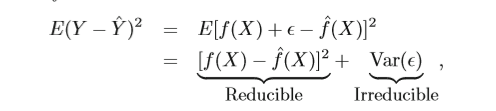
**Prediction**

Y = f(X) + error/epsilon 🡪 Y = alpha + Beta \* X + error

In order to predict Y, we use estimates Y-hat = f-hat(X) (error term averages to 0 - +ve and -ve errors cancel each other). Here f-hat is treated as a black box bcoz we are not concerned about f-hat as long as Y-hat is accurate.

The accuracy of Yhat as a prediction for Y depends on 2 quantities – reducible and irreducible error.

* **Reducible error:** f-hat will not be a perfect estimate for f, and this inaccuracy will introduce some error. This error is reducible because we can potentially improve the accuracy of f-hat by using the most appropriate statistical learning technique to estimate f.
* **Irreducible error:** Even if we form a perfect estimate for f such that Y-hat = f(X), there will still be some error because Y is also a function of error that cannot be predicted using X. This is known as the irreducible error, because no matter how well we estimate f, we cannot reduce the error introduced by error term.
  + There may be some unmeasured variables that are useful in predicting Y which have not been included. That becomes part of the error variability.
  + Also, unmeasurable variation in error term. For example, the risk of an adverse reaction might vary for a given patient dep on patients’ wellbeing on that day
  + These factors contribute to error > 0



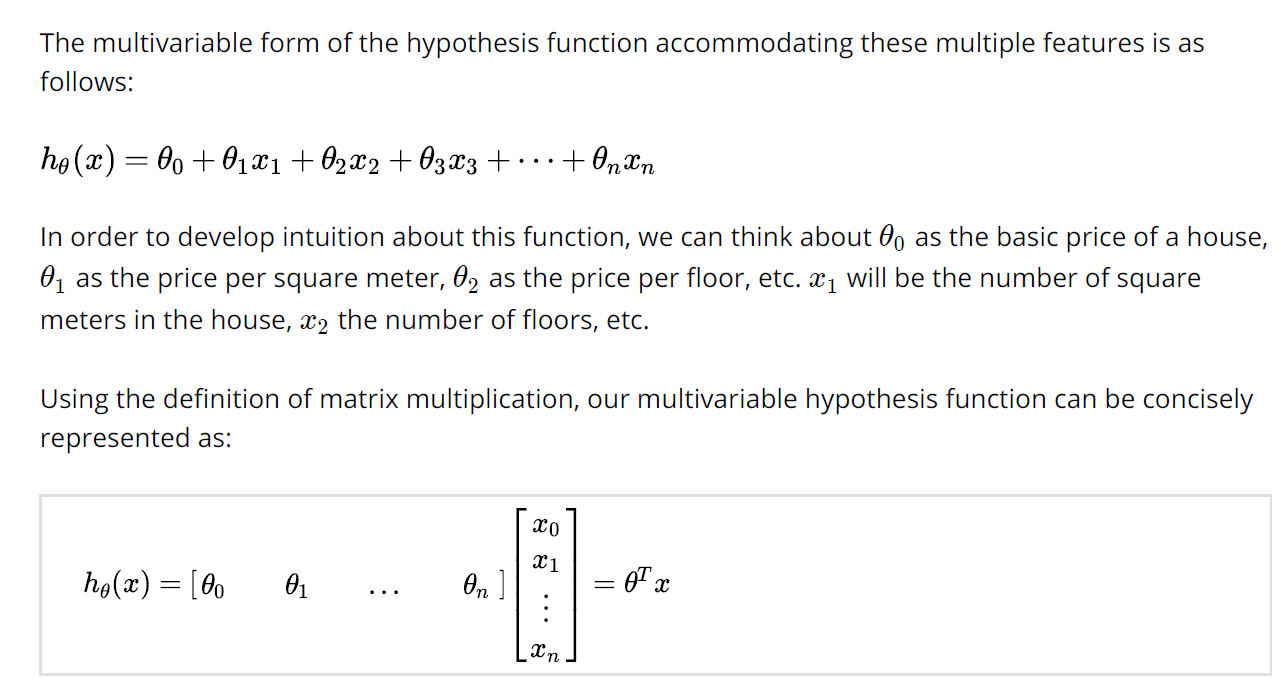
where E(Y – Y-hat )^2 represents the average, or expected value, of the squared expected value diﬀerence between the predicted and actual value of Y , and Var(Epsilon) represents the variance associated with the error term epsilon.

