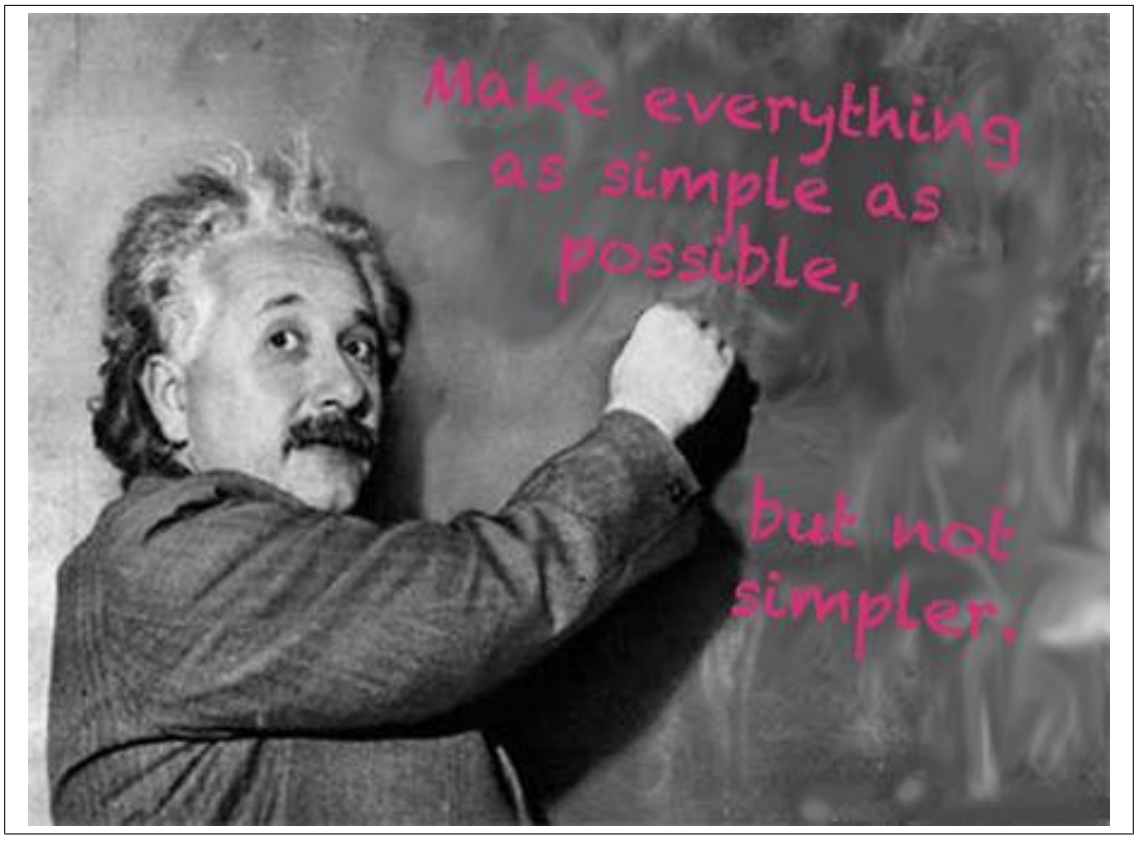
SimplifyingStats

Saptak Narula, Yatin Nandwani, Mansi Goel, Akshat Shankar, Manuj Goel

July 9, 2012



Contents

1 Descriptive Statistics 3
- 1.1 Data Types . . . 3
- 1.2 Visualization of Data . . . 3
- 1.3 Measures of Data . . . 3

2 Probability 4
- 2.1 Basic Probability . . . 4
- 2.2 Advanced Probability . . . 4

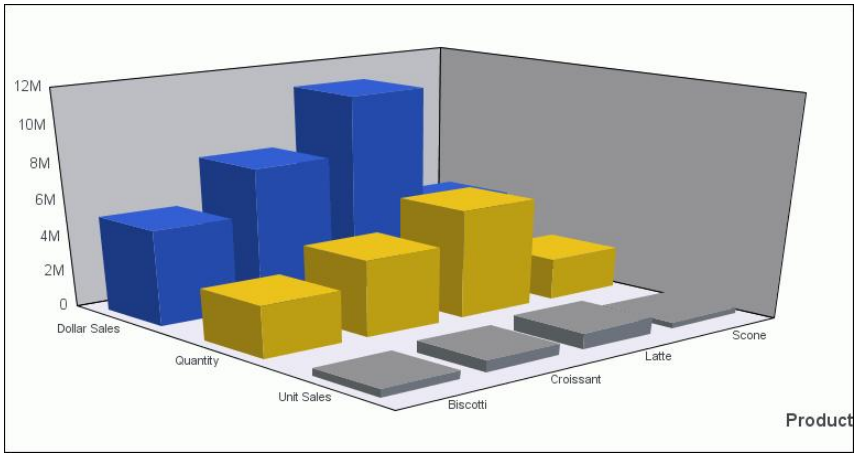
3 Inferential Statistics 5
- 3.1 Estimation Theory . . . 5
- 3.2 Hypothesis Testing . . . 5
- 3.3 Confidence Interval . . . 6

