Universidad de Alcalá

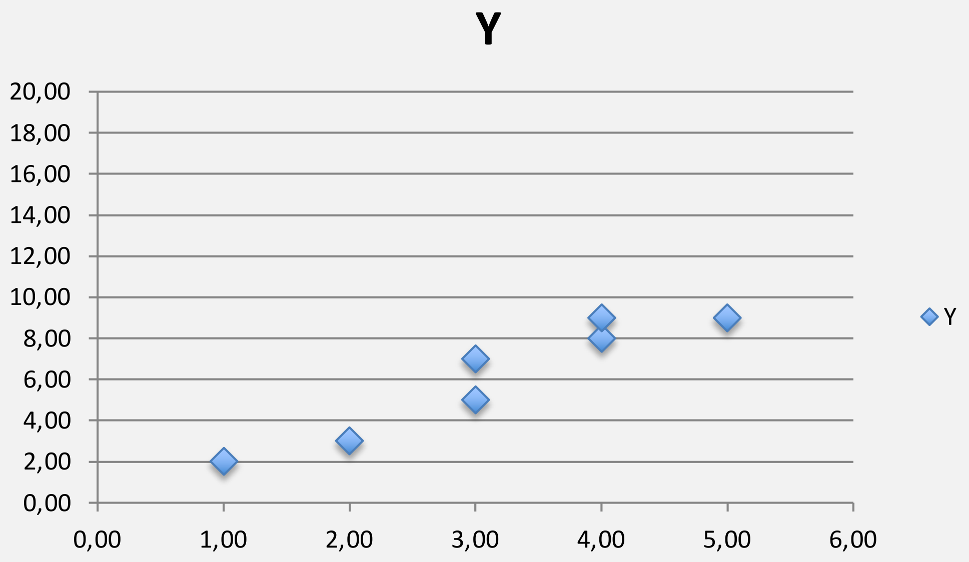
Basic Models

Prof. Ignacio Olmeda
AI LAB

LINEAR MODELS

- As we have mentioned ML allows to create models in a supervised, semi-supervised, unsupervised or reinforced Learning fashion.
- Nevertheless most of the models need to be pre-specified in some sense, that is, we can not simply tell the machine "to learn", we need to formulate some particular hypothesis or, intuitively, to provide some "hints".
- For example, asssume that, after observation, we record and plot the behavior of two variables X and Y, where Y is the variable we are interested in and X the variables that we want to use to predict Y, let us assume that we have:

X	Y
3.00	7.00
2.00	3.00
4.00	8.00
5.00	9.00
4.00	9.00
1.00	2.00
3.00	5.00



- We see that whenever X is bigger Y is bigger so we can postulate that there may be a proportional relationship between X and Y.

$$Y = \alpha X$$

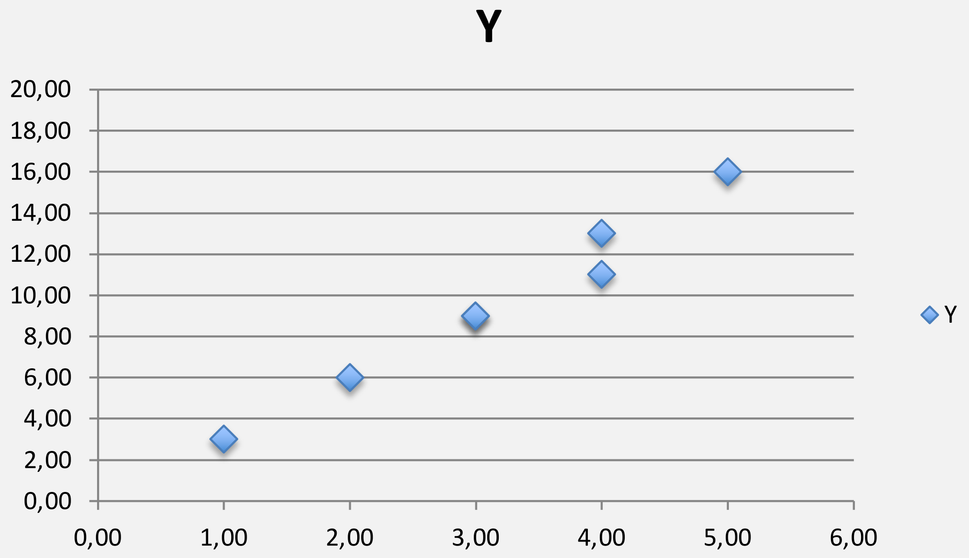
- This is a *model*, that is, a mathematical entity that allows to explain some particular fact, in this case, the relationship between X and Y.

