ANALYSIS OF PSYCHOMETRIC DATA USING STATISTICAL AND MACHINE LEARNING METHODS

by

KRISHNAPRIYA SUBRAMANIAN

M.E (ECE), Anna University, India, 2011

A project
presented to Ryerson University
in partial fulfillment of the
requirement for the degree of
Master of Engineering
in the Program of
Electrical and Computer Engineering

Toronto, Ontario, Canada, 2016

Author's Declaration

I hereby declare that I am the sole author of this project. This is a true copy of the project, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this project to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this project by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my project may be made electronically available to the public.

Abstract

ANALYSIS OF PSYCHOMETRIC DATA USING STATISTICAL AND MACHINE LEARNING METHODS

Master of Engineering 2016

KRISHNAPRIYA SUBRAMANIAN

Electrical and Computer Engineering

Ryerson University

The objective of this thesis is to analyse the psychometric data using statistical and machine learning methods. Psychological data are analysed to predict illness and injury of athletes. Regression technique, one of the statistical processes for estimating the relationship among variables is used as basis of this thesis. We apply the linear regression, time series and logistics regression to predict illness and well-being. Our linear regression simulation results are mainly used, to understand the data well. By reviewing the results of linear regression, time series model is developed which predicts sickness one day ahead. The predicted values of this time series model are continuous. However, logistic regression can be used, to provide a probabilistic approach to predict the future levels as a categorical value. Hence we have developed a binomial logistics regression model, when observation variable is the type of dichotomous. Our simulation results show that this prediction model performs well. Our empirical studies also show that our method can act as early warning system for athletes.

Acknowledgments

I am using this opportunity to express my gratitude to everyone who supported me throughout this project. I am sincerely thankful to my supervisor, Dr. Xiao-Ping Zhang, whose encouragement, guidance and support from the initial to the final level enabled me to develop an understanding of the subject. Besides providing constructive discussions, his kind help was crucial for me to build my technical writing skills and the right attitude. I will definitely benefit from those research skills and the knowledge he shared with me.

I would also like to thank every member of my thesis defense committee. I appreciate their time, efforts, and contributions to this work.

Special thanks go to my colleague Dr. Adnan Gavili at the Communications and Signal Processing Applications Laboratory (CASPAL). It was always a pleasure to discuss technical issues and exchange ideas with him.

Finally, I would like to give my deepest gratitudes to my husband for the care and support he has provided me. My parents have always been the driving force for my accomplishments and will always be for my future endeavors. I dedicate this thesis to them.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and Objectives

Prediction of athletes injury/illness is an important issue in the sports field. This has led to many researches [1] in predictive work. Predicting an individual athlete injury/illness based upon his/her past record can be critical in the selection of team members in international competitions. This process is highly subjective usually requires much expertise and negotiating decision making. This project deals foremost with predictions of future wellness of athlete based upon historical data. The goal is to develop the model that can forecast risks of illnesses and injuries. It enables sports organizations and trainers to monitor wellness and health risk of their athletes. It acts as an early warning system which gives support staff and supervisors actionable insights that they can use to prevent health problems from serious effects, effectively stopping such issues in their tracks.

The traditional method of its estimation employs Ordinary Least Squares (OLS) regression. This thesis explores various methods of prediction like regression, time series analysis and logistics regression. This study also gives the introduction about machine learning/clustering. The prediction is based on various regression methods derived from everyday athletes data.

The primary objective of this thesis is to predict one day and one week ahead illness/wellness of athletes using the past history of their data. In order to implement this model, we have conducted several studies to understand how well is the data set followed

by several regression analyses and finally came out with the logistics regression model which gives prediction in terms of probability.

1.2 Problem Formulation

Wellness is really important for athletes to perform well. The prediction of illness will always help them to take more care on their health in advance which avoids getting sick, that affects their performance. Coaches may examine an athlete from many points of view to determine if they will be successful in their sports. Their assessment mostly includes an examination of physical ability. Athletes success is not only based on this, but also based on the psychological attributes and mental skills. It is better to consider all these factors when predicting their chances of getting injured or sick. For example, when it is really near to tournaments, they have to take extra care in their health by taking rest, having a good sleep, etc.There are several attributes that affect health condition of athletes. More specifically, one of the hydration factor, which is a level of intake of water that affects their health. If they dont drink water properly that will lead to several health problems. Therefore, this study considered not only the health status additionally considered several psychological factors to predict illness.

1.3 Literature Review

The World Health Organization(WHO) [2], defines term wellness is best defined as an individual's "physical, mental, and social well being and not merely absence of disease". In one of the previous studies [3], theorists and college students seem to agree that wellness is more than just a physical issue. They mentioned example of two separate surveys, found that college students believed emotional and social dimensions of wellness were just as important as the physical dimension.

In [4], shown the model of wheel of wellness. The model proposed with five life tasks. These five tasks were spirituality, self regulation, work, friendship, and love. They mentioned based on research using an assessment instrument developed from the model (the Wellness

Evaluation of Lifestyle, or WEL), the life task of work was subdivided into the two tasks of work and leisure (rest). The life task of self regulation included seven components. 12 subtasks were clearly defined in [4]. These 12 subtasks are as follows: (a) sense of worth, (b) sense of control, (c) realistic beliefs, (d) emotional awareness and coping, (e) problem solving and creativity, (f) sense of humour, (g) nutrition, (h) exercise, (i) self-care, (j) stress management, (k) gender identity, and cultural identity. The wheel of wellness is shown in figure 1.1, which is showing the health is surrounded by different psychological factors.

In [5], stated that researchers have found injury rates to be between 65 % and 95 % per year. Pre-injury research has identified a number of factors that could increase injury risk among athletes. They [5] also said that Bahrand Krosshaug [6], developed a theoretical model in which internal risk factors (such as health, physical fitness, skill level, and psychological factors) are associated with an athletes predisposition towards increased injury risk.

Time series analysis tools that are used for modelling and forecasting time series datasets are widely used in various fields, including economic field (business, finance, foreign exchange and stock problems), investment, engineering energy, internet, and network traffic as stated in [7]. An efficient prediction ability is required in order to assist the process of decision making. In the past studies, many strategies have been established regarding the time series prediction. The methods are grouped into two categories are statistical and machine learning (ML) methods. There are several types of statistical methods: Autoregressive (AR), moving average (MA), autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA). Statistical methods are good enough to be used for forecasting purposes if the amount of data is not too much with linear data types.

Therefore, to model prediction of illness well known regression analysis is used. It is a great challenge to achieve accuracy in prediction. In order to understand the data, histogram analysis and correlation analysis are used. The author in [4], clears that health is dependent

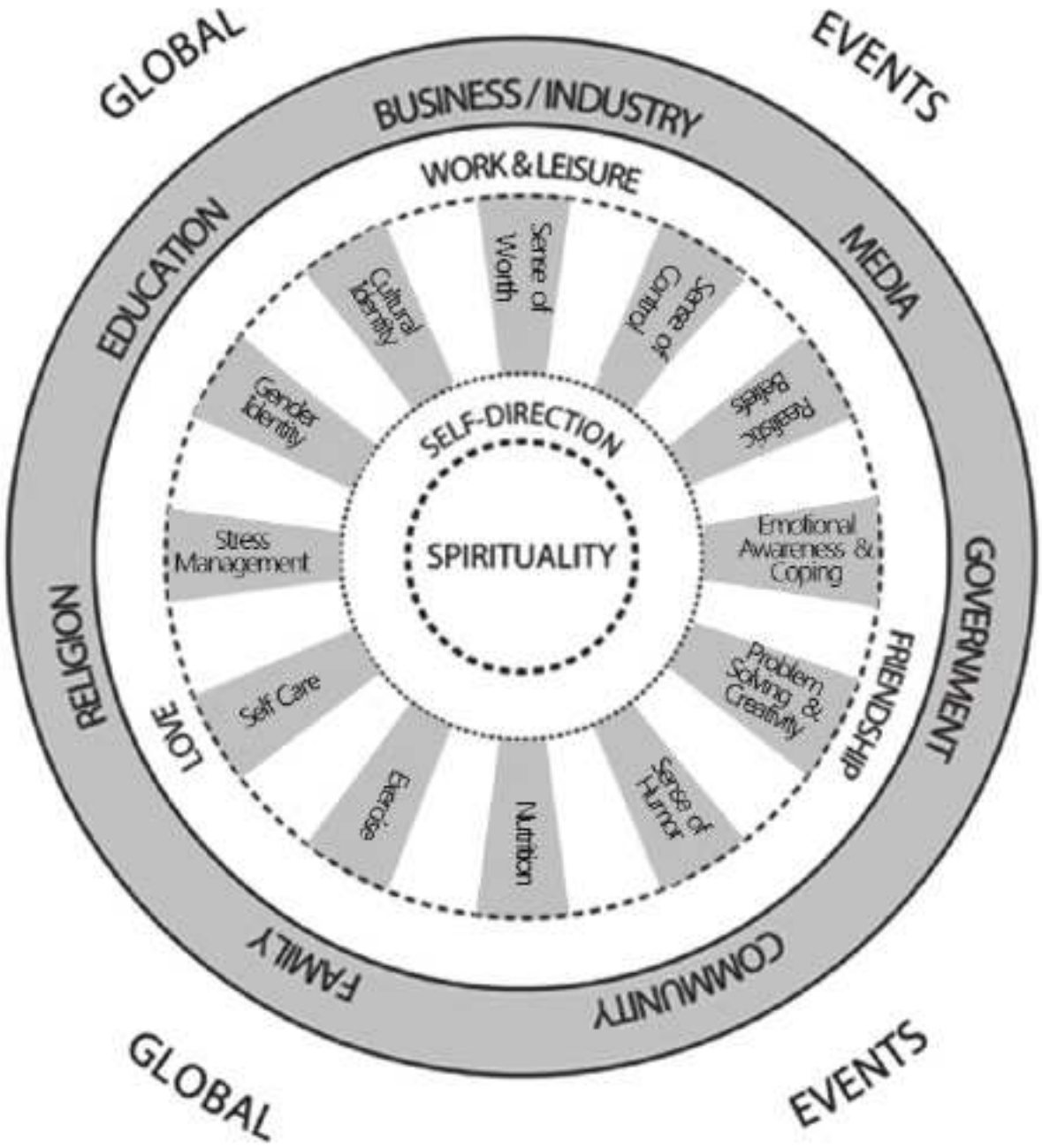


Figure 1.1: Wheel of Wellness, Source: Reference [4]

on several features like stress, rest, irritation,nutrition, etc.

1.4 Data Set and Feature Description

This research used the dataset from various sports. This project is analysing prediction of 8 sports domains. All the data for all 8 domains are compiled into excel file which has several features includes health, stress, nutrition,hydration, sleep, energy, soreness, irritability, rest, enjoyment, exertion(physical or mental effort) additionally with fields of athlete name, ID, group name and date. For the training and testing purposes Sports-1 domain is taken, as it has a good history of data.

Athletes enter the data for several features on everyday basis.They enter each feature ranges from 1 to 7. For example, in health feature, if he/she feels good for the day, they enter 7 and when they are sick, the range given by them is 1. When they feel moderate in health, the score may differ from 2 to 6. The dataset includes thoughts of the day in which athletes enter how they feel for the day will help us to predict their feelings from the keywords like nausea, confusion, dizziness or did more work out. The athlete data have following fields, each range from 1 to 7 which are called as features/variables in the following sections:

1. Health (1=very very bad, 7=very very good)
2. Stress (1=overwhelmed, 7=no stress)
3. Nutrition (1=very very bad, 7=very very good)
4. Hydration (1=very very bad, 7=very very good)
5. Sleep (1=very very bad, 7=very very good)
6. Energy (1=very very low, 7=very very high)
7. Soreness (1=very very sore, 7=no soreness)
8. Irritability (1=Get away from me, 7=I love everyone).
9. Rest (1= No rest, 7= Took good rest)

10. Enjoyment (1= No enjoyment, 7= Enjoyed well)

11. Exertion (1=Low effort, 7=High effort).

1.5 Organization of Thesis

In chapter 2, we briefly discuss the methods of regression analysis, which include ordinary least squares, time series and logistics regression. We also discuss the theory of machine learning methods which includes support vector machine, feed forward neural network and clustering.

In chapter 3, we present the linear regression and explained about residuals and R-squared. We discuss the various steps involved in variable selection and model validation procedure. This section also explains the terminology used during the analysis.

In chapter 4, we present the time series regression analysis and discuss the auto regressive model with exogenous inputs (ARX).

In chapter 5, we present the logistics regression with the simple and binary complex logistics model. We also present the model fit of the logistics regression.

In chapter 6, results of linear regression, time series and logistics regressions are presented. We discuss the simulation results and also calculated the error rate to validate the model.

Chapter 7 is the conclusion of this thesis.

Chapter 2

Methodology

Forecasting is the process of making predictions of the future state of a parameter based on the past and the present data available for that parameter and analysis of trends. There exist numerous statistical and analytical methods for forecasting data or parameters related to data, to name a few, time series, cross-sectional or longitudinal data. The type of forecasting can also be categorized into two general cases, namely, qualitative and quantitative approaches. Qualitative forecasting techniques are subjective, based on the opinion and judgement of consumers, experts and etc. This type of prediction is subject to change due to different understanding of the prediction problem as two experts may have different viewpoints regarding how to predict the future state of a parameter. Quantitative approaches, however, provides a solid and objective method such that the result of quantitative predictions are reproducible. Following sections summarizes the most applicable prediction approaches to numerical data by considering the example of predicting health feature.

2.1 Time Series analysis

2.1.1 Average approach

In this approach [8], the predictions of all future values are equal to the mean of the past data. This approach can be used with any sort of data where past data is available. In time series notation (2.1),

$$\hat{Y}_{N+h|N} = \bar{Y} = (Y_1 + \cdots + Y_N)/N \tag{2.1}$$

where, $(Y_1 + \cdots + Y_N)$ is the past data, $\hat{Y}_{N+h}$ is the future feature in time series notation, T is the total number of observations and h is the forecast horizon.

Even though this method used time series notation, this can also be used for cross-sectional data (when we are predicting a value not included in the data set). Then the prediction for values, not observed is the average of those that have been observed.

2.1.2 Naive approach

Naive forecasts [8], are the most cost-effective forecasting model. In time series data, using a naive approach would produce forecasts that are equal to the last observed value. This method works quite well for economic and financial time series, which often have patterns that are difficult to reliably and accurately predict. If the time series is believed to have seasonality, the seasonal naive approach may be more appropriate where the forecasts are equal to the value from last season. In time series notation (2.2),

$$\hat{Y}_{N+h|T} = Y_N \tag{2.2}$$

where, $\hat{Y}_{N+h}$ is the future feature in time series notation, Y_N is the last observed health feature.

2.1.3 Drift Method

A variation on the naive method [8], is to allow the forecasts to increase or decrease over time, where the amount of change over time (called the drift) is set to be the average change seen in the historical data. So the forecast for time T+h is given by (2.3),

$$\hat{Y}_{N+h|N} = Y_N + \frac{h}{N-1}\sum_{t=2}^{N}(Y_t - Y_{t-1}) = Y_N + \frac{h(Y_N - Y_1)}{N-1} \tag{2.3}$$

where, $(Y_t - Y_{t-1})$ is the difference between two consecutive days of health feature, Y_t is the observation at time t, Y_N is the last observed health feature, $\hat{Y}_{N+h}$ is the future feature in time series notation, N is the total number of observations and h is the forecast horizon.

2.1.4 Moving Average

Moving average [9], is a calculation to analyse data points by creating a series of averages of different subsets of the full data set. It is also called a moving mean (MM) or rolling mean and is a type of finite impulse response filter. Given a series of numbers and a fixed subset size, the first element of the moving average is obtained by taking the average of the initial fixed subset of the number series. Then the subset is modified by "shifting forward", that is, excluding the first number of the series and including the next number following the original subset in the series. This creates a new subset of numbers, which is averaged. This process is repeated over the entire data series. The plot line connecting all the fixed averages is the moving average.

A moving average is a set of numbers, each of which is the average of the corresponding subset of a larger set of datum points. A moving average may also use unequal weights for each datum value in the subset to emphasize particular values in the subset. A moving average is commonly used with time series data to smooth out short-term fluctuations and highlight longer-term trends or cycles. The threshold between short-term and long-term depends on the application, and the parameters of the moving average will be set accordingly. SMA is the simple moving average, which is the unweighed mean of the previous n data. An example of a simple equally weighted running mean for a n-day sample of the closing price is the mean of the previous n days closing prices. If those prices are $P_M + P_{M-1} + \cdots + P_{M-(N-1)}$ then the formula is given by (2.4),

$$\text{SMA} = \frac{P_M + P_{M-1} + \cdots + P_{M-(N-1)}}{N} \tag{2.4}$$

When calculating successive values, a new value comes into the sum and an old value drops out, meaning a full summation, each time is unnecessary for this simple case and this is given by (2.5),

$$\text{SMA}_{Today} = \text{SMA}_{Yesterday} = \frac{P_M}{N} + \frac{P_{M-N}}{N} \tag{2.5}$$

A major drawback of the SMA is that, it lets through a significant amount of the signal shorter than the window length. Worse, it actually inverts it. This can lead to unexpected

artifacts, such as peaks in the smoothed result appearing where there were troughs in the data. It also leads to the result being less smooth than expected since some of the higher frequencies are not properly removed.

2.1.5 Cumulative Moving Average

On a cumulative moving average [9], the data arrive in an ordered datum stream, and it is the average of all data until a current datum point. Moreover, cumulative moving average filtering can be written as in (2.6),

$$\mathrm{CMA}_N = \frac{X_1 + \cdots + X_N}{N} \tag{2.6}$$

Where $X_1 + \cdots + X_N$ is the health feature from day 1 to N, N is the history of data.The brute-force method to calculate this, store all of the data and calculate the sum and divide by the number of datum points every time a new datum point arrived. However, it is possible to simply update cumulative average as a new value X_{N+1} then the formula is given by (2.7),

$$\mathrm{CMA}_{N+1} = \frac{X_{N+1} + N \cdot \mathrm{CMA}_N}{N+1} \tag{2.7}$$

Finally the equation solved as (2.8),

$$\mathrm{CMA}_{N+1} = \mathrm{CMA}_N + \frac{X_{N+1} - \mathrm{CMA}_N}{N+1} \tag{2.8}$$

2.1.6 Weighted Moving Average

A weighted average [9], is an average that is multiplying factors to give different weights to data at different positions in the sample window. Mathematically, the moving average is the convolution of the datum points with a fixed weighting function that is given by equation (2.9),

$$\mathrm{WMA}_N = \frac{N\mathrm{p}_M + (N-1)\mathrm{p}_{M-1} + \cdots + 2\mathrm{p}_{(M-N+2)} - \mathrm{p}_{(M-N+1)}}{N + (N-1) + \cdots + 2 + 1} \tag{2.9}$$

Where $N + (N-1) + \cdots + 2 + 1$ is the N-day WMA, the latest day has weight n, the second latest $N-1$, etc., down to one. Weighted moving average (WMA) has the specific meaning of weights that decrease in arithmetical progression.

2.1.7 Autoregression and Moving average(ARMA)

Given a time series of data X_t, the ARMA model [10], is a tool for understanding and perhaps, predicting future values in this series. The model consists of two parts, an autoregressive (AR) part and a moving average (MA) part. The notation AR(p) refers to the autoregressive model of order p. The AR(p) model is written as in (2.10),

$$\mathrm{X}_t = \beta + \sum_{i=1}^{p} \beta_i X_{t-i} + \varepsilon_t \tag{2.10}$$

Where β is the constant, $\beta_1 \cdots \beta_p$ are the coefficients and random variable ε is the white noise. The notation MA(q) refers to the moving average model of order q and it is given by (2.11),

$$\mathrm{X}_t = \mu + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \tag{2.11}$$

Where μ is the expected value of the parameter to be predicted, $\theta_1 \cdots \theta_q$ are the parameters of the model and random variable ε_t is the white noise error terms.

