segment.
- The pressure is higher in DFA compared to CFA and SFA.
- The wall shear stress is greater in DFA than in SFA and CFA.
- The velocity variation is more significant in the DFA than in the SFA; similarly, the variation in wall shear stress is higher in CFA, followed by SFA and DFA. There is almost zero variation for pressures in all three components.
- From the skewness and kurtosis, it is evident that more observations are concentrated near the average value than on the tails and the extreme values are present towards the right side of the distribution.
- The data also indicates the absence of outliers.

Bivariate Analysis

Step 2: Analysis of a simple linear relationship between variables

Given the quantitative nature of the observations, we conducted simple correlations using Karl Pearson's method, subsequently evaluating their significance, as represented in Table (2).

Table 2. Representation of correlation and its significance.

Response Variable	Explanatory variable	Correlation value	Extent and direction of correlation	p-value	Significance (Y/N)
Vel_SFA	Den_CFA	0.0039	Low degree positive correlation	0.8525	N
	Vis_CFA	-0.0328	Low degree negative correlation	0.1252	N
	Den_SFA	-0.0021	Low degree negative correlation	0.9235	N
	Vis_SFA	-0.2095	Low degree negative correlation	0.0	Y
	YS1	0.0171	Low degree positive correlation	0.4233	N
	Mp1	-0.0	Low degree negative correlation	0.9999	N

	Invel	0.9769	High degree positive correlation	0.0	Y
Vel_DFA	Den_CFA	0.0013	Low degree positive correlation	0.9498	N
	Vis_CFA	-0.0206	Low degree negative correlation	0.3359	N
	Den_SFA	0.0017	Low degree positive correlation	0.9367	N
	Vis_SFA	0.0819	Low degree positive correlation	0.0001	Y
	YS1	-0.0197	Low degree negative correlation	0.3561	N
	Mp1	-0.0	Low degree negative correlation	0.9999	N
	Invel	0.9962	High degree positive correlation	0.0	Y
Pre_CFA	Den_CFA	0.9354	High degree positive correlation	0.0	Y
	Vis_CFA	0.0705	Low degree positive correlation	0.0009	Y
	Den_SFA	0.0024	Low degree positive correlation	0.9099	N
	Vis_SFA	0.111	Low degree positive correlation	0.0	Y
	YS1	0.018	Low degree positive correlation	0.399	N
	Mp1	0.0	Low degree positive correlation	0.9999	N
	Invel	0.3273	Low degree positive correlation	0.0	Y
Pre_SFA	Den_CFA	0.0029	Low degree positive correlation	0.8937	N
	Vis_CFA	0.0161	Low degree positive correlation	0.4529	N
	Den_SFA	0.9851	High degree positive correlation	0.0	Y
	Vis_SFA	0.0555	Low degree positive correlation	0.0094	Y
	YS1	0.0057	Low degree positive correlation	0.7892	N
	Mp1	-0.0	Low degree negative correlation	0.9999	N
	Invel	0.1619	Low degree positive correlation	0.0	Y
Pre_DFA	Den_CFA	-0.0067	Low degree negative correlation	0.7536	N
	Vis_CFA	0.0989	Low degree positive correlation	0.0	Y
	Den_SFA	0.0056	Low degree positive correlation	0.7925	N
	Vis_SFA	0.2707	Low degree positive correlation	0.0	Y
	YS1	0.1045	Low degree positive correlation	0.0	Y

