Data Analytics and Intelligence COMP8811– Lecture 2

Data Preparation, Exploration and Regression

Lecturer: Dr Neda Sakhaee

Outline

0. Assignment 1

1. Data Preparation

- Data validation
- Data transformation
- Data reduction

2. Data Exploration

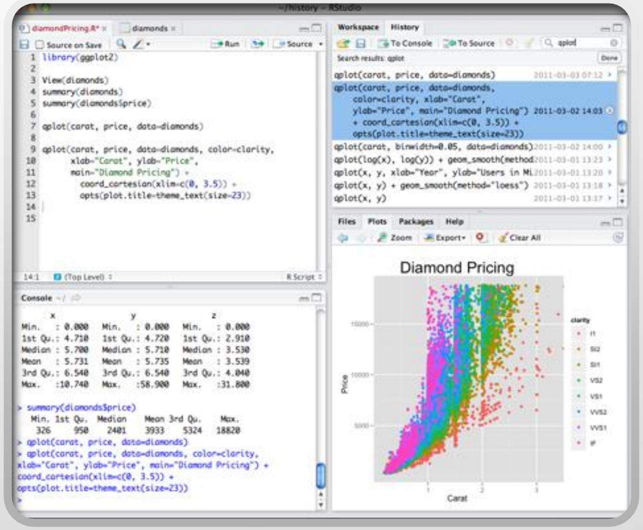
- Univariate analysis
- Bivariate analysis
- Multivariate analysis

3. Regression Analysis

- Simple linear regression
- Multiple linear regression
- Validation of regression models
- Selection of predictive variables