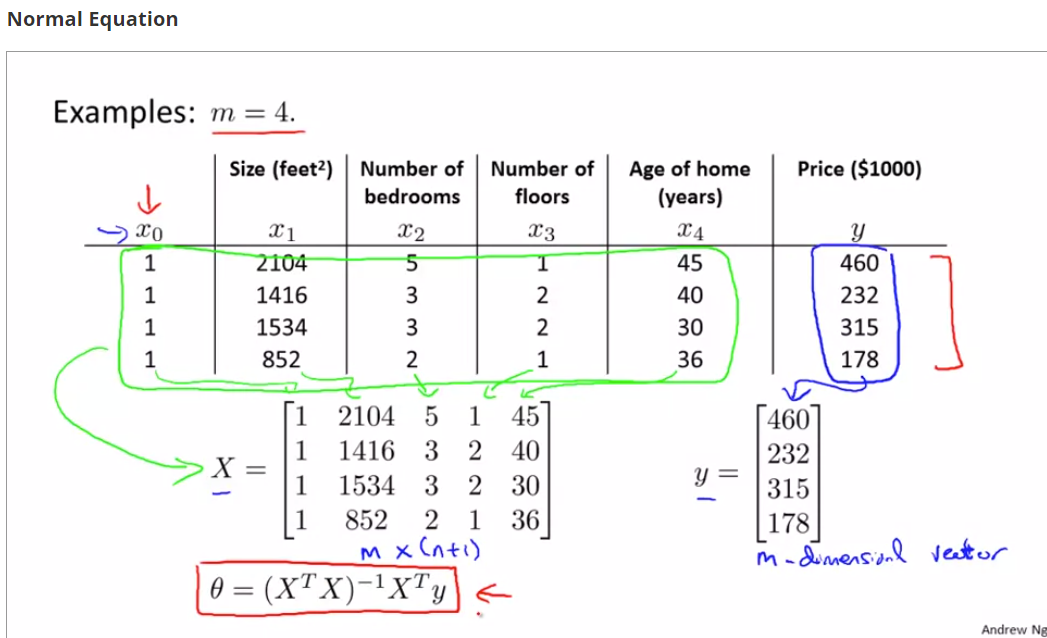
**Inference**

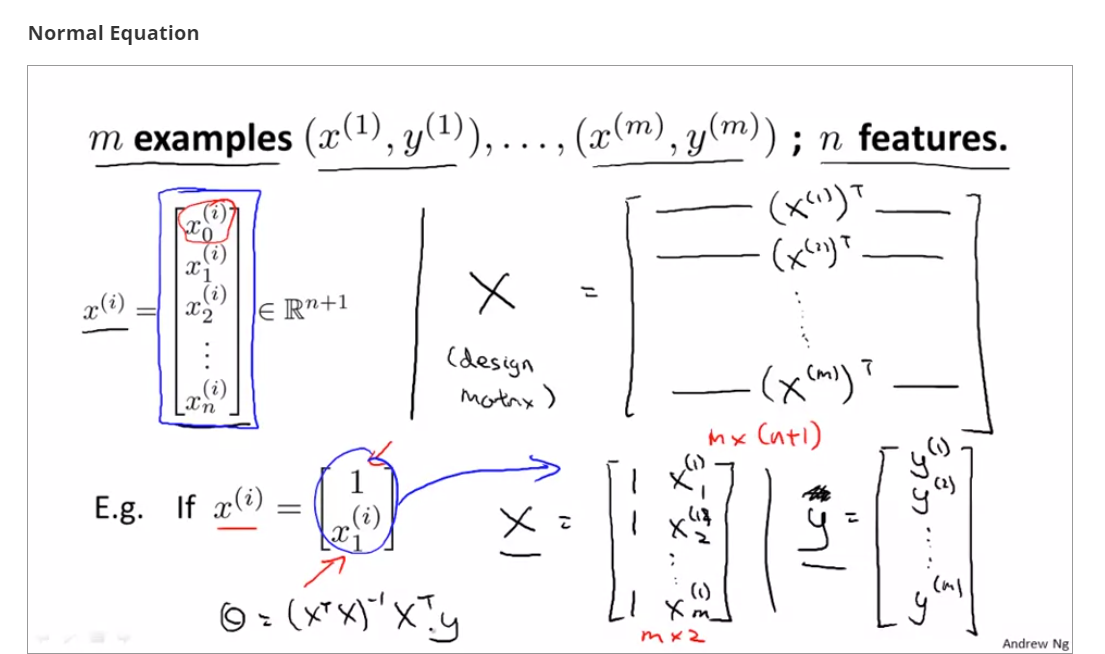
* How Y changes as a function of predictor variables.
* F-hat cannot be treated as a black box, because we need to know its exact form.
  + Which predictors are associated with the response?
  + What is the relationship between the response and each predictor?
  + Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

**Some examples**

|  |  |
| --- | --- |
| **Prediction** | **Inference** |
| The goal is to identify individuals who will respond positively to a mailing, based on observations of demographic variables measured on each individual | Which media contribute to sales?  – Which media generate the biggest boost in sales? or  – How much increase in sales is associated with a given increase in TV advertising? |
| Non-linear models make accurate predictions | linear models are best suited for inferences |







**Estimation of f using parametric and non-parametric approaches**

**Parametric:** The model-based approach just described is referred to as parametric; it reduces the problem of estimating f down to one of estimating a set of parameters

We give a functional form and shape to f(X). Let’s assume linear form



* Estimate β0,β1,...,βp using Least Square or other estimations such that the cost function is minimized.
* **Advantage:** Assuming a parametric form for f simpliﬁes the problem of estimating f because it is generally much easier to estimate a set of parameters, such as β0,β1,...,βp (in this example, it’s the linear model), than it is to ﬁt an entirely arbitrary function f
* **Disadvantage:** The functional form used to estimate f is very diﬀerent from the true f, in which case the resulting model will not ﬁt the data well**.** The model we choose will usually not match the true unknown form of f which may result in poor estimate. But in general, ﬁtting a more ﬂexible model requires estimating a greater number of parameters and may potentially lead to overfitting the data.