4 Unsupervised Learning 6
- 4.1 Clustering . . . 6
- 4.2 Principal Component Analysis . . . 6

5 Supervised Learning 7
- 5.1 Regression . . . 7
  - 5.1.1 Single Variable Regression . . . 7
  - 5.1.2 Multiple Variable Regression . . . 8
  - 5.1.3 Problems with Regression . . . 8
- 5.2 Classification . . . 9
  - 5.2.1 Logistic Regression . . . 9
  - 5.2.2 Decision Trees . . . 9

References 10

1 Descriptive Statistics



1.1 Data Types

- Types of Data: Nominal, Ordinal, Interval and Ratio Variables.
- Handling Data: Plotting different Data Types, Finding Measures on different Data Types and Examples of different Data Types

1.2 Visualization of Data

- Displaying the Data: Column Charts, Bar Charts, Line Charts, Pie Charts, Scatter Plots, Area Plots.
- Frequency Distribution: Histograms, Relative Frequencies, Cumulative Frequencies

1.3 Measures of Data

- Measures of Central Tendency: Arithmetic Mean, Weighted Average, Geometric Mean, Median, Mode.
- Measures of Position: Quartiles, Deciles and Percentiles
- Measures of Dispersion: Range, Interquartile Range, Variance, Standard Deviation and Mean Absolute Deviation.
- Measures of Relation: Covariance, Correlation and Linear Regression.

2 Probability

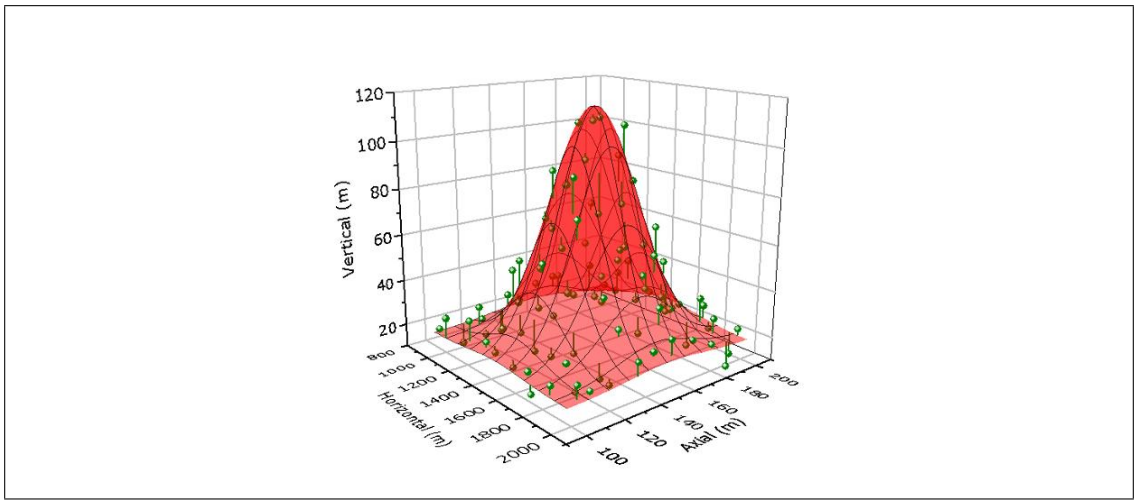
2.1 Basic Probability

- Basic Definitions: Experiment, Sample Space and Events.
- Meaning of Probability: Axioms, Examples and Intuitive understanding using the Frequentist Approach.
- Conditional Probability: Conditional Sample Space, Computation of Conditional Probability, Independence and Bayes Formula.

2.2 Advanced Probability

1. Random Variables: Definition of Random Variable, Discrete and Continuous Random Variables, Probability Mass function, Probability Density Function and Cumulative Distribution Function.
2. Examples of Discrete Random Variables: Bernoulli Random Variable, Binomial Random Variable, Geometric Random Variable and Poisson Random Variable.
3. Examples of Continuous Random Variables: Uniform Random Variable, Normal Random Variable and Exponential Random Variable.
4. Attributes of Random Variable: Expectation of Random Variables, Variance of Random Variables and Properties of Expectation and Variance.
5. Jointly Distributed Random Variables: Independence of Random Variables, Covariance, Correlation and Properties of Covariance.
6. Limit Theorems: Law of Large Numbers and Central Limit Theorem.

