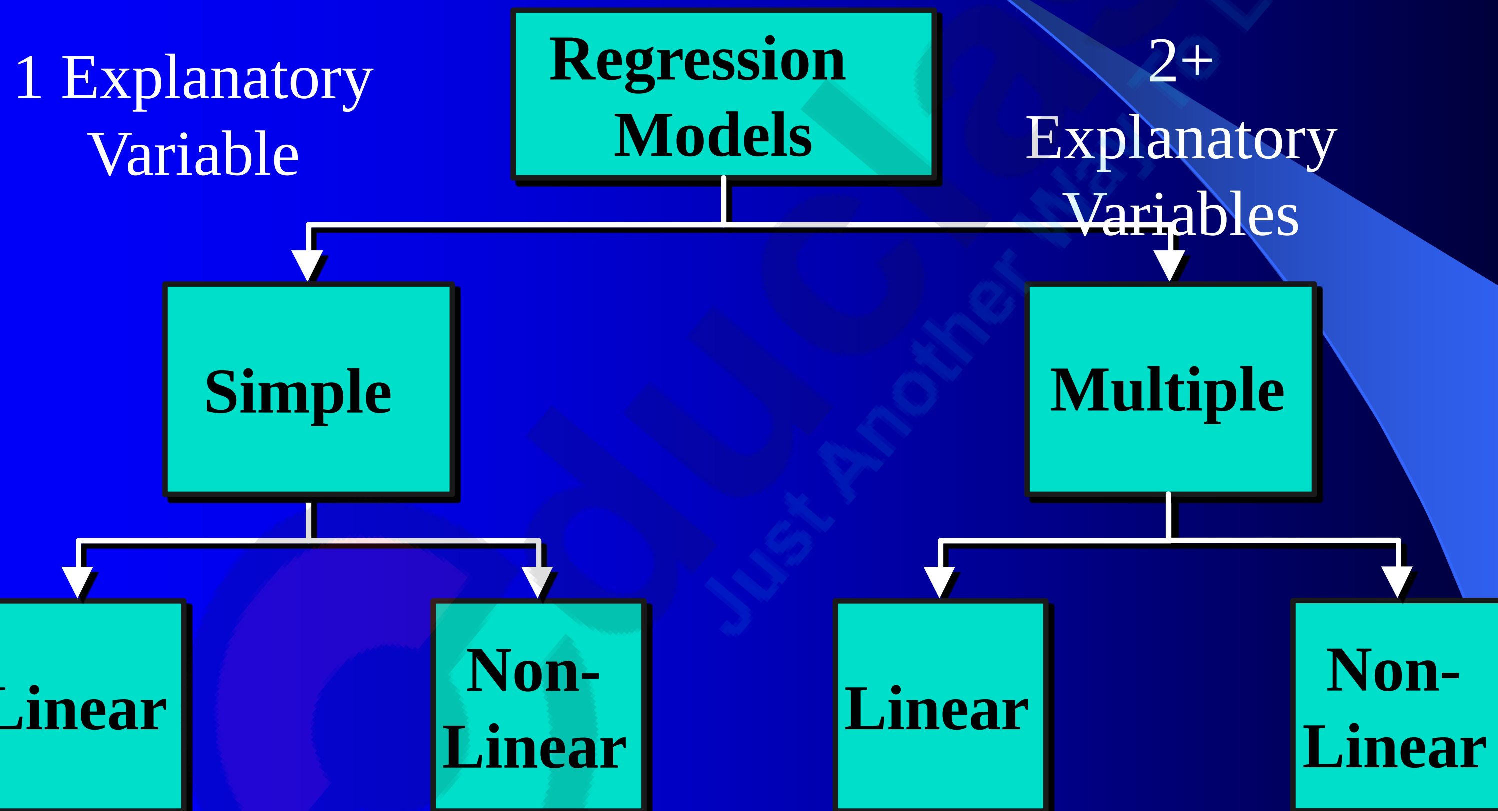
Background

- **Supervised:** We are given input samples (X) and output samples (y) of a function $y = f(X)$. We would like to "learn" f, and evaluate it on new data. Types:
  - **Classification:** y is discrete (class labels).
  - **Regression:** y is continuous, e.g. linear regression.

