Machine Learning

Unit 2 – Linear Models

M.E. (CSE) – I Semester

Linear model – Introduction

- Models that can be understood in terms of lines and planes, commonly called linear models.
- Linear models are parametric – means they have a fixed form with a small number of numeric parameters that need to be learned from data.
- This is different from tree/rule models, where the structure of the model is not fixed in advance.

Linear model – Introduction

- Linear models are stable – small variations in the training data have only limited impact on the learned model.
- Linear models are less likely to overfit the training data – they have relatively few parameters.
- To summarize: linear models have low variance but high bias.
- Linear models are preferred when you have limited data and want to avoid overfitting.
- Linear models exists for all predictive tasks, including classification, probability estimation and regression.

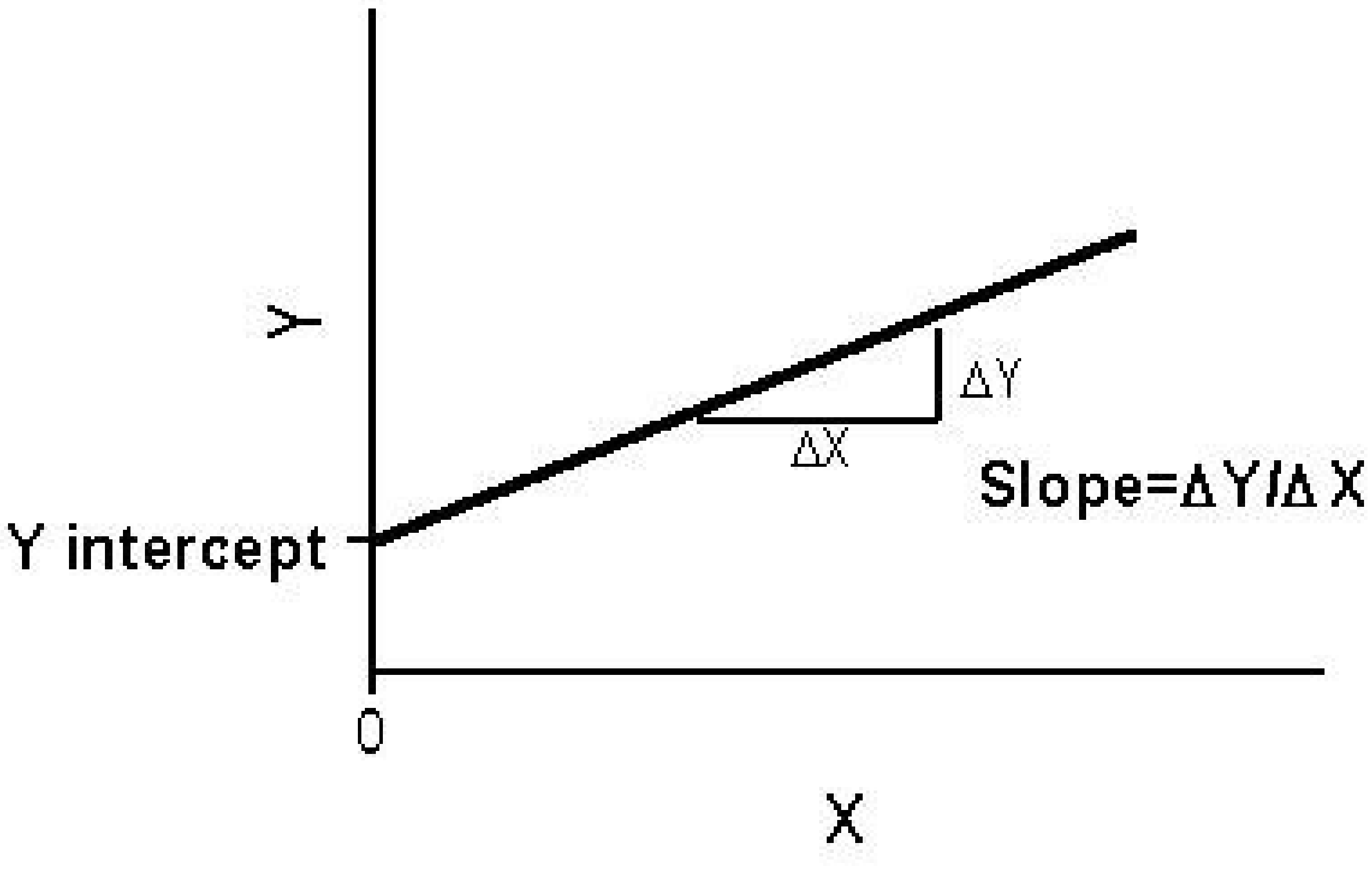
Linear model – Introduction

- Regression problem is to learn learn a function estimator $f: X \rightarrow R$ from examples $(x_i, f(x_i))$.
- The differences between actual and estimated function values on training examples are called residuals $\epsilon_i = f(x_i) - f'(x_i)$.
- The least-square method by Carl Friedrich Gauss consists in finding f' such that the sum of squared residuals is minimised.

