



RASHTREEYA SIKSHANA SAMITHI TRUST
R V INSTITUTE OF MANAGEMENT
IT And Business Analytics
Department



Course Outline

Programme:	MBA
Batch:	2019-2021
Semester:	3
Subject Name:	Predictive Analytics Using R Programming
Subject Code:	3.7.2
Credits:	4 (56 sessions)
Course Instructors:	Ms. Shreya Shankar and Ms. Vandana Gablani

PART A

INTRODUCTION:

The amount of data in the world is increasing exponentially as time passes. It is estimated that the total amount of data produced in 2020 will be 20 zettabytes (Kotov, 2014), that is, 20 billion terabytes. Many businesses invest lots of money and efforts for collecting the data and most of it is not analysed fully and / or not analysed appropriately. The main reason to analyse the data is to predict the future i.e. to construct actionable knowledge. This course will help and allow the students to do data analysis and build models while learning various tools & techniques. The prerequisite of the course is students must have undergone basic courses on Statistical modelling.

The Tool utilized will be R Studio. This is an open-source software which is widely used for predictive analytics in the market.



POs	Program Outcomes
1	Apply knowledge of management theories and practices to solve business problems
2	Foster Analytical and critical thinking abilities for data-based decision making
3	Ability to develop Value based Leadership
4	Ability to understand, analyze and communicate global, economic, societal, cultural, legal and ethical aspects of business
5	Ability to lead themselves and others in the achievement of organizational goals, contributing effectively to a team environment
6	Ability to identify business opportunities, frame innovative solutions and launch new business ventures or be an entrepreneur
7	Ability to deal with contemporary issues using multi-disciplinary approach with the help of advanced Management and IT tools and techniques
8	Ability to apply domain specific knowledge and skills to build competencies in their respective functional area
9	Ability to engage in research and development work with cognitive flexibility to create new knowledge and be a lifelong learner
10	Ability to understand social responsibility and contribute to the community for inclusive growth and sustainable development of society through ethical behavior
11	Ability to function effectively as individuals and in teams through effective communication and Negotiation skills

COURSE OUTCOMES (CO):

At the successful completion of this course the students should be able to;

CO1	Understand the data and develop the skill set to clean the data.
CO2	Analyze and create visualizations.
CO3	Develop and Evaluate predictive models and derive business insights for higher business efficiency and effectiveness

Mapping of Course Outcomes to Program Outcomes:

COs/POs	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	P10	P11
CO1	1	3	-	2	-	2	3	3	1	-	-
CO2	1	3	-	1	-	2	3	3	1	-	-
CO3	3	3	1	2	-	3	3	3	2	-	1

LEVEL 3-Substantial 2-Moderate 1-Slight "-" No Correlation

KEY CONCEPTS:

MODULE 1: INTRODUCTION TO PREDICTIVE ANALYTICS

- Introduction to predictive analytics
- Evolution of Data Analytics
- Applications of predictive analytics
- Predictive models

MODULE 2: EXPLORATORY DATA ANALYSIS

- Organizing and processing of data
- Data Cleaning – Missing values, Outlier treatment
- Univariate Analysis
- ARIMA model for forecasting.

MODULE 3: PREDICTION- LINEAR REGRESSION

- Understanding simple regression model
- Computing the intercept and slope coefficient
- Obtaining the residuals
- Computing the significance of the coefficient.
- Correlation & R^2
- Multiple Linear Regression

MODULE 4: DECISION TREES & LOGISTIC REGRESSION

- Introduction to Decision trees and Logistic regression
- Interpreting the model parameters
- assessing the impact of predictors on the probability of outcome.

MODULE 5: NEURAL NETWORKS



- Introduction and Structure of neural networks
- Information flow
- Types of layers
- Training a neural network
- Back Propagation

MODULE 6: INTRODUCTION TO OTHER REGRESSION ANALYSIS

- Introduction to Polynomial
- Multiple linear
- Poisson
- Nonlinear and
- Nonparametric

MODULE WISE OUTCOMES:

MODULE 1: INTRODUCTION TO PREDICTIVE ANALYTICS

Post completion of this module, students should be able to attain the following module outcomes;

MO 1: Understand the meaning and importance of predictive analytics.

MO 2: Understand Application of predictive analytics.

MODULE 2: EXPLORATORY DATA ANALYSIS

After completion of module 2, students should be able to attain the following module outcomes;

MO 3: Organize and pre-process the data for further analysis.

MO 4: Understand and implement missing values and outlier treatment.

MO 5: Evaluate and interpret univariate data.

MO 6: Understand time series data and evaluate it for forecasting.

MODULE 3: PREDICTION- LINEAR REGRESSION

Post completion of module 3, students should be able to attain the following module outcomes;

MO 7: Understand OLS for simple linear regression.

MO 8: Interpret and evaluate outcomes of simple linear regression.

MO 9: Understand Multiple linear regression.

MODULE 4: DECISION TREES & LOGISTIC REGRESSION

Post completion of Corporate restructuring, students should be able to attain the following module outcomes;

MO 10: Understand Decision tree and logistic regression.

MO 11: Interpret Model parameters and outcomes.

MODULE 5: NEURAL NETWORKS

Post completion of Financial evaluation, students should be able to attain the following module outcomes;

MO 12: Understand structure of neural network.

MO 13: Implement Neural network.

MODULE 6: INTRODUCTION TO OTHER REGRESSION ANALYSIS

Post completion of module 6, students should be able to attain the following module outcomes;

MO 14: Understand Non-linear and Non-parametric methods for data analysis.

**COURSE EVALUATION PLAN:
(a) END - TERM**

Evaluation	Weightage (%)	Duration (in Minutes)	Open / Close Book	CLO Tested
End Term Exam	70	180	Close book	All

(b) OTHER ASSESSMENT:

Sl. No.	Evaluation Item*	Unit of Evaluation	Marks Allotted	TIME	CO
1	Attendance	Individual	5	Every session	None
2	One Internal	Individual	10	Once in the semester	All
3	Capstone Project	Individual	15	Once in the semester	All

REFERENCES

1. Business Intelligence for Dummies, Swain Scheps, Wiley Publication
2. Successful Business Intelligence by Cindi Howson, McGraw Hill
3. Business Intelligence by David Leshin, Elsevier, second edition
4. Data mining for Business Intelligence, GalitShmueli, Nitin R Patel & Peter C Bruce, Wiley Publication
5. Business Intelligence, Practices, Technologies and Management, Rajiv Sabherwal, Irma Becerra-Fernandes, Wiley Publication
6. Business Intelligence Guide book, Rick Sherman, Elsevier
7. Business Intelligence Strategy & Big data analytics, Steve Williams, Elsevier
8. Statistics for management, David S. Rubin and Levin.
9. R Cookbook , Paul Teetor

Websites

1. www.statquest.com
2. www.kaggle.com
3. www.vidhyaanalytics.com
4. www.medium.com

COURSE FACILITATORS:

Dr. Santhosh M

Associate Professor

Email Id: santhoshm.rvim@rvei.edu.in



Ms. Shreya Shankar

Teaching Associate

Email ID: shreyashankar.rvim@rvei.edu.in

Prof. Dileep S

Assistant Professor

Email ID: dileep.rvim@rvei.edu.in

Ms. Vandana Gablani

Assistant Professor

Email ID: vandanag.rvim@rvei.edu.in


Director
R.V. INSTITUTE OF MANAGEMENT
C.A. 17, 36th Cross, 26th Main,
4th 'T' Block, Jayanagar,
BANGALORE - 560041





PART B
SESSION PLAN

MODULE 1: INTRODUCTION TO PREDICTIVE ANALYTICS

Session	Coverage of the Key Concept	Pedagogy/Activity (Discussion Points)	Reading material to be Referred
1	Introduction to R <ul style="list-style-type: none">• Applications of R• Basic Programing Terminology• R Environment WalkThrough	Classroom discussion with PPT	Rstudio Documentation Article https://data-flair.training/blogs/r-applications/
2	Installing R software	Classroom discussion with PPT Lab Session	Video https://youtu.be/NZxSA80IF11
3	Vector Manipulation	Classroom discussion with PPT Lab Session	Website www.datacamp.com
4	Factors - Categorical Variable Analysis Business Decision Matrix Manipulation of a Business Decision Matrix	Classroom discussion with PPT Lab Session	Website www.datacamp.com
5	Lists and Data Frames	Classroom discussion with PPT Lab Session	Website www.datacamp.com
6	Introduction to Data Visualization Ggplot2 package Types of Graphical Representation.	Classroom discussion with PPT Lab Session	Rstudio Documentation ggplot2 official documentation Data Inbuilt





MODULE 2: EXPLORATORY DATA ANALYSIS

Session	Coverage of the Key Concept	Pedagogy/Activity (Discussion Points)	Reading material to be Referred
8	Data Cleaning	Classroom discussion with PPT	Book R Cookbook Paul Teetor
9	Data Cleaning: Simple Method of Imputation	Classroom discussion with PPT Lab Session	RStudio Documentation Book R Cookbook Paul Teetor
10	Data Cleaning: Mean Imputation	Classroom discussion with PPT Lab Session	RStudio Documentation Book R Cookbook Paul Teetor
11	Data Cleaning: Median Imputation	Classroom discussion with PPT Lab Session	RStudio Documentation Book R Cookbook Paul Teetor
12	Data Cleaning: Mode Imputation	Classroom discussion with PPT Lab Session	RStudio Documentation Book R Cookbook Paul Teetor
13	Data Cleaning: Imputation Through Package- MICE	Classroom discussion with PPT Lab Session	RStudio Documentation Book R Cookbook Paul Teetor
14	Outliers : Graphical Method of Identification of Outliers	Classroom discussion with PPT Lab Session	RStudio Documentation Book R Cookbook Paul Teetor





15	Outliers : Min-Max Method of Outliers Manual Treatment of Outliers	Classroom discussion with PPT Lab Session	RStudio Documentation Book R Cookbook Paul Teetor
16	Univariate Analysis	Classroom discussion with PPT Lab Session	RStudio Documentation Book R Cookbook Paul Teetor Data Kaggle
17	Time Series Modeling	Classroom discussion with PPT Lab Session	RStudio Documentation Book R Cookbook Paul Teetor Data Inbuilt
18	Autocorrelation Factor Partial Autocorrelation Factor	Classroom discussion with PPT Lab Session	RStudio Documentation Book R Cookbook Paul Teetor Data Inbuilt
20	ARIMA Modelling	Classroom discussion with PPT Lab Session	RStudio Documentation Book R Cookbook Paul Teetor Data Inbuilt





MODULE 3: PREDICTION- LINEAR REGRESSION

Session	Coverage of the Key Concept	Pedagogy/Activity (Discussion Points)	Reading material to be Referred
21	Linear Regression	Classroom discussion with PPT Lab Session	RStudio Documentation Books Statistics For Management David S. Rubin and Levin R Cookbook Paul Teetor Data Kaggle
22	Residuals	Classroom discussion with PPT Lab Session	RStudio Documentation Books Statistics For Management David S. Rubin and Levin R Cookbook Paul Teetor Data Kaggle
23	Ordinary Least Squares	Classroom discussion with PPT Lab Session	RStudio Documentation Books Statistics For Management David S. Rubin and Levin R Cookbook Paul Teetor Data Kaggle
24	P Value	Classroom discussion with PPT Lab Session	RStudio Documentation Books Statistics For Management David S. Rubin and Levin Data Kaggle
25	T Test	Classroom discussion with PPT Lab Session	RStudio Documentation Books Statistics For Management David S. Rubin and Levin R Cookbook Paul Teetor Data Kaggle
26	F Test	Classroom discussion with PPT Lab Session	RStudio Documentation Books Statistics For Management David S. Rubin and Levin R Cookbook





			Paul Teetor Data Kaggle
27	R Squared	Classroom discussion with PPT Lab Session	RStudio Documentation Books Statistics For Management David S. Rubin and Levin R Cookbook Paul Teetor Data Kaggle
28	R Squared	Classroom discussion with PPT Lab Session	RStudio Documentation Books Statistics For Management David S. Rubin and Levin R Cookbook Paul Teetor Data Kaggle
29	Multiple Regression Analysis	Classroom discussion with PPT Lab Session	Books Statistics For Management David S. Rubin and Levin
30	Residual Analysis	Classroom discussion with PPT Lab Session	RStudio Documentation Data Kaggle
31	Residual Analysis	Classroom discussion with PPT Lab Session	RStudio Documentation Data Kaggle
32	Variance Inflation Factor	Classroom discussion with PPT Lab Session	RStudio Documentation Data Kaggle
33	Variance Inflation Factor	Classroom discussion with PPT Lab Session	RStudio Documentation Data Kaggle





MODULE 4: DECISION TREES & LOGISTIC REGRESSION

Session	Coverage of the Key Concept	Pedagogy/Activity (Discussion Points)	Reading material to be Referred
34	Decision Trees	Classroom discussion with PPT Lab Session , Board Work	RStudio Documentation
35	ID3 method	Classroom discussion with PPT Lab Session , Board Work	RStudio Documentation
36	ID3 method	Classroom discussion with PPT Lab Session , Board Work	RStudio Documentation
37	Gini Index	Classroom discussion with PPT Lab Session , Board Work	RStudio Documentation Video https://youtu.be/7VeUPuFGJHk
38	Logistic Regression	Classroom discussion with PPT Lab Session , Board Work	RStudio Documentation Video https://youtu.be/yIYKR4sgzI8
39	Akaike Information Criteria	Classroom discussion with PPT Lab Session , Board Work	RStudio Documentation Book R Cookbook Paul Teetor
40	Interpretation	Classroom discussion with PPT Lab Session , Board Work	RStudio Documentation Book R Cookbook Paul Teetor
41	Interpretation	Classroom discussion with PPT Lab Session , Board Work	RStudio Documentation Book R Cookbook Paul Teetor





42	P Value	Classroom discussion with PPT Lab Session . Board Work	RStudio Documentation Book R Cookbook Paul Tector
43	F statistic	Classroom discussion with PPT Lab Session , Board Work	RStudio Documentation Book R Cookbook Paul Tector
44	Predicting Logistic Equations	Classroom discussion with PPT Lab Session . Board Work	RStudio Documentation Book R Cookbook Paul Tector

MODULE 5: NEURAL NETWORKS

Session	Coverage of the Key Concept	Pedagogy/Activity (Discussion Points)	Reading material to be Referred
45	Artificial Neural Network	Classroom discussion Board Lab Session	RStudio Documentation Book Machine Learning: A Guide to Current Research Tom M .Mitchelle Video https://youtu.be/aircArvnKk
46	Structure of neural networks	Classroom discussion Board Lab Session	RStudio Documentation Book Machine Learning: A Guide to Current Research Tom M .Mitchelle





47	Information Flow	Classroom discussion Board Lab Session	RStudio Documentation Book Machine Learning: A Guide to Current Research Tom M .Mitchelle
48	Hidden Layers	Classroom discussion Board Lab Session	RStudio Documentation Book Machine Learning: A Guide to Current Research Tom M .Mitchelle
49	Training a Neural Network	Classroom discussion Board Lab Session	RStudio Documentation Book Machine Learning: A Guide to Current Research Tom M .Mitchelle
50	Backpropagation	Classroom discussion Board Lab Session	RStudio Documentation Book Machine Learning: A Guide to Current Research Tom M .Mitchelle
51	Interpreting a Neural Network	Classroom discussion Board Lab Session	RStudio Documentation Book Machine Learning: A Guide to Current Research Tom M .Mitchelle





52	Interpreting a Neural Network	Classroom discussion Board Lab Session	RStudio Documentation Book Machine Learning: A Guide to Current Research Tom M .Mitchelle
----	-------------------------------	-------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------

MODULE 6:INTRODUCTION TO OTHER REGRESSION ANALYSIS

Session	Coverage of the Key Concept	Pedagogy/Activity (Discussion Points)	Reading material to be Referred
53	Introduction to other regression anal	Classroom discussion, Study Material	Book Statistics For Management David S. Rubin and Levin
54	Polynomial Regression	Classroom discussion, Study Material	Book Statistics For Management David S. Rubin and Levin
55	Poisson Regression	Classroom discussion, Study Material	Book Statistics For Management David S. Rubin and Levin
56	Non Linear Regression	Classroom discussion, Study Material	Book Statistics For Management David S. Rubin and Levin
57	Non Parametric Tests	Classroom discussion, Study Material	Book Statistics For Management David S. Rubin and Levin
58	Non Parametric Tests	Classroom discussion, Study Material	Book Statistics For Management David S. Rubin and Levin



Syllabus

3.7.2. PREDICTIVE ANALYTICS USING R

1. GENERAL INFORMATION

No. of Credits: 04

No. of Hours per Week: 04

2. COURSE PERSPECTIVE:

The amount of data in the world is increasing exponentially as time passes. It is estimated that the total amount of data produced in 2020 will be 20 zettabytes (Kotov, 2014), that is, 20 billion terabytes. Many businesses invest lots of money and efforts for collecting the data and most of it is not analysed fully and / or not analysed appropriately. The main reason to analyse the data is to predict the future i.e. to construct actionable knowledge. This course will help and allows the students to do data analysis and build models while learning various tools & techniques. The prerequisite of the course is students must have undergone basic courses on Statistical modelling.

1. COURSE OBJECTIVES AND OUTCOMES OBJECTIVES

- To enable the students to be able to understand the predictive analytics in present scenario and its applications by the industry.
- Formulate the regression models for prediction

OUTCOMES

By successfully completing the course the students will be able to examine the data for model fitness and ETL process and to formulate and evaluate the prediction using regression, time series analysis, neural networks and decision tree model.

2. COURSE CONTENT AND STRUCTURE

MODULE 1: INTRODUCTION TO PREDICTIVE ANALYTICS 06 HOURS

Introduction to predictive analytics, definition, Evolution of Data Analytics, Applications of predictive analytics, Predictive models:

Propensity model, Clustering Model & Collaborative filtering; used cases on predictive analytics.

MODULE 2: EXPLORATORY DATA ANALYSIS 12 HOURS

Time Series Data Analysis : Organizing and processing of data with R, Data Cleaning - Missing values, Outlier treatment,

Pre-processing and cleaning and Univariate Analysis, ARIMA model for forecasting.

MODULE 3: PREDICTION- LINEAR REGRESSION 14 HOURS

Understanding simple regression in R, Scenarios for using OLS regression, Computing the intercept and slope coefficient, Obtaining the residuals, Computing the significance of the coefficient. Correlation & R^2 , Multiple Linear Regression in R, Model building.

MODULE 4: DECISION TREES & LOGISTIC REGRESSION 10 HOURS

Introduction to Decision trees, Data pre-processing, Model building in R, Model comparison. Introduction to Logistic Regression:

Interpreting the model parameters and assessing the impact of predictors on the probability of outcome.

MODULE 5: NEURAL NETWORKS 08 Hours

Introduction, Structure of neural networks, Information flow, Types of layers, Training a neural network, Back Propagation, Neural networks in R

Syllabus

MODULE- 6: INTRODUCTION TO OTHER REGRESSION ANALYSIS 06 HOURS

Introduction to other regression analysis Polynomial, Multiple linear, Poisson, Nonlinear and Nonparametric.

Predictive Analytics Using R – Laboratory driven course.

1. Demonstration of Reading data from files and working with datasets
2. Demonstration of Graphs: Basic high-level plots, Modifications of scatter plots, Modifications of histograms, parallel box plots
3. Demonstration of Exploratory Data Analysis :Missing values, Outlier treatment
4. Demonstration of Univariate Analysis
5. Demonstration of Time Series Data Analysis
6. Demonstration of Linear Regression
7. Demonstration of Multiple Regression
8. Demonstration of Decision Trees
9. Demonstration of Logistic Regression
10. Demonstration of Neural Networks

Note:

- For all the above Exercises Students can use their own datasets or Used Cases
- For all Exercises Writing programs, taking output print and writing Interpretation is compulsory in the Lab Journal
- In the practical Examination student are expected to execute the R program and write Interpretation for two exercises out of 10 exercises.
- Change of exercise is not permitted in the Practical Examination.

3. University Examination for 70 marks Evaluation Pattern

EXAMINER PROFILE

A person should be proficient in R and should have some experience in analytics using R. The BOE Chairperson has the discretion to select the Examiner.

2. PEDAGOGY

1. The course pedagogy includes; Lab driven sessions, Programme Writing, Lecturers, Real world Case studies, Capstone Projects, Individual and group projects, Working with Cross sectional data, Time series data and Panel data, Preparing Econometrics Model. Analyzing and interpreting the results.
2. Talk by the industry experts and industry visit.

7.TEACHING /LEARNING RESOURCES

1. Evans, J. R. (2013). Business Analytics: Methods, Models, and Decisions
2. Robert Stine, Dean Foster, "Statistics for Business: Decision Making and Analysis", Pearson Education, 2nd edition, 2013.
3. Turban, E., Aronson, J. E., Liang, T. P., & Sharda, R. (2010). Decision support and business intelligence systems (9th ed., p. 720). Prentice-Hall.
4. Berson, A., Smith, S. J., & F. (1997). Data Warehousing, Data Mining and OLAP (1st ed., p 640). Computing McGraw-Hill.
5. Han, J., & Kamber, M. (2000). Data Mining : Concepts and Techniques (1st ed., p. 550). Morgan Kaufmann
6. Robert Kabacoff, Second Edition (2015), Manning publications: R in Action Data analysis and graphics with R
7. U Dinesh Kumar, "Business Analytics" Wiley India Pvt. Ltd publication, 2017

Syllabus

8. Dr. Umesh R. Hodeghatta and Umesha Nayak, Apresspublication : Business Analytics Using R - A Practical Approach
9. Jeffrey S. Strickland, Simulation Educators (2014) Predictive Analytics using R
10. Subhashini Sharma Tripathi, Apress publication, Learn Business Analytics in Six Steps Using SAS and R

R – Programming

1. Wickham H., Grolemond G. (2016). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O' Reilly Media.
2. Cotton, R. (2013). Learning R: A Step-by-Step Function Guide to Data Analysis 1st Edition [Kindle Version]. Retrieved from <http://www.amazon.in>.
3. Knell, R. (2013) Introductory R: A Beginner's Guide to Data Visualisation, Statistical Analysis and Programming in R. [Kindle Version]. Retrieved from <http://www.amazon.in>.
4. Murray, S. (2013) Learn R in a Day. [Kindle Version]. Retrieved from <http://www.amazon.in>.

Applications of news analytics in finance: A review

Leela Mitra ^{*†} Gautam Mitra ^{*†}

June 17, 2010

Contents

^{*}CARISMA (Centre for Analysis of Risk and Optimisation Modelling Applications), Brunel University, Uxbridge, United Kingdom, UB8 3PH

[†]OptiRisk Systems, OptiRisk R&D House, One Oxford Road, Uxbridge, Middlesex, UB9 4DA UNITED KINGDOM

Abstract

A review of news analytics and its applications in finance is given in this chapter. In particular we review the multiple facets of current research and some of the major applications. It is widely recognised news plays a key role in financial markets. The sources and volumes of news continue to grow. New technologies that enable automatic or semi-automatic news collection, extraction, aggregation and categorisation are emerging. Further machine learning techniques can be used to process the textual input of news stories to determine quantitative sentiment scores. We consider the various types of news available and how these can be processed to form inputs to financial models. We report applications of news, for prediction of abnormal returns, for trading strategies, for diagnostic applications as well as the use of news for risk control.

1 Introduction

General value of news analysis for the asset management process

News (North, East, West, South) streams in from all parts of the globe. There is a strong yet complex relationship between market sentiment and news. The arrival of news continually updates investor's understanding and knowledge of the market and influences investor sentiment. There is a growing body of research literature that argues media influences investor sentiment, hence asset prices, asset price volatility and risk (Tetlock 2007, Da, Engleberg and Gao 2009, Odean and Barber 2008, diBartolomeo and Warrick 2005, Mitra, Mitra and diBartolomeo 2009, Dzielinski, Rieger and Talpsepp 2010). Traders and other market participants digest news rapidly, revising and rebalancing their asset positions accordingly. Most traders have access to newswires at their desks. As markets react rapidly to news, effective models which incorporate news data are highly sought after. This is not only for trading and fund management, but also for risk control. Major news events can have a significant impact on the market environment and investor sentiment resulting in rapid changes to the risk structure and risk characteristics of traded assets. Though the relevance of news is widely acknowledged how to incorporate this effectively, in quantitative models and more generally within the investment decision making process, is a very open question.

In considering how news impacts markets, Odean and Barber (2008) note "significant news will often affect investors' beliefs and portfolio goals heterogeneously, resulting in more investors trading than is usual" (high trading volume). It is well known volume increases on days with information releases (Bamber, Barron & Stober 1997, Karpoff 1987, Busse & Green 2002). Important news frequently results in large positive or negative returns. Ryan & Taffler (2002) find for large firms a significant portion (65%) of large price changes and volume movements can be linked to publicly available news releases. Sometimes investors may find it difficult to interpret news resulting in high trading volume without significant price change.

Financial news can be split into regular synchronous announcements (expected news) and event driven asynchronous announcements (unexpected news). Textual news is frequently unstructured, qualitative data. It is characterised as being non-numeric and hard to quantify. Unlike analysis based on quantified market data textual news data contains information about the effect of an event and the possible causes of an event. It is natural to expect that the application of this news data will lead to improved analysis (such as, predictions of returns and volatility) . However, extracting this information in a form that can be applied to the investment decision making process is extremely challenging.

News has always been a key source of investment information. The volumes and sources of news are growing rapidly. In increasingly competitive markets investors and traders need to select and analyse the relevant news, from the vast amounts available to them, in order to make "good" and timely decisions. A human's (or even a group of humans) ability to process this news is limited. As computational capacity grows, technologies are emerging which allow us to extract, aggregate and categorise large volumes of news effectively. Such technology might be applied for quantitative model construction for both high frequency trading and low frequency fund rebalancing. Automated news analysis can form a key component driving algorithmic trading desks' strategies and execution, and the traders who use this technology can shorten the time it takes them to react to breaking stories (that is, reduce latency times). News analytics (NA) technology can also

be used to aid traditional non-quantitative fund managers in monitoring the market sentiment for particular stocks, companies, brands and sectors. These technologies are deployed to automate filtering, monitoring and aggregation of news. These technology aids free managers from the minutae of repetitive analysis, such that they are able to better target their reading and research. These technologies reduce the burden of the routine monitoring for fundamental managers.

The basic idea behind these NA technologies is to automate human thinking and reasoning. Traders, speculators and private investors anticipate the direction of asset returns as well as, the size and the level of uncertainty (volatility) before making an investment decision. They carefully read recent economic and financial news to gain a picture of current situation. Using their knowledge of how markets behaved in the past, under different situations, people will implicitly match the current situation with those situations in the past most similar to the current one. News analytics seeks to introduce technology to automate or semi-automate this approach. By automating the judgement process, the human decision maker can act on a larger, hence more diversified, collection of assets. These decisions are also taken more promptly (reducing latency). Automation or semi-automation of the human judgement process widens the limits of the investment process. Leinweber(2009) refers to this process as Intelligence Amplification (IA).

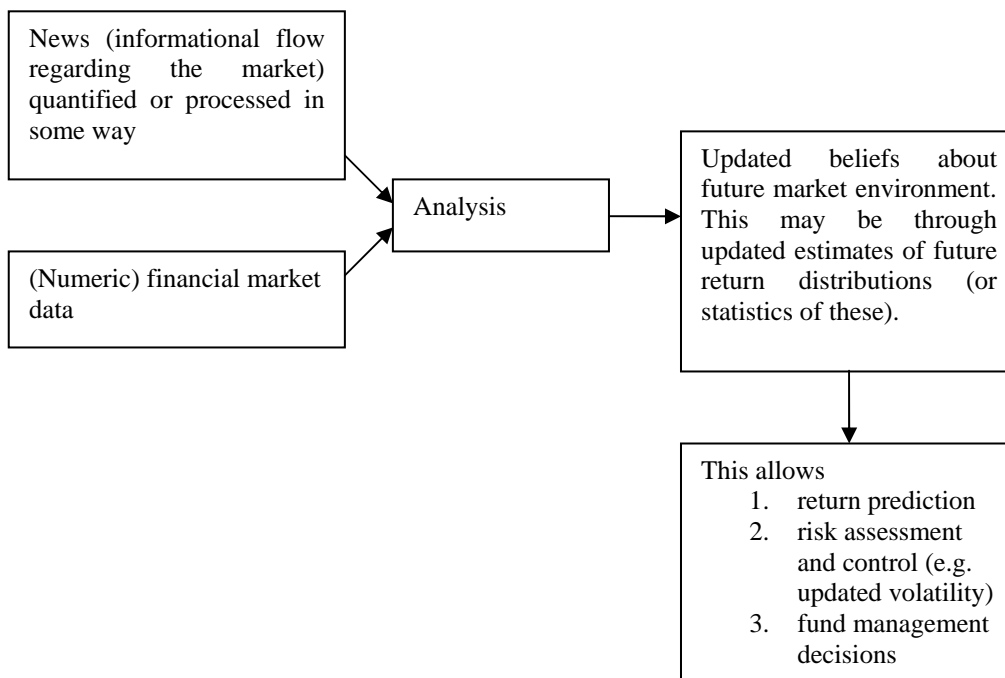


Figure 1: A simple representation of news analytics in financial decision making

As shown in Figure 1 news data is an additional source of information that can be harnessed to enhance (traditional) investment analysis. Yet it is important to recognise that NA in finance is a multi disciplinary field which draws on financial economics, financial engineering, behavioural finance and Artificial Intelligence (in particular Natural Language Processing). Expertise in these respective areas need to be combined effectively for the development of successful applications in this area. Sophisticated machine learning algorithms applied without an understanding of the structure and dynamics of financial markets and the use of realistic trading assumptions can lead to applications with little commercial use (See Mittermayer and Knolmayer 2006).

The remainder of the chapter is organised as follows. In section 2.1 we consider the different sources of

news and information flows which can be applied for updating (quantitative) investor beliefs and knowledge. Section 2.2 covers several aspects of pre-analysis to be considered when using news in trading systems and quantitative models. In section 3 we consider how qualitative text can be converted to quantified metrics which can form inputs to quantitative models. In section 4 we present news based models; In particular we consider the computational architecture (section 4.1), applications for trading and fund management (section 4.2) and applications for risk management (section 4.3). In section 4.4 desirable industry applications are outlined. The appendix ?? contains an annotated bibliography of selected papers.

2 News data

2.1 Data sources

In this section we consider the different sources of news and information flows which can be applied for updating (quantitative) investor beliefs and knowledge. Leinweber (2009) distinguishes four broad classifications of news (informational flows).

1. **News** This refers to mainstream media and comprises the news stories produced by reputable sources. These are broadcast via newspapers, radio and television. They are also delivered to traders' desks on newswire services. Online versions of newspapers may also exist.
2. **Pre-News** This refers to the source data that reporters research before they write news articles. It comes from primary information sources such as, Securities and Exchange Commission reports and filings, court documents and government agencies. It also includes scheduled announcements such as macro economic news, industry statistics, company earnings reports and other corporate news.
3. **Rumours** These are blogs and websites that broadcast "news", and are less reputable than news and pre-news sources. The quality of these vary significantly. Some may be blogs associated with highly reputable news providers and reporters (for example, Robert Peston of the BBC's blog!). At the other end of the scale some blogs may lack any substance and may be entirely fueled by rumour.
4. **Social media** These websites fall at the lowest end of the reputation scale. Barriers to entry are extremely low and the ability to publish "information" easy. These can be dangerously inaccurate sources of information. However, if carefully applied (with consideration of human behaviour and agendas) there may be some value to be gleaned from these. At a minimum they may help us identify future volatility.

Individual investors pay more attention to the second two sources of news than institutional investors. (Dzielinski, Rieger and Talpsepp 2010 and Das & Chen 2007). Information from the web may be less reliable than mainstream news. However there may be "Collective Intelligence" information to be gleaned. That is, if a large group of people have no ulterior motives, then their collective opinion may be useful (Leinweber Ch 10 2009). The SEC does monitor message boards. So there is some, though perhaps far from perfect, checking of information published. This should constrain message board posters actions to some extent.

There are services which facilitate retrieval of news data from the web. For example the Google trends is a free but limited service which provides historical weekly time series of the popularity of any given search term. This search engine reports the proportion of positive, negative and neutral stories returned for a given search.

The Securities and Exchange Commission (SEC) provides a lot of useful pre news. It covers all publicly traded companies (in the US). The Electronic Data Gathering, Analysis and Retrieval (EDGAR) system was introduced in 1996 giving basic access to filings via the web. (See <http://www.sec.gov/edgar.shtml>) Premium access gave tools for analysis of filing information and priority earlier access to the data. In 2002 filing information was released to the public in real time. Filings remain unstructured text files without semantic Web and XML output, though the SEC are in the process of upgrading their information dissemination. High end resellers electronically dissect and sell on relevant component parts of filings. Managers are obliged to disclose a significant amount of information about a company via SEC filings. This information

is naturally valuable to investors. Leinweber introduces the term “molecular search: the idea of looking for patterns and changes in groups of documents”. Such analysis/information are scrutinized by researchers/analysts to identify unusual corporate activity and potential investment opportunities. However mining the large volume of filings, to find relationships, is challenging. Engleberg and Sankaraguruswamy (2007) note the EDGAR database has 605 different forms and there were 4,249,586 filings between 1994 and 2006. Connotate provides services which allows customised automated collection of SEC filing information for customers (fund managers and traders). Engleberg and Sankaraguruswamy (2007) consider how to use a SAS web crawler to mine SEC filing information through EDGAR.

As stated in section 1, financial news can be split into *regular synchronous announcements (scheduled or expected news)* and *event driven asynchronous announcements (unscheduled or unexpected news)*. Main stream news, rumours and social media normally arrive asynchronously in an unstructured textual form. A substantial portion of pre news arrive at pre scheduled times and generally in a structured form.

Scheduled (news) announcements often have a well defined numerical and textual content, and may be classified as structured data. These include macro economic announcements and earnings announcements. Macro economic news, particularly economic indicators from the major economies are widely used in automated trading. They have an impact in the largest and most liquid markets, such as foreign exchange, government debt and futures markets. Firms often execute large and rapid trading strategies. These news events are normally well documented, thus thorough backtesting of strategies is feasible. Since indicators are released to a precise schedule, market participants can be well prepared to deal with them. These strategies often lead to firms fighting to be first to the market; speed and accuracy are the major determinants of success. However the technology requirements to capitalise on events is substantial. Content publishers often specialise in a few data items and hence trading firms often multi source their data. Thomson Reuters, Dow Jones and Market News International are a few leading content service providers in this space.

Earnings are a key driving force behind stocks’ prices. Scheduled earnings announcement information are also widely anticipated and used within trading strategies. The pace of response to announcements has accelerated greatly in recent years. (See p.104-105 Leinweber) Wall Street Horizon and Media Sentiment (see Munz 2010) provide services in this space. These technologies allow traders to respond quickly and effectively to earnings announcements.

Event driven asynchronous news streams in unexpectedly over time. These news items usually arrive as textual, unstructured, qualitative data. It is characterised as being non-numeric and difficult to process quickly and quantitatively. Unlike analysis based on quantified market data textual news data contains information about the effect of an event and the possible causes of an event. However, to be applied in trading systems and quantitative models it needs to be converted to a quantitative input time series. This could be a simple binary series where the occurrence of a particular event or the publication of a news article about a particular topic is indicated by a one and the absence of the event can be indicated by a zero. Alternatively we can try to quantify other aspects of news, over time. For example, we could measure news flow (volume of news) or we could determine scores (measures) based on the language sentiment of text or determine scores (measures) based on the market’s response to particular language.

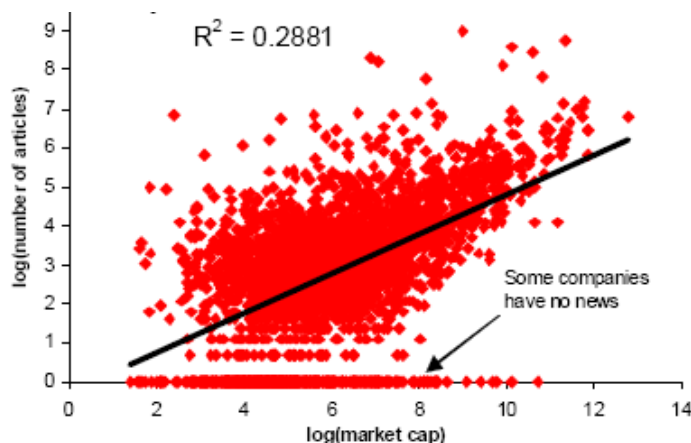
It is important to have access to historical data for effective model development and backtesting. Commercial news data providers normally provide large historical archives for this purpose. The details of historic news data for global equities provided by RavenPack and Thomson Reuters NewsScope are summarized in Appendix 1. They are taken from the RavenPack NewsScores User Guide (RavenPack 2010) and Thomson Reuters NewsScope Sentiment engine (2009).

2.2 Pre analysis of news data

Collecting, cleaning and analysing news data is challenging. Major news providers collect and translate headlines and text from a wide range of worldwide sources. For example, the Factiva database provided by

Dow Jones holds data from 400 sources ranging from electronic newswires, newspapers and magazines.

We note *there are differences in the availability of news data for different companies*. Larger companies (with more liquid stock) tend to have higher news coverage/news flow. Moniz, Brar and Davis (2009) sort companies by $\frac{\log \text{market cap}}{6 \text{ month ADV}}$. The top quintile accounts for 40% of all news articles and the bottom quintile for only 5%. Cahan, Jussa and Luo (2009) also find news coverage is higher for larger cap companies. See figure 2.



Source: RavenPack, Macquarie Capital (USA), May 2009

Figure 2: Number of news items versus log market capitalisation: Taken from Cahan, Jussa and Luo 2009

Classification of news items is important. Major newswire providers tag incoming news stories. A reporter entering a story, on to the news systems, will often manually tag it with relevant codes. Further machine learning algorithms may also be applied to identify relevant tags for a story. These tags turn the unstructured stories into a basic machine readable form. The tags are often stored in XML format. They reveal the stories' topic areas and other important meta data. For example, they may include information about which company a story is about. Tagged stories held by major newswire providers will also be accurately time stamped. The SEC is pushing to have companies file their reports using eXtensible Business Reporting Language (XBRL). Rich Site Summary (RSS) feeds (an XML format for web content) allow customised, automated analysis of news events from multiple online sources.

Tagged news stories provide us with hundreds of different types of events. We need to distinguish what types of news are relevant to our model (application). Further the market may react differently to different types of news. For example, Moniz et. al. (2009) finds the market seems to react more strongly to corporate earnings related news than corporate strategic news. They postulate that it is harder to quantify and incorporate strategic news into valuation models, hence it is harder for the market to react appropriately to such news.

Machine readable XML news feeds can turn news events into exploitable trading signals since they can be used relatively easily to backtest and execute event study based strategies. See Kothari and Warner (2005) and Campbell, Lo, and MacKinlay (1996) for in depth reviews of event study methodology. Leinweber (2010) uses Thomson Reuters tagged news data to investigate several news based event strategies. Elementised news feeds mean the variety of event data available is increasing significantly. News providers also provide archives of historic tagged news which can be used for backtesting and strategy validation. News event algorithmic trading is reported to be gaining acceptance in industry. (Schmerken 2006)

To apply news effectively in asset management and trading decisions *we need to be able to identify*

news which is both relevant and current. This is particularly true for intraday applications, where algorithms need to respond quickly to accurate information. We need to be able to identify an “information event”, that is, we need to be able to distinguish those stories which are reporting on old news (previously reported stories), from genuinely “new” news. As would be expected, Moniz et. al. (2009) finds markets react strongly when “new” news is released.

Tetlock, Saar-Tsechansky and Macskassy (2008) undertake an event study which illustrates the impact of news on cumulative abnormal returns (CAR). They use 350,000 news stories about S&P 500 companies appearing in the Wall Street Journal and Dow Jones News Service from 1984 - 2004. Each story’s (language) sentiment is determined using the General Inquirer and a story is classified as either positive or negative. The CAR for each story classification type, relative to the date of the news release is shown in Figure 3. There seems to be a connection between a news story’s release and the CAR. However, there also seems to be some “information leakage” since CAR seem to react before the date of the story’s release. Leinweber (2009) considers that this may be due to the inclusion of me-too stories that refer back to an original release of “new” news. This highlights that though textual news may have an obvious connection with returns it needs to be processed carefully and effectively.

Reuters identify relevance scores for different news articles. This measures by how much a measure of relevance the article is about a particular company. They also measure article novelty(uniquness) which determines the repetition among articles and how many similar articles there are for a particular company. RavenPack (2010) also apply machine learning techniques to extract similar pertinent information for incoming newswire stories. In particular, they distinguish stories which are events. These are stories which carry the first mention of a particular theme. Stories which are not events are excluded. This is done to minimise the number of duplicate stories.