Assignment 1

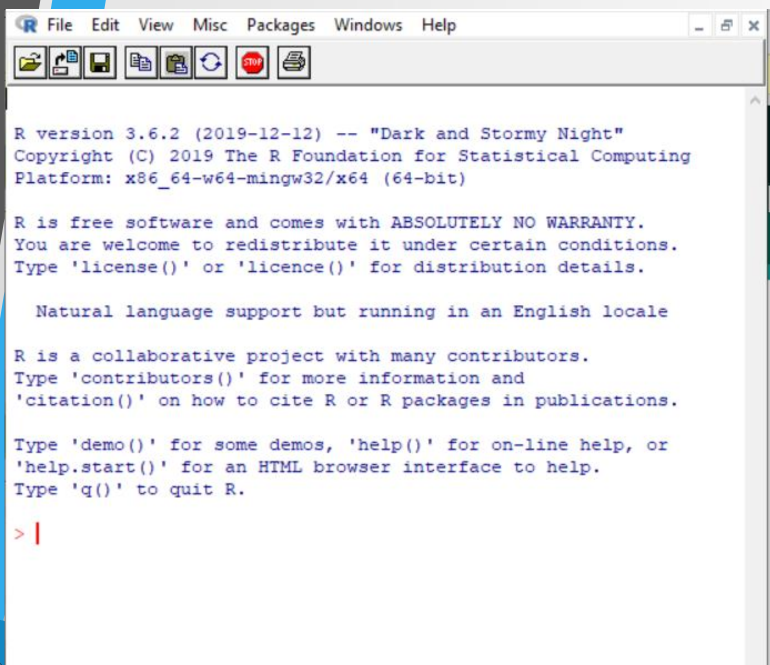
Data Mining Tools

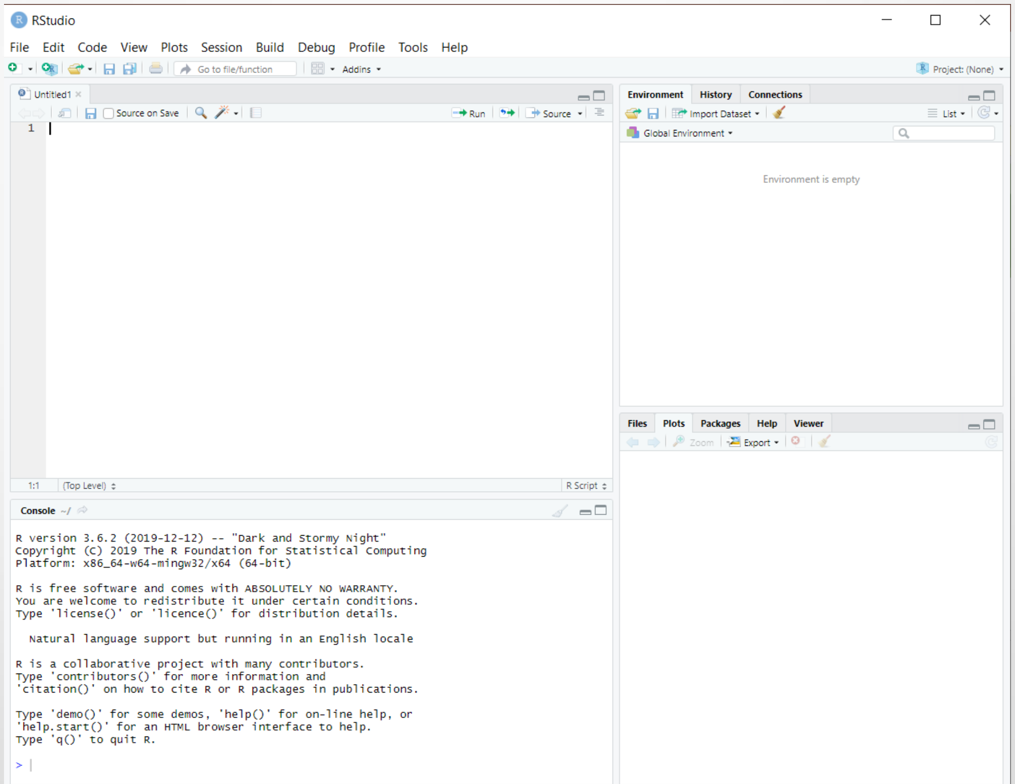


https://www.r-project.org/

https://rstudio.com/products/rstudio/#rstudio-desktop/

R and R studio





R – Lecture 1

R – Part one Exercise
15 min

Break
10 min

Outline

0. Assignment 1

1. Data Preparation

- Data validation
- Data transformation
- Data reduction

2. Data Exploration

– Univariate analysis
– Bivariate analysis
– Multivariate analysis

3. Regression Analysis

– Simple linear regression
– Multiple linear regression
– Validation of regression models
– Selection of predictive variables

Data Preparation

- Data extracted from the primary sources may have several anomalies.
- Business intelligence systems and mathematical models for decision making can achieve accurate and effective results only when the input data are highly reliable.
- The purpose of data preparation is to create a highly quality dataset for subsequent use for business intelligence and data mining analysis.

Data Preparation

- Data Preparation Techniques:.
  - Data validation
    - Identify and remove anomalies and inconsistencies
  - Data transformation and integration
    - To improve the accuracy and efficiency of learning algorithms
  - Data size reduction and discretization
    - To obtain a dataset with a lower number of attributes and records (as informative as the original data set)

Data Validation

- Unsatisfactory data quality due to
  - Incompleteness
  - Noise
  - Inconsistency
- The purpose of data validation techniques is to identify and implement corrective actions in case of incomplete and inconsistence data or data affected by noise.

Data Validation

- How to handle noisy data:
  - Identify all outliers in a data set.
  - Eliminate or Replace abnormal values by a suitable value obtained through identification or substitution.

Data Validation

- To partially correct incomplete data:
  - Elimination
    - Problem: substantial loss of information
  - Inspection (by an expert)
    - Problem: time-consuming for large dataset
  - Identification (assuming a different value for missing data e.g. -1)
    - Problem: no correction
  - Substitution (with mean, maximum likelihood value, etc.)
    - Problem: complex and time-consuming

Data Transformation

- In most data mining analyses it is appropriate to apply a few transformations to the data set in order to improve the accuracy of the learning models developed. Examples are
  - Outlier correction
  - Normalization
  - New Variables

Data Transformation

- The most popular normalization techniques are:
  - Decimal scaling
  - Min-max method
  - Z-index

