

# Advanced Data & Network Mining

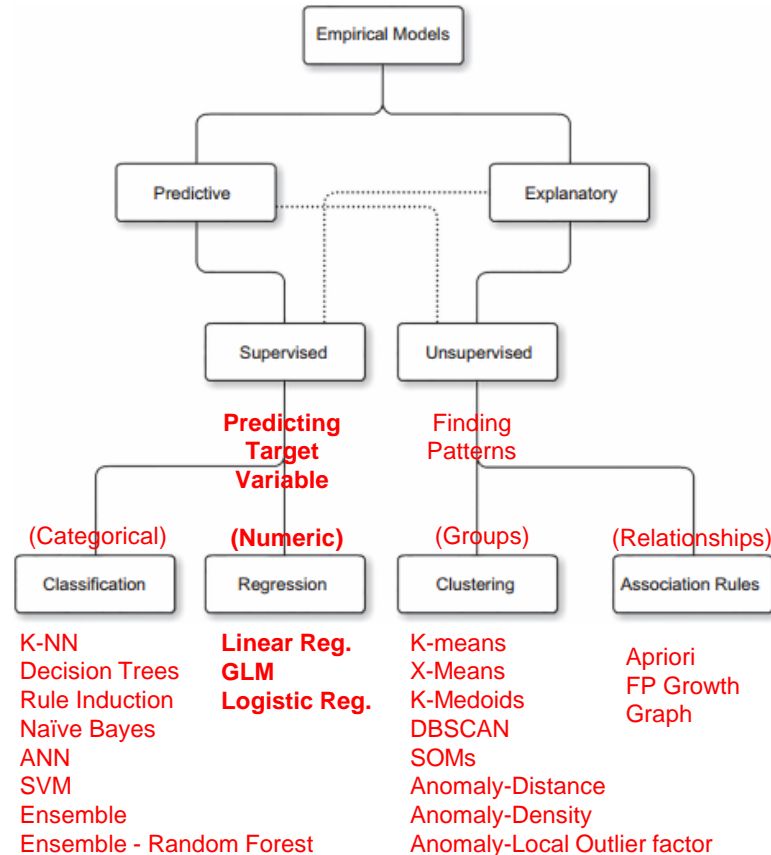
## Modelling - Supervised Learning Regression

*2023-24*

*terri.hoare@dbs.ie*

# Data Mining

## Taxonomy of Algorithms



# Data Mining Regression

Predicts the value (or class) of a dependent variable by combining the predictor variables into a function  $y = f(X)$ .

One of the three most common analytics techniques (along with decision trees and clustering) used by practitioners.

- **Linear Regression**

Classical predictive model that expresses the relationship between inputs and an output parameter in the form of an equation. Pioneered by Sir Francis Galton (1888) who observed a strong tendency for tall parents to have children slightly shorter than themselves and short parents to have children slightly taller than themselves. Even if the parents' heights were at the tail ends of a bell curve (normal distribution), all samples of the children's heights “**regressed**” towards the mean

- **Logistic Regression**

Technically this is a classification method closer in application to decision trees or Bayesian methods, however structurally similar to linear regression in its **function fitting** methodology

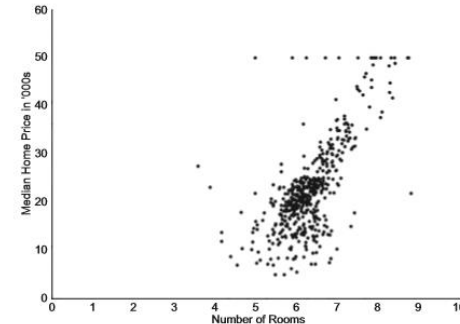
# Regression Models

## Linear Regression : How it Works : Home Prices

Boston Housing Data Set (05\_Regression\_5.1\_bos\_housing)

Based on an Urban Study in 1978. Data set contains physical attributes number rooms, age, tax, location and neighbourhood attributes schools, industries, crime, zoning. Median price in thousands of dollars.

1. CRIM per capita crime rate by town
2. ZN proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS proportion of nonretail business acres per town
4. CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. NOX nitric oxides concentration (parts per 10 million)
6. RM average number of rooms per dwelling
7. AGE proportion of owner-occupied units built prior to 1940
8. DIS weighted distances to five Boston employment centers
9. RAD index of accessibility to radial highways
10. TAX full-value property-tax rate per \$10,000
11. PTRATIO pupil-teacher ratio by town
12.  $B_1000(B_k - 0.63)^2$  where  $B_k$  is the proportion of blacks by town
13. LSTAT % lower status of the population
14. MEDV Median value of owner-occupied homes in \$1000's

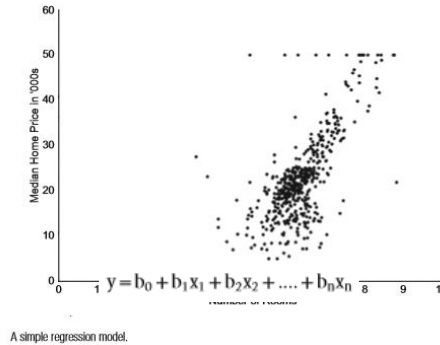


A simple regression model.