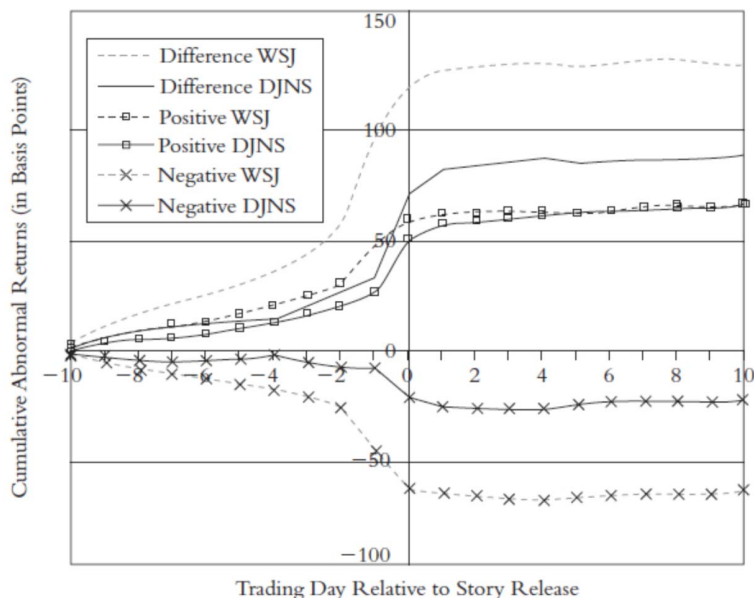


Figure 3: CAR start to respond several days before relevant news being published

Several studies also report *strong seasonality in newsflow* at hourly, daily and weekly levels. (Lo 2008 Hafez 2009, Moniz et. al. 2009) A valuable aspect of pre-analysis of news data is to identify periods of

unexpected newsflow levels, from periods of variation due to seasonality, in order to identify periods where significant levels of information are flowing into the market. Hafez (2009) investigates the seasonality patterns of news arrival. Figure 4 shows the intra day pattern. He notes that larger volumes of newsflow arrive just before the opening of the European, US and Asian trading sessions. On the intra week level we can see little newsflow takes place on the weekends. In the week, the peak of newsflow occurs on Wednesday and Thursday, while the trough falls on Friday. Lo also notes that the median number of weekday Reuters news alerts falls between 1,500 and 2,000, while the median for the entire weekend is 130.

The *time of the day when news is released*, has also been found to be relevant in understanding the connection between market variables and news. Robertson, Geva and Wolff (2006) find that there is a greater likelihood of events that lead to rising volatility at the start of the day. Boyd, Hu, and Jagannathan (2005) find that *market conditions* can influence the types of news that are reported. They report that interest rate information dominates in expansionary periods. In contrast information about future corporate dividends dominates when the markets are contracting.

As would be expected the *informational content of news* has a large influence on how markets react to news (Blasco, Corredor, Del Rio and Santamaria 2005, Boyd, Hu, and Jagannathan 2005 Liang 2005 and Tetlock 2007). We discuss how to extract the informational content of news (that is the sentiment) in section 3. It has been recognised that stock returns react more strongly to “negative” news than “positive” (Tetlock 2007). There also tends to be a positive sentiment bias, that is there is a larger volume of “positive” news to “negative” news. Das and Chen (2007) find that a histogram of normalised stock message board sentiment is positively skewed. There are days when messages about a stock are extremely optimistic but there is not a similar level of expression of pessimistic views. RavenPack (2010) also find a positive sentiment bias in their sentiment classifiers. This bias is more marked in bull markets than bear markets. They report a ratio of 2:1 of positive sentiment to negative sentiment stories in bull markets.

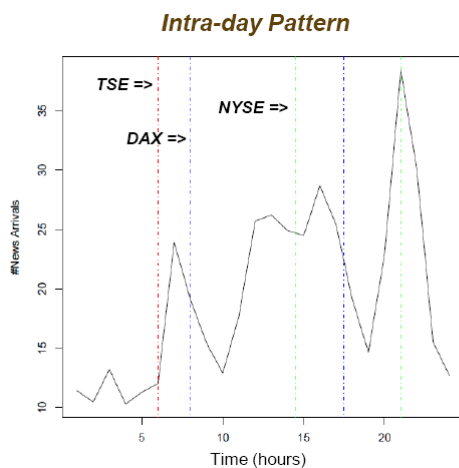


Figure 4:

The relationship of different news stories to each other is also an important consideration. Companies may make several announcements that fall under different classifications on the same day. These may or may not be related and may be related to varying degrees. For example a company may announce a profit warning, resignation of their CEO and provide guidance on its sales outlook. The dependence or independence between different news stories is a consideration.

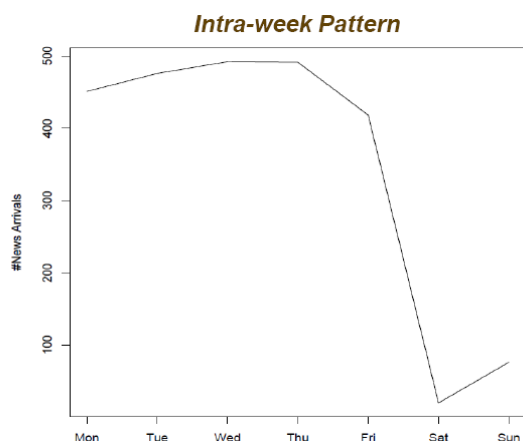


Figure 5:

3 Turning qualitative text into quantified metrics and time series

A salient aspect of news analysis is to discover the *informational content of news*. Converting qualitative text into a machine readable form is a challenging task. We may wish to distinguish whether a story’s informational content is positive or negative, that is, determine its sentiment. We may go further and try to identify “by how much” the story is positive or negative. In doing this we may try to assign a quantified sentiment score or index to each story. A major difficulty in this process is identifying the context in which a story’s language is to be judged. Sentiment may be defined in terms of how positively or negatively a human (or group of humans) interprets a story. That is the emotive content of the story for that human. In particular standards can be defined using experts to classify stories. Some of RavenPack’s classifiers are calibrated using finance experts to define the context of language. Further dictionary based algorithms which use psychology based interpretations of words may be used. Since different groups of people are effected by events differently and have different interpretations of the same events, conflicts may arise. Moniz et. al. (2009) gives an example of the term “dividend cuts”. This may be classified as a negative term by a dictionary based algorithm. In contrast, it may be interpreted positively by market analysts who may believe this indicates the company is saving money and more able to repay its debts. Loughran and McDonald (2010) also consider how context effects how the tone of text is interpreted. They note a Psychological dictionary like the Harvard-IV-4 may classify words as negative when they do not have a negative financial meaning. They develop an alternative negative word list that better reflects the tone of financial text.

An attractive alternative is to use market based measures to interpret and define the importance of news. The markets’ relative change in returns or volatility, for a particular asset or asset class lagged against a relevant news story, can be used to define the sentiment (informational content) of the news story. This approach intrinsically assumes that the market has responded to the news story. Lo (2008) uses this approach for creating the Reuters Newscope Event Indices. He creates separate indices for market responses to news, in terms of returns and volatility. So he assumes that sentiment measured in the context of these two variables are different. This approach is quite pragmatic and is focussed on using the news content directly in the context that the modeller is interested in. Lavernko, Schmill, Lawrie, Ogilvie, Jensen and Allan (2000), Moniz et. al. (2009), Peramunetilleke and Wong (2002) and Luss and d’Aspremont (2009) also use market based measures in determining the “sentiment” of news. SemLab (see Vreijling 2010) provide a

tool which allows the user to filter news items and examine each item's impact on market variables. Using this interactive tool, the user is able to define their own tailored context of "sentiment".

Given a definition of sentiment, machine learning and natural language techniques are frequently used to determine the sentiment of new incoming stories. Hence we can determine sentiment scores over time as news arrives. Such sentiment indices then allow us to develop systematic investment and risk management processes. Once a sentiment index is constructed, to use it effectively, we must be able to find evidence of a relationship with relevant asset returns, trading volumes or volatility.

The definition of market sentiment is very much context dependent. In general we are interested in discovering the "informational context of news" In this review paper for the purpose of (quantitative) modelling applications we use these two terms "news sentiment" and "informational content of news" interchangeably, and in this section we discuss some of the leading methods of computing / quantifying "sentiment" and other related measures.

We review below Das and Chen (2007) and Lo (2008). Both paper cover the following items.

1. A definition of the context of sentiment is given.
2. Application of algorithms (natural language, machine learning and linear regression) to calibrate and define sentiment scores.
3. Validation of the effectiveness of the scores by comparing their relationship with relevant asset returns, volumes or volatility.

Das and Chen (2007) use statistical and natural language techniques to extract investor sentiment from stock message boards and generate sentiment indices. They apply their method for 24 technology stocks present in the Morgan Stanley High Tech (MSH) Index. A web scraper program is used to download tech sector message board messages. Five algorithms, each with different conceptual underpinnings are used to classify each message. A voting scheme is then applied to all five classifiers.

Three supplementary databases are used in the classification algorithms.

1. "*Dictionary*" is used for determining the nature of the word. For example, is it a noun, adjective or adverb?
2. "*Lexicon*" is a collection of hand picked finance words which form the variables for statistical inference within the algorithms.
3. "*Grammar*" is the training corpus of base messages used in determining the in-sample statistical information. This information is then applied for use on the out-of-sample messages.

The lexicon and grammar jointly determine the context of the sentiment. Each of the classifiers relies on a different approach to message interpretation. They are all analytic, hence computationally efficient.

1. *Naive classifier* (NC) is based on a word count of positive and negative connotation words. Each word in the lexicon is identified as being positive, negative or neutral. A parsing algorithm negates words if the context requires it. The net word count of all lexicon matched words is taken. If this value is greater than one, we sign the message as a buy. If the value is less than one the message is a sell. All others are neutral.
2. *Vector distance classifier* Each of the D words in the lexicon is assigned a dimension in vector space. The full lexicon then represents a D -dimensional unit hypercube and every message can be described as a word vector in this space ($m \in \mathbb{R}^D$). Each hand tagged message in the training corpus (grammar) is converted into a vector G_j (grammar rule). Each (training) message is pre classified as positive,

negative or neutral. We note that Das and Chen use the terms Buy/Positive, Sell/Negative and Neutral/Null interchangeably. Each new message is classified by comparison to the cluster of pretrained vectors (grammar rules) and is assigned the same classification as that vector with which it has the smallest angle. This angle gives a measure of closeness.

3. *Discriminant based classification* NC weights all words within the lexicon equally. The discriminant based classification method replaces this simple word count with a weighted word count. The weights are based on a simple discriminant function (Fisher Discriminant Statistic). This function is constructed to determine how well a particular lexicon word discriminates between the different message categories ($\{ \text{Buy, Sell, Null} \}$). The function is determined using the pre classified messages within the grammar. Each word in a message is assigned a signed value, based on its sign in the lexicon multiplied by the discriminant value. Then as for NC a net word count is taken. If this value is greater than 0.01, we sign the message as a buy. If the value is less than -0.01 the message is a sell. All others are neutral.
4. *Adjective - adverb phrase classifier* is based on the assumption that phrases which use adjectives and adverbs emphasize sentiment and require greater weight. This classifier also uses a word count but uses only those words within phrases containing adjectives and adverbs. A “tagger” extracts noun phrases with adjectives and adverbs. A lexicon is used to determine whether these significant phrases indicate positive or negative sentiment. The net count is again considered to determine whether the message has negative or positive overall sentiment.
5. *Bayesian classifier* is a multi variate application of Bayes Theorem. It uses the probability a particular word falls within a certain classification and is hence indifferent to the structure of language. We consider three categories $C = \{ c_i \mid i = 1, \dots, C \}$. Denote each message $m_j \quad j = 1, \dots, M$. The set of lexical words is $F = \{ w_k \}_{k=1}^D$. (The total number of lexical words is D) We can determine a count of the number of times each lexical item appears in each message $n(m_j, w_k)$. Given the class of each message in the training set we can determine the frequency with which a lexical word appears in a particular class. We are then able to compute the conditional probability of an incoming message j falling in category i , $Pr(m_j|c_i)$, from the word based frequencies. $Pr(c_i)$ is set to the proportion of messages in the training set classified in class c_i . For a new message we are able to compute the probability it falls within class c_i given its component lexicon words, that is $P(c_i|m_j)$, through an application of Bayes Theorem. The message is classified as being from the category with the highest probability.

A voting scheme is then applied to all five classifiers. The final classification is based on achieving a majority amongst the five classifiers. If there is no majority the message is not classified. This reduces the number of messages classified but enhances the classification accuracy.

Das and Chen also introduce a method to detect message ambiguity. Messages posted on stock message boards are often highly ambiguous. The grammar is often poor and many of the words do not appear in standard dictionaries. They note “Ambiguity is related to the absence of “aboutness””. The General Inquirer has been developed by Harvard University for content analyses of textual data. They use it to determine an independent optimism score for each message. By using a different definition of sentiment it is ensured there is no bias to a particular algorithm. The optimism score is the difference between the number of optimistic and pessimistic words as a percentage of the total words in the body of the text. This score allows us to rank the relative sentiment of all stories within a classification group. For example, they can rank the relative optimism of all stories which have been classified by their scheme as positive. The mean and standard deviation of the optimism score for different classification types ($\{ \text{Buy, Sell, Null} \}$) can be calculated. They filter *in* and consider only highly optimistically scored stories in the positive category. For example only those stories with optimism scores above the mean value plus one standard deviation are considered. Similarly they filter *in* and consider only the most highly pessimistic scores in the negative category. Once the classified stories are further filtered for ambiguity, it is found that the number of false positives dramatically declines.

Once the sentiment for each message is determined using the voting algorithm, a daily sentiment index is compiled. The classified messages up to 4pm each day are used to create the aggregate daily sentiment for each stock. A buy (sell) message increments (decrements) the index by one. These indices are further aggregated across all stocks to obtain an aggregate sentiment for the technology portfolio. A disagreement measure is also constructed

$$DISAG = \left| 1 - \left| \frac{B - S}{B + S} \right| \right| \quad (1)$$

B (S) is the number of buy (sell) messages. This measure lies between 0 (no disagreement) and 1 (high disagreement) and is computed as a daily time series. The daily MSH index and component stock values are also collected. Trading volatility and volume of stocks are also calculated and message volume is also recorded. All the time series are normalised.

Das and Chen check that the constructed sentiment indices have a relationship with relevant asset variables. The relationship between the MSH index and the aggregate sentiment index is investigated. Fig 2 plots the two against each other. These two series do seem to track each other. The sentiment index is found to be highly autocorrelated out to two trading weeks. Regression analysis is undertaken to investigate the relationship. They conclude sentiment does offer some explanatory power for the level of the index. However, the autocorrelation makes it difficult to establish the empirical nature of the relationship.

Das and Chen undertake regression analysis between the individual stock level and the individual stock sentiment level and find there is a significant relationship. (t-statistic of coefficient falls within a significant level) The relationship between first differences is much weaker. We cannot conclude there is a strong predictive ability on forecasting individual stock returns. Sentiment and stock levels are not unrelated, but determining the precise nature of the relationship is difficult.

Fig 4 summaries the relationship between the sentiment measure, disagreement measure, message volume, trading volume and volatility. Sentiment is inversely related to disagreement. As disagreement increases the sentiment falls. Sentiment is correlated to high posting volume. As discussion increases this indicates optimism about that stock is rising. There is a strong relationship between message volume and volatility. This is consistent with Antweiler and Frank (2002). Trading volume and volatility are strongly related to each other.

Lo (2008) develops the Reuters NewsScope Event Indices (NEI) which are constructed to have “predictive” power for particular asset returns and (realised) volatility. They are constructed in an integrated framework where news, returns and volatility are used in calibrating the indices. The white paper (dated November 2007) considers specifically indices for foreign exchange. However, the method can be applied to other asset classes.

Lo uses news alerts in developing his sentiment indices. These are quick news flashes which are issued when a newsworthy event occurs. They are both timely and relevant. An example of Reuters NewsScope Alert

TimeStamp 02 AUG 2007 04:44:26.155

Alert Tsunami Warning Issued for Japan’s Western Hokkaido Coast

Tags JP ASIA NEWS DIS LEN RTRS

The alerts comprises three items (i) TimeStamp (ii) A short headline and (iii) Tags and meta data. The tags are machine readable and will often contain information about the topic area. The headlines lend themselves well to machine analysis since they are concise and formed from a small vocabulary. Lo notes the purpose of the indices is to rapidly identify and report market moving information. Once constructed he undertakes (event study) experiments to validate their quality, developing metrics which have the potential to indicate whether the indices are able to predict significant market movements.

Framework for real time news analytics

We consider here the framework for developing the Reuters NEI. For a given asset class and related topic area the following parameters are used.

- (1) List of keywords and phrases with real valued weights; $(W_1, \gamma_1), \dots, (W_k, \gamma_k)$.
- (2) A rolling “sentiment” window of size r (say 5/10 minutes).
- (3) A rolling calibration window of size R (say 90 days).

Initially a *raw score* is created.

We have $(W_1, \gamma_1), \dots, (W_k, \gamma_k)$, where W_1 is the first keyword and γ_1 is the weighting for the first keyword.

The raw score at time t is assigned by considering the time period $(t - r, t]$. (w_1, \dots, w_k) is the vector of keyword frequencies in $(t - r, t]$, that is, w_i is the number of times keyword W_i occurred in the last r minutes. The raw score is defined as

$$s_t \equiv \sum_i \gamma_i w_i \quad (2)$$

The raw score will tend to be high when the news volume is high. A *normalised score* is therefore produced using the rolling calibration window. At all times t for the R days in the calibration window, we record

- (i) the raw score s_t that would have been assigned,
- (ii) the news volume; $n_{[t-r,t]}$ the number of words that were observed in the time interval $[t - r, t)$.

The normalised score is determined by comparing the current raw score against the distribution of raw scores in the calibration window, where the news volume equalled the current news volume. This means we only consider those raw scores where the news volume equals the current news volume.

$$S_t \equiv \frac{|\{t' \in [t - R, t) : n_{[t'-r,t']} = n_{[t-r,t]} \& s_{t'} < s_t\}|}{|\{t' \in [t - R, t) : n_{[t'-r,t']} = n_{[t-r,t]}\}|} \quad (3)$$

We notice the numerator is a subset of the denominator, hence $S_t \leq 1$. If $S_t = 0.92$, we can say 92% of the time when news volume is at the current level, the raw score is less than it currently is. Lo creates an alternative score based on topic codes. Instead of counting word frequencies, the fraction of news alerts (in the last r minutes) tagged with particular topic codes, are used.

Naturally the scoring method is dependent on the list of keywords/topic areas (W_1, \dots, W_k) and the real valued weights $(\gamma_1, \dots, \gamma_k)$. The lists of keywords/topics were created by selecting the major news categories that related to the asset class (foreign exchange) and creating lists, by hand, of words and topic areas that suggest news relevant to the categories. A tool was created to extract news from periods where high scores were assigned. This news was then manually inspected, so that the developer could determine whether the keywords (topics) were legitimate or needed adjusting.

The optimal weights $(\gamma_1, \dots, \gamma_k)$ for the intraday return sentiment index were determined by regressing the word (topic) frequencies against the intraday asset returns. Similarly the (optimal) weights for the intraday volatility sentiment index were determined by regressing the word (topic) frequencies against the intraday (de-seasonalised) realised volatility. Volatility was observed to show strong seasonality on intraday timescales, hence this series was de-seasonalised prior to derivation of the weights. Returns did not exhibit any seasonality. The time series are given on an intraday basis, hence to keep the data manageable a random subset of the observations are used in calibration. Lo notes the determination of the weights can be expressed as a more general classification problem. Other techniques might be applied, in particular machine

learning algorithms such as the perceptron algorithm or support vector machines. He suggests further study is required to find the best approach, but the standard linear regression approach does perform well.

To establish that the final NEI have empirical significance, Lo undertakes detailed event study analysis. He uses the NEI series to define an event. An event is defined to take place when the index exceeds a certain threshold (say 0.995). He then removes any events that follow in less than one hour of another event. This guards against identifying “new” events which are actually based on old news. The behaviour of exchange rates before and after these events are then studied. Two time series are considered; the log returns and the deseasonalised squared log returns. He then tests the null hypothesis that the distribution of log returns / deseasonalised squared log returns are the same before and after the events. He uses samples of one hour centered on the events.

We can visually assess the impact of events on the volatility of EUR/USD exchange rate.

(1) Figure 6 shows the averaged volatility event window. The pre event (averaged) volatility is shown in blue, and the post event (averaged) volatility is shown in red. There is a peak at the centre where there is a significant increase in volatility.

(2) Figure 7 shows the density function of pre event samples and post event samples of deseasonalised squared log returns. The shift to the right indicates an upward shift in volatility on average.

As well as visual inspection, statistical tests can be introduced to compare the pre and post event samples. A t-test can be used to test equality of the means in the two samples. Levene’s test can be used to determine whether there has been a change in the standard deviation. The χ^2 goodness of fit test can be used to determine whether the two samples are likely to have come from different distributions.

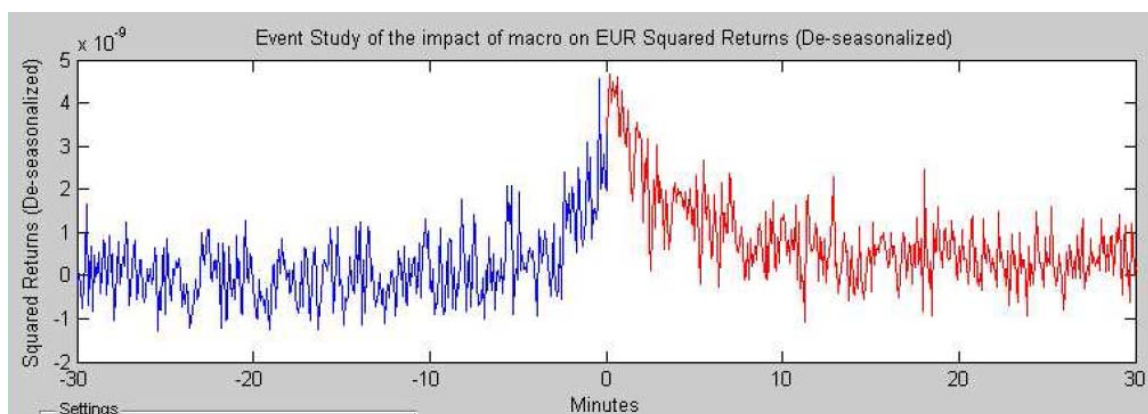


Figure 6: Pre and post event squared returns

Indices and FX Implied Volatility

Lo finds that the event studies confirm the constructed event indices, on average, impact the *realised* foreign exchange volatility. He further considers the relationship of the indices to *implied* volatility. The NEI volatility indices are constructed to predict volatility over 30 minute periods. Implied volatility gives the markets’ expectations of volatility over a much longer horizon, typically 30 days. Event study analysis

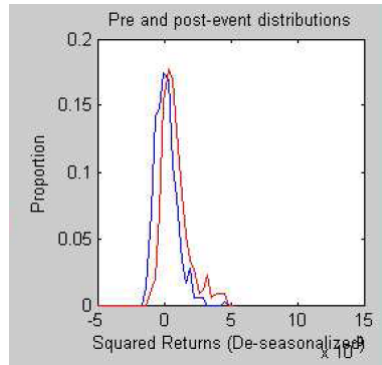


Figure 7: Distribution of pre and post event squared returns

between implied volatility and the NEI volatility indices shows no evidence of a relationship. Lo feels that implied volatility and the indices may function as complementary sources of information for risk management, since they intrinsically focus on different time horizons.

4 Models and applications

News Analytics in finance is the use of technology and algorithms to process news, within the investment management process. It allows investors to update their beliefs about the future market environment more effectively. This technology may be geared towards human decision support or it may be used to create automated quantitative strategies. The use of news data in addition to historic market data makes models more proactive and less reactive. The applications broadly fall into two areas; trading and risk control.

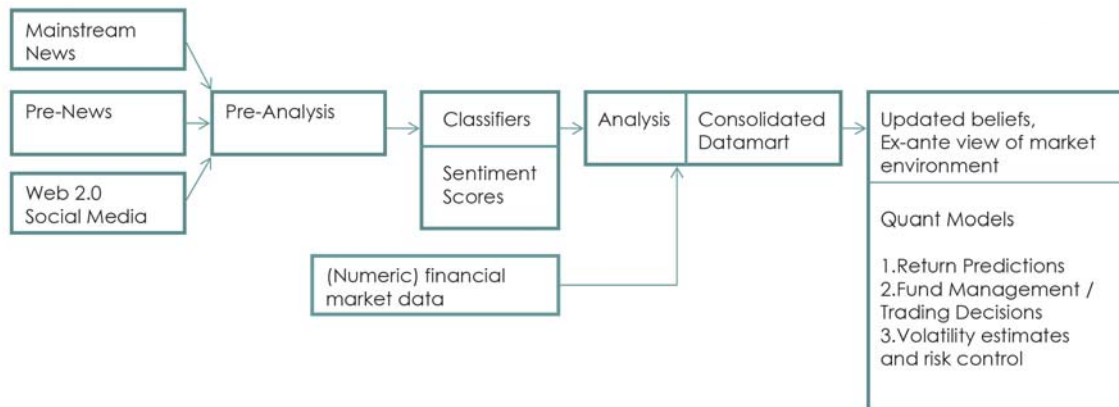


Figure 8: Information flow and computational architecture

4.1 Information flow and computational architecture

News analytics in finance focus on improving IT based legacy system applications. These improvements come through research and development directed to automating/semi automating programmed trading, fund rebalancing and risk control applications.

The established good practice of applying these analytics in the traditional manual approach are as follows. News stories and announcements arrive synchronously and asynchronously. In the market assets' (stocks, commodities, FX rates, etc) prices move (market reactions). The professionals digest these items of information and accordingly make trading decisions, investment decisions, recompute their risk exposures.

The information flow and the (semi) automation of the corresponding IS architecture is set out in figure 8. There are two streams of information which flow simultaneously, news data and market data. Pre-analysis is applied to news data; it is further filtered and processed by classifiers to relevant metrics. This is consolidated with the market data of prices together they constitute the classical datamart which feed into whatever relevant model based applications are sought. A key aspect of these applications is that they set out to provide technology enabled support to professional decision makers thereby achieve intelligence amplification (Leinweber 2009).

4.2 Trading and fund management

Generally traders and quantitative fund managers seek to identify and exploit asset mispricings, before they correct, in order to generate alpha. Most simply they may use (quantified) news data to *rank stocks* and identify which stocks are relatively attractive (unattractive). They may then buy (sell) the highest (lowest) ranking stocks, thereby rebalancing a portfolio composed of desired weights on the selected stocks. Similarly the news data may be used to identify trading signals for particular stocks. Alternatively analysts may use *factor models* to process new sources of news data. (Factor models, which are applied to give updated estimates of future asset returns and volatility, allow us to determine an optimal future portfolio to hold. That is, they tell us which assets to hold and also in what proportions.) Analysts may also use news data to identify and exploit *behavioural biases* in investor behaviour arising due to the market and analysts' misreaction to new information. In particular this can arise due to delayed information diffusion or due to investors' inattention and limited ability to process all relevant information instantaneously.

Stock picking and ranking

Li (2006) uses a simple ranking procedure to identify stocks with positive and negative (financial language) sentiment. He examines form 10-K Securities and Exchange Commission (SEC) filings for non-financial firms between 1994-2005. He creates a "risk sentiment measure" which is formed by counting the number of times the words risk, risks, risky, uncertain, uncertainty and uncertainties occur within the management discussion and analysis sections. A strategy which goes long in stocks with a low risk sentiment measure and short stocks with a high risk sentiment measure is found to produce a reasonable level of returns. Leinweber (2009) notes it is rumoured similar approaches are being applied. The performance of the strategy deteriorates in recent years, possibly due to wider use of such strategies.

Moniz et. al. (2009) focuses on turning news signals into a trading strategy. Equity analysts collect, process and disseminate information on companies to investors. In particular they use their research to form earnings forecasts for companies. Earnings momentum strategies, thus proxy for corporate newsflow. Moniz notes these strategies do not explicitly identify the piece of information that has triggered the change in earnings forecast. He investigates whether news leads earnings revisions. He finds news data can be used to reinforce proxies for news already incorporated in models and a strategy based on earnings momentum reinforced by newsflow is found to be effective.

Event studies based on news events can also allow fund managers to identify potentially under/over priced stocks. See the discussion in section 2.2.

Factor models

The Efficient Market Hypothesis (EMH) asserts that financial markets are “informationally efficient” so prices of traded assets reflect all known information and update instantaneously to reflect new information. Further it is assumed that agents act rationally. It is widely accepted, within the fund management and trading community, that the EMH, particularly in its strong form, does not hold. In the long term markets may be efficient. But “The long run is a misleading guide to current affairs. In the long run we are all dead.” as John Maynard Keynes said. In the shorter term traders and quantitative fund managers seek to identify and exploit asset mispricings, before these prices correct themselves, in order to generate alpha. In undertaking this process they often seek to gain a competitive advantage by applying improved and differentiating sources of data and information.

The capital asset pricing model (CAPM) is the classical approach to pricing equities (Sharpe 1964 and Lintner 1965). Any asset’s return can be split into a component that is correlated with the market’s return and a residual component that is uncorrelated with the market. Under the CAPM, it is assumed that, the expected return for the residual component is zero and any stock’s expected return is dependent only on the expected return of the market. The CAPM states that only risk (uncertainty) due to market variability should be priced. Residual risk can be diversified and therefore should not be compensated.

The arbitrage pricing theory (APT) (introduced by Ross (1976)) extends the CAPM to a more general linear model where additional sources of information to market returns are considered. Under the APT (multifactor models) an asset’s expected return is represented as a linear sum of several “risk” (uncertainty) factors that are common to all assets and an asset specific component. The APT states the investor should be compensated for their exposure to all sources of (non-diversifiable) risk.

Active portfolio managers seek to incorporate their investment insight to “beat the market”. An accurate description of asset price uncertainty is key to the ability to outperform the market. Tetlock, Saar-Tsechansky and Macskassy (2008) note that an investor’s perceptions about the future asset returns are determined by their knowledge about the company and its prospects, that is, by their “information sets”. They note that these are determined from three main sources: analysts forecasts, quantifiable publicly disclosed accounting variables and linguistic descriptions of the firm’s current and future profit generating activities. If the first two sources of information are incomplete or biased, the third may give us relevant information for equity prices.

Multi factor models are now widely used by fund managers in constructing alpha generating strategies (Rosenberg, Reid and Lanstein 1985). Identifying the relevant factors (and betas) is a measure of skill. Fund managers are always seeking new sources of advantage. This can be data and factors which translate to “quantitative knowledge”. “Profits may be viewed as the economic rents which accrue to [the] competitive advantage of ... superior information, superior technology, financial innovation” (Lo 1997). A “quantcentration” effect is frequently observed. That is, since most fund managers have access to the same sources of data, it is difficult to distinguish between their models and performance. Cahan, Jussa and Luo (2009) find that news sentiment scores provided by RavenPack act as an orthogonal factor to traditional quantitative factors currently used. Hence they add a diversification benefit to traditional factor models. In particular they note the value of this source of information during the credit crisis, when determining fundamentals (which traditional quant factors are based on) was problematic.

Behavioural biases

Behavioural economists challenge the assumption that market agents act rationally. Instead they propose that individuals display certain biased behaviour, such as, loss aversion (Kahneman and Tversky 1979), overconfidence (Barber and Odean 2001), overreaction (DeBondt and Thaler 1986) and mental accounting (Tversky and Kahneman 1981). Due to individual behavioural biases investors systematically deviate from optimal trading behaviour (Daniel, Hirshleifer and Teoh 2002, Hirshleifer 2001, Odean and Barber 1998). Behavioural economists use these biases to explain abnormal returns, rather than risk based explanations.

Naturally investor behaviour is dependent on individual and group psychology. Some of the research within behavioural finance seeks to understand the mechanisms of human investor behaviour, drawing heavily on the fields of neuroscience and psychology (See for example Peterson 2007). Lo (2004) proposes a new framework the Adaptive Market Hypothesis (AMH) which seeks to reconcile market efficiency with behavioural alternatives. This is an evolutionary model, where individuals adapt to a changing environment via simple heuristics.

As noted before the relationship between news and markets is complex. A number of studies consider how investors react to news releases, in particular, the behavioural and cognitive biases in their reactions to news. Quantitative investors often seek to systematically exploit the anomalies observed in prices arising from investors' behavioural biases (Moniz et. al. 2009 Barber and Odean 2007 Seasholes and Wu 2004). There is a commercial fund called MarketPsy which employs strategies that exploit "collective investor misbehaviour" (see <http://marketpsy.com/>).

Barber and Odean (2007) consider evidence for the behavioural bias that individual investors have a tendency to buy attention-grabbing stocks. Attention grabbing stocks are defined as ones that display abnormal trading volumes, extreme one day returns or are mentioned on the Dow Jones New Service. In contrast professional managers who are better equipped to assess a wider range of stocks are less prone to buying attention grabbing stock. In particular institutional investors, who use computers to manage their searches, normally specialise in a particular sector and may consider only those stocks that meet certain criteria. For every buyer there must be a seller. So if one group incurs losses the other group profits. If individual investors fail to react appropriately to news and attention, there is scope for institutional investors to profit. Seasholes and Wu (2004) find individual investors tend to buy stocks that hit an upper price limit. They find an impact on the prices of these attention grabbing stocks, which reverses to pre event levels within ten working days. Further they find a group of professional investors who profit from the biased behaviour of the individual investors.

Fang and Peress (2009) consider whether media coverage can help predict the cross-section of future stock returns. They find stocks with no media coverage outperform widely covered stocks even after allowing for well-known risk factors. This is contrary to the findings of Barber and Odean (2007). But this finding supports Merton's (1987) investor recognition hypothesis. Da, Engleberg and Gao (2009) also consider how the amount of attention a stock receives affects its cross-section of returns. They use the frequency of Google searches for a particular company as a measure of the amount of attention a stock receives. They find some evidence that changes in investor attention can predict the cross-section of returns. This is most pronounced amongst the small cap stocks.

Some researchers consider how informational flows cause investors to update their expectations, in order to explain momentum and reversal effects. DeBondt and Thaler (1986) suggest that investors overreact to recent earnings placing less emphasis on long term averages. Daniel, Hirshleifer, and Subrahmanyam (1998) suggest price momentum is a result of investors overacting to private information causing prices to be pushed away from fundamentals. In contrast Hong and Stein (1999) suggest price momentum occurs due to investors underreacting to new information. They suggest information diffuses slowly and is gradually incorporated into prices. Hirshleifer, Lim and Teoh (2010) find that when there is a significant number of earnings announcements in the market, investors are distracted and underact to relevant new information and the post announcement drift is strong. Investors fail to price the information efficiently, leaving an opportunity for quantitative investors. Scott, Xu and Stumpp (2003) conclude that price momentum is caused by under reaction of stocks to earnings related news. This is contrary to prior literature which suggested that price momentum was connected to trading volume.

Chan (2003) finds stocks with major public news exhibit momentum over the following month. In contrast stocks with large price movements, but an absence of news, tend to show return reversals in the following month. This would support a trading strategy based on momentum reinforced with news signals. Da, Engleberg and Gao (2009) extend their analysis of Google searches to consider the debate on how momentum works. They find price momentum is stronger in stocks with high levels of Google (SVI) searches. This

supports Daniels et al. (1998) view since one would expect investors to overreact to stocks they are paying close attention to. Gutierrez and Kelley (2007) Hou, Peng and Xiong (2009) also investigate the relationship between news(information flows) and momentum.

4.3 Monitoring risk and risk control

For effective financial risk control companies need to identify, understand and quantify potential (adverse) outcomes, their related probabilities and the severity of their impacts. This knowledge allows them to assess how best to manage and mitigate risk. Traditionally historic asset price data has been used to estimate risk measures. These traditional approaches have the disadvantage that they provide ex post retrospective measures of risk. They fail to account for developments in the market environment, investor sentiment and knowledge. Incorporating measures or observations of the market environment within the estimation of future portfolio return distributions is important, since the market conditions are likely to vary from historic observations. This is particularly important when there are significant changes in the market. In these cases risk measures, calibrated using historic data alone, fail to capture the true level of risk (See Mitra, Mitra and diBartolomeo 2009 and diBartolomeo and Warrick 2005). Recent technological developments have enabled the creation of data-mining tools that can interpret live news feeds. (See section 3 also RavenPack 2010; Brown/Thomson Reuters 2010; Vreijling/SemLab 2010) Mitra, Mitra and diBartolomeo 2009 find that updating risk estimates using news data can provide dynamic (adaptive) measures that account for the market environment. Further these measures may be useful in identifying and giving early ex-ante warning of extreme risk events.

The risk structure of assets may change over time, in response to news. Patton and Verardo (2009) investigate whether the systematic risk (beta) of stocks increases in response to firm specific news (in the form of earnings announcements). They undertake an event study on the beta of stocks around their earnings announcement dates. The change in beta on announcement date is decomposed into the change due to an increase in volatility of that stock and the change due to an increase in the covariance with the index. They find that news releases do have an important impact on the risk of stocks. Further much of the beta increase arises from an increase in covariance with other stocks. This suggests there could be a contagion effect in the information releases for one stock on the price movements of other stocks. This supports anecdotal evidence that investors will monitor earnings of related stocks when investigating the earnings of a particular stock. They suggest the credit crisis (2008) could be viewed as a negative earnings surprise for the market. Correlations were observed to increase during this period.

The relationship between public information release and asset price volatility has been widely investigated and noted. Ederington and Lee (1993) find a relationship between macroeconomic announcements and foreign exchange and interest rate futures return volatility. Graham, Nikkinen and Sahlstrom (2003) find stock prices on S&P500 are also influenced by macro economic announcements. Kalev, Liu, Pham, and Jarnecic (2004) find that a GARCH model for equity returns which incorporates asset specific news gives improved volatility forecasts. This study is extended in Kalev and Duong (2010). Robertson, Geva and Wolff (2007) also consider a GARCH model which accounts for “content aware” measures of news.

It is observed that volatility is higher in down markets. This is sometimes referred to as the “*leverage effect*”. Dzielinski et al. (2010) refer to it as *volatility asymmetry*. Their investigation concludes it is likely to be driven by the over reaction of private investors to bad news. In line with this theory, they find an increase in private investor attention to negative news can predict a rise in volatility. Increased private investor attention to negative news, is measured by a change in the level of Google searches for negative words related to the macro economy, such as recession.

The relationship between equity price volatility and web activity has also been widely investigated. Wysocki (1999) finds that spikes in Yahoo! Message board activity are good predictors of equity volatility (also volume and excess returns). Antweiler and Frank (2004) also have similar findings for equity volatility. An application for traffic analysis from the web was developed by Codexa for Bear Wagner to aid their risk

management strategy in predicting (unexpected) high volatility (Leinweber (2009) Ch10 p.237).

As discussed in section 3 Lo (2008) creates event indices (scores) that are constructed to predict changes in (foreign exchange) volatility. Empirical event studies show these are effective at converting incoming qualitative text (textual news) into quantitative signals that do indicate changes in volatility.

News data (flows) can also be used for non-quantitative risk control. Wolf detectors (circuit breakers) are a risk control feature for algorithmic trading built on machine readable news. Essentially they “break the circuit” stopping an automated algorithm from trading on a certain asset when particular types of news are released. It is important to try not to shout “Wolf!” when no wolf has actually appeared. These risk control features can be customized to only be tripped when substantive news events have occurred. Alternatively the algorithms can be turned back after the nature of the news has been programmatically analysed. This can be done using different features of machine readable news data (A Team 2010).

4.4 Desirable industry applications

Stock picking, trading and fund management(section 4.2) and risk control (section 4.3) are the established application areas in finance industry and the use of NA is researched to achieve improved performance.

- Market surveillance

Responding to the state of the market and taking into consideration the preoccupation of the watch-dogs, that is, the regulators market surveillance is becoming an important application area of quant models. It is gaining in importance because managers through internal control functions as much as external compliance requirement wish to have surveillance in place to catch rogue trading, insider information based trading. An innovative application of NA is to spot patterns which captures these.

- Trader decision support

News data can aid traders in making decisions. News data signals may confirm traders existing analysis or it may cause them to reconsider their analysis.

- Wolf detection / circuit breaker

Wolf detectors (circuit breakers) are a risk control feature for algorithmic trading built on machine readable news. Essentially they “break the circuit” stopping an automated algorithm from trading on a certain asset when particular types of news are released. It is important to try not to shout “Wolf!” when no wolf has actually appeared. These risk control features can be customised to only be tripped when substantive news events have occurred. Alternatively the algorithms can be turned back after the nature of the news has been programmatically analysed. This can be done using different features of machine readable news data. (See A-Team report)

- News flow algorithms

It is widely recognized that newsflow is a good indicator of volume and volatility. As the flow of news about a company rises, the volume traded rises resulting in more stock price volatility. If news flow can be used effectively to predict volume or volatility spikes then algorithms based on News VWAP versus VWAP may add value for trade execution strategies.

- Post trade analysis

Assist in proving best execution and trader performance??

News data is likely to add value for investors trading at all frequencies from volatility based strategies to equity trading.

- Alpha generating signal

News data can be used in alpha generation at various trading frequencies. News sentiment data may be used within factor models. Cahan, Jussa and Luo (2009) consider such an application. Their results

are positive and they find that such an approach does add value. In particular they note the value of this source of information during the credit crisis, when determining fundamentals (which traditional quant factors are based on) was problematic. News data can also aid quant investors to identify non rational biased behaviour of investors. These can then be exploited.

