COMP 333 Data Analytics

Exploratory Data Analysis

Greg Butler

Data Science Research Centre

and

Centre for Structural and Functional Genomics

and

Computer Science and Software Engineering
Concordia University, Montreal, Canada

gregb@cs.concordia.ca

Models

Definition

A *model* is a representation of some part of the real world.

Simplify Real World by Abstraction

A model is very simple compared to the whole real world.
Omit details that are not useful.
Keep important information.

Model for a Purpose

Model is ...
... Simple enough to understand and construct.
... But not so simple that it is not useful
for whatever purpose you need the model.

General Tasks for Models

- regression
- classification
- prediction
- simulation

Regression

Regression analysis is a form of predictive modelling technique which investigates the relationship between

a dependent (target) and independent variable (s) (predictor).

Classification

Classification is the process of predicting

the class of given data points.

Classes are sometimes called as targets/ labels or categories.

Prediction

Predictive modeling is a process that uses data and statistics to predict outcomes with data models.

The predictions are *numbers* such as scores, probabilities, amounts.

These models can be used to predict anything
from sports outcomes and TV ratings
to technological advances and corporate earnings.

Simulation

Simulation refers to the representation of a system or process that is defined by known relationships.

Simulation, allows us to build a mathematical model of the world and run it several times on a computer.

It allows us to evaluate various decisions and choose between them.

The *importance of simulation* is that it allows parameters to be changed in the models to understand *cause and effect* at a level which is not possible in other ways.

It also permits phenomena to be studied which might be too expensive or dangerous for conventional experimental methods.

Regression: Curve Fitting

Regression Analysis

a set of statistical processes for estimating the relationships among variables

helps one understand how the typical value of the dependent variable changes
when any one of the independent variables is varied,s
while the other independent variables are held fixed.

Linear Regression

fit a line to (x,y) data
y is dependent variable, x is independent variable

Curve Fitting

Can fit other forms of curves to data

Regression: Curve Fitting

Anscombe's Quartet