- Note that the model involves not only the variables (X and Y, in our case) but *parameters* (α i our case) that give flexibility so that if either X or Y changes then the model still holds in some sense.

- Note also that the parameters may differ even if the relationship still holds, e.g.

- Assume now the following data and look at the corresponding plot:

X	Y
3.00	9.00
2.00	6.00
4.00	11.00
5.00	16.00
4.00	13.00
1.00	3.00
3.00	9.00



- Note that the relationship still holds (Y is proportional to X) but it is not exactly the same, Y is more responsive to X of, in geometric terms, the slope of the points has increased.

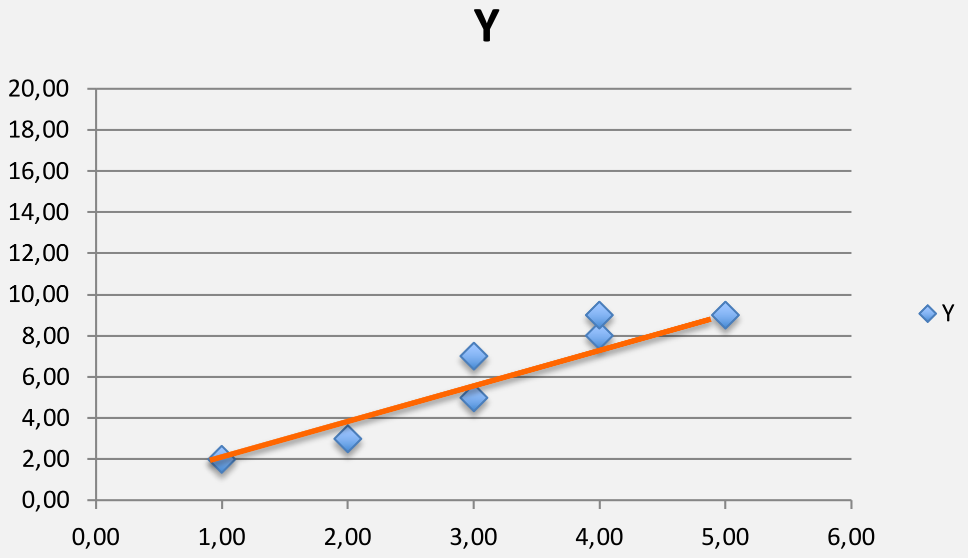
- Notice that even though the model is valid, the value of the parameter previous parameter is not: it should be higher in the second case.

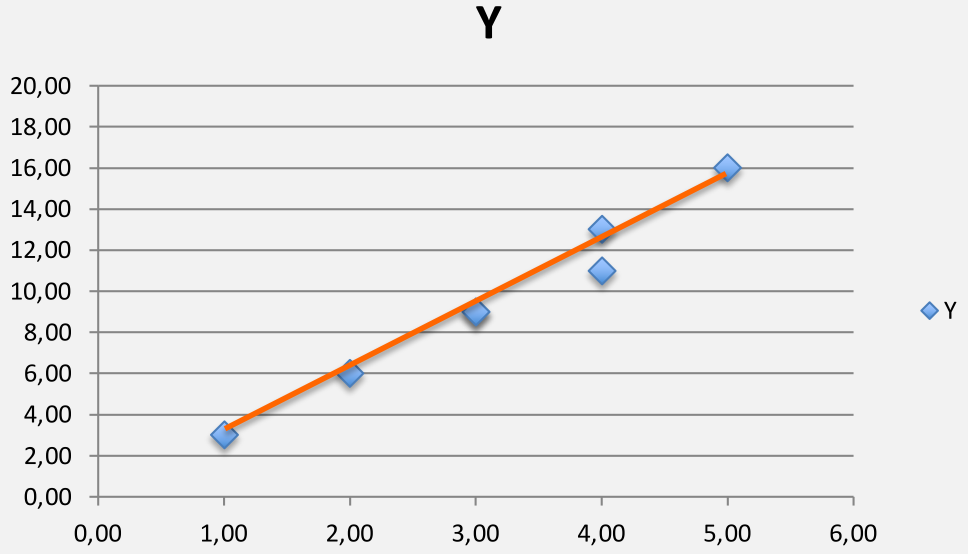
- The process of finding the optimal parameters for a model is what we call *estimation* (in statistical terms) or *learning* (in the ML argot).