[? Tetlock, Saar-Tsechansky, and Macskassy (2008) note that an investor's perceptions about the future asset returns are determined by their knowledge about the company and its prospects, that is, by their "information sets". They note that these are determined from three main sources: analysts forecasts, quantifiable publicly disclosed accounting variables and linguistic descriptions of the firm's current and future profit generating activities. If the first two sources of information are incomplete or biased, the third may give us relevant information. ?]

- Stock screening tool

News data can be used to aid stock screening. In particular sentiment data may be used to try to guess the directional movement of future returns. Very good news stocks (for example top sentiment quintile) might be selected to be held long and very bad news stocks (for example bottom sentiment quintile) might be selected to be held short.

- Fundamental research

News analysis tools may aid traditional non-quant managers, by allowing them to undertake market research more efficiently.

- Risk Management

The use of news data within risk forecasting can allow for dynamic (adaptive) risk management strategies that are forward looking and are based on changing market environments. Further this risk analysis applied using news data can help investors understand event risk and how different kinds of events can impact their portfolio risk profile.

We may use certain news data within quantitative models. We may use it simply to forecast the directional impact of news on asset prices. In more sophisticated models we might wish to determine return predictions. Models which forecast volatility and volume on the basis of news will also find important applications within the investment management process. Volatility prediction for volatility traders?? Volume prediction for factor models that use volume as a factor?

5 Summary and discussions

The development of news analytics and its applications to finance through sentiment analysis is gaining progressive acceptance within the investment community. A growing number of academic studies have been conducted; in this paper we have reviewed these in a summary form. Research by service providers of data and content for the finance industry is also discussed in this paper and we have identified the applications of News Analytics to high frequency and low frequency trading as well as in risk control and compliance. The study of News Analytics draws upon research from a number of disciplines including natural language processing, AI pattern recognition and classifiers, text mining, information engineering as well as financial engineering; we believe News Analytics will soon become an important area of study within financial analytics.

6 References

References

- [1] W. Antweiler and M. Frank. Is All That Talk Just Noise? The Information Content of Stock Message Boards. *Journal of Finance*, 59(3), 2004.
- [2] L.S. Bamber, O.E. Barron, and T.L. Stober. Trading volume and different aspects of disagreement coincident with earnings announcements. *The Accounting Review*, 72:575–597.
- [3] B.M. Barber and T. Odean. Boys Will be Boys: Gender, Overconfidence, and Common Stock Investment. *Quarterly Journal of Economics*, 116(1):261–292, 2001.
- [4] B.M. Barber and T. Odean. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *Review of Financial Studies*, 21(2):785–818, 2008.
- [5] B.M. Barber and T. Odean. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors UPDATED. *In this volume: Chapter 5*, 2010.
- [6] N. Blasco, P. Corredor, C. Del Rio, and R. Santamaria. Bad news and dow jones make the spanish stocks go round. *European Journal of Operational Research*, 163(1):253 – 275, 2005.
- [7] J.H. Boyd, J. Hu, and R. Jagannathan. The stock market’s reaction to unemployment news: Why bad news is usually good for stocks. *The Journal of Finance*, 60(2):649–672, 2005.
- [8] Brown. *CARISMA annual conference: Incorporating news analytics into quantitative investment and trading strategies* <http://www.optirisk-systems.com/papers/RichardBrown.pdf>, 2010.
- [9] J. A. Busse and T. Clifton Green. Market efficiency in real time. *Journal of Financial Economics*, 65(3):415–437, 2002.
- [10] R. Cahan, J. Jussa, and Y. Luo. Breaking news: How to use news sentiment to pick stocks. *MacQuarie Research Report*.
- [11] J.Y. Campbell, A.W. Lo, and A.C. MacKinlay. The econometrics of financial markets. *Chapter 4: Event study analysis*.
- [12] W.S. Chan. Stock Price Reaction to News and No-News: Drift and Reversal After Headlines. *Journal of Financial Economics*, 70(2):223–260, 2003.
- [13] Z. Da, J. Engelberg, and P. Gao. In Search of Attention. Working Paper: Available on SSRN http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1364209, 2009.
- [14] K. Daniel, D. Hirshleifer, and A. Subrahmanyam. Investor psychology and security market under-and overreactions. *The Journal of Finance*, 53(6):1839–1885, 1998.
- [15] K. Daniel, D. Hirshleifer, and S.H. Teoh. Investor psychology in capital markets: Evidence and policy implications. *Journal of Monetary Economics*, 49(1):139–209, 2002.
- [16] S. Das. News analytics metrics: Desirable properties TBC. *In this volume: Chapter 3*, 2010.
- [17] S.R. Das and M.Y. Chen. Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388, 2007.
- [18] W. De Bondt and R. Thaler. Does the stock market overreact? *The Journal of Finance*, 40(3):793–805, 1985.
- [19] D. diBartolomeo. Using news as a state variable in assessment of financial market risk. *In this volume: Chapter 9*, 2010.

- [20] D. diBartolomeo and S. Warrick. Making covariance based portfolio risk models sensitive to the rate at which markets reflect new information. In Knight J. and Satchell. S., editors, *Linear Factor Models*. Elsevier Finance, 2005.
- [21] M. Dzielinski, M.O. Rieger, and T. Talpsepp. Volatility, asymmetry, news and private investors. *In this volume: Chapter 10*, 2010.
- [22] L.H. Ederington and J.H. Lee. How markets process information: News releases and volatility. *Journal of Finance*, 48:1161–1191, 1993.
- [23] J. Engleberg and S. Sankaraguruswamy. How to gather data using a web crawler: An application using SAS to search EDGAR. Available: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1015021&r...
- [24] E.F. Fama and K.R. French. The cross-section of expected stock returns. *Journal of Finance*, 47(2):427–466, 1992.
- [25] L.H. Fang and J. Peress. Media coverage and the cross-section of stock returns. *Forthcoming in Journal of Finance*, 2009.
- [26] M. Graham, J. Nikkinen, and P. Sahlstrom. Relative Importance of Scheduled Macroeconomic News for Stock Market Investors. *Journal of Economics and Finance*, 27(2):153–165, 2003.
- [27] R.C. Gutierrez Jr, E. Kelley, and M.C. Hall. The long-lasting momentum in weekly returns. *Journal of Finance*, Forthcoming.
- [28] Hafez. *CARISMA annual conference: The role of news in financial markets* <http://www.optirisk-systems.com/papers/PeterAgerHafez.pdf>, 2010.
- [29] P. Hafez. Detection of seasonality in newsflow. *White Paper available from RavenPack*, 2009.
- [30] D. Hirshleifer. Investor psychology and asset pricing. *The Journal of Finance*, 56(4):1533–1597, 2001.
- [31] D. Hirshleifer, S.S. Lim, and S.H. Teoh. Driven to Distraction: Extraneous Events and Underreaction to Earnings News (Digest Summary). *CFA Digest*, 40(1), 2010.
- [32] H. Hong and J.C. Stein. A unified theory of underreaction, momentum trading, and overreaction in asset markets. *The Journal of Finance*, 54(6):2143–2184, 1999.
- [33] K. Hou, L. Peng, and W. Xiong. A tale of two anomalies: The implications of investor attention for price and earnings momentum. Available <http://ssrn.com/abstract=976394>.
- [34] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.
- [35] P.S. Kalev and H.N. Duong. Firm specific news arrival and the volatility of intraday stock index and futures returns. *In this volume: Chapter 11*, 2010.
- [36] J.M. Karpoff. The relation between price changes and trading volume: A survey. *Journal of Financial and Quantitative Analysis*, 22:109–126, 1987.
- [37] J. Kittrell. Sentiment reversals as buy signals. *In this volume: Chapter 8*, 2010.
- [38] SP Kothari and J.B. Warner. Econometrics of event studies. In *Handbook of Empirical Corporate Finance*. Elsevier Finance, 2005.
- [39] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan. Language models for financial news recommendation. In *Proceedings of the ninth international conference on Information and knowledge management*. ACM, 2000.
- [40] D. Leinweber. *Nerds on Wall Street*. John Wiley, 2009.
- [41] D. Leinweber and J. Sisk. Relating news analytics to stock returns. *In this volume: Chapter 4*, 2010.

- [42] F. Li. Do stock market investors understand the risk sentiment of corporate annual reports? *University of Michigan Working Paper Available http://papers.ssrn.com/sol3/papers.cfm?abstract_id=898181*.
- [43] F. Li. Do stock market investors understand the risk sentiment of corporate annual reports? UPDATED. *In this volume: Chapter 6*, 2010.
- [44] X. Liang. Impacts of internet stock news on stock markets based on neural networks. In *Advances in neural networks*. Springer Berlin/Heidelberg, 2005.
- [45] J. Lintner. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics*, 47(1):13–37, 1965.
- [46] A. Lo. Reuters NewsScope Event Indices. *AlphaSimplex research report produced in partnership with Thomson Reuters*.
- [47] A. Lo. *Market efficiency: Stock market behaviour in theory and practice*. Edward Elgar Pub, 1997.
- [48] A. Lo. The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *The Journal of Portfolio Management*, 2004.
- [49] A. Lo and A. Healy. Reuters NewsScope Event Indices. *In this volume: Chapter 2*, 2010.
- [50] T. Loughran and B. McDonald. When is a liability not a liability? *Journal of Finance*, Forthcoming, 2010.
- [51] R. Luss and A. d’Aspremont. Predicting abnormal returns from news using text classification. *Working Paper from ORFE Princeton*, 2009.
- [52] R.C. Merton. A simple model of capital market equilibrium with incomplete information. *Journal of Finance*, pages 483–510, 1987.
- [53] L. Mitra, G. Mitra, and D. diBartolomeo. Equity portfolio risk (volatility) estimation using market information and sentiment. *Quantitative Finance*, 9(8):887–895, 2009.
- [54] M.A. Mittermayer and G. Knolmayer. Text mining systems for market response to news: A survey. *Working Paper University of Bern: Available on SSRN <http://www2.ie.iwi.unibe.ch/publikationen/berichte/resource/WP-184.pdf>*, 2006.
- [55] A. Moniz, G. Brar, and C. Davis. Have I Got News for You. *MacQuarie Research Report*.
- [56] A. Moniz, G. Brar, and C. Davis. Have I Got News for You. *In this volume: Chapter 7*, 2010.
- [57] Munz. *CARISMA annual conference: US markets: Earnings news release - an inside look <http://www.optirisk-systems.com/papers/MarianMunz.pdf>*, 2010.
- [58] T. Odean and B. Barber. Are investors reluctant to realize their losses? *The Journal of Finance*, 53(5):1775–1798, 1998.
- [59] A. Patton and M. Verardo. Does Beta Move with News? Systematic Risk and Firm-Specific Information Flows. *FMG Discussion Papers available from <http://eprints.lse.ac.uk/24421/1/dp630.pdf>*, 2009.
- [60] D. Peramunetilleke and R. K. Wong. Currency exchange rate forecasting from news headlines. In Xiaofang Zhou, editor, *Thirteenth Australasian Database Conference (ADC2002)*, Melbourne, Australia, 2002. ACS.
- [61] R.L. Peterson. Affect and financial decision-making: How neuroscience can inform market participants. *The Journal of Behavioral Finance*, 8(2):70–78, 2007.
- [62] Kalev P.S., W.M Liu, P.K. Pham, and E. Jarnecic. Public information arrival and volatility of intraday stock returns. *Journal of Banking and Finance*, 28(6):1441–1467, 2004.
- [63] RavenPack. RavenPack News Scores User Guide. *February 11, 2010, Version 1.3.1*, 2010.

- [64] C. Robertson, S. Geva, and R. Wolff. What types of events provide the strongest evidence that the stock market is affected by company specific news? In *Proceedings of the fifth Australasian conference on Data mining and analytics-Volume 61*, page 153. Australian Computer Society, Inc., 2006.
- [65] C.S. Robertson, S. Geva, and R.C. Wolff. News aware volatility forecasting: Is the content of news important? In *Proceedings of the sixth Australasian conference on Data mining and analytics-Volume 70*, pages 161–170. Australian Computer Society, Inc., 2007.
- [66] B. Rosenberg, K. Reid, and R. Lanstein. Persuasive evidence of market inefficiency. *Journal of Portfolio Management*, 11(3):9–16, 1985.
- [67] S.A. Ross. The Arbitrage Pricing Theory of Capital Asset Pricing. *Journal of Economic Theory*, 13(3):341–360, 1976.
- [68] P. Ryan and R.J. Taffler. Are economically significant stock returns and trading volumes driven by firm-specific news releases? *Journal of Business Finance & Accounting*, 31(1-2):49–82, 2004.
- [69] I. Schmerken. Trading off the news. *Wall Street and Technology*, Available from http://www.wallstreetandtech.com/technology-risk-management/showArticle.jhtml;jsessionid=ZYNMF1D4EJ4LHQE1GHRSKHWATMY32JVN?articleID=185302817&_requestid=532279.
- [70] J. Scott, M. Stumpp, and P. Xu. News, not trading volume, builds momentum. *Financial Analysts Journal*, 59(2):45–54, 2003.
- [71] M. Seasholes and G. Wu. Profiting from predictability: Smart traders, daily price limits, and investor attention. *University of California, Berkeley, working paper available <http://www.nd.edu/~pschultz/SeasholesWu.pdf>*, 2004.
- [72] W.F. Sharpe. Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk. *Journal of Finance*, 19(3):425–442, 1964.
- [73] A Team. Machine readable news and algorithmic trading. *Whitepaper produced for Thomson Reuters and Market News International*, 2010.
- [74] P.C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62:1139–1168, 2007.
- [75] P.C. Tetlock, M. Saar-Tsechansky, and S. Mackassy. More than words: Quantifying language to measure firms’ fundamentals. *Journal of Finance*, 63(3):1437–1467, 2008.
- [76] Thomson-Reuters. Reuters NewsScope Sentiment Engine: Guide to Sample Data and System Overview. *Dec, 2009*, Version 3, 2009.
- [77] A. Tversky and D. Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453, 1981.
- [78] Vreijling. *CARISMA annual conference: Practical use of news in equity trading strategies <http://www.optirisk-systems.com/papers/MarkVreijling.pdf>*, 2010.
- [79] Wysocki. Cheap Talk on the Web: The Determinants of Postings on Stock Message Boards. *Working Paper No. 98025 University of Michigan*, 1999.

A Appendix I: Structure and Content of News Data

In this appendix we summarise the services by two leading providers of news analytics, namely Thomson Reuters (NewsScope) and RavenPack (NewsScore). The information is presented under three main headings, (i) Coverage, (ii) Method and Types of Cores, (iii) Example of News Data in a tabular form.

Details of Thomson Reuters News Analytics Coverage

(i) Coverage

Real-time and historical equity coverage

Commodities and Energy: 39 C&E Topics

Equity:

All Equities	34,037	100.00%
Active companies	32,719	96.1%
Inactive companies	1,318	3.9%

Equity Coverage by region

Americas: 14,785

APAC: 11,055

EMEA: 8,197

Equity coverage updates: Bi-Weekly upgrade to recent changes

History: Available from January, 2003 (history kept for de-listed companies; symbology change tracked).

Data Fields: 82 metadata fields including Timestamp (GMT Millisec), linked counts over various time periods which measure repetition, linked item cross references, language, topics, prevailing sentiment, detailed sentiment, relevance, size of item, broker action, market commentary, number of companies mentioned, position of first mention, news intensity, news source, story type, headline, company identifier, among others.

Delivery Mechanisms: Internet/VPN, co-lo, dedicated circuits, deployed on-site, FTP, Thomson Reuters Quantitative Analytics / Market QA.

News Sources: Reuters host of third party sources standard; Able to process customer specific sources including internet feeds, PDF files, and text from databases.

(ii) Method and Types of Scores

(Thomson Reuters : Please supply a short description of column headings)

(iii) Example of News Data in a tabular form

TIMESTAMP	R/C	RELEVANCE	SENTIMENT	POSITIVE	NEUTRAL	NEGATIVE	LINKED COUNTS	ITEMTYPE	HEADLINE	TOPIC CODES
00:34:28.944	IBM.N	0.29	1	0.538	0.454	0.008	0,0,0,0,0	ARTICLE	Arrow to buy smaller rival for \$485 million	US WHO LEN RTRS MRG SPWR H
11:14:04.042	IBM.N	1	1	0.842	0.133	0.025	0,0,0,0,0	ALERT	UBS RAISES IBM «IBM.N» TO BUY FROM NEUTRAL - THEFLYONTHEW	RCH US CA LEN RTRS
11:16:55.812	IBM.N	1	1	0.850	0.119	0.031	1,1,1,1,1	ARTICLE	US RESEARCH NEWS-UBS raises IBM to buy - theflyonthewall.com	RCH US CA LEN RTRS
11:20:50.082	IBM.N	1	0	0.247	0.614	0.138	1,1,1,1,1	ARTICLE	RESEARCH ALERT-UBS upgrades IBM to buy - theflyonthewall.com	RCH DPR HDWR SPWR US LEN R'
12:22:43.689	IBM.N	1	1	0.842	0.133	0.025	0,0,0,0,0	ALERT	IBM «IBM.N» SHARES RISE 1.1 PCT TO \$98.50 BEFORE THE BELL. AFTER	RCH US CA LEN RTRS
12:36:50.695	IBM.N	1	1	0.542	0.450	0.008	3,3,3,3,3	ARTICLE	Before the Bell - Bed Bath & Beyond, IBM rise early	DPR HDWR US STX HOT LEN RTR
14:59:02.943	IBM.N	1	1	0.701	0.164	0.135	1,1,1,1,1	ARTICLE	UPDATE 1-RESEARCH ALERT-UBS upgrades IBM to buy from neutral	US RCH DPR HDWR SPWR BUS LI
15:05:53.790	IBM.N	0.13	-1	0.056	0.125	0.819	1,1,1,1,1	ARTICLE	US RESEARCH NEWS-Credit Suisse recommends trading buy on GM	RCH US CA LEN RTRS
15:06:13.000	IBM.N	0.08	-1	0.056	0.125	0.819	2,2,2,2,2	APPEND	US RESEARCH NEWS-Credit Suisse recommends trading buy on GM	RCH US CA LEN RTRS
16:31:45.041	IBM.N	0.25	0	0.218	0.612	0.170	4,4,4,4,4	APPEND	HEADLINE STOCKS - U.S. stocks on the move on Jan 8	US STX FIN RESF RES BUS HOT L
16:31:55.631	IBM.N	0.25	0	0.218	0.612	0.170	6,6,6,6,6	APPEND	HEADLINE STOCKS - U.S. stocks on the move on Jan 8	US STX FIN RESF RES BUS HOT L
18:49:48.004	IBM.N	0.32	0	0.221	0.613	0.166	7,7,7,7,7	APPEND	HEADLINE STOCKS - U.S. stocks on the move on Jan 8	US STX FIN RESF RES BUS HOT L
19:18:14.726	IBM.N	0.20	-1	0.180	0.251	0.568	0,0,0,0,0	ARTICLE	UPDATE 1-Sears aims to drive sales with virtual showroom	RET US WWW LEN RTRS
20:09:19.547	IBM.N	0.34	1	0.830	0.128	0.042	0,0,0,0,0	ARTICLE	US STOCKS-Indexes higher, upgrades boost tech sector	US STX BUS MUNI FIN NEWS LEN
20:09:54.796	IBM.N	0.14	1	0.830	0.128	0.042	1,1,1,1,1	APPEND	US STOCKS-Indexes higher, upgrades boost tech sector	US STX BUS MUNI FIN NEWS LEN
04:09:34.780	IBM.N	1	1	0.512	0.382	0.107	0,0,0,0,0	ARTICLE	IBM appoints new Greater China CEO	CN ASIA ELI HK TW F EMRG LEN I
19:13:02.511	IBM.N	0.17	0	0.216	0.611	0.174	0,0,0,0,0	ARTICLE	CES-Visa, Nokia turn mobile phones into mobile wallets	WEU EUROPE WWWV DE NORD US
19:13:59.476	IBM.N	0.17	0	0.216	0.611	0.174	1,1,1,1,1	APPEND	CES-Visa, Nokia turn mobile phones into mobile wallets	WEU EUROPE WWWV DE NORD US
11:55:22.595	IBM.N	1	-1	0.188	0.112	0.700	0,0,1,1,1	ALERT	AG EDWARDS CUTS IBM «IBM.N» TO HOLD FROM BUY - THEFLYONTHEW	RCH US DPR HDWR SPWR LEN R'
12:02:25.855	IBM.N	1	-1	0.137	0.217	0.645	1,1,3,3,3	ARTICLE	RESEARCH ALERT-AG Edwards cuts IBM to hold - theflyonthewall.com	RCH US DPR HDWR SPWR LEN R'
15:20:49.892	IBM.N	1	-1	0.311	0.145	0.544	1,1,3,3,3	ARTICLE	UPDATE 1-RESEARCH ALERT-AG Edwards downgrades IBM	US RCH DPR HDWR SPWR LEN R'
11:29:20.729	IBM.N	0.18	1	0.841	0.123	0.036	0,0,0,0,0	APPEND	FACTBOX-UK companies cut, close final-salary pensions	GB WEU EUROPE FUND FIN RTM F
12:57:47.150	IBM.N	1	1	0.552	0.441	0.007	0,0,0,0,0	ALERT	BANC OF AMERICA RAISES IBM «IBM.N» PRICE TARGET TO \$110 FROM	RCH DPR US LEN RTRS ENT HDW
13:24:15.667	IBM.N	1	1	0.565	0.342	0.093	3,3,5,7,7	ARTICLE	RESEARCH ALERT-BofA raises price targets on IBM, Apple, EMC	RCH DPR US LEN RTRS ENT HDW

Figure 9: Thomson Reuters NewsScope Sentiments

Details of RavenPack NewsScores: Equity coverage and available data

(i) Coverage

Real-time and historical equity coverage

Equity coverage	10,742	100.00%
Active companies	9,934	92.48%
Inactive companies	808	7.42%

Coverage by region

Americas:	4,141
Asia:	3,642
Europe:	2,431
Oceania:	353
Africa:	175

Available data

Historical:

Data format:	Comma separated values (CSV) file
Archive range:	Single .csv files comprising coverage of all companies in their same sector 41 sector files plus 1 uncategorised = 42 .csv files compressed in .zip files on a per year basis
Data fields:	10 fields: Date/time, ID, Sentiment (5), Market Impact, Category and Company relevance
Download:	Secure web download

Real-time:

Connection:	Over the internet
Software:	Local installation of RavenPack Data Gateway(Windows Client) plus API
Access:	Push feed for real-time plus historical query mechanism to fill gaps as required
API:	Java

(ii) Method and Types of Scores

TIMESTAMP_UTC The Date/Time (yyyy-mm-dd hh:mm:ss.000) at which the news item was published in Coordinated Universal Time (UTC).

COMPANY_IDENTIFIER A unique and permanent company identifier assigned by RavenPack that consistently identifies companies throughout the historical archive and in real-time. Company identifiers are mapped to common securities identifiers such as ISINs, CUSIPs, TICKERs, etc.

COMPANY_RELEVANCE A score between 0-100 that indicates how strongly related the company is to the underlying news story, with higher values indicating greater relevance.

EVENT_CATEGORIES An element or "tag" representing a company-specific news announcement or formal event. Highly relevant stories about companies are classified into a set of predefined event categories. When applicable, the role played by the company in the story is also detected and tagged.

EVENT_SENTIMENT A score between 0 and 100 that represents the news sentiment for a given company by measuring various sentiment proxies sampled from the news. The score is determined by systematically matching stories typically categorized by financial experts as having short-term positive or negative share price impact. The algorithm produces a score for more than 160 types of news events - scheduled and

unscheduled.

EVENT _NOVELTY A score between 0 and 100 that represents how "new" or novel a news story is within a given time window.

EVENT _NOVELTY _KEY An identifier that provides a way to chain or relate similar stories about a similar company-specific event. Its a powerful way to build a model based on relationships between companies and event categories like product recalls, layoffs, corporate or legal issues, earnings announcements, analyst and credit revisions, and many more.

COMPOSITE _SENTIMENT A sentiment score between 0 and 100 that represents the news sentiment of a given story by combining various sentiment analysis techniques. The direction of the score is determined by looking at emotionally charged words and phrases and by matching stories typically rated by experts as having short-term positive or negative share price impact. The strength of the score (values above or below 50, where 50 represents neutral strength) is determined from intraday stock price reactions modeled empirically using tick data.

STORY _LEVEL _SENTIMENT A collection of sentiment scores between 0 and 100 that represent the sentiment of a given story. RavenPack includes various metrics, each calculated using a different linguistic classifier or technique. A classifier is an algorithm designed to assess the overall sentiment of a company by detecting events and language within stories that are likely to drive stock prices upwards or downwards over a given period of time.

(iii) Example of News Data in a tabular form

TIMESTAMP_UTC	RP_STORY_ID	WLE	PCM	ECM	RCM	VCM	NIP	COMPANIES
2009-05-13 18:55:33	761C6BB4D362C6106FC51A5DA7F39074	50	0	50	50	50	41	US/IHS:2
2009-05-13 18:55:44	C5281C388E3AFE07D6AB866216034AE5	50	50	100	50	50	27	US/RAI:90
2009-05-13 18:55:56	3450143335BD6A9BF1DD6F0D6C5CC47	50	50	0	50	50	41	US/BAC:90
2009-05-13 18:56:03	B62C859C48AE29FA81586CC95A764B33	50	50	100	100	50	42	US/IBM:100
2009-05-13 18:56:39	75871161C64B263C495642152B0302CF	50	50	50	50	50	55	US/BAC:12,GB/BARC:2,US/RF:2
2009-05-13 18:56:39	75871161C64B263C495642152B0302CF	50	50	50	50	50	55	US/BAC:12,GB/BARC:2,US/RF:2
2009-05-13 18:56:39	75871161C64B263C495642152B0302CF	50	50	50	50	50	55	US/TOL:2,US/DHI:2
2009-05-13 18:56:39	75871161C64B263C495642152B0302CF	50	50	50	50	50	55	US/TOL:2,US/DHI:2
2009-05-13 18:56:39	75871161C64B263C495642152B0302CF	50	50	50	50	50	55	US/BA:12

Source: RavenPack, Macquarie Capital (USA), May 2009

Figure 10: RavenPack NewsScores

A.1 Appendix II: Annotated bibliography of selected papers

A.1.1 Reuters NewsScope Event Indices (NEI)

Author:

Andrew Lo, AlphaSimplex

Focus:

The creation of event indices (Reuters NewsScope Event Indices NEI) which reflect issuance of market moving news. The indices are constructed to have “predictive” power for (realised) volatility and returns, since they are constructed in an integrated framework where news, returns and (realised) volatility are considered in creating the indices. The indices are designed to form inputs into systematic investment and risk management protocols.

Method:

The framework for developing the Reuters NEI is as follows. For a given asset class and related topic area the following parameters are used.

- (1) List of keywords and phrases with real valued weights; $(W_1, \gamma_1), \dots, (W_k, \gamma_k)$.
- (2) A rolling “sentiment” window of size r (say 5/10 minutes).
- (3) A rolling calibration window of size R (say 90 days).

Initially a *raw score* is created.

We have $(W_1, \gamma_1), \dots, (W_k, \gamma_k)$, where W_1 is the first keyword and γ_1 is the weighting for the first keyword.

The raw score at time t is assigned by considering the “sentiment” window $(t - r, t]$. (w_1, \dots, w_k) is the vector of keyword frequencies in $(t - r, t]$, that is, w_i is the number of times keyword W_i occurred in the last r minutes. The raw score is defined as

$$s_t \equiv \sum_i \gamma_i w_i \quad (4)$$

The raw score will tend to be high when the news volume is high. A *normalised score* is therefore produced using the rolling calibration window. At all times t for the R days in the calibration window, we record

- (i) the raw score s_t that would have been assigned,
- (ii) the news volume; $n_{[t-r, t]}$ the number of words that were observed in the time interval $[t - r, t)$.

The normalised score is determined by comparing the current raw score against the distribution of raw scores in the calibration window, where the news volume equalled the current news volume. This means we only consider those raw scores where the news volume equals the current news volume.

$$S_t \equiv \frac{|\{t' \in [t - R, t) : n_{[t'-r, t']} = n_{[t-r, t]} \ \& \ s_{t'} < s_t\}|}{|\{t' \in [t - R, t) : n_{[t'-r, t']} = n_{[t-r, t]}\}|} \quad (5)$$

The numerator is a subset of the denominator, hence $S_t \leq 1$. If $S_t = 0.92$, we can say 92% of the time when news volume is at the current level, the raw score is less than it currently is. Lo creates an alternative score based on topic codes. Instead of counting word frequencies, the fraction of news alerts (in the last r

minutes) tagged with particular topic codes, are used.

Naturally the scoring method is dependent on the list of keywords/topic areas (W_1, \dots, W_k) and the real valued weights ($\gamma_1, \dots, \gamma_k$). The lists of keywords/topics were created by selecting the major news categories that related to the asset class (foreign exchange) and creating lists, by hand, of words and topic areas that suggest news relevant to the categories. A tool was created to extract news from periods where high scores were assigned. This news was then manually inspected, so that the developer could determine whether the keywords (topics) were legitimate or needed adjusting.

The optimal weights ($\gamma_1, \dots, \gamma_k$) for the intraday return sentiment index were determined by regressing the keyword/topic frequencies against the intraday asset returns. Similarly the (optimal) weights for the intraday volatility sentiment index were determined by regressing the keyword/topic frequencies against the intraday (de-seasonalised) realised volatility. Volatility was observed to show strong seasonality on intraday timescales, hence this series was de-seasonalised prior to derivation of the weights. Returns did not exhibit any seasonality. The time series are given on an intraday basis, hence to keep the data manageable a random subset of the observations are used in calibration. Lo notes the determination of the weights can be expressed as a more general classification problem. Other techniques might be applied, in particular machine learning algorithms such as the perceptron algorithm or support vector machines. He suggests further study is required to find the best approach, but the standard linear regression approach does perform well.

Data source:

Reuters NewsScope Alerts which are news flashes issued when newsworthy events occur. These items are both timely and relevant. They are tagged with machine readable codes. The alerts' text is concise and formed from a relatively small vocabulary, hence lends itself well to applications of machine learning algorithms.

Results and conclusions:

Lo undertakes detailed event study analysis to establish that the final NEI have empirical significance. He uses the NEI series to define an event. An event is defined to take place when the index exceeds a certain threshold (say 0.995). He then removes any events that follow in less than one hour of another event. This guards against identifying "new" events which are actually based on old news. The behaviour of exchange rates before and after these events are then studied. Two time series are considered; the log returns and the deseasonalised squared log returns. He then tests the null hypothesis that the distribution of log returns / deseasonalised squared log returns are the same before and after the events. He uses samples of one hour centered on the events. Lo finds that the event studies confirm the constructed event indices, on average, impact the *realised* foreign exchange volatility.

A.1.2 Yahoo! for Amazon: Sentiment extraction from small talk on the web

Author:

Sanjiv Das and Mike Chen

Focus:

Das and Chen (2007) use statistical and natural language techniques to extract investor sentiment from stock message boards and generate sentiment indices. They apply their method for 24 technology stocks present in the Morgan Stanley High Tech (MSH) Index.

Method:

A web scraper program is used to download tech sector message board messages. Five algorithms, each with different conceptual underpinnings are used to classify each message. A voting scheme is then applied to all five classifiers.

Three supplementary databases are used in the classification algorithms.

1. “*Dictionary*” is used for determining the nature of the word. For example, is it a noun, adjective or adverb?
2. “*Lexicon*” is a collection of hand picked finance words which form the variables for statistical inference within the algorithms.
3. “*Grammar*” is the training corpus of base messages used in determining the in-sample statistical information. This information is then applied for use on the out-of-sample messages.

The lexicon and grammar jointly determine the context of the sentiment. Each of the classifiers relies on a different approach to message interpretation. They are all analytic, hence computationally efficient.

1. *Naive classifier* (NC) is based on a word count of positive and negative connotation words. Each word in the lexicon is identified as being positive, negative or neutral. A parsing algorithm negates words if the context requires it. The net word count of all lexicon matched words is taken. If this value is greater than one, we sign the message as a buy. If the value is less than one the message is a sell. All others are neutral.
2. *Vector distance classifier* Each of the D words in the lexicon is assigned a dimension in vector space. The full lexicon then represents a D -dimensional unit hypercube. Every message can be described as a word vector in this space ($m \in \mathbb{R}^D$). Each hand tagged message in the training corpus (grammar) is converted into a vector G_j (grammar rule). Each (training) message is pre classified as positive, negative or neutral. We note that Das and Chen use the terms Buy/Positive, Sell/Negative and Neutral/Null interchangeably. Each new message is classified by comparison to the cluster of pretrained vectors (grammar rules) and is assigned the same classification as that vector with which it has the smallest angle. This angle gives a measure of closeness.
3. *Discriminant based classification* NC weights all words within the lexicon equally. The discriminant based classification method replaces this simple word count with a weighted word count. The weights are based on a simple discriminant function (Fisher Discriminant Statistic). This function is constructed to determine how well a particular lexicon word discriminates between the different message categories. These categories are { Buy, Sell, Null }. The function is determined using the pre classified messages within the grammar. Each word in a message is assigned a signed value, based on its sign in the lexicon multiplied by the discriminant value. Then as for NC a net word count is taken. If this value is greater than 0.01, we sign the message as a buy. If the value is less than -0.01 the message is a sell. All others are neutral.

4. *Adjective - adverb phrase classifier* is based on the assumption that phrases which use adjectives and adverbs emphasize sentiment and require greater weight. This classifier also uses a word count but uses only those words within phrases containing adjectives and adverbs. A “tagger” extracts noun phrases with adjectives and adverbs. A lexicon is used to determine whether these significant phrases indicate positive or negative sentiment. The net count is again considered to determine whether the message has negative or positive overall sentiment.
5. *Bayesian classifier* is a multi variate application of Bayes Theorem. It uses the probability a particular word falls within a certain classification and is hence indifferent to the structure of language. We consider three categories $C = \{c_i \mid i = 1, \dots, C\}$. Denote each message $m_j \quad j = 1, \dots, M$. The set of lexical words is $F = \{w_k\}_{k=1}^D$. (The total number of lexical words is D) We can determine a count of the number of times each lexical item appears in each message $n(m_j, w_k)$. Given the class of each message in the training set we can determine the frequency with which a lexical word appears in a particular class. We are then able to compute the conditional probability of an incoming message j falling in category i , $Pr(m_j|c_i)$, from the word based frequencies. $Pr(c_i)$ is set to the proportion of messages in the training set classified in class c_i . For a new message we are able to compute the probability it falls within class c_i given its component lexicon words, that is $P(c_i|m_j)$, through an application of Bayes Theorem. The message is classified as being from the category with the highest probability.

A voting scheme is then applied to all five classifiers. The final classification is based on achieving a majority amongst the five classifiers. If there is no majority the message is not classified. This reduces the number of messages classified but enhances the classification accuracy.

Das and Chen also introduce a method to detect message ambiguity. Messages posted on stock message boards are often highly ambiguous. The grammar is often poor and many of the words do not appear in standard dictionaries. [They note “Ambiguity is related to the absence of “aboutness””. The General Inquirer has been developed by Harvard University for content analyses of textual data. They use it to determine an independent optimism score for each message. By using a different definition of sentiment it is ensured there is no bias to a particular algorithm. The optimism score is the difference between the number of optimistic and pessimistic words as a percentage of the total words in the body of the text. This score allows us to rank the relative sentiment of all stories within a classification group. For example, they can rank the relative optimism of all stories which have been classified by their scheme as positive. The mean and standard deviation of the optimism score for different classification types ($\{\text{Buy, Sell, Null}\}$) can be calculated. They filter *in* and consider only highly optimistically scored stories in the positive category. For example only those stories with optimism scores above the mean value plus one standard deviation are considered. Similarly they filter *in* and consider only the most highly pessimistic scores in the negative category.] Once the classified stories are further filtered for ambiguity, it is found that the number of false positives dramatically declines.

Once the sentiment for each message is determined using the voting algorithm, a daily sentiment index is compiled. The classified messages up to 4pm each day are used to create the aggregate daily sentiment for each stock. A buy (sell) message increments (decrements) the index by one. These indices are further aggregated across all stocks to obtain an aggregate sentiment for the technology portfolio.

Data source:

Results and conclusions:

Human Resource Predictive Analytics (HRPA) for HR Management in Organizations

Sujeet N. Mishra, Dev Raghvendra Lama, Yogesh Pal

Abstract— Human resource predictive analytics is an evolving application field of analytics for HRM purposes. The purpose of HRM is measuring employee performance and engagement, studying workforce collaboration patterns, analyzing employee churn and turnover and modelling employee lifetime value. The motive of applying HRPA is to optimize performances and produce better return on investment for organizations through decision making based on data collection, HR metrics and predictive models. The paper is divided into three sections to understand the emergence of HR predictive analytics for HRM. Firstly, the paper introduces the concept of HRPA. Secondly, the paper discusses three aspects of HRPA: (a) Need (b) Approach & Application (c) Impact. Lastly, the paper leads to the conclusion on HRPA.

Index Terms— Predictive Analytics, Talent Analytics, HR Analytics, Human Resource Management, Modelling, Return on Investment (ROI), Decision Making.

1 INTRODUCTION

HR analytics is a multidisciplinary approach to integrate methodology for improving the quality of people-related decisions in order to improve individual and organizational performance. There are interchangeable terms used for HR analytics are talent analytics, people analytics, and workforce analytics. HR analytics plays a role in every aspect of the HR function, including recruiting, training and development, succession planning, retention, engagement, compensation, and benefits. HR analytics are those that involve “high-end” predictive modelling where what-if scenarios forecast the consequences of changing policies or conditions. Traditional HR analytics focuses on the present, that is, items such as turnover and cost per hire. But most organizations lacked a consistent and general view of the workforce and thus needed HR analytics to perform workforce optimization and hence it became important for HR to develop IT and finance analytical skills and capabilities to produce better Return on Investment (ROI). [1], [2], [3]

Further advancement of technologies when combined with predictive analytics exponentially enhanced HR purposes in last decade. HRPA generates insights that cannot be achieved through traditional benchmarking as HR analytics is reactive and an evidence-based decision system whereas HRPA is proactive and fact-based decision system. [4]

Three significant changes that have really created a hunger for predictive analytics in HR and these are: [5]

- i. Major boost in computing power and its affordability
- ii. HR big data digitally accessible via cloud storage for

- iii. processing
Global talent war to protect and pursue talent streams.

Predictive analytics is unlike descriptive analysis which considers external benchmarking data and involves tables, reports, ratios, metrics, dashboards or complex maths; it is about data-derived insights that drive better decisions. It includes statistical techniques, machine learning methods, and data mining models that analyse and extract existing and historical facts to make predictions. It enables organizations to analyse the past and look forward to spot trends in key factors related to voluntary termination, absences and other sources of risk. Predictive analytics involves models of organizational systems for prediction of future outcomes and realize the significances of hypothetical changes in organizations. Predictive analytics have led to prescriptive analytics where HR gets decision options to optimize performance and reshape entire HRM decision making. [6]

Being an evolving phenomenon HRPA has much scope for HR purposes in future. Predictive analytics might be unexplored zone for HR, therefore to fully realize its profits; HR personnel need to team up with other business units and customer-facing functions to understand how they pull data and analytics to create value. [7] HRPA faces constraints of training and resistant to adapt from HR Personnel for e.g. a score of an employee engagement survey may mean different things to different lines of business, and regions around the world, depending on business purposes, economic actualities and workforce size. Collin says, [8] HRPA is an art and a critical skill to bring out the business insights of data analysis. There are opportunities for HRPA in HRM to expand due to necessary boost provided to enhance HR functions, to better business outcomes and to improve ROI. TimesJobs.Com COO Vivek Madhukar indicated to the fact that the ability to move from gut-based judgements to data-driven decision-making is making HRPA the future of HRM in India. Over 55 per cent of organizations feel that HRPA predictions help to secure quality hires. [9]

- *Sujeet N. Mishra is currently pursuing master's degree program in computer science engineering in SRMU, India, PH-+918898181778. E-mail: er.sujeetnmishra@gmail.com*
- *Dev Raghvendra Lama is currently working as assistant professor in faculty of computer science engineering in SRMU, India. E-mail: dev.lama007@gmail.com*
- *Yogesh Pal is currently pursuing master's degree program in computer science engineering in SRMU, India, E-mail: er.yogeshpal15@gmail.com*

2 ASPECTS OF HRP A

2.1 Need

Organizations make sure the right people are in the right place at the right time by means of analytics. [10] To remain commercially relevant HRM needs to provide senior executives with a predictive analytics based justification for important talent related decisions. No organization is identical in terms of workforce, talent, environment, strategies, and market type. And hence one successful but fixed model cannot be applied to any function of HR. Only past data of the particular organization or its identical culture have ability to provide right decision for HRM. Thus HRP A becomes essential for industries which desire for bringing unique decision policies. The HR requires skills of technology and management both where technology is not limited to analytics. HR should be able to create insights into data and produce predictive models that optimize the organizational performance. Advent of advance machine learning programs and HR expert systems [11] has eased to achieve organizational objectives of human capital management (HCM), workforce planning, employee management, and performance management etc.