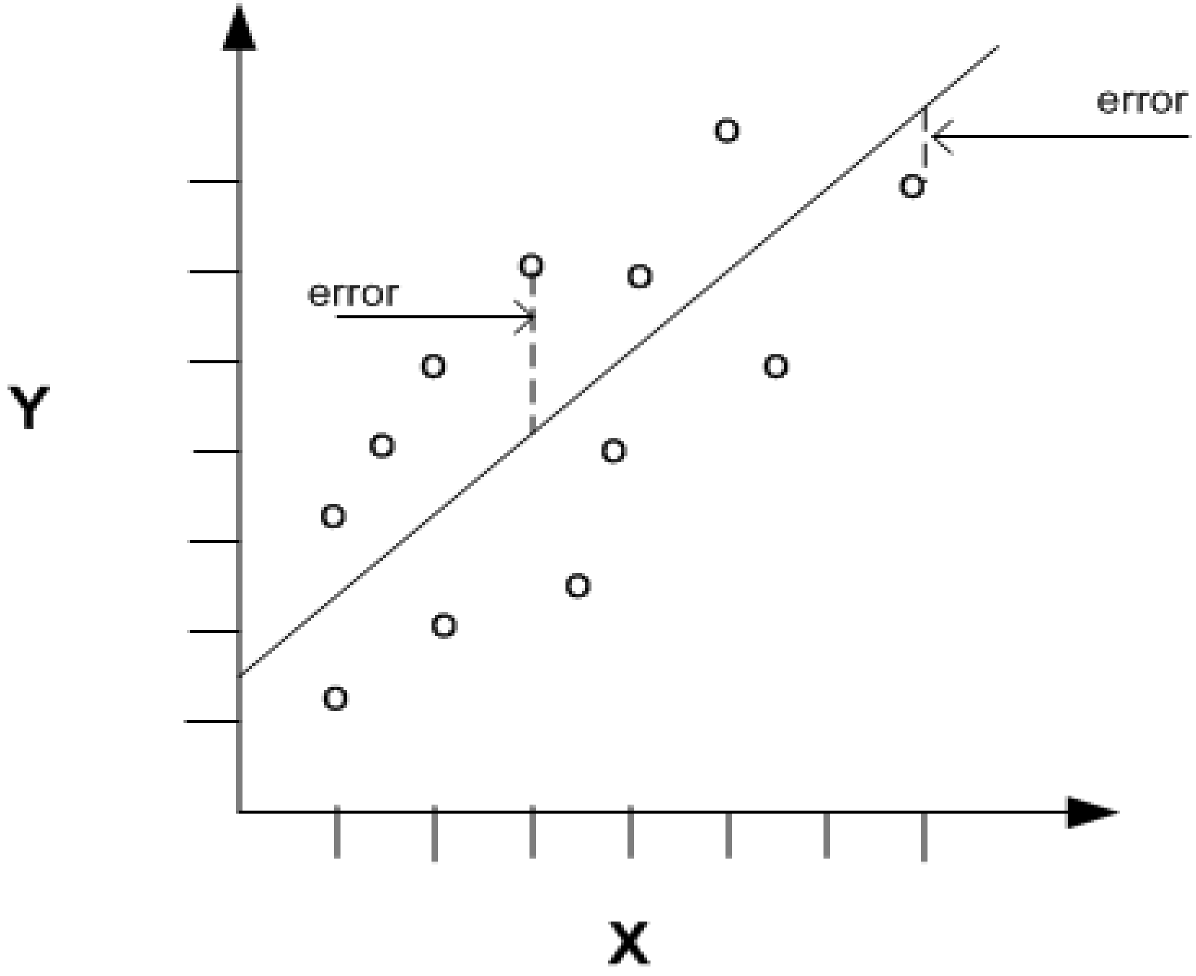
Linear regression

- Line is created by the equation y = a + bx
  - where b is the slope of the line, and a is the intercept i.e. where the line cuts the y axis.
- Suppose we have a dataset which is strongly correlated and so exhibits a linear relationship, how would we draw a line through this data so that it fits all points best?
- We use the principle of least squares, draw a line through the dataset so that the **sum of squares of the deviations of all points from the line is minimised.**