The notation ARMA (p,q) refers to the model with p autoregressive terms and q moving-average terms. This model contains the AR(p) and MA(q) models and this is represented as (2.12),

$$\mathrm{X}_t = \beta + \sum_{i=1}^{p} \beta_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \tag{2.12}$$

2.1.8 Linear Prediction

Linear prediction [11], is a mathematical operation where future values of a discrete-time signal are estimated as a linear function of previous samples. The most common representation is given by (2.13),

$$\hat{X}(n) = \sum_{i=1}^{p} \beta_i X(n-i) \tag{2.13}$$

Where $\hat{X}(n)$ is the predicted signal value, $X(n-i)$ is the observed values (actual values) and β_i is the predictor coefficients. The error generated by this estimate is given by (2.14),

$$e(n) = X(n) - \hat{X}(n) \tag{2.14}$$

Where $X(n)$ is the actual value and $\hat{X}(n)$ is the predicted value. Most common choice in the optimization of parameters β_i is the root mean square criterion which is also called the autocorrelation criterion. In this method, we minimize the expected value of the squared error $E[e^2(n)]$, which yields the below equation (2.15),

$$\sum_{i=1}^{p} \beta_i R(j-1) = -R(j) \tag{2.15}$$

for $1 \leq j \leq p$, where R is the autocorrelation of signal x(n), defined as in equation (2.16)

$$R(i) = E[X(n)X(n-i)] \tag{2.16}$$

Where $E[X(n)X(n-i)]$ is the expectation of true signal and previous observed value.

2.2 Cross Sectional Prediction/Regression

2.2.1 Linear Regression

In statistics, regression analysis [10], is a statistical process for estimating the relationships among variables. It includes many techniques for modelling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or predictors). More specifically, regression analysis helps to understand how the typical value of the dependent variable (or criterion variable) changes when any one of the independent variables is varied, while the other independent variables are held fixed. The response variable y is a scalar. Simple linear regression is the least squares estimator of linear regression with single explanatory variable. It fits a straight line through the set of n points as shown in figure 2.1, in such a way that makes the sum of squared residuals of the model as small as possible.

In more general multiple regression model, there are p independent variables. The model with p variables is given by (2.17),

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i \tag{2.17}$$

Where β is the unknown parameter, X_{ip} is the i^{th} observation on the p^{th} independent variable, where the first independent variable takes the value 1 for all i (so β_1 is the regression

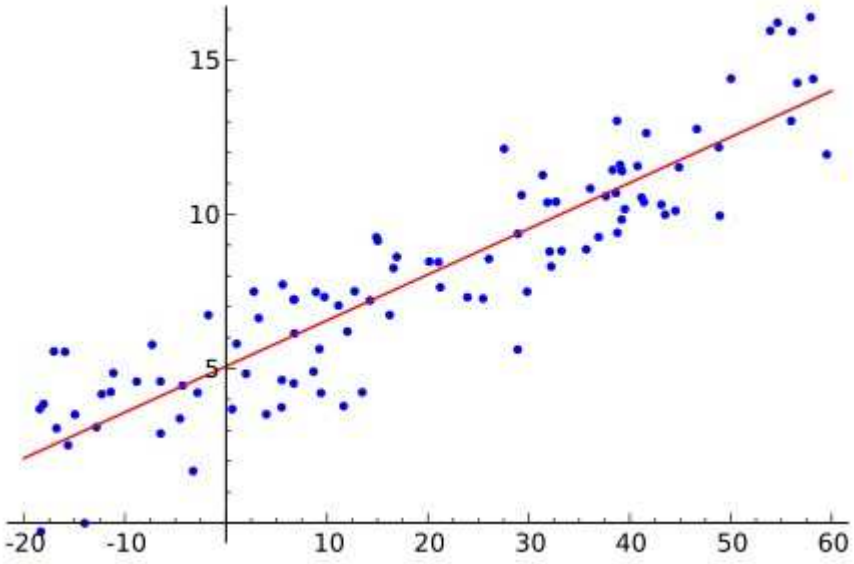


Figure 2.1: Linear Regression , Source: en.wikipedia.org

intercept), ε is the error term. The least squares parameter estimates are obtained from p normal equations. The residual can be written as below (2.18),

$$\varepsilon_i = Y_i - \hat{\beta}_1 X_{i1} - \cdots - \hat{\beta}_p X_{ip} \tag{2.18}$$

The normal equation is given by (2.19),

$$\sum_{i=1}^{N} \sum_{k=1}^{p} X_{ij} X_{ik} \hat{\beta}_k = \sum_{i=1}^{N} X_{ij} Y_j, j = 1 \cdots p \tag{2.19}$$

In matrix notation, the normal equations are written as in (2.20),

$$(X^T X)\hat{\beta} = X^T Y \tag{2.20}$$

Where the ij element of X is X_{ij}, the i element of the column vector Y is Y_i, and the j element of $\hat{\beta}$ is $\hat{\beta}_j$ and X^T is the transpose matrix of X. Thus X is $n \times p$, Y is $n \times 1$, and $\hat{\beta}$ is $p \times 1$. The solution is given by (2.21),

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{2.21}$$

2.2.2 Nonlinear Regression

In statistics, non-linear regression [10], is a form of regression analysis in which observational data is modelled by a function which is a non-linear combination of the model parameters and depends on one or more independent variables. The data are fitted by a method of successive approximations. One simple implementation of this method using the linear regression is to linearise the regression function in the parameters that are to be estimated. More specifically, the assumption underlying this procedure is that the model can be approximated by a linear function as shown in equation (2.22),

$$Y = f(X_i, \beta) + \varepsilon_i \tag{2.22}$$

Where $f(X_i, \beta)$ is the regression function, β is the unknown parameters, which may represent a scalar or a vector,X_i are the given features and ε_i is the error whose distribution may or may not be normal.

2.3 Machine Learning

Machine learning [12], explores the study and construction of algorithms that can learn from and make predictions on data. Such algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions, rather than following strictly static program instructions. Machine learning focuses on prediction, based on known properties learned from the training data.

2.3.1 Types of Problems

Machine learning tasks are typically classified into three broad categories, depending on the nature of the learning "signal" or "feedback" available in a learning system. These are:

Supervised learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.

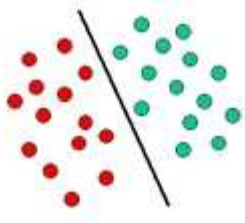


Figure 2.2: Example of SVM
Source: http://www.statsoft.com/

Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

Reinforcement learning: A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle), without a teacher explicitly telling it whether it has come close to its goal. Another example is learning to play a game by playing against an opponent.

2.3.2 Supervised Learning Algorithm

Support vector Machine (SVM)

Support vector machines [12], are supervised learning models with associated learning algorithms that analyse data and recognize patterns, used for classification. It defines decision boundaries. Decision boundary that separates the different set of actions. An example is shown in figure 2.2. The boundary line that separates the green and red objects.

Feed forward Neural Network (FFNN)

The simplest kind of FFNN [12], is a single perceptron network which has a single layer of output nodes. The inputs are fed directly to the output via some series weights. The sum of the products of the weights and the inputs is calculated at each node, and if the value is above some threshold (typically 0) the neuron fires and takes the activated value (typically 1); otherwise it takes the deactivated value (typically -1). There is no feedback between layers. It is suitable for linearly separable data. Feed forward neural network is shown in

figure 2.3.

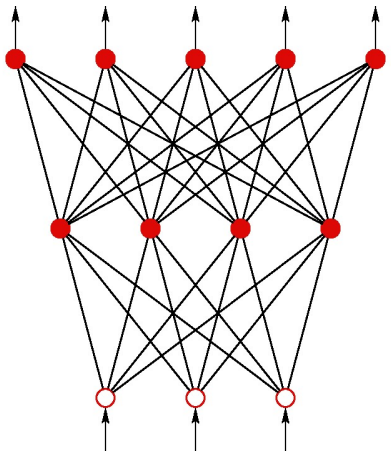


Figure 2.3: Feed Forward Neural Network
Source:http://www.fon.hum.uva.nl/

2.3.3 Unsupervised Learning: K-means Clustering

K-means [12], is one of the unsupervised learning algorithms that solve the well-known clustering problem. It classifies a given data set through a certain number of clusters. The main aim is to define k centroids, one for each cluster. The better way is to place them as much as possible far away from each other. The points in the dataset are arranged by calculating the distance between each center of clusters and it takes the minimum distance group. After the first stage of grouping is done, it will recalculate the center of the clusters again. The process will be repeated until no more changes can be done. An example of cluster assignment with two centroids is shown in figure 2.4.

Finally, this algorithm aims at minimizing an objective function like in (2.23),

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||(\mathbf{x}_i^{(j)} - \mathbf{c}_j)^2|| \tag{2.23}$$

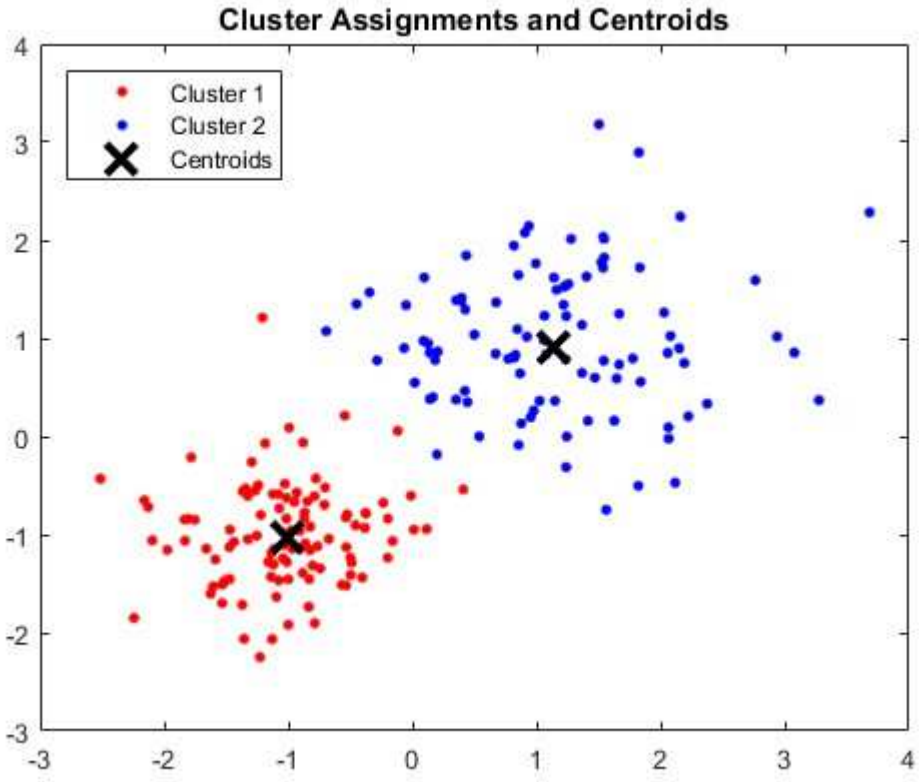


Figure 2.4: Cluster Assignments with Two Centroids, red and blue dots showing two groups
Source: http://www.mathworks.com/

Where $||\mathbf{x}_i^{(j)} - \mathbf{c}_j||$ is a chosen distance measure between a data point $\mathbf{x}_i^{(j)}$ and the cluster center $\mathbf{c}_j$ is an indicator of the distance of the n data points from their respective cluster centers.

2.3.4 Deep Learning

Deep learning [13] is a branch of machine learning based on a set of algorithms that attempt to model high-level abstractions in data by using model architectures, with complex structures or otherwise, composed of multiple non-linear transformations. Deep learning is part of a broader family of machine learning methods based on learning representations of data. An observation (e.g., an image) can be represented in many ways, such as a vector of intensity values per pixel, or in a more abstract way as a set of edges, regions of particular shape,etc.

2.3.5 Outlier Method

There exist numerous methods to obtain the outlier in a multivariate distribution,for example from estimated marginal distributions, estimated joint distribution, fitted marginal distribution and fitted joint distribution. Obtaining outliers from joint distribution is more accurate than the marginal distribution. Moreover, estimating the distribution for a multivariate distribution function is time consuming, hence using a fitting method is beneficial. An approach to finding outliers is to consider the distance of each point to some central location. Data points that are unreasonably far away are considered as outliers. Mahalanobis test [12], is a powerful test which measure a modified version of distance as that has F-distribution shown in (2.24),

$$d(X_i) = d_{\hat{\Sigma}^{-1}}(X_i - \mu) = (X_i - \mu)^T \hat{\Sigma}^{-1} (X_i - \mu) \tag{2.24}$$

If $d(X_i) >$ threshold (we choose the threshold) then we consider that measurement as outlier.μ is the mean of the data, X_i is the given data, and $\hat{\Sigma}$ is the estimate of the covariance matrix. Outliers can remove unwanted data on both sides of the tails. This will help to perform proper classification of clusters.

2.4 Chapter Summary

In this chapter, we reviewed the several methods to perform prediction. This chapter gives a brief intro of regression, time series regression, linear prediction and machine learning methods.

Chapter 3

Linear Regression Estimation

In this chapter, we apply ordinary least squares regression technique to identify the significant changes in beta and examine if any events are associated with such a change. We develop a methodology to identify the significant features in the developed model.

3.1 Ordinary Least Squares

The most widely used statistical methods for fitting a regression line is the method of least squares regression. In statistics, ordinary least squares (OLS) is a method for estimating the unknown parameters in a linear regression model [14], with the goal of minimizing the differences between the observed responses in some arbitrary dataset and the responses predicted by the linear approximation of the data that is visually this is seen as the sum of the vertical distances between each data point in the set and the corresponding point on the regression line - the smaller the differences, the better the model fits the data. The equation of linear regression is shown in (3.1),

$$Y = \beta X + \varepsilon \tag{3.1}$$

Where Y is the independent variable, X is the matrix of predictors and random variable ε is the error term. β is given by $\beta_0 + \beta_1 + \cdots + \beta_i$ where β_0 is the constant(intercept) and $\beta_1 + \cdots + \beta_i$ are the regression coefficients for i predictors.

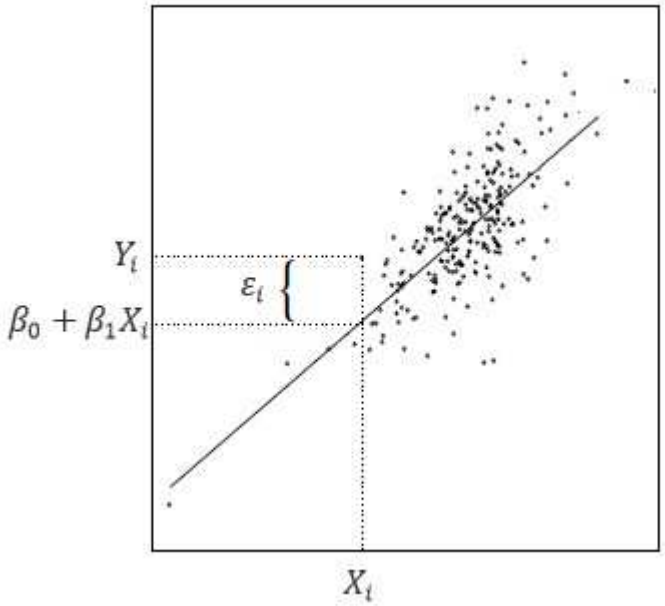


Figure 3.1: Scatter Diagram with Fitted Line
Source: Book of Regression Analysis by Example.

3.1.1 Fitting a Line to a Scatter of Data

The starting point is a set of points in a scatter diagram corresponding to a "n" paired observations (X_i, Y_i),$i = 1, 2, \cdots n$. The aim is to find the line that gives the best fit to these points [15]. The formula for the line is given by (3.2),

$$Y_i = \beta_0 + \beta_1 X_i, i = 1, 2 \cdots n \tag{3.2}$$

Where Y is the variable to be explained/Endogenous variable/Dependent variable, β_0 is the intercept , β_1 is the slope of the line and X is the explanatory variable/ Independent variable/Regressor/Covariate/ Exogenous variable.The idea is to explain the differences in the outcomes of the variable Y in terms of differences in the corresponding values of the variable X. The scatter diagram with fitted line is shown in fig 3.1

Scatter diagram with observed data (X_i, Y_i) regression line $(\beta_0 + \beta_1 X_i)$ and residual ε_i is given by (3.3),

$$\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i, i = 1, 2 \cdots n \tag{3.3}$$

Where ε_i is the error that occurs due to the prediction of Y_i by means of the variable X_i using the linear relation.

3.1.2 Residuals and R^2

Least Square Residuals

Given the observations $(X_1, Y_1) \cdots (X_n, Y_n)$, the residuals are given by (3.4),

$$\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i, i = 1, 2 \cdots n \tag{3.4}$$

The two properties of residuals are given by (3.5),

$$\sum \varepsilon_i = 0; \sum (X_i - \bar{X}) = 0 \tag{3.5}$$

Where $\bar{X}$ is the mean of the variable X. $\bar{X} = \sum \frac{X_i}{n}$. The residuals have zero mean and they are correlated with the explanatory variable.

Three Sum of Squares

A traditional way to measure the performance of least squares [15] is to compare the sum of Squared residuals with the sum of squares of $(Y_i - \bar{Y})$. Rewrite equation (3.4) as in (3.6),

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i (Y_i - \bar{Y}) = (X_i - \bar{X}) + \varepsilon_i \tag{3.6}$$

Where $\bar{Y}$ is the mean of the variable Y. $\bar{Y} = \sum \frac{Y_i}{n}$. The differences from the mean $(Y_i - \bar{Y})$ can be decomposed as a sum of two components, a component corresponding to the difference from the mean of the explanatory variable $(X_i - \bar{X})$ and an unexplained component described by the residual ε_i. The sum of squares of $(Y_i - \bar{Y})$ consists of two components are (3.7) and (3.8).

$$\sum (Y_i - \bar{Y})^2 = \beta_1^2 \sum (X_i - \bar{X})^2 + \sum (\varepsilon_i)^2 \tag{3.7}$$

$$SST = SSE + \beta_1 SSR \tag{3.8}$$

Here SST is called the total sum of squares, SSE is the explained sum of squares and SSR is the Sum of squared residuals.

Coefficient of Determination: R^2

The three sums of squares depend on the scale of measurement of the variable Y. To get a performance measure that is independent of scale, it is divided through by SST. The coefficient of determination, denoted by the symbol R^2, is defined as the relative explained sum of squares as (3.9),

$$R^2 = \frac{SSE}{SST} = \frac{\beta_1^2 \sum (X_i - \bar{X})^2}{\sum (Y_i - \bar{Y})^2} \tag{3.9}$$

Where R^2 is equal to the square of the correlation coefficient between X and Y and it is given by equation (3.10),

$$R^2 = 1 - \frac{\sum \varepsilon_i^2}{\sum (Y_i - \bar{Y})^2} \tag{3.10}$$

Where SSE is the explained sum of squares and SST is the total sum of squares, β_1 is the coefficient, $\bar{Y}$ is the mean of Y , $\bar{X}$ is the mean of X and ε is the error term.

Quality of R-Squared

The coefficient determination (R^2) is a regression statistic that shows how closely the predictor variable X are related to the response variable. It tells how well the regression model fits the data. It lies in the range of (-1, 1). When $R^2 = +1$, exists a positive correlation between X and Y. When R^2 = -1, there exists a negative correlation between X and Y. A limitation with the coefficient of determination is that, it is not the best thing to access the quality of the model. Therefore, additional tests of significance must be performed. One of these tests, the analysis of variance test (ANOVA test).

Terminology Used

The implementation of this thesis is done by using MATLAB. Some of the terminology used during regression analysis are:

Estimate: The estimate is the coefficient estimation for each corresponding term in the model.

Standard Error (SE): The standard error of the estimate is a measure of the accuracy of

predictions. It is given by equation (3.11),

$$\sigma = \sqrt{\frac{\sum(Y_i - Y_i')^2}{N}} \tag{3.11}$$

Where σ is the standard error of the estimate, Y is the actual value of the feature, Y' is the predicted value of the corresponding feature. $\sum(Y_i - Y_i')^2$ is the sum of squared differences between the actual and predicted value.

T- Statistics (tstat): T- Statistics is the ratio of the estimate and standard error [16]. In regression analysis, the t-statistics is useful for making inferences about the regression coefficients. The hypothesis test on the coefficient, tests the null hypothesis that is equal to zero (that is the term is not significant) versus the alternate hypothesis that the coefficient is different from zero.

F- Statistics: In linear regression, the F-statistic is the test statistic for the analysis of variance (ANOVA) approach to test the significance of the model or the components in the model.

P- Value: P-value for the F statistic of the hypothesis test that the corresponding coefficient is equal to zero or not. If a p-value is less than 0.05, indicates the strong evidence against the null hypothesis, so it is rejected means features are significant. A large P-value greater than 0.05, indicates weak evidence against the null hypothesis means features are not significant.

Error degrees of freedom (DF): Error degrees of freedom is given by, $n - p$, where n is the number of observations, and p is the number of coefficients in the model, including the intercept. For example, the model has eight predictors (p) and n=1182, so the Error degrees of freedom is $1182 - 8 = 1174$.

Root Mean Squared Error (RMSE): The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed.

F-statistic vs. constant model: Test statistic for the F-test on regression model. It tests for a significant linear relationship between the response variable and the predictor variables.