**Non-parametric methods:**

Non-parametric methods do not make explicit assumptions about the functional form of f. Instead they seek an estimate of f that gets as close to the data points as possible.

* **Advantage:** By avoiding the assumption of a functional form for f, they have the potential to accurately ﬁt a wider range of possible shapes for f. There’s no danger of estimated f being very diff from true f as no functional form assumed.
* **Disadvantage:** Since they do not reduce the problem of estimating f to a small number of parameters, a very large number of observations (far more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for f.

**Restricted vs flexible methods:**

**Restricted methods:**

* **Linear regression** is a relatively inﬂexible approach, because it can only generate linear functions such as the lines or the plane.
* **Lasso** relies upon the linear model but uses a more restrictive procedure for estimating coefficients and sets number of coeff to 0. It is a less ﬂexible approach than linear regression but more interpretable than linear regression. This is because the response variable will only be related to a small subset of the predictors, those with nonzero coeﬃcient estimates
* **GAM(Generalized Additive Models),** the relationship between each predictor and the response is modeled using a curve. More ﬂexible than linear regression and less interpretable than linear regression.
* If we are mainly interested in inference, then restrictive models are much more interpretable. For instance, when inference is the goal, the linear model may be a good choice since it will be quite easy to understand the relationship between Y and X1, X2,...,Xp.