	Mp1	0.0	Low degree positive correlation	0.9999	N
	Invel	0.9502	High degree positive correlation	0.0	Y
WSS_CFA	Den_CFA	0.0037	Low degree positive correlation	0.8631	N
	Vis_CFA	0.6461	Moderate degree positive correlation	0.0	Y
	Den_SFA	0.0005	Low degree positive correlation	0.9821	N
	Vis_SFA	0.0243	Low degree positive correlation	0.2556	N
	YS1	-0.0033	Low degree negative correlation	0.878	N
	Mp1	0.0	Low degree positive correlation	0.9999	N
	Invel	0.7599	High degree positive correlation	0.0	Y
WSS_SFA	Den_CFA	0.0086	Low degree positive correlation	0.6859	N
	Vis_CFA	-0.0051	Low degree negative correlation	0.8122	N
	Den_SFA	0.0018	Low degree positive correlation	0.9312	N
	Vis_SFA	0.3875	Low degree positive correlation	0.0	Y
	YS1	0.0099	Low degree positive correlation	0.6409	N
	Mp1	-0.0	Low degree negative correlation	0.9999	N
	Invel	0.9204	High degree positive correlation	0.0	Y
WSS_DFA	Den_CFA	0.0057	Low degree positive correlation	0.7911	N
	Vis_CFA	-0.021	Low degree negative correlation	0.3259	N
	Den_SFA	0.0002	Low degree positive correlation	0.9914	N
	Vis_SFA	0.0188	Low degree positive correlation	0.3804	N
	YS1	0.0454	Low degree positive correlation	0.0339	Y
	Mp1	0.0	Low degree positive correlation	0.9999	N
	Invel	0.9985	High degree positive correlation	0.0	Y

Observations

- Table (2) shows that all the response variables depend on inlet velocity. However, the pressures in CFA and SFA are not highly correlated with the inlet velocity.
- Almost all the response variables depend on the blood viscosity in SFA except for the wall shear stress in CFA and DFA.

- Pressure and wall shear stress in DFA are dependent on the yield stress.
- Pressures are dependent on the blood density except for the pressure in DFA.

The significant predictor variables obtained in step (2) are used to construct multiple linear regression discussed in step (3).

Multivariate Analysis

Step 3: Multiple Linear Regression Analysis

The statistical evaluation of medical data aims to reveal the relationships between multiple variables, typically accomplished through regression analysis. It is known that Regression analysis is a widely used statistical technique that investigates and models the interrelation between variables. These variables are categorized into two types based on their role in the study: independent (or predictor), used to estimate the nature of other variables, and dependent (or response), derived from known information.

Regression analysis is typically employed when the response variable is continuous. Its primary objective is to ascertain and estimate the parameters of a model that best fits a given dataset. This is achieved through the principle of least squares, where the sum of squares of errors for a sample is minimized. Predictor variables are determined likewise, adhering to this principle. For regression analysis to yield reliable results, certain assumptions must be met by the dataset:

- Dependent and independent variables exhibit a linear relationship.
- There should be minimal correlation among independent variables.
- The observations for the explained variable are drawn from a normal and independent population.
- Residuals, representing the difference between observed and predicted values, adhere to a normal distribution with a mean of zero and constant variance.

We conducted multiple linear regression for each response variable using the Ordinary Least Squares (OLS) method, incorporating significant regressor variables identified in Table (2). Subsequently, the accuracy of the model was evaluated using the F-test and coefficient of determination. The significant variables for each response variable are presented in Table (3), accompanied by the corresponding model accuracy metrics.

Table 3. Multiple linear regression model's coefficients with constants and the evaluation metrics

Response Variable	Using Multiple Linear Regression			R^2 Value	Adjusted R^2 value	F – test statistic p-value
	Explanatory variable	Coefficient Value	P - value			
Vel_SFA	Constant	0.0187	0.0	0.998	0.998	0.000
	Vis_SFA	-5.385	0.0			
	Invel	1.2554	0.0			
Vel_DFA	Constant	-0.0341	0.0	0.999	0.999	0.000
	Vis_SFA	2.4341	0.0			
	Invel	1.4812	0.0			
Pre_CFA	Constant	7998.79	0.0	0.999	0.999	0.000