3.1.3 Initial Variable Selection

One way to choose variables, called forward selection, [15] is to do a linear regression for each of the X variables, one at a time, then pick the X variable that had the highest R^2. Add the X variable that increases the R^2 by the greatest amount, if the P value of the increase in R^2 is below the desired cut-off (p-value which may or may not be 0.05, depending on how is the feeling about extra variables in regression model). Continue adding X variables by adding another X variable does not significantly increase the R^2.

A second technique, called backward elimination, is to start with a multiple regression using all of the X variables, then perform multiple regressions with each X variable removed in turn. Eliminate the X variable whose removal causes the smallest decrease in R^2, if the P value is greater than the "P-to-leave". Continue removing X variables until getting significant in all X variables.

1) As a base level, choose the number of independent variables that included in the model. Having more independent variables does not guarantee to have best model. It is important to choose the significant independent variables corresponds to the dependent variable. Check the linear regression with the less number of features. Compare R- squared value of this model with previous model. If this is better, having more feature is useful. Health is taken as the dependent variable. Independent features are Stress, Nutrition, Hydration, Sleep, Energy, Soreness and Irritability. Initially the model is designed with 4 independent features are stress, nutrition, hydration and sleep. The coefficients are shown in table 3.1, where N is the total number of samples used.

Linear Regression Model:Health $\sim$ 1+ Stress + Nutrition + Hydration + Sleep

The model is tested by adding additional feature, one feature at a time with the existing four features.Achieved R- squared value of 0.109 with 4 features stress, nutrition, hydration and sleep. Achieved R- squared value is 0.117 when another feature energy is included. R- Squared value is 0.118 with stress, nutrition, hydration, sleep, energy and soreness. R-Squared value is 0.138 with stress, nutrition, hydration, sleep, energy, soreness and Irritability. R- Squared values are getting improve for the addition of features. More features are

Table 3.1: Linear Regression Model with Four Independent Features,N=481

	Estimate	SE	tstat	pvalue
Intercept	2.921	0.3641	8.022	8.16e-15
Stress	0.084	0.0425	1.984	0.0478
Nutrition	0.093	0.0671	1.387	0.1661
Hydration	0.158	0.0648	2.443	0.0149
Sleep	0.190	0.0432	4.401	1.33e-05
Number of observations : 481 , Error degrees freedom : 476				
Root Mean Squared Error : 0.653				
R- squared: 0.109 , Adjusted R- squared: 0.102				
F-statistic Vs constant model: 14.6 , p- value : 3.09e-11				

giving better results. The final model with all 7 independent features are shown in table 3.2, where N is the total number of samples used.

Linear Regression Model: Health $\sim$ 1+ Stress + Nutrition + Hydration + Sleep +Energy + Soreness + Irritability

Table 3.2: Linear Regression Model with Seven Independent Features , N=481

	Estimate	SE	tstat	pvalue
Intercept	2.4643	0.3989	6.1769	1.41e-09
Stress	-0.0648	0.0586	-1.1048	0.26981
Nutrition	0.0669	0.0669	0.9998	0.31793
Hydration	0.1252	0.0646	1.9369	0.05335
Sleep	0.1270	0.0499	2.5438	0.01128
Energy	0.1054	0.0635	1.6594	0.09769
Soreness	0.0219	0.0372	0.5906	0.55505
Irritability	0.2290	0.0688	3.3275	0.00094
Number of observations : 481 , Error degrees freedom : 473				
Root Mean Squared Error : 0.644				
R- squared: 0.138, Adjusted R- squared: 0.125				
F-statistic Vs constant model: 10.8, p- value : 1.1 e-12				

2) Having high R^2 value is not enough to get the good fit of the model. Backward selection is used to select the significant features. Initially the model is created with all features. Check the significance using t-stat (T- test). If this value is significant its good to keep that feature otherwise remove the non- significant features. The null hypothesis

$H_0 : \beta = 0$ is rejected against the alternative $H_1 : \beta \neq 0$, if b (least square estimator) is too far from zero that is if |t-statistics| $> C$, where C is the constant variable. Check the t-statistics with rough rule of thumb when |t-statistics| $> 2(C = 2)$, the feature is significant. From table 3.2, check the non-significant features. Start removing the features from the lowest value in turn. Stress has t-Stat value of -1.1048. Hence first, remove the stress feature as shown in table 3.3, where N is the number of samples used.

Linear Regression Model: Health $\sim$ 1+ Nutrition + Hydration + Sleep +Energy + Soreness + Irritability

Table 3.3: Linear Regression Model after removing Stress Features , N= 481

	Estimate	SE	tstat	pvalue
Intercept	2.4594	0.3990	6.1638	1.52e-09
Nutrition	0.0628	0.0669	0.9396	0.34788
Hydration	0.1231	0.0646	1.9040	0.05752
Sleep	0.1283	0.0499	2.5714	0.01043
Energy	0.1047	0.0635	1.6474	0.10015
Soreness	0.0172	0.0370	0.4648	0.64225
Irritability	0.1772	0.0504	3.5182	0.00048
Number of observations : 481 , Error degrees freedom : 474				
Root Mean Squared Error : 0.644				
R- squared: 0.136, Adjusted R- squared: 0.125				
F-statistic Vs constant model: 12.4, p- value : 5.2e-13				

Next soreness has least value in non- significant features. Try removing soreness feature from the model and check the significant features again and continue until getting the significant results. Finally the model is significant with three features hydration, sleep and irritability for dependent variable of health. The significant model of the regression of health is shown in table 3.4, where N is the number of samples.

Linear Regression Model:Health $\sim$ 1+ Hydration + Sleep + Irritability

3.1.4 Analysis of Variance(ANOVA)

Analysis of variance is used to analyse the significance of effect of one or more predictor variables on a response variable [17]. This is used to test the quality of fitted model. ANOVA

Table 3.4: Significant Linear Regression Model, N= 481

	Estimate	SE	tstat	pvalue
Intercept	2.7841	0.3479	8.0026	9.34e-15
Hydration	0.1654	0.0572	2.8945	0.003972
Sleep	0.1786	0.0426	4.1960	3.24e-05
Irritability	0.2034	0.0487	4.1735	3.57e-05
Number of observations : 481 , Error degrees freedom : 477				
Root Mean Squared Error : 0.645				
R- squared: 0.128, Adjusted R- squared: 0.123				
F-statistic Vs constant model: 23.4, p- value : 3.93e-14				

is a test for checking the hypothesis that $\beta_i = 0$ or not. When ANOVA test is applied to the regression, it allows the model to be tested for statistical significance. It can be used to test the quality of each betas. It uses F- test to test the significance. F- Test statistics is ratio of between group variability and within group variability. ANOVA determines if the predicted regression line is equivalent to a slope of zero depends on F statistics generated. Using a probability density function for the appropriate F distribution, as defined by the degrees of freedom, a probability of equivalence can be determined. This is called P- value. The level of type- I error acceptable is chosen to be 1% or 5%. If the probability of the given F statistics is less than 1% or 5%, then the assumptions are that the relationship between X and Y does exists and they are statistically significant. If the probability is greater than the considered limit, the null hypothesis cannot be rejected and they are not significant in nature. For the model, Health $\sim$ ANOVA(1+ Stress + Nutrition + Hydration + Sleep +Energy + Soreness + Irritability) coefficients are shown in table 3.5.

The first column shows the features included in the model [18]. The second column SumSq is the sum of squared error for each term except constant. Third column DF is Degrees of freedom. DF is 1 for each feature and for the error term it is n-p, where n is the number of observations and p is the number of coefficients in the model including intercept. Here n = 481 and p is 8 (including intercept) hence n - p = 473. The fourth column MeanSq is the mean squared error for each feature. The fifth column is the F-values for each coefficient. The F value is the ratio of mean squared of each term and mean squared error.

Table 3.5: ANOVA Test with Seven Features

	SumSq	DF	MeanSq	F	pValue
Stress	0.5066	1	0.5066	1.2206	0.26981
Nutrition	0.4149	1	0.4149	0.9996	0.31793
Hydration	1.5572	1	1.5572	3.7517	0.05335
Sleep	2.6858	1	2.6858	6.4707	0.01128
Energy	1.1430	1	1.1430	2.7537	0.09769
Soreness	0.1448	1	0.1448	0.3488	0.55505
Irritability	4.5956	1	4.5956	11.072	0.00095
Error	196.33	473	0.4151		

Each F-statistic has an F distribution. The last column is the p-value for each hypothesis test on the coefficient of the corresponding feature in the linear model. By checking the significance of p- value at 0.05, only irritability and sleep features are significant while others are not which p value is greater than 0.05. The lack of fit test is conducted for the model in table in 3.5 and it is shown in table 3.6.

Table 3.6: Lack of Fit F Test

	SumSq	DF	MeanSq	F	pValue
Total	227.78	480	0.4745	1.2206	0.26981
Model	31.447	7	4.4925	10.823	1.10e-12
Residual	196.33	473	0.4151		
Lack of fit	95.64	209	0.4576	1.1998	0.08092
Pure error	100.69	264	0.3814		

The model has 7 features. Degrees of freedom is set to 7. The DF of residual is n- p, n is the number of observations and p is the number of coefficients. There are n= 481 observations and c= 217 distinct values. The lack of fit degrees of freedom c-8 = 217 -8 = 209 and the pure error degrees of freedom is n-c= 481-217 =264, sum to the error degrees of freedom (residual degrees of freedom) n-8 =481-8 = 473. The corresponding F- statistics in the F-column are for testing the significance of the model. The residual term is separated into two parts: first is the error due to the lack of fit, and the second is the pure error independent of the model obtained from replicated observations. The F- statistics is calculated by the ratio

of mean square error of lack of fit and mean square error of pure error. The corresponding F- statistics in the F-column are for testing, the lack of fit, that is, whether the proposed model is an adequate fit or not. The standard hypothesis test is followed in the procedure of doing lack of fit F test [17].

H_0 : There is no lack of linear fit

H_A: There is lack of linear fit. The decision is based upon the following procedure:

If p-value is smaller than significant level $\alpha = 0.05$, the null hypothesis is rejected. There exists sufficient evidence to conclude there is lack of linear fit.

If p-value is larger than significant level 0.05, the null hypothesis is not rejected and there is not enough evidence to conclude for lack of linear fit. In this example, the F- statistics is 1.1998 and p- value is 0.080919. P-value is larger than significance level $\alpha = 0.05$, fail to reject the null hypothesis. There is not enough evidence at the level to conclude that there is lack of linear fit.

3.1.5 Key Assumptions to Validate Model

Validation of model [15] is really important in regression analysis, which is used to find the type of regression analysis that best fit to the data. Following are the key assumptions in regression analysis:

1. Linearity of the relationship between dependent and independent variable
2. Independence of the errors(no correlation between consecutive errors in time series data)
3. Homoscedasticity (constant variance of the errors)
4. Normality of the error distribution

1) Linearity: Once the model is set with significant features as in table 3.4, check for linearity/non linearity. If the relationship between independent variable (IV) and the dependent variable (DV) is not linear, the results will underestimate the true relationship.

This underestimation carries two risks: increased risk of Type II error (fail to reject the null hypothesis when it is false) for the independent variables. In multiple regression analysis an increased risk of Type I errors (over- estimation that is a rejection of the null- hypothesis) for other independent variables that share variance with that independent variable. The non-linearity can be viewed in a graph of observed versus predicted values or a plot of residuals versus predicted values. The points should be symmetrically distributed around a diagonal line in the former plot or around the horizontal line in the latter plot, with a roughly constant variance. In multiple regression models, non linearity or non-additivity can be viewed by systematic patterns in plots of the residuals versus individual independent variables. Another way of looking non-linearity is to add $(Predictedvalue/Fittedvalue)^2$ as the additional feature. Check this feature is significant or not. This table is shown in 3.7, where N is the number of samples. If this feature is significant add non-linear terms with the existing model. Non-linear terms can be cross multiplication of independent features or it can be higher order terms of independent features.

Linear Regression Model: Health $\sim$ 1+ Hydration + Sleep + Irritability + FittedVarSquared

Table 3.7: Significant Model with $(Fittedvalue)^2$, N= 481

	Estimate	SE	tstat	pvalue
Intercept	3.0605	0.35242	8.6843	6.13e-17
Hydration	1.9800	0.51357	3.8554	0.00013
Sleep	2.0901	0.53935	3.8751	0.00012
Irritability	2.4504	0.63394	3.8654	0.00013
FittedVarSquared	-0.9873	0.27774	-3.5548	0.00042
Number of observations : 481 , Error degrees freedom : 476				
Root Mean Squared Error : 0.637				
R- squared: 0.151, Adjusted R- squared: 0.144				
F-statistic Vs. constant model: 21.1, p- value : 4.79e-16				

tStat of (Fitted value)^2 is significant hence add non- linear term with the model. Number of non-linear term added with significant model can be done by using step function in Matlab by defining number of terms added. This is shown in table 3.8.

Linear Regression Model: Health ∼ 1+ Hydration * Irritability + Sleep * Irritability.

Table 3.8: Model with Non Linear Terms

	Estimate	SE	tstat	pvalue
Intercept	-5.5895	1.9307	-2.8950	0.00397
Hydration	1.2825	0.3647	3.5161	0.00048
Sleep	0.7291	0.2811	2.5933	0.00980
Hydration:Irritability	-0.2132	0.0682	-3.1249	0.00189
Sleep: Irritability	-0.1141	0.0540	-2.1155	0.03491
Number of observations : 479 , Error degrees freedom : 473				
Root Mean Squared Error : 0.586				
R- squared: 0.162, Adjusted R- squared: 0.154				
F-statistic Vs. constant model: 18.3, p- value : 1.18e-16				

2) Independence: Each observations should be independent with each other. Y_i should be independent of one another or ε_i should be independent of one another. The graphs are shown for significant model with independent features hydration, sleep and irritability for the prediction of health.

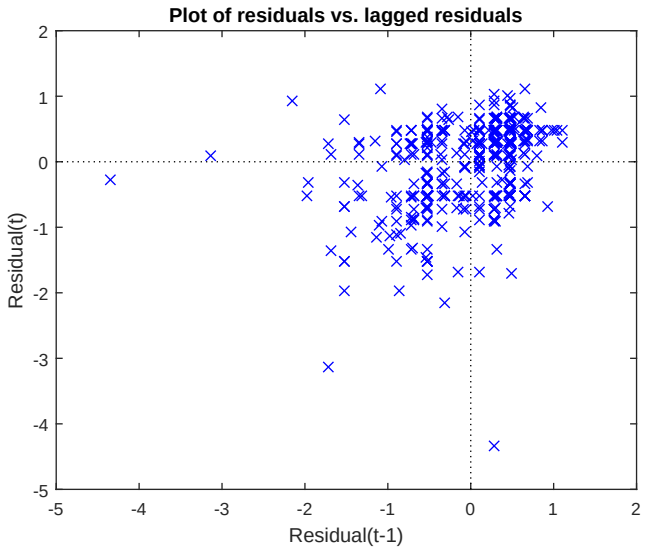


Figure 3.2: Plot of Residual versus Lagged Residual shows there is a possible correlation

From the figure of residuals versus lagged residuals shown in 3.2,we can visualize that there is a possible correlation. Especially in the lower left corner and upper right corner.

If there is serial correlation, there is a room for improvement in the model. Ideally, most of the autocorrelation bounds should fall within the 95% confidence levels around zero. For example, for the sample size of 50, it should fall +/- 0.3. For the dataset taken for 481 samples,autocorrelations should be fall in the range +/- 0.1 around zero. Autocorrelations of the significant model is shown in figure 3.3.

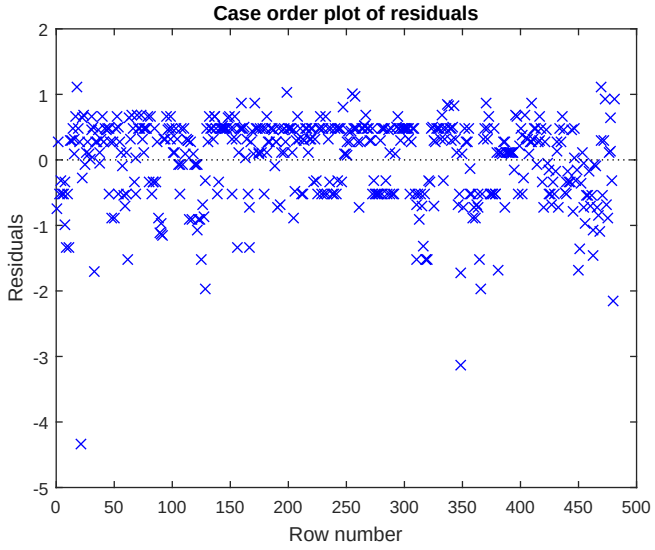


Figure 3.3: Case order plot of residuals - Autocorrelation should fall in the range +/- 0.1 which indicates there is room for improvement in the model

3) Check normality of error using Jarque- Bera test. The Jarque-Bera test is used to check hypothesis about the fact that a given X is a sample of a normal random variable with unknown mean and dispersion. This test is based on the fact that skewness and kurtosis of normal distribution equal zero. In other words, we can say this is goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution.

The Jarque-Bera test for normality is testing the null hypothesis stated in [15]:

H_0 : Normal distribution,skewness is zero and excess kurtosis is zero, against the alternative hypothesis.

H_1 : Non-normal distribution. The Jarque-Bera test statistic is shown in equation 3.12,

$$R^2 = Jb = n\left[\frac{S^2}{6} + \frac{EK^2}{24}\right] \tag{3.12}$$

Where S is the sample skewness, K is the sample kurtosis and E is the excess kurtosis(K-3). For the significant model, the results of Jarque-Bera test is, h = 1 ; p = 0 ; Jbstat = 70.9048 ; critval = 6.2740.

Jbstat is the Jarque-Bera statistics value and critval is the critical value. Jbstat is greater than critval. Hence the normality is rejected.

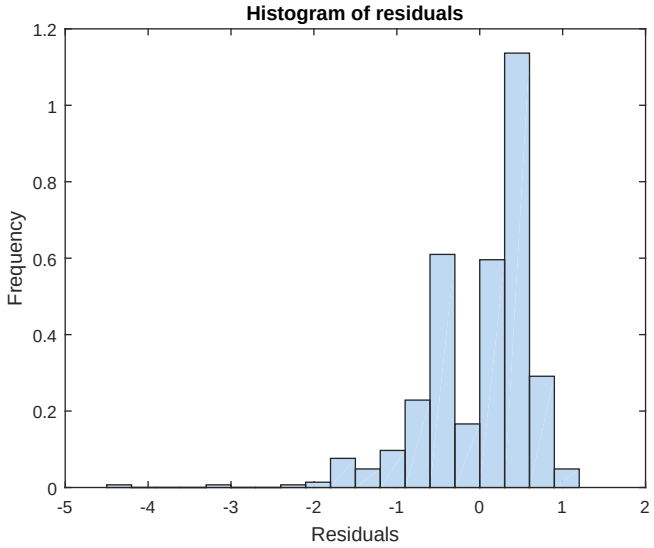


Figure 3.4: Histogram of Residuals for N=481, shows there is skewness in the left hand side distributions

Histogram of residuals in figure 3.4, shows there is skewness in the left hand side distributions. The figure 3.5 , shows deviation from normality and skewness on the left hand tail of residual distributions. Residuals are not equally distributed around their median that is shown in figure 3.6, failing to satisfy normality.

4) Homoscedasticity: If the data points are homoscedastic, the variance of Xi are equal. If the data points are heteroscedastic, the variance of Xi is not equal. If the variance increases with fitted value, suggests that exists possible heteroscedasticity as shown in figure 3.7.