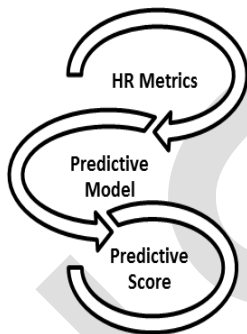


Fig. 1: HRP A Approach for Decision Making

2.2 Approach & Application

HR metrics are clearly defined to assess and collect essential data and then predictive models are built for each HR function which is needed to produce insights from historical data for future decision making in HRP A. [12]

2.2.1 HR Metrics:

The organizations use metric values to scale HR functions in terms of recruitment time, attrition level, employee turnover, and probability of success. Having concrete metrics is crucial to exhibit senior leaders and executives how strategic HR initiatives can help affect an organization's bottom line. There are 13 famous HR metrics [13] as follows;

- i. Monthly turnover rate
- ii. Revenue per Employee
- iii. Yield Ratio
- iv. Human capital cost
- v. HR to staff ratio

- vi. Return on investment
- vii. Promotion rate
- viii. Percentage female at management level
- ix. Employee absent rate
- x. Worker's compensation cost per employee
- xi. Worker's compensation incident rate
- xii. Overtime per individual contributor headcount
- xiii. Average employee age.

2.2.2 Predictive Modelling

Predictive model combines analytical algorithms and provide results in form of a value or probability scores based on which decisions are taken. HCM: 21 Model [14] is one such predictive model proposed for HRM strategies inspired from Dr. Jac Fitz-Enz which states four stages of HRM, 1.Scan 2.Plan 3.Produce 4.Predict and HRP A applies the same.

Predictive modelling [4], [15], [16] is applied on HR functions for decision making in various field of HRM as follows;

- i. Turnover Modelling: Predicts future turnover of business in specific functions, business units, geographies and countries by looking at factors such as commute time, time since last role change, and performance over time. Thus this can scale hiring efforts accordingly, reducing empty desk time and panic hiring, which can lead to lower cost, higher quality hiring.
- ii. Response Modelling: Use old advertising jobs data from previous campaigns to avoid contacting candidates or using channels that don't yield a response and focus on those channels that do work.
- iii. Predictive Retention Modelling: Identify high-risk employees, build profiles, predict vacancies and leadership needs, and understand how risk is distributed throughout the organization.
- iv. Risk Modelling: Develop a profile of candidates with a higher risk of leaving prematurely or performing below standard.
- v. Talent Forecasting: Being able to predict which new hires, based on their profile, and are likely to be high fliers and then moving them in to your high potential programs.

2.3 Impact

Vestas [17], a US wind turbine maker, changed recruitment and retaining policies after HRP A suggested women employees performs 5% better than men. Gallup, a global consulting company, in their recent survey found that employee engagement can help companies survive, and possibly even succeed, in tough economic times. It promotes that companies with high engagement have a 20% boost in efficiency and effectiveness. Cisco [18] used IBM SPSS analytics tool to transform the relationship between its HR analysts and its executive leaders. It predicted and prevented attrition level.

3 CONCLUSION

It is evident that industries cannot survive in the long run if they do not possess predictive analytics skills from the human resource management. The usefulness of predictive analytics is wider and hence application in all related areas of HRM is essential. HRPAs help organizations contain HR-related costs while optimizing business performance as well as employee engagement and satisfaction. HRPAs are rapidly changing and growing technologies which have the potential to achieve 100% accuracy in decision making for HR. By 2020, HRPA will fully take over traditional analytics in organizations.

ACKNOWLEDGMENT

The authors wish to thank Dr. Neeraj Kumar Tiwari and Dr. Bineet Kumar Gupta from faculty of Computer Science and Engineering, SRMU, India for their continuous support and guidance for the research work.

REFERENCES

- [1] L. Bassi, "Raging Debates in HR Analytics", *People & Strategy*, Vol. 34, Issue 2, 2011
- [2] M. Molefe, "From Data to Insights : HR Analytics in Organizations," Gordon Institute of Business Science, University of Pretoria, 11 Nov. 2013
- [3] L. Bassi and D. McMurrer, "A Quick Overview of HR Analytics: Why, What, How, and When?" Association for talent development, March 04, 2015
- [4] D. Handa and Garima, "Human Resource (HR) Analytics: Emerging Trend In HRM (HRM)", *IJRCM*, Vol. No. 5, Issue No. 06, June 2014, ISSN 0976-2183
- [5] K. Ladimeji, "5 Things that HR Predictive Analytics will Actually Predict." *Recruiter* (Jan. 23, 2013), sec. 1 p.1.
- [6] J Fitz-enz and J. R. Mattrox II, "Predictive Analytics for Human Resource." Wiley Publication, SAS Institute Inc., Cary, North America, USA, 2014 pp. 2-3
- [7] D. Ulrich, B. Schiemann and L. Sartain, "The Rise of HR: Wisdom from 73 Thought Leaders," HR Certification Institute, Alexandria, VA, Ed. 2015 pp. 19-23
- [8] C. Waxer, "HR Executives: Analytics Role Needs Higher Profile," *Data Informed*, 13 March, 2013
- [9] "Predictive Talent Analytics the Future of HR," Press Trust of India, Aug 28, 2015, http://www.business-standard.com/article/pti-stories/predictive-talent-analytics-the-future-of-hr-in-india-115082500643_1.html
- [10] T. S. Dey, & P. De, "Predictive Analytics in HR: A Primer," TCS White Paper, 2015, <http://www.tcs.com/SiteCollectionDocuments/White-Papers/Predictive-Analytics-HR-0115-1.pdf>
- [11] "Applying Advanced Analytics to HR Management Decisions," James C. Sesil, Pearson Publication, New Jersey, March 2014, pp. 13-25
- [12] B. Khatri, "Talent Analytics: Toolkit for Managing HR Issues," *Sai Om Journal of Commerce & Management*, Vol. 1, Issue 5 , May 2014, Online ISSN-2347-7571
- [13] J. Miller-Merrell, "13 Best HR & Workforce Metrics Formula Examples," *Blogging4Jobs* (3 April 2012)
- [14] E. Muscalu and A. Şerban, "HR Analytics for Strategic Human Resource Management", in Proc. 8th International Management Conference on Management Challenges for Sustainable Development, Bucharest, Romania, November 6th -7th, 2014, pp. 939.
- [15] M. Blankenship, "Managing and Measuring Talent Risk," SHRM Foundation Thought Leaders Conference, Oct 2011.
- [16] L. Smeyers, "7 Benefits of Predictive Retention Modeling (HR analytics)."

iNostix (May 6, 2013)

- [17] K. Mølgaard-Pedersen, "The Power of HR Analytics in Strategic Planning," SVP Global HR Competence Centre, Vestas, 2010
- [18] IBM, "Cisco Success Story", 2015, ibm.com/business-analytics

Stock Market Prediction System with Modular Neural Networks

Takashi Kimoto and Kazuo Asakawa
Computer-based Systems Laboratory
FUJITSU LABORATORIES LTD., KAWASAKI
1015 Kamikodanaka, Nakahara-Ku, Kawasaki 211, JAPAN

Morio Yoda and Masakazu Takeoka
INVESTMENT TECHNOLOGY & RESEARCH DIVISION
The Nikko Securities Co., Ltd.
3-1 Marunouchi 3-Chome, Chiyoda-Ku Tokyo 100, JAPAN

Abstract: This paper discusses a buying and selling timing prediction system for stocks on the Tokyo Stock Exchange and analysis of internal representation. It is based on modular neural networks[1][2]. We developed a number of learning algorithms and prediction methods for the TOPIX(Tokyo Stock Exchange Prices Indexes) prediction system. The prediction system achieved accurate predictions and the simulation on stocks trading showed an excellent profit. The prediction system was developed by Fujitsu and Nikko Securities.

1. Introduction

Modeling functions of neural networks are being applied to a widely expanding range of applications in addition to the traditional areas such as pattern recognition and control. Its non-linear learning and smooth interpolation capabilities give the neural network an edge over standard computers and expert systems for solving certain problems.

Accurate stock market prediction is one such problem. Several mathematical models have been developed, but the results have been dissatisfying. We chose this application as a means to check whether neural networks could produce a successful model in which their generalization capabilities could be used for stock market prediction.

Fujitsu and Nikko Securities are working together to develop TOPIX's a buying and selling prediction system.

The input consists of several technical and economic indexes. In our system, several modular neural networks learned the relationships between the past technical and economic indexes and the timing for when to buy and sell. A prediction system that was made up of modular neural networks was found to be accurate. Simulation of buying and selling stocks using the prediction system showed an excellent profit. Stock price fluctuation factors could be extracted by analyzing the networks.

Section 2 explains the basic architecture centering on the learning algorithms and prediction methods. Section 3 gives the outcome of the prediction simulation and results of the buying and selling simulation. Section 4 compares statistical methods and proves that stock prices can forecast by internal analysis of the networks and analysis of learning data.

2. Architecture

2.1 System Overview

The prediction system is made up of several neural networks that learned the relationships between various technical and economical indexes and the timing for when to buy and sell stocks. The goal is to predict the best time to buy and sell for one month in the future.

TOPIX is an weighted average of market prices of all stocks listed on the First Section of the Tokyo Stock Exchange. It is weighted by the number of stocks issued for each company. It is used similar to the Dow-Jones average.

Figure 1 shows the basic architecture of the prediction system. It converts the technical indexes and economic indexes into a space pattern to input to the neural networks. The timing for when to buy and sell is a weighted sum of the weekly returns. The input indexes and teaching data are discussed in detail later.

2.2 Network Architecture

2.2.1 Network Model

Figure 2 shows the basic network architecture used for the prediction system. It consists of three layers: the input layer, the hidden layer, and the output layer. The three layers are completely connected to form a hierarchical network.

Each unit in the network receives input from low-level units and performs weighted addition to determine the output. A standard sigmoid function is used as the output function. The output is analog in the [0,1] section.

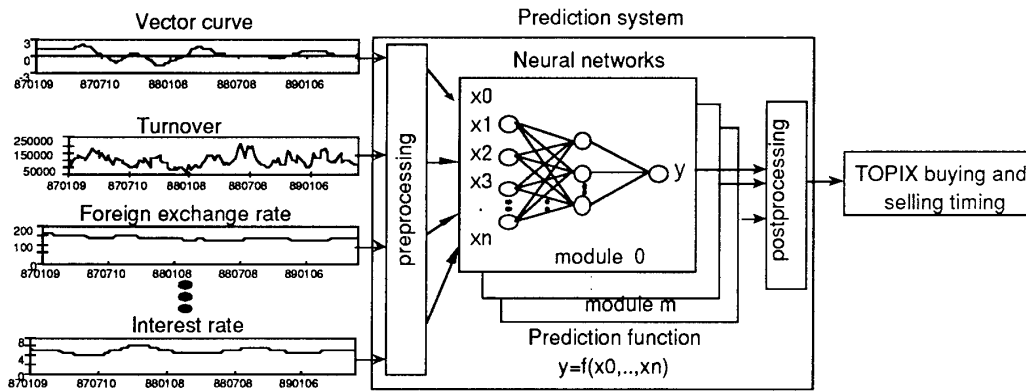


Figure 1 Basic architecture of prediction system

2.2.2 High-speed Learning Algorithm

The error back propagation method proposed by Rumelhart[3] is a representative learning rule for hierarchical networks. For high-speed learning with a large volume of data, we developed a new high-speed learning method called supplementary learning[4].

Supplementary learning, based on the error back propagation, automatically schedules pattern presentation and changes learning constants.

In supplementary learning, the weights are updated according to the sum of the error signals after presentation of all learning data. Before learning, tolerances are defined for all output units. During learning, errors are back-propagated only for the learning data for which the errors of output units exceed the tolerance. Pattern presentation is automatically scheduled. This can reduce the amount calculation for error back propagation.

As learning progresses, learning data for which tolerances are exceeded are reduced. This also reduces the calculation load because of the decreased amount of data that needs error back propagation. High-speed learning is thus available even with a large amount of data.

Supplementary learning allows the automatic change of learning constants depending on the amount of learning data. As the amount of learning data changes and learning progresses, the learning constants are automatically updated. This eliminates the need for changing learning parameters depending on the amount of learning data.

With supplementary learning, the weight factor is updated as follows:

$$\Delta w(t) = -(\epsilon / \text{learning_patterns}) \partial E / \partial W + \alpha \Delta w(t-1)$$

Where

ϵ : learning rate

α : momentum

learning_patterns : number of learning data items that require error back propagation

The value of ϵ is divided by the number of learning data items that actually require error back propagation. The required learning rate is automatically reduced when the amount of learning data increases. This allows use of the constants ϵ regardless of the amount of data.

As learning progresses, the amount of remaining learning data decreases. This automatically increases the

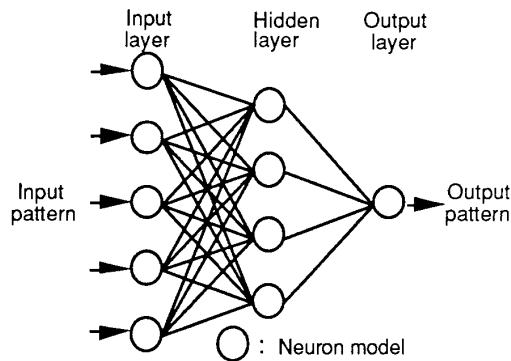


Figure 2 Neural network model

learning rate. Using this automatic control function of the learning constants means there is no need to change the constants ($\epsilon = 4.0$, $\alpha = 0.8$) throughout simulation and that high-speed learning can be achieved by supplementary learning.

2.3 Learning Data

2.3.1 Data Selection

We believe stock prices are determined by time-space patterns of economic indexes such as foreign exchange rates and interest rates and of technical indexes such as vector curves and turnover.

The prediction system uses a moving average of weekly average data of each index for minimizing influence due to random walk.

Table 1 lists some of the technical and economic indexes used. The time-space pattern of the indexes were converted into space patterns. The converted indexes are analog values in the [0,1] section.

Table 1 Input indexes

- | | |
|-------------------------------|--------------------------|
| 1. Vector curve | 2. Turnover |
| 3. Interest rate | 4. Foreign exchange rate |
| 5. New York Dow-Jones average | 6. Others |

2.3.2 Teaching Data

The timing for when to buy and sell is indicated as an analog value in the [0,1] section in one output unit. The timing for when to buy and sell used as teaching data is weighted sum of weekly returns. When the TOPIX weekly return is r_t , teaching data $r_N(t)$ is defined as:

$$r_t = \ln(\text{TOPIX}(t) / \text{TOPIX}(t-1)) \quad \text{TOPIX}(t) : \text{TOPIX average at week } t$$

$$r_N(t) = \sum_i \phi^i r_{t+i} \quad \phi : \text{Weight}$$

2.4 Preprocessing

Input indexes converted into space patterns and teaching data are often remarkably irregular. Such data is preprocessed by log or error functions to make them as regular as possible. It is then processed by a normalization function which normalizes the [0,1] section, correcting for the irregular data distribution.

2.5 Learning Control

In the TOPIX prediction system, we developed new learning control. It automatically controls learning iterations by referring to test data errors, thereby preventing overlearning. The learning control allows two-thirds of data in the learning period to be learned and uses the rest as test data in the prediction system. The test data is evaluation data for which only forward processing is done during learning, to calculate an error but not to back propagate it.

Our learning control is done in two steps. In the first step, learning is done for 5,000 iterations and errors against test data are recorded. In the second step, the number of learning iterations where learning in the first step suggests a minimum error against the test data is determined, and relearning is done that number of iterations. This prevents overlearning and acquires a prediction model involving learning a moderate number of times. In the second step, learning is done for at least 1,000 iterations.

2.6 Moving Simulation

For prediction of an economic system, such as stock prices, in which the prediction rules are changing continuously, learning and prediction must follow the changes.

We developed a prediction method called moving simulation. In this system, prediction is done by simulation while moving the objective learning and prediction periods. The moving simulation predicts as follows.

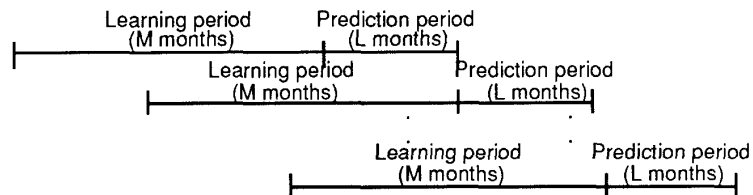


Figure 3 Moving simulation

As shown in Figure 3, the system learns data for the past M months, then predicts for the next L months. The system advances while

repeating this.

3. Result of Simulations

3.1 Prediction Simulation

We verified the accuracy of the prediction system by simulating prediction of the timing for when to buy and sell. We used historical data of stock prices, technical indexes and economical indexes.

The TOPIX prediction system improves its prediction accuracy by averaging prediction results of modular networks that learn for different learning data items. Four independent modular networks learn for four types of different learning data. Moving simulation is used with L as one month. The average of prediction outputs from these networks became the prediction output from the system. Prediction was thus repeated by moving simulation for each month to verify accuracy. Prediction was done for 33 months from January 1987 to September 1989.

Figure 5 shows the correlation coefficient between the predictions and teaching data and those of individual networks and prediction system. The prediction system uses the average of the predictions of each network. Thus the prediction system could obtain a greater correlation coefficient for teaching data than could be obtained with neural network prediction.

Table 2 Correlation coefficient

	Correlation Coefficient
Network1	0.435
Network2	0.458
Network3	0.414
Network4	0.457
System	0.527

3.2 Simulation for Buying and Selling Simulation

To verify the effectiveness of the prediction system, a simulation of buying and selling of stock was done. Buying and selling according to the prediction system made a greater profit than the buying and holding.

Buying and selling was simulated by the one-point buying and selling strategy, so performance could be clearly evaluated. One-point buying and selling means all available money is used to buy stocks and means all stocks held are sold at a time. In the prediction system, an output of 0.5 or more indicates buy, and an output less than an 0.5 indicates sell. Signals are intensified as they get close to 0 or 1.

The buying and selling simulation considered "buy" to be an output above some threshold and "sell" to be below some threshold. Figure 4 shows an example of the simulation results. In the upper diagram, the buy-and-hold performance (that is, the actual TOPIX) is shown as dotted lines, while the prediction system's performance is shown as solid lines. The TOPIX index of January 1987 was considered as 1.00, it was 1.67 by buy-and-hold at the end of September 1989. It was 1.98 by the buying and selling operation according to the prediction system. Use of the system showed an excellent profit.

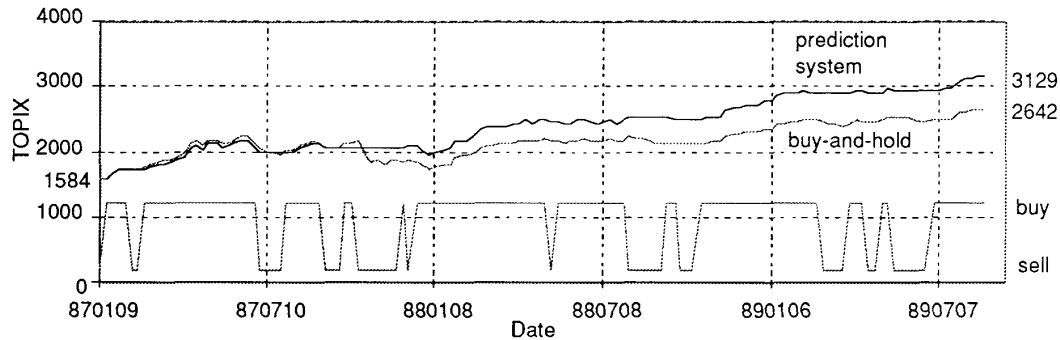


Figure 4 Performance of the prediction system

4. Analysis

4.1 Comparison with Multiple Regression Analysis

The timing for when to buy and sell stocks is not linear, so statistical methods are not effective for creating a model. We compared modeling with the neural network and with multiple regression analysis. Weekly learning data from January 1985 to September 1989 was used for modeling. Since the objectives of this test were comparison of learning capabilities and internal analysis of the network after learning, the network learned 100,000 iterations.

The hierarchical network that had five units of hidden layers learned the relationships between various

economic and technical indexes and the timing for when to buy and sell. The neural network learned the data well enough to show a very high correlation coefficient, the multiple regression analysis showed a lower correlation coefficient. This shows our method is more effective in this case. Table 3 shows the correlation coefficient produced by each method. The neural network produced a much higher correlation coefficient than multiple regression.

Table 3 Comparison of multiple regression analysis and neural network

	Correlation coefficient with teaching data
Multiple regression analysis	0.543
Neural network	0.991

4.2 Extraction of Rule

The neural network that learned from January 1985 to September 1989 (Section 4.1) was analyzed to extract information on stock prices stored during that period.

Cluster analysis is often used to analyze internal representation of a hierarchical neural network [5][6]. In 1987 stock prices fluctuated greatly. The hidden layer outputs were analyzed to cluster learning data. The cluster analysis was applied to the output values in the [0,1] sections of the five units of hidden layers. Clustering was done with an inter-cluster distance determined as Euclidian, and the clusters were integrated hierarchically by the complete linkage method. Figure 5 shows the cluster analysis results of the hidden layers in 1987. It indicates that bull, bear, and stable markets each generate different clusters.

From cluster analysis, characteristics common to data that belong to individual clusters were extracted by analyzing the learning data. Figure 6 shows the relationships between TOPIX weekly data and six types of clusters in 1987.

This paper analyzes the factors for the representative bull (2)(3) and bear (6) markets in 1987 as follows.

Data (2) and (3) belong to different clusters but have similar characteristics. Figure 7 shows learning data corresponding to clusters (2) and (3). The horizontal axis shows some of the index-

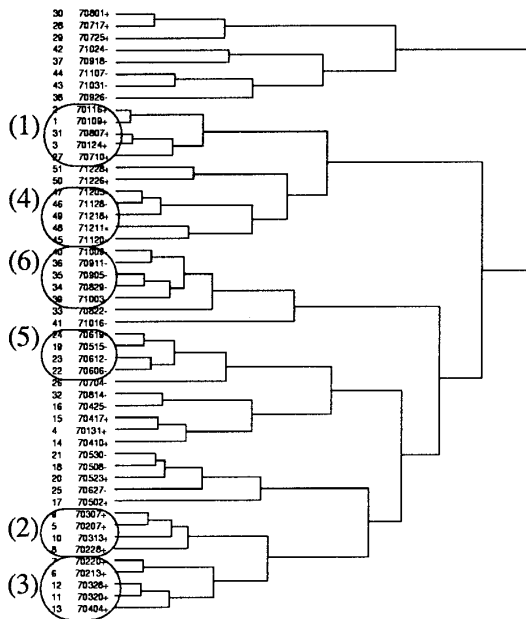


Figure 5 Cluster analysis results from 1987

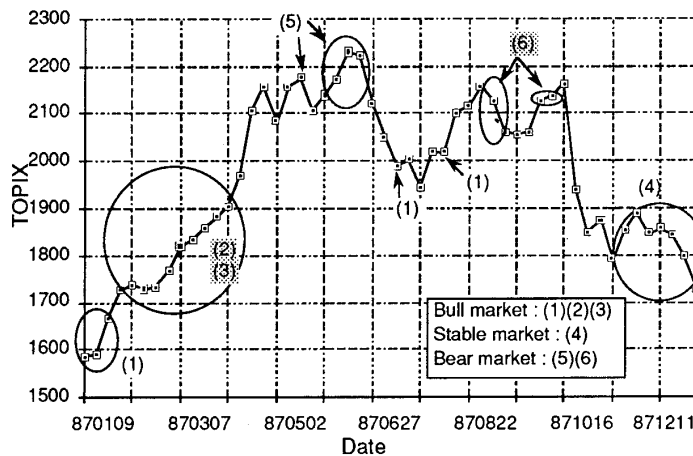


Figure 6 TOPIX in 1987

es of the neural network. The vertical axis show the value of each index. For example, New York Dow-Jones average is low when it is close to 0 and is high when close to 1.

This diagram suggests that the vector curve in the bull market during February to the beginning of April in 1987 were high enough to indicate a high-price zone. At the same time, however, the high turnover kept the market going up. Also, the low interest rates and high New York Dow-Jones average helped pushed up stock prices.

Figure 8 shows the learning data corresponding to (6) in Figure 6. It is obvious that the high interest rate pulled prices down.

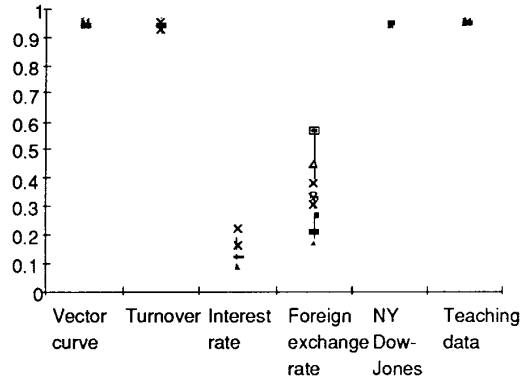


Figure 7 Input indexes for bull market(2)(3)

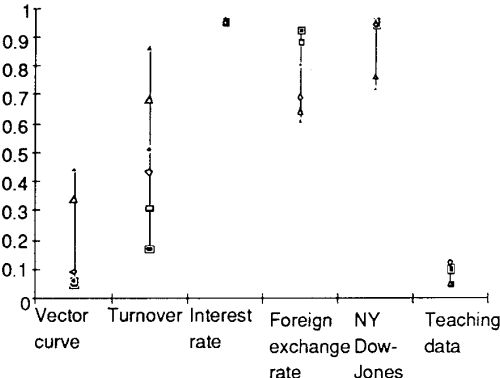


Figure 8 Learning data in bear market (6)

Part of the learning data in 1987 was analyzed. It was proved that the causes of stock price fluctuation could be analyzed by extracting the characteristics common to the learning data in the clusters obtained by cluster analysis of the neural network.

5.0 Further Research

The following subjects will be studied.

- Using for actual stock trading

The current prediction system uses future returns to generate teaching data. A system in which teaching data is generated in combination with a statistical method must be developed.

- Adaptation of network model that has regressive connection and self-looping

The current prediction system requires much simulation to determine moving average. Automatic learning of individual sections requires building up a prediction system consisting of network models fit to time-space processing.

6.0 Summary

This paper has discussed a prediction system that advises the timing for when to buy and sell stocks. The prediction system made an excellent profit in a simulation exercise. The internal representation also was discussed and the rules of stock price fluctuation were extracted by cluster analysis.


For developing the prediction system, Nikko Securities offered investment technology and know-how of the stock market and Fujitsu offered its neural network technology. Fujitsu and Nikko Securities are studying further to build up more accurate economic prediction systems.

References:

- [1] S. Nagata, T. Kimoto and K. Asakawa, Control of Mobile Robots with Neural Networks, INNS, 1988, 349
- [2] H. Sawai, A. Waibel, et al., Parallelism, Hierarchy, Scaling in Time-Delay Neural Networks for Spotting Japanese Phonemes/CV-Syllables, IJCNN vol II, 1989, 81-88
- [3] D. E. Rumelhart, et al., Parallel Distributed Processing vol. 1, The MIT Press, 1986
- [4] R. Masuoka, et al., A study on supplementary learning algorithm in back propagation, JSAI, 1989, 213-217.
- [5] T. J. Sejnowski, C. R. Rosenberg, Parallel Networks that Learn to Pronounce English Text, Complex Systems, 1, 1987
- [6] R. Paul Gorman, T. J. Sejnowski, Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets, Neural Networks, Vol 1, No 1, 1988, 75-90

The goal is business effectiveness through 'verticalization,' usability, and integration with operational systems.

EMERGING TRENDS IN BUSINESS ANALYTICS

The field of business analytics has improved significantly over the past few years, giving business users better insights, particularly from operational data stored in transactional systems. An example is e-commerce data analysis, which has recently come to be viewed as a killer app for the field of data mining [5, 6]. The data sets created by integrating clickstream records generated by Web site activity with demographic and other behavioral data dwarf, in size and complexity, the largest data warehouses of just a few years ago [4]. The result is massive databases requiring a mix of automated analysis techniques and human effort to give business users strategic insight about the activity on their sites, as well as about the characteristics of the sites' visitors and customers. With many millions of clickstream records generated every day, aggregated to customer-focused records with hundreds of attributes, there is a clear need for automated techniques for finding patterns in the data. Here, we discuss the technology and enterprise-adoption trends associated with business analytics.

The key consumer is the business user, whose job, possibly in merchandising, marketing, or sales, is not directly related to analytics per se, but who typically uses analytical tools to improve the results of some business process along one or more dimensions (such as profit and time to market). Fortunately, data mining,¹ analytic applications, and business intelligence

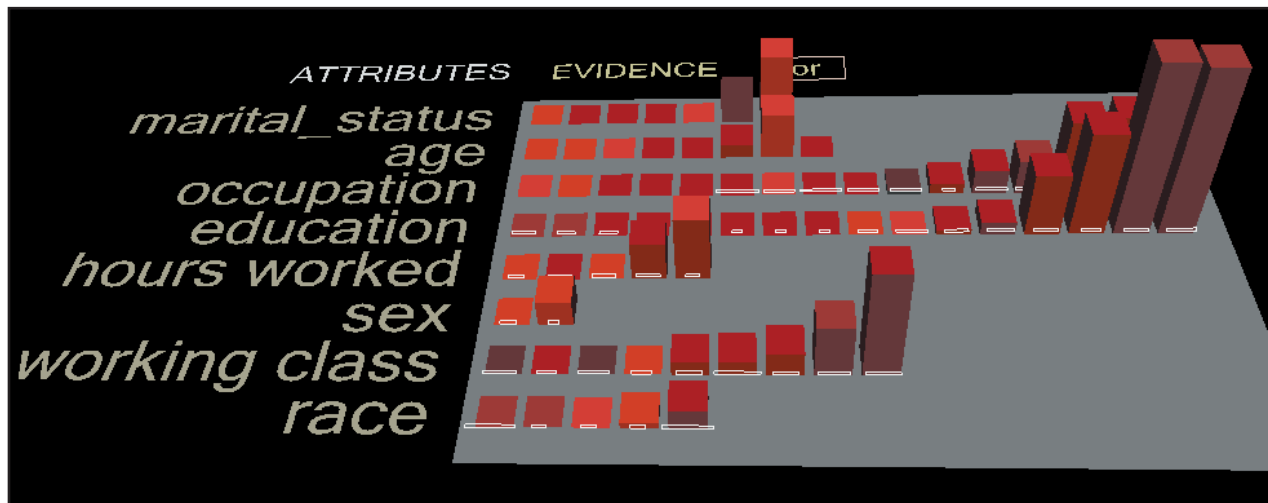
systems are now better integrated with transactional systems than they once were, creating a closed loop between operations and analysis that allows data to be analyzed and the results reflected quickly in business actions. Mined information today is deployed to a broader business audience taking advantage of business analytics in its everyday activities. Analytics are now routinely used in sales, marketing, supply chain optimization, and fraud detection [2, 3].

Business Users

Even with these advances, business users, while expert in their particular areas, are still unlikely to be expert in data analysis and statistics. To make decisions based on the data collected by and about their organizations, they must either rely on data analysts to extract information from the data or employ analytic applications that blend data analysis technologies with task-specific knowledge. In the former, business users impart domain knowledge to the analyst, then wait for the analyst to organize and analyze it and communicate back the results. These results typically raise further questions, hence several iterations are necessary before business users can actually act on the analysis. In the latter, analytic applications incorporate not only a variety of data mining techniques but provide recommendations to business users as to how to best analyze the data and present the extracted information. Business users are expected to use it to improve performance along multiple metrics. Unfortunately, the gap between relevant analytics and users' strategic business needs is significant. The gap is characterized by several challenges:

¹The terms data mining and analytics are used interchangeably here to denote the general process of exploration and analysis of data to discover new and meaningful patterns in data. This definition is similar to those in [2, 3] where it's referred to as knowledge discovery.

By RON KOHAVI, NEAL J. ROTHLEDER, AND EVANGELOS SIMOUDIS



A visualization of a Naive Bayes model for predicting who in the U.S. earns more than \$50,000 in yearly salary. The higher the bar, the greater the amount of evidence a person with this attribute value earns a high salary.

Cycle time. The time needed for the overall cycle of collecting, analyzing, and acting on enterprise data must be reduced. While business constraints may impose limits on reducing the overall cycle time, business users want to be empowered and rely less on other people to help with these tasks.

Analytic time and expertise. Within the overall cycle, the time and analytic expertise necessary to analyze data must be reduced.

Business goals and metrics. Unrealistic expectations about data mining “magic” often lead to misguided efforts lacking clear goals and metrics.

Goals for data collection and transformations. Once metrics are identified, organizations must collect and transform the appropriate data. Data analysis is often an afterthought, limiting the possible value of any analysis.

Distributing analysis results. Most analysis tools are designed for quantitative analysts, not the broader base of business users who need the output translated into language and visualizations appropriate for business needs.

Integrating data from multiple sources. The extract-transform-load (ETL) process is typically complex, and its cost and difficulty are often underestimated.

The Driving Force

The emerging trends and innovations in business analytics embody approaches to these business challenges. Indeed, it is a very healthy sign for the field that regardless of the solution-process, technology, system integration, or user interface, business problems remain the driving force.

“*Verticalization.*” In order to reduce discovery cycle time, facilitate the definition and achievement of business goals, and deploy analysis results to a wider audience, developers of analytical solutions started verticalizing their software, or customizing applications within specific industries. The first step toward verticalization was to incorporate task-specific knowledge; examples include: knowledge about how to analyze customer data to determine the effectiveness of a marketing campaign; knowledge of how to analyze clickstream data generated by a Web site to reduce shopping cart abandonment and improve ad effectiveness; knowledge about how an investment bank consolidates its general ledger and produces various types of forecasts; and how an insurance company analyzes data in order to provide an optimally priced policy to an existing customer.

In the process of incorporating industry-specific knowledge, companies are also able to optimize the performance of their applications for specific industries. For example, a company that developed an analytic application for budgeting and forecasting targeted at the financial services industry determined that its online analytical processing, or OLAP, engine’s execution speed could be optimized by limiting to nine the number of dimensions it had to handle, a number deemed sufficient for the particular application in that industry.

The use of industry-specific knowledge is not limited to the data mining components of analytic applications but also affects how the extracted information is accessed and presented. For example, organizations in the financial services, retail, manufacturing, utilities, and telecommunications industries increasingly want their field personnel to have access to business analytic information through wireless devices. Analytic application vendors are now developing technologies to automatically detect wireless devices and their form

factors, automatically tailoring analysis results to fit the capabilities of a particular device. For example, if the information is to be displayed on a phone supporting the Wireless Access Protocol (implying small screen size), it may be necessary to automatically summarize text, abbreviate words, and limit the use of graphics by automatically selecting only the most relevant figures.

Comprehensible models and transformations. In light of the need to let business users analyze data and quickly gain insight, and aiming for the goal of reducing reliance on data mining experts, comprehensible models are more popular than opaque models. For example, in the KDD-Cup 2000 [5], a data mining competition in which insight was important, the use of decision trees, generally accepted as relatively easy to understand, outnumbered other methods more than two to one.

Business users do not want to deal with advanced statistical concepts; they want straightforward visualizations and task-relevant outputs. The figure outlines a Naive Bayes model for predicting who in the U.S. earns more than \$50,000 in yearly salary. Instead of the underlying log conditional probabilities the model actually manipulates, the visualization uses bar height to represent evidence for each value of a contributing factor listed on the left and color saturation to signify confidence of that evidence [1]. For example, evidence for higher salaries increases with age, until the last age bracket, when it drops off; evidence for higher salaries increases with years of education, number of hours worked, and certain marital status and occupations. Note also the visualization shows only a few attributes determined by the mining algorithm to be the most important ones, highlighting to business users the most critical attributes from a larger set. Other examples of visualizing data and data mining models are in [7, 9].

Part of the larger system. The needs of data analysis are being designed into systems, instead of being an afterthought, typically addressing the following areas:

Data collection. You cannot analyze what you do not collect, so collecting rich data is critical. For example, e-commerce systems can collect attributes ranging from the user's local time, screen resolution (useful for determining the quality of images to send), and network bandwidth.

Generation (and storage) of unique identifiers. In order to help merge information from several records and remove duplicate records, systems must generate unique keys to join data and store them. For example, all clickstream records in the same session should store the session IDs so they can be joined later to session records stored in other tables.

Integration with multiple data sources. Analysis is more effective when data is available from multiple sources. For example, in customer analytics, data should be merged from multiple touchpoints, including the Web, call centers, physical stores, wireless access, and ad campaigns (both direct and online). Behavioral data can be more powerful when overlaid with demographic and socioeconomic data from other sources.

Hardware sizing. Analysis requires hardware capable of dealing with large amounts of data. Some organizations have traditionally underestimated the need for sophisticated IT infrastructure and the hardware needed to make timely analysis feasible.

In new areas. During the past few years, recognition of the strategic value of business analytics has led to significant developments in business applications that analyze customer data. They've been used to reduce customer attrition, improve customer profitability, increase the value of e-commerce purchases, and increase the response of direct mail and email marketing campaigns.

This success has paved the way for new applications; three are particularly promising: supply chain visibility, price optimization, and work force analysis. Organizations have automated portions of their supply chains, enabling collection of significant data about inventory, supplier performance, and logistics of materials and finished goods. Newer applications analyze this data to provide insights about the performance of suppliers and partners, material expenditures, accuracy of sales forecasts for controlling materials inventory, accuracy of production plans, and accuracy of plans for order delivery.

The wide adoption of customer relationship management, or CRM, and supply chain management software has allowed enterprises to fully interface and integrate their demand and supply chains. Based on this integration, they are better able to capture up-to-the-minute data about demand for a particular product, as well as data of similar granularity about the supply of corresponding data. Analyzing these two data streams, organizations optimize the price of a particular product along several dimensions so demand meets available supply; for example, the price of a product may be different through one channel (such as the Web) than through another (such as a retail store). Price optimization allows any type of organization to maximize profit margins for each item sold while reducing inventory.

Once organizations are able to analyze data about their customers and their suppliers, they begin analyzing data about their employees, too. A new generation

of analytic applications allows enterprises to identify work force trends (such as attrition rates) and perform HR management tasks (such as compensation and benefits analyses). Companies whose cost or revenue model is dependent on hourly models (such as contact centers and systems integrators) use it to optimize staffing levels and skill requirements while minimizing the number of employees who are not able to bill.

Integration with action and measurement. With increased understanding of and experience in analytics, business users become more demanding and discerning, particularly when it comes to action based on insight and return on investment (ROI) [8]. Increasingly, analytics users ask two key questions: How do I turn discovered information into action? and How can I determine the effect of each action on my organization's business performance? Tales of data mining applications used to end with some novel analytical result; today, however, it is increasingly necessary that solutions use analytic results as a starting point toward the critical next steps of action and measurement. It is no longer enough for, say, cluster-discovery algorithms to uncover interesting groups of customers. The successful analytic solution must make it easier for the user to grasp the significance of these clusters in the context of a business action plan; for example, these people have a propensity for purchasing new fashions. Achieving these results requires nontrivial transformations from the base statistical models. Traditionally, achieving these results necessitated the participation of expert human analysts.

Integrating analytics with existing systems is a key to both action and measurement. For example, if the analytic application identifies customers likely to respond to a promotion, but it takes a cadre of IT specialists to incorporate the relevant data into the advertising system to run the promotion, the results are unlikely to be used, as IT specialists are likely to be in short supply. Similarly, if promotion-targeting solutions enable distribution of catalogs with optimized promotions, but the order submission system isn't closely tied back into the customer analytics, the resulting lag in ROI reports inhibits timely adjustment in the next catalog mailing. Efforts to integrate operations and analytic systems have seen major initiatives over the past five years, including entire product lines whose value proposition is the optimization of the collect-analyze-act-measure cycle.

Conclusion

Recent innovations and trends in business analytics—spanning organizations and technical processes, new technologies, user interface design, and system integration—are all driven by business value. Business

value is measured in terms of progress toward bridging the gap between the needs of the business user and the accessibility and usability of analytic tools. In order to make analytics more relevant and tangible for business users, solutions increasingly focus on specific vertical applications tailoring results and interfaces for these users, yielding human-level insight. For ease of use, simpler and more effective deployment, and optimal value, analytics are also increasingly embedded in larger systems. Consequently, data collection, storage, processing, and other issues specific to analytics are incorporated into overall system design.