	Den_CFA	0.2677	0.0			
	Vis_CFA	201.7766	0.0			
	Vis_SFA	317.8147	0.0			
	Invel	46.8337	0.0			
Pre_SFA	Constant	7999.50	0.0	1.000	1.000	0.000
	Den_SFA	0.4559	0.0			
	Vis_SFA	256.9304	0.0			
	Invel	37.4557	0.0			
Pre_DFA	Constant	8485.46	0.0	0.997	0.997	0.000
	Vis_CFA	42.8188	0.0			
	Vis_SFA	117.2284	0.0			
	YS1	45.2629	0.0			
	Invel	20.5721	0.0			
WSS_CFA	Constant	-0.4175	0.0	0.995	0.995	0.000
	Vis_CFA	89.0494	0.0			
	Invel	5.2357	0.0			
WSS_SFA	Constant	-0.6424	0.0	0.997	0.997	0.000
	Vis_SFA	95.0873	0.0			
	Invel	11.2935	0.0			
WSS_DFA	Constant	-0.2866	0.0	0.999	0.999	0.000
	YS1	11.4085	0.0			
	Invel	12.5530	0.0			

Observations

- The significant regressor variables obtained using simple correlation are found to be significant parameters for multiple linear regression.
- To describe the influence of predictor factors on response variables, R^2 and modified R^2 values are employed. It was found that the significant predictor variables could provide more than 99% of the predictions for all the response variables.
- Using the F-test statistic, the model's significance for estimations is determined, and the model is significant for the predictor variables listed in Table (3).

Step 4: Analysing residuals

Next, we computed the residuals and analyzed them to assess the model's accuracy by predicting the response values. A plot depicting the relationship between the response variable and its prediction is employed to verify linearity. Subsequently, the Durbin-Watson test and Karl Pearson's correlation coefficient are utilized to evaluate autocorrelation among the predictor variables. It is known that the Durbin-Watson test aids in determining the nature of autocorrelation, while Karl Pearson's method identifies correlated variables. Anderson-Darling test is employed to assess the normality of residuals. The reason for using the Anderson-Darling test is that it is a modified version of the Kolmogorov-Smirnov test, placing more emphasis on the tails where the null and alternative hypotheses are given by

H_0: Residuals adhere to a normal distribution.

H_1: Residuals deviate from a normal distribution.

Furthermore, the Breusch Pagan test, is employed to examine constant variance, or homoscedasticity, in residuals for which the null and alternative hypotheses are

H_0: Residuals exhibit constant variance.

H_1: Residuals display non-constant variance.

The results of these tests are presented in Table (4).

Table 4. Performance measure of multiple linear regression model.

Response Variable	Linearity Test	Normality Test		Test for autocorrelation		Test for Homoskedasticity	
		P – value	Inference	Statistic value	Type of autocorrelation	P value	inference
Vel_SFA	Linear	0.000	Not normal	2.4742	Little or no autocorrelation	0.000	Heteroscedasticity
Vel_DFA	Linear	0.000	Not normal	0.2672	Positive autocorrelation	0.000	Heteroscedasticity
Pre_CFA	Linear	0.000	Not normal	1.2606	Positive autocorrelation	0.309	Homoscedasticity
Pre_SFA	Linear	0.000	Not normal	0.4206	Positive autocorrelation	0.000	Heteroscedasticity
Pre_DFA	Linear	0.000	Not normal	2.7944	Negative autocorrelation	0.004	Heteroscedasticity
WSS_CFA	Linear	0.000	Not normal	2.6226	Negative autocorrelation	0.435	Homoscedasticity
WSS_SFA	Linear	0.000	Not normal	2.5666	Negative autocorrelation	0.005	Heteroscedasticity
WSS_DFA	Linear	0.000	Not normal	0.4138	Positive autocorrelation	0.068	Homoscedasticity

Observations

- From Table (4), it is noted that all the response variables are linear.
- From the normality test, it is evident that all the response variables do not follow the normal distribution.
- Autocorrelation exists in the data where Vel_DFA, Pre_CFA, Pre_SFA, and WSS_DFA exhibit positive autocorrelation while other variables exhibit negative autocorrelation. The response variable Vel_SFA has little or no autocorrelation.
- Almost all the response variables exhibit heteroscedasticity except for Vel_DFA, WSS_CFA, and WSS_DFA, which exhibit homoscedasticity.