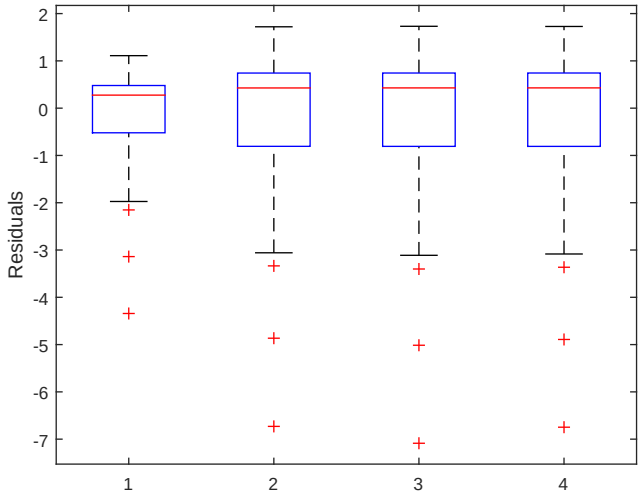


Figure 3.5: Box plot of Residuals, shows there is skewness in the left hand side distributions

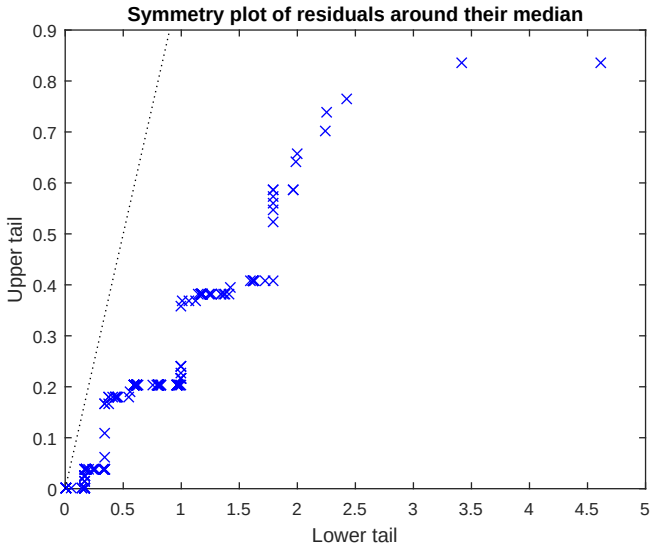


Figure 3.6: Symmetry Plot of Residuals- Residuals are not equally distributed around their median failing to satisfy normality

3.2 Chapter Summary

This chapter summarizes the ordinary least square method. The analysis of the data set is started with linear regression. Because the linear relationships are non trivial that can be imagined easily and easier to work also. When the data does not fit with linear type

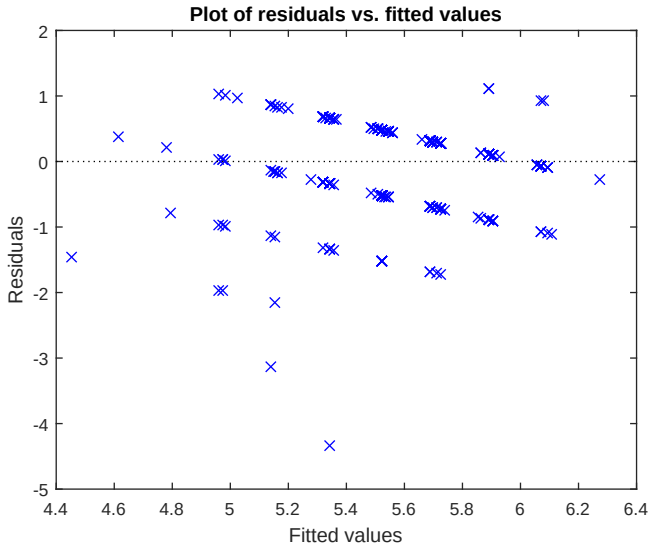


Figure 3.7: Residual versus fitted value shows variance increases with fitted value that suggests that exists possible heteroscedasticity

we can easily modify into non-linear regression by adding non linear terms with that. The residuals and R^2 are discussed and how to choose the variables also discussed. Because having significant features in the model is really important. The key assumptions of regression is also tested. Several graphs are plotted to visualize the key assumptions like normality,linearity and heteroscedasticity. ANOVA test is conducted to find the linear fit of the model and to check the significant features.

Chapter 4

Time Series Regression

This chapter introduces time series regression, AR model and explaining about the AR model with exogenous (ARX) inputs. ARX model has an independent variable with the past values of response variable as well lagged values of other features that has significant effect in the model. Partial auto correlation is explained to choose the order of auto regression model which makes the model simple by adding only particular time series lags.

A time series is a sequence of data points, consists of successive measurements or observations of quantifiable variables prepared over a certain time interval. Time series can be represented as a function of specific time [19], $X_T, T = 1, 2 \cdots t$. When time series has a defined trend, it can be represented as a function of previous time values. When the feature to predict is Y at time t, the form of the predicted model is given by (4.1)

$$Y_t = Y_{t-1} + Y_{t-2} + \cdots + Y_{t-P} + \varepsilon_t \tag{4.1}$$

There are essentially two types of variables employed in regression analysis: endogenous and exogenous. An endogenous variable (Y) is one whose values are explained by the model; it is often referred as the dependent variable. Exogenous variables (X) are those whose values are determined outside the confines of the present model; they are often called as explanatory or independent variable. There are two basic types of time series regression models: non-lagged and lagged. A non-lagged model captures the relationship of variable over time when both the endogenous and exogenous variables are observed at the same point in time. For example, present day endogenous health feature is dependent on present day

exogenous features. Lagged time regression model captures the relationship of present day health is dependent on the past level of exogenous and/or endogenous features. (Past values of health and/or other features).

Time series can be classified into two classes, namely: Univariate and multivariate time regression. When time regression involves sequences of measurements with single variable collected over time is called univariate time regression. Time series with more than one variable is called a multivariate time regression. Most of the data analysis is multivariate time series. [14]. Changes in one element in the observation vector of one variable imply corresponding changes in other variables of that model.

4.1 Auto Regressive Models

Auto regressive model is used to forecast the variable of interest using a linear combination [10], of past values of the variable. The term auto regression indicates that it is a regression of the variable against it. Let consider the variable health (Y) is measured as a function of time series. An autoregressive model is when a value from a time series is regressed on previous values from that same time series. For example Y_t on Y_{t-1} that is today's health is dependent on previous day health feature and represented by equation (4.2)

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t \tag{4.2}$$

.

The previous values of the response variable become as a predictor for this model. The order of an auto regression is the number of immediately preceding values in the series that are used to predict the value at present time. Hence this is a first order auto regression represented as AR(1). The prediction of today's health feature dependent on past two days health feature, then the autoregressive model is given by equation (4.3)

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \varepsilon_t \tag{4.3}$$

This model in (4.3) is called as second order auto regression represented as AR (2). Generally, P^{th} order auto regression represented as AR(P) in which value of time series at

any time t is a function of the values at times $t-1, t-2 \cdots t-P$. The aim of this paper is to predict for future days. The prediction of next day health feature (Y) that is one day ahead in time is given by (4.4),

$$Y_{t+1} = \beta_0 + \beta_1 Y_t + \varepsilon_t \tag{4.4}$$

Where $Y_{(t+1)}$ is one day ahead in time, $Y_t$is the present day feature, $Y_{(t-1)}$ and $Y_{(t-2)}$ are past two days feature, β_0 is the constant, β_1 and β_2 are coefficients and ε_t is the error.

4.2 Autocorrelation and Partial Autocorrelation

The coefficient of correlation between two values in a time series is called the autocorrelation function (ACF) [10]. The ACF for a time series Y_t is given by, $Corr(Y_t, Y_{t-P})$. The value of P is the time gap being considered and is called lag. A lag 1 autocorrelation means P=1, is the correlation between values that are one time period apart. Lag P autocorrelation means is the correlation between values that are P time periods apart. ACF is used to measure the linear relationship between an observation at time t, and observations of previous values. For AR(P) model, it is not necessary to add all times from $t-1, t-2 \cdots t-P$. It is better to filter out the linear influence lags that lies in between. This is called partial autocorrelation(PACF). This is useful in identifying the order of an auto regression model and which makes model simpler by adding only particular time series lags.

Graphical approach is used to assess the lag of an autoregressive model by looking at the PACF values versus the lag. If the graph has large ACF values and a non-random pattern, then the values are serially correlated. In a PACF versus lag graph it usually appear in a random manner. PACF with larger values lag may be added as a choice if it makes sense. Figure 4.1 shows the partial autocorrelation for the prediction of health one day ahead that is autocorrealtion of one day ahead health feature with other health feature time lags. The PACF cuts off mostly after the second lags. Hence it represents the AR(2) model.

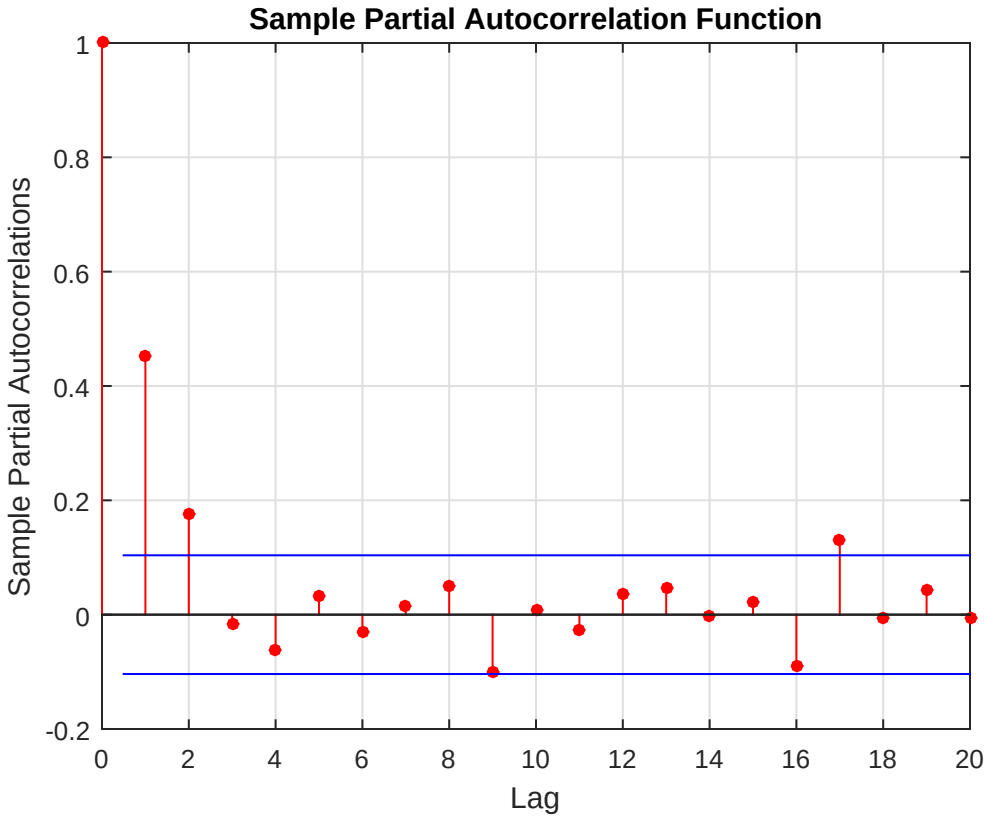


Figure 4.1: Partialautocorrelation to access time series lags which shows significance at +/- 0.1. This is a second order model. Adding larger lag 17th, depends on the model only if it makes sense.

4.3 Multivariable Autoregressive with Exogenous Inputs Model (MVARX)

In the autoregressive model, the predictors consisting past values of the response variable and it of the error term. Autoregressive model with exogenous inputs means the model consists of past values of the response variable as well lagged values of other features and error term in [20]. The ARX model one day ahead prediction of variable Y represented by the equation (4.5),

$$Y_{t+1} = \beta_0 + \beta_1 Y_t + \beta_2 Y_{t-1} + \cdots + \beta_p Y_{t-p} + \beta_{11} X_t + \beta_{12} X_{t-1} + \cdots + \beta_{1p} X_{t-p} + \varepsilon_t \qquad (4.5)$$

Where $Y_{(t+1)}$ is the one day ahead prediction, $Y_t, Y_{t-1} \cdots Y_{t-P}$ are past values of the same time series, $X_t, X_{(t-1)} \cdots X_{(t-P)}$ are the externally determined features that influences the series of interest, p is the number of predictors, β_0 is the intercept,β are the coefficients and ε_t is the error term.

This paper aims to create the model with past values of the same series as well, including significant features by testing t-statistics and performing add/removal of features. For one day ahead prediction of health feature (Y_{t+1}) the model is designed by ARX (1) is given in equation (4.6),

$$Y_{(t+1)i} = \beta_0 + \beta_1 Y_{(t)i} + \beta_2 X_{(1t)i} + \cdots + \beta_p X_{(pt)i} + \varepsilon_{(t)i} \tag{4.6}$$

Where $i - 1, 2, 3 \cdots N$ (History of data), X are the exogenous variable and ε is the error term p is the number of predictors, β_0 is the intercept, β are the coefficients and t is the time.

4.4 Chapter Summary

This chapter describes regarding the time series regression, AR model and choosing the AR model with external inputs that has a serious effect on the model. Partial auto correlation helps in finding the lags which makes the model easier instead of using all lags, the graph tells us which lags will give significant results.

Chapter 5

Logistics Regression

The most popular regression method is a linear regression using the method of least squares also referred to as conventional regression analysis (CRA). It is applicable if the dependent variable is continuous, independent and identically distributed only. In cases where the dependent variable is categorical, conventional regression analysis is not appropriate. Logistics regression [17], is a type of predictive analysis that is used to estimate the relationship between one or more predictor variables and a single response variable and is extensively used in numerous disciplines such as medical, bioinformatics and social science fields [21]. The difference between logistics and linear regression is that logistics regression response variable is binary in nature. The relationship between dependent and independent variable need not to be linear in logistics regression. It has two possible outcomes such as sick/healthy or injured/not injured in the case of binary logistics regression, more than two levels of output such as sick/ not too healthy/ healthy is called as multinomial logistics regression. Logistics regression is suitable when the response variable is not a real number. Logistics regression is a method of developing prediction probabilities.

The binary logit model can be applied to the tables in which the response variable is dichotomous, it is also possible to use the equivalent binomial logit model; the binomial logit model is based on the frequency counts of success and failure for each combination of explanatory variable values. When it is applicable, the binomial logit model offers several advantages, including efficient computation, a test of the fit of the model based on its residual

deviance and better behaved diagnostics.

Odds and Odds Ratio

Odds are the ratio of probability of an event will occur divided by the probability of it will not occur stated in [22]. The odds equation is given in (5.2),

$$\text{Odds} = \frac{P(Success)}{P(Failure)} = \frac{P}{(1-P)} \tag{5.1}$$

Odds have values greater than zero. If odds value is larger than 1 means that success will occur more likely than failure. For example, odds= 3 means, its possible to have 3 success events for every one failure. If odds= 1/3, reverse process will take place. Odds ratio, is the ratio of two odds given in equation (5.2),

$$\text{Odds Ratio} = \frac{\frac{P_1}{1-P_1}}{\frac{P_2}{1-P_2}} \tag{5.2}$$

Where P_1 and P_2 refers to the probability in group 1 and 2. If the odds ratio is greater than 1, it refers that the odds of the outcome in group 1 is larger than in group 2. The number of successes in group 1 is larger than group 2. If the odds ratio less than value one, the reverse will occur. When odds ratio is equal to one, odds of group 1 and group 2 are equal.

5.1 Simple Logistics Regression Model

The simple logistics regression model [17], usually called as the logit model. It has relationship between one predictor variable X and a binary variable Y. The logistics equation can be used to examine how the probability of an event changes as the predictor variable changes.

The simple logistic regression equation is given in (5.3),

$$\ln\left[\frac{P}{1-P}\right] = \beta_0 + \beta_1 X \tag{5.3}$$

Solved for probability (P) is given in (5.4),

$$P = \frac{exp(\beta_0 + \beta_1 X)}{1 + exp(\beta_0 + \beta_1 X)} \tag{5.4}$$

Here P is the probability of the response variable, β_1 is the regression coefficient and β_0 is the intercept. β_1 and β_0 are calculated using maximum likelihood equation using iterative technique.

5.2 Binary Complex Logistics Regression Model

The binary logistics regression model is a type of regression analysis where the dependent variable is dummy variable. The logistics regression model use logit transform and it is represented as in (5.5),

$$\ln\left[\frac{P_i}{1-P_i}\right] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} \tag{5.5}$$

Where ($P_i = P(Y_i = 1) = 1 - P(Y_i = 0)$), $P(Y_i = 1)$ and $P(Y_i = 0)$ are probability of healthy and sick of an observation i respectively. β_0 is the log-odds when all are 0. β_j is the increase in log-odds when is increased by one unit $j = 1 \cdots k$, X are the independent variables, k is the number of predictors.

5.2.1 Selection of Variables

The procedures for choosing variables are basically the same as for multiple linear regression. There are different objective method: forward selection, backward elimination, stepwise selection or it is possible to use a careful examination of the data and understanding of the biology to subjectively choose the best variables. The main difference from OLS is that instead of using the change of R^2 to measure the difference in fit between an equation with or without a particular variable, change in likelihood is used in logistics regression.

5.3 Model Fit of Logistics Regression

5.3.1 Deviance

With logistic regression, instead of R^2 as the statistic for overall fit of the model, deviance is used instead [23]. Deviance of a model M_1 is twice the difference between the log likelihood of that model and the saturated model, M_S. The saturated model is the model with the

maximum number of parameters that can be estimated. For example, if there are "n" observations $y_i, i = 1, 2 \cdots n$, with potentially different values for $X_i^T \beta$, then saturated model is defined with "n" parameters. Let L(b,y) denote the maximum value of the likelihood function for a model. Then the deviance of model M_1 is given in (5.6),

$$\text{Deviance} = -2(\text{Log}L(b_1, y) - \text{Log}L(b_S - y)) \tag{5.6}$$

Where $L(b_1, y)$ is the likelihood of the model, $L(b_S - y)$ is the likelihood of the saturated model, b_1 is the estimated parameters of the model M_1, b_S is the estimated parameters of the saturated model M_S.

The bigger the difference (or "deviance") of the observed values from the expected values, the poorer the fit of the model. Hence small deviance is expected. When we add more variables to the equation the deviance should get smaller, indicating an improvement in fit. The deviance has a chi-square distribution with n - p degrees of freedom, where n is the number of parameters in the saturated model and p is the number of parameters in model M_1. If M_1 and M_2 are two different generalized linear models, then the fit of these can be assessed by comparing the deviances D_1 and D_2 of these models. The difference of the deviances is in (5.7),

$$D = D_2 - D_1 = -2(\text{Log}L(b_2, y) - \text{Log}L(b_1 - y)) \tag{5.7}$$

This difference has a chi-square distribution with degrees of freedom V equal to the number of parameters that are estimated in one model but fixed (typically at 0) in the other. That is, it is equal to the difference in the number of parameters estimated in M_1 and M_2. The p-value for this test is obtained by equation (5.8),

$$\text{P-value} = 1 - chi2cdf(D, V) \tag{5.8}$$

Where $D = D2 - D1$ is the difference of deviances,V is the degrees of freedom.

5.3.2 Maximum Likelihood

Instead of finding the best fitting line by minimizing the squared residuals like in OLS regression, Maximum Likelihood (ML) approach is used in logistics regression. ML is a way

of finding the smallest possible deviance between the observed and predicted values that is kind of like finding the best fitting line using calculus. With ML, the computer uses different "iterations" in which it tries different solutions until it gets the smallest possible deviance or best fit. Once it has found the best solution, it provides a final value for the deviance, which is usually referred as "negative two log likelihood" [21].

The likelihood ratio test, G: A chi-square difference test using the "null" or constant-only model

Instead of using the deviance (-2LL) to judge the overall fit of a model, however, another statistic is usually used that compares the fit of model with and without the predictor(s). This is similar to the change in R^2 when another variable has been added to the equation. Decrease in the deviance is expected, because the degree of error in prediction decreases when another variable is added. To do this, the deviance with just the intercept is compared to the deviance when the new predictor or predictors have been added. The difference between these two deviance values is often referred to as G for goodness of fit.

G=χ^2=D(for the model without the variable)-D(for the model with variable)

or rewrite as shown in (5.9) and (5.10),

$$G = \chi^2 = D_{Null} - D_k \tag{5.9}$$

$$G = \chi^2 = -2LL_{Null} - 2LL_k \tag{5.10}$$

Where D_{Null} is the deviance for the constant only model an D_k is the deviance for the model containing k number of predictors. An equivalent formula is given by (5.11),

$$G = \chi^2 = -2ln\frac{LL_{Null}}{2LL_k} \tag{5.11}$$

Where the ratio of the ML values is taken before taking the log and multiplying by -2. This gives rise to the term "likelihood ratio test" to describe G. The likelihood ratio is the one of the most commonly used fit tests.

5.3.3 Wald Test

A Wald test [21], is used to test the statistical significance of each co- efficient β of the model. Wald test statistics is given by equation (5.12),

$$Z = \frac{\text{Co-efficients}}{\text{Standard error}} \tag{5.12}$$

If this Z value is squared, which is yielding (Z^2) and it is called as Wald statistics with a chi-square distribution. In the Wald test, the null hypothesis is rejected if $|Z| > c$ where c is a pre-determined critical value. As a rough rule of thumb when $|Z| > 2$, the variable is considered as significant in nature.

Advantage

Easy to calculate and confidence interval has a closed form.