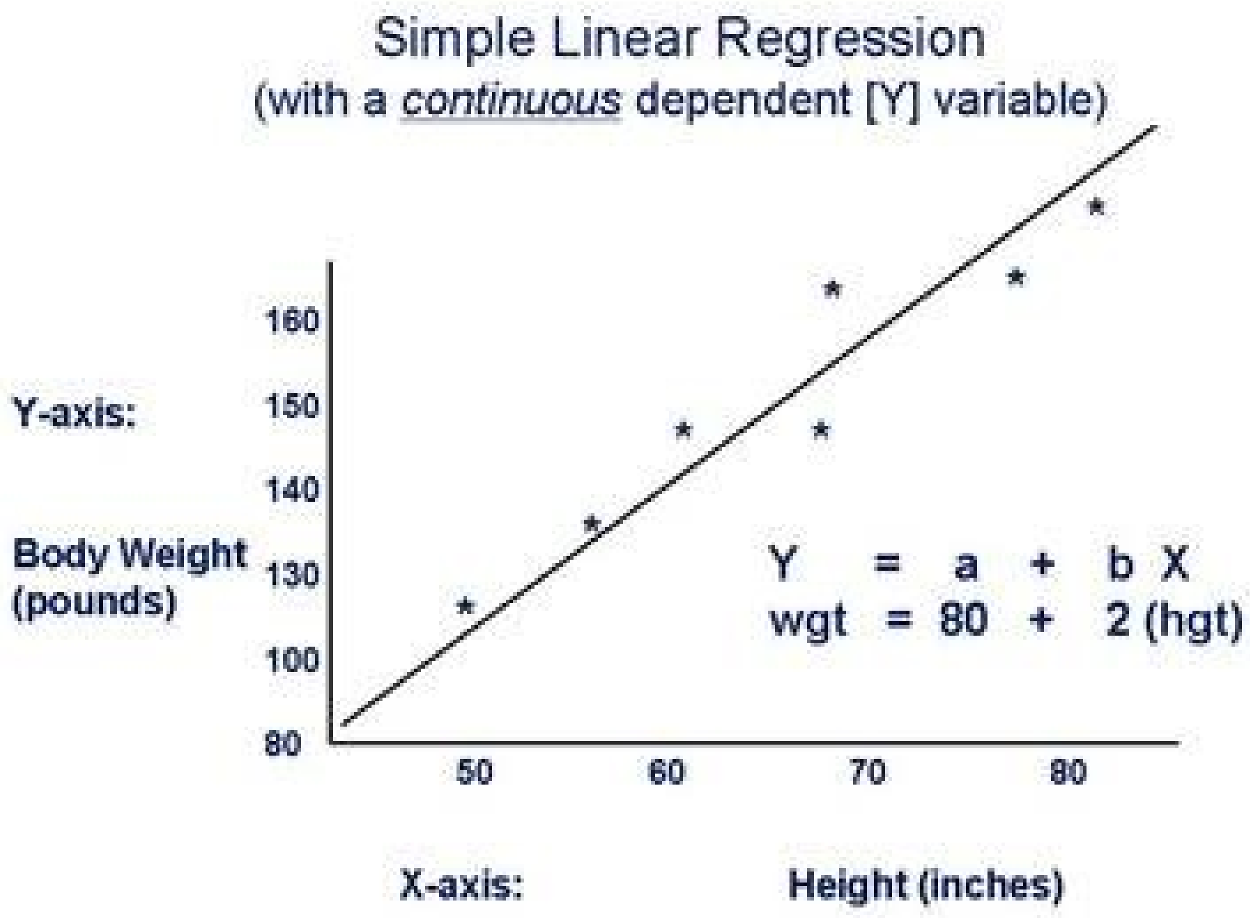
Linear regression – slope(b), a



Linear regression – error



Linear regression: Y'=a+bX



Linear regression

- For each point in the dataset, `y-(a+bx)` measures the vertical deviation (vertical distance) from the point to the line.
- For points above the line, `y-(a+bx)`will be positive.
- For points below the line, `y-(a+bx)`will be negative.
- Square these deviations to make them all positive.
- Calculate `[y-(a+bx)]2` for each point (x,y), and add them up, we get the sum of the squared distances of all the points from the line.