Regression Models

- Relationship between one dependent variable and explanatory variable(s)
- Use equation to set up relationship
  - Numerical Dependent (Response) Variable
  - 1 or More Numerical or Categorical Independent (Explanatory) Variables
- Used Mainly for Prediction & Estimation

Types of Regression Models



Regression

- Regression is to build a function of independent variables (also known as predictors) to predict a dependent variable (also called response).
- For example: banks assess the risk of home-loan applicants based on their age, income, expenses, occupation, number of dependents, total credit limit, etc.
- Types:
  - Linear regression models
  - Generalized linear models (GLM)

Linear Regression

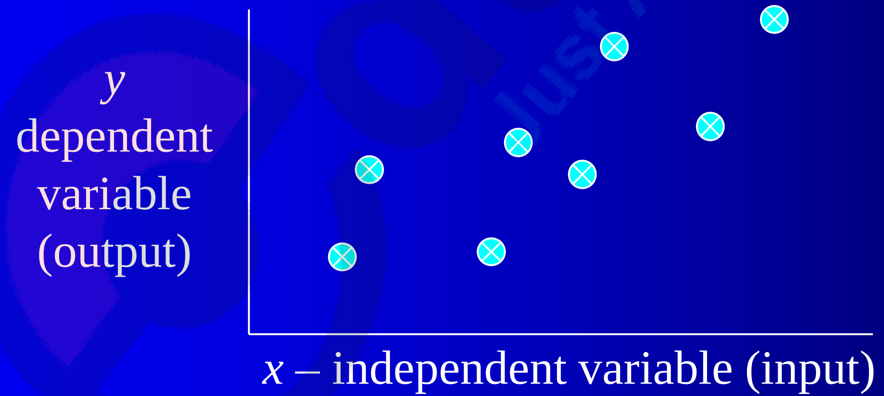
- Linear regression is to predict response with a linear function of predictors as follows:

- where x1; x2;… xk are predictors and y is the response to predict.

$$y = c_0 + c_1 x_1 + c_2 x_2 + \cdots + c_k x_k,$$

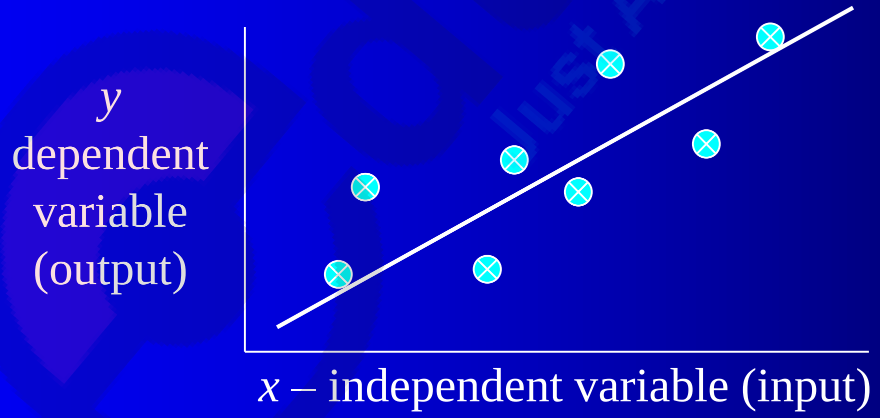
Regression

- For classification the output(s) is nominal
- In regression the output is continuous
  - Function Approximation
- Many models could be used – Simplest is linear regression
  - Fit data with the best hyper-plane which "goes through" the points