# Regression Models

## Linear Regression : How it Works : Home Prices cont.

Boston Housing Data Set (05\_Regression\_5.1\_bos\_housing.csv)



What is the effect of number of rooms in a house on its median sale price?

(If we have two predictors the problem is to find a surface and for more than two predictors the problem can be expressed as finding the function)

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n + \epsilon$$

## Data Mining

### Linear Regression : How it Works : Home Prices cont.

Linear Regression uses the method of least squares error to fit a function. In the case of a single predictor, suppose we look to fit a straight line through the data to minimise error. If Predicted value  $\hat{y} = b_0 + b_1x$ , then  $e = y - \hat{y} = y - (b_0 + b_1x)$  defines the error at a single location  $(x,y)$  in the data set. By squaring the difference to eliminate the sign bias, the average error for a given fit is given by

$$\sum e^2 = 1/n * (y_i - \hat{y}_i)^2 = 1/n \sum (y_i - b_0 - b_1x_i)^2$$

Calculus is applied to solve for the values of  $b_0$  and  $b_1$  that define the straight line that minimises the error. This method is generalisable to multiple predictors to solve for the values of  $b_0, b_1, \dots$  that will minimise the error using **Multiple Linear Regression**. Practical linear regression algorithms use an optimization technique known as gradient descent

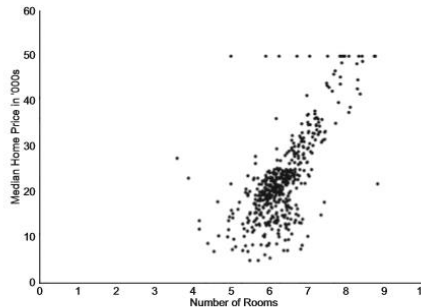
# Data Mining

## Linear Regression : How it Works : Home Prices cont.

For the Boston house price data set one predictor number of rooms

**Median Price =  $9.1 * \text{Number of Rooms} - 34.7$**  is the best fit

Where **Median price = 50**, the median price appears to independent of number of rooms. This could be that there are other factors influencing price indicating use of the **Multiple Linear Regression method**



A simple regression model.



# Data Mining

## Linear Regression : How to Implement

- **Objective**

1. Identify which of several attributes are required to accurately predict the median price of a house
2. Build a multiple linear regression model to predict the median price using the most important attributes


- **Learning Map**

1. Building a linear regression model
2. Measuring the performance of the model
3. Understanding the commonly used options for the Linear Regression operator
4. Applying the model to predict MEDV (Median Price) prices for unseen data

# Data Mining

## Linear Regression : How to Implement cont.

- **Step 2 – Model Building (Validation ports “mod” and “ave” to output ports)**




Data

Description

Annotation

**LinearRegression**

- 0.119 \* CRIM  
+ 0.050 \* ZN  
+ 0.018 \* INDUS  
+ 2.433 \* CHAS  
- 16.919 \* NOX  
+ 3.670 \* RM  
- 0.002 \* AGE  
- 1.575 \* DIS  
+ 0.296 \* RAD  
- 0.012 \* TAX  
- 0.920 \* PTRATIO  
+ 0.009 \* B  
- 0.554 \* LSTAT  
+ 37.281



Data

Description

Annotation

Attribute	Coefficient	Std. Error	Std. Coeffi...	Tolerance	t-Stat	p-Value	Code
ZN	0.050	0.014	0.128	0.876	3.494	0.001	****
NOX	-16.919	3.973	-0.215	0.804	-4.258	0.000	****
RM	3.670	0.393	0.277	0.574	9.348	0	****
DIS	-1.575	0.212	-0.362	0.839	-7.428	0	****
RAD	0.296	0.069	0.281	0.772	4.301	0.000	****
PTRATIO	-0.920	0.136	-0.214	0.789	-6.775	0	****
LSTAT	-0.554	0.053	-0.431	0.464	-10.520	0	****
(Intercept)	37.281	4.880	?	?	7.639	0	****
CRIM	-0.119	0.038	-0.103	0.843	-3.115	0.002	***
CHAS	2.433	0.894	0.068	0.991	2.721	0.007	***
TAX	-0.012	0.004	-0.216	0.744	-3.097	0.002	***
B	0.009	0.003	0.084	0.905	3.061	0.003	***
INDUS	0.018	0.064	0.014	0.683	0.288	0.776	
AGE	-0.002	0.014	-0.007	0.803	-0.164	0.871	