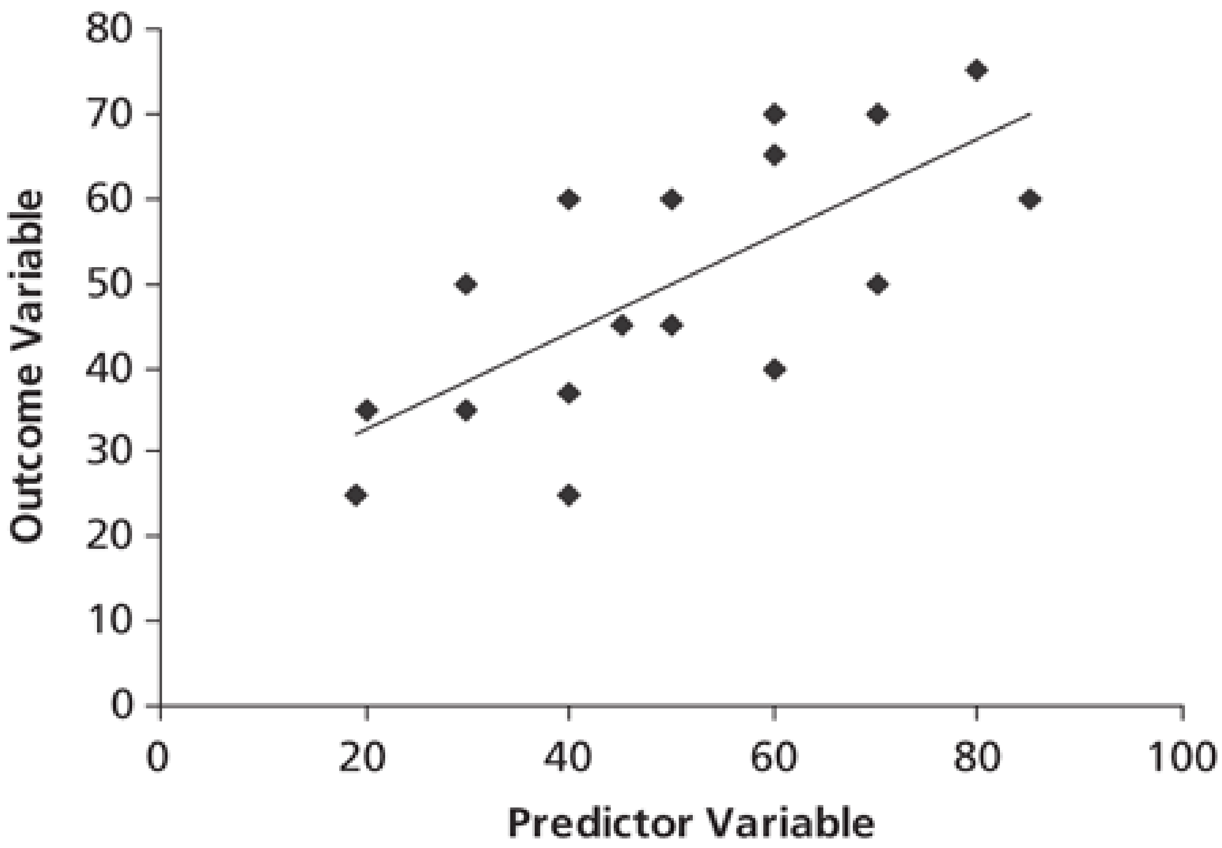
Linear regression

- The line which minimises this sum of squared distances is the line which fits the data best and we call it Least Squares Line.
- What is Linear Regression, what does it tell you?
- Linear regression uses the fact that there is a statistically significant correlation between two variables to allow you to make predictions about one variable based on your knowledge of the other.
- For linear regression to work there needs to be a linear relationship between the variables.