- *Algorithms* are computational procedures -most of the times *iterative* procedures- that allow to find the parameters of the model which are optimal under some criterion.

- For example, we may try to find the value of the parameter β so that the model "fits" the data as close as possible.

- Intuitively (no formal method used for the moment) we may suggest a value of β=2 in the first case and β=3 in the second.





- The model is still the same but the *calibration* is different.

- Even generative models that modify their structure adding complexity to capture the relationship between independent and dependent variables need to be calibrated.

- For the moment we can assume that the parameters are found by trial and error.

- Notice that, after finding the optimal value (optimality is defined later) we can employ the model to forecast unseen examples, e.g.

$$\hat{Y}' = \hat{\beta} X'$$

- Models may have some inertial component (in models such as artificial neural networks it is called the *bias* and in Statistics the *intercept*) so that there is a response even if the input is zero:

$$Y = \alpha + \beta X$$

- And similarly we can forecast for new observations:

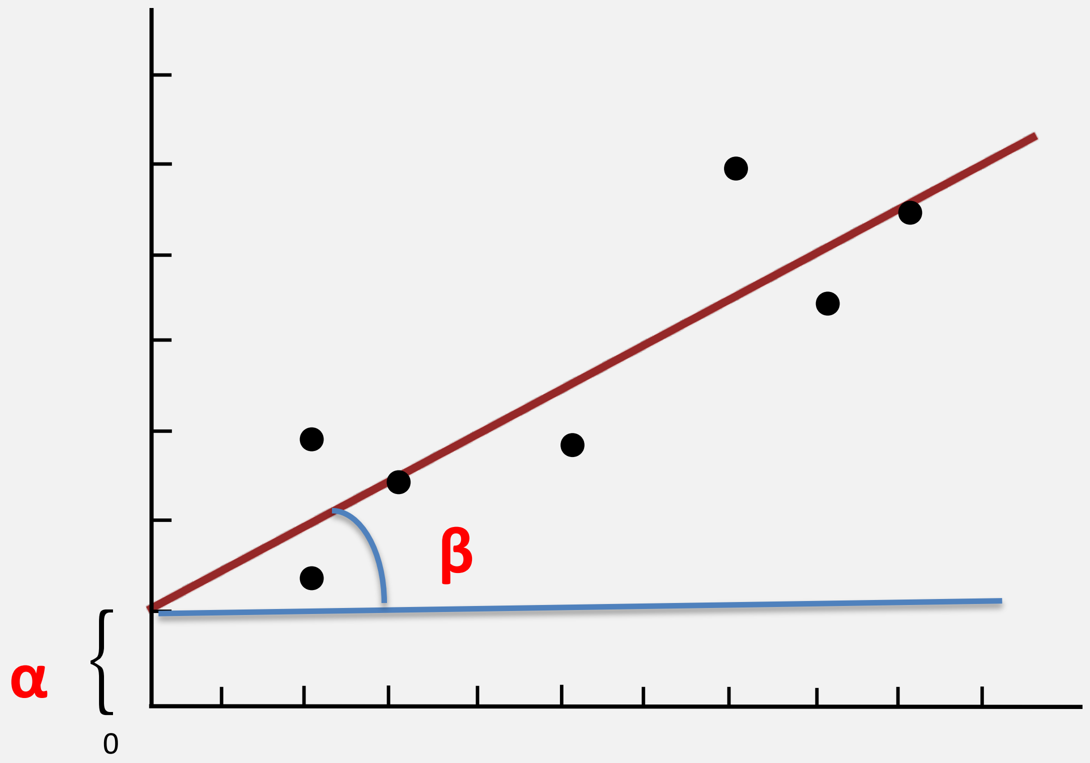
$$\hat{Y}' = \hat{\alpha} + \hat{\beta} X'$$

- The model that we have just described is called a *linear* model for obious reasons, if we plot Y against X the result is a line (which passes through the origin if α=0) and we say that there is a *linear relationship* between Y and X.