05\_Regression\_5.1\_LinReg\_bos\_housing.rmp

# Data Mining

## Linear Regression : How to Implement cont.

- **Step 2 – Model Building** (changing feature selection to “**greedy**” removes least significant features INDUS and AGE)

**LinearRegression**

- 0.119 \* CRIM  
+ 0.049 \* ZN  
+ 2.444 \* CHAS  
- 16.784 \* NOX  
+ 3.646 \* RM  
- 1.579 \* DIS  
+ 0.292 \* RAD  
- 0.011 \* TAX  
- 0.916 \* PTRATIO  
+ 0.009 \* B  
- 0.556 \* LSTAT  
+ 37.258

using 'greedy' feature selection

Attribute	Coefficient	Std. Error	Std. Coeffi.	Tolerance	t-Stat	p-Value	Code
RM	3.646	0.378	0.046	0.581	9.633	0	****
LSTAT	-0.556	0.049	-0.129	0.490	-11.408	0	****
DIS	-1.579	0.197	-0.119	0.823	-8.029	0	****
(Intercept)	37.258	4.785	?	?	7.787	0	****
PTRATIO	-0.916	0.132	-0.868	0.793	-6.935	0	****
NOX	-16.784	3.706	-0.469	0.812	-4.529	0.000	****
RAD	0.292	0.066	0.894	0.769	4.431	0.000	****
ZN	0.049	0.014	0.128	0.877	3.509	0.001	****
TAX	-0.011	0.003	-0.003	0.749	-3.329	0.001	***
CRIM	-0.119	0.038	-0.103	0.843	-3.128	0.002	***
B	0.009	0.003	0.161	0.905	3.059	0.003	***
CHAS	2.444	0.891	1.836	0.991	2.743	0.007	***

# Data Mining

## Linear Regression : How to Implement cont.

- **Step 2 – Model Building (squared correlation and goodness of fit)**

Takes values  $[0,1]$  with values closer to 1 indicating a good model



05\_Regression\_5.1\_LinReg\_bos\_housing.rmp

## Data Mining

### Linear Regression : How to Implement cont.

- **Step 3 – Execution and Interpretation**
  - Note the useful **Data View Tab** which shows the coefficients of the linear regression function together with coefficient significance. Double-click on “Code” to sort descending on significance. RapidMiner assigns (\*\*\*\*) to highly significant factors
  - Observing the low significance of some attributes, re-run Step 2 with parameter **feature selection** set to “**greedy**” in order to remove the least significant factors INDUS and AGE
  - A handy check of goodness of fit is the **squared correlation** taking values [0,1] with values closer to 1 indicating a better model. Both runs of the MLR model with different values for **feature selection** gave a value of 0.67 with different values for **squared error output**

## Data Mining

### Linear Regression : How to Implement cont.

- **Step 3 – Execution and Interpretation**
  - The **t-stat** and **p-values** are the result of hypothesis tests on the coefficients. A higher **t-stat** signals that the **NULL Hypothesis**, which assumes the coefficient is zero, can safely be rejected. The corresponding **p-value** indicates the probability of wrongly rejecting the null hypothesis.
  - The linear regression model using **Size of Rooms** had an  $R^2$  (**squared correlation**) of 0.405 and squared error of 45. The MLR model is a better model for predicting Median Price with an  $R^2$  of 0.676 and squared error of 25. This affirms the decision to use multiple factors

## Data Mining

### Linear Regression : How to Implement cont.

- **Step 4 – Application to Unseen Data**
  - Once the model has been built, it can be tested against the unseen data created using **Filter Examples** in Step (1). When we apply the model we will be able to compare the predicted MEDV with the actual MEDV to test how well the model would perform on new data. The difference between predicted MEDV and actual MEDV is termed “residuals”. **Generate Attributes** is used to create and calculate the value for “residuals”

05\_Regression\_5.1\_LinReg\_bos\_housing.rmp

# Data Mining

## Linear Regression : How to Implement cont.

- **Step 4 – Application to Unseen Data**

- Statistics for “residuals” indicate that the mean is close to 0 (-0.2) but that the standard deviation (and hence variance) at 4.35 is not small. Also the histogram suggests that the residuals are not quite normally distributed which would be a motivation to continue to improve the model



05\_Regression\_5.1\_LinReg\_bos\_housing.rmp



## Data Mining

### Linear Regression : How to Implement cont.

- **Regression Model Checkpoints**

1. Quantify  $R^2$  [0,1] to effectively explain how much variability in the dependent variable is explained by the independent variables. Generally very low values ( $<0.2$ ) indicate that the variables in your model do not explain the outcome satisfactorily
2. Ensure that all error terms in the model are normally distributed. (In RapidMiner, use **Generate Attributes** to generate an error attribute that is the difference between the predicted and actual median prices in the Test data set)
3. Highly **non-linear** relations may fail the above checks and need more advanced analytical techniques applied. Also remember correlation is not causation!