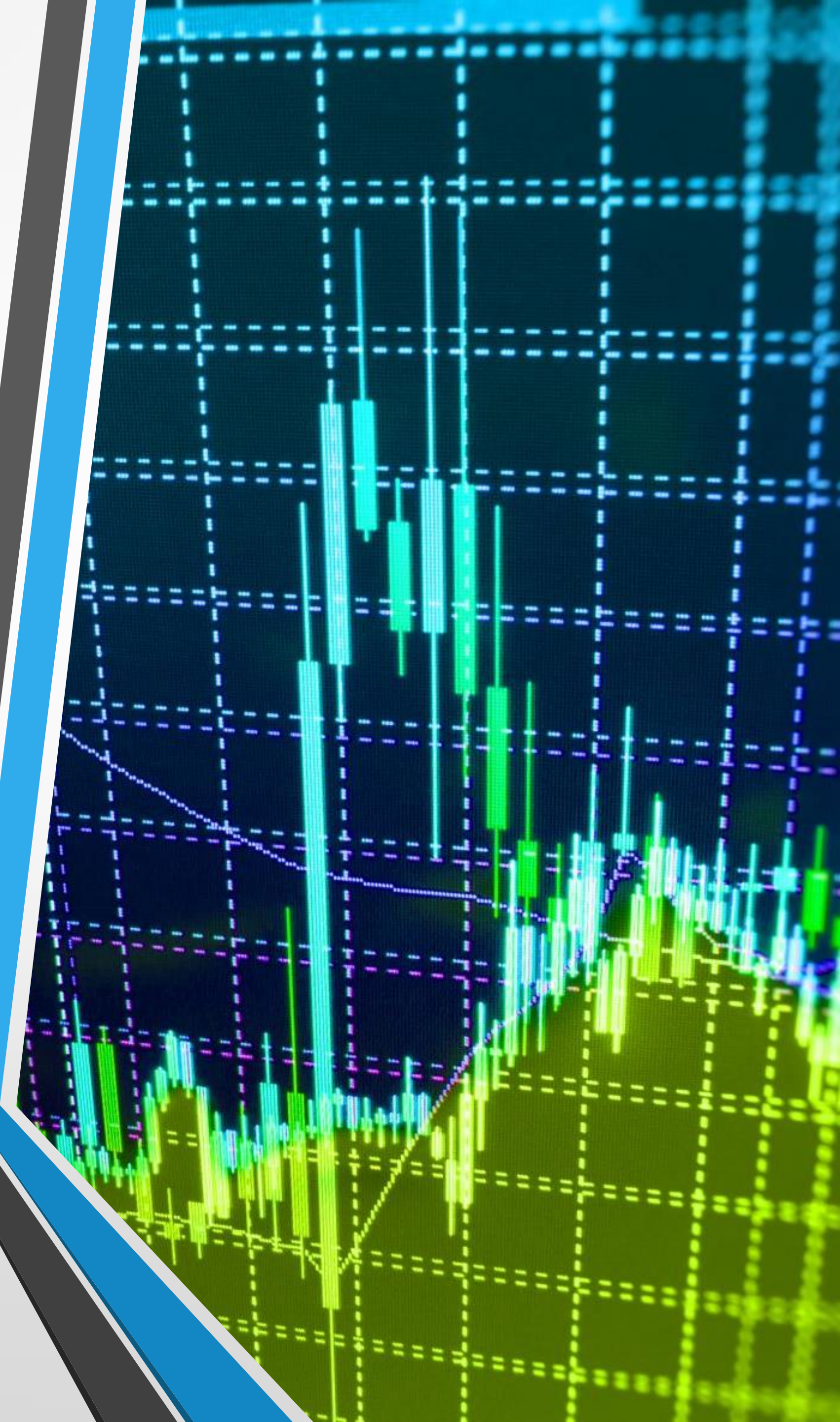
Data Transformation

- Data Scaling Method:

$$X_{New} = \frac{X_{Old}}{H}$$

X_{New} is the normalized data and H is a given parameter that determines the scaling factor.

In general H , is fixed at a value that gives transformed values in the range [-1 1].



Data Transformation

- Min-max Method

$$X_{New} = MIN_{NEW} + \frac{X_{Old} - MIN}{MAX - MIN}(MAX_{NEW} - MIN_{NEW})$$

MIN and MAX are the minimum and maximum values before the transformation and MIN_{NEW} and MAX_{NEW} are the minimum and maximum values that we wish to obtain.

For example, $MAX_{NEW} = 1$ and $MIN_{NEW} = 0$ results in

$$X_{New} = \frac{X_{Old} - MIN}{MAX - MIN}$$

Data Transformation

- Z-index Method

$$X_{New} = \frac{X_{Old} - \mu}{\sigma}$$

μ is the mean of data and σ is the standard deviation of data.

Mean = ?

Standard deviation = ?