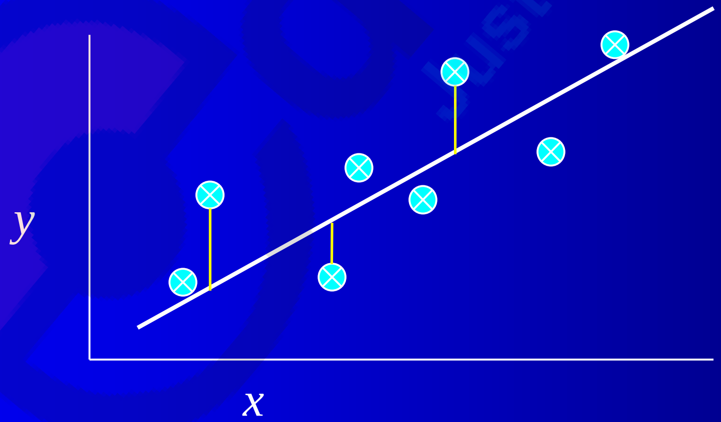
Regression

- For classification the output(s) is nominal
- In regression the output is continuous
  - Function Approximation
- Many models could be used – Simplest is linear regression
  - Fit data with the best hyper-plane which "goes through" the points



Regression

- For classification the output(s) is nominal
- In regression the output is continuous
  - Function Approximation
- Many models could be used – Simplest is linear regression
  - Fit data with the best hyper-plane which "goes through" the points
  - For each point the differences between the predicted point and the actual observation is the *residue*



Simple Linear Regression

- For now, assume just one (input) independent variable *x*, and one (output) dependent variable *y*
  - Multiple linear regression assumes an input vector **x**
  - Multivariate linear regression assumes an output vector **y**
- We will "fit" the points with a line (i.e. hyper-plane)
- Which line should we use?
  - Choose an objective function
  - For simple linear regression we choose sum squared error (SSE)
    - $\Sigma\ (predicted_i - actual_i)^2 = \Sigma\ (residue_i)^2$
  - Thus, find the line which minimizes the sum of the squared residues (e.g. least squares)

Regression Trees for Prediction

- Used with continuous outcome variable
- Procedure similar to classification tree
- Many splits attempted, choose the one that minimizes impurity

Advantages of trees

- Easy to use, understand
- Produce rules that are easy to interpret & implement
- Variable selection & reduction is automatic
- Do not require the assumptions of statistical models
- Can work without extensive handling of missing data

Disadvantages

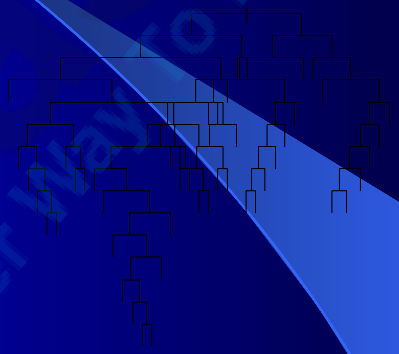
- May not perform well where there is structure in the data that is not well captured by horizontal or vertical splits
- Since the process deals with one variable at a time, no way to capture interactions between variables

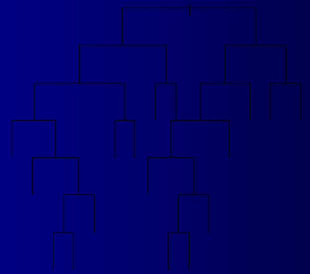
How CART Selects the Optimal Tree

- Use **cross-validation** (CV) to select the optimal decision tree.
- Built into the CART algorithm.
  - Essential to the method; not an add-on
- Basic idea: "grow the tree" out as far as you can…. Then "prune back".
- CV: tells you when to stop pruning.

Growing & Pruning

- One approach: stop growing the tree early.
  - But how do you know when to stop?
- CART: just grow the tree all the way out; then prune back.
  - Sequentially collapse nodes that result in the smallest change in purity.
  - "weakest link" pruning.