## Regression Models

### Linear Regression : Discussion Points

- Curse of dimensionality. Too many attributes limit ability to obtain a good model as well as increasing computational and interpretational complexity
- Feature selection is required in order to reduce the number of features / factors to a minimum and still obtain a good model
- Remember the checkpoints.  $R^2$  or the amount of variability in the dependent model that is explained by the independent variables, ensuring error terms are normally distributed, remembering correlation does not imply causation, and that if linear regression is not suitable, there may still be highly non-linear relationships present which will need other advanced techniques

# Regression Models

## Linear Regression : Summary

- **Model**

The model consists of coefficients for each input predictor and their statistical significance. A bias (intercept may be optional)

- **Input**

All attributes should be numeric

- **Output**

The label may be numeric or binomial

- **Pros**

The workhorse of most predictive modelling techniques. Easy to use and explain to non technical business users

- **Cons**

Cannot handle missing data. Categorical data are not directly usable, but require transformation into numeric

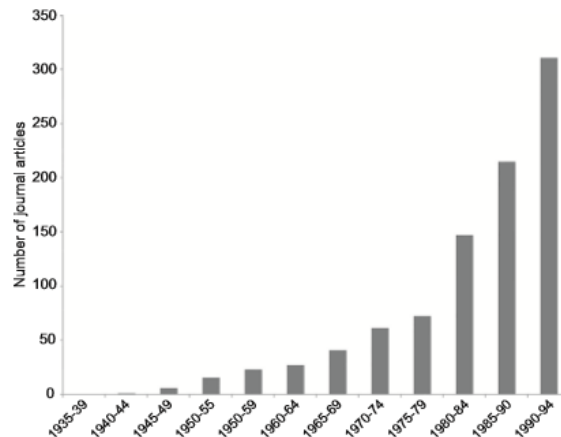
- **Use Cases**

Any scenario that requires predicting a continuous numeric variable

# Regression Models

## Logistic Regression

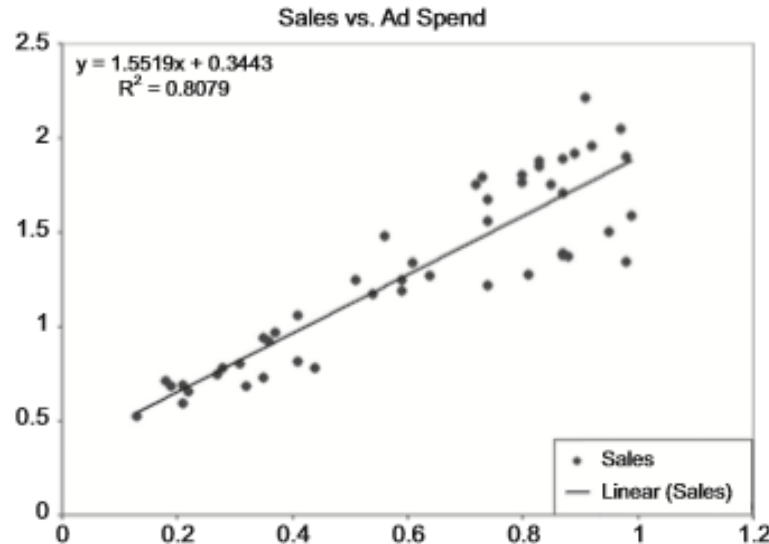
Logistic Regression arose in the mid-twentieth century as a result of the simultaneous development of the concept of the **logit** in the field of biometrics and the advent of the digital computer which made computations of such terms easy. Logistic Regression has become increasingly important in a variety of scientific and business applications over the last few decades.



# Regression Models

## Logistic Regression : Concept of Logit

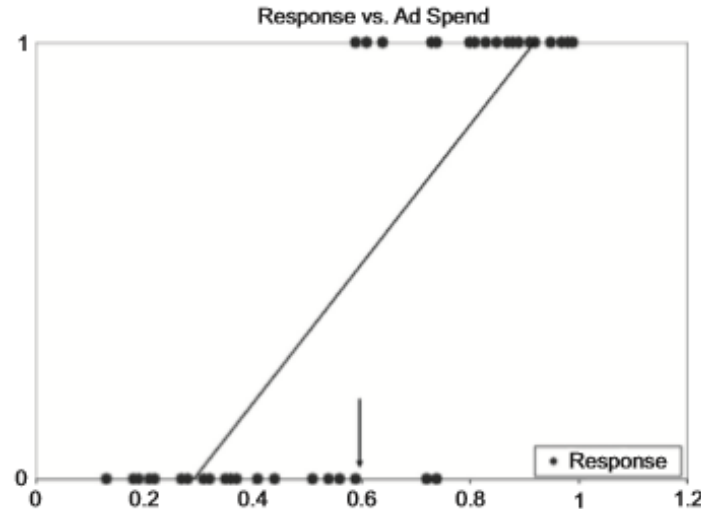
Recall Linear Regression is the process of finding a function to fit the x's that vary linearly with y. Applying a Linear Regression Model. E.g. we can make an intuitive assessment that increase in Ad Spend also increases Sales.