To address the deviation in assumptions of linear models, we proceeded to study using regularized linear models, as discussed in step (5).

Step 5: Regularized Linear Models

Considering the association among regressor variables, as evident from Table (5), we opted for regularized linear models. These models aim to mitigate bias in linear

models by applying a penalty to the loss function, thus preventing overfitting. The analysis is conducted while considering the optimal penalty. Furthermore, evaluation metrics are computed for all models, enabling the selection of the best model to streamline the regression model.

The significant regressor variables for estimations using the regularized linear models are identified in Table (5).

Table 5. Significant parameters of regularized linear models

Response Variable	Linear model without regularization	Regularized Linear Models		
		Ridge	Lasso	Elastic Net
Vel_SFA	Vis_SFA, Invel	Vis_SFA, Invel	None	Invel
Vel_DFA	Vis_SFA, Invel	Vis_SFA, Invel	None	Invel
Pre_CFA	Den_CFA, Vis_CFA, Vis_SFA, Invel	Den_CFA, Vis_CFA, Vis_SFA, Invel	Den_CFA, Invel	Den_CFA, Vis_SFA, Invel
Pre_SFA	Den_SFA, Vis_SFA, Invel	Den_SFA, Vis_SFA, Invel	Den_SFA, Invel	Den_SFA, Invel
Pre_DFA	Vis_CFA, Vis_SFA, YS1, Invel	Vis_CFA, Vis_SFA, YS1, Invel	Invel	Invel
WSS_CFA	Vis_CFA, Invel	Vis_CFA, Invel	None	Invel
WSS_SFA	Vis_SFA, Invel	Vis_SFA, Invel	None	Invel
WSS_DFA	YS1, Invel	YS1, Invel	None	Invel

As depicted in Table (5), the Lasso regression failed to identify significant regressor variables for the model, except in cases of pressures of all three components, where Invel, Den_CFA, and Den_SFA also proved significant. In contrast, the ridge regression model outperforms the elastic net model in estimating the model's parameters. The elastic net model oversimplifies the model by considering only the inlet velocity as the significant parameter.

Table (6) illustrates the performance of these models using evaluation metrics such as coefficient of determination (R^2), mean square error (MSE), mean absolute error (MAE), and root mean square error (RMSE).

Table 6. Evaluation metrics of regularized linear models

Response variable	Regression Model	Evaluation metrics			
		R^2	MAE	MSE	RMSE
Vel_SFA	Linear	0.9983	0.0004	0.0	0.0004
	Ridge	0.9746	0.014	0.0	0.0017
	Lasso	0.0	0.009	0.0001	0.0105
	Elastic Net	0.6138	0.0056	0.0	0.0065
Vel_DFA	Linear	0.9991	0.0003	0.0	0.0004
	Ridge	0.9955	0.0077	0.0	0.0008
	Lasso	0.0	0.0102	0.0001	0.0121
	Elastic Net	0.7380	0.0053	0.0	0.0062
Pre_CFA	Linear	0.9995	0.0221	0.0007	0.0268
	Ridge	0.9902	0.0962	0.0134	0.1158
	Lasso	0.9712	0.1623	0.0393	0.1983
	Elastic Net	0.9835	0.1246	0.0226	0.1501
Pre_SFA	Linear	0.9997	0.0295	0.0012	0.0351
	Ridge	0.9979	0.0722	0.0071	0.0845
	Lasso	0.9924	0.1351	0.0272	0.1650