Broadening the effects of analytics in the business process, solutions go beyond customer-centric applications to support sales, marketing, supply chain visibility, price optimization, and work force analysis. Finally, in order to achieve the greatest possible business value, analytic solutions have to produce results that are actionable, along with ways to measure the effects of key changes. ■

REFERENCES

1. Becker, B., Kohavi, R., and Sommerfield, D. Visualizing the simple Bayesian classifier. In *Information Visualization in Data Mining and Knowledge Discovery*, chapt. 18, U. Fayyad, G. Grinstein, and A. Wierse, Eds. Morgan Kaufmann Publishers, San Francisco, 2001, 237–249.
2. Berry, M. and Linoff, G. *Mastering Data Mining*. John Wiley & Sons, Inc., New York, 2000.
3. Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery and Data Mining*, chapt. 1, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. AAAI Press, Menlo Park, CA, and the MIT Press, Cambridge, MA, 1996, 1–34.
4. Kimball, R. and Merz, R. *The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse*. John Wiley & Sons, Inc., New York, 2000.
5. Kohavi, R., Brodley, C., Frasca, B., Mason, L., and Zheng, Z. KDD-Cup 2000 organizers' report: Peeling the onion. *SIGKDD Explor.* 2, 2 (Dec. 2000), 86–98; see www.ecn.purdue.edu/KDDCUP.
6. Kohavi, R. and Provost, F. Applications of data mining to electronic commerce. *Data Min. Knowl. Disc.* 5, 1/2 (Jan.-Apr. 2001); see robotics.stanford.edu/users/ronnyk/ecommerce-dm.
7. Lee, J., Podlaseck, M., Schonberg, E., and Hoch, R. Visualization and analysis of clickstream data of online stores for understanding Web merchandising. *Data Min. Knowl. Discov.* 5, 1/2 (Jan.-Apr. 2001).
8. Souza, R., Manning, H., and Gardiner, K. How to measure what matters. *Forrester Rep.* (May 2001).
9. Thearling, K., Becker, B., DeCoste, D., Mawby, B., Pilote, M., and Sommerfield, D. Visualizing data mining models. In *Information Visualization in Data Mining and Knowledge Discovery*, U. Fayyad, G. Grinstein, and A. Wierse, Eds. Morgan Kaufmann Publishers, San Francisco, 2001.

RON KOHAVI (ronnyk@cs.stanford.edu) is senior director of data mining at Blue Martini Software, San Mateo, CA.

NEAL J. ROTHLEDER (nealr@digimine.com) is director of analytic technology at DigiMine, Inc., Bellevue, WA.

EVANGELOS SIMOUDIS (evangelos.simoudis@apax.com) is a partner at Apax Partners, Palo Alto, CA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

© 2002 ACM 0002-0782/02/0800 \$5.00

THE USE OF REGRESSION ANALYSIS IN MARKETING RESEARCH

DUMITRESCU Luigi

Lucian Blaga University of Sibiu, Romania

STANCIU Oana

Lucian Blaga University of Sibiu, Romania

TICHINDELEAN Mihai

Lucian Blaga University of Sibiu, Romania

VINEREAN Simona

Lucian Blaga University of Sibiu, Romania

Abstract:

The purpose of the paper is to illustrate the applicability of the linear multiple regression model within a marketing research based on primary, quantitative data. The theoretical background of the developed regression model is the value-chain concept of relationship marketing. In this sense, the authors presume that the outcome variable of the model, the monetary value of one purchase, depends on the clients' expectations regarding seven dimensions of the company's offer. The paper is structured in two parts. In the first part, a brief literature review enumerates the main multivariate data analysis methods used in marketing research and describes the general linear multiple regression model and its assumptions. The second part explains a set of procedures specific for the regression analysis.

Keywords: *Linear Multiple Regression, Marketing Research, Multivariate Data Analysis Methods, Relationship Marketing*

1. Introduction

The core of decision making is information. Information is mainly the result of data analysis or data interpretation, thus it represents a result of applying different methods or techniques for creating value out of the gathered raw data. The

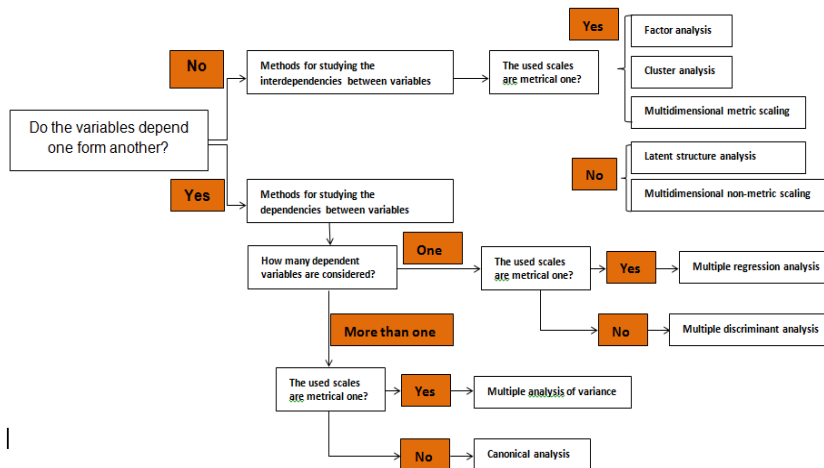
instruments used to filter the raw data into a valuable piece of information differ according to the nature of the underlying data: qualitative or quantitative data. Qualitative data is unstructured data, collected through methods and techniques like: focus group interviews, depth interviews and projective techniques. Quantitative data is measurable data obtained through methods like: survey, observation, experiment and simulation. Measurement and scaling are the two main concepts used in defining and analyzing quantitative data. Thus, measurement represents the assignment of numbers or other symbols to characteristics of objects according to certain prespecified rules and scaling a generation of continuum upon which measured objects are located (Malhotra, 2010).

There are four main criteria upon which the methods used for data analysis differ within a marketing research: the used scale type (metric or non-metric scales), the number of researched samples (one, two or more samples), the relation between the researched samples (dependent or independent) and the number of used variables (one, two or more variables) (Catoiu, 2010). According to this classification, the multivariate data analysis techniques (techniques which use more than two variables) can differ if:

- a. the considered variables are dependent one from another (causal relation) or if there is only a associative relation between them
- b. there are one, two or more dependent variables within a causal relation
- c. a metric scale or a non-metric scale is used for the considered variables.

The following figure contains the most known and used multivariate data analysis techniques:

Figure 1 – Multivariate Data Analysis Techniques



Source: Catoiu (2009, pg.552)

As seen in the figure above, multiple regression analysis is a technique that studies the dependency relation between one dependent variable and two or more independent variables, all measured on metric scales.

2. Literature Review

The combining of variables into a dependency relation is done by the economic theory in general and by the marketing theory in particular (e.g. one company's sales depend on the spent advertising money, different forms of price-bundles, the number of distribution channel and the size and quality of the general offer). Through empirical investigation, the theory is confirmed or not by the actual data or underlying behavior. One such instrument or method that confronts the theory with the actual behavior is the classical multiple linear regression model. This powerful instrument is used in various areas of economic research, such as international economics (Burnete, 2012) and macroeconomics (Opreana, 2010) and business research (Pop, 2010). The classical linear multiple regression model studies the linear relationship between one dependent variable and two or more independent variables.

The general form of the model is:
$$Y = f(X_1, X_2, X_3, \dots, X_k) + \varepsilon$$
$$= X_1 \beta_1 + X_2 \beta_2 + \dots + X_k \beta_k + \varepsilon ,$$

where Y is the dependent or explained variable and X_j , with $j=1, k$ are the independent or explanatory variables. The symbol ε represents the disturbance of the model or the residual variable. The set of independent variables explain to a certain extend the dependent variable; the rest of it is contained within the disturbance. The disturbance arises for several reasons, primary because it is not possible to capture every influence on a specific economic variable in a model, no matter how elaborate the model is and due to errors in measurement (Green, 2003).

Because of several constrains (time, cost, human resource, etc.), most of the marketing researches use statistical representative samples from which the data is collected. In this context, the linear multiple regression model has the following form:

$$y_i = x_{i1} \beta_1 + x_{i2} \beta_2 + \dots + x_{ik} \beta_k + \varepsilon ,$$

with $i=1, n$, where n is the used sample size.

The purpose of the model is to estimate the underlying unknown parameters of the model used further to check the validity of the stated theory and infer the sample's results on population level.

Because the regression model is based on two parts (a deterministic part: the independent variables and a random part: disturbance), the used data should respect some assumptions for proper model estimation and inference procedures (Green, 2003). The following section contains the six standard assumptions regarding the

multiple regression model that will be explained through SPSS procedures in the research part of the paper.

According to Green (2003) the assumptions of the linear multiple regression model are:

1. Linearity: the model specifies a linear relationship between the outcome variable and the independent variables. Therefore, the mean value of the outcome variable for each increase of the predictors lies on a straight line.
2. Full rank (multicollinearity): there is no exact linear relationship between any of the independent variables in the model.
3. Exogeneity of the independent variables: any estimated residual value is not a function of the independent variables. This means that the predictors do not contain useful information about the residual variable.
4. Homoscedasticity and nonautocorrelation: the residuals have a constant variance σ^2 within every level of the predictors and are not correlated one with each other.
5. Exogenously generated data: the process generating the data operates outside the model assumptions, thus independently of the process that generates the residuals.
6. Normal distribution: The residuals are normally distributed.

3. Research methodology

The purpose of the paper is to illustrate the application of the linear multiple regression model within a marketing research that uses primary, quantitative data as input. In this sense, five research objectives (O1 – O5) were developed accordingly to the assumptions of the linear multiple regression model.

Objective 1: To check if the developed model is a linear model.

Objective 2: To check the existence of multicollinearity between the predictors.

Objective 3: To check the independence of the residuals.

Objective 4: To check the homoscedasticity and non-autocorrelation of the residuals.

Objective 5: To check the normality of the residuals.

The developed regression model is based on the value chain concept within the relationship marketing theory (Bruhn, 2010). The company's outputs are considered the results of a complex interaction between consumer specific factors (observed or deduced by the company) which, themselves are generated by the company's actions. One of the deduced, customer specific dimensions is customer satisfaction. Customer satisfaction is defined by Oliver (1999) as a complex process of the information analysis. Bloemer (1998) describes the customers' satisfaction regarding a certain brand as an experience felt by them after consuming that brand, as a subjective evaluation of the way in which the brand's performance equals or not,

performs or not to their initial expectations. Generally, satisfaction is the pleasure or disappointment felt by someone in the moment when he compares his initial expectations regarding the brand with its actual performance obtained through consumption (Kotler, 2009). The result of the evaluation process can be favorable for the client, if the brand's performance exceeds his initial expectations regarding that certain brand (the client is satisfied), or not, if the brand performance is under the initial expectation level (the client is unsatisfied or even annoyed, and feels abused). A latent or fragile satisfaction is felt if the brand performance equals the initial expectations. From the mentioned satisfaction theory, the authors assume that the dimensions which form the satisfaction regarding a company's offer influence the monetary value spent by the customer within his last commercial contact with the company. By considering satisfaction as a subjective evaluation process (simply said as a mental difference), a further presumption could consider that the initial expectations of a customer's regarding the company's offer (perceived through its dimensions) may influence in some measure the spent monetary value.

In our particular case, we have chosen a fuel distributing company which has one sell point in the town of Sibiu, Romania. Moreover, we have considered the following seven underlying dimensions of clients' satisfaction: time spent within the fuel company, fuel quality, personnel politeness, shop diversity, personnel response to the clients' problems, auxiliary services that can be used by the clients and personnel readiness to help the clients. The clients' expectations regarding the mentioned dimensions were measured through nine-point interval scales with the scale poles representing the minimum/maximum expectation for the considered dimension (table 1).

Table 1 – Example of scale used for measuring the expectation regarding the fuel quality

E1	Fuel quality	Expected level of performance								
		Low				High				
		1	2	3	4	5	6	7	8	9

These seven variables are considered independent variables or predictors which influence one client's spent monetary value within his last contact with the fuel-company. The monetary value was measured in Lei (Romanian currency) using a ratio scale. As information source, end-customers of the fuel station were selected; they represent an external information source (clients who are part of the external environment of the fuel company), a primary information source (the obtained data is analyzed for the first time by the authors), and also they represent a free information source. A survey was used as research method and a questionnaire as research instrument. The sample size was represented by 10 respondents, for which the selection variables were two demographic variables (age [at least 18 years old] and possession of a car) and a behavioral variable (people using the car for at least four times per week). The data collection period was from March 1 until April 8, 2011. The

small sample size of 10 will surely affect the regression model consistency and assumptions, but, nevertheless, not accepting the model or rejecting some model assumptions don't dilute the value of the regression analysis.

The data was analyzed using the statistical software package SPSS V.19. Regression analysis requires a set of procedure which are automatically computed and displayed (through outputs) by the mentioned software.

A first output (table 2) of the analysis includes some descriptive statistics (mean and standard deviation) of the used variables (explained variable and explanatory variables).

Table 2 – Descriptive Statistics of the used variables

Variable name	Mean	Standard deviation	Number of cases
Monetary value (MV)	152	84,564	10
Time spent within the fuel-station (TS)	5,7	1,494	10
Fuel quality (FQ)	7,5	0,707	10
Personnel politeness (PP)	6,5	1,509	10
Shop diversity (SD)	4,9	1,912	10
Personnel response to the clients' problems (PR)	6,2	1,398	10
Auxiliary services that can be used by the clients (AS)	5,3	1,16	10
Personnel readiness to help the clients (PRead)	6,4	1,35	10

Source: own computation

The results reveal that a Rompetrol customer has spent a mean value of 152 Lei within his last commercial contact with the company. The variation of the monetary value variable (MV) is pretty high within the customer base, a value of $R^2 = 0,556$ indicates that the values are dispersed from the mean. Further information are contained by the high mean value (7,5) and low standard deviation (0,707) of the variable *fuel quality* (Q), which can be interpreted as a high level of expectations for fuel quality. A high level may be the result of good past experience of the customer with the company and, more important, of the intrinsic value of this central attribute – fuel. The expectations regarding the shop diversity have the lowest tolerance level (4,9), showing that this offer attribute (shop diversity) displays little importance for the customer.

A second output (table 3) contains the values of the Pearson correlation coefficient and their statistical significance for each pair of the model variables.

Table 3 – Correlation Matrix between the used Model Variables

Pearson Correlation	MV	TS	FQ	PP	SD	PR	AS	PRead
MV	1	0,058	0,186	-0,235	0,311	0,306	0,073	0,021
TS	0,058	1	0,053	0,517	-0,012	0,510	0,250	0,672
FQ	0,186	0,053	1	0,469	0,123	-0,112	-0,203	0,233
PP	-0,235	0,517	0,469	1	-0,096	0,158	0,413	0,545
SD	0,311	-0,012	0,123	-0,096	1	0,175	-0,386	-0,026
PR	0,306	0,510	-0,112	0,158	0,175	1	0,233	0,718
AS	0,073	0,250	-0,203	0,413	-0,386	0,233	1	0,483
PRead	0,021	0,672	0,233	0,545	-0,026	0,718	0,483	1
Sig. (1-tailed)	MV	TS	FQ	PP	SD	PR	AS	PRead
MV	.	0,437	0,304	0,257	0,191	0,195	0,421	0,477
TS	0,437	.	0,443	0,063	0,487	0,066	0,243	0,017
FQ	0,304	0,443	.	0,086	0,367	0,379	0,287	0,259
PP	0,257	0,063	0,086	.	0,396	0,331	0,118	0,051
SD	0,191	0,487	0,367	0,396	.	0,315	0,135	0,472
PR	0,195	0,066	0,379	0,331	0,315	.	0,259	0,010
AS	0,421	0,243	0,287	0,118	0,135	0,259	.	0,079
PRead	0,477	0,017	0,259	0,051	0,472	0,010	0,079	.

Source: own computation

The values within the correlation matrix are Pearson coefficients computed for each pair of the model variables. These values have no informational power if there is no statistical significance of the relationship they represent. It seems that there is *more* correlation between the explanatory variables than between the outcome and the predictors. Relative intense and statistical significant correlations (for a significance level of $\alpha=0,1$) are found between PP-TS (0,517; p-value=0,063); PR-TS (0,51; p-value=0,066); PRead-TS (0,672; p-value=0,017); PP-FQ (0,468, p-value=0,086); PRead-PP (0,545, p-value=0,051); PRead-PR (0,718, p-value = 0,01) and PRead-AS (0,483, p-value=0,079). Although the theory suggests (Field, 2006) that substantial correlation between the predictors (multicollinearity) might appear if the Pearson coefficient value is $>0,9$, the obtained results represent a first sign of **multicollinearity** within the specified model,

The overall model summary is presented in the following table (table 4):

Table 4 – Overall Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
0,973	0,947	0,761	41,334	1,914

Source: own computation

The value of the multiple correlation coefficient R (0,973) certifies a strong linear correlation between the outcome variable (MV) and the predictors. Highly important is the coefficient of determination value (R Square = 0,947) which indicates the proportion of the outcome variable variance which is *determined* by the variation of the predictors. Thus, 94,7% of the variability in MV (monetary value) is influenced by the variability of the predictors, a great result which should awake a lot of question marks in every researcher's mind.

The adjusted R-Square value (0,761) gives us a clue regarding the possibility of generalizing the computed model from the used sample to the entire target population. This value should be as close as possible to the initial R-Square value (0,947). The difference between R-Square and adjusted R-Square ($0,947 - 0,761 = 0,186$) is interpreted as the additional variance (18,6%) induced within the variance of the outcome variable by the sample model relative to a population model. This high percentage states that the obtained model is mostly valid for the used sample, its generalization will bias the outcome. The last idea is an assumption which can be confirmed or not by the results of a model validation process. In most cases, a model validation process presumes a training set and a test set, a presumption that makes not the purpose of the present paper.

The Durbin-Watson statistic indicates whether the residuals are independent one from another (**nonautocorrelation of the residuals**). A residual is the difference between the observed value and the predicted value, thus a correlation of every two residuals would mean a correlation of the underlying observations. An independence of residuals is observed if the Durbin-Watson value is near to 2. In our case (D-W = 1,914), thus the assumption of residual autocorrelation is rejected.

Another output represents the ANOVA-table (table 5) which explains the variance of the outcome variable through the influence of the computed model and residuals.

Table 5 – ANOVA Statistics

ANOVA	Sum of Squares	df	Mean Square	F	Sig.
Regression	60943,004	7	8706,143	5,096	0,174
Residual	3416,996	2	1708,498		
Total	64360,000	9			

Source: own computation

The Total Sum of Squares determines the quantity of variance obtained if we have used the mean value of MV for making predictions. The Sum of Squares of the Residuals indicates the quantity of variance which is not explained or generated by the computed model. Thus, the Regression Sum of Squares stands for the variance of the predicted outcome variable generated by the computed model. All three variances are relative small, the main cause is represented by the reduced number of cases (10) upon which the regression model was built.

A high F-value shows the prediction consistency of the computed model relative to the residual influence. Its value is calculated as a proportion between the Regression Mean Square and Residual Mean Square (5,096). This F-value is positively related to the Regression and Residual Sum of Square, therefore high absolute Sum of Square values (more observation) and significant difference between Regression and Residuals Sum of Squares imply a high F-value. The statistical significance of 0,174 shows the probability that the computed F-value is obtained by chance. The present variance analysis reveals a high prediction consistency of the computed model relative to the residuals, but the small F-value indicates the lack of variance in the used variables (observed and predicted outcome variables) due to the reduces number of observations.

The estimated model coefficients and their statistical significance are presented in table 6:

Table 6 – Estimated Model Coefficients

	Unstandardized Coefficients		Standardized Coefficient	t	Sig.	95% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
(Constant)	-1126,5	252,799		-4,456	0,047	-2214,22	-38,805
TS	39,398	14,5	0,696	2,717	0,113	-22,993	101,788
FQ	149,47	31,07	1,250	4,81	0,041	15,763	283,18
PP	-56,16	15,085	-1,002	-3,723	0,065	-121,064	8,744
SD	13,725	8,24	0,31	1,666	0,238	-21,731	49,18
PR	57,104	17,656	0,944	3,234	0,084	-18,866	133,073
AS	83,814	19,302	1,149	4,342	0,049	0,763	166,865
PRead	-88,68	24,687	-1,416	-3,592	0,07	-194,901	17,541

Source: own computation

The model coefficients were estimated using the least squares method. If we include the estimated values, the regression line will have the following equation:

$$\hat{MV}_i = -1126,5 + 39,398TS_i + 149,47FQ_i - 56,16PP_i + 13,725SD_i + 57,104PR_i + 83,814AS_i - 88,68PRead_i$$

B-values (estimated coefficients) represent one of the most important aspects of regression analysis. Their sign indicate the sense of the relation between the

outcome variable and the associated predictor and their value the change which occurs in the outcome variable if the other predictors are constant or excluded from the analysis. Inverse relationships can be detected between the monetary value and personnel politeness and between monetary value and personnel readiness to solve the customers' problems. A simple interpretation of these findings is that low levels of expectations regarding the personnel politeness and readiness imply high levels of the spent value within the last contact with the company. It should be clear that the company's seven perceived attributes are measured through the customers' expectations regarding them. In this sense, the two identified inverse relationships may contain a clue that the attributes (personnel politeness and readiness) present little importance in the customer's decision making process. The assumption that the obtained results are pure mathematical findings that have no correspondence in reality should not be excluded.

T-values are computed values of the t-student test. They contain two kinds of information regarding the estimated coefficients: (1) their measure or magnitude indicate the contribution the associated predictor has on the model and (2) their statistical significance (Sig.) represents the probability that the estimated coefficient doesn't differ significant from 0. As expected, the fuel quality has a substantial influence on the outcome variable (t-value = 4,81). The statistical significance of the estimated coefficients is pretty good (all values are under the statistical coefficient of $\alpha=0,09$) with two exceptions: time spent (Sig.=0,113) and shop diversity (Sig.=0,238).

A confidence interval was computed for each estimated coefficient. It is used to predict (with a certain probability) the true value of the estimated coefficients on population level. In our case, the probability that the estimated coefficient on population level is included in the computed interval is of 0,95. The theory suggests that the narrower the confidence interval is, the more precise the estimated coefficient on sample level is (Field, 2006). The data in the last two columns of table 6 present large confident intervals, therefore it is not recommended to use the estimated coefficients on sample level for predications at population level. Most of the confidence intervals (5 of 7) cross the point 0, indicating that in some samples the predictor has a negative relationship to the outcome variable whereas in others in has a positive relationship (Field, 2006). This case suggests a lack of consistency in the computed model. The small number of samples observations reveals, once again, the impossibility of inferring the sample results on population level.

Table 7 – Zero-order, Part and Partial Correlations

	Correlations		
	Zero-order	Partial	Part
(Constant)			
TS	0,058	0,887	0,443
FQ	0,186	0,959	0,784
PP	-0,235	-0,935	-0,607

SD	0,311	0,762	0,271
PR	0,306	0,916	0,527
AS	0,073	0,951	0,707
PRead	0,021	-0,930	-0,585

Source: own computation

The table above (table 7) contains the three sets of correlation coefficients computed by SPSS. Zero-order correlations are simple correlation between the outcome variable and each predictor. Their interpretation shows the fact that if we would use simple linear regressions between the outcome variable and each predictor, there would be none or a low correlation between the mentioned pairs of variables.

Partial correlations identify the pure or unique correlation between two variables (e.g. the outcome variable MV and a predictor TS) by excluding the effect (or the variance) of one or more control variables (e.g. FQ, PP, SD, PR, AS, PRead) on both initial variables. High values are obtained for the correlation between each pair of the outcome variable and a predictor, by eliminating the effects of the other predictors. A more in depth analysis of these results certifies that these high partial correlations can be explained by a minimum of three other arguments: (1) Interval scales were used to measure the seven explanatory variables which imply (2) relative small variation within the variables and (3) a sign of **multicolliniarity** can be detected.

Part correlations (semi-partial correlations) detect the association between two variables (e.g. the outcome variable MV and a predictor TS) by eliminating the effect of one or more variables (control variables) on one of the two initial variables (e.g. eliminate the effect of PP on MV, but not on TS). All the values are smaller relative to the partial correlation values because the mass of variation is *more* within the part correlation than the partial correlation. The interpretation of these values is similar to the partial correlation, with the exception that the common variance of the control variables and one of the initial variables is not eliminated.

Under the mentioned consideration, it can be concluded that high expectations regarding fuel quality, auxiliary services and personnel response exert a high influence on the outcome variable – monetary value.

Table 8 – Collinearity Statistics

Model	Collinearity statistics	
	Tolerance	VIF
(Constant)		
TS	0,404	2,474
FQ	0,393	2,544
PP	0,366	2,73
SD	0,765	1,308
PR	0,311	3,211

AS	0,379	2,639
PRead	0,171	5,85

Source: own computation

For every estimated coefficient, the variation inflation factor measures how much of this variation is *inflated* (overestimated) due to the existence of **multicollinearity** between the predictors.

$$VIF_k = \frac{1}{1 - R_k^2}$$

From its formula: $VIF_k = \frac{1}{1 - R_k^2}$, it can be concluded that big values of this inflation factor are determined by the existence of multicollinearity between the predictors. The value of R_k^2 represents the determination coefficient computed for the multiple regression between the k^{th} predictor and the other remaining predictors. Table 8 contains big values of the VIF for all the 7 predictors, indicating the clear existence of **multicollinearity** between the predictors.

The second collinearity indicator is the invers of the VIF. Tolerance below 0.2 indicate serious problem of multicollinearity (Field, 2006).

Table 9 – Collinearity diagnostics

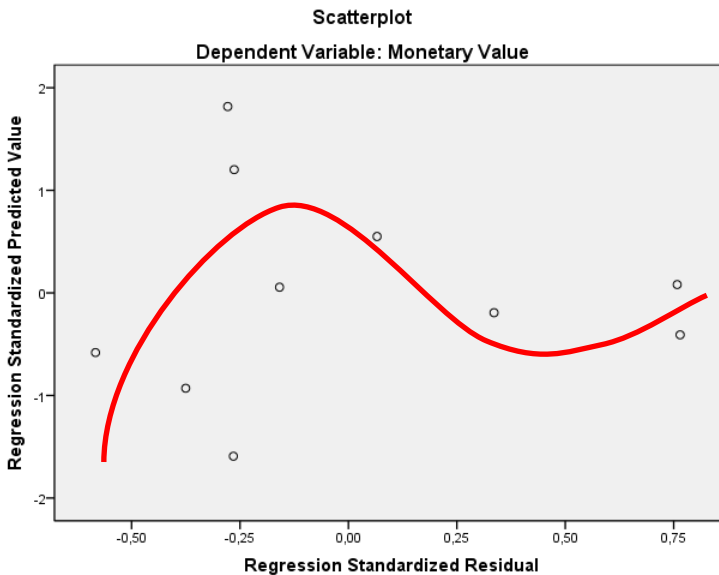
Dimension	Eigenvalue	Condition Index	Variance proportion							
			(Constant)	TS	FQ	PP	SD	PR	AS	PRead
1	7,735	1,000	,00	,00	,00	,00	,00	,00	,00	,00
2	0,135	7,558	,00	,01	,00	,01	,53	,00	,01	,00
3	0,048	12,654	,01	,13	,01	,02	,00	,07	,03	,01
4	0,036	14,726	,00	,14	,00	,16	,00	,08	,08	,00
5	0,021	19,150	,02	,00	,04	,06	,41	,02	,24	,00
6	0,016	22,010	,01	,49	,00	,22	,02	,10	,11	,06
7	0,007	32,974	,01	,00	,02	,31	,03	,48	,00	,50
8	0,001	78,133	,95	,22	,93	,22	,00	,25	,52	,43

Source: own computation

Further information regarding the possible existence of **multicollinearity** between the predictors is contained by the data within the above table. A lack of multicollinearity would mean that for each of the seven dimensions (we exclude the dimension associated with the constant) a high proportion of the variance of only one predictor should account for. It is not our case; multicollinearity of the predictors has once again been proven.

A final part of the regression analysis contains graphical representations based on which some other assumptions of the regression model can be checked (**linearity of the model, homoscedasticity of the residuals and normality of the residuals**).

Figure 2 – Scatterplot of the Regression Standardized Residuals and Regression Standardized Predicted Value

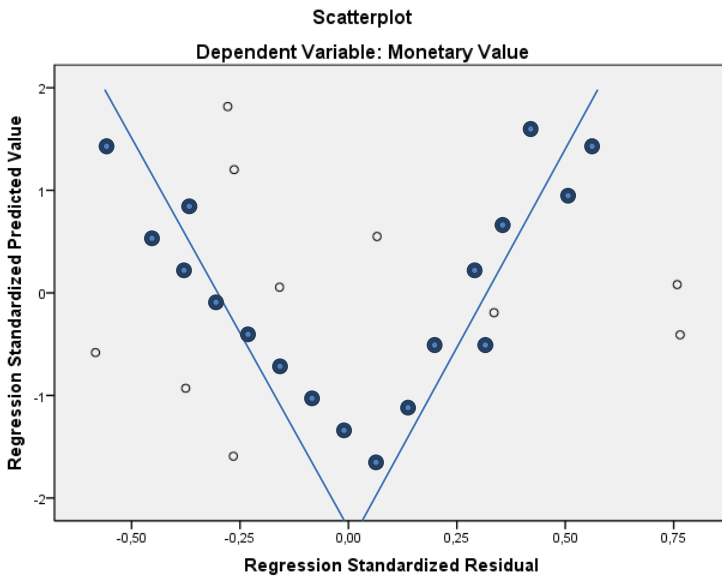


Source: own computation

Residuals are computed as the difference between the predicted value and the observed value for each observation of the database. The figure above overlaps the standardized values of the predicted values with the ones of the residuals. By analyzing it, the two assumption of **model linearity and homoscedasticity of the residuals** can be checked. For the first assumption to be confirmed (**model linearity**), the dots of the scatterplot should converge to a line. Their arrangement follows rather a non-linear path (denoted in the figure 2 by the red curve), signaling for the non-linearity of the model.

The second assumption of the **residuals' homoscedasticity** is confirmed if the dots of the scatterplot are randomly distributed in the bi-dimensional space. Homoscedasticity of the residuals should be understood as a constant variance of the residuals within each level of the predictors. A non-constant variance of the residuals would be graphically represented as a funnel which could display an increase of the residuals' variance related to an increase of the predicted values (Figure 3 – the displayed blue lines and dots are a typical example of heteroscedasticity).

Figure 3 – An example of heteroscedasticity of the residuals



Source: own computation

The distribution of the dots in figure 2 has no particular pattern, the dots are randomly distributed, and thus the assumption of constant variance of the residuals (**homoscedasticity**) is confirmed.

The next set of figures contains information regarding the assumption of the residuals' normality.

Figure 4 – Histogram of the Standardized Residuals

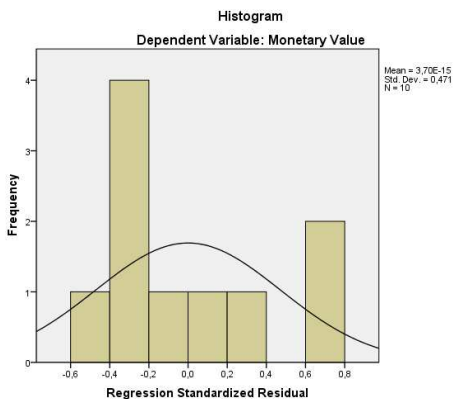


Figure 5 – P-P Plot of the Standardized Residuals

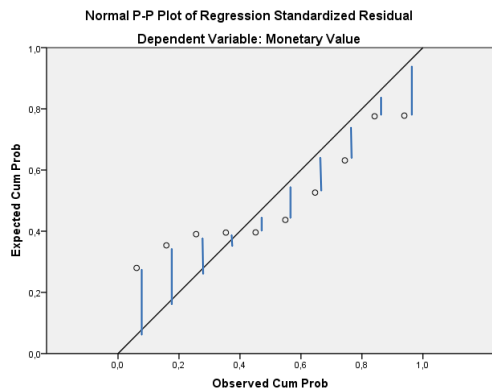


Figure 4 represents the histogram of the obtained residuals and the normality curve. It can be easily observed that the residuals do not follow a normal distribution; peaks that differ from the mean value can be observed for the intervals [-0,4 -0,2] and [0,6 0,8]. The same conclusion is drawn after analyzing figure 5. The vertical segments

between the dots and the normality line indicate the absence of a normal distribution for the observed residuals. If the residuals were normally distributed; the dots would be identical with the normality line or would vary slightly from it. Thus, the normality assumption of the residuals is not confirmed.

4. Conclusions

The purpose of the present paper is to illustrate how to apply and interpret the linear multiple regression analysis within a marketing research based on primary, quantitative data. The model is based on the value-chain concept of relationship marketing theory. In this sense, the monetary value spent by a customer within his last purchase (dependent variable) is influenced by his expectations regarding the dimensions of the company's offer. The fuel-station offer was taken as study-object and the expectation for seven dimensions were measured through a nine-point interval scale.

After applying a set of regression specific procedures (using the statistical software SPSS v.19), two of the five regression assumptions were confirmed. Thus, the developed regression model is not a linear model, signs of multicolliniarity within the predictors are in every part of the analysis and the residuals do not follow a normal distribution. The computed residual values present constant variance at every level of the predictors (homoscedasticity) and do not correlate between each other (nonautocorrelation).

These results demonstrate the lack of consistency of the developed model. Several factors could have biased the desired results. All predictors are basically dimensions that form the company's offer. These dimensions are semantically and practically different; they can satisfy different customer needs. If a customer is asked to evaluate (on a nine-point interval scale) his expectations regarding the mentioned dimensions, a process of *mental smoothing* can appear because of everyone's desire of high expectation in general and regarding a company's offer in particular. The small variation within and between the predictors is the main cause for multicolliniarity. Another possible cause for the obtained results is the small sample size. Although the theoretical rule states that $n \geq k$ (number of observations should be greater than the number of predictors), the sample size of $n=10$ observations generates small variations within the variable, diminishing the prediction power of the computed model relative to the residuals. Thus, measurement and scaling techniques require essential improvement.

Further research will be orientated towards the applicability of other multivariate analysis methods and models (factor analysis, principle components, MANOVA, etc.) within the relationship marketing theory.

5. References

- Bloemer, J., de Ruyter, K., 1998. On the relationship between store image, store satisfaction and store loyalty, *European Journal of Marketing*, Vol.32, No.5/6, 499-513
- Bruhn M., 2010. *Relationship Marketing: Das Management von Kundenbeziehungen*, München, VahlenVerlag
- Burnete S., 2012. *New EU Member Countries are Phasing out Labor-intensive Activities – an Econometric Approach*, Working Paper
- Cătoi I. (Coordinator), 2009. *Tratat de Cercetări de Marketing*. București. Editura Uranus
- Field A., 2006. *Discovering Statistics using SPSS*. Second Edition. London. Sage Publications.
- Greene W.H., 2002. *Econometric Analysis*. Fifth Edition. New Jersey. Prentice Hall.
- Kotler Ph., Keller J., 2009. *Managementul Marketingului*, Ediția a V-a, București, Editura Teora
- Malhotra N., 2010. *Marketing Research. An Applied Approach*. Sixth Edition. New Jersey. Pearson Education.
- Neagoe C., 2012
- Oliver, R.L., Swan, J.E., 1989. Customer Perceptions of Interpersonal Equity and Satisfaction in Transactions: A Field Survey Approach, *Journal of Marketing*, Vol.53, 21-35
- Opreana, A., 2010. The Long-Run Determinants of Investment: A Dynamic Approach for the Future Economic Policies, *Studies in Business and Economics*, Vol.5, Issue 3, 227-237.
- Pop N. Al., Neagoe C., Vlădoi A.-D., 2010. Comunicarea integrată în marketingul intern. Studiu de caz pentru întreprinderile mici și mijlocii. *Proceedings of the 5th International Conference on Business Excellence*, Vol. 2, 89 - 92.

What is Regression Analysis?

Regression analysis is a predictive modelling technique that analyzes the relation between the target or dependent variable and independent variable in a dataset. The different types of regression analysis techniques get used when the target and independent variables show a linear or non-linear relationship between each other, and the target variable contains continuous values. The regression technique gets used mainly to determine the predictor strength, forecast trend, time series, and in case of cause & effect relation.

Regression analysis is the primary technique to solve the regression problems in machine learning using data modelling. It involves determining the best fit line, which is a line that passes through all the data points in such a way that distance of the line from each data point is minimized.

Types of Regression Analysis Techniques

There are many types of regression analysis techniques, and the use of each method depends upon the number of factors. These factors include the type of target variable, shape of the regression line, and the number of independent variables.

Below are the different regression techniques:

1. Linear Regression
2. Logistic Regression
3. Ridge Regression
4. Lasso Regression
5. Polynomial Regression
6. Bayesian Linear Regression

The different types of regression in machine learning techniques are explained below in detail:

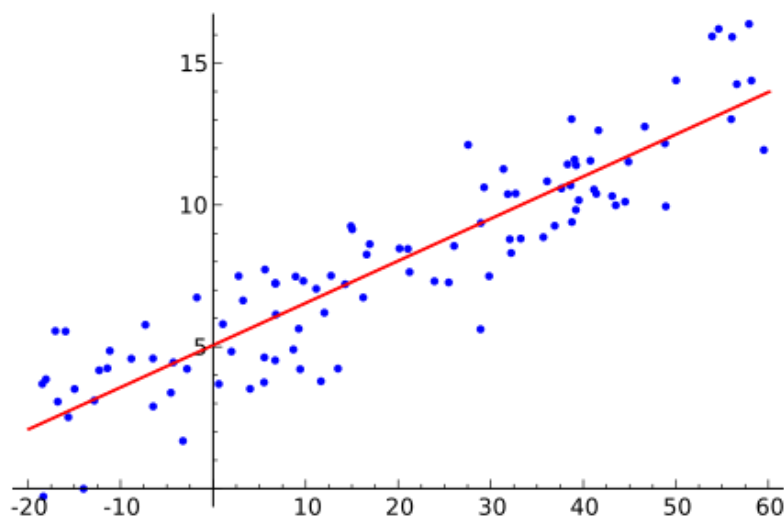
1. Linear Regression

Linear regression is one of the most basic types of regression in machine learning. The linear regression model consists of a predictor variable and a dependent variable related linearly to each other. In case the data involves more than one independent variable, then linear regression is called multiple linear regression models.

The below-given equation is used to denote the linear regression model:

$$y=mx+c+e$$

where m is the slope of the line, c is an intercept, and e represents the error in the model.



The best fit line is determined by varying the values of m and c . The predictor error is the difference between the observed values and the predicted value. The values of m and c get selected in such a way that it gives the minimum predictor error. It is important to note that a simple linear regression model is susceptible to outliers. Therefore, it should not be used in case of big size data.

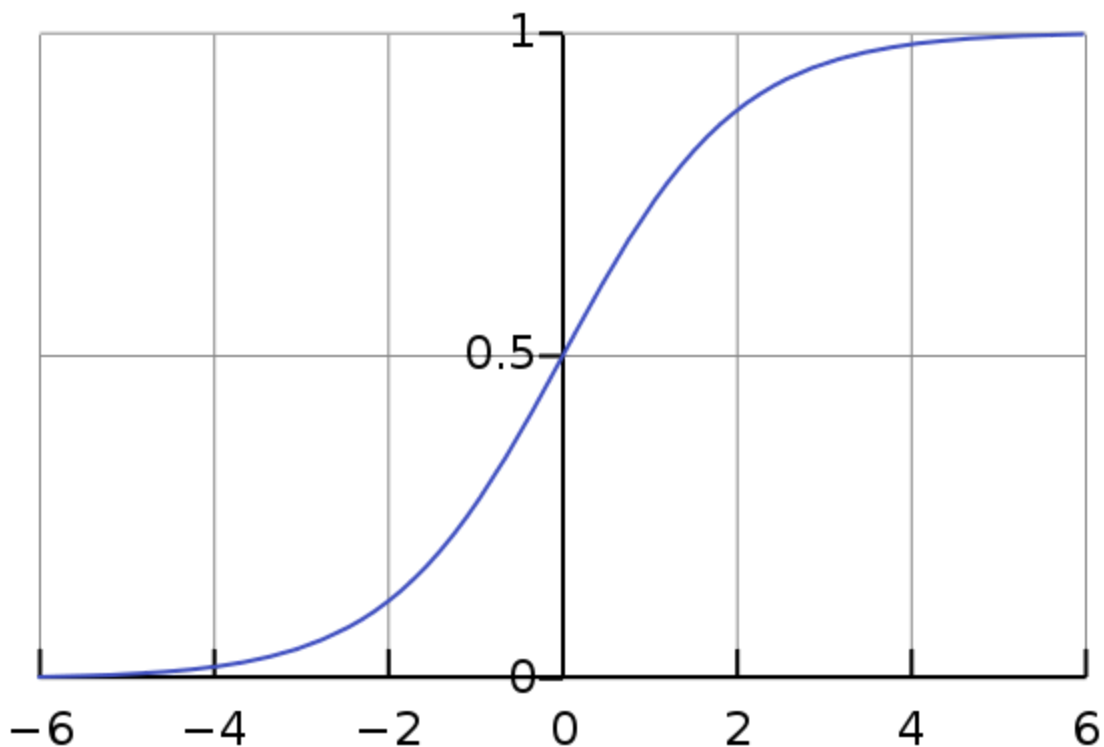
2. Logistic Regression

Logistic regression is one of the types of regression analysis technique, which gets used when the dependent variable is discrete. Example: 0 or 1, true or false, etc. This means the target variable can have only two values, and a sigmoid curve denotes the relation between the target variable and the independent variable.

Logit function is used in Logistic Regression to measure the relationship between the target variable and independent variables. Below is the equation that denotes the logistic regression.

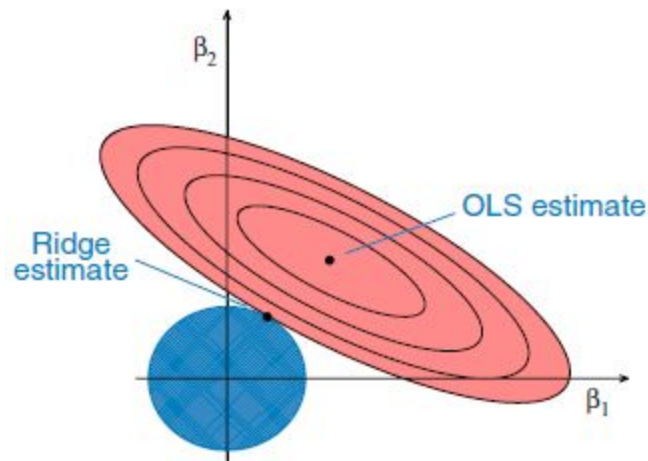
$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \dots + b_kX_k$$

where p is the probability of occurrence of the feature.