3 Inferential Statistics



3.1 Estimation Theory

- Introduction: Meaning of an Estimator and an Estimate.
- Desirable Qualities of an estimator: Unbiased Estimator, Efficient Estimator and Consistent Estimator.
- Attributes of an Estimator Mean Squared Error and Cramer Rao Inequality.
- Approaches of constructing an estimator: Maximum Likelihood Estimation.

3.2 Hypothesis Testing

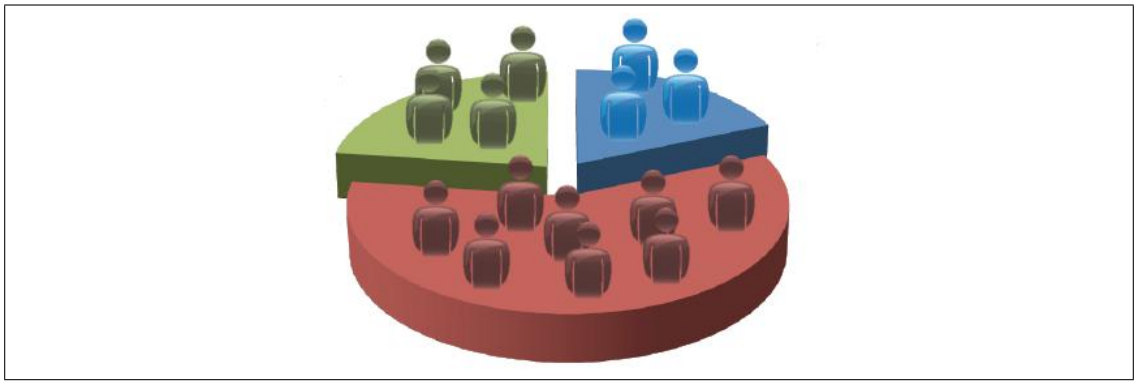
- Defining Hypothesis: Steps of hypothesis testing, Choice of the null and alternative hypothesis, Difference between one-tailed and two-tailed tests of hypothesis.
- Testing Hypothesis: Type I and Type II errors, Developing Test Statistic, Significance Level, Acceptance Region and Critical Region.
- Output of a Hypothesis Test: Definition of PValue, Interpretation of PValue and Decision Rules.
- Different Statistical Tests: Test for Population Mean with known/unknown variances, Test for equality of Population Mean, Test for variance of a Normally Distributed Population.

3.3 Confidence Interval

- Interval Estimation: Difference between Point Estimation and Interval Estimation.
- Interpretation of Confidence Intervals: Understanding that Confidence Intervals are random and not static.
- Examples of Confidence Interval: Confidence Interval for Population Mean of Normally Distributed Random Variable with known/unknown variances.

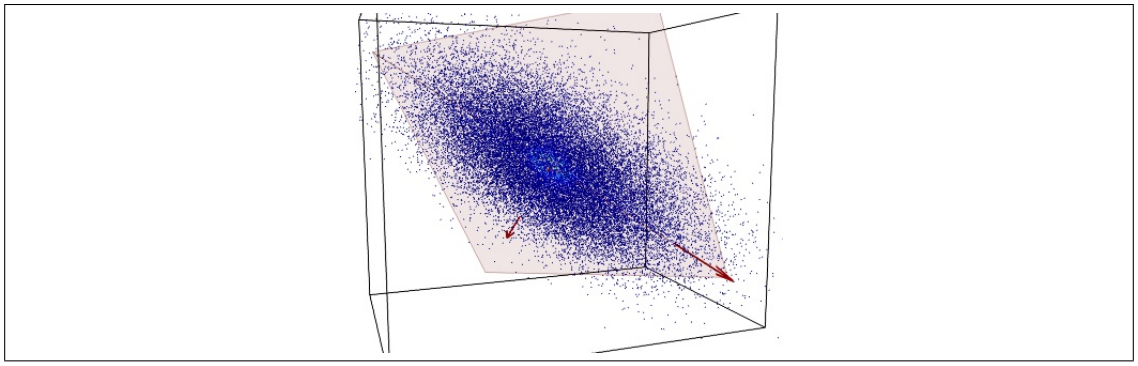
4 Unsupervised Learning

4.1 Clustering



- Types of Clustering Algorithms: Connectivity Models and Centroid Models
- K Means Clustering: Algorithm of K Means Clustering.
- K Means Clustering: Implementation of K Means Clustering, Random Initialization, Choice of number of Clusters