# Regression Models

## Logistic Regression : Concept of Logit cont.

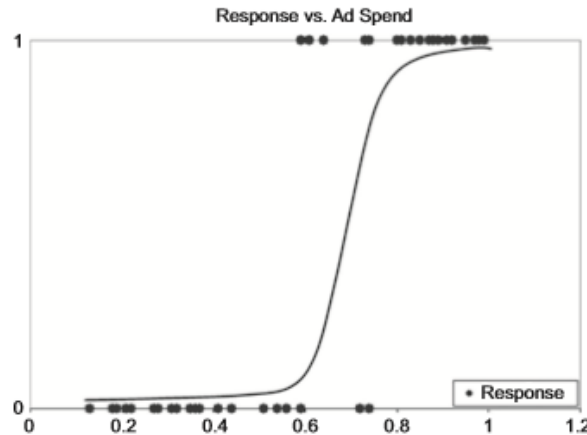
Applying a Linear Fit for a Binary (e.g. yes/no) outcome. Although we can make an intuitive assessment that increase in Ad Spend increases Response, the switch is abrupt – around 0.6. Using the straight line, we cannot really predict the outcome.



# Regression Models

## Logistic Regression : Concept of Logit cont.

Applying a Logistic Regression Model. The S-shaped curve is clearly a better fit for *most* of the data. We can state Ad Spend increases Sales **and** we may also be able to predict using this model



Logistic Regression is the process of obtaining an appropriate nonlinear curve to fit the data when the target variable is discrete. How is the sigmoid curve obtained? How does it relate to the predictors?

## Regression Models

### Logistic Regression : How it Works

- Logistic Regression is a mathematical modelling approach in which a best-fitting, yet least restrictive model is selected to describe the relationship between several independent explanatory variables and a dependent binomial response variable
- If we transform the target variable  $y$  to the **logarithm of the odds of  $y$** , then the **transformed** target variable is **linearly** related to the predictors  $X$ 
  - If  $y$  is an event (response, pass/fail, etc.),
  - and  $p$  is the probability of the event happening ( $y=1$ ),
  - then  $(1-p)$  is the probability of the event *not* happening ( $y=0$ ),
  - And  $p/(1-p)$  are the *odds* of the event happening



## Regression Models

### Logistic Regression : How it Works

The logarithm of the odds,  $\log(p/(1-p))$  is linear in predictors  $X$

$$\text{logit} = \log p/(1-p) = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n \text{ and } p = e^{\text{logit}} / (1 + e^{\text{logit}})$$

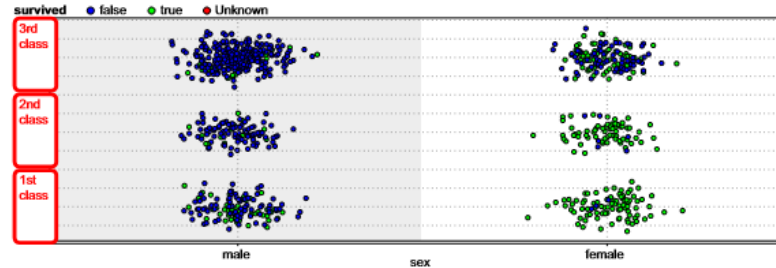
- The coefficients  $b$  are determined from a trial set of  $b$  such that the quantity  $p^y(1-p)^{(1-y)}$  is maximised over a summation across all training examples. This quantity, a simplified form of a **likelihood** function is **maximised for good estimates** and **minimised for poor estimates**. In practice more sophisticated formulations of likelihood are used as well as optimisation techniques such as gradient search to search for coefficients  $b$  with the objective of maximising the likelihood of correct estimation

# Regression Models

## Logistic Regression : How it Works : An Example

<http://www.kaggle.com/c/titanic-gettingStarted>

In the 1912 shipwreck of the Titanic, 75% of the women and 63% of first class passengers survived. If a passenger was a woman and if she travelled first class, her probability of survival was 97%!



pclass	sex	survived?
3.0	male	0.0
1.0	female	1.0
3.0	female	1.0
1.0	female	1.0
3.0	male	0.0
3.0	male	0.0

## Regression Models

### Logistic Regression : How it Works : An Example cont.

<http://www.kaggle.com/c/titanic-gettingStarted>

A data mining competition challenged analysts to develop an algorithm that could classify the passenger list (891 samples) into survivors and non-survivors.