- Mor specifically, this model is called a *univariate* linear model because there is only one idependent variable.

- The parameter β is said to be the *slope*, that is the change in Y associated with a one-unit change in X (β=ΔY/ΔX).

- Geometrically:



- In most real applications, the relationship between Y and X will not be perfect it may be affected by unpredictable components that are called *noise*, in such cases we have:

$$Y = \alpha + \beta X + \varepsilon$$

- Note that, by definition, ε is unpredictable so it will be useless trying to make any guess about values of ε, all that can be "learnt" of the relationship between Y and X is sumarized in the parameters of the model, α and β in our case.

- Finally, note that we may have not just one independent variable but a set of them, in such case we can proceed similarly and consider a *multivariate linear model*:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

LINEAR MODEL ESTIMATION
(Optional)

- As mentioned, the parameters that index any model need to be found, we have also mentioned that they should be the "best" ones under some particular criterion.

- One of the commonest criteria is *mean squared error*, there are technical details why this criterion is optimal in many applications but for the moment we will rely on an intuitive interpretation.

- Note that for any value of the parameters we can calculate the difference between the actual data and our prediction using those parameters:

$$\varepsilon = g(Y, \hat{Y}) = (Y - \hat{Y}) = (Y - (\hat{\alpha} + \hat{\beta} X))$$

- Intuitively, it is obvious that a "good" model would provide foerecasts that are as close as possible to the true observations, that is, a good model would try to minimize ε, i.e.

$$find\ \alpha, \beta \quad to\min \varepsilon$$

- Of course, one would like to do this for ALL the observations in the dataset, assume we have a *sample*

$$\left\{X_i, Y_i\right\}$$

- Naturally one woud like:

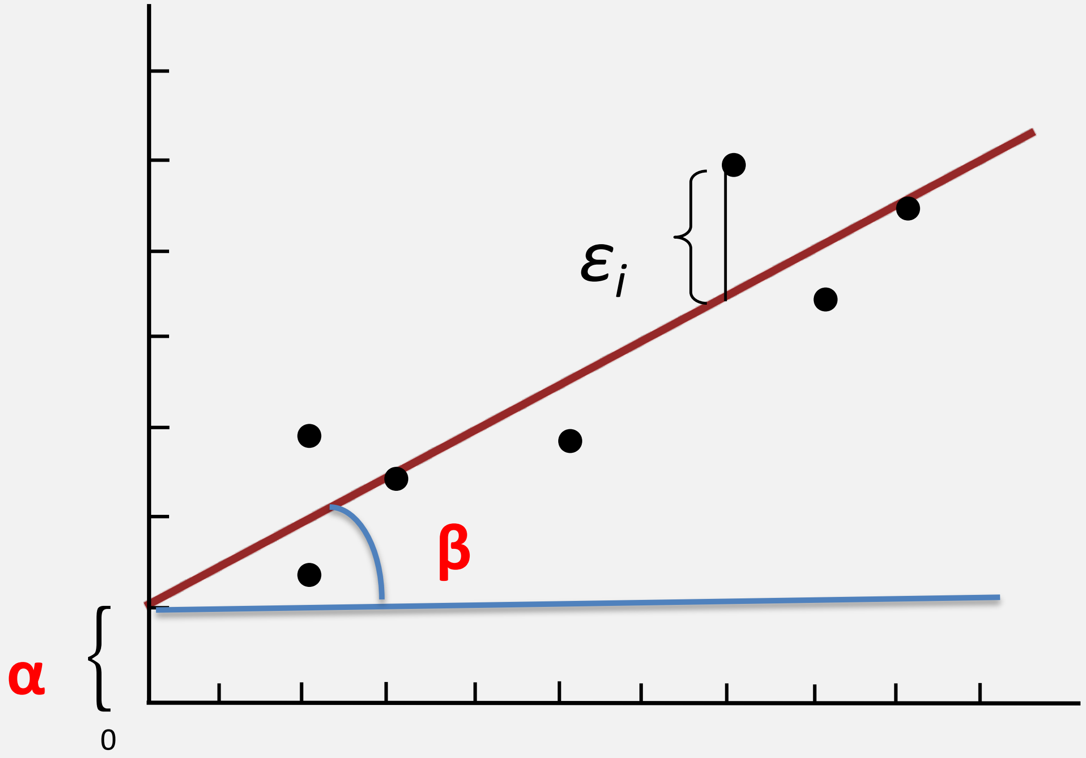
$$\min_{\alpha,\beta} \sum \varepsilon_i = (Y_i - (\alpha + \beta X_i))$$