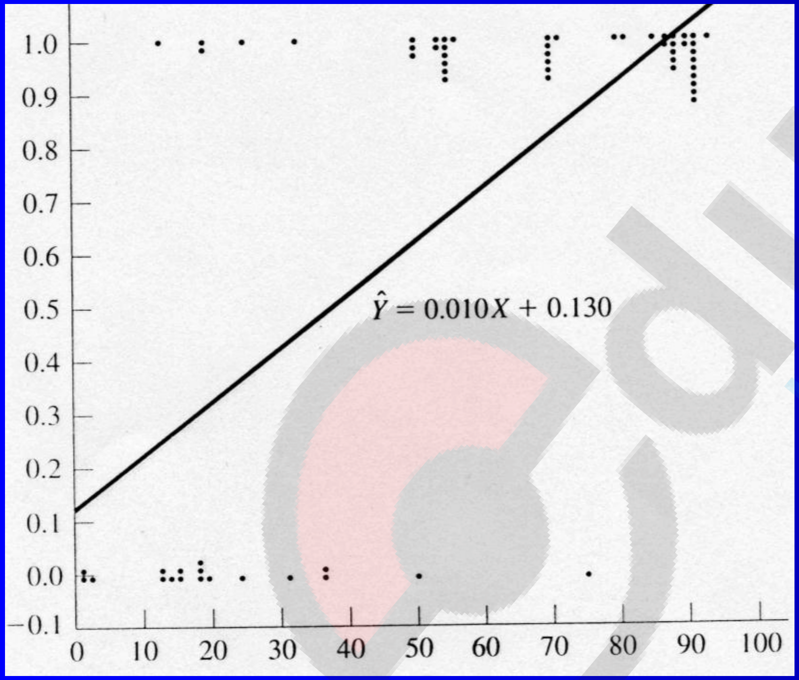


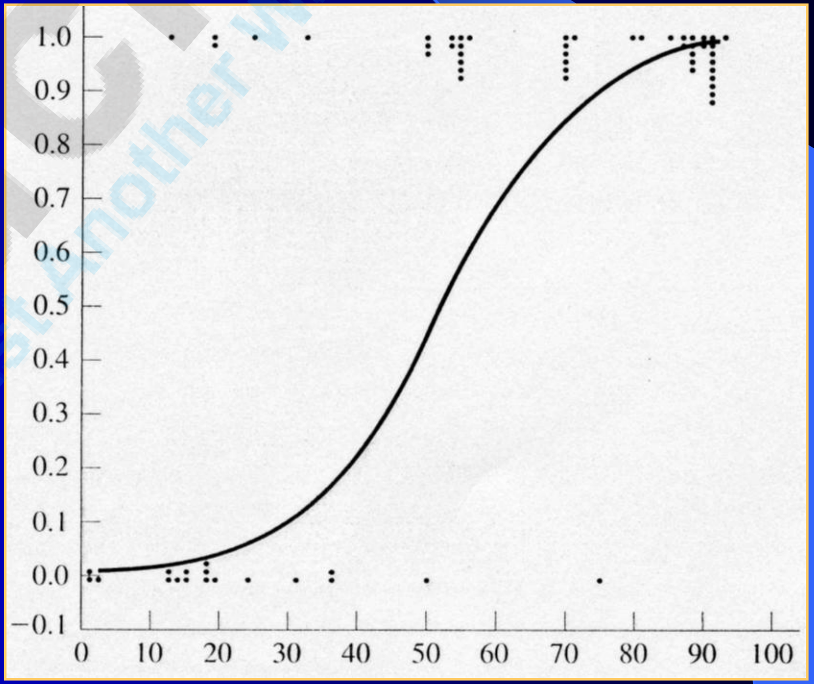
Logistic Regression

- One commonly used algorithm is Logistic Regression
- Assumes that the dependent (output) variable is binary which is often the case in medical and other studies. (Does person have disease or not, survive or not, accepted or not, etc.)
- Like Quadric, Logistic Regression does a particular non-linear transform on the data after which it just does linear regression on the transformed data
- Logistic regression fits the data with a sigmoidal/logistic curve rather than a line and outputs an approximation of the probability of the output given the input