0-male; 1-female. Fitting a model to the data yielded the following equation for predicting the class “survived = “false” :-

$$\text{logit} = -0.6503 - 2.6417 * \text{sex} + 0.9595 * \text{pclass}$$

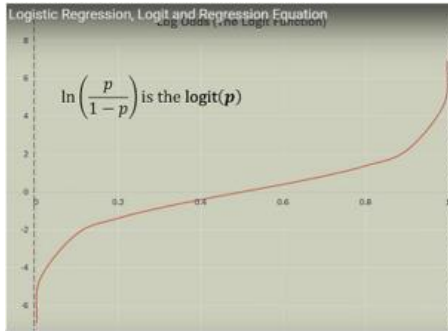
# Regression Models

## Logistic Regression : How it Works : An Example cont.

<http://www.kaggle.com/c/titanic-gettingStarted>

### **Interpreting the coefficients**

Recall  $p = e^{\text{logit}} / (1 + e^{\text{logit}})$ . As  $\text{logit} \rightarrow -\infty$   $p \rightarrow 0$  and as  $\text{logit} \rightarrow +\infty$   $p \rightarrow 1$



The negative coefficient on variable “sex” indicates that this probability reduces for females and the positive coefficient on variable “pclass” indicates that the probability of not surviving increases the higher the number of the travel class

## Regression Models

### Logistic Regression : How it Works : An Example cont.

<http://www.kaggle.com/c/titanic-gettingStarted>

#### **Interpreting the odds form of the Logistic Regression Model**

$$\text{logit} = -0.6503 - 2.6417 \cdot \text{sex} + 0.9595 \cdot \text{pclass}$$

$$\text{Odds}(\text{survived}=\text{false}) = e^{-0.6503} \cdot 2.6103^{\text{pclass}} \cdot 0.0712^{\text{sex}}$$

Note:- Logit = log(odds),  $2.6103 = e^{0.9595}$  and  $0.0712 = e^{-2.6417}$

A positive coefficient in the logit model translates into a coefficient higher than 1 in the odds model. A negative coefficient in the logit model translates into coefficients smaller than 1 in the odds model

Again it is clear that odds of not surviving increases with travel class and reduces with gender = female

## Regression Models

### Logistic Regression : How it Works : An Example cont.

<http://www.kaggle.com/c/titanic-gettingStarted>

#### **An analysis of the odds ratio**

$$\text{Odds}(\text{survived}=\text{false}) = e^{-0.6503 * 2.6103^{pclass} * 0.0712^{sex}}$$

Consider a female passenger (sex="female" / sex=1). We can calculate the survivability of this passenger if she was in 1<sup>st</sup> class versus if she was in second class as an odds ratio.

$$\begin{aligned} &\text{Odds}(\text{survived}=\text{false}2\text{ndclass}) / \text{odds}(\text{survived}=\text{false}1\text{stclass}) \\ &= 2.6103^2 / 2.6103^1 = 2.6103 \end{aligned}$$

Based on the Titanic data set, the odds that a female passenger would not survive if she was in second class increases by a factor of 2.6 compared to her odds if she was in first class. Similarly the odds that a female would not survive increases by nearly seven times if she was in third class!

## Regression Models

### Logistic Regression : Discussion Points

- Logistic Regression can be considered equivalent to using linear regression for situations where the target (or dependent) variable is discrete (not continuous). The response variable can take on two categories (binary decisions) Yes/No, Accept/Not Accept, Default/Not Default
- Logistic Regression comes from the concept of the “logit”. The logit is the logarithm of the odds of the response  $y$  expressed as a function of independent or predictor variables  $x$  and a constant term.

For example: -  $\ln(\text{odds of } y = \text{“Yes”}) = b_1x + b_0$

- The above **logit** gives the odds of the “yes” event, it has to be transformed algebraically to give a probability

$$P(y = \text{“Yes”}) = 1/(1 + e^{-b^1x - b_0})$$

- The predictors can be either numerical or categorical for standard logistic regression. However, in RapidMiner, the predictors can only be numerical because it is based on the SVM formulation

## Regression Models

### Logistic Regression : Summary

- **Model**

The model consists of coefficients for each input predictor that relate to the “logit”. Transforming the logit into probabilities of occurrence (of each class) completes the model

- **Input**

All attributes should be numeric

- **Output**

The label may only be binomial

- **Pros**

One of the most common classification methods. Computationally efficient

- **Cons**

Cannot handle missing data. Not very intuitive when dealing with a large number of predictors

- **Use Cases**

Marketing scenarios (e.g. will click or not click), any general two class problem



## Regression Models

### Generalised Linear Model (GLM)

In statistics, the **generalized linear model (GLM)** is a flexible generalization of ordinary linear regression that allows for response variables that have error distributions other than normal.

**GLM** is a generalization of linear regression in cases where the linear model is related to the response variable via a **link function** that allows the magnitude of the variance of each measurement to be a function of its predicted value.