- If the data has a normal distribution, z-index generates values that are almost certainly within the range (-3,3)

Data Reduction

- When dealing with a large dataset, it is often necessary to reduce its size, in order to make learning algorithms more efficient, without sacrificing the quality of the results. There are 3 main criteria to determine whether a data reduction technology should be used:
  - Efficiency
  - Accuracy
  - Simplicity

Data Reduction

- Some popular data reduction techniques are
  - Sampling
  - Feature selection
  - Principal component analysis
  - Data discretization

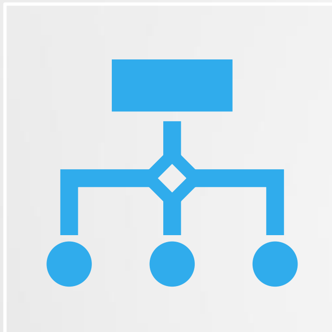
Data Reduction

- Sampling:
  - Data reduction can be achieved by extracting a sample of observations that is significant from a statistical standpoint.
  - Care must be taken to preserve in the sample an enough percentage of the original dataset with respect to a categorical attribute.
  - Generally, a sample comprising a few thousand observations is adequate to train most learning models.
  - It is also useful to set up several independent samples, each of a predetermined/fixed size.
  - Computation time increases linearly with the number of samples determined.

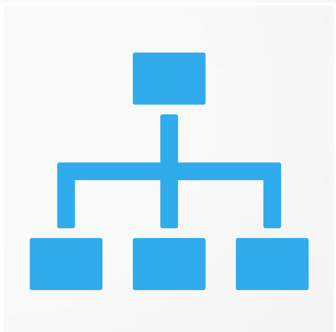
Data Reduction

- Feature Selection (or Feature Reduction)
  - The purpose is to eliminate a subset of variables which are not relevant for data mining.
  - Feature Selection Methods:
    - Filter Method: the attributes deemed most significant are selected for learning, while the rest are excluded.
    - Wrapper Method: Each time, a different set of attributes will be used for model training. A search engine is used to identify the best possible combination of attributes that guarantees high accuracy
    - Embedded Method: the attribute selection process lies inside the learning algorithm (classification tree).

Data Reduction

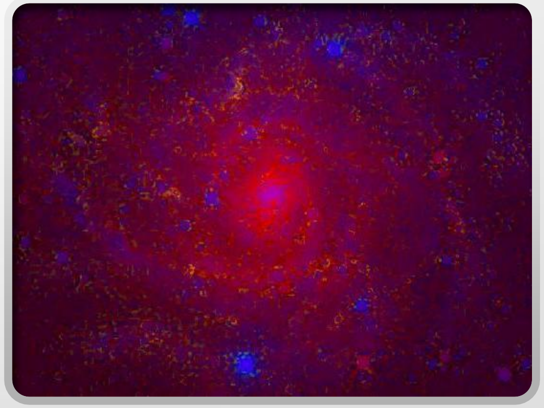


Principal Component Analysis (PCA) is a mathematical procedure that uses a transformation to convert a set of data into sub-sets (new data) called **principal components**.



Each sub-set has lower number of attributes.