4.2 Principal Component Analysis

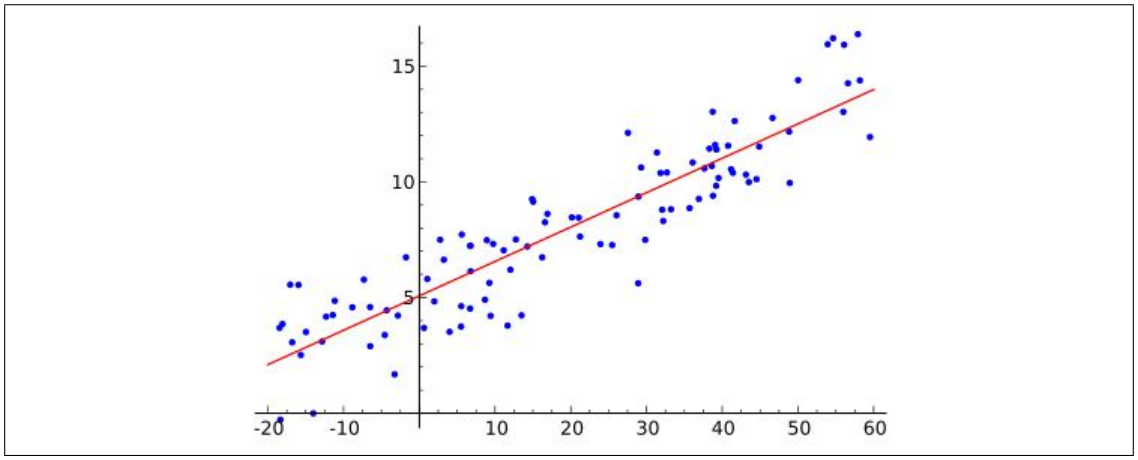


- Introduction to Principal Component Analysis: Intuition behind the technique, Application in Visualization of Data and Data Compression.
- Theory of Principal Component Analysis: Changing the Axis, Projection of data on Principal Components and Variances Across Different Dimensions.
- Algorithm of PCAs: Creating the Variance Covariance Matrix, Eigen Values and Eigen Vectors of the Matrix.
- Choosing the number of Principal Components: Explained Variance, Loss in Information and Reconstruction of the Compressed data.

5 Supervised Learning

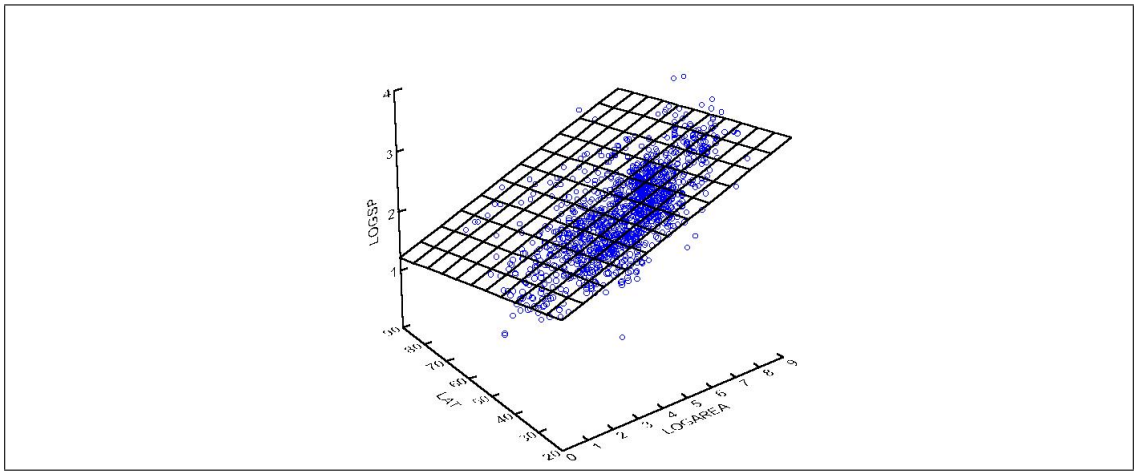
5.1 Regression

5.1.1 Single Variable Regression



- Assumptions of Linear Regression: Difference between the dependent and independent variables, Regression Equation, List of Assumptions and the reason for the Assumptions.
- Linear Regression Estimation: Estimates for the regression coefficients and interpretation of regression coefficients.
- Test of Coefficients: Formulation of null and alternative hypothesis about a population value of a regression coefficient,determination of the appropriate test statistic and testing the null hypothesis at a given level of significance.
- Analysis of Variance: Elements of ANOVA table in regression analysis, Rsquare, Relation of RSquare with Correlation and Computation of T- statistic.