## Regression Models

### Generalised Linear Model (GLM)

#### Example 1

A prediction model might predict that a 10 degree temperature decrease would lead to 1,000 fewer people visiting the beach. This is unlikely to generalize well over both small beaches (e.g. those where the expected attendance is 50 at a particular temperature) and large beaches (e.g. those where the expected attendance is 10,000 at a low temperature). A temperature drop of 10 degrees would lead to 1,000 fewer people visiting a small beach, that is, the impossible predicted attendance value of -950.

A more realistic model would predict a constant **rate** of increased beach attendance (e.g. an increase in 10 degrees leading to a doubling in beach attendance, and a drop in 10 degrees leading to a halving in attendance).

For such a model, it is the logarithm of the response that is predicted to vary linearly.

## Regression Models

### Generalised Linear Model (GLM)

#### Example 2

A model that predicts the probability of making a **yes/no** choice is even less suitable as a linear-response model, since probabilities are bounded on both ends (between 0 and 1).

For example, where a model predicts the likelihood of a given person going to the beach as a function of temperature, a reasonable model might predict that a change in 10 degrees makes a person two times more or less likely to go to the beach.

What does "twice as likely" mean in terms of a probability? It cannot literally mean to double the probability value (e.g. 50% becomes 100%, 75% becomes 150%, etc.). Rather, it is the odds that are doubling: from 2:1 odds, to 4:1 odds, to 8:1 odds, etc. Such a model is a **log-odds model**.

## Regression Models

### Generalised Linear Model (GLM)

Generalized linear models make allowance for response variables that have arbitrary distributions (rather than simply normal distributions). They allow an arbitrary function (the **link function**) of the response variable to vary linearly with the predicted values (rather than assuming that the response itself must vary linearly).

The generalized linear model typically fits a model to the data to maximize the log-likelihood.

# Regression Models

## Summary

- Function fitting methods are one of the earliest predictive modelling techniques based on concept of supervised learning
- Multiple linear regression works with numeric predictors and a numeric label. Logistic regression works with numerical or categorical predictors and a categorical (typically binomial) label
- A simple linear regression model is developed using methods of calculus. Feature selection impacts on the coefficients of a model. Looked at use of t-stat and p-values and checkpoints for developing good quality models

## Regression Models

### Summary

- A sigmoid curve can better fit predictors to a binomial label. Introduced the concept of logit to map a complex function to a recognisable linear form. Discussed how coefficients of logistic regression can be interpreted and how to measure and improve classification performance
- the generalized linear model (GLM) allows for response variables that have arbitrary distributions (rather than simply normal distributions)

## Base Academic Papers & References

- Black, K. (2008). *Multiple regression analysis*. In K. Black (Ed.), *Business statistics* (pp. 601610). Hoboken, NJ: John Wiley and Sons.
- Cramer, J. (2002). *The origins of logistic regression* (pp. 115). Tinbergen Institute Discussion Paper.
- Eliason, S. (1993). *Maximum likelihood estimation: Logic and practice*. Newbury Park, CA: Sage Publishers.
- Fletcher, R. A. (1963). *A rapidly convergent descent method for minimization*. The Computer Journal, 6(2), 163168.
- Galton, F. (1888). *Co-relations and their measurement, chiefly from anthropometric data*. Proceedings of the Royal Society of London, 45(273-279), 135145.