For selecting logistic regression, as the regression analyst technique, it should be noted, the size of data is large with the almost equal occurrence of values to come in target variables. Also, there should be no multicollinearity, which means that there should be no correlation between independent variables in the dataset.

3. Ridge Regression



This is another one of the types of regression in machine learning which is usually used when there is a high correlation between the independent variables. This is because, in the case of multi collinear data, the least square estimates give unbiased values. But, in case the collinearity is very high, there can be some bias value. Therefore, a bias matrix is introduced in the equation of Ridge Regression. This is a powerful regression method where the model is less susceptible to overfitting.

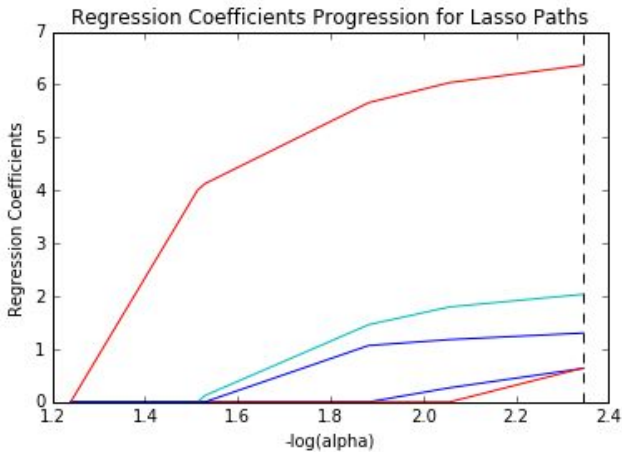
Below is the equation used to denote the Ridge Regression, where the introduction of λ (lambda) solves the problem of multicollinearity:

$$\beta = (X^T X + \lambda * I)^{-1} X^T y$$

4. Lasso Regression

Lasso Regression is one of the types of regression in machine learning that performs regularization along with feature selection. It prohibits the absolute size of the regression coefficient. As a result, the coefficient value gets nearer to zero, which does not happen in the case of Ridge Regression.

Due to this, feature selection gets used in Lasso Regression, which allows selecting a set of features from the dataset to build the model. In the case of Lasso Regression, only the required features are used, and the other ones are made zero. This helps in avoiding the overfitting in the model. In case the independent variables are highly collinear, then Lasso regression picks only one variable and makes other variables to shrink to zero.



Below is the equation that represents the Lasso Regression method:

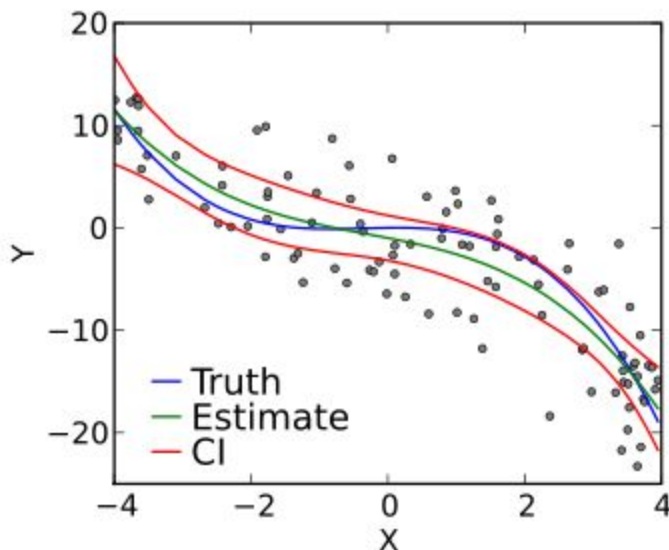
$$\sum_{i=1}^N f(x_i, y_i, \alpha, \beta)$$

5. Polynomial Regression

Polynomial Regression is another one of the types of regression analysis techniques in machine learning, which is the same as Multiple Linear Regression with a little modification. In Polynomial Regression, the relationship between independent and dependent variables, that is X and Y, is denoted by the n-th degree.

It is a linear model as an estimator. Least Mean Squared Method is used in Polynomial Regression also.

The best fit line in Polynomial Regression that passes through all the data points is not a straight line, but a curved line, which depends upon the power of X or value of n.



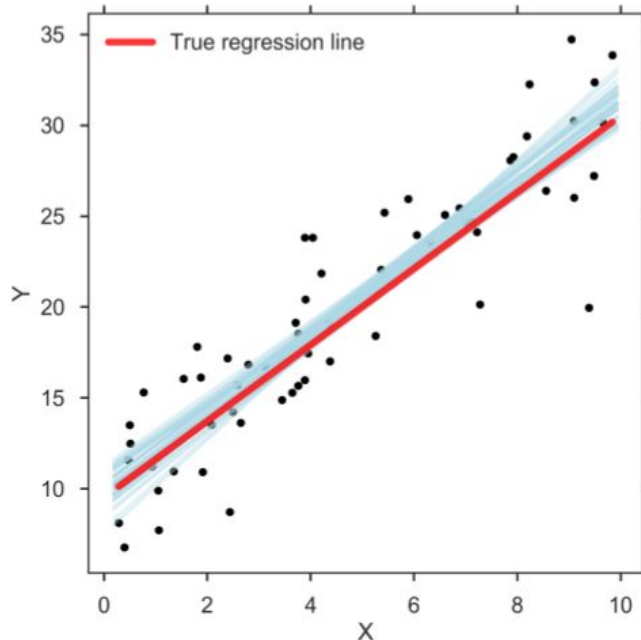
While trying to reduce the Mean Squared Error to a minimum and to get the best fit line, the model can be prone to overfitting. It is recommended to analyze the curve towards the end as the higher Polynomials can give strange results on extrapolation.

Below equation represents the Polynomial Regression:

$$l = \beta_0 + \beta_1 x + \epsilon$$

6. Bayesian Linear Regression

Bayesian Regression is one of the types of regression in machine learning that uses the Bayes theorem to find out the value of regression coefficients. In this method of regression, the posterior distribution of the features is determined instead of finding the least-squares. Bayesian Linear Regression is like both Linear Regression and Ridge Regression but is more stable than the simple Linear Regression.



INTRODUCTION TO NONPARAMETRIC STATISTICS

The majority of hypothesis tests discussed so far have made inferences about population parameters, such as the mean and the proportion. These parametric tests have used the parametric statistics of samples that came from the population being tested. To formulate these tests, we made restrictive assumptions about the populations from which we drew our samples.

For example, we assumed that our samples either were large or came from normally distributed populations. But populations are not always normal. And even if a goodness-of-fit test indicates that a population is approximately normal, we cannot always be sure we're right because the test is not 100 percent reliable.

Clearly, there are certain situations in which the use of the normal curve is not appropriate. For these cases, we need alternatives to the parametric statistics and the specific hypothesis tests we've been using so far. Fortunately, in recent times statisticians have developed useful techniques that do not make restrictive assumptions about the shape of population distributions. These are known as distribution-free or, more commonly, nonparametric tests. The hypotheses of a nonparametric test are concerned with something other than the value of a population parameter.

1. The sign test for paired data, where positive or negative signs are substituted for quantitative values.

2. A rank sum test, often called the Mann-Whitney U test, which can be used to determine whether two independent samples have been drawn from the same population. It uses more information than the sign test.

3. Another rank sum test, the Kruskal-Wallis test, which generalizes the analysis of variance to enable us to dispense with the assumption that the populations are normally distributed.

4. The one-sample runs test, a method for determining the randomness with which sampled items have been selected.

5. Rank correlation, a method for doing correlation analysis when the data are not available to use in numerical form, but when information is sufficient to rank the data first, second, third, and so forth.

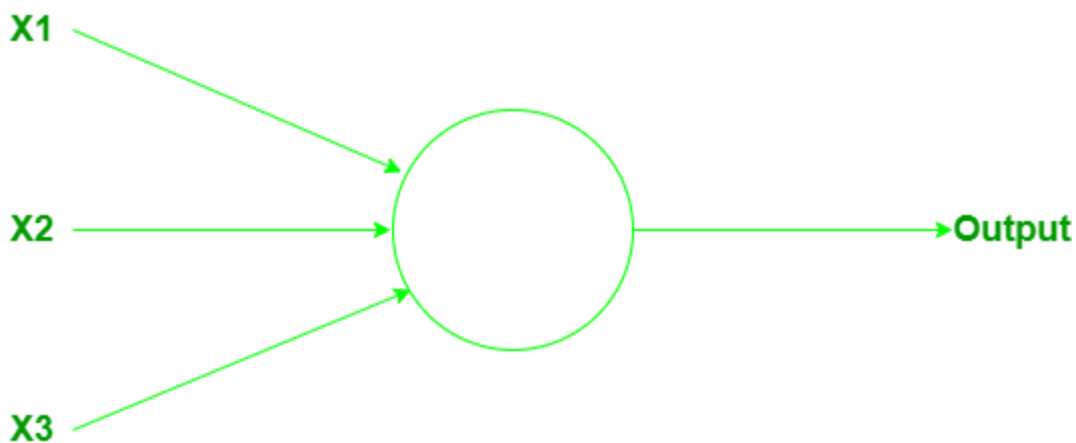
6. The Kolmogorov–Smirnov test, another method for determining the goodness of fit between an observed sample and a theoretical probability distribution.

Neural Network

The term Neural Networks refers to the system of neurons either organic or artificial in nature. In artificial intelligence reference, neural networks are a set of algorithms that are designed to recognize a pattern like a human brain. They interpret sensory data through a kind of machine perception, labeling, or clustering raw input. The recognition is numerical, which is stored in vectors, into which all real-world data, be it images, sound, text, or time series, must be translated. A neural network can be pictured as a system that consists of a number of highly interconnected nodes, called 'neurons', which are organized in layers that process information using dynamic state responses to external inputs. Before understanding the working and architecture of neural networks, let us try to understand what artificial neurons actually are.

Artificial Neurons

Perceptron: Perceptrons are a type of artificial neurons developed in the 1950s and 1960s by the scientist Frank Rosenbalt, inspired by earlier work by Warren McCulloch and Walter Pitts. So, how do perceptron works? A perceptron takes several binary outputs x_1, x_2, \dots , and produces a single binary output.



It could have more or fewer inputs. To calculate/compute the output weights play an important role. Weights w_1, w_2, \dots , are real numbers expressing the importance of the respective inputs to the outputs. The neuron's output(0 or 1) totally depends upon a threshold value and is computed according to the function:

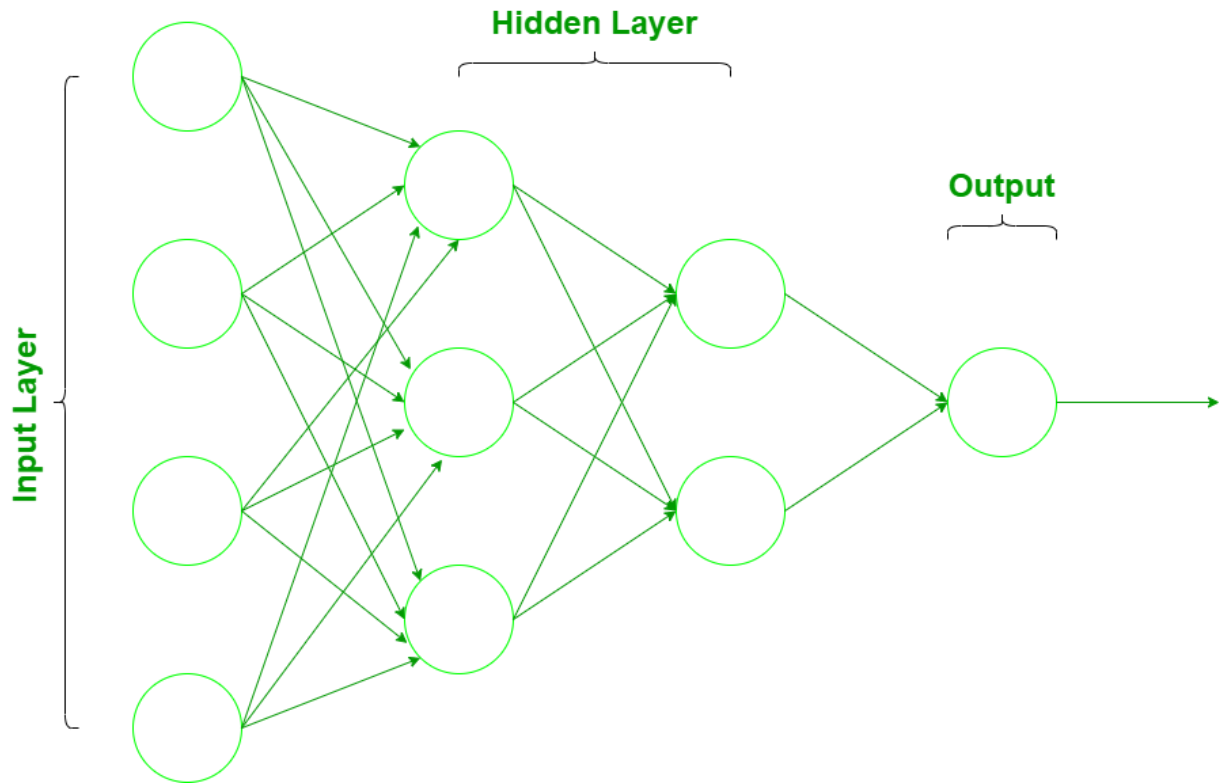
$$output = \begin{cases} 0 & \text{if } \sum_j w_j \cdot x_j \leq t_0 \\ 1 & \text{if } \sum_j w_j \cdot x_j > t_0 \end{cases}$$

Here t_0 is the threshold value. It is a real number which is a parameter of the neuron. That's the basic mathematical model. The perceptron is that it's a device that makes decisions by weighing up the evidence. By varying the weights and the threshold, we can get different models of decision-making.

The Architecture of Neural Networks

A neural network consists of three layers:

1. Input Layer: Layers that take inputs based on existing data.
2. Hidden Layer: Layers that use backpropagation to optimise the weights of the input variables in order to improve the predictive power of the model.
3. Output Layer: Output of predictions based on the data from the input and hidden layers.



The input data is introduced to the neural network through the input layer that has one neuron for each component present in the input data and is communicated to hidden layers (one or more) present in the network. It is called 'hidden' only because they do not constitute the input or output layer. In the hidden layers, all the processing actually happens through a system of connections characterized by weights and biases (as discussed earlier). Once the input is received, the neuron calculates a weighted sum adding also the bias and according to the result and an activation function (the most common one is sigmoid), it decides whether it should be 'fired' or 'activated'. Then, the neuron transmits the information downstream to other connected neurons in a process called 'forward pass'. At the end of this process, the last hidden layer is linked to the output layer which has one neuron for each possible desired output.

Steps to build a NN in R

1. Scaling of the data

To set up a neural network to a dataset it is very important that we ensure a proper scaling of data. The scaling of data is essential because otherwise, a variable may have a large impact on the prediction variable only because of its scale. Using unscaled data may lead to meaningless results. The common techniques to scale data are min-max normalization, Z-score normalization, median and MAD, and tan-h estimators. The min-max normalization transforms the data into a common range, thus removing the scaling effect from all the variables.

Step 2: Sampling of the data

Now divide the data into a training set and test set. The training set is used to find the relationship between dependent and independent variables while the test set analyses the performance of the model. We use 60% of the dataset as a training set. The assignment of the data to training and test set is done using random sampling. We perform random sampling on R using `sample()` function. Use `set.seed()` to generate some random sample every time and maintain consistency. Use the index variable while fitting neural network to create training and test data sets. The R script is as follows:

3. Fitting a Neural Network

Now fit a neural network on our data. We use neuralnet library for the same. `neuralnet()` function helps us to establish a neural network for our data. The `neuralnet()` function we are using here has the following syntax.


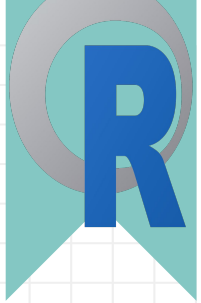
Syntax:

neuralnet(formula, data, hidden = 1, stepmax = 1e+05, rep = 1, lifesign = "none", algorithm = "rprop+", err.fct = "sse", linear.output = TRUE)

Parameters:

Argument	Description
formula	a symbolic description of the model to be fitted.
data	a data frame containing the variables specified in formula.
hidden	a vector of integers specifying the number of hidden neurons (vertices) in each layer
err.fct	a differentiable function that is used for the calculation of the error. Alternatively, the strings 'sse' and 'ce' which stand for the sum of squared errors and the cross-entropy can be used.
linear.output	logical. If act.fct should not be applied to the output neurons set linear output to TRUE, otherwise to FALSE.

lifesign	a string specifying how much the function will print during the calculation of the neural network. 'none', 'minimal' or 'full'.
rep	the number of repetitions for the neural network's training.
algorithm	a string containing the algorithm type to calculate the neural network. The following types are possible: 'backprop', 'rprop+', 'rprop-', 'sag', or 'slr'. 'backprop' refers to backpropagation, 'rprop+' and 'rprop-' refer to the resilient backpropagation with and without weight backtracking, while 'sag' and 'slr' induce the usage of the modified globally convergent algorithm (grprop).
stepmax	the maximum steps for the training of the neural network. Reaching this maximum leads to a stop of the neural network's training process.



PREDICTION WITH R SOFTWARE







CONTENTS OF THE CLASS



1. INTRODUCTION TO R
2. APPLICATIONS OF R
3. EXPLAINING PROGRAMMING BASICS
4. R STUDIO ENVIRONMENT WALKTHROUGH
5. BASIC CODES



THE UNIVERSITY OF
AUCKLAND
Auckland University of Education & Research
NEW ZEALAND



Prof. Ross Ihaka ..
**HE CREATED A LANGUAGE
USED BY MILLIONS**

A language called R. And as those millions use it to predict shifts in the stock market, track the behaviour of marine life, or search for a cure for cancer, the effects of this statistical programming language – developed by Associate Professor Ross Ihaka and his colleagues at the University of Auckland – reach far beyond millions.

That said, what Ross is focused on now is the future. As datasets grow larger, so too does the need for R to become more powerful.

So whether it's being used by an internet giant in California, or a little-known institute near Cape Town, this adaptive, open-source platform known by just a single letter, will allow the new challenges we face to be understood more deeply, and in turn bring about amazing new breakthroughs.

For more about Ross and R, or how others are achieving the amazing at the University of Auckland, visit achievetheamazing.ac.nz/research

ACHIEVE THE
AMAZING

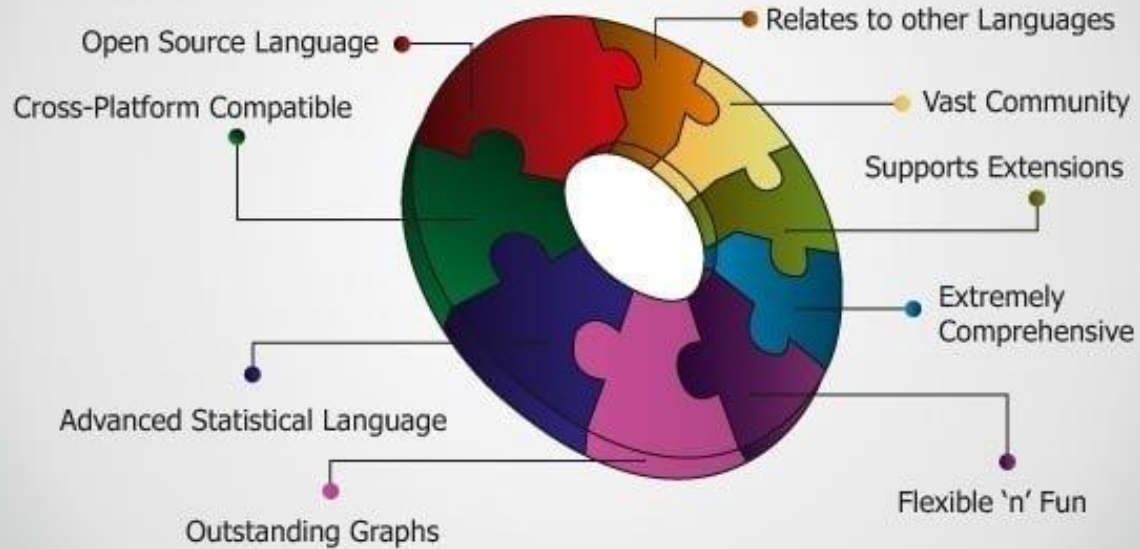
Ross Ihaka

Robert Gentleman

1993



Why Learn R?



Major Applications of R Software



Applications of



Industry wise Applications

Finance

- Risk Management(candlestick charts, density plots)
- Time Series Analysis
- Credit Risk Management and Portfolio Management (ANZ)
- QUANTMOD

Banking

- Mortgage Hair Model
- Customer Segmentation
- Financial Reporting (BOA)

Healthcare

- BIOCONDUCTOR
- Genomic Analysis
- Bioinformatics (Institute of Bioinformatics -Kshitish Acharya)
- Drug Discovery
- Epidemiology
- Clinical Trials

Social Media

- SOCIALMEDIAMINE R
- Market Analysis
- Sentiment Analysis
- Datamining

Real World Examples

- Facebook – Facebook uses R to update status and its social network graph. It is also used for predicting colleague interactions with R.
- Google – Google uses R to calculate ROI on advertising campaigns and to predict economic activity and also to improve the efficiency of online advertising.
- Mozilla – It is the foundation behind the Firefox web browser and uses R to visualize web activity.
- New York Times – R is used in the news cycle at The New York Times to crunch data and prepare graphics before they go for printing.
- Trulia – Trulia, the real-estate analysis website uses R for predicting house prices and local crime rates.



Companies that use R for Analytics



Lab 1 : Demonstration of Reading data from files and working with
datasets

Introduction To R programming



PROGRAMMERS JARGON



Variable

Containers for storing
data values

Array

Cluster of similar items

Vectors

One dimensional arrays,
that hold data

Matrices

two-Dimensional Array

Factors

Statistical Data type to
store categorical data

Lists

N-dimensional arrays,
holds all kinds of data



PROGRAMMERS JARGON



DATA FRAMES

A data frame has the variables of a data set as columns and the observations as rows.

	mpg	cyl	dis
Mazda RX4	21.0	6	
Mazda RX4 Wag	21.0	6	
Datsun 710	22.8	4	
Hornet 4 Drive	21.4	6	
Plymouth Sportabout	18.7	8	
Valiant	18.1	6	

Agenda for the session (6/1/2020)

1. Understand how to build numeric and character vectors
2. Conduct a small analysis

Basic Analytics with Vectors

Functions :

1. `c()` - concatenation
2. `names()`- change the name of vectors
3. `Mean ()` - find the mean values
4. `sum()`- find the summation

Topics :

1. How to build Vectors
2. Basic Vector Manipulation
3. How to interpret results

Case Study I

Days	Poker	Roulette
1	Made 140	Lost 24
2	Lost 50	Lost 50
3	Made 20	Made 100
4	Lost 120	Lost 350
5	Made 240	Made 10

You have gone to Las Vegas to try your luck and these are your readings.

```
poker_vector <- c(140,-50,20,-120,240)  
poker_vector = c(140,-50,20,-120,240)
```

In the previous exercise, we created a vector with your earnings over the week. Each vector element refers to a day of the week but it is hard to tell which element belongs to which day. It would be nice if you could show that in the vector itself.

You can give a name to the elements of a vector with the `names()` function. Have a look at this example:

```
some_vector <- c("John Doe", "poker player")
names(some_vector) <- c("Name", "Profession")
```

This code first creates a vector `some_vector` and then gives the two elements a name. The first element is assigned the name `Name`, while the second element is labeled `Profession`. Printing the contents to the console yields following output:

```
  Name  Profession
"John Doe" "poker player"
```

Agenda 11/2020

1. Indexing Vectors
2. Factors and Order

Indexing : The process of selecting values from a vector or data frame.

12/1/2020

1. Build a Business Decision Matrix and generate insights

What's a matrix?

In R, a matrix is a collection of elements of the same data type (numeric, character, or logical) arranged into a fixed number of rows and columns. Since you are only working with rows and columns, a matrix is called two-dimensional.

You can construct a matrix in R with the `matrix()` function. Consider the following example:

```
matrix(1:9, byrow = TRUE, nrow = 3)
```

In the `matrix()` function:

- The first argument is the collection of elements that R will arrange into the rows and columns of the matrix. Here, we use `1:9` which is a shortcut for `c(1, 2, 3, 4, 5, 6, 7, 8, 9)`.
- The argument `byrow` indicates that the matrix is filled by the rows. If we want the matrix to be filled by the columns, we just place `byrow = FALSE`.
- The third argument `nrow` indicates that the matrix should have three rows.

What is a Business Decision Matrix ?

Decision Matrix Analysis is a useful technique to use for making a decision. It's particularly powerful where you have a number of good alternatives to choose from, and many different factors to take into account. This makes it a great technique to use in almost any important decision where there isn't a clear and obvious preferred option.

Being able to use Decision Matrix Analysis means that you can take decisions confidently and rationally, at a time when other people might be struggling to make a decision.

Steps to Build a Decision Matrix

1. List out the Row values or the categories you have to decide from
2. Populate the row values with data that you have gathered
3. Add Totals (Row or Column Values)
4. Add weights according criteria of your choice
5. Combine all the values to create a matrix

Case Study 1 :

Consider that you are the manager of a Car Showroom that exclusively deals with three cars and you want to analyze which car brings you the highest profit. For this you have decided to conduct a survey and you have gathered the following intel

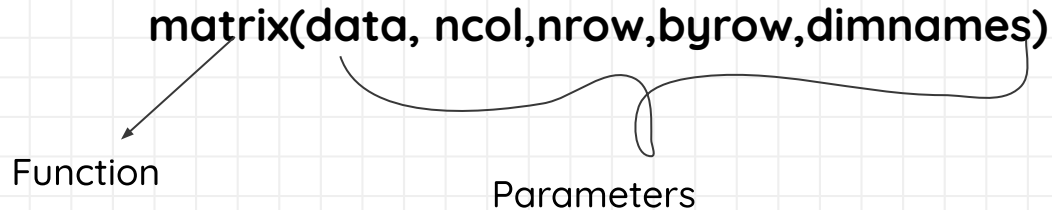
Criteria :

'Cost','Practicality','Performance','Reliability','Fuel Economy'

You also have data on the scores the customers have given for the cars out of 5.

Car A	Car B	Car C
5	3	3
2	4	3
4	2	5
1	2	4
2	3	3

Matrix function in R



matrix(data , ncol (specify the number of columns), nrow (specify the number of rows), byrow (specify how R should fill the row and columns), dimnames (add names to rows and columns of the matrix))

```
m3 <- matrix(w_all_car, ncol = 3, nrow=5, byrow = FALSE, dimnames = list(criteria, option))  
print(m3)
```

13.11.2020

MATRIX COMMANDS

colSums() :

rbind():

cbind():

rowSums():

Indexing A Matrix

M1[row value, column value]

Plot function

plot(x,y,main=,sub=,xlab=,ylab=)

CASE STUDY QUESTIONS :

1. Which brand is the most cost effective ?
2. Which brand has the highest practicality ?
3. If the fuel prices were to increase by 10% would that affect the brands and in turn the sales? (show numerically)
4. Are the sales of Benz and Mini Cooper correlated ? (show graphically)
5. As the manger which car would you keep an increased inventory for ? (detailed answer)

18.11.2020

1. Introduction to Dataframe
2. Introduction to Lists
3. Reading Data from computer
4. Introduction R packages

Case Study : Building a Dataframe

Reading a dataframe

```
Data <- read.csv ("location of the file with / and file name with extention")
```

Creating a dataframe

```
df <- data.frame( names of variables separated by commas)
```

Indexing a Dataframe

01	Through Column Names	<code>Required_col <- df['name of the column']</code>
02	Through Matrix indexing	<code>Required_col <- df[row , column]</code>
03	Through \$ symbol	<code>Required_col <- df\$column_name</code>

Conditional Indexing : `df[df$column_name == "condition",]`

Case Study Queries

You are the manager of an Ice Cream Parlor and you have decided to build a dataframe to analyze sales and quantities sold.

1. Which Product generates the most revenue ?
2. What is the customers preferred payment method ?
3. Which product is sold the most ?

Lists

R commands :

`list()` - to convert any array into a list

Indexing a list

`List[[list row value]][matrix/df row and col values]`

Quiz Time

What is the difference between a list and a dataframe ?

What is the difference between a matrix and a Dataframe ?

Give three examples of categorical variables and continuous variables ?

What is a function and Parameter in R ?

19.11.2020

Lab 2

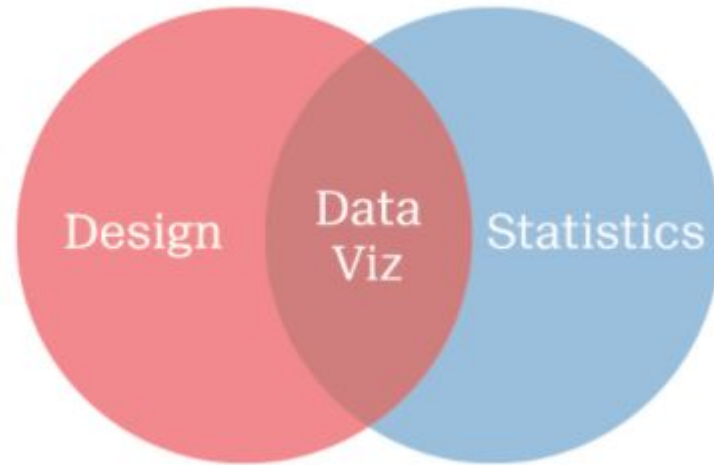
Demonstration of Graphs: Basic high-level plots,
Modifications of scatter plots, Modifications of
histograms, parallel box plots

Introduction to data Visualization with R

What is Data Visualization ?

Data visualization & data science

- A core skill in Data Science.



Exploratory versus explanatory



Exploratory

Easily made
Data Heavy
Specialized
Eg , Data analysts

Explanatory

Labour intensive
Not specialized
Larger audience
Eg, Graphs in Publications

What are packages in R ?

R packages are collections of functions and data sets developed by the community. They increase the power of R by improving existing base R functionalities, or by adding new ones.

What are Repositories ?

A repository is a place where packages are located so you can install them from it. Three of the most popular repositories for R packages are:

1. CRAN
2. Bioconductor
3. Github

How to install packages ?

R command : `install.packages("package name")`

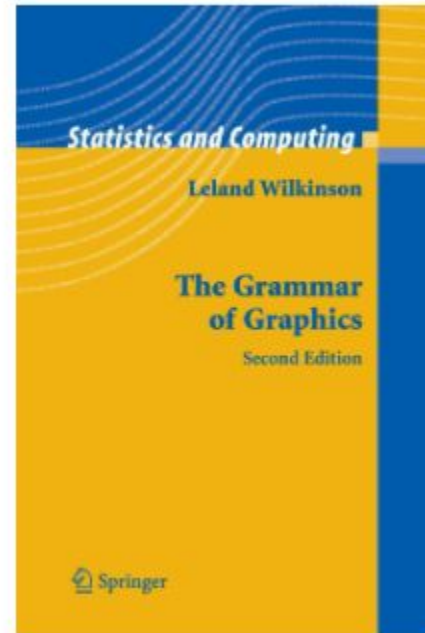
ggplot : Grammar of Graphics

The quick brown fox jumps over the lazy dog

Article	<i>The</i>	<i>A</i>	<i>The</i>
Adjective	<i>quick brown</i>	<i>rabid red</i>	
Noun	<i>fox</i>	<i>fox</i>	<i>Hunter</i>
Verb	<i>jumps</i>	<i>bit</i>	<i>shot</i>
Preposition	<i>over</i>		
Article	<i>the</i>	<i>the</i>	<i>the</i>
Adjective	<i>lazy</i>	<i>friendly</i>	<i>rabid red</i>
Noun	<i>dog.</i>	<i>dog.</i>	<i>fox.</i>

Grammar of graphics

- Plotting framework
- Leland Wilkinson, Grammar of Graphics, 1999
- 2 principles
 - Graphics = distinct layers of grammatical elements
 - Meaningful plots through aesthetic mappings



20.11.2020

The three essential grammatical elements

Element	Description
Data	The data-set being plotted.
Aesthetics	The scales onto which we <i>map</i> our data.
Geometries	The visual elements used for our data.

Jargon for each element

	Data	<i>{variables of interest}</i>				
Aesthetics	<i>x-axis</i> <i>y-axis</i>	<i>colour</i> <i>fill</i>	<i>size</i> <i>labels</i>	<i>alpha</i> <i>shape</i>	<i>line width</i> <i>line type</i>	
Geometries	<i>point</i>	<i>line</i>	<i>histogram</i>	<i>bar</i>	<i>boxplot</i>	

Ggplot2 layers



Typical aesthetic mappings

Aesthetic	Description
x	X axis position
y	Y axis position
fill	Fill color
color	Color of points, outlines of other geoms
size	Area or radius of points, thickness of lines

Aesthetic	Description
alpha	Transparency
linetype	line dash pattern
labels	Text on a plot or axes
shape	Shape

Using GGplot package to explore the data

Case Study :

Aim : To find out if wt field correlates to mpg

Source Data :

We used the mtcars data set that is built-in to the R distribution. mtcars data comes from the 1974 Motor Trend magazine. The data includes fuel consumption data, and ten aspects of car design for then-current car models.

Method :

Correlation Analysis using Scatterplots



[1]	mpg	Miles/(US) gallon
[2]	cyl	Number of cylinders
[3]	disp	Displacement (cu.in.)
[4]	hp	Gross horsepower
[5]	drat	Rear axle ratio
[6]	wt	Weight (1000 lbs)
[7]	qsec	1/4 mile time
[8]	vs	Engine (0 = V-shaped, 1 = straight)
[9]	am	Transmission (0 = automatic, 1 = manual)
[10]	gear	Number of forward gears

23.11.2020

Geom Layers

48 geometries

geom_*						
abline	contour	dotplot	jitter	pointrange	ribbon	spoke
area	count	errorbar	label	polygon	rug	step
bar	crossbar	errorbarh	line	qq	segment	text
bin2d	curve	freqpoly	linerange	qq_line	sf	tile
blank	density	hex	map	quantile	sf_label	violin
boxplot	density2d	histogram	path	raster	sf_text	vline
col	density_2d	hline	point	rect	smooth	

Attributes vs Aesthetics

Attributes of a graph are the changes that can be made in the geometric layers.

Essential	Optional
x,y	alpha, color, fill, shape, size, stroke

Shape attribute values

□ 0 ○ 1 △ 2 + 3 × 4

◇ 5 ▽ 6 ⊠ 7 * 8 ⬠ 9

⊕ 10 ⊗ 11 ⊞ 12 ⊗ 13 ⊞ 14

■ 15 ● 16 ▲ 17 ◆ 18 ● 19

● 20 ● 21 ■ 22 ◆ 23 ▲ 24 ▽ 25

Shape attribute values

□ 0 ○ 1 △ 2 + 3 × 4

◇ 5 ▽ 6 ⊠ 7 ✱ 8 ⬠ 9

⊕ 10 ⚡ 11 ⊞ 12 ⊗ 13 ⊞ 14

■ 15 ● 16 ▲ 17 ◆ 18 ● 19

● 20 ● 21 ■ 22 ◆ 23 ▲ 24 ▽ 25

Summary

What are scatter plots used for ?

To measure correlation between the data points

Name the advanced Scatter plots that we used ?

Scatter plots with two datasets in one

Scatter plot with the line of best fit or linear reg line

Scatter plot with Ellipse

3d Plots

What is a linear relationship ?

Increase in one , leads to increase in an other.

Eg Hours spent studying R and acing your exams

Plotting Hierarchy

High Level Plots

Eg , 3d Scatter plots, scatterplot matrix



Plots with 4 values

Eg, Scatter plots , Line Graphs , density plots



Low Level Plots

Eg, Bar plots, Histograms

BARPLOTS

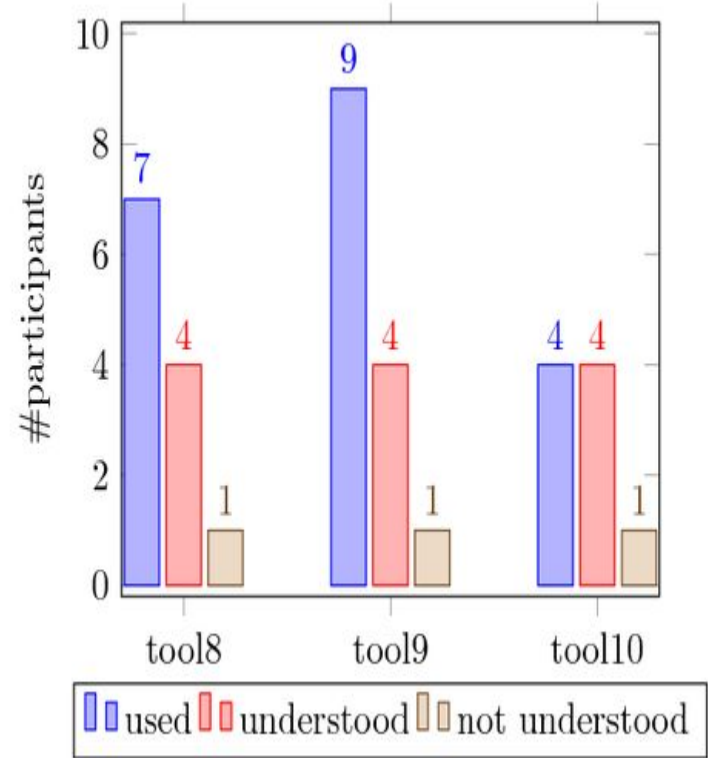
A bar graph (also known as a bar chart or bar diagram) is a visual tool that uses bars to compare data among categories.

A bar graph may run horizontally or vertically. The important thing to know is that the longer the bar, the greater its value.

Bar graphs consist of two axes. On a vertical bar graph, as shown above, the horizontal axis (or x-axis) shows the data categories. In this example, they are years. The vertical axis (or y-axis) is the scale. The colored bars are the data series.

Bar graphs have three key attributes:

- A bar diagram makes it easy to compare sets of data between different groups at a glance.
- The graph represents categories on one axis and a discrete value in the other. The goal is to show the relationship between the two axes.
- Bar charts can also show big changes in data over time.



Barplots in Rstudio

Bar Plots, with a categorical X-axis

- Use `geom_bar()` or `geom_col()`

Geom	Stat	Action
<code>geom_bar()</code>	"count"	Counts the number of cases at each x position
<code>geom_col()</code>	"identity"	Plot actual values

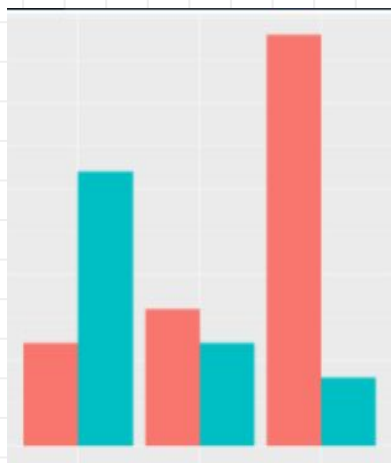
- All positions from before are available
- Two types
 - Absolute counts
 - Distributions

Bar plots in Rstudio

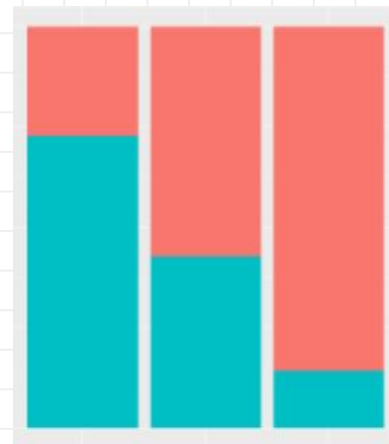
Stacked



Dodge



Fill



Ggplot - Barplot

Case Study :


Aim : to conduct univariate analysis using the “cyl” column

Source Data :

We used the mtcars data set that is built-in to the R distribution. mtcars data comes from the 1974 Motor Trend magazine. The data includes fuel consumption data, and ten aspects of car design for then-current car models.