	Elastic Net	0.9966	0.0952	0.0123	0.1108
Pre_DFA	Linear	0.9968	0.0082	0.0001	0.0099
	Ridge	0.9463	0.0341	0.0017	0.0409
	Lasso	0.4229	0.1154	0.018	0.1343
	Elastic Net	0.9016	0.0465	0.0031	0.0554
WSS_CFA	Linear	0.9949	0.0032	0.0	0.004
	Ridge	0.7706	0.0222	0.0007	0.0269
	Lasso	0.0	0.0449	0.0032	0.0563
	Elastic Net	0.5655	0.0315	0.0014	0.0371
WSS_SFA	Linear	0.9973	0.0044	0.0	0.0052
	Ridge	0.9166	0.0239	0.0008	0.0289
	Lasso	0.0	0.0859	0.01	0.1002
	Elastic Net	0.8434	0.0335	0.0016	0.0396
WSS_DFA	Linear	0.9991	0.0026	0.0	0.0031
	Ridge	0.9979	0.0039	0.0	0.0047
	Lasso	0.0	0.0853	0.0105	0.1026
	Elastic Net	0.9934	0.0068	0.0001	0.0083

Table (6) demonstrates that although Lasso regression performs satisfactorily in estimating pressures, it struggles to predict response variables for velocities and wall shear stress (WSS). Moreover, it is observed that the performance ridge regression outperforms the elastic net.

Further remarks:

The statistical analysis of velocity, pressure, and wall shear stress in the CFA, SFA, and DFA components underscores the significance of inlet velocity and the viscosity of CFA as key determinants. Notably, wall shear stress has more significant variability than velocities across the three arterial segments, while pressures remain relatively close to the average value.

The clinical data showcases the velocity, pressure, and wall shear stress profiles across the three segments, as detailed in Table (7).

Table 7. Parameters of the Clinical Data[11]

Artery	CFA	SFA	DFA
Mean velocity (cm/s)	14.1±5.4	8.9±3.9	10.7±5.0
Mean WSS (Pa)	0.35±0.18	0.49±0.15	
Mean pressure (mmHg)		70.9±6.7	

The velocity, pressure, and WSS profiles in three segments are computed and presented in Tables (8) and (9) from the developed statistical models.

Table 8. Predictions of the multiple linear regression model

Artery	CFA	SFA	DFA
Mean velocity (cm/s)	-	11.02± 0.01	10.89±0.012
Mean WSS (Pa)	0.41±0.06	0.75±0.10	0.89±0.10
Mean pressure (mmHg)	62.16±1.17	63.64±1.89	63.67±0.18

Table 9. Predictions of the ridge regression model.

Artery	CFA	SFA	DFA
Mean velocity (cm/s)	-	11.02±0.01	10.89±0.01
Mean WSS (Pa)	0.41±0.04	0.75±0.09	0.89±0.1

Mean pressure (mmHg)	62.16±1.16	63.64±1.89	63.67±0.17

Tables (7), (8), and (9) highlight a noticeable discrepancy between the expected values derived from linear models and the clinical data. While linear models proved significant in predicting parameters, they fall short as the best prediction models due to the data's non-normality, autocorrelation, and heteroskedasticity. Among the regularized linear models, it is observed that ridge regression outperforms Lasso and Elastic Net in terms of velocity and wall shear stress prediction. Also, it is seen that the Elastic Net regressor oversimplified the model by only considering the inlet velocity near CFA as the significant variable, thereby suppressing the contribution of other regressor variables.

5. Conclusions

This paper provides a comprehensive guide for analyzing medical data using statistical methods. It systematically explains each procedure, simplifying complex techniques for better understanding. By offering clear, step-by-step instructions, the paper enables readers to confidently navigate the analytical process, ensuring they can effectively apply statistical methods in their medical research and practice.

Funding Declaration: No funds, grants, or other support was received.

Clinical Trial Number: Not Applicable.

References

1. Simpson SH (2015) Creating a Data Analysis Plan: What to Consider When Choosing Statistics for a Study. *Can J Hosp Pharm* 68(4):3 11-7. DOI: 10.4212/cjhp.v68i4.1471.
2. Li R, Tao Q, Luo Y and Yan L (2020) Research on Practical Application of Data Analysis and Visualization. *International Conference on Virtual Reality and Intelligent Systems (ICVRIS)* pp.78-81. DOI: 10.1109/ICVRIS51417.2020.00026.
3. Yan F, Robert M and Li Y (2017) Statistical methods and common problems in medical or biomedical science research. *Int J Physiol Pathophysiol Pharmacol* 9(5):157-163.
4. Cooksey RW (2020) Descriptive Statistics for Summarising Data. *Illustrating Statistical Procedures: Finding Meaning in Quantitative Data* 15:61–139. DOI: 10.1007/978-981-15-2537-7_5.
5. Schneider A, Hommel G and Blettner M (2010) Linear regression analysis: part 14 of a series on evaluation of scientific publications. *Dtsch Arztebl Int* 107(44):776-82. DOI: 10.3238/arztebl.2010.0776.