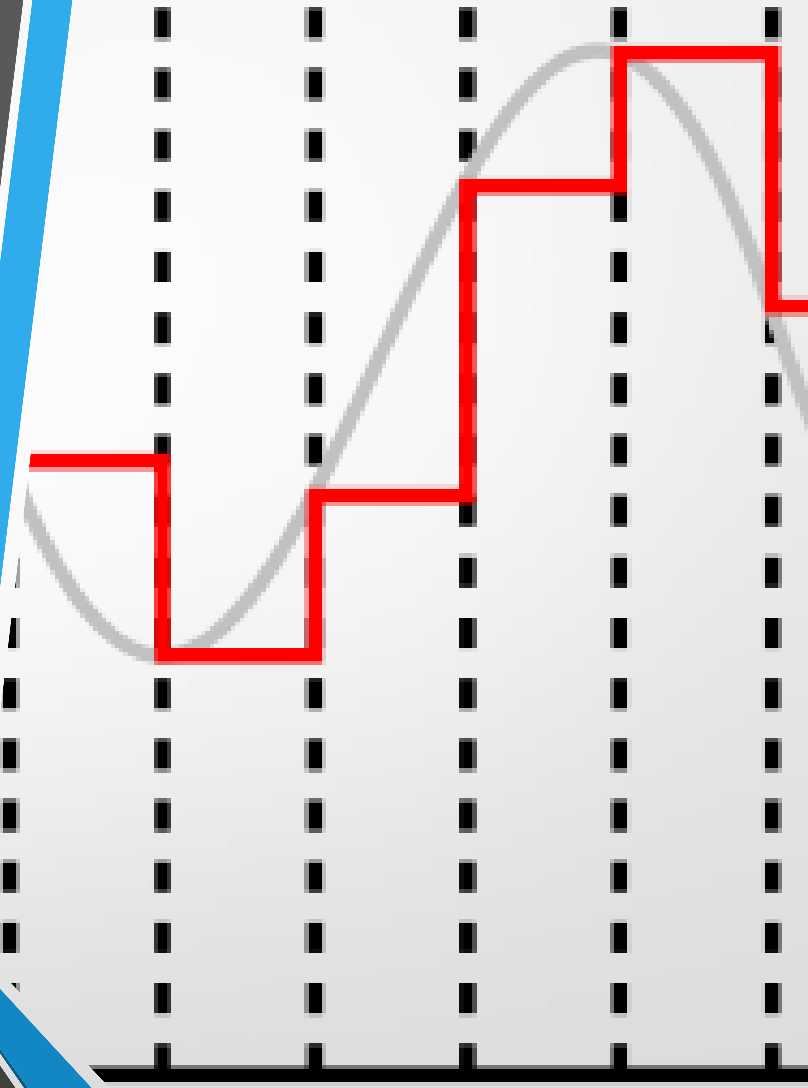
Data Reduction



- PCA Applications:

Data Reduction

- Data Discretisation
  - to decrease the number of distinct values of attributes.
  - reduce the problem complexity
  - improve generalization capability



Data Reduction

- Example for data discretization
  - The weekly spending of a mobile phone customer is a numerical attribute. The attribute can be discretized into several classes
    - Low [0,10)
    - Medium low [10,20)
    - Medium [20,30)
    - Medium high [30, 40)
    - High [40, ∞)

R – Lecture 2

R – Part two Exercise
15 min

Break

15 min

start 3:15 pm

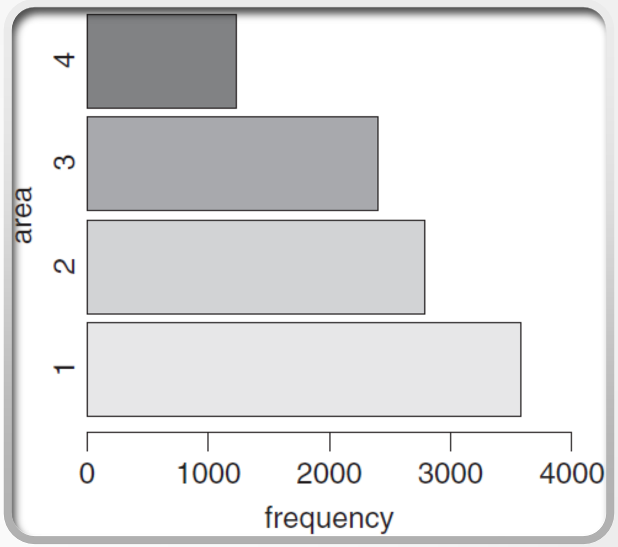
Data Exploration

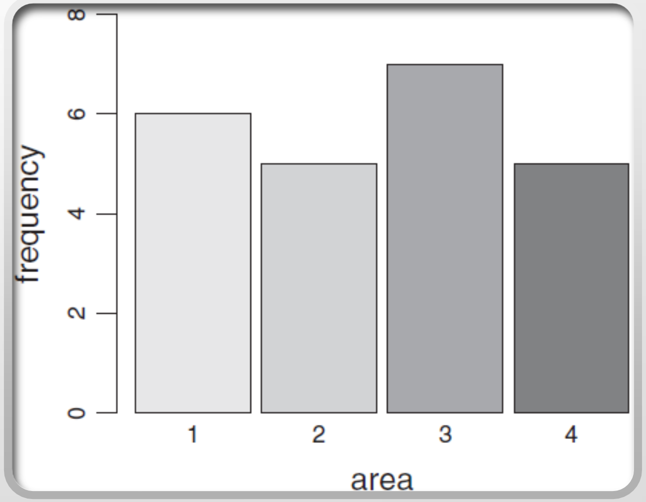
Data Exploration

- The primary purpose of data exploration is to highlight the relevant features of attributes using
  1) Graphical methods
  2) Calculating summary statistics