Disadvantage

The Wald statistics is currently the main logistics regression metric of variable coefficient significance calculated by statistical packages. The reliability of Wald statistics is questionable. If the large coefficient is produced, the standard error of the coefficient can be inflated. This will produce undersized Wald statistic, which give the result of significant coefficient is not significant. There is one more error, when sample size is too small. The Wald statistic is biased for small sample sizes.

5.3.4 Pearson Residuals and Deviance Residuals

Pearson and deviance residuals [21], are useful in identifying observations that are not explained well by the model. Pearson residuals are components of the Pearson chi-square statistic and deviance residuals are components of the deviance. The Pearson residual for the cell j is given by equation (5.13),

$$r_j = \frac{(X_j - n\hat{\pi}_j)}{\sqrt{(n\hat{\pi}_j)}} \tag{5.13}$$

Pearson residual compares the observed with the expected counts. The sign (positive or negative) indicates whether the observed frequency in cell j is higher or lower than the value fitted under the model, and the magnitude indicates the degree of departure. When the data do not fit a model, examination of the Pearson residuals often helps to diagnose where the model has failed.

The Pearson chi-square statistic is the sum of squares of the Pearson residuals. The Pearson goodness of fit statistics is given by equation (5.14),

$$\chi^2 = \sum_{j=1}^{k} \frac{(X_j - n\hat{\pi}_j)^2}{n\hat{\pi}_j} \tag{5.14}$$

The deviance residual for the jth observation is given by equation (5.15),

$$d_j = \sqrt{|2X_j log\frac{X_j}{n\hat{\pi}_j}|}.sign(X_j - n\hat{\pi}_j) \tag{5.15}$$

The sign function can take three values as shown in equation (5.16) to (5.18)

$$sign(X_j - n\hat{\pi}_j) = -1 \; if \; (X_j - n\hat{\pi}_j) < 0 \tag{5.16}$$

$$sign(X_j - n\hat{\pi}_j) = 0 \; if \; (X_j - n\hat{\pi}_j) = 0 \tag{5.17}$$

$$sign(X_j - n\hat{\pi}_j) = 1 \; if \; (X_j - n\hat{\pi}_j) > 0 \tag{5.18}$$

The deviance statistics is given by equation (5.19),

$$G^2 = \sum_{j=1}^{k} 2X_j log\frac{X_j}{n\hat{\pi}_j} \tag{5.19}$$

Where X_j is the observed count in cell j, $n\hat{\pi}_j$ is the expected count in cell j under the assumption that null hypothesis is true that is the assumed model is a good one. $\hat{\pi}_j$ is the estimated (fitted) cell proportion j under the null hypothesis. χ^2 and G^2 both measure how closely the model fits the observed data.

5.4 Stepwise Regression

Stepwise regression [17], is a semi-automated process of building a model by successively adding or removing variables based on the t-statistics of their estimated coefficients. It gives information at our fingertips than the multiple regression. It is useful in selecting significant variables when the model has more number of independent variables by adding/removing the variables. If it is improperly used, it may converge a poor model. In the multiple regression procedure in most statistical software packages, we can choose the stepwise variable selection with specifying the method as "Forward" or "Backward", and also define the threshold values for enter and to get removed from the model. Stepwise selection allows to begin the model with no variables and proceed adding one variable at a time which is called forward selection. The model starts with all variables and proceed removing one variable at a time, which is called backward selection.

Method of stepwise regression used in MATLAB [24], is bidirectional elimination, it performs both forward and backward selection to determine the final model.At each step, the function searches for terms to add or remove from the model based on the value of cut-off thresholds. This model uses the two cut-off thresholds named as PEnter and PRemove. The model adds one variable at a time, if the p-value of F or chi-squared statistic of this variable is less than the threshold usually 0.05(PEnter) or for specified value it allows the variable to enter into the model. Sometimes adding one variable makes other variable non significant. Hence it performs a backward selection. If the p-value of F or chi-squared statistic is larger than PRemove (usually it is set to the value larger than PEnter to avoid the infinite loop) it removes the term from the model.

5.5 Chapter Summary

This chapter briefly explains about the simple and complex logistics regression model. Binomial means observation variable has only two classes (sick/healthy) and multinomial logistics regression is used to design the model when there are more than two classes in the obser-

vation variable. Stepwise regression method is discussed to choose the significant variables from the model. This is a semi automated process.Several tests are conducted to check the fit of a logistics model using likelihood, Wald test and Deviance. The plot of Pearson and deviance residuals gives the idea of outliers of designed model which help to improve the model by removing the outliers.

Chapter 6

Simulations and Results

6.1 Dataset

Sports-1 domain data set is taken for the simulations. Among all data domains this has the large data set. Most of the athletes have a good history of data. This domain has 208 athletes. 33 athletes reported more than 450 days. 49 athletes are reported more than 300 days. 8 athletes reported more than 250 days. More specifically, sports-1 domain has 50994 days of samples. Each day has 8 features. These are health, stress, nutrition, hydration, sleep, energy, soreness and irritability.

In time series regression analysis, the user is considered with the following samples: The data set is divided into two sets which is named as training data set and testing data set. The training data set has 70% of data and its past history of the data. Hence it has 394 days of samples. Testing dataset is used to validate the model and this is 30% of a data set which is a recent history of data. This has 168 days of sample.

Logistics regression analysis used two datasets like in time series regression named as training data set and testing data set. As our objective is the prediction of illness the user is chosen in such a way having more sick data. Here sick data refers , his/her health feature value is less than 3.5. The training data set has 381 days of sample and testing data set has 163 days of sample. In the following sections, "N" represents the total number of days and all the simulations are done using MATLAB.

6.2 Histogram

A histogram is a graphical method for displaying the shape of a distribution. It is useful when there are a large number of observations in the data set. Histograms are useful data summaries that convey the following information:

1) The general shape of the frequency distribution(Normal, chi-square)

2) The symmetry of the distribution(Skewed or un-skewed)

3) Modality- Unimodal, Bimodal, Multimodal.

The histogram is plotted for all athletes that is for 50994 days of samples and for a single user which has 374 days of samples. The histogram plotted for two features health and soreness. Features range from "1 to 7". In health feature "7" considered as well in health and score "1" considered as sickness. X-axis of the histogram has a range of health feature and Y-axis indicates the frequency of occurrence. Histogram of all 206 athletes for health feature with 7 bins is shown in figure 6.1 and it makes clear that most of the good scores

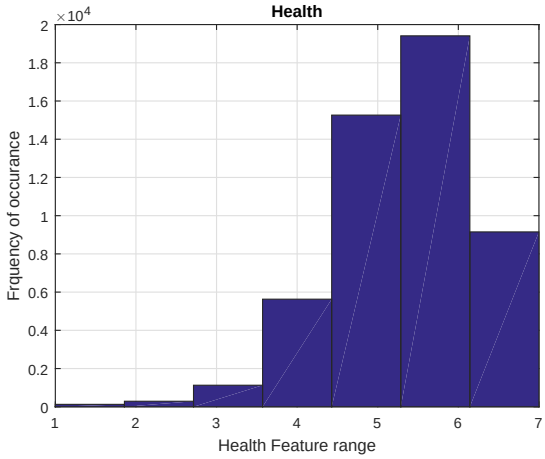


Figure 6.1: Histogram of health feature of all athletes(N= 50994) skewed left hand side of the distribution and has high score "6" which reveals good in health condition repeated most of the times in his/her history of data

are on the right hand side of the histogram.

Histogram of health feature of one athlete is shown in figure 6.2, from this it is visible that

lower scores of health feature are in the left hand side of the distribution. More specifically, the score "6" has around 200 times of occurrence in his/her history of data.

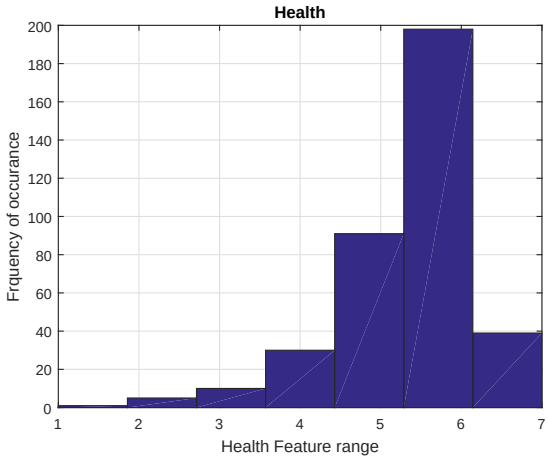


Figure 6.2: Histogram of health feature of one athlete (N= 374) skewed left hand side of the distribution and has high score "6" with more frequency of occurrence which reveals good in health condition most of the times in his/her history of data

The histogram for the feature soreness for all athletes is shown in figure 6.3, shows that soreness feature has more entry on the score "5" where "1" refers for very very sore and score "7" indicates no soreness. The histogram for the feature soreness for one athlete (N=374)

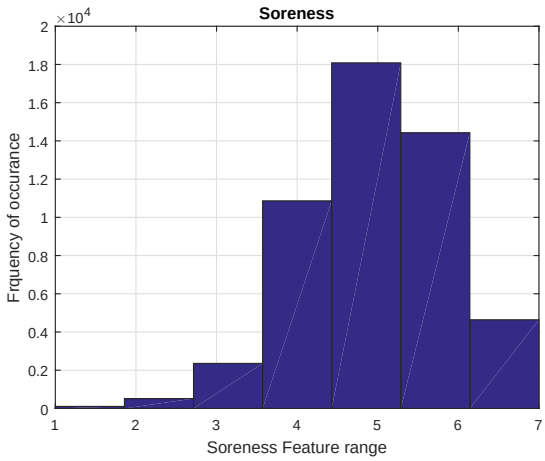


Figure 6.3: Histogram of Soreness feature of all athletes (N=50994) shows skewed on the left hand side of distribution and has more entry of score "5" which indicates the athlete was not that sore in his/her history of data

is shown in figure 6.4, shows that soreness feature histogram mostly looks like a normal distribution.

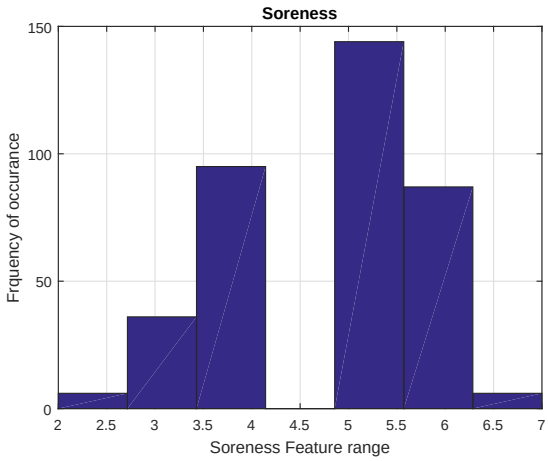


Figure 6.4: Histogram of soreness feature of one athlete (N=374) shows that it follows a normal distribution and this athlete has more entry of score "5"

6.3 Correlation Matrix

Correlation matrix is shown in figure 6.5. A correlation matrix is used to investigate the dependence between multiple variables at the same time. Result containing the correlation coefficients between each variable and the others. It shows the relationship between one feature with other features. Table 6.1 explaining how well the features are correlated with each other.

6.4 Linear Regression Results

Each domain has 8 features and thoughts of the day. By using linear regression one of the feature is predicted when the other independent features are fixed. For example, health is predicted using linear regression function where the other features are inputs of the system as shown in figure 6.6. Sensitivity of the predicted variable with respect to the other variables is also shown. Histogram of residues is calculated to find the error of predicted and actual

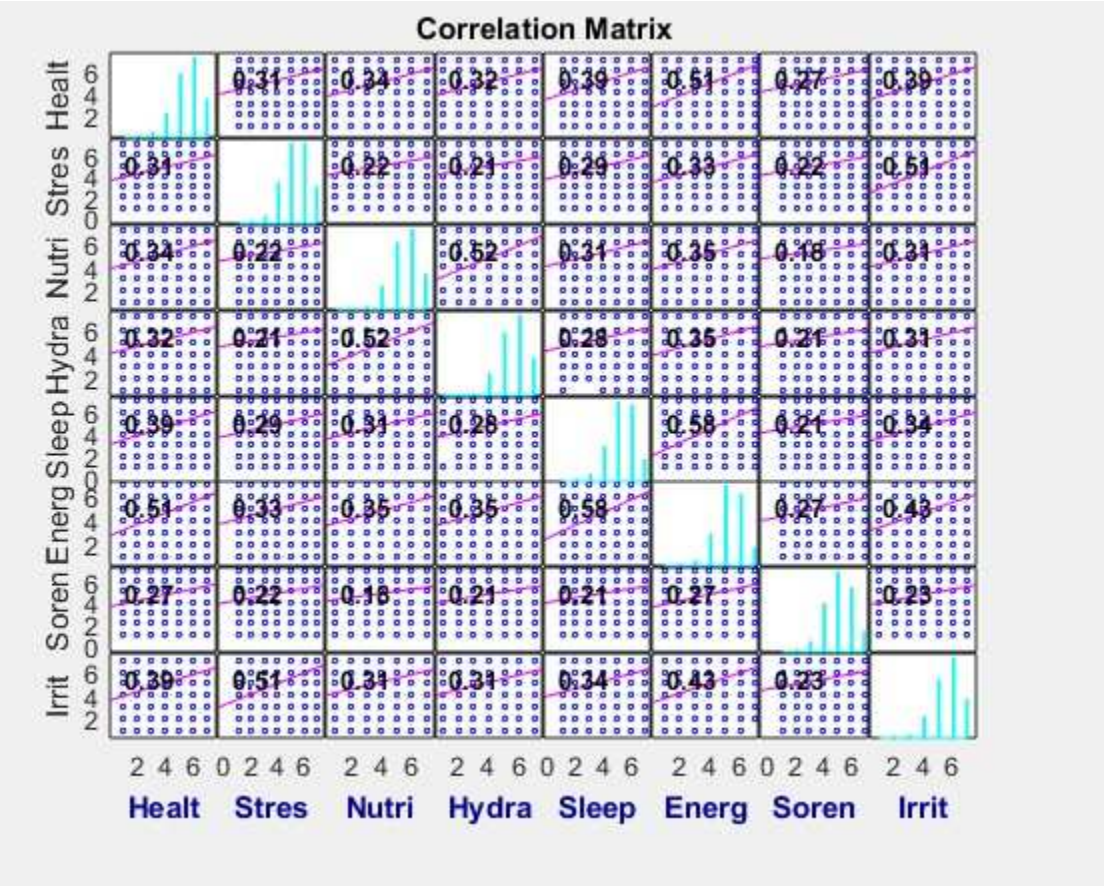


Figure 6.5: Correlation Strength of All Athletes Data in Sports-1 Domain

Table 6.1: Correlation Strength Between Eight Features

Features	Mostly correlated	Less correlated
Health	Energy	Soreness
Stress	Irritability	Hydration, Nutrition and Soreness
Nutrition	Hydration	Soreness
Hydration	Nutrition	Soreness and Stress
Sleep	Energy	Soreness
Energy	Sleep and Health	Soreness
Soreness	-	All
Irritability	Stress	Soreness

value and this is shown in figure 6.7. On X axis range of the error is taken and in Y axis probability of error is taken.

The model is designed with response variable health and other features are considered as

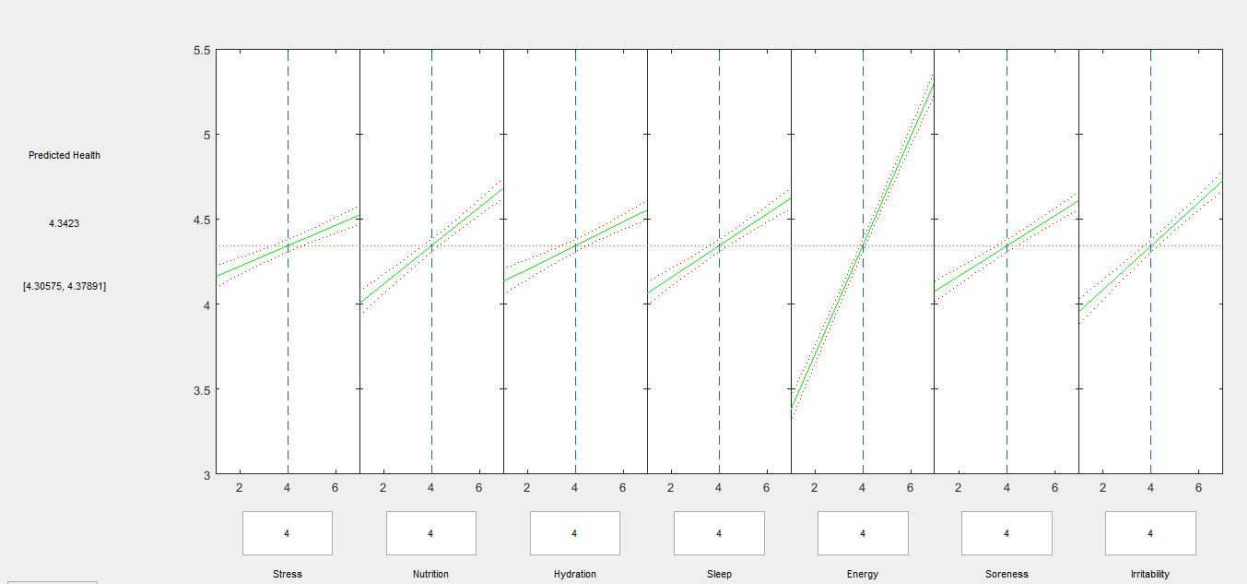


Figure 6.6: Slice Plot for all athletes with the response variable of health with the sensitivity of +/- 0.5

independent variables which is shown in table 6.2, where N is the number of samples.

Linear Regression Model: Health $\sim$ 1+ Stress + Nutrition + Hydration + Sleep +Energy + Soreness + Irritability

Table 6.2: Linear Regression Model of All Athletes Data, N=50994

	Estimate	SE	tstat	pvalue
Intercept	0.8441	0.0318	26.479	1.84e-153
Stress	0.0602	0.0040	14.939	2.36e-50
Nutrition	0.1128	0.0046	24.304	9.68e-130
Hydration	0.0700	0.0048	14.698	8.38e-49
Sleep	0.0934	0.0045	20.774	1.86e-95
Energy	0.3205	0.0052	61.519	0
Soreness	0.0893	0.0037	24.174	2.19e-128
Irritability	0.1283	0.0048	26.906	2.40e-158
Number of observations: 50994, Error degrees of freedom: 50986				
Root Mean Squared Error: 0.846				
R-squared: 0.335, Adjusted R-Squared 0.335				
F-statistic vs. constant model: 3.67e+03, p-value = 0				

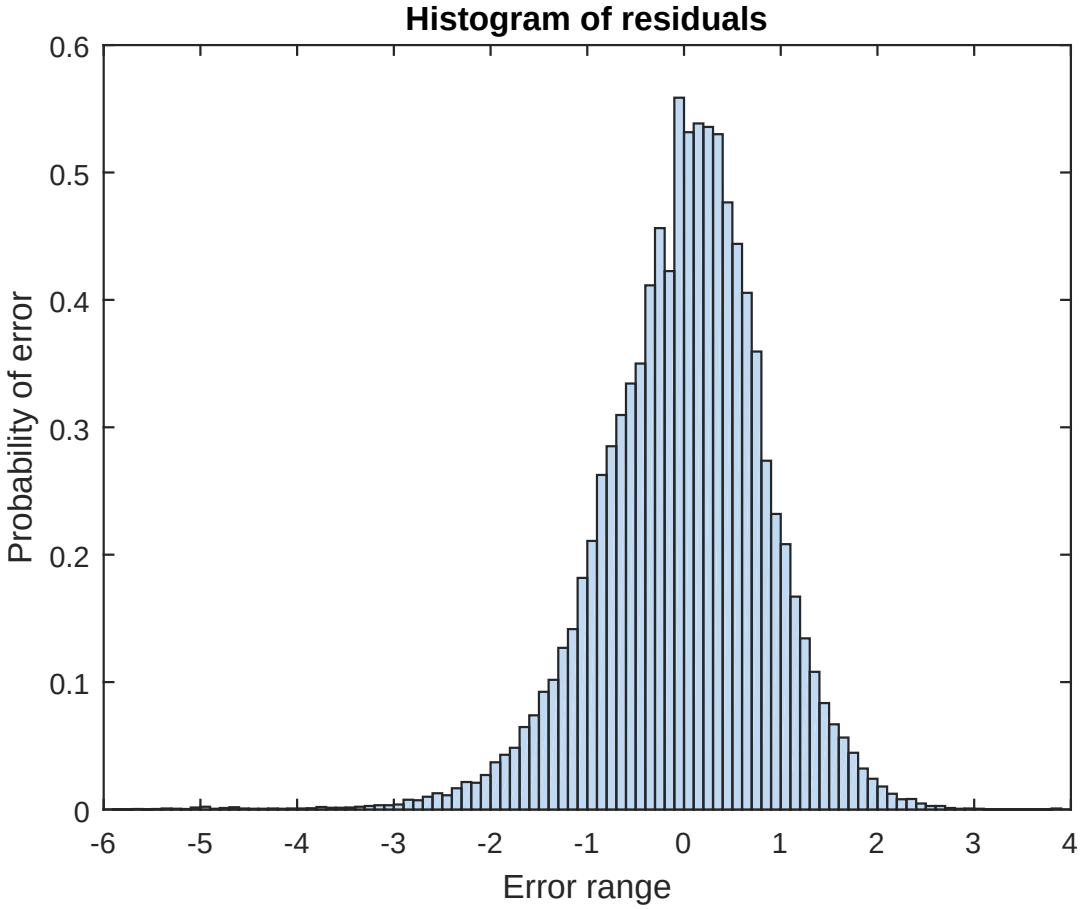


Figure 6.7: Histogram of Residuals for All Athletes Data of Sports-1 Domain

6.4.1 Outlier Detection

Detecting the outliers and removing them from the data set, then obtaining the linear regression for the remaining data set will improve the R- squared value. For the histogram of residuals shown in figure 6.7,the mean value is 1.5123e-12 and standard deviation is 0.8461. The rejection rule for this residual can be chosen as below,

|Residual Error| $> \alpha \times$ Standard deviation.