**Flexible methods:**

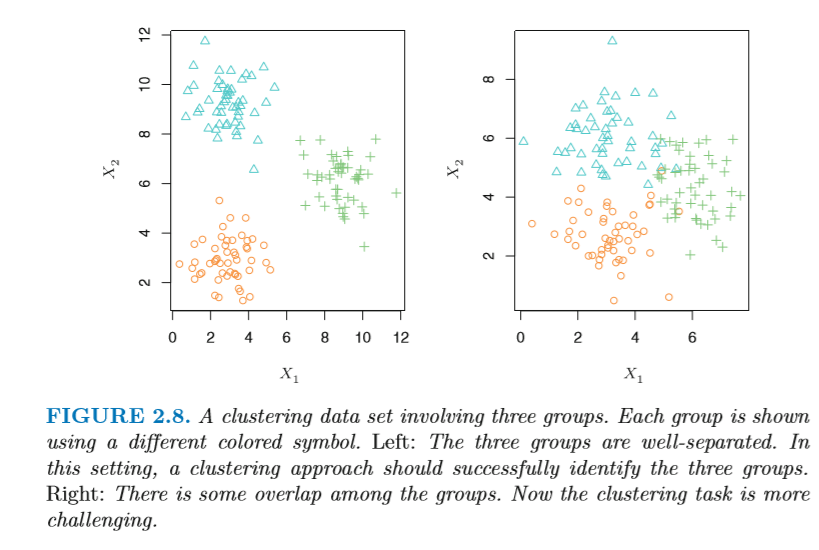
* Methods such as the thin plate **splines** are considerably more ﬂexible because they can generate a much wider range of possible shapes to estimate f. Very ﬂexible approaches, such as the splines and the boosting methods can lead to such complicated estimates of f that it is diﬃcult to understand how any individual predictor is associated with the response.
* **bagging, boosting, and support vector machines** with non-linear kernels are highly ﬂexible support vector machine approaches that are harder to interpret.
* If we are interested in prediction where interpretability is not of interest, flexible models are suited but many times it leads to overfitting, we prefer less flexible models.

**Supervised learning vs unsupervised learning:**

**Supervised:** For each observation of the predictor measurement(s) xi, i =1,...,n there is an associated response measurement yi. Example, linear regression and logistic regression as well as more modern approaches such as GAM, boosting, and support vector machine follow supervised approach

**Unsupervised:** For each observation of the predictor measurement(s) xi, i =1,...,n there is no associated response measurement yi. It is not possible to ﬁt a linear regression model, since there is no response variable to predict.

* **Cluster analysis, or clustering – ascertain if the observations fall into distinct cluster groups.** **A clustering method could not be expected to assign all the overlapping points to their correct group (blue, green, or orange). If there are small number of predictor variables, a scatter plot might be able to identify the clusters, but with p predictor variables, scatter plot is not a viable option. For this reason, automated clustering methods are important**



**Semi supervised learning:** For instance, suppose that we have a set of n observations. For m of the observations, where m<n, we have both predictor measurements and a response measurement. For the remaining n − m observations, we have predictor measurements but no response measurement. Such a scenario can arise if the predictors can be measured relatively cheaply but the corresponding responses are much more expensive to collect

**Regression vs Classification problems:**

**Regression:** Where the response is quantitative or numerical. For ex, person’s age, height, or income. ***Least squares linear regression*** is used with a quantitative response.

**Classification:** Where the response is qualitative or Categorical. For ex, person’s gender (male or female), the brand of product purchased (brand A, B, or C). ***Logistic regression*** is typically used with a qualitative (two-class, or binary) response. Since it estimates class probabilities, it can be thought of as a regression as well.

***K-nearest neighbors*** and ***boosting*** can be used in regression as well as classification problems.