Logistic Regression Example

- Age (X axis, input variable) – Data is fictional
- Heart Failure (Y axis, 1 or 0, output variable)
- Could use value of regression line as a probability approximation
  - Extrapolates outside 0-1 and not as good empirically
- Sigmoidal curve to the right gives empirically good probability approximation and is bounded between 0 and 1





Logistic Regression Approach

Learning

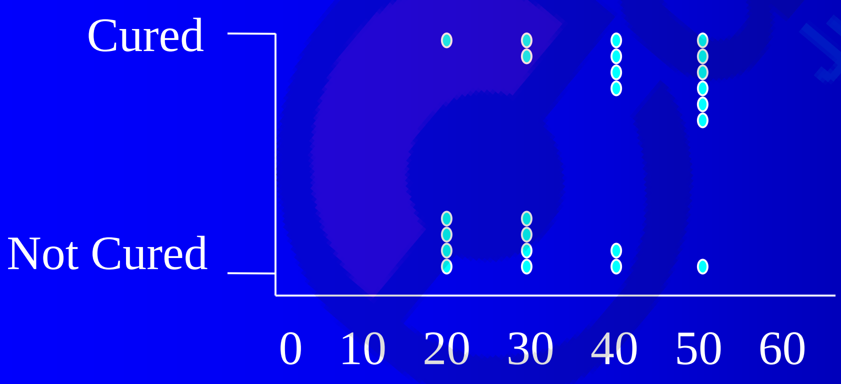
1. Transform initial input probabilities into log odds (logit)
2. Do a standard linear regression on the logit values
   - This effectively fits a logistic curve to the data, while still just doing a linear regression with the transformed input (ala quadric machine, etc.)

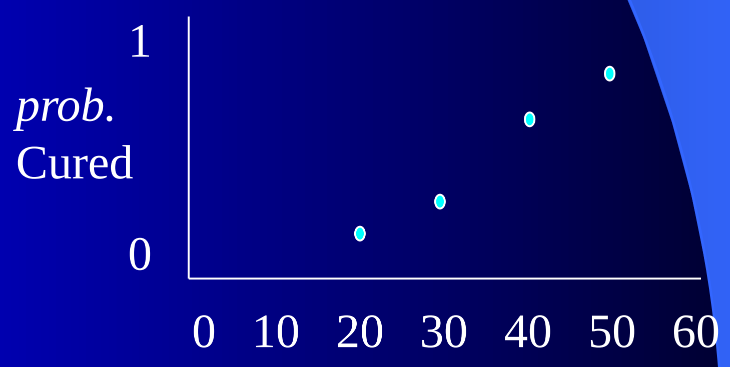
Generalization

1. Find the value for the new input on the logit line
2. Transform that logit value back into a probability

Non-Linear Pre-Process to Logit (Log Odds)

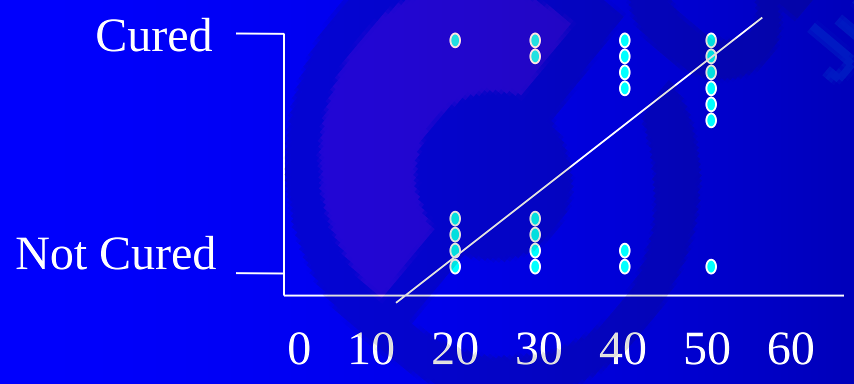
Medication Dosage	# Cured	Total Patients	Probability: # Cured/Total Patients
20	1	5	.20
30	2	6	.33
40	4	6	.67
50	6	7	.86