Linear regression

- A simple linear regression, predict scores on one variable from the scores on a second variable.
- The variable we are predicting is called the criterion variable and is referred to as Y.
- The variable we are basing our predictions on is called the predictor variable and is referred to as X.
- When there is only one predictor variable, the prediction method is called simple regression.

Linear regression

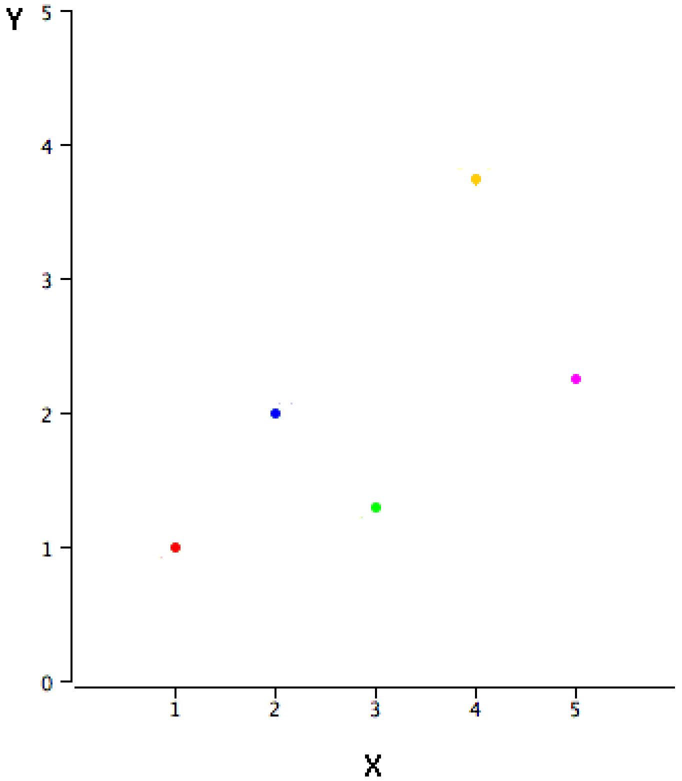


Linear regression – Ex:1

- In simple linear regression, the predictions of Y when plotted as a function of X form a straight line.

Table 1. Sample data.

X	Y
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25



Linear regression – Ex:1

- The formula for a regression line is $Y' = bX + A$

  where Y' is the predicted score, b is the slope of the line, and A is the Y intercept.

- The equation for the line in Figure 2 is $Y' = 0.425X + 0.785$
- For $X = 1$, $Y' = (0.425)(1) + 0.785 = 1.21$.
- For $X = 2$, $Y' = (0.425)(2) + 0.785 = 1.64$.
- For $X = 3$, $Y' = (0.425)(3) + 0.785 = 2.06$.

  …....