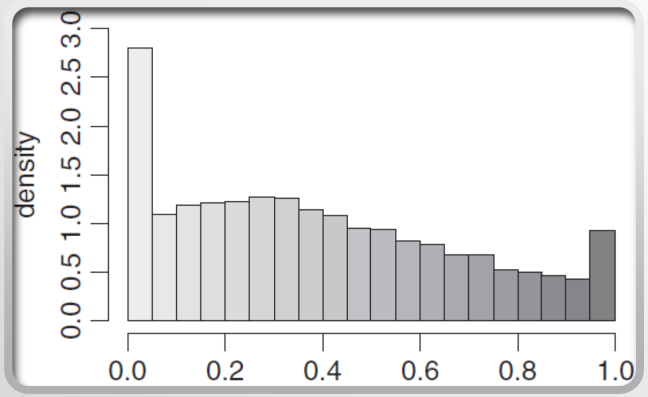
Data Exploration

- Graphical Analysis of Categorical Attributes (**Bar Chart**)
- An attribute is categorical if it assumes a finite set of different values that general mathematical calculations does not apply.





Data Exploration



- Graphical analysis of numerical attributes (**Histogram**)
- A numerical attribute take its value from a continuous range of values.

Data Exploration

- Measures of central tendency (location)
  - Mean
  - Median
  - Mode
  - Midrange
  - Geometric Mean

Data Exploration

- The best-known measure of location used to describe a numerical attribute is certainly the mean.

$$\mu = \frac{x_1 + x_2 + \cdots + x_m}{m} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

Data Exploration

- The **median** of m observation can be defined as the central value assuming that the observations have been ordered in non-decreasing way.

If m is an odd number, the median is the observation occupying the position $(m+1)/2$:

$$x^{\text{med}} = x_{(m+1)/2}.$$

If m is an even number, the median is the middle point in the interval between the observations of position $m/2$ and $(m+2)/2$:

$$x^{\text{med}} = \frac{x_{m/2} + x_{(m+2)/2}}{2}.$$

Data Exploration

- Mode
  - The value that corresponds to the peak of the empirical density curve of an attribute.
- Midrange
  - Midpoint in the interval between the minimum and maximum values

$$MIDR = \frac{MIN + MAX}{2}$$

Data Exploration

- Geometric Mean
  - m-th root of the product of the m observations:

$$\mu = \sqrt[m]{x_1 x_2 \ldots x_m}$$

Data Exploration

- Measures of central tendency (location):
  - Range

$$RANGE = MAX - MIN$$

  - Mean Absolute Deviation

$$MAD = \frac{1}{m}\sum_{i=1}^{m}|x_i - \mu|$$

  - Variance

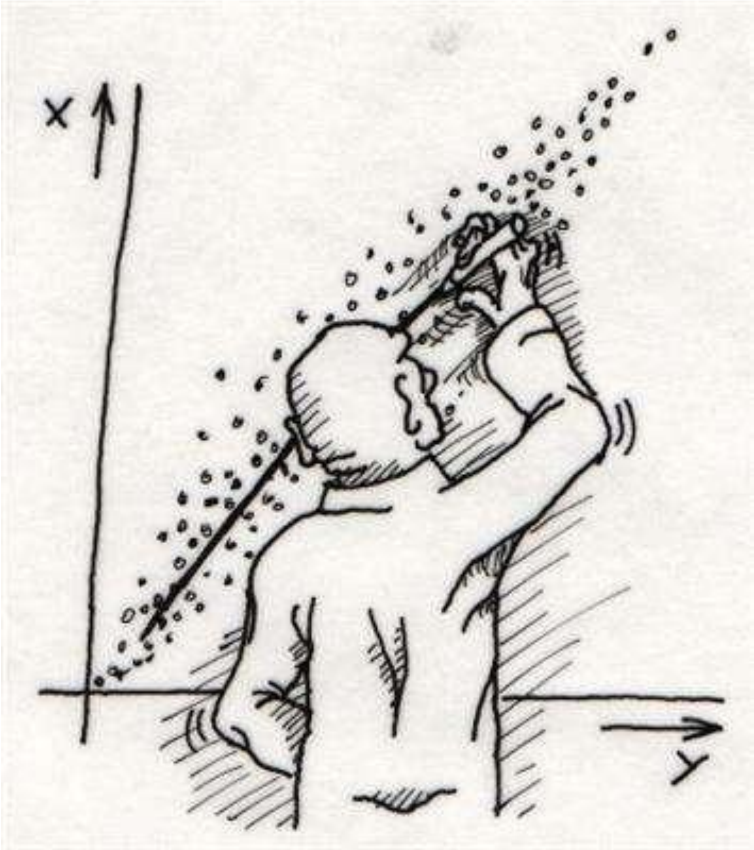
$$\sigma^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu)^2$$