Where α depends on the level of confidence. For example, 99.99% confidence level is equivalent to α=3.6. Now the obtained histogram of residual is shown in figure 6.8. The value of $R^2 = 0.369$ which is better than before removing the outliers.

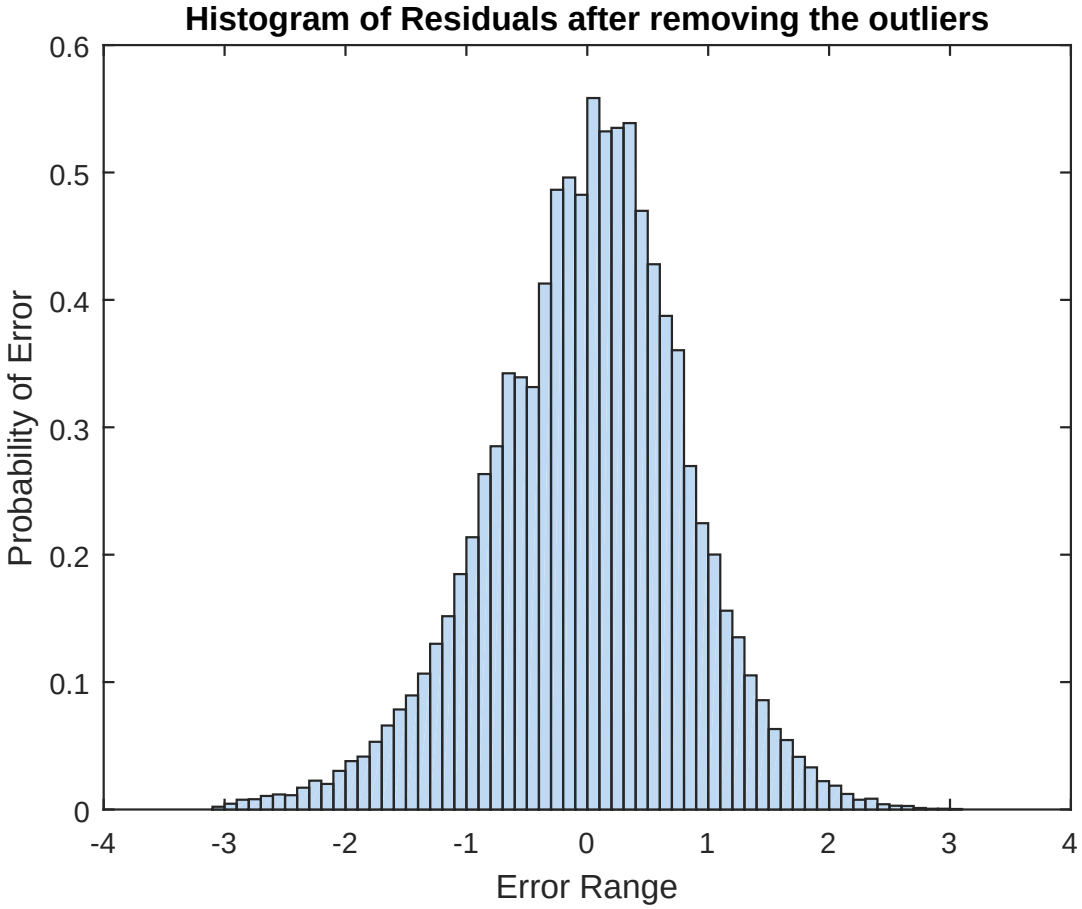


Figure 6.8: Histogram of Residuals after removing the outliers with α=3.6

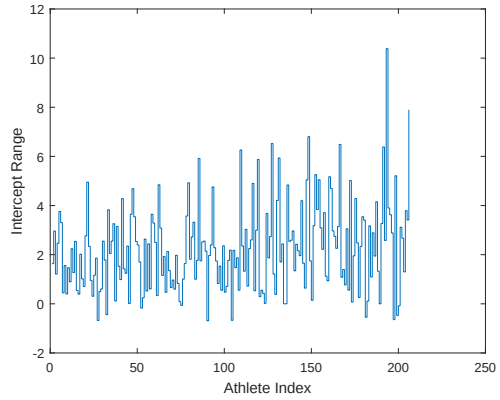
6.4.2 Regression Graphs

The model consists of 8 features. Health, stress, nutrition, hydration, sleep, energy, soreness and irritability. The regression of health as a function of other features that is feature health is calculated as a function of other 7 features. Compared beta (coefficient) values and R-squared values for 206 athletes. The Linear regression model is given in (6.1),

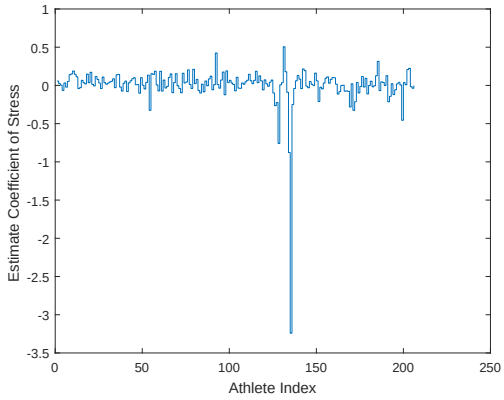
$$\begin{aligned} Health_t \sim 1 + Stress_t + Nutrition_t + Hydration_t + Sleep_t \\ + Energy_t + Soreness_t + Irritability_t + Error_t \end{aligned} \tag{6.1}$$

The estimate coefficients of intercept and other 7 features are shown below in figure 6.9 and 6.10 for 206 athletes. These graphs are useful to clearly visualize the coefficients β of different athletes. The β (Estimate coefficient) values are different for every individual

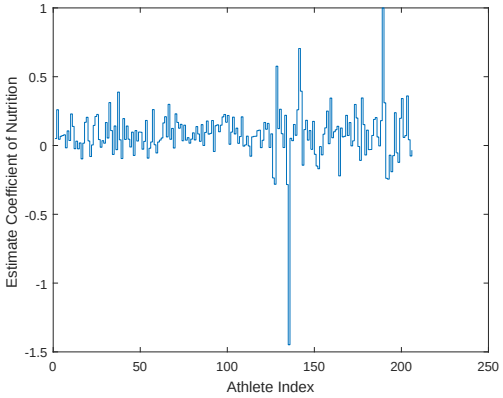
athlete.There is no correlation between them.



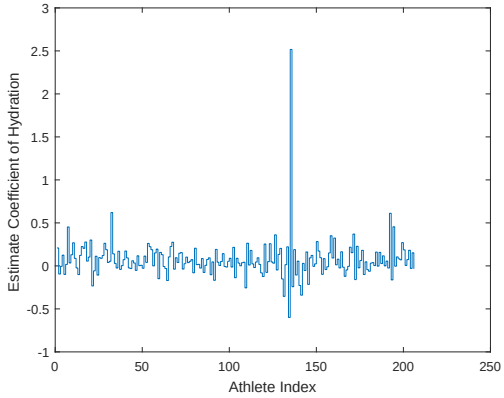
(a) Intercept



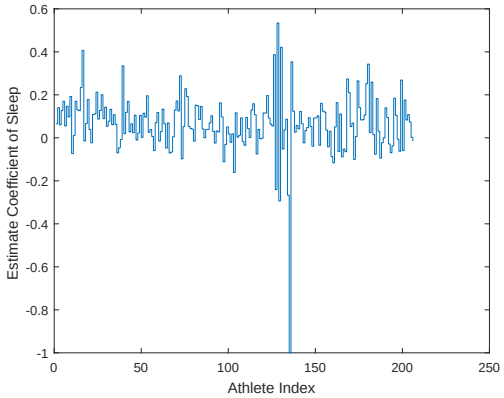
(b) Estimate coefficient of stress



(c) Estimate coefficient of nutrition

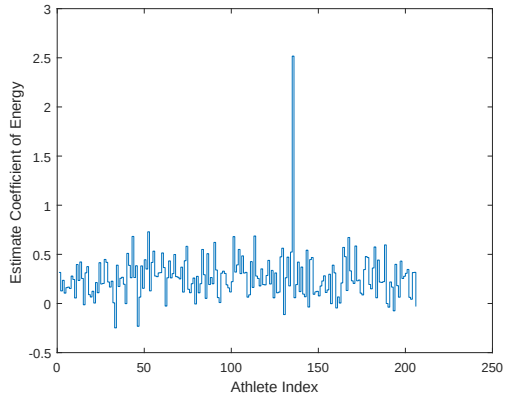


(d) Estimate coefficient of hydration

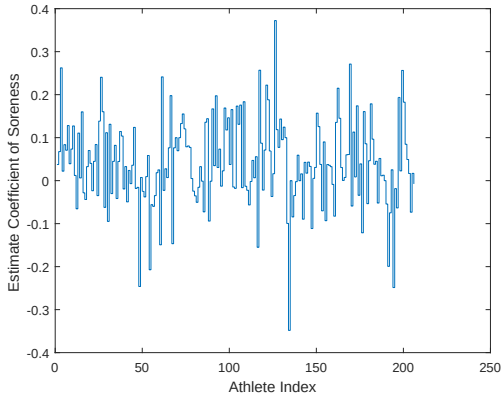


(e) Estimate coefficient of sleep

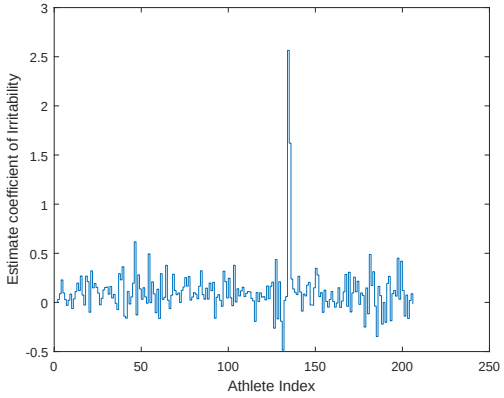
Figure 6.9: Estimate Coefficients for the Linear Model



(a) Estimate coefficient of energy



(b) Estimate coefficient of soreness



(c) Estimate coefficient of irritability

Figure 6.10: Estimate Coefficients for the Linear Model

In figure 6.9 and 6.10, athletes index range from 1 to 206 is taken in X axis and on the Y axis estimate coefficients are taken. The R-Squared value for 206 athletes of Sports-1 domain is shown in figure 6.11. In this some of the athletes have high R- squared value and some of the athletes have low R- squared value. The R^2 =1 is due to the corresponding user has a poor history of data. More specifically this user entered the features only for 5 days. Another example of $R^2 = 0.7104$ is due to the corresponding athlete has 129 days of history, but he/she entered the score almost "7" for every single feature.

As mentioned earlier, there is no correlation between coefficients values and every individual athlete has different R-Squared values. All the athletes are performing in an antithetical

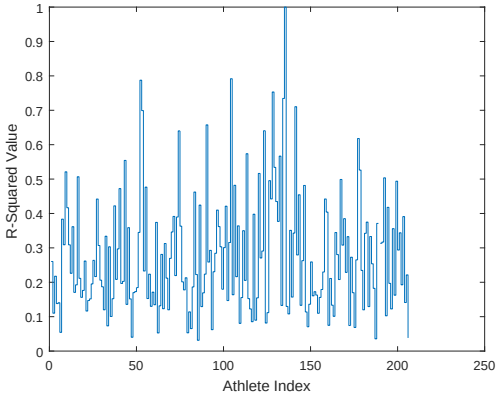


Figure 6.11: R-Squared value for all athletes for the response variable of health

way. It is suggested that overall athletes regression is not the best choice. The individual athlete regression analysis will give the best results.

6.5 Time Series Regression Results(ARX Model)

6.5.1 ARX Model-1(8 Features)

The model-1 consists of 8 features. Health, stress, nutrition, hydration, sleep, energy, soreness and irritability. Here exogenous variables are stress, nutrition, hydration, sleep, energy, soreness and irritability. The response variable is health one day in advance. Feature health one day ahead is predicted. In first model all exogenous features and the past value of response variable are considered with order 1 that is an ARX model with order 1. The time series regression model-1 design equation is given in (6.2) and the coefficients are shown in table 6.3, where N is the number of samples.

$$\begin{aligned} Health_{t+1} \sim 1 + Health_t + Stress_t + Nutrition_t + Hydration_t + Sleep_t + \\ Energy_t + Soreness_t + Irritability_t + Error_t \end{aligned} \tag{6.2}$$

6.5.2 Time Series ARX Model-2(11 Features)

The model-2 consists of 11 features. Health, stress, nutrition, hydration, sleep, energy, soreness and irritability. Here exogenous variables are stress, nutrition, hydration, sleep,

Table 6.3: Time Series Regression Model-1 with Eight Features, N=373

	Estimate	SE	tStat	pValue
Intercept	3.1566	0.47704	6.6172	1.31e-10
Health	0.4967	0.05270	9.4253	5.12e-19
Stress	-0.0497	0.07199	-0.6907	0.49020
Nutrition	0.0193	0.05350	0.3600	0.71909
Hydration	-0.0003	0.06575	-0.0043	0.99658
Sleep	-0.0143	0.04835	-0.2959	0.76750
Energy	-0.0846	0.06175	-1.3707	0.17130
Soreness	-0.0310	0.04511	-0.6864	0.49284
Irritability	0.0845	0.07771	1.0874	0.27757
Number of observations : 373 , Error degrees freedom : 364				
Root Mean Squared Error : 0.872				
R- squared: 0.226, Adjusted R- squared: 0.209				
F-statistic Vs. constant model: 13.3, p- value : 7.1e-17				

energy, soreness and irritability. The response variable is health one day in advance. Feature health one day ahead is predicted. In this model all exogenous features and past value of the response variable are considered with lag 1. Hence this is an ARX model with order 1. Mean square error, mean absolute error and error rate are calculated to evaluate the performance of the model. The dataset is spilt into training and testing data set. 70% of the data set is considered as a training data set. 394 days are trained. Remaining 30% of data is taken to validate the model. 168 days of history of data taken as testing data set.

Mean Square Error(MSE)

MSE measures the average of the squares of the error. It is the average of the square of the difference between the actual value and the predicted value. MSE is given by (6.3),

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - Y_i')^2 \tag{6.3}$$

Mean Absolute Error(MAE)

MAE is the average of the absolute value of error. Error is the difference between actual value and predicted value. MAE equation is given by (6.4),

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|(Y_i - Y_i')| \tag{6.4}$$

Error Rate

To find out the error rate, predicted value is rounded to nearest integer denoted as Y_r'. (rounded predicted value). The error rate is calculated between actual value and rounded predicted value. (health shifted feature). The total number of errors is calculated to know about how many days are predicted wrong. The error rate equation is given by (6.5)

$$\text{Error Rate} = \frac{1}{N}\sum_{N}(|(Y - Y_r')| > 0) \tag{6.5}$$

Where N is the total number of days , $Y_i, i = 1 \cdots N$ is the actual value of health feature, $Y_i', i = 1 \cdots N$ is the predicted value of health feature and Y_r' is the rounded predicted value of health feature. The time series regression model-2 design equation is given in (6.6) and the coefficients are shown in table 6.4, where N is the number of days in the training set.

$$\begin{aligned} Health_{t+1} \sim 1 + Health_t + Stress_t + Nutrition_t + Hydration_t + Sleep_t + Energy_t + \\ Soreness_t + Irritability_t + Rest_t + Exertion_t + Enjoyment_t + Error_t \end{aligned} \tag{6.6}$$

For testing data set(30% of data), the model is validated to predict the health feature one day ahead using the trained model. The following parameters are obtained using the testing data set,

Mean square error = 0.2617.

Mean absolute error = 0.3862.

Error rate= 0.2814.

The number of errors are 47 when 167 days are tested. The error rate is shown in figure 6.12.

Table 6.4: Time Series Regression Model-2 with 11 Features, N=394

	Estimate	SE	tStat	pValue
(Intercept)	1.9828	0.6837	2.9002	0.00394
Nutrition	-0.0111	0.0664	-0.1668	0.86761
Sleep	0.0130	0.0628	0.2069	0.83617
Irritability	0.0957	0.0536	1.7857	0.07495
Hydration	-0.1260	0.0496	-2.5396	0.01146
Stress	0.0285	0.0570	0.5004	0.61711
Rest	0.1541	0.1237	1.2453	0.21379
Energy	-0.1414	0.0660	-2.1428	0.03276
Soreness	0.0811	0.0636	1.2754	0.20293
Health	0.4904	0.0482	10.173	1.16e-21
Enjoyment	-0.0333	0.0347	-0.9605	0.33741
Exertion	0.0145	0.0099	1.4627	0.14437
Number of observations : 393 , Error degrees freedom : 381				
Root Mean Squared Error : 0.498				
R- squared: 0.262, Adjusted R- squared: 0.24				
F-statistic Vs. constant model: 12.3, p- value : 7.84e-20				

6.5.3 Time Series ARX Model- 3

Time series ARX Model- 3 is same like ARX model- 2 but only significant features are considered. Significant features are taken using t-Test. Backward variable selection is used. All the variables are included in the model. The variables are removed one at a time by checking t-statistics value. The variables are removed from the model when |t-statistics| values are less than 2. This procedure is continued until all the variables gets a significant result in t-stat. Achieving the significant model with equation given in (6.7) and the coefficients are shown in table 6.5, where N is the number of samples in the training set.

$$Health_{t+1} \sim 1 + Hydration_t + Energy_t + Health_t + Error_t \tag{6.7}$$

The same procedure of ARX model- 2 is followed here. The data set is split into two sets. 70% of data is taken as training set. 30% of data is taken as testing data set. For testing data set, the model is validated. The following results are obtained,
Mean square error = 0.2727.