- Since positive errors could compensate negative errors it seems more reasonable

$$\min_{\alpha,\beta} \sum \varepsilon_i = (Y_i - (\alpha + \beta X_i))^2$$

- Note that the above formula explains the name: *minimum* *<u>squared</u>* *errors.*

- Geometrically:



- Since, as mentioned, random guess is generally unfeasible, the problem now becomes on using an efficient algorithm to find the parameters.

- There are several ways to do that but the simplest method is to employ the *normal equation*, in the multivariate case assume we have

$$X = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2n} \\ \dots & \dots & \dots & \dots \\ X_{m1} & X_{m2} & \dots & X_{mn} \end{pmatrix} \qquad Y = \begin{pmatrix} Y_{11} \\ Y_{12} \\ \dots \\ Y_{1m} \end{pmatrix}$$

- The optimal parameters, under the *mse* criterion, can be found solving the equation:

$$\beta = (X^t X)^{-1} X^t Y$$

- Note, thet if we have an intercept the above formula holds by considering

$$X = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1n} \\ 1 & X_{21} & X_{22} & \dots & X_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & X_{m1} & X_{m2} & \dots & X_{mn} \end{pmatrix}$$

- In this case: $\alpha, \beta = (X^t X)^{-1} X^t Y$

- In order to reduce notation complexity it is common to simplify as

$$\theta = (X^t X)^{-1} X^t Y$$

- Where

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \ldots + \theta_n X_n$$

- Note that these parameters found are optimal under the mse criterion, but they do not need to be if we change the *performance* function, e.g.

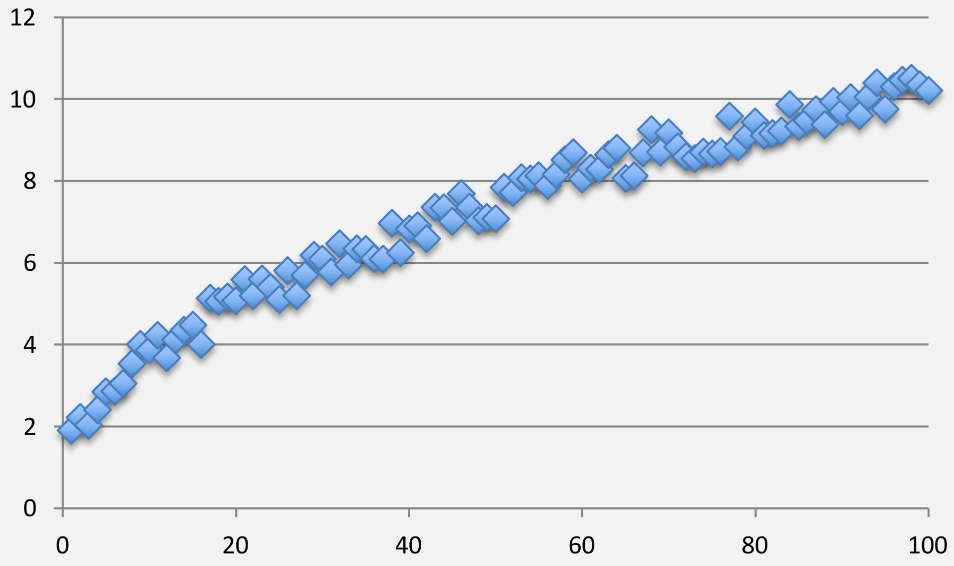
$$\varepsilon = g(Y, \hat{Y}) = \left| Y - \hat{Y} \right|$$