5.1.2 Multiple Variable Regression

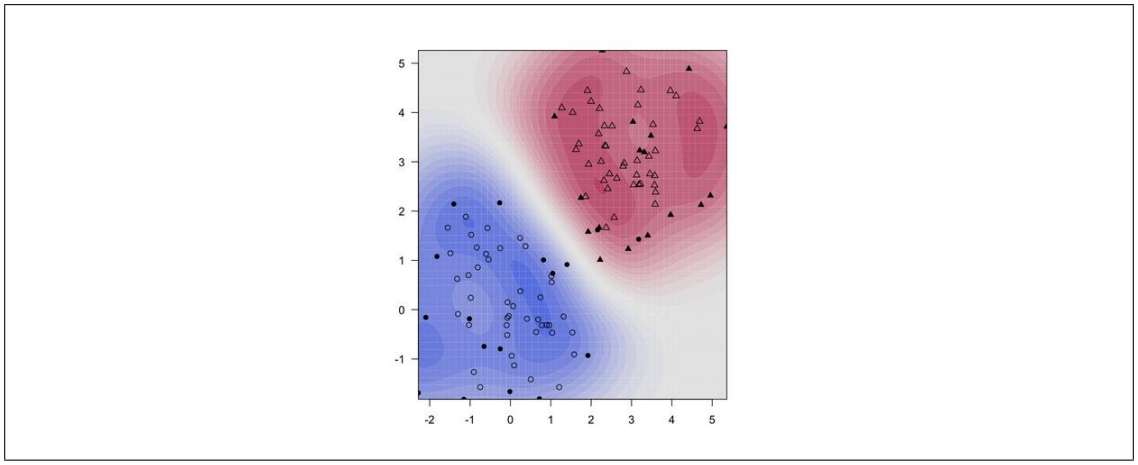


- Assumptions of Multiple Variable Regression: Regression Equation, Additional Assumptions for multiple variable case and the reason for the Assumptions.
- Regression Estimation: Estimates for the regression coefficients and interpretation of regression coefficients.
- Test of Coefficients: Formulation of null and alternative hypothesis about a population value of a regression coefficient,determination of the appropriate test statistic and testing the null hypothesis at a given level of significance.
- Dummy Variables: Representation of qualitative factor in Regression and Interpretation of dummy variables
- Analysis of Variance: Elements of ANOVA table in regression analysis, Rsquare, Relation of RSquare with Correlation, Computation of T- statistic and Computation of F-Statistic.
- Adjusted RSquare: Difference between RSquare and Adjusted RSquare.

5.1.3 Problems with Regression

- Multicollinearity: Meaning of Multicollinearity, Detection and Correction of Multicollinearity using Correlation Matrix, Principal Component Analysis Approach and Variance Inflation Factor Approach.
- Serial Correlation: Detection and Correction of Serial Correlation.
- Heteroskedasticity: Meaning of Heteroskedasticity, Types of Heteroskedasticity and the effects of Heteroskedasticity
- Limitations of regression analysis: Cases where Regression should not be used.

5.2 Classification



5.2.1 Logistic Regression

- Introduction to Logistic Regression: Difference between Linear Regression and Logistic Regression, Logistic Curve, Odds Ratio and Logistic equation.
- Assumptions of Logistic Regression: List of Assumptions and the reason for the Assumptions.
- Intuition behind Logistic Regression: Decision Boundary, Objective Function and Gradient Descent.
- Diagnostics for Logistic Regression: Different Tests for the Validity of Logistic Regression.
- Multiclass Classification: One vs All method for Multiclass Classification.

5.2.2 Decision Trees

- Introduction to Classification: Difference between Regression, Classification and Clustering.
- Introduction to Decision Trees: Basic Structure of Decision Trees and Interpretation of Decision Trees.
- Algorithm for Decision Trees: Gini Coefficient, Iterative step for Decision Trees and Stopping Criteria.

References

[1] Aczel, Amir D., Sounderpandian, Jayavel.*Complete business statistics.* 2002.

[2] Ross, Sheldon. *Introduction to Probability Models.* 2000.

[3] Wasserman, Larry. *All of Statistics.* 2012.

[4] NG, Andrew. *Online Machine Learning Course, Stanford University* 2012.