R – Lecture 3

R – Part three Exercise
15 min

Break

10 min



Regression Analysis

Regression Analysis

- Identify the functional relationship between the target attribute (numerical) and a subset of the remaining attributes contained in the dataset.
- Predict the future value of the target attribute.

$$Y = f(X_1, X_2, \ldots, X_n)$$

- Suppose that a dataset contains m observations and n+1 attributes including n explanatory attributes and 1 target attribute

- Function f is often called the hypothesis.

Regression Analysis

- Forms of hypothesis
  - Linear, quadratic, logarithmic, exponential
  - Note: most types of nonlinear relationships may be reduced to the linear case by means of appropriate preliminary transformations to the original observations.
- Example:
  - Suppose that $Y = b + wX + dX^2$
  - Transformation: $Z = X^2$
  - Resulting relation:

$$Y = b + wX + dZ$$

Regression Analysis

- Linear regression (general form)

$$Y = w_1 X_1 + w_2 X_2 + \cdots + w_n X_n + b = \sum_{j=1}^{n} w_j X_j + b$$

- Linear regression (simple form)

$$Y = wX + b$$

- The probabilistic model

$$Y = wX + b + \varepsilon$$

Regression Analysis

- Define residue

$$e_i = y_i - f(x_i) = y_i - wx_i - b, \quad i \in \mathcal{M}$$

- Sum of squared errors

$$\mathrm{SSE} = \sum_{i=1}^{m} e_i^2 = \sum_{i=1}^{m} [y_i - f(x_i)]^2 = \sum_{i=1}^{m} [y_i - wx_i - b]^2$$

- Objective (minimum MSE, Mean Squared Value):
  - Find suitable values of *w* and *b* such that SSE can be reduced to its global minimum.

$$\frac{\partial\, \mathrm{SSE}}{\partial b} = -2 \sum_{i=1}^{m} [y_i - wx_i - b] = 0,$$

$$\frac{\partial\, \mathrm{SSE}}{\partial w} = -2 \sum_{i=1}^{m} x_i [y_i - wx_i - b] = 0$$

Regression Analysis

- Solve the equations

$$\begin{pmatrix} m & \sum_{i=1}^m x_i \\ \sum_{i=1}^m x_i & \sum_{i=1}^m x_i^2 \end{pmatrix} \begin{pmatrix} b \\ w \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^m y_i \\ \sum_{i=1}^m x_i y_i \end{pmatrix}$$

- Results:

$$\boxed{\begin{aligned} \hat{w} &= \frac{\sigma_{xy}}{\sigma_{xx}}, \\ \hat{b} &= \bar{\mu}_y - \hat{w}\bar{\mu}_x \end{aligned}}$$

  - where

$$\sigma_{xy} = \sum_{i=1}^{m} (x_i - \bar{\mu}_x)(y_i - \bar{\mu}_y),$$

$$\sigma_{xx} = \sum_{i=1}^{m} (x_i - \bar{\mu}_x)^2,$$

$$\bar{\mu}_x = \frac{\sum_{i=1}^m x_i}{m}, \qquad \bar{\mu}_y = \frac{\sum_{i=1}^m y_i}{m}$$

Iris data set

Calculating Descriptive Statistics Using

Quantitative Bivariate Analysis

- Reading data set
- Creating scatter plots
- Calculating correlation
- Calculating linear model
- Plotting linear model
- Calculating the accuary

```
iris <- read.csv("data set 4.csv")
```

```
plot(x=iris$Petal.Length,y=iris$Petal.Width)
```

```
cor(iris$Petal.Length,iris$Petal.Width)
```

```
model<-
lm(iris$Petal.Width~iris$Petal.Length)
```

```
lines(iris$Petal.Length,model$fitted.values)
error = Model$fitted.values - iris$Petal.Width
MSE  = mean(error^2)
```

Q&A