$$\varepsilon = g(Y, \hat{Y}) = \left(Y^+ - \hat{Y}^+ \right)^2$$

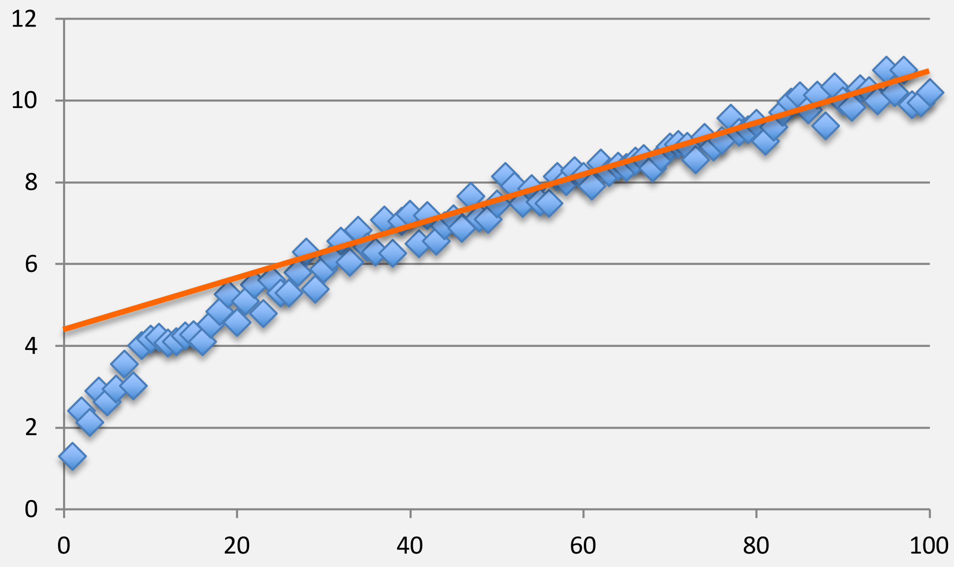
NONLINEAR MODELS

- In many situations the relationship between dependent and independent variables can not be captured using the preceeding models.

- For example, let us suppose that we want to estimate the number of individuals of some population, as time passes the population increases fast but after some point it increases at a lower pace because individuals compete for scarce resources.



- Note that a linear model would fail to capture the relationship between time and population size.



$$Y \neq \alpha + \beta t$$

- For example we could suggest

$$Y = \alpha + \beta t + \gamma t^2$$

- Which is a *quadratic regression model*.

- After a nonlinear model has been proposed, the procedure to fit it to the data is very similar as for the linear case, again we have to find the optimal value of the parameters that index the model

$$Y = \hat{\alpha} + \hat{\beta} t + \hat{\gamma} t^2$$

- The complication comes from the fact that, in most of the cases, the normal equation can not be used because of the nonlinear structure of the relationship, and an iterative procedure must be used.

- There are many algorithms to estimate such parameters being the *Gauss-Newton method* one of the most commonly employed.

- Notice, also, that some non-linear models can be made linear one using some transformation.

- For example, consider the *multiplicative model* (in contrast with the preceeding one, which is and *additive model*):

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2} \ldots . X_n^{\beta_n}$$

- Taking logarithms

$$\log(Y) = \log(\beta_0) + \beta_1 \log(X_1) + \beta_2 \log(X_2) + \ldots + \beta_n \log(X_n)$$

- And renaming log(Y)=Y’, log(β_0)=β_0’, log (X_i) = X'_i, we obtain:

$$Y' = \beta_0' + \beta_1 X_1' + \beta_2 X_2' + \ldots + \beta_n X_n'$$

- Which is linear, and so parameters can be estimated in the usual way.