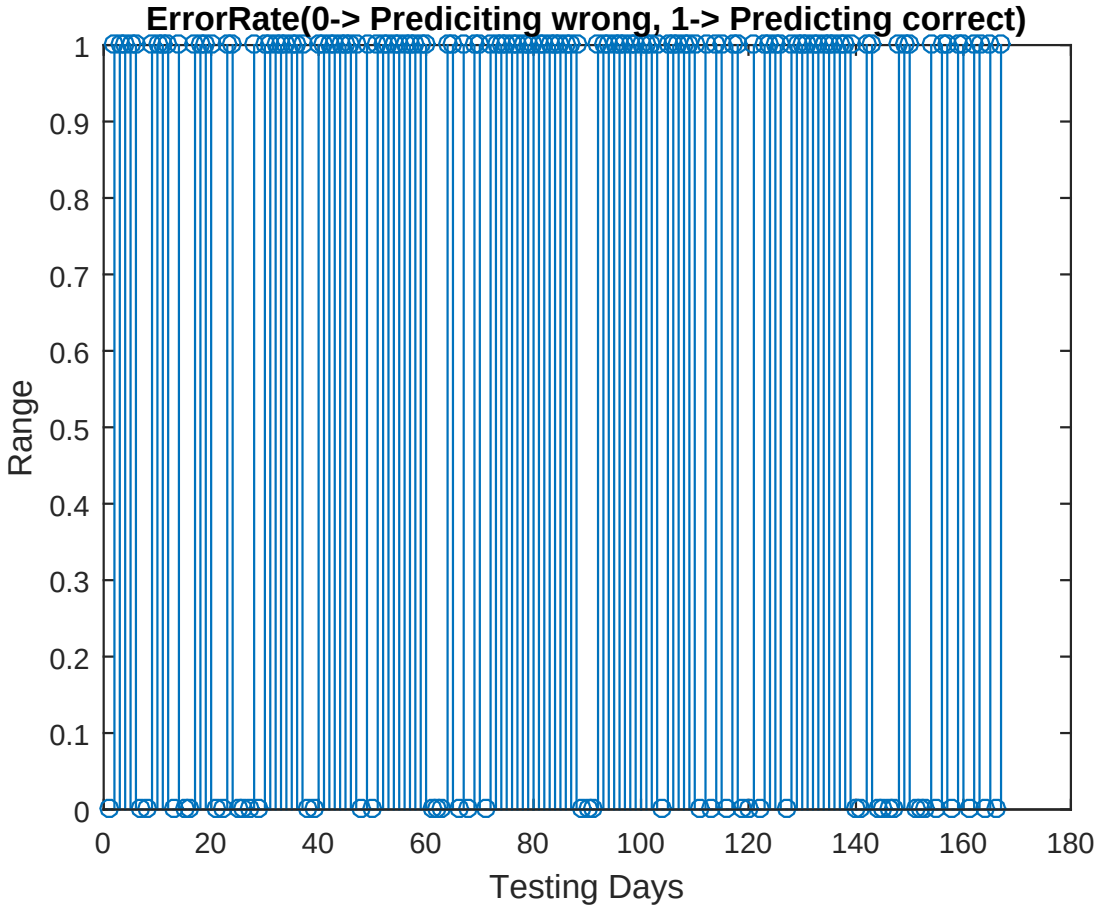


Figure 6.12: Error rate of time series regression model-2 with 11 features, predicted 47 days of events wrong out of 167 days tested

Table 6.5: Time Series Regression Model-3 with Significant Features, N=394

	Estimate	SE	tStat	pValue
(Intercept)	3.1302	0.3246	9.6427	7.16e-20
Hydration	-0.0954	0.0475	-2.0085	0.045280
Energy	-0.1079	0.0497	-2.1731	0.030376
Health	0.5062	0.0454	11.1600	2.81e-25
Number of observations : 393 , Error degrees freedom : 389				
Root Mean Squared Error : 0.499				
R- squared: 0.245, Adjusted R- squared: 0.239				
F-statistic Vs. constant model: 42, p- value : 1.55e-23				

Mean absolute error = 0.4162.

Error rate= 0.2695.

Total number of errors are 45. The error rate is shown in figure 6.13.

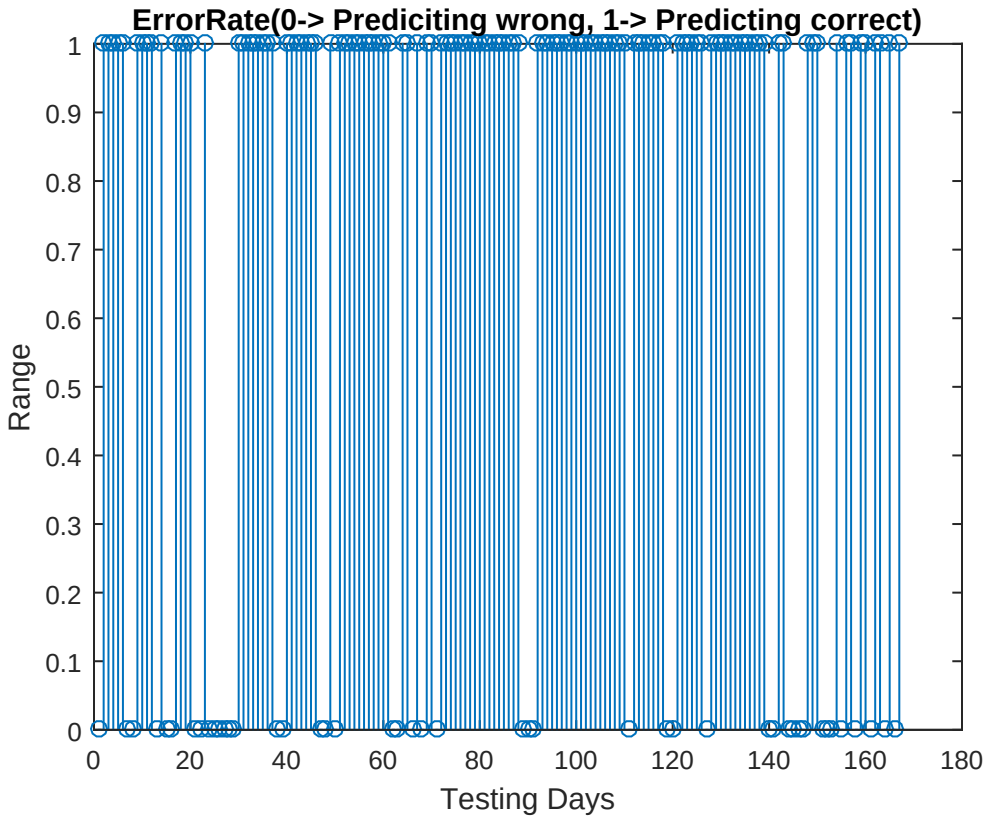


Figure 6.13: Error rate of time series regression model-3 with significant features (Hydration ,Energy and Health), predicted 45 days of events wrong out of 167 days tested

6.5.4 Time Series ARX model-4 with Error Rate by Optimizing Beta

The model 2 and model 3 validation is good in terms of producing less error rate. This project aims for prediction of illness/ injury with most accuracy. Model 3 gives the error rate of 26% which is really good. There exists drawback in terms of calculating the error function. The error function is not optimized according to the objective function. This can be improved to get better results. The error function is not evaluated using out-of-sample

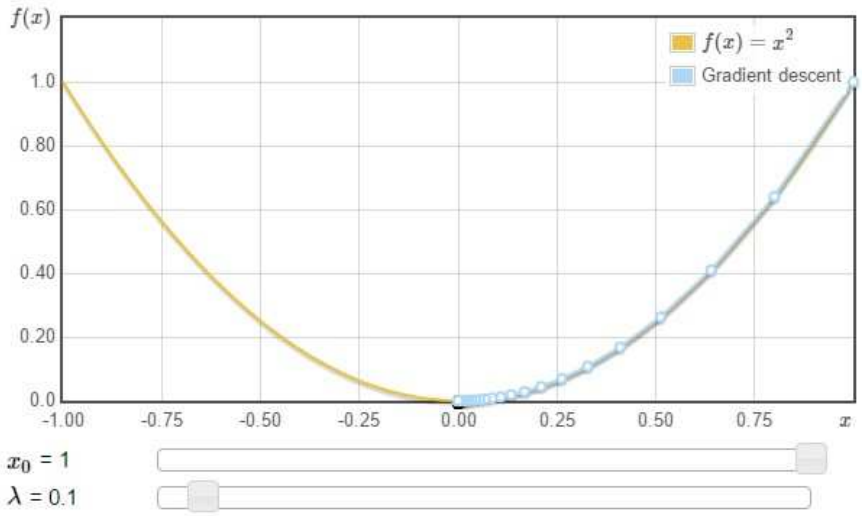
data. For in-sample data, with the more model parameters, it always gives the better results. There is a method to overcome this drawback:

1) Finding min (J(BX-Y)) instead of the ordinary least squares: $\min(||BX - Y||^2)$.

2) Beta values should be optimized to make sure as many points fall ON the line (or a linear zone) as possible with the error rate function.

3) A gradient descent technique is used as an optimization algorithm which is used to find a local minimum of beta values.

4) The way it works, start with an initial guess of the solution and take the gradient of the function at that point. Step the solution in the negative direction of the gradient and repeat the process. The algorithm will eventually converge where the gradient is zero (which correspond to a local minimum). The gradient descent diagram is shown in figure 6.14. The



The sliders above control the initial point x_0 and the constant λ. In the above plot you can see the function to be minimized and the points at each iteration of the gradient descent. If you increase λ too much, the method will not converge.

Figure 6.14: Gradient Descent Example
source: http://www.onmyphd.com/?p=gradient.descent & ckattempt=1

model is designed with all features for the prediction of health feature one day ahead. The

time series regression model-4 design equation is given in (6.8),

$$Health_{t+1} \sim 1 + Health_t + Stress_t + Nutrition_t + Hydration_t + Sleep_t + Energy_t + Soreness_t + Irritability_t + Rest_t + Exertion_t + Enjoyment_t + Error_t \tag{6.8}$$

The data set is split into training data set and testing data set. The model is trained to predict the health feature with a training data set. Beta values (Estimate coefficients) are taken from this trained linear model as initial points to gradient descent method optimization. Gradient descent method is performed to find the global minimum. Hence the beta values are optimized according to the error function given in (6.9),

$$\text{Error}, e_n = Y - Y'_n = Y_n - \beta Z_n \tag{6.9}$$

Where, Y_n is the actual value at time n, Y'_n is the predicted value at time "n".The error discriminator is given by equation (6.10),

$$\begin{aligned} \text{Error discriminator}, \zeta(e_n) = \{1, |e_n| > \lambda \\ \{0, |e_n| < \lambda \end{aligned} \tag{6.10}$$

Where λ is the parameter (Example: 0.5). The error rate function is given by equation (6.11) and sigmoid function which is approximating the $\zeta(e_n)$ is given by equation (6.12),

$$\text{Error Rate Function}, J(\beta) = \sum_{n=2}^{N} \frac{\zeta(e_n)}{N-1} \tag{6.11}$$

$$\zeta(e_n), \text{approximated by sigmoid function} = min_\beta \sum_{n=2}^{N} \frac{2}{1 + e^{-(Y_n - \beta Z_n)}} - 1 \tag{6.12}$$

Where β is the optimizing parameter given by $[\beta_0, \beta_1, \beta_2 \cdots \beta_k]$, k is the number of features, Z_n is given by $[1\ Y_{(n-1)}\ X_{1(n-1)}\ X_{2(n-1)} \cdots X_{k(n-1)}]$, N is the history of data, Y is the variable to predict, Y_{n-1} is the variable to predict at time n-1, X are the exogenous features.

Once the beta values are optimized, these are taken to predict the health feature for testing data set. Hence the new predicted health feature is calculated by using the equation (6.13),

$$Y'_n = \beta_{new} Z_{ntest} \tag{6.13}$$

The error rate is calculated for the new predicted value with optimized function. The predicted value is rounded to nearest integer (rounded predicted value). The error rate is calculated between actual value and rounded predicted value(health shifted feature). Obtained the error rate value of 0.2515 and the number of errors are 42 in 167 days, which is less than ARX model 2 when all the features are considered. Optimizing the beta values will improve the accuracy in prediction.

6.5.5 Summary of Time Series Models Results

The time series models are developed to predict one day ahead health condition. Model-1 is developed with 8 features. Only R^2 value is calculated for this model, which is 0.226. Model-2 has 11 independent features. Here, the error rate is 0.2814. Model-3 is implemented using the backward selection method that has significant features hydration, energy and health. The achieved error rate is 0.2695, which is less than model-2. Model-4 is obtained by optimizing error function. The error rate is calculated based upon the new predicted value of health obtained using new beta value. For this model, the error rate is 0.2515, which is better than other models. It shows optimizing error function helps to improve the model performance. The results of all time series models are shown in table 6.6

Table 6.6: Results of Time Series Regression Models

Name of the Model	R^2	MSE	MAE	Error rate	Number of errors
Model-2 (11 Features)	0.262	0.2617	0.3862	0.2814	47/167
Model-3 (Significant Features)	0.245	0.2727	0.4162	0.2695	45/167
Model-4 (Optimizing Error function)	0.262	0.2534	0.3734	0.2515	42/167

6.6 Logistics Regression

6.6.1 Multinomial Logistics Regression Model-1

In binary logistics regression only two levels (success/failure) are considered in the response variable. Response variable with more than two levels are called as multinomial logistics regression. The model is trained with response variable which has four classes in health feature: Very poor, poor, average and good. Lagged predictors are used to predict one day ahead value of health feature. Features are range from 1 to 7. The following levels are considered which is shown in table 6.7.

Table 6.7: Range of Sick and Healthy for Multinomial model

Range	Levels
1 - 1.9	Very poor
2 - 3.9	Poor
4 - 4.9	Average
5 - 7.9	Good

The model is trained with 70% of training data. For the prediction of health feature with hydration, energy and health with one day lag are used. Significant features are used in the model design. The model is validated using 30% of testing data. For example, for the given value [Hydration, Energy, Health] = [3, 5, 2],the achieved probabilities is shown in table 6.8.

Table 6.8: Achieved Probabilities of Multinomial Model

Levels	Achieved Probability	Lower Error Bound	Upper Error Bound
Very poor	0.0000	0.0000	0.0000
Poor	0.9487	0.1442	0.1442
Average	0.0504	0.1420	0.1420
Good	0.0008	0.0025	0.0025

6.6.2 Binomial Logistics Regression Model-2 with 11 Features

Model is trained with dependent variable health having two classes healthy and sick. Considered all features (totally 11 features) as independent variables, X = [Nutrition, Sleep, Irritability, Hydration, Stress, Rest, Energy, Soreness, Health, Enjoyment ,Exertion]. X is the binary response variable of health feature. The range that determines the sick and healthy is given in table 6.9 and the coefficients are shown in table 6.10,where N is the number of days in the training set.

Table 6.9: Range of Sick and Healthy for Binomial Model

Range	Levels
1 to 3	Sick
4 to 8	Healthy

Table 6.10: Binomial Model with all Features Coefficients Values, N=381

	Estimate	SE	tStat	pValue
Intercept	0.8334	6.0554	0.1376	0.8905
Nutrition	-0.8417	0.5635	-1.4936	0.1353
Sleep	0.2085	0.5943	0.3509	0.7256
Irritability	-0.2689	0.5956	-0.4515	0.6516
Hydration	0.7859	0.6942	1.1320	0.2576
Stress	0.0695	0.4514	0.1539	0.8777
Rest	-0.8619	1.2039	-0.7160	0.4740
Energy	-0.0570	0.5528	-0.1031	0.9178
Soreness	-0.5512	0.6166	-0.8940	0.3713
Health	1.8410	0.4980	3.6964	0.0002
Enjoyment	0.2718	0.3953	0.6877	0.4917
Exertion	0.0038	0.1556	0.0247	0.9803

Value of health <= 3 considered as "sick". Value from 4 -7 considered as "healthy" prediction. Predicted one day from now. This model can be used to predict n'th day from now.Calculated average error rate as the average of the probabilities of wrong predictions. Probabilities are compared with the actual value of health feature. The threshold is selected as mid point 3.5. If a health feature is greater than 3.5, the prediction should belong to a

healthy condition, hence sick probability taken as an error. If a health feature is less than 3.5, the prediction should belong to sick, hence healthy probability taken as an error. Average of this error is calculated as shown below,

Average Probability of error = $\frac{\text{Error Probability}}{\text{History of data}}$

The model is validated using testing data set. That is 30% of the data set which has 163 days of sample. For example, of the given data of independent features,

Testing set=[Nutrition,Sleep,Irritability,Hydration, Stress, Rest,Energy,Soreness,Health,Enjoyment,Exertion]

Testing set = [3, 2, 2, 3, 4, 5, 2, 2, 3, 3, 2]

Deviance = 66.2982.

Healthy Probability = 0.8380.

Sick Probability = 0.1620.

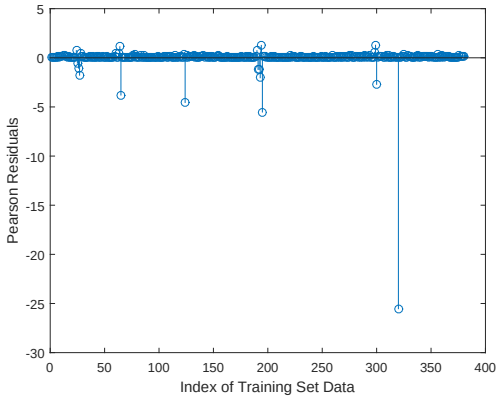
For testing data set,

Average Probability of error = 0.0301.

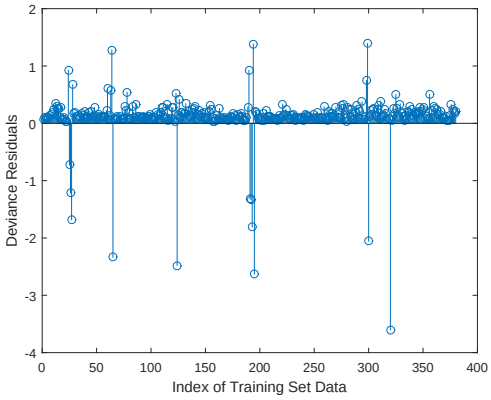
Deviance is the difference between the considered model and saturated model(model with maximum number of features). Lower the deviances are better. Additionally the plot of Pearson and Deviance residuals are used to find the outliers. This explains how well the model fits with the data. The training set has 381 days (History of data). This is taken in X axis as an index. In Y axis Pearson /Deviance residuals are taken. It is easy to compare the deviances of the observations in order to see which particular days are deviating from the model.This is shown in figure 6.15.

6.6.3 Variable Selection in Binomial Logistics Regression Model-3

In the previous model all features are considered. In the following model only significant features are considered. For a selection of variables backward selection method is used. All the features are included in the model. Significance is tested using t-statistics, this is defined as Wald statistical variable Z. If |t- statistics| > 2, the feature is considered as significant. Variable removal starts from the feature which is having least t- statistics value in turn.



(a) Pearson Residuals



(b) Deviance Residuals

Figure 6.15: Pearson and Deviance Residuals for Binomial Model with 11 independent Features showing outliers, which helps to improve the model by removing them.

Finally achieved the significant model with only health feature. Here prediction is one day ahead. Hence the model is designed in such a way that the prediction of health feature is dependent on previous day health feature. This equation is shown in 6.14 and the coefficients are shown in table 6.11,where N is the number of days in the training set.

$$ln\left[\frac{P_i}{1-P_i}\right] = \beta_0 + \beta_1 Health_t + Error \tag{6.14}$$

Where β_0 is the intercept , β_1 is the coefficient, $i = 1, 2 \cdots N$, N is the history of data, t -time, $P_i = P(Y_i = 1) = 1 - P(Y_i = 0)$, $P(Y_i = 1)$ and $P(Y_i = 0)$ are probabilities of healthy and sick of an observation i respectively.

Deviance= 72.1808.

Table 6.11: Binomial Model with Significant Features Coefficients Values, N=381

	Estimate	SE	tStat	pValue
Intercept	3.3925	1.2377	2.7410	0.0061
Health	-1.4028	0.2740	-5.1183	3.08e-07

The model is validated using testing data set that is for 163 days of sample.For example, for the given data of independent feature which is health,

X= [3]

Healthy Probability = 0.6933.

Sick Probability = 0.3067.

For testing data set,

Average Probability of error = 0.0339.

6.6.4 Stepwise Binomial Logistics Regression Linear Model-4

Stepwise regression is used to do variable selection. One day ahead prediction is used to predict the health feature. It uses both forward and backward selection of variables. This gives the same results like the backward selection. It gives the fitted model with only health feature. The p-value of health is less than threshold value 0.05. Hence it is considered as significant. Logistics regression uses chi-square statistics rather in linear regression uses F-statistics. Generalized Linear Regression model: $ln\left[\frac{P_i}{1-P_i}\right] \sim 1 + x9$, the coefficients for this model is shown in table 6.12, where N is the number of days in the training set and the distribution is binomial,$P_i = P(Y_i = 1) = 1 - P(Y_i = 0)$, $P(Y_i = 1)$ and $P(Y_i = 0)$ are probabilities of healthy and sick of an observation $i = 1, 2 \cdots N$ respectively.

Table 6.12: Stepwise Binomial Constant Model with Significant Features Coefficients Values, N=381

	Estimate	SE	tStat	pValue
(Intercept)	-3.3925	1.2377	-2.741	0.0061256
Health	1.4028	0.2741	5.1183	3.0827e-07
380 observations , 378 error degrees of freedom				
Dispersion: 1				
Chi^2-statistics vs. constant model: 27.4 , p-value = 1.63e-07				

Deviance= 72.1808.

R-Squared: 0.1578.

The model is validated using testing data set which has 163 days of sample. For example, for the given data of independent feature which is health,

X= [3]

Healthy Probability = 0.6933.

Sick Probability = 0.3067.

For testing data set,

Average Probability of error = 0.0339.