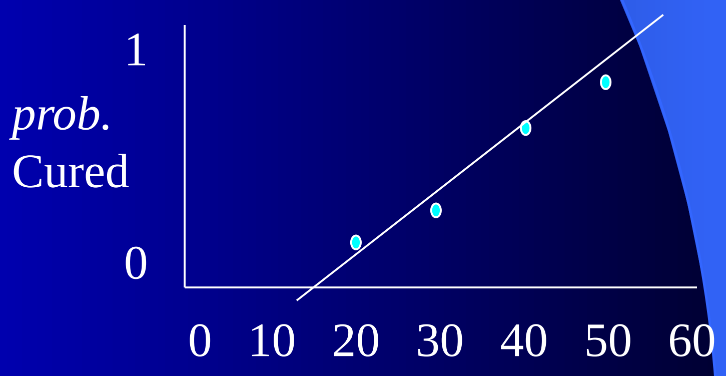




Non-Linear Pre-Process to Logit (Log Odds)

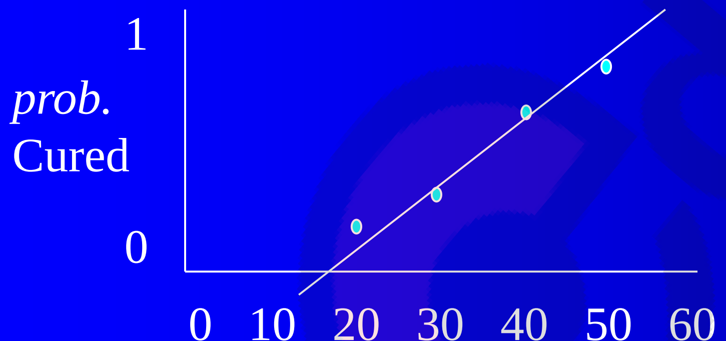
Medication Dosage	# Cured	Total Patients	Probability: # Cured/Total Patients
20	1	5	.20
30	2	6	.33
40	4	6	.67
50	6	7	.86

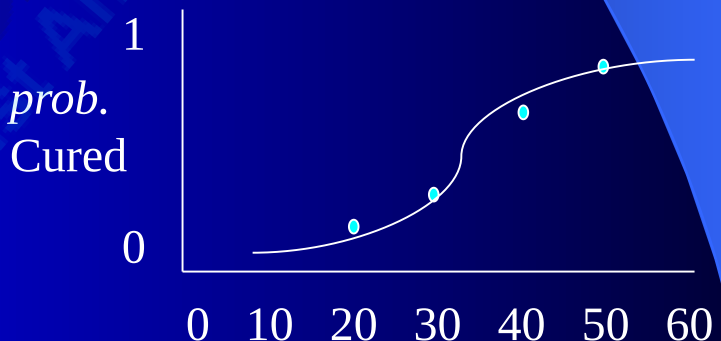




Logistic Regression Approach

- Could use linear regression with the probability points, but that would not extrapolate well
- Logistic version is better but how do we get it?
- Similar to Quadric we do a non-linear pre-process of the input and then do linear regression on the transformed values – do a linear regression on the log odds - Logit





Summary

- Linear Regression and Logistic Regression are nice tools for many simple situations
  - But both force us to fit the data with one shape (line or sigmoid) which will often underfit
- These models are commonly used in data mining applications and also as a "first attempt" at understanding data trends, indicators, etc.