- After the best values of the parameters are found, one can transform them back to find the model of interest, i.e.

$$\hat{\beta}_0 = e^{\hat{\beta}_0 '}$$

- In practice, these transformations rarely do exist (except in very particular settings) and one has to consider the original functional form and estimate its parameters directly.

- Out of the infinite number of nonlinear models one can find some that are particularly useful and that are extensively used in machine learning.

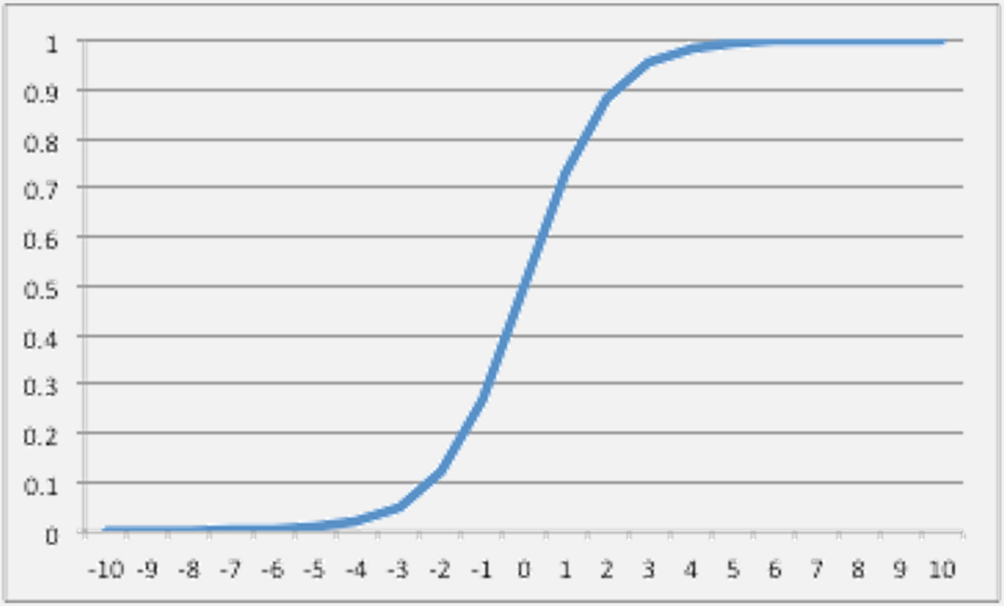
- One of them is the *univariate logistic model*:

$$Y = \frac{1}{1+e^{-\beta_0-\beta_1 X}}$$

- Note that this model has a relevant property, when X is possitive and very large Y=1 and when X is negative and very large then Y=0.

- The response variable (output) is, then, a bounded number between [0,1], and so it can be interpreted as a probability.

$$P(Y=1 \mid X)$$



- Of course we can also consider the *multivariate logistic model*:

$$Y = \frac{1}{1 + e^{-\beta_0 - \beta_1 X_1 - \beta_2 X_2 - \ldots - \beta_n X_n}}$$

$$P(Y = 1 \mid X_1, X_2, \ldots, X_n)$$

- As mentioned, logistic models can be used in situations where one wants to determine the probability that some event happens (Y=1) given some predictor variables.

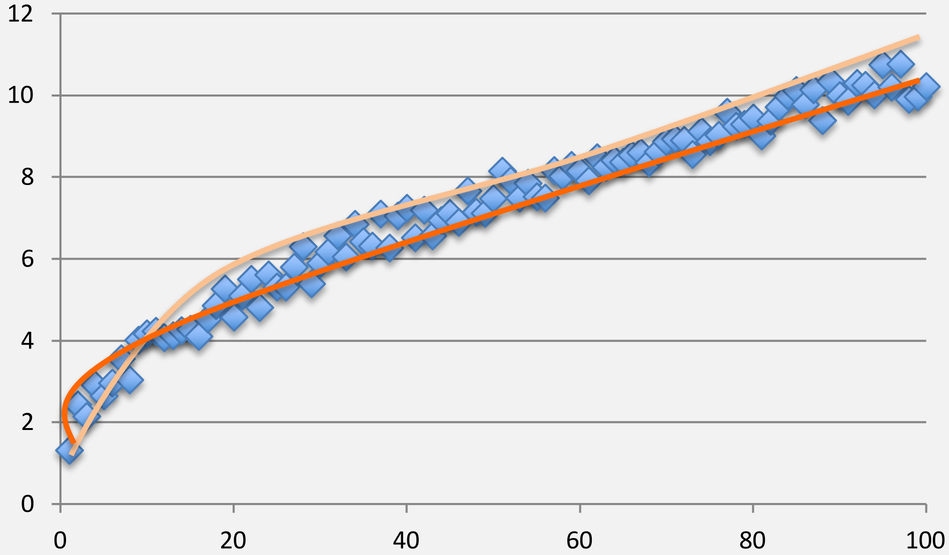
- For example, one might be interested in calulating the probability that one individual develops some particular symptom given that he has taken some combination of drugs by using some database of symptoms-drugs intake.