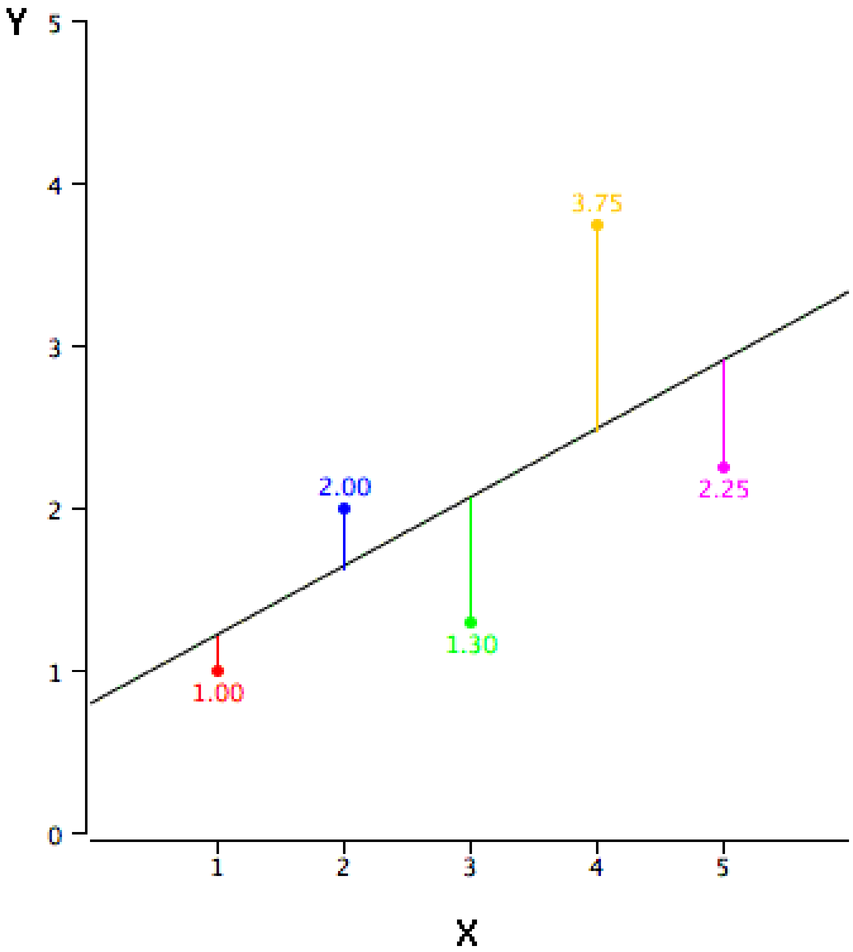
Linear regression – Ex:1

- The error of prediction for a point is the value of the point minus the predicted value (the value on the line).
- Table 2 shows the predicted values (Y') and the errors of prediction (Y-Y').

X	Y	Y'	Y-Y'	sqr(Y-Y')
1.00	1.00	1.210	-0.210	0.044
2.00	2.00	1.635	0.365	0.133
3.00	1.30	2.060	-0.760	0.578
4.00	3.75	2.485	1.265	1.600
5.00	2.25	2.910	-0.660	0.436

best-fitting line is the line that minimizes the sum of the squared errors of prediction

Linear regression – Ex:1



- The black diagonal line is the regression line and consists of the predicted score on Y for each possible value of X.
- The red point is very near the regression line; its error of prediction is small.
- Yellow point is much higher than the regression line; its error of prediction is large.

Linear regression – Ex:1

- MX is the mean of X, MY is the mean of Y, sX is the standard deviation of X, sY is the standard deviation of Y, and r is the correlation between X and Y.
- Table 3. Statistics for computing the regression line.