Method :

Analysis using barplots



[1]	mpg	Miles/(US) gallon
[2]	cyl	Number of cylinders
[3]	disp	Displacement (cu.in.)
[4]	hp	Gross horsepower
[5]	drat	Rear axle ratio
[6]	wt	Weight (1000 lbs)
[7]	qsec	1/4 mile time
[8]	vs	Engine (0 = V-shaped, 1 = straight)
[9]	am	Transmission (0 = automatic, 1 = manual)
[10]	gear	Number of forward gears

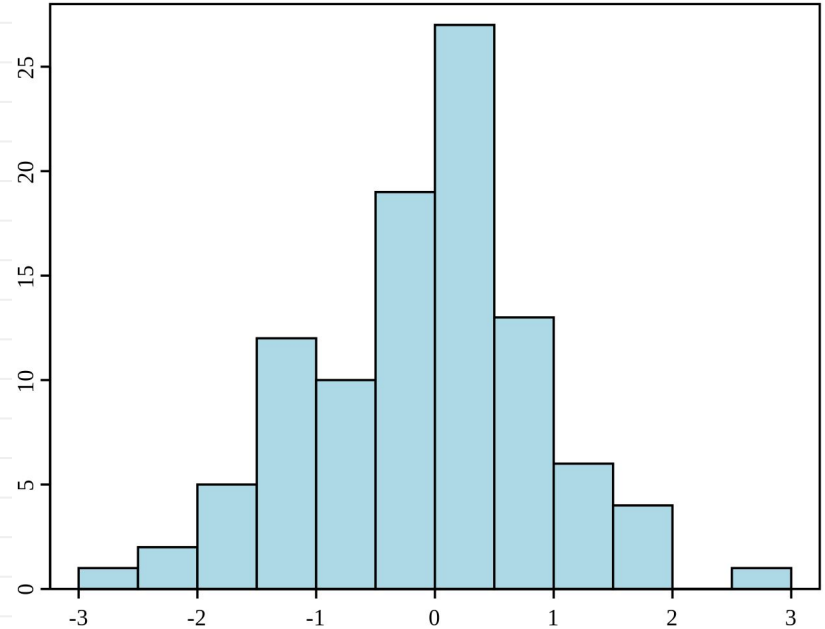
HISTOGRAMS

A histogram is a plot that lets you discover, and show, the underlying frequency distribution (shape) of a set of continuous data.

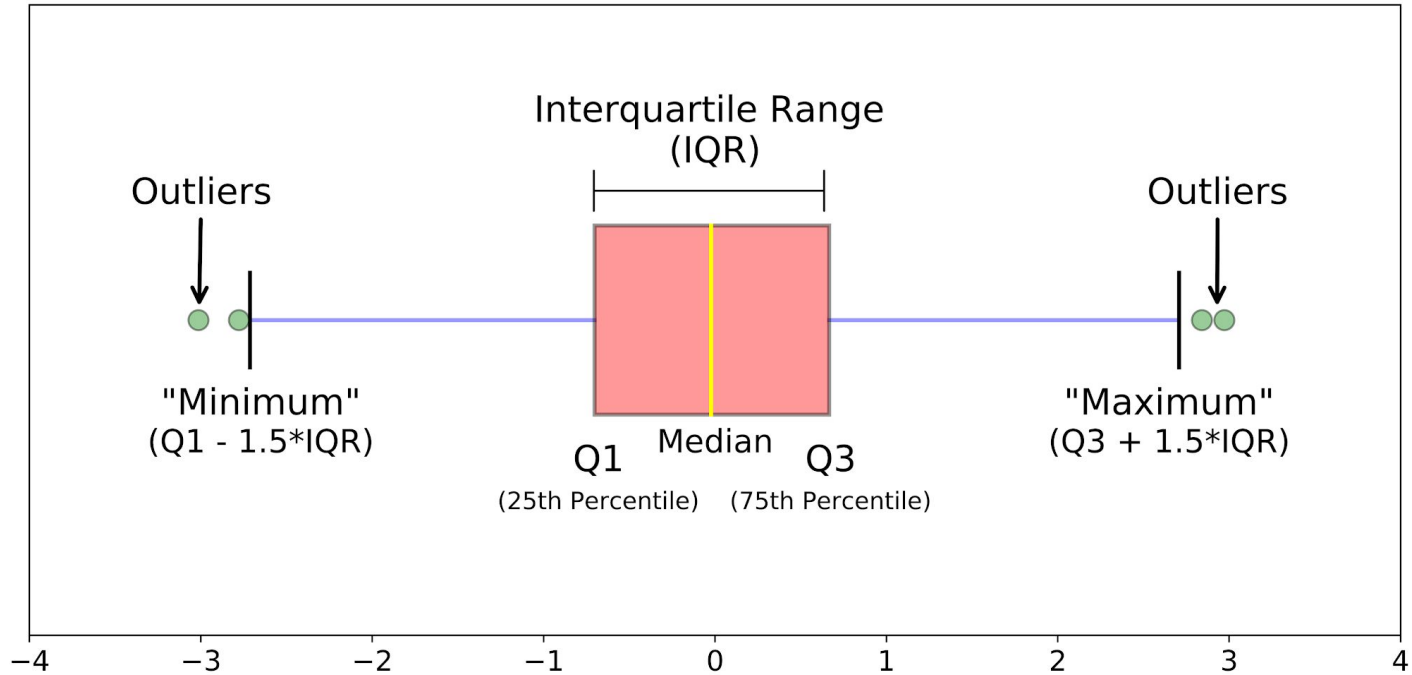
This allows the inspection of the data for its underlying distribution (e.g., normal distribution), outliers, skewness, etc.

In a histogram, it is the area of the bar that indicates the frequency of occurrences for each bin.

This means that the height of the bar does not necessarily indicate how many occurrences of scores there were within each individual bin. It is the product of height multiplied by the width of the bin that indicates the frequency of occurrences within that bin.



Box Plots



30.11.2020

Theme Layer

This is the non data layer of ggplot. Here you can change the look of the graph using predefined themes .

- `theme_bw()`: a variation on `theme_grey()` that uses a white background and thin grey grid lines.
- `theme_linedraw()`: A theme with only black lines of various widths on white backgrounds, reminiscent of a line drawing.
- `theme_light()`: similar to `theme_linedraw()` but with light grey lines and axes, to direct more attention towards the data.
- `theme_dark()`: the dark cousin of `theme_light()`, with similar line sizes but a dark background. Useful to make thin coloured lines pop out.
- `theme_minimal()`: A minimalistic theme with no background annotations.
- `theme_classic()`: A classic-looking theme, with x and y axis lines and no gridlines.
- `theme_void()`: A completely empty theme.

Recap

Concepts :

1. **Data Visualization** :Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.
2. **Ggplot layers**
3. **Correlation** :Correlation is used to test relationships between quantitative variables or categorical variables. In other words, it's a measure of how things are related.
4. **Continuous data** : Continuous Data can take any value (within a range)
5. **Discrete Data** : Discrete Data can only take certain values.
6. **Categorical Data** : Categorical variables represent types of data which may be divided into groups.
7. **Skewed Distribution** : Right , left, uniform and Normal

Different Types of Graphs :

1. **Scatter plots**
 - `geom_point()`
 - used for analysing correlation
2. **Barplot**
 - `geom_bar()`
 - used for comparing categorical Data
3. **Histogram**
 - `geom_histogram()`
 - used for understanding the distribution of data
4. **Box Plots**
 - `geom_boxplot()`
 - 5 values statistical plot

4.12.2020

Lab 3

Demonstration of Exploratory Data Analysis: Missing values, Outlier treatment

Why are missing values a problem ?

1. Absence of data reduces statistical power.
2. Bias in estimation of parameters
3. Representativeness of samples will be flawed
4. Complicate the analysis of study.

How to identify Missing values ?

1. Observation
2. Graphical Representation
3. R commands

`is.na()`

`is.null()`

Types of Missing Data

Missing at Random (MAR): Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data

Missing Completely at Random (MCAR): The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.

Missing not at Random (MNAR): Two possible reasons are that the missing value depends on the hypothetical value (e.g. People with high salaries generally do not want to reveal their incomes in surveys) or missing value is dependent on some other variable's value (e.g. Let's assume that females generally don't want to reveal their ages! Here the missing value in age variable is impacted by gender variable)

Imputation Measures

1. Simple Imputation Methods

- a. Method of Deletion
- b. Imputation through Mean/Median/Mode

2. Prediction Method

We can use regression, ANOVA, Logistic regression, and various modeling techniques to perform this.

3. KNN Imputation

In this method of imputation, the missing values of an attribute are imputed using the given number of attributes that are most similar to the attribute whose values are missing.

3. Package Imputation Methods

There are several packages that are designed to help solve imputation problems.

Eg . MICE, Amelia, missForest, Hmisc, mi

MICE PACKAGE

MICE (Multivariate Imputation via Chained Equations) is one of the commonly used package by R users. Creating multiple imputations as compared to a single imputation (such as mean) takes care of uncertainty in missing values.

MICE assumes that the missing data are Missing at Random (MAR), which means that the probability that a value is missing depends only on observed value and can be predicted using them. It imputes data on a variable by variable basis by specifying an imputation model per variable.

For example: Suppose we have X_1, X_2, \dots, X_k variables. If X_1 has missing values, then it will be regressed on other variables X_2 to X_k . The missing values in X_1 will be then replaced by predictive values obtained. Similarly, if X_2 has missing values, then X_1, X_3 to X_k variables will be used in prediction model as independent variables. Later, missing values will be replaced with predicted values.

MICE PACKAGE

By default, linear regression is used to predict continuous missing values. Logistic regression is used for categorical missing values. Once this cycle is complete, multiple data sets are generated. These data sets differ only in imputed missing values. Generally, it's considered to be a good practice to build models on these data sets separately and combining their results.

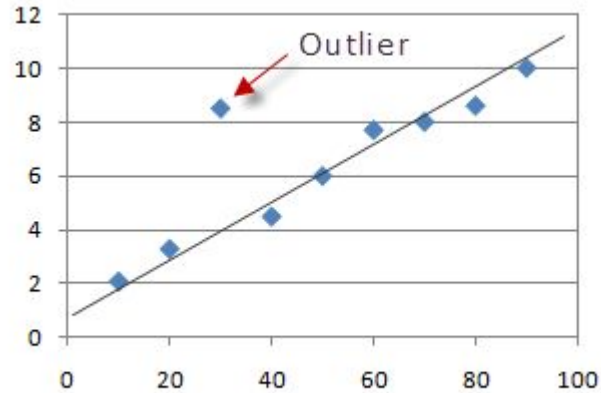
Precisely, the methods used by this package are:

1. PMM (Predictive Mean Matching) - For numeric variables
2. logreg(Logistic Regression) - For Binary Variables(with 2 levels)
3. polyreg(Bayesian polytomous regression) - For Factor Variables (≥ 2 levels)
4. Proportional odds model (ordered, ≥ 2 levels)

Outliers

What are outliers ?

A simple example of an outlier is here, a point that deviates from the overall pattern.



Outlier Detection Methods

- Extreme Value Analysis
- Z-score method/ Statistical Testing
- K Means clustering-based approach
- Visualizing the data

Outlier Treatment Methods

- Mean/Median or random Imputation
- Trimming
- Top, Bottom, and Zero Coding
- Discretization

Statistical Tests for finding outliers.

1. Grubb's Test
2. Rosner's Test
3. Dixon's Test

Grubbs Test

Grubbs' test is used to find a single outlier in a normally distributed data set. The test finds if a minimum value or a maximum value is an outlier.

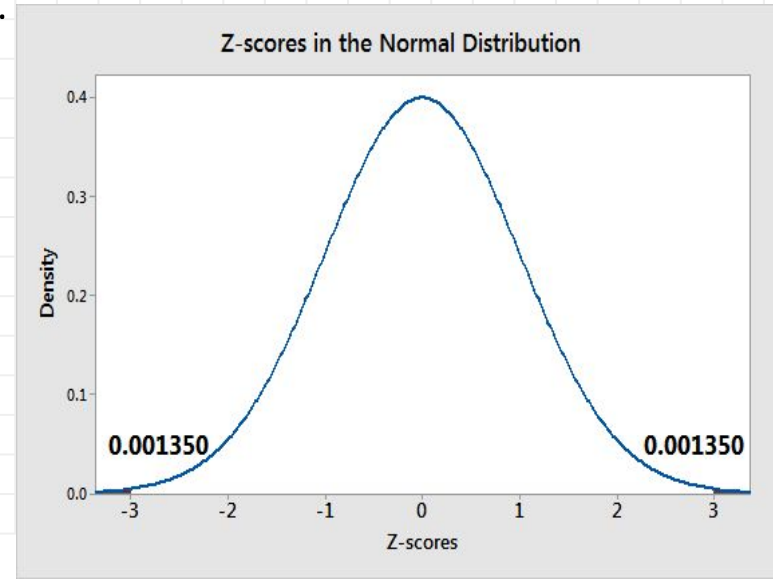
The test is a deceptively simple one to run. It checks for outliers by looking for the maximum of the absolute differences between the values and the mean.

It is used to answer two questions :

1. Is the maximum value an outlier
2. Is the minimum value an outlier

H_0 : there are no outliers in the data

H_a : the maximum value is an outlier



Dataset Information

Format

Dataset containing 474 observations across 6 variables

job

Ordered factor indicating job category, with levels "custodial", "admin" and "manage".

education

Education in years.

gender

Factor indicating gender.

minority

Factor. Is the employee member of a minority?

Salary

Numeric how much does the employee earn

Age

Numeric the age of the employee

Lab 4 : Univariate Analysis

Why Univariate Statistics?

Univariate analysis explores each variable in a data set, separately. It looks at the range of values, as well as the central tendency of the values. It describes the pattern of response to the variable. It describes each variable on its own.

Descriptive statistics describe and summarize data. Univariate descriptive statistics describe individual variables.

Steps in Data Exploration and Preprocessing:

1. Identification of variables and data types
2. Analyzing the basic metrics
3. Non-Graphical Univariate Analysis
4. Graphical Univariate Analysis
5. Bivariate Analysis
6. Variable transformations
7. Missing value treatment
8. Outlier treatment
9. Correlation Analysis
10. Dimensionality Reduction

Lab 5

Decision Trees

Why Do we Need Algorithms ?

There are two major solutions to all business problems.
ie Regression and Classification.

To solve those problems we have two different methods -
Supervised Learning Models and Unsupervised Learning Models.

Parameters	Supervised machine learning technique	Unsupervised machine learning technique
Process	In a supervised learning model, input and output variables will be given.	In unsupervised learning model, only input data will be given
Input Data	Algorithms are trained using labeled data.	Algorithms are used against data which is not labeled
Algorithms Used	Support vector machine, Neural network, Linear and logistics regression, random forest, and Classification trees.	Unsupervised algorithms can be divided into different categories: like Cluster algorithms, K-means, Hierarchical clustering, etc.

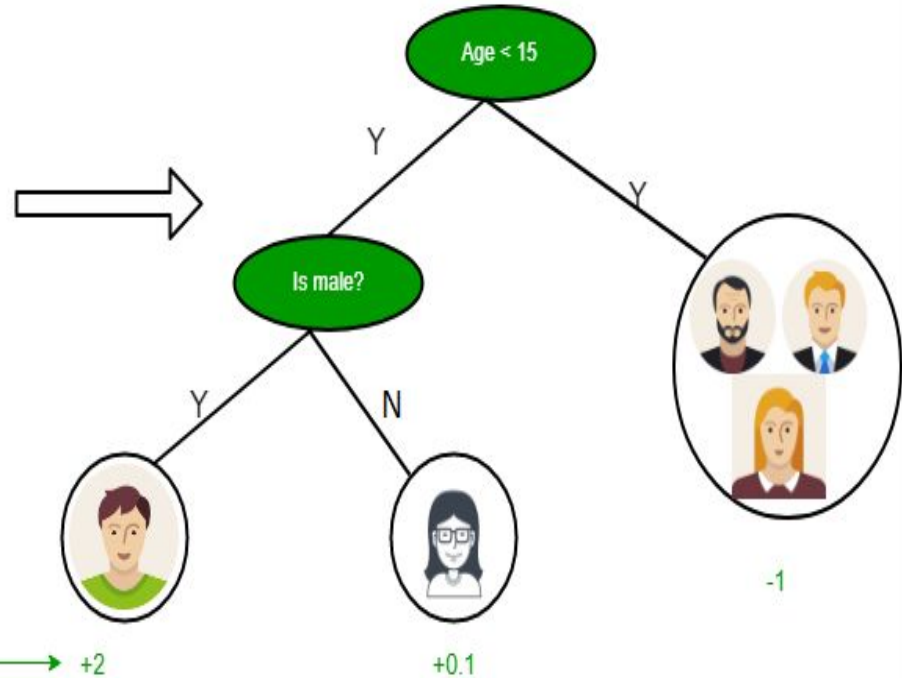
Decision Trees

- Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems.
- Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree.
- We can represent any boolean function on discrete attributes using the decision tree.

Input: Age, Gender, Occupation, . .



Does the person likes computer games



Prediction score in each leaf → +2

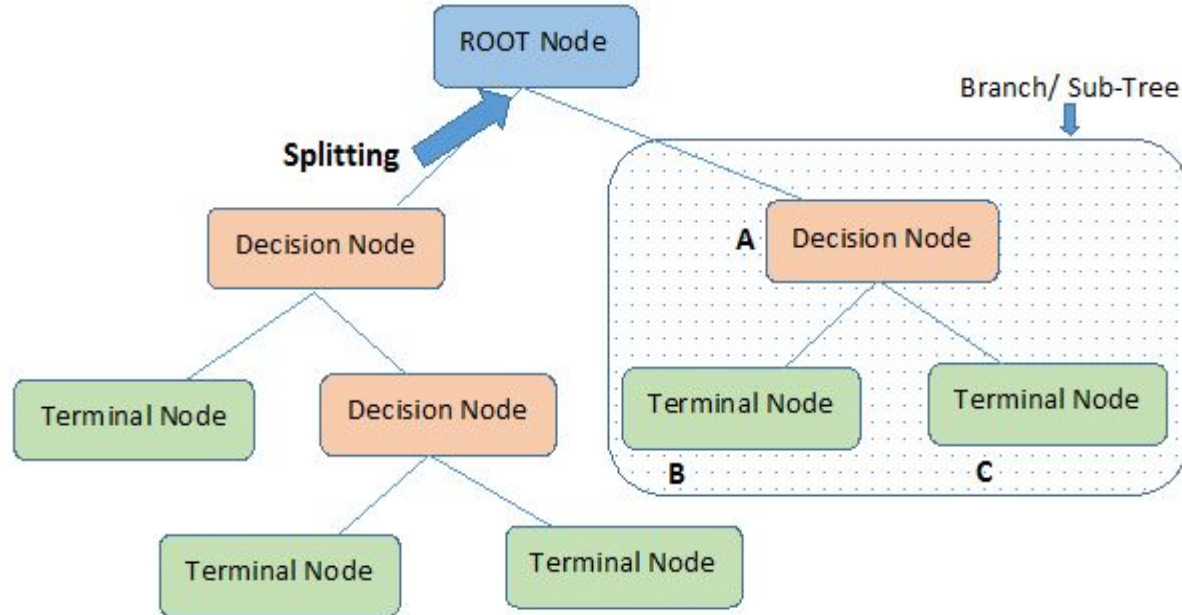
+0.1

-1

Applications of Decision Trees

- ❖ decision tree modelling is widely used in customer relationship management and fraud detection
- ❖ decision trees to investigate the relationships between the customers' needs and preferences and the success of online shopping (google scholar)
- ❖ decision trees are widely used in energy consumption and fault diagnosis
- ❖ Chang (2007) has developed a decision tree model on the basis of 516 pieces of data to explore the hidden knowledge located within the medical history of developmentally-delayed children. The created model identifies that the majority of illnesses will result in delays in cognitive development, language development, and motor development, of which accuracies are 77.3%, 97.8%, and 88.6% respectively.

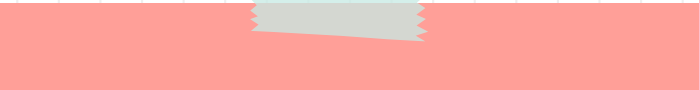
Decision Tree

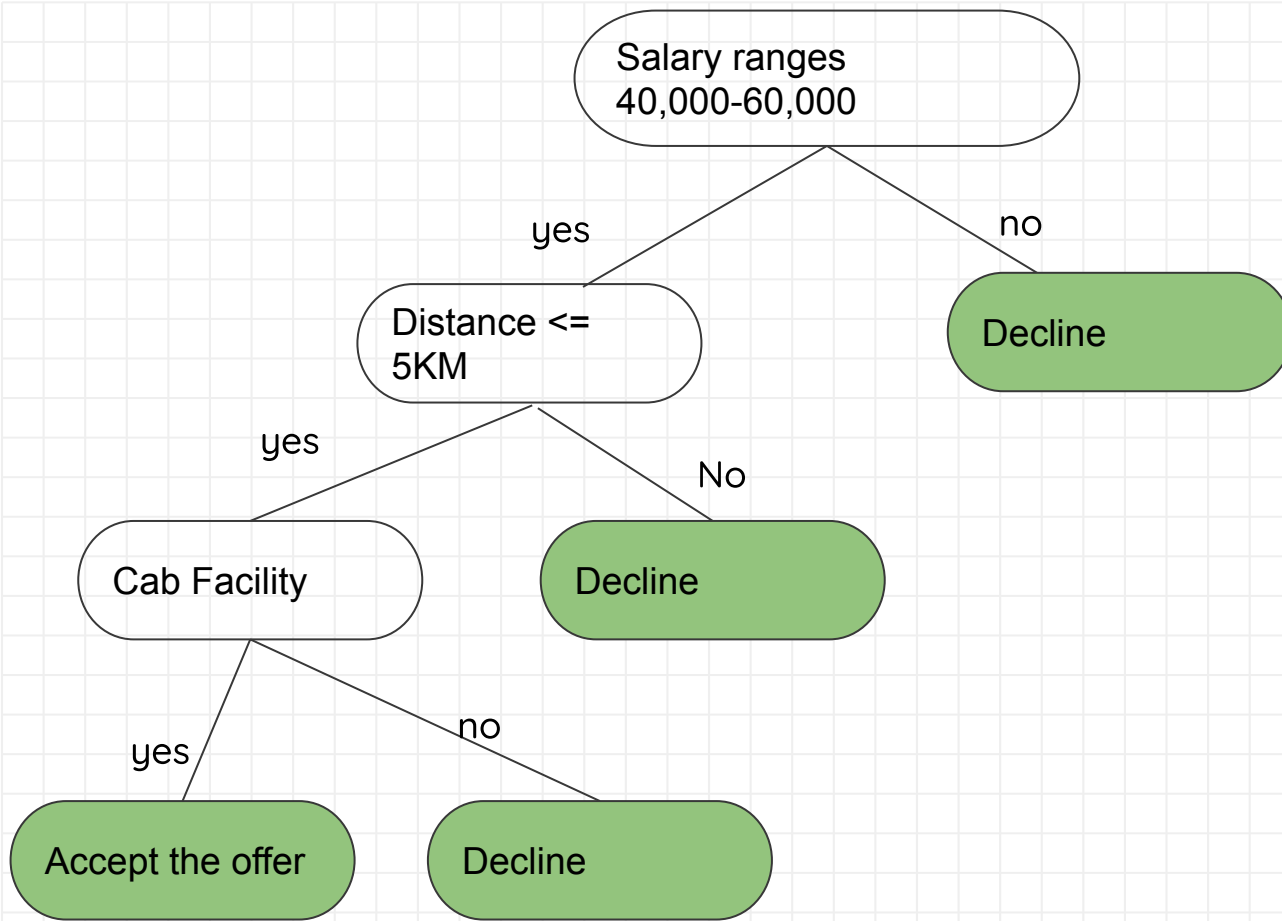


Decision Trees Terminology

- Root Node: Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.
- Leaf Node: Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.
- Splitting: Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.
- Branch/Sub Tree: A tree formed by splitting the tree.
- Pruning: Pruning is the process of removing the unwanted branches from the tree.**
- Parent/Child node: The root node of the tree is called the parent node, and other nodes are called the child nodes.

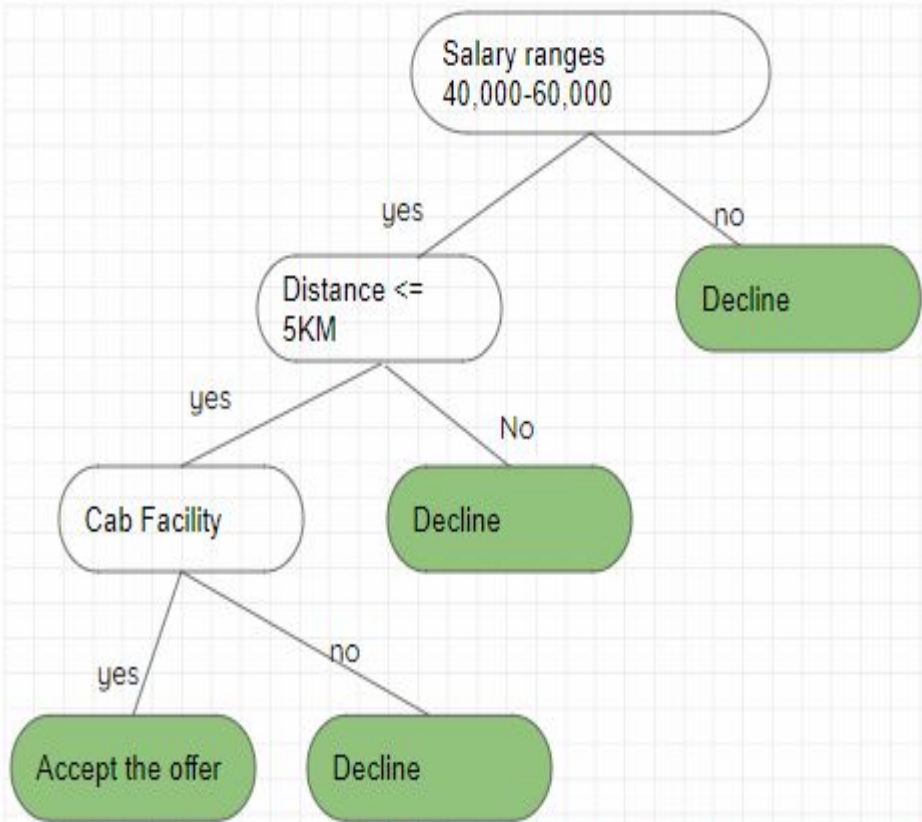
How does the Decision Tree algorithm Work?

- Step-1: Begin the tree with the root node, says S, which contains the complete dataset.
 - Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).
 - Step-3: Divide the S into subsets that contains possible values for the best attributes.
 - Step-4: Generate the decision tree node, which contains the best attribute.
 - Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.
- 



Build a DT to negotiate the right salary.

- Salary must be root node
- Distance
- Cab facility

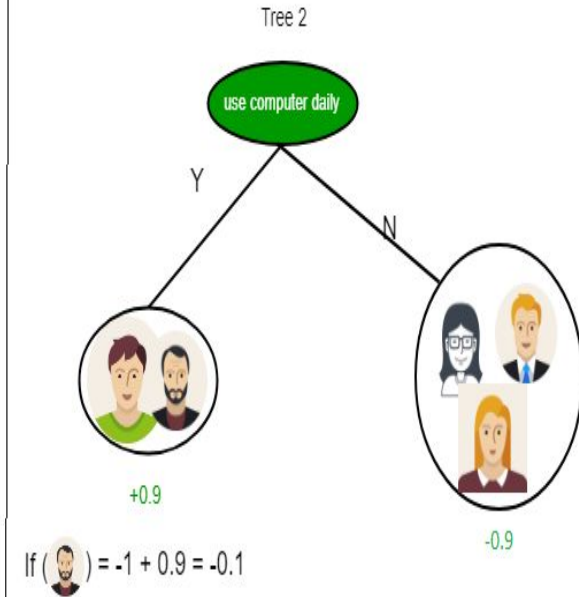
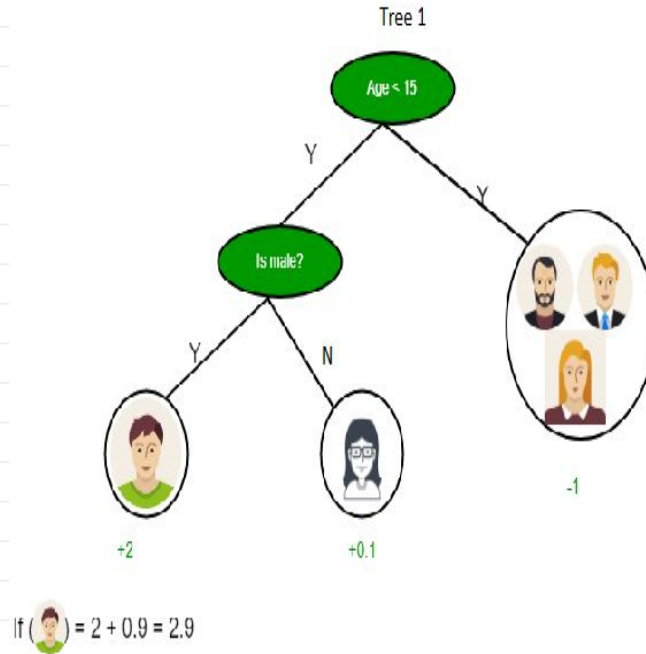


Name	Salary	Distance	Cab Facility
Raj	45,000	3 km	Yes
Sid	55,000	7 km	Yes

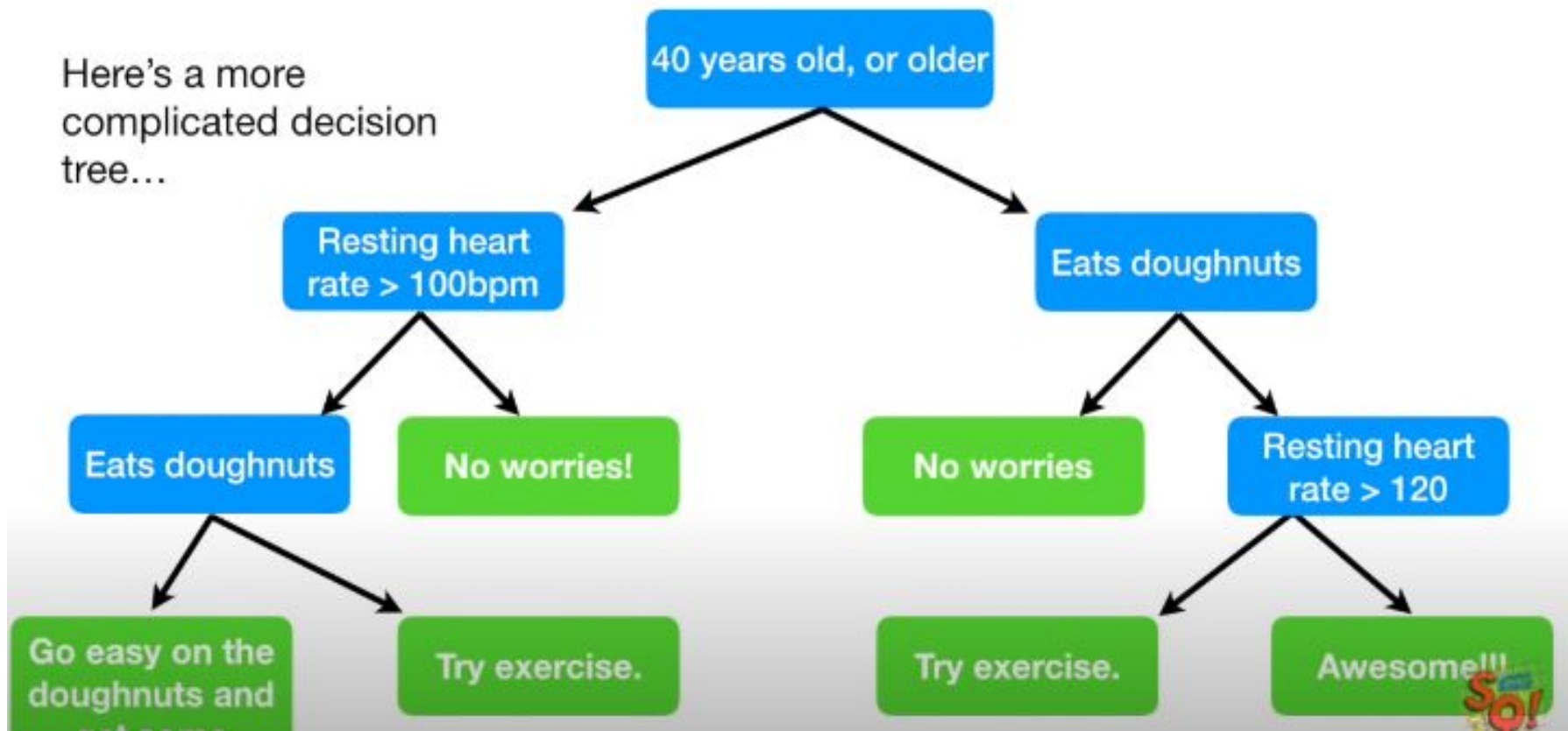
In conclusion we will procede to hire Raj

Assumptions of a Decision Tree

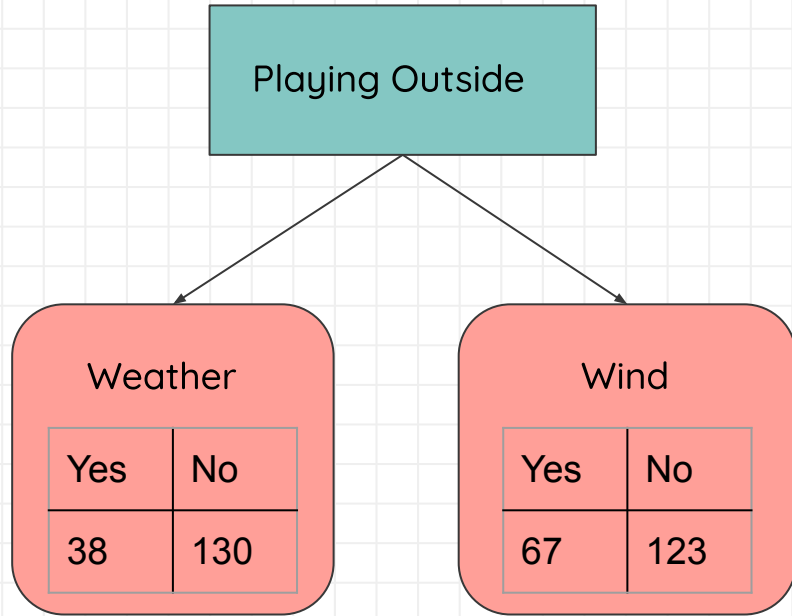
- At the beginning, we consider the whole training set as the root.
- Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.
- On the basis of attribute values records are distributed recursively.
- We use statistical methods for ordering attributes as root or the internal node.



Here's a more complicated decision tree...



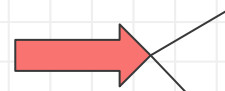
What is Impurity ?



Because none of these leaf nodes are 100% yes Playing our or 100% no playing outside they are all Impure.

Algorithms to build a DT

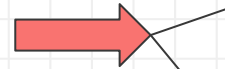
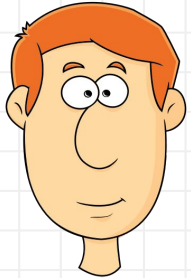
1. CART (Classification and Regression Trees) — This makes use of Gini impurity as the metric.
2. ID3 (Iterative Dichotomiser 3) — This uses entropy and information gain as metric.



50% - 100

50% - (-45)

Expected Value is 27.5



50% - 75

50% - (-10)

Expected value is 32.5

Other Terminology related to DT

Entropy

In machine learning, entropy is a measure of the randomness in the information being processed. The higher the entropy, the harder it is to draw any conclusions from that information.

$$H(X) = - \sum_{i=1}^n P(x_i) \log_b P(x_i)$$

Other Terminology related to DT

Information Gain

Information gain can be defined as the amount of information gained about a random variable or signal from observing another random variable. It can be considered as the difference between the entropy of parent node and weighted average entropy of child nodes.

$$IG(S, A) = H(S) - H(S, A)$$

Alternatively,

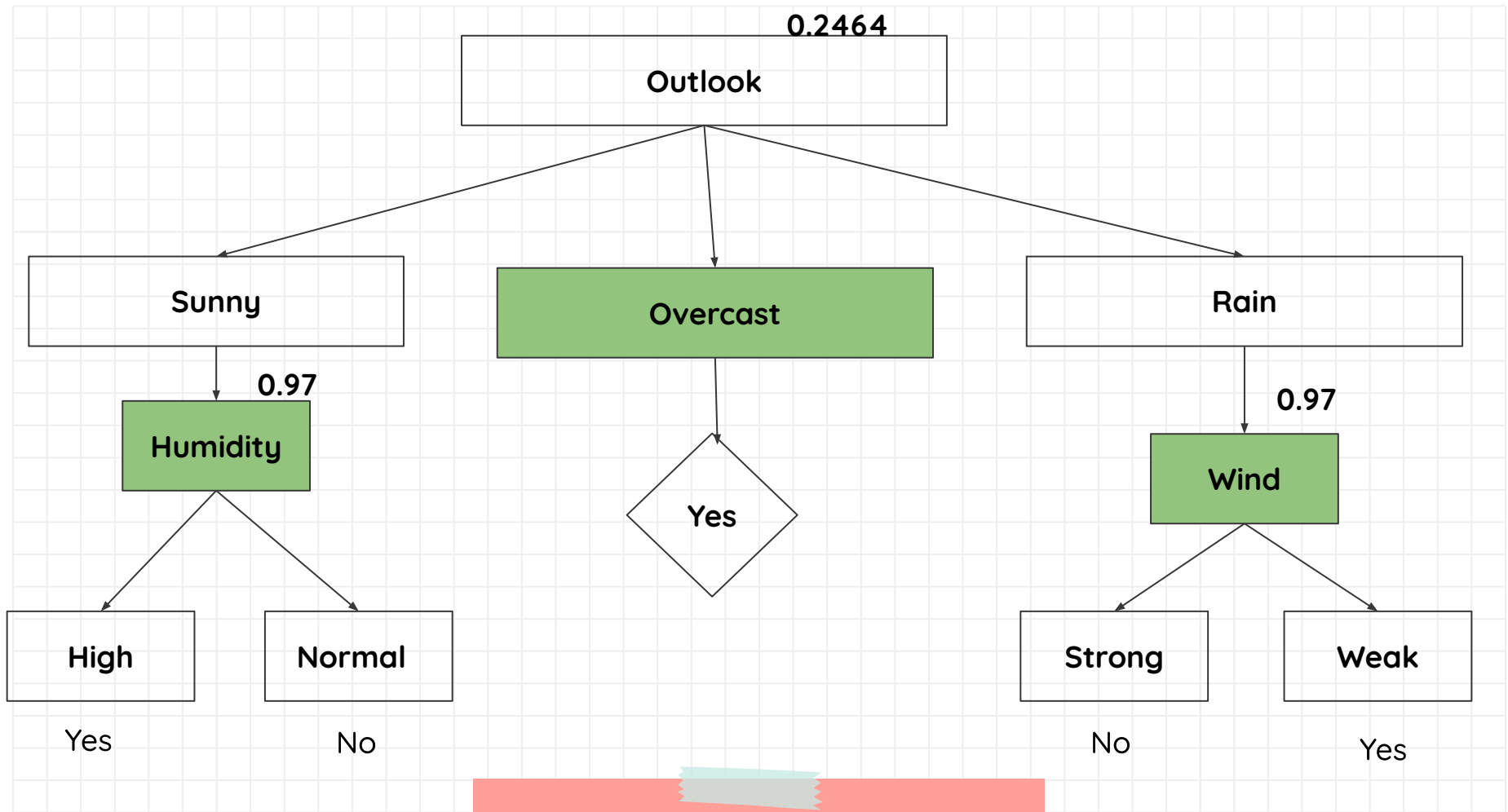
$$IG(S, A) = H(S) - \sum_{i=0}^n P(x) * H(x)$$

Other Terminology related to DT

Gini Impurity

Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.

$$Gini(E) = 1 - \sum_{j=1}^c p_j^2$$



Advantages of a DT

- Decision trees are able to generate understandable rules.
- Decision trees perform classification without requiring much computation.
- Decision trees are able to handle both continuous and categorical variables.
- Decision trees provide a clear indication of which fields are most important for prediction or classification.

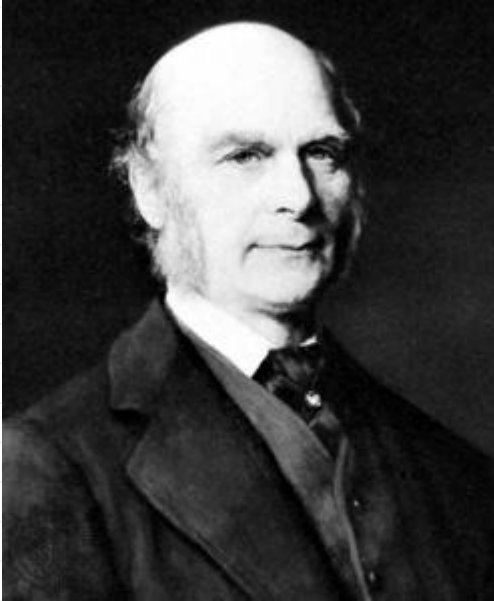
Disadvantages of a DT

- Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
- Decision trees are prone to errors in classification problems with many class and relatively small number of training examples.
- Decision tree can be computationally expensive to train. The process of growing a decision tree is computationally expensive. At each node, each candidate splitting field must be sorted before its best split can be found. In some algorithms, combinations of fields are used and a search must be made for optimal combining weights. Pruning algorithms can also be expensive since many candidate sub-trees must be formed and compared.

Lab 6

Linear Regression

Introduction



Sir Francis Galton (1877)

Regression and Correlation analysis will show us how to determine both nature and strength of the relationship between two variables.