6. Ali P and Younas A (2021) Understanding and interpreting regression analysis. *Evid Based Nurs* 24(4):116-118. DOI: 10.1136/ebnurs-2021-103425. Epub 2021.
7. Bzovsky S, Phillips MR, Guymer RH, et al. (2022) The clinician's guide to interpreting a regression analysis. *Eye* 36, pp. 1715–1717. Available at: https://doi.org/10.1038/s41433-022-01949-z
8. Alexopoulos EC (2010) Introduction to multivariate regression analysis. *Hippokratia*. 14(Suppl 1):23-8.
9. Bryan K Slinker and Stanton A Glantz (2008) Multiple Linear Regression, Accounting for Multiple Simultaneous Determinants of a Continuous Dependent Variable. *Circulation* 117(13):1732-7. DOI: 10.1161/CIRCULATIONAHA.106.654376.
10. Hao Kang and Hailong Zhao (2020) Description and Application Research of Multiple Regression Model Optimization Algorithm Based on Data Set Denoising. *J. Phys.: Conf. Ser.* 1631 012063. DOI: 10.1088/1742-6596/1631/1/012063
11. Huei and Lung Liang (2020) Doppler Flow measurement of lower extremity arteries adjusted by pulsatility index. *AJR*. 214. doi.org/10.2214/AJR.19.21280.
12. Noora Shrestha (2020) Detecting Multicollinearity in Regression Analysis. *American Journal of Applied Mathematics and Statistics* pp. 39-42. DOI: 10.12691/ajams-8-2-1.
13. Jamal I Daoud (2017) Multicollinearity and regression analysis. *Journal of Physics: Conference Series.* 949 (2017) 012009. DOI: 10.1088/1742-6596/949/1/012009
14. Vatcheva KP, Lee M, McCormick JB, Rahbar MH (2016) Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology (Sunnyvale)* 6(2):227. DOI: 10.4172/2161-1165.1000227.
15. Breusch TS and Pagan AR (1979) A simple test for heteroscedasticity and random coefficient variation. *Econometrica*. Vol. 47, No.5, pp. 1287–1294. DOI: https://doi.org/10.2307/1911963.
16. Osborne Jason W and Waters Elaine (2019) Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research, and Evaluation*. DOI: https://doi.org/10.7275/r222-hv23
17. Hickey GL, Kontopantelis E, Takkenberg JJM, Beyersdorf F (2019) Statistical primer: checking model assumptions with regression diagnostics. *Interact CardioVasc Thorac Surg* 28:1–8.
18. Nayak Debismita & Tantravahi Sai (2024) On building machine learning models for medical datasets with correlated features. *Computational and Mathematical Biophysics* 12. 10.1515/cmb-2023-0124.
19. Ogutu JO, Schulz Streeck T and Piepho HP (2012) Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC Proc 6 (Suppl 2)*. Available at: https://doi.org/10.1186/1753-6561-6-S2-S10.
20. Tibshirani R (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society* 58(1). pp. 267–288. Available at: http://www.jstor.org/stable/2346178.

21. Finch WH and Finch ME (2016) Regularization Methods for Fitting Linear Models with Small Sample Sizes: Fitting the Lasso Estimator using R. *Practical Assessment, Research, and Evaluation* 21(1): 7. DOI: https://doi.org/10.7275/jr3d-cq04
22. Hui Zou, Trevor Hastie (2005) Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology.* pp. 301–320.
Available at: https://doi.org/10.1111/j.1467-9868.2005.00503.x