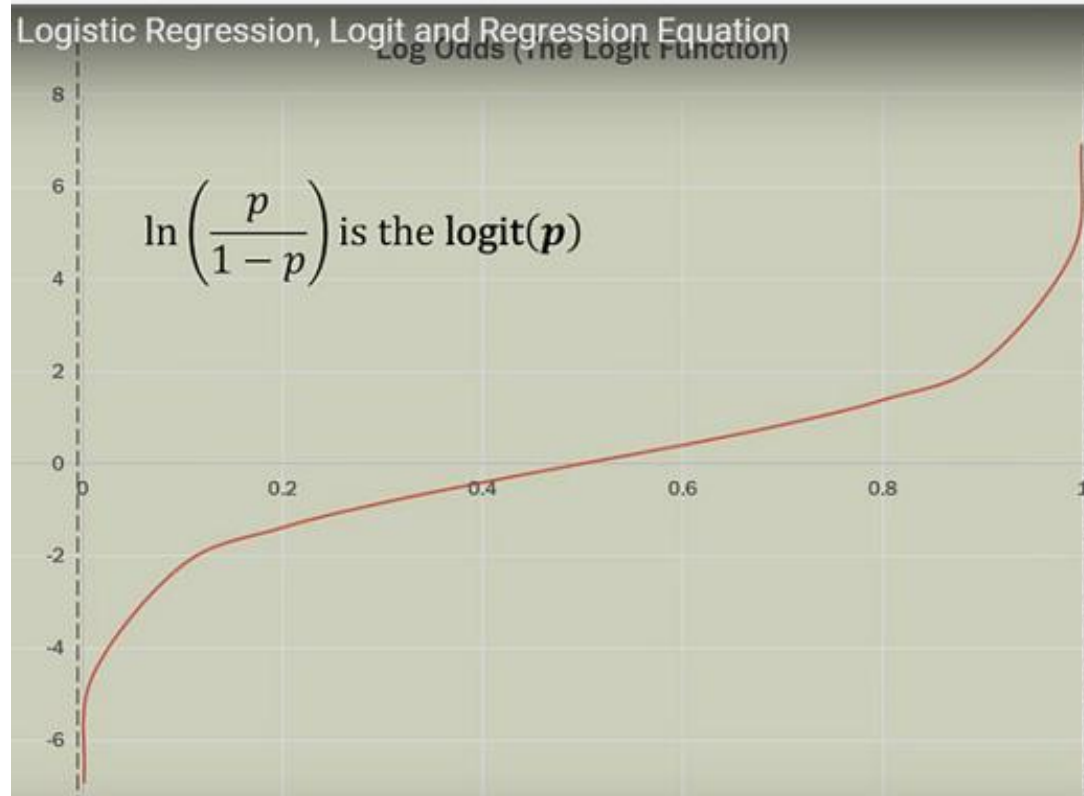
## Base Academic Papers & References

- Harrison, D. A. (1978). *Hedonic prices and the demand for clean air*. Journal of Environmental Economics and Management, 5(1), 81102.
- Marquardt, D. (1963). *An algorithm for least-squares estimation of nonlinear parameters*. Journal of the Society for Industrial and Applied Mathematics, 11(2), 431441.
- Stigler, S. (1999). *Statistics on the table: The history of statistical concepts and methods*. Cambridge: Harvard University Press



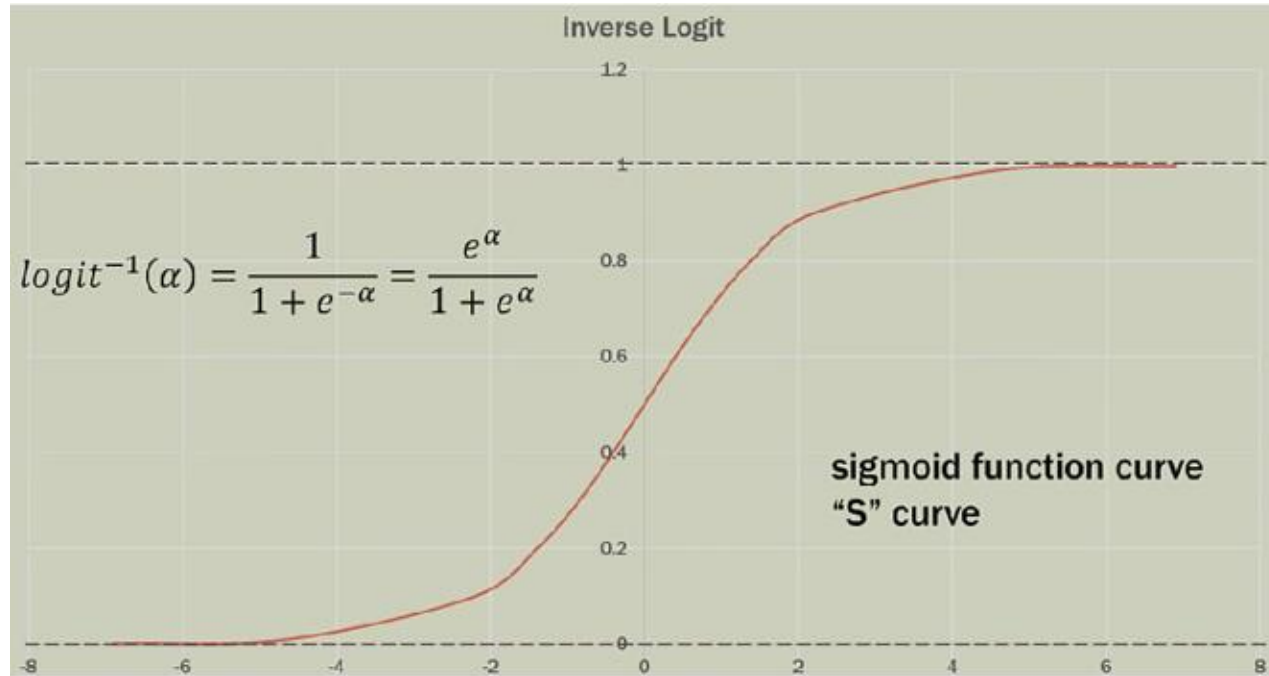
# Regression Models

## Logistic Regression Appendix : Logit



# Regression Models

## Logistic Regression Appendix : Inverse Logit



# Regression Models

## Logistic Regression Appendix : Inverse Logit