Identifying the X and the Y

Target variable

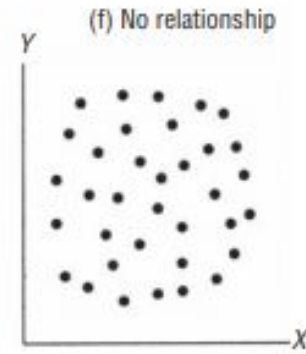
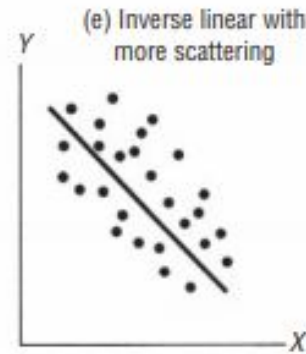
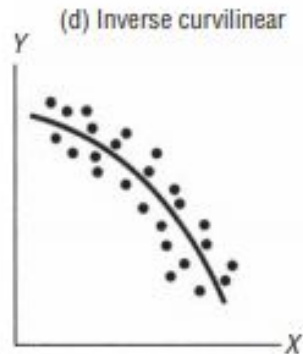
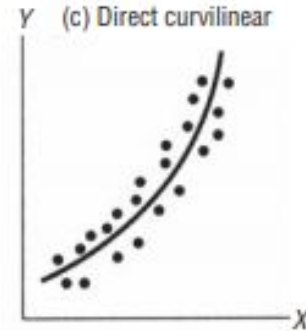
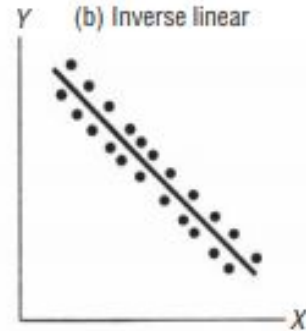
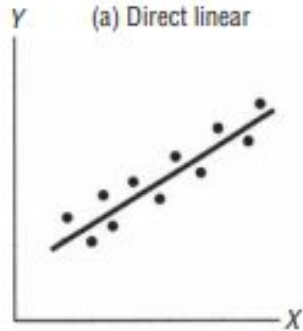
The “target variable” is the variable whose values are to be modeled and predicted by other variables. It is analogous to the dependent variable (i.e., the variable on the left of the equal sign) in linear regression. There must be one and only one target variable in any analysis.

Predictor variable

A “predictor variable” is a variable whose values will be used to predict the value of the target variable. It is analogous to the independent variables (i.e., variables on the right side of the equal sign) in linear regression. There must be at least one predictor variable specified; there may be many predictor variables

Correlation analysis

A scatter diagram can give us two types of information. Visually, we can look for patterns that indicate that the variables are related. Then, if the variables are related, we can see what kind of line, or estimating equation, describes this relationship



How does the Regression Algorithm Work ?

Identify the known and unknown Variable



Estimating equation

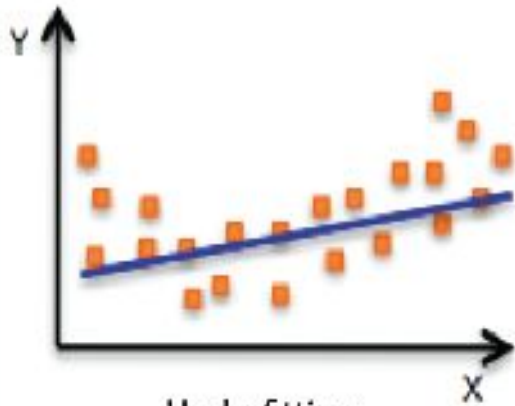


Fitting a line

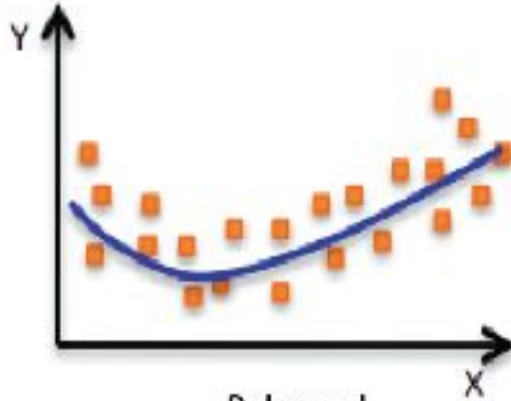


Predict

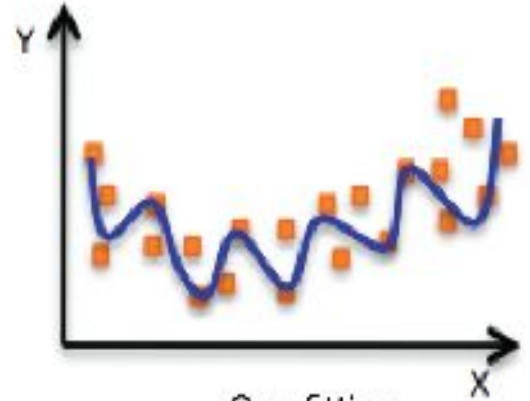
Fitting a Line



Underfitting



Balanced

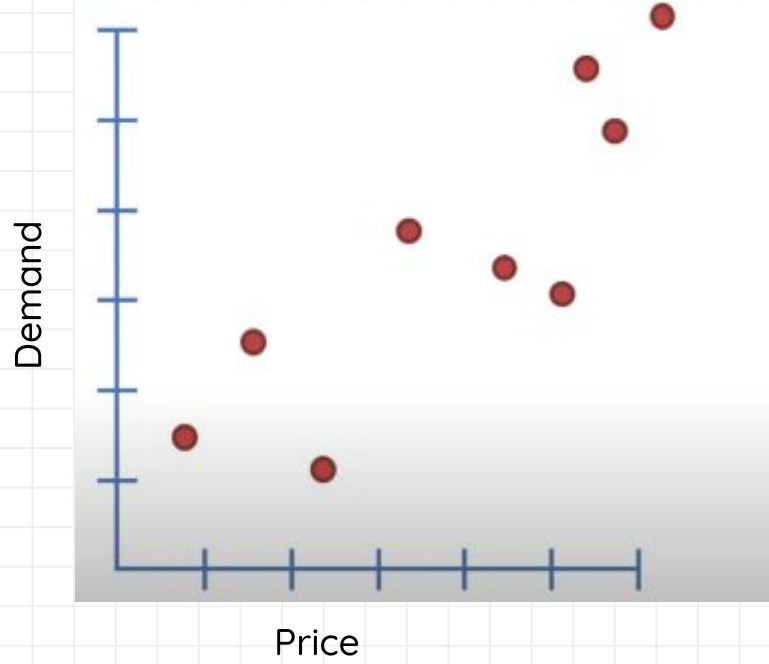


Overfitting

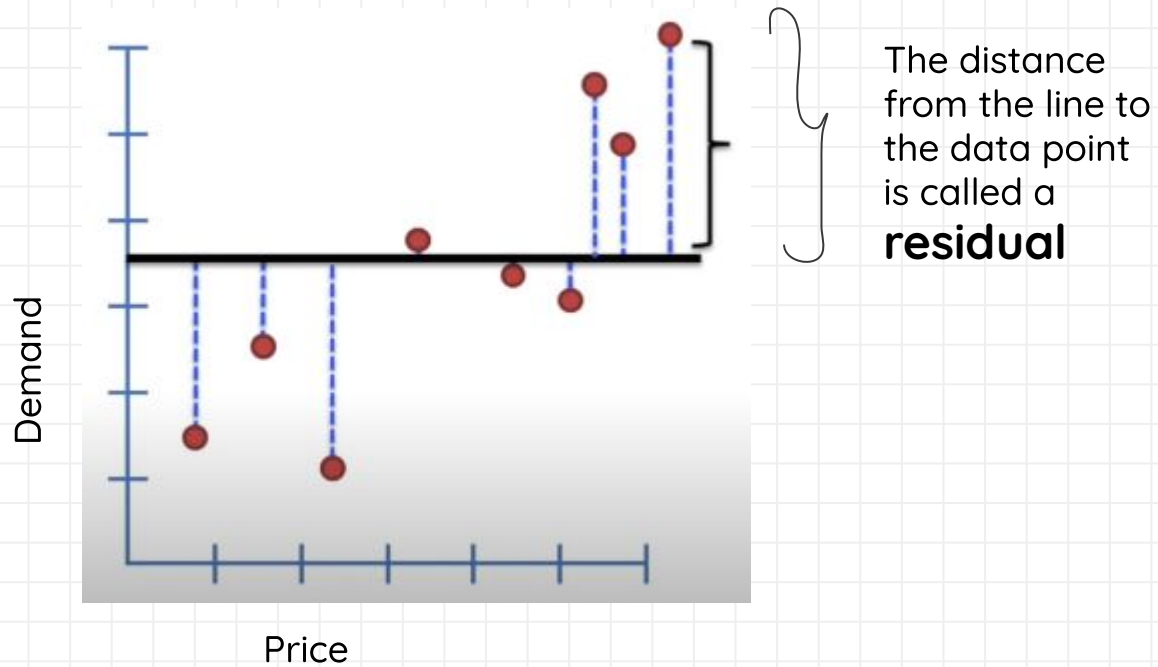
3 main ideas of linear regression

1. We use least squares to fit a line
2. Calculate R-squared R^2
3. Calculate the p-value

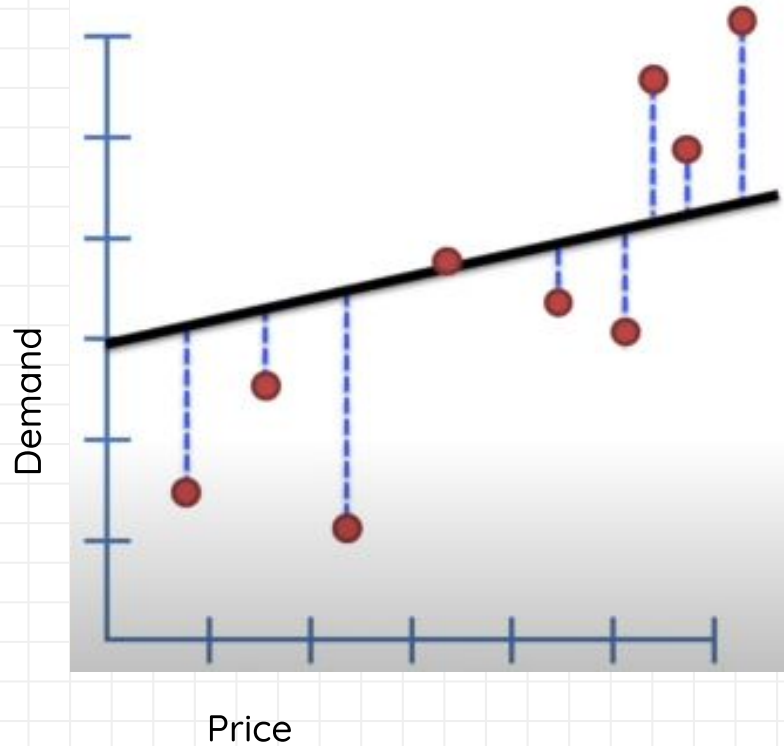
We are part of Auctioning group where we sell antiques, collectables etc

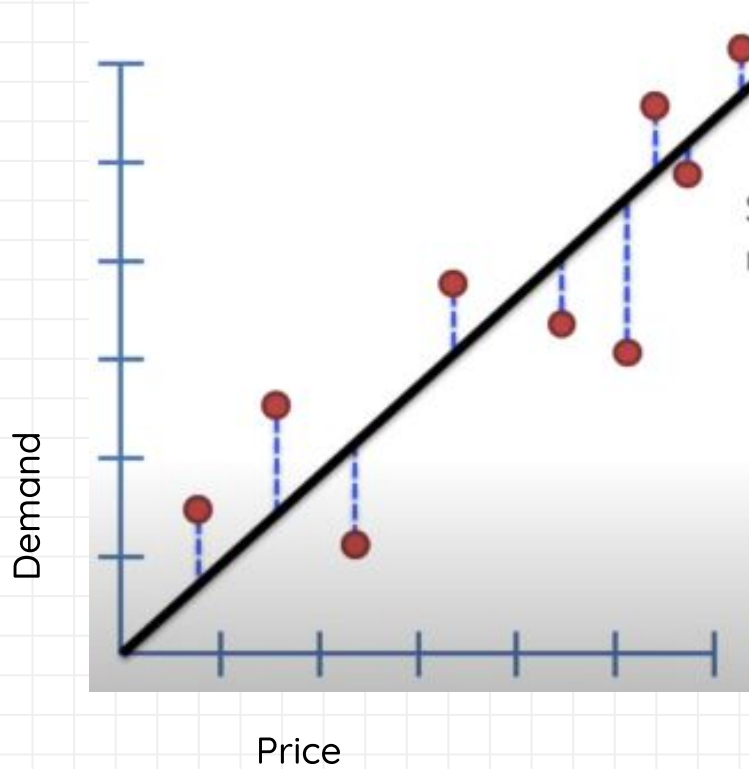


1. We will fit one line randomly through the data
2. Measure the distance from the line to the data, square the value and add them up

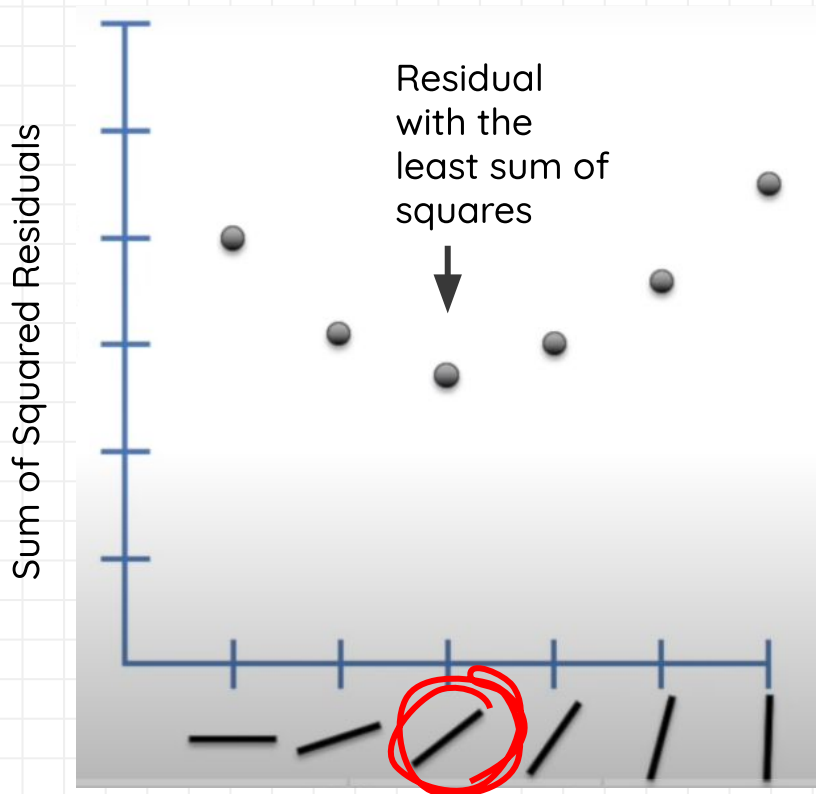


3. Rotate the line and find out the residuals , square them and sum up the squares.





Find the residuals
, square them and
sum them up

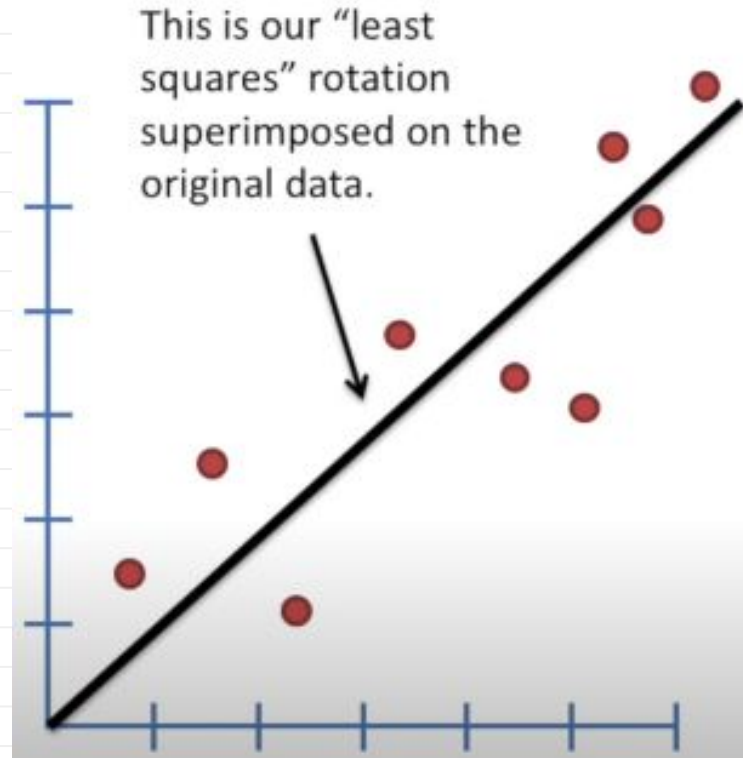


Rotation

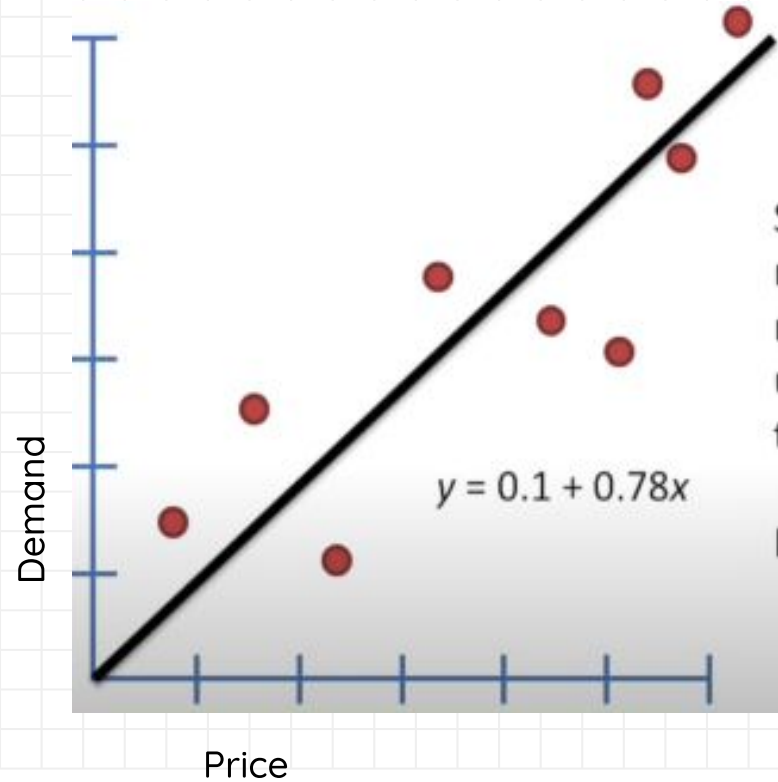
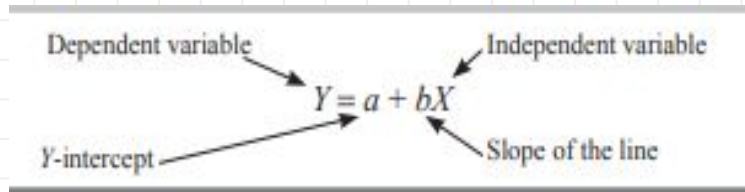
This is the rotation with least squares so this is the one that will fit our data

This is why we call the method as **least squares** method.

If this method is done for linear Regression analysis we call it **Ordinary Least Squares (OLS)**

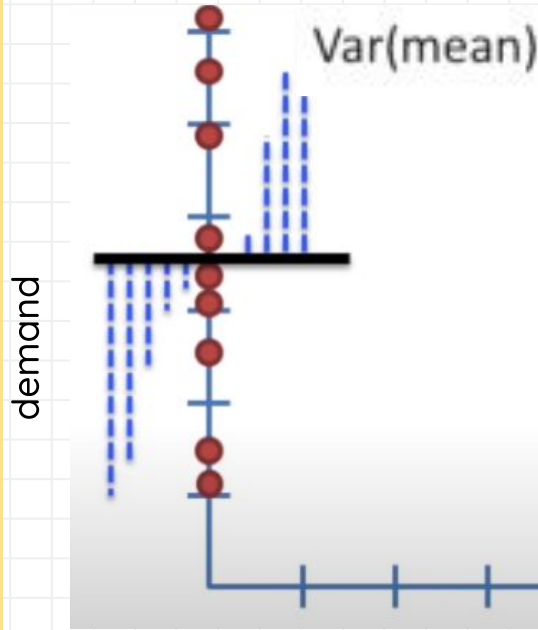


The OLS method will give us the estimation in the form an equation and we call this equation as the **regression equation**.



How good is our guess ?

To find out how good our guess is we have to calculate R^2



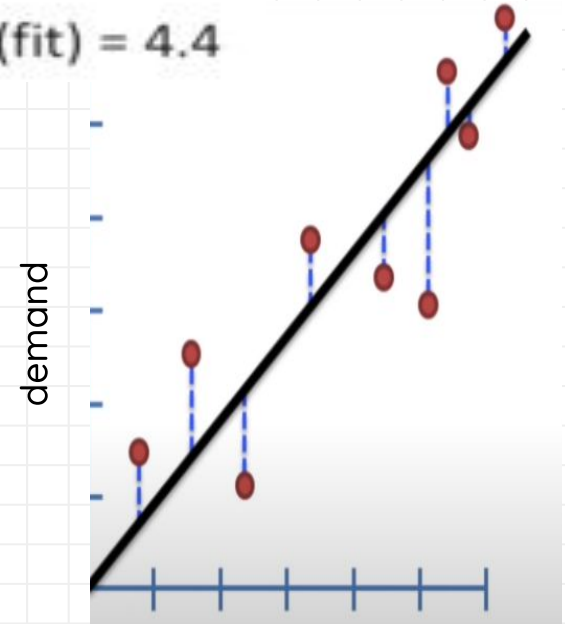
Price



$$R^2 = \frac{\text{Var}(\text{mean}) - \text{Var}(\text{fit})}{\text{Var}(\text{mean})}$$

$$R^2 = \frac{11.1 - 4.4}{11.1}$$

$$R^2 = 0.6 = 60\%$$



Price

$$R^2$$

Explains the percentage of variation caused caused by the independent variables.

What is the P value ?

When you perform a hypothesis test in statistics, a p -value helps you determine the significance of your results. **Hypothesis tests** are used to test the validity of a claim that is made about a population. This claim that's on trial, in essence, is called the *null hypothesis*.

The *alternative hypothesis* is the one you would believe if the null hypothesis is concluded to be untrue. The evidence in the trial is your data and the statistics that go along with it. All hypothesis tests ultimately use a p -value to weigh the strength of the evidence (what the data are telling you about the population). The p -value is a number between 0 and 1 and interpreted in the following way:

- A small p -value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.
- A large p -value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.
- p -values very close to the cutoff (0.05) are considered to be marginal (could go either way). Always report the p -value so your readers can draw their own conclusions.

For example, suppose a pizza place claims their delivery times are 30 minutes or less on average but you think it's more than that.

Frame your hypothesis

Ho : The pizza delivery is on time

P-value = 0.001

H1 : Not on time .

Since $0.001 < 0.05$ we reject the null hypothesis and conclude that the pizza place delivery on an average is more than 30 mins.

Regression Hypothesis

What do we measure while conducting a Regression Analysis ?

We measure the strength of the relationship between two variables.
In other the terms the significance of the independent variable wrt the dependent variable.

H_0 : variables are not significant ie X doesn't influence Y

H_1 : variables are significant

$$H_0 : \beta = 0$$

$$H_1 : \beta \neq 0$$

Case Study

Edith and Co is a publishing firm that uses various forms of advertising to boost its sales. For each book released they promote the book in a combination of advertisements through Newspapers, TV and Radio. They have hired you to help them identify which advertising technique is the most effective. You are provided with a dataset containing the cost of advertising and sales generated through the same.

Interpreting Linear Regression Results in Rstudio

Call:

```
lm(formula = df$Sales ~ df$Newspaper)
```

Residuals:

Min	1Q	Median	3Q	Max
-11.2272	-3.3873	-0.8392	3.5059	12.7751

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	12.35141	0.62142	19.88	< 2e-16	***
df\$Newspaper	0.05469	0.01658	3.30	0.00115	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.092 on 198 degrees of freedom

Multiple R-squared: 0.05212, Adjusted R-squared: 0.04733

F-statistic: 10.89 on 1 and 198 DF, p-value: 0.001148

Call:

```
lm(formula = df$Sales ~ df$Newspaper)
```

$$Y = \beta_0 + \beta_1 \cdot X$$

Sales = y intercept + Slope * cost of newspaper advertising.

Residuals:

Min	1Q	Median	3Q	Max
-11.2272	-3.3873	-0.8392	3.5059	12.7751

Coefficients:

```
Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.35141 0.62142 19.88 < 2e-16 ***
df$Newspaper 0.05469 0.01658 3.30 0.00115 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

These are coefficients of our parameters.

Intercept estimate : Average value of Y when X=0.

Slope estimate interpretation : Given one unit increase in X this is the expected change in Y on average.

Coefficients:

```

Estimate Std. Error t-value Pr(>|t|)
(Intercept) 12.000
df$Newspaper (

```

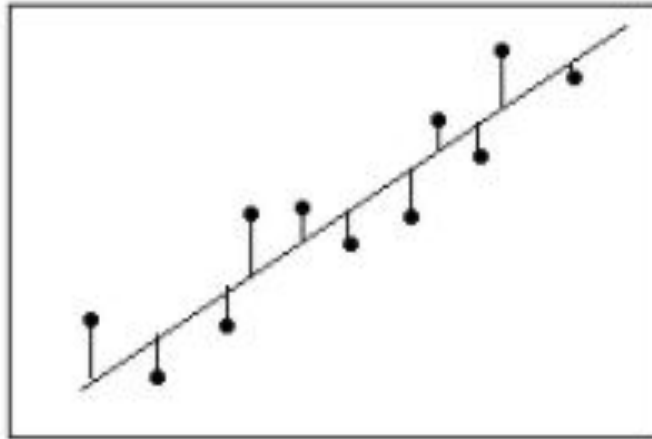
Signif. codes:

Standard errors

The standard error represents the variability of coefficients

Std Error represents the variability of coefficients

because it indicates that the observations are closer to the fitted line.



S becomes smaller when the data points are closer to the line.

```

***
**
'. ' 0.1 ' ' 1

```

coefficient due to SAMPLING variability of the coefficients and standard error of the

observed values fall from the regression model is on

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	12.35141	0.62142	19.88	< 2e-16 ***
df\$Newspaper	0.05469	0.01658	3.30	0.00115 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

T-value for each of the coefficients .

$$t\text{-value} = \text{Estimate} / \text{Standard Error}$$

These are the p values for all the t values on each of the coefficients , given n-2 degrees of freedom.

p value measures how statistically significant each of our estimates is.

Coefficients:

```
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.35141    0.62142   19.88 < 2e-16 ***
df$Newspaper  0.05469    0.01658    3.30 0.00115 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 5.092 on 198 degrees of freedom
Multiple R-squared:  0.05212,    Adjusted R-squared:  0.04733
F-statistic: 10.89 on 1 and 198 DF,  p-value: 0.001148
```

This is the Residual standard error of the model it quantifies how well or how poorly the model does at predicting Y values in the data , on average.

This is like the average error of the model.

$$\sigma_e = \frac{\sqrt{\text{sum of squared error}}}{\sqrt{\text{Deg of freedom error}}}$$

Coefficients:

```
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.35141    0.62142   19.88 < 2e-16 ***
df$Newspaper  0.05469    0.01658    3.30 0.00115 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 5.092 on 198 degrees of freedom

Multiple R-squared: 0.05212,

Adjusted R-squared: 0.04733

F-statistic: 10.89 on 1 and 198 DF, p-value: 0.001148

This is the R squared of the model .

It is interpreted as the percentage of variation in the Dependent variable that is explained by the Variation in the independent variables.

$$R^2 = \frac{SSM}{SST}$$

This is the Adjusted R squared of the model.

It is calculated the same way as R squared but it is adjusted to variables in the model.

If you add more variables in the model R sq will go up but if these variables aren't statistically significant then Adj R sq will tell us that .

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	12.35141	0.62142	19.88	< 2e-16	***
df\$Newspaper	0.05469	0.01658	3.30	0.00115	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.092 on 198 degrees of freedom
Multiple R-squared: 0.05212, Adjusted R-squared: 0.04733
F-statistic: 10.89 on 1 and 198 DF, p-value: 0.001148

This is the value for the F statistic of the model , It measures the overall significance of the model .

It is useful when there are more than one explanatory variables in the model.

$$F = \frac{\text{Mean Sq Model}}{\text{Mean Sq Error}}$$

Extra questions

A study by the Atlanta, Georgia, Department of Transportation on the effect of bus-ticket prices on the number of passengers produced the following results:

Ticket price (cents)	25	30	35	40	45	50	55	60
Passengers per 100 miles	800	780	780	660	640	600	620	620

- Plot these data.
- Develop the estimating equation that best describes these data.
- Predict the number of passengers per 100 miles if the ticket price were 50 cents. Use a 95 percent approximate prediction interval.

Extra questions

William C. Andrews, an organizational behavior consultant for Victory Motorcycles, has designed a test to show the company's supervisors the dangers of oversupervising their workers. A worker from the assembly line is given a series of complicated tasks to perform. During the worker's performance, a supervisor constantly interrupts the worker to assist him or her in completing the tasks. The worker, upon completion of the tasks, is then given a psychological test designed to measure the worker's hostility toward authority (a high score equals low hostility). Eight different workers were assigned the tasks and then interrupted for the purpose of instructional assistance various numbers of times (line X). Their corresponding scores on the hostility test are revealed in line Y .

X (number of times worker interrupted)	5	10	10	15	15	20	20	25
Y (worker's score on hostility test)	58	41	45	27	26	12	16	3

- Plot these data.
- Develop the equation that best describes the relationship between the number of times interrupted and the test score.
- Predict the expected test score if the worker is interrupted 18 times.

Why do our SLR models fail ?

1. Distribution of errors
2. Variance of Errors
3. Independence of Errors

Finding a good SLR model

To see if our model is good it is safe to assume certain rules :

1. The Error terms are to be normally distributed
2. Homoscedasticity (constant variance assumption)
3. Error Terms are Independent.

Homoscedasticity

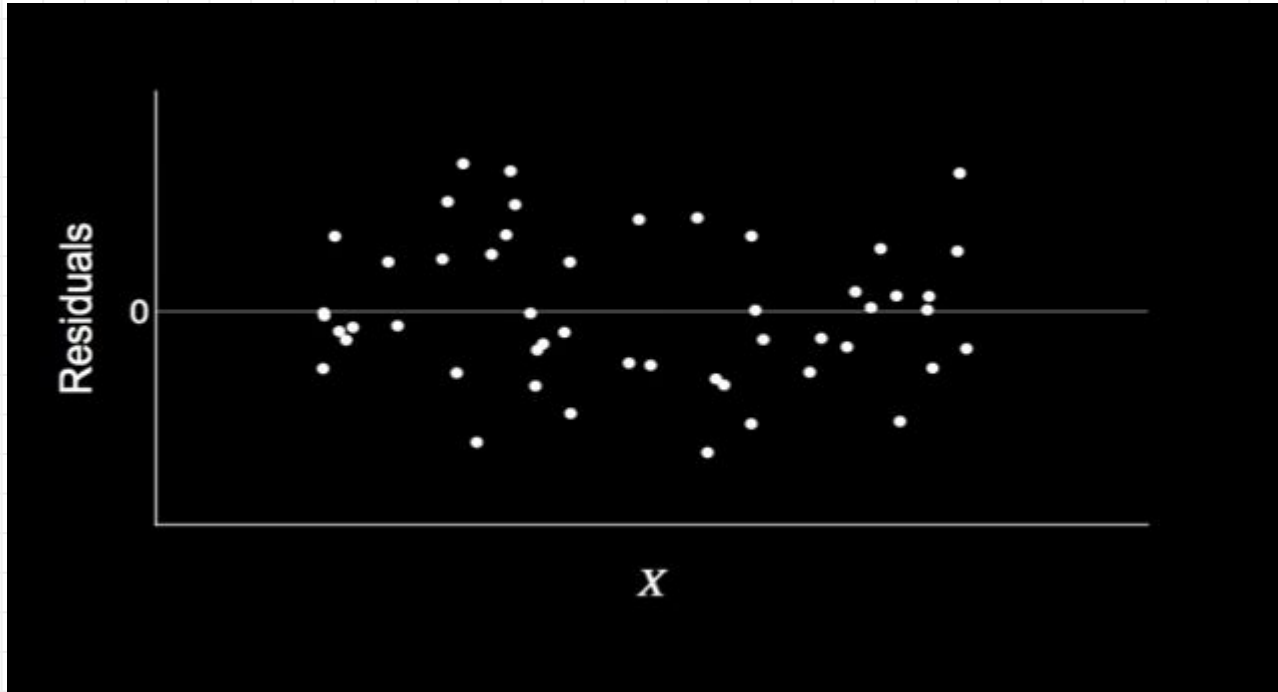
This assumption says that the errors should have the same variance at every point on x

Formulas

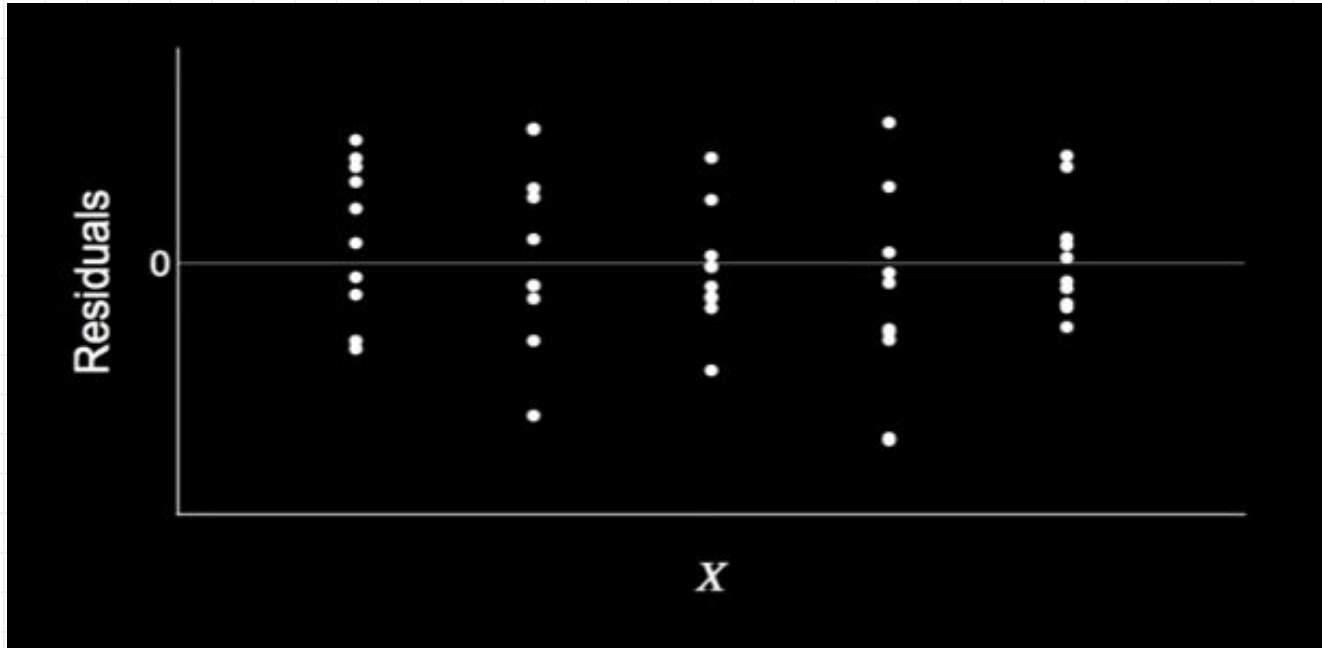
$$\varepsilon_i = Y_i - \widehat{Y}_i$$

$$\sum e_i = 0$$

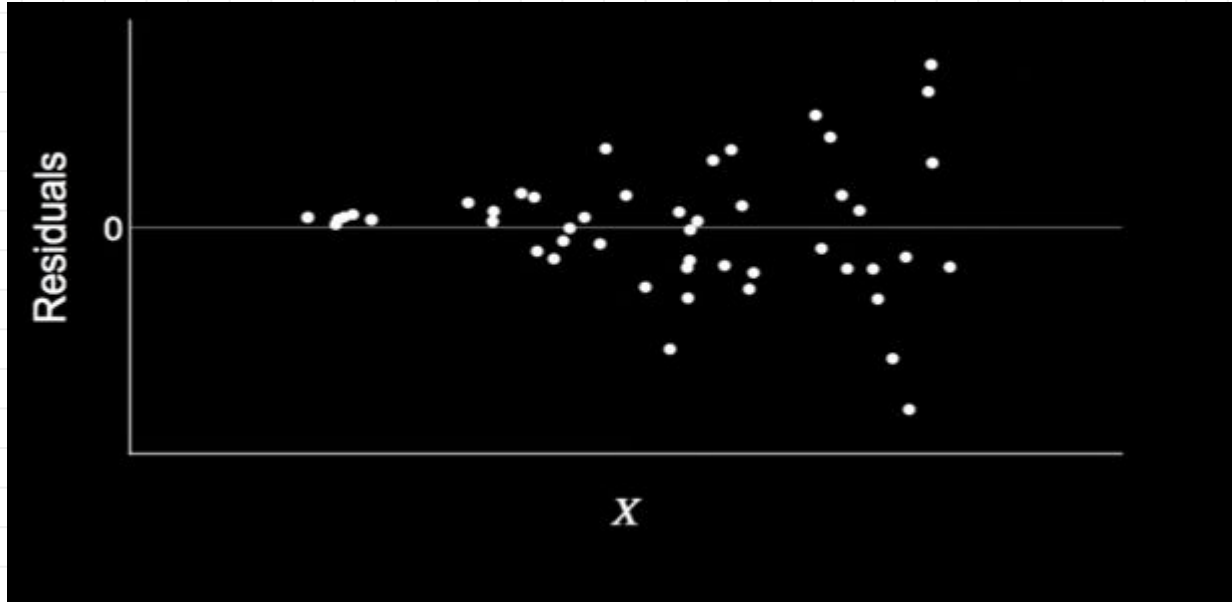
Example 1



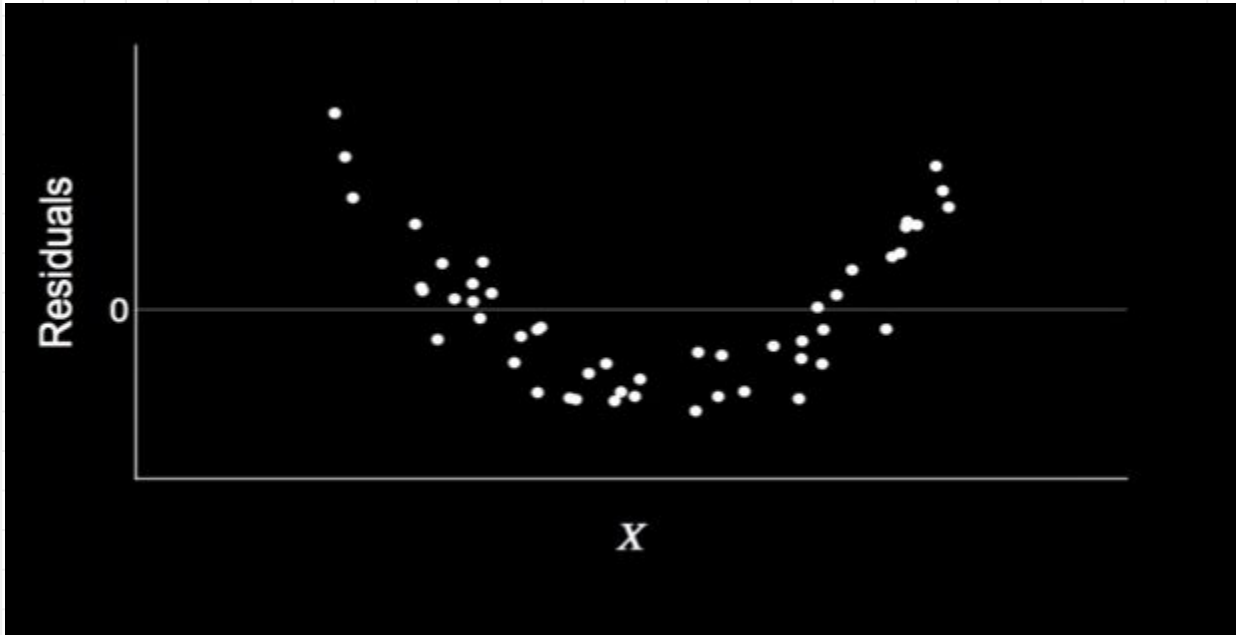
Example 2



Example 3



Example 4



How to deal with heteroskedasticity ?

- Transforming the data
(weighted, log, x-square)
- Reframing the equations