- After the model is built and tested, doctors could use it to predict whether or not some combination will provoke some reaction.

- The logistic function has been particularly important in the development of Artificial Neural Networks models since, for some time, it was the most widely employed transfer function used in artificial neurons.

PARAMETRIC AND NONPARAMETRIC MODELS

- The problem is that all linear models are the same all but all the non-linear models are different, that is, there is a huge number of models that can be tried.
- In the preceeding example we could propose any of these two models



$$Y = \alpha t + \beta t^2 + \gamma t^3 + \delta t^4$$

$$Y = \alpha \sqrt{t}$$

- In fact there is an infinite number of non-linear models that can be fitted.

- Obviously it would be infeasible to try many different models, or even a relatively moderate number of them, we need to employ one single parameterization that can be used regardless of the relationship between Y and X.

- The models that we have seen up to now (*univariate* and *multivariate linear* models and *logistic* model) are *parametric* models, they employ a limited number of parameters which are estimated trying to fit the data.

- In contrast, a *nonparametric model* is the one which can not be characterized by a bounded number of parameters, that is, the number of parameters is not pre-determined.

- Notice that nonparametric models still use parameters, the name is a bit confusing since it can be interpreted that these models do not employ parameters, which is not the case.

- Some nonparametric models have the important property that they can approximate any funtion to any desired level just given enough complexity, this is called the *universal approximation property*.

- For example, asume that we want to find the relationship between one variable Y and a two variables X_1,X_2, that is:

$$Y = f(X_1, X_2)$$

- Assume that we do not know the specific functional form of *f* so that e.g. we do not know whether *f* is linear, logistic etc.

- We may use a nonparametric model so that regardless which is *f* it can replicate the relationship between Y and X.

- One candidate is a *polinomial model of order n*, where *n* is indeterminate:

$$Y = \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_1^2 + \alpha_4 X^2 + \alpha_5 X_1 X_2 + ...$$

- It can be demonstrated that there exist parameters

$$\hat{\alpha}_1,\hat{\alpha}_2,\hat{\alpha}_3,\hat{\alpha}_4,\hat{\alpha}_5.....$$

such that

$$Y = f(X_1,X_2) \approx \hat{\alpha}_1 X_1 + \hat{\alpha}_2 X_2 + \hat{\alpha}_3 X_1^2 + \hat{\alpha}_4 X^2 + \hat{\alpha}_5 X_1 X_2 + ...$$

to any desired level.

- Notice that to the extent that we employ nonparametric models with the approximation property it is unimportant which is the specific functional form that we are trying to find, since the nonparametric models will "behave" exactly the same as the functional form of interest, which is unknown.

- There are several nonparametric models with the universal approximation property, being feedforward neural networks with (e.g.) sigmoid units one of the most powerful ones.

- Nonparametric mdels have a number of advantages against parametric ones being the most important the possibility to use a single model to find any particular relationship.

- Nevertheless they have several disadvantages too:
    - Their complexity needs to be controlled to avoid *overfitting* (more on this latter)
    - The computational load is generaly important
    - Models are not easily understandable
    - Formal testing (e.g. significativity of the parameters) is difficult or mostly impossible.
    - Performance degrades when there is not enough data