Stepwise with constant model gives the probability of health and sickness as well average probability of error all are same as backward selection model. This model is designed to take only the constant term. Deviance test is conducted to check the goodness fit of the model. Deviance is twice the difference between the log likelihoods of the corresponding model and the saturated model. Error degrees of freedom (DFE) is the number of observations minus the number of parameters in the corresponding model. Chi-squared statistic (chi2stat) is the difference between the deviance of the constant model and the deviance of the full model. P-value is the chi-squared statistic with (number of coefficients in the model minus one) degrees of freedom as explained in section 5.3. In this model the p-value is less than 0.05 which indicating the significance of the model, it is shown in table 6.13.

Table 6.13: Deviance test in Stepwise Binomial Constant Model with Significant Features

	Deviance	DFE	Chi2stat	pValue
Logit(y) $\sim$ 1	99.609	379		
Logit(y) $\sim$ 1+Health	72.181	378	27.428	1.6308e-07

6.6.5 Stepwise Logistics Regression Non-Linear Model -5

In this section, stepwise binomial logistics regression with linear model for the prediction of health one day ahead is presented. The following shows the stepwise process regarding the adding or removing the variables with linear models. It gives constant, linear terms and product pairs for each predictor.

1. Adding Soreness: Exertion, Deviance = 61.5756. Chi2Stat =4.72251 , Pvalue = 0.0297702.

2. .Adding Hydration: Health, Deviance = 51.3531. Chi2Stat = 10.2226 , Pvalue = 0.00138733.

3. Adding Sleep: Irritability , Deviance = 45.8408 ,Chi2Stat = 5.5123, Pvalue = 0.0188832.

4. Removing Stress , Deviance = 45.857 ,Chi2Stat = 0.01665 , Pvalue = 0.89733.

5. Removing Enjoyment, Deviance =46.045, Chi2Stat = 0.18783 , Pvalue = 0.66473.

6. Removing Rest , Deviance =47.551, Chi2Stat = 1.5061 , Pvalue = 0.21974.

7. Removing Energy, Deviance =49.053, Chi2Stat = 1.5012 , Pvalue = 0.22049.

Features stress, enjoyment, rest and energy are removed as it has p- value greater than 0.05. Additionally it, checks the improvement in R-squared also. If the increase in R-squared of the model is larger than default value 0.1, then the particular variable is added to the model.Hence the final design for a stepwise linear model is,

Generalized linear regression model: $ln\left[\frac{P_i}{1-P_i}\right] \sim 1+$ Nutrition + Sleep*Irritability + Hydration * Health + Soreness * Exertion.

The coefficients for this model is given in table 6.14, where N is the number of days in training set, $P_i = P(Y_i = 1) = 1 - P(Y_i = 0)$, $P(Y_i = 1)$ and $P(Y_i = 0)$ are probabilities of healthy and sick of an observation $i = 1, 2 \cdots N$ respectively

R-squared= 0.4539.

Deviance = 49.05.

The model is validated using testing data set that is with 163 days of sample. For example, for the given data,

Testing Set = [Nutrition. Sleep, Irritability, Hydration, Soreness, Health, Exertion]

Testing Set = [3, 2, 3, 1, 1, 3, 2]

Healthy Probability = 0.3840.

Sick Probability = 0.6160.

For testing data set,

Average Probability of error = 0.0401.

Table 6.14: Coefficients of Stepwise Binomial Linear Model, N=381

	Estimate	SE	tStat	pValue
(Intercept)	-20.105	26.162	-0.7685	0.44220
Nutrition	-1.4017	0.8047	-1.7419	0.08153
Sleep	-6.7697	3.6086	-1.8760	0.06066
Irritability	-6.7535	3.8297	-1.7635	0.07782
Hydration	12.6030	4.3992	2.8649	0.00417
Soreness	-4.8038	1.6403	-2.9286	0.00340
Health	18.5050	6.3188	2.9286	0.00340
Exertion	-3.9300	1.4256	-2.7568	0.00584
Sleep: Irritability	1.3760	0.7409	1.8571	0.06330
Hydration: Health	-2.5041	0.9328	-2.6845	0.00726
Soreness: Exertion	0.7198	0.2571	2.7995	0.00512
380 observations , 369 error degrees of freedom				
Dispersion: 1				
Chi^2 -statistics vs. constant model: 50.6 , p-value = 2.11e-07				

Stepwise regression with linear model has a high R-squared value which is 0.4539 compared with a constant term model which is 0.1578. If we compare the average probability of error of stepwise linear model which is higher than the constant model.

6.6.6 Summary of Logistics Regression Models Results

Model 1 is implemented for multinomial logistics regression. Model 2 to 5 designed with binomial logistics regression. From all of them, stepwise linear model-4 gives the better results, which considers the constant and linear terms of predictors. The average probability of error is 0.0339. This model considered only significant features hence avoiding the chances of making misclassification. The results of binomial logistics models are shown in table 6.15.

6.6.7 Binomial Logistics Regression One Week Ahead Prediction

Logistics regression is used in the analysis of prediction of one week ahead. Health feature is predicted as a function of other independent features. We define the index of health to be a binary number representing the status of health. More specifically, if the value of the health

Table 6.15: Results of Binomial Logistics Regression Models

Name of the Model	Average Probability of error
Model-2 (11 Features)	0.0301
Model-3 (Significant Features)	0.0339
Model-4 (Stepwise-Constant and Linear Features)	0.0339
Model-5 (Stepwise-Constant,Linear Features and Product pairs)	0.0401

feature is larger than 3.5, we assign "1" to the value of health index, otherwise "0" will be assigned. Moreover, the criteria of predicting the health status of an athlete are decided to be the following statement, if an athlete is sick at least one day during a period of one week, we say the corresponding athlete is sick, otherwise we assume he/she is healthy. This gives us the tool to use the logical OR function over the health index of an athlete in a time period of one week. Such a variable can then be used in a predictive model as the observation variable. Following the logic that we have presented before in the logistic regression model, presented earlier, we aim to obtain a logistic regression function of health-related features that can explain the observation variable that was introduced earlier in this chapter. The more precise equation is shown in (6.15),

$$\begin{aligned}\text{Index(Health)}_t\text{Sick} = OR\ Function(&\text{Health}_{t-7} < 3.5, \text{Health}_{t-6} < 3.5, \text{Health}_{t-5} < 3.5,\\ &\text{Health}_{t-4} < 3.5, \text{Health}_{t-3} < 3.5, \text{Health}_{t-2} < 3.5, \text{Health}_{t-1} < 3.5)\\ &\text{otherwise}(\text{Index(Health)}_t) \text{ is Healthy}\end{aligned} \tag{6.15}$$

The range of sick and healthy conditions is shown in table 6.16, Index of health is obtained by using equation (6.15). Therefore, logistics regression is used to train and test the developed model. As we mentioned earlier, the value of health index is taken as the dependent variable.

Table 6.16: Range of Sick and Healthy for One week Prediction

Range	Levels
1 to 3	Sick
4 to 8	Healthy

Independent variables are chosen according to the well-known backward selection procedure in classical regression. We note that we have included all features in logistic regression model, however, using the significance test "Wald Test", one can omit the insignificant parameters within a certain range. If | t-statistics| $>$ 2, the feature is considered to be significant. The successive variable removal procedure starts from a feature with the lowest t-value and remove it from the model. This procedure is repeated enough such that all the remaining features in the model are significant. For our specific data set for a certain user, only the health feature turns out to be significant. Therefore, the proposed model is given by equation (6.16),

$$ln\left[\frac{P_i}{1-P_i}\right] = \beta_0 + \beta_1 Health_{i(t-7)} + Error \tag{6.16}$$

Where β_0 is the intercept , β_1 is the constant,$i = 1, 2 \cdots N$, N is the history of data, (t-7) is one week behind in time, $P_i = P(Y_i = 1) = 1 - P(Y_i = 0)$, $P(Y_i = 1)$ and $P(Y_i = 0)$ are probabilities of healthy and sick of an observation $i = 1, 2 \cdots N$ respectively. The model is validated with 30% of testing data set with trained model. Achieved average probability of error in the testing data set is 0.1531.

False Positive and False Negative Calculation

Confusion matrix [25],is used to calculate the false positive and false negative values. A false negative is defined as predicting healthy when the athlete is really sick. A false positive is defined as predicting sick when the athlete is really healthy. The model is validated using trained logistics regression model of one week ahead prediction. For the validation, testing dataset is used which has 156 days of sample. The general confusion matrix is shown in table 6.17, where PT is probability threshold.

N1- Actual (Sick) versus Predicted (Sick)

Table 6.17: Confusion Matrix

		Predicted Probabilities (Y)	
		Sick ($Y < PT$)	Healthy ($Y > PT$)
Actual index of health (I)	Sick(I=1)	N1(TP)	N2(FN)
	Healthy(I=0)	N3(FP)	N4 (TN)

N2- Actual (sick) versus Predicted (Healthy)

N3- Actual (Healthy) versus Predicted (Sick)

N4- Actual (Healthy) versus Predicted (Healthy)

I- Index of health feature of testing data set (156 days)

Y- Predicted health probability of testing data set.

FN- False negative (Predicting healthy when the actual health index is sick)

FP- False positive (Predicting sick when the actual health index is healthy).

N1 is the number of counts of predicted sick events when the athlete is actually sick. This is called true positive. N2 is the number of counts of predicted healthy events when the athlete is really sick. This is a false negative. N3 is the number of counts of sick predicted events when the athlete actually is healthy. This is a false positive. N4 is the number of counts of predicted healthy events when the athlete is actually healthy. This is called as true negative. Hence N1 and N4 are the correct number of positive and negative predictions. The true positive rate (TP) is the proportion of positive cases that were correctly identified, as calculated using the equation (6.17),

$$\text{True Positive Rate} = \frac{N1}{N1 + N2} \tag{6.17}$$

The false positive rate (FP) is the proportion of negative cases that were incorrectly classified as positive, as calculated using the equation (6.18),

$$\text{False Positive Rate} = \frac{N3}{N3 + N4} \tag{6.18}$$

The true negative rate (TN) is defined as the proportion of negative cases that were classified correctly, as calculated using the equation (6.19),

$$\text{True Negative Rate} = \frac{N4}{N3 + N4} \tag{6.19}$$

The false negative rate (FN) is the proportion of positive cases that were incorrectly classified as negative, as calculated using the equation (6.20),

$$\text{False Negative Rate} = \frac{N2}{N1 + N2} \tag{6.20}$$

The total number of predicted healthy events is given by equation 6.21,

$$\text{Total Predicted Healthy=N2+N4} \tag{6.21}$$

The total number of predicted sick events is given by equation 6.22,

$$\text{Total Predicted Sick=N1+N2} \tag{6.22}$$

The error rate is proportion of the total number of predictions that were wrong.It is calculated by equation 6.23,

$$\text{Error Rate} = \frac{N2 + N3}{N1 + N2 + N3 + N4} \tag{6.23}$$

The figure shown in 6.16, is the ROC curve for different thresholds. False positive rate is taken in X axis. True positive rate is taken in Y axis. Threshold is set to start from "0" and ends with "1" by incrementing 0.01. Hence this is the ROC plot of 100 thresholds values.

Training samples:

Actual training data set - 381 days

One week ahead training data set - 374 days of sample

Total number of sick in actual training data set - 11 samples

Total number of sick in index of health - 48 samples.

Total number of healthy in index of health - 326 samples.

Testing samples:

Actual testing data set - 163 days

After removing 7 days - 156 days of samples

Total number of sick in actual testing data set - 4 samples

Total number of sick in index of health - 10 samples.

Total number of healthy in index of health - 146 samples.

For example, for the threshold of 0.85 the confusion matrix is shown below in table 6.18,

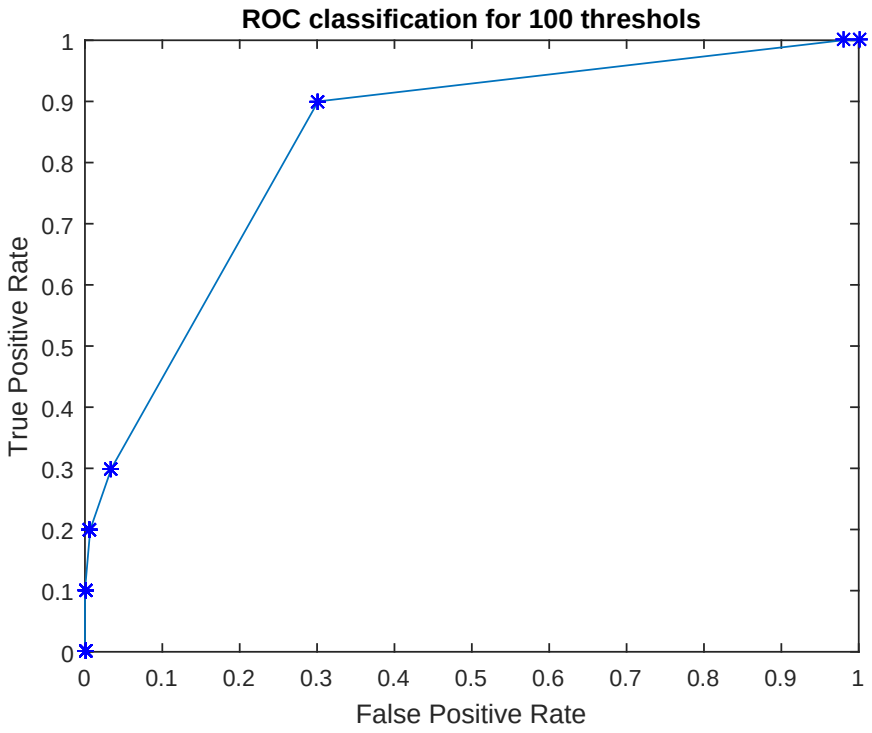


Figure 6.16: ROC curve: This figure is the plot of ROC (Receiver operating characteristics), meaning that we have plotted the true-positive against the false-positive rate. Here changed the level of threshold from 0 to 1, with step-size of "0.01" and obtained the above graph.

Table 6.18: Confusion Matrix for the Probability Threshold(PT)= 0.85

PT= 0.85 Testing samples:156 Training samples:374		Predicted Probabilities(Y)	
		Sick ($Y < PT$)	Healthy ($Y > PT$)
Actual Index of health (I)	Sick (I=1)	N1(TP):9	N2 (FN) :1
	Healthy(I=0)	N3(FP): 44	N4(TN):102

where PT is probability threshold. The obtained results are: Total actual sick = 10 ; Total actual healthy = 146 ; Total predicted sick = 53 ; Total predicted healthy = 103.

False positive rate= 0.3014. (Predicted sick when he is really healthy).

False negative rate = 0.1000 (Predicted healthy when he is really sick)

Error Rate = 0.2885.

Accuracy = 0.7115.

6.7 Chapter Summary

This chapter shows the simulation results of several regression analysis like linear, time series and logistics regression. The analysis of the data set is started with linear regression. Because the linear relationships are non trivial that can be imagined easily and easier to work also. When the data does not fit with linear type we can easily modify into non-linear regression by adding non linear terms with that as mentioned earlier in chapter 3. Time series are used to predict one day ahead. The error rate is calculated to validate the models. In among 5 models, time series with stepwise linear model performs better. Optimizing the β improves the model performance by reducing the error rate. Logistics regression is used to predict the health feature in terms of probability when the dependent variable is dichotomous. The stepwise constant model performs better.One day and one week ahead predictions of health are implemented. In one week ahead prediction,false negative and positive values are calculated using confusion matrix which gives the accurate error rates.

Chapter 7

Conclusion and Future Work

In the sports field, prediction of illness/wellness is a very helpful analysis, which help them to take care of them in advance when they come to know factors are affecting their health condition. Therefore, it is necessary to predict accurately. Regression analysis is used as a fundamental model to implement prediction analysis. In a literature review, it is shown that health is dependent on several psychological attributes.

This study analysed various regression methods to predict athlete illness/wellness. The same methods are applicable for the prediction of injury as well using soreness feature. It is really challenging that to apply specific regression analysis to specific user as every individual athlete has a different history of data and they behave differently. As our objective is the prediction of illness, the analysis is done with a user who is having more sick days.

Among various methods time series regression method with optimized function gives better results in terms of error rate. Time series regression is done for one day ahead prediction of illness. To obtain the output in terms of probability we developed the logistics regression model. Logistics regression analysis is implemented in time series data. The logistics stepwise regression model gives the better prediction in terms average probability of error. In logistics regression one day ahead and one week ahead predictions are implemented. This can be modified to "n" number of day's prediction. The accuracy is also depends on goodness of data. When some of the users just give the scores randomly for several features that affects the prediction level.

In future work, this thesis can be extended to implement one month prediction of health. The model with more inputs can be implemented. We can examine the performance on data sets where athlete does not log as frequently, to see whether the same model performs better or need to adjust the model. This premise will be the guidelines to do the prediction of injury.

Bibliography

[1] Mark B. Andersen and Jean M. Williams, "A Model of Stress and Athletic Injury: Prediction and Prevention," *Journal of Sport and Exercise Psychology*, vol. 10, pp. 294–306, 1988.

[2] "Mental Health: Strengthening Our Response," 2011.

[3] David A. Hermon and Richard J. Hazler, "Adherence to a wellness model and perceptions of psychological well-being," *Journal of Counseling and Development*, vol. 77, pp. 339–343, 1999.

[4] J Melvin Witmer, Myers Jane E and Sweeney Thomas J, "The Wheel of Wellness Counseling for Wellness: A Holistic Model for Treatment Planning," *Journal of Counseling and Development*, vol. 78, no. 3, pp. 251–266, 2000.

[5] Magnus Lindwall, Henrik Gustafsson, Mats Altemyr, Andreas Ivarsson and Urban Johnson, "Research Psychosocial Stress as a Predictor of Injury in Elite Junior Soccer: A Latent Gowth Curve Analysis," *Journal of Science and Medicine in Sport*, vol. 17, pp. 366370, 2014.

[6] Krosshaug T and Bahr R, "Understanding Injury Mechanisms: A Key Component of Preventing Injuries in Sport.," *British Journal of Sports Medicine*, vol. 39, no. 6, pp. 39:324329, 2005.

[7] Rayner Alfred and Haviluddin, "Performance of Modeling Time Series Using Nonlinear Autoregressive with exogenous input (narx) in the Network Traffic Forecasting," in

International Conference on Science in Information Technology (ICSITech), 2015, pp. 164–168.

[8] Rob J Hyndman and George Athanasopoulos, *Forecasting: Principles and Practice*, OTexts, 2013.

[9] Popovic Dobrivoje and Palit Ajoy K., *Computational Intelligence in Time Series Forecasting : Theory and Engineering Applications.*, Springer, 2005.

[10] Jianqing Yao, Fan and Qiwei, *Nonlinear Time Series*, Springer, 2003.

[11] Vaidyanathan P.P, *The Theory of Linear Prediction*, Morgan & Claypool, 2008.

[12] Peter E. Hart, Richard O. Duda and David G. Stork, *Pattern Classification*, John Wiley & Sons, second edition, 2001.

[13] Deng Li and Dong Yu, *Deep Learning Methods and Applications*, vol. 7: 3-4, 2014.

[14] Hadi Ali S and Chatterjee Samprit, *Regression Analysis by Example*, WileyInterscience, 2006.

[15] Christiaan Heij, Paul de Boer, Philip Hans Franses, Teun Kloek and Herman K. van Dijk, *Econometric Methods with Applications in Business and Economics*, Oxford university press, 2004.

[16] "Http://www.mathworks.com/help/stats/f-statistic-and-t-statistic.html," Accessed on Mar. 2016.

[17] John Fox, *Applied Regression Analysis and Generalized Linear Models*, Sage,Second edition, 2008.

[18] "Http://www.mathworks.com/help/stats/f-statistic-and-t-statistic.html," Accessed on Mar. 2016.

[19] J. H. Cochrane, "Time Series for Macroeconomics and Finance," Tech. Rep., Graduate School of Business, University of Chicago, 2005.

[20] Bo Qi and Taiyao Wang, "Strong Consistency of Parameter Estimates for Purely Explosive Autoregressive Models with Exogenous Inputs," in *33rd Chinese Control Conference*, July, 2014, pp. 6588–6592.

[21] Lemeshow and Hosmer, *Applied Logistics Regression*, vol. Second Edition, Wiley Series in Probability and Statistics, 2004.

[22] Ronald Christensen, *Log- Linear Models and Logistic Regression*, vol. 2nd edition, Springer, 1997.

[23] "Http://www.mathworks.com/help/stats/generalizedlinearmodel.deviancetest.html# btc _f6i," Accessed on Mar. 2016.

[24] "Http://www.mathworks.com/help/stats/stepwiselm.html," Accessed on Mar.2016.

[25] Debbie I. Craig and Scot Raab, *Evidence-Based Practice in Athletic Training*, Human Kinetics., 2016.