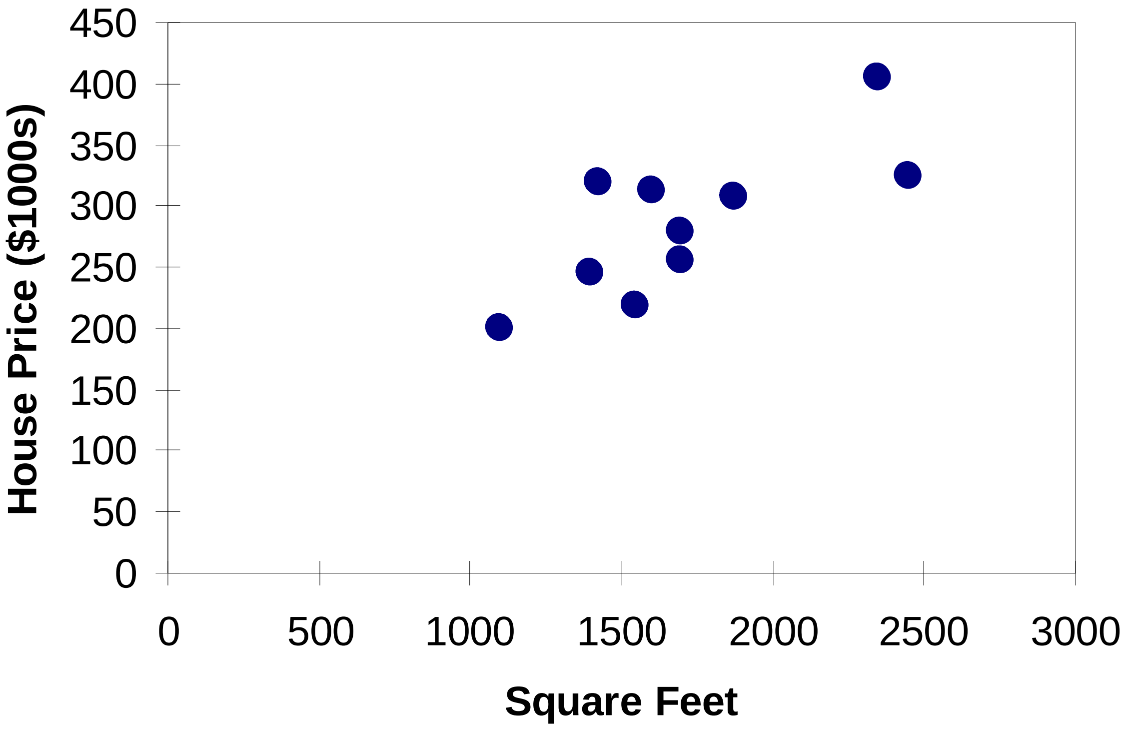
MX	MY	sX	sY	r
3	2.06	1.581	1.072	0.627

- The slope (b) can be calculated as follows: b = r sY/sX
- and the intercept (A) can be calculated as A = MY – b MX
- b = (0.627)(1.072)/1.581 = 0.425
- A = 2.06 - (0.425)(3) = 0.785

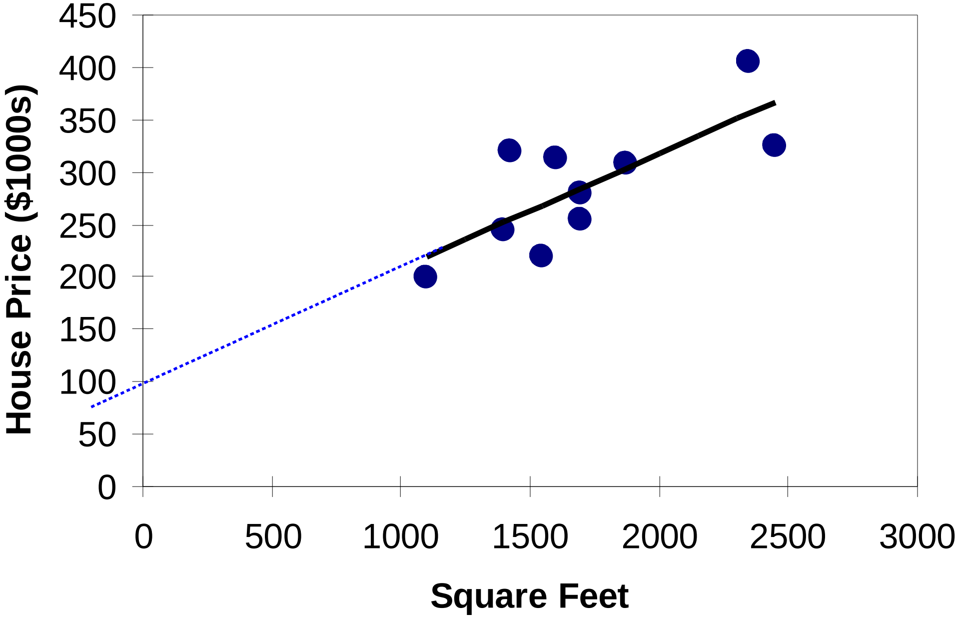
Linear regression – Ex:2

House Price in $1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

House Price: Scatter plot



House Price: Linear regression



Intercept
= 98.248

Slope
= 0.10977

$$\text{house price} = 98.24833 + 0.10977\ (\text{square feet})$$

House Price: Linear regression

Predict the price for a house with 2000 square feet:

$$\text{house price} = 98.25 + 0.1098\,(\text{sq.ft.})$$

$$= 98.25 + 0.1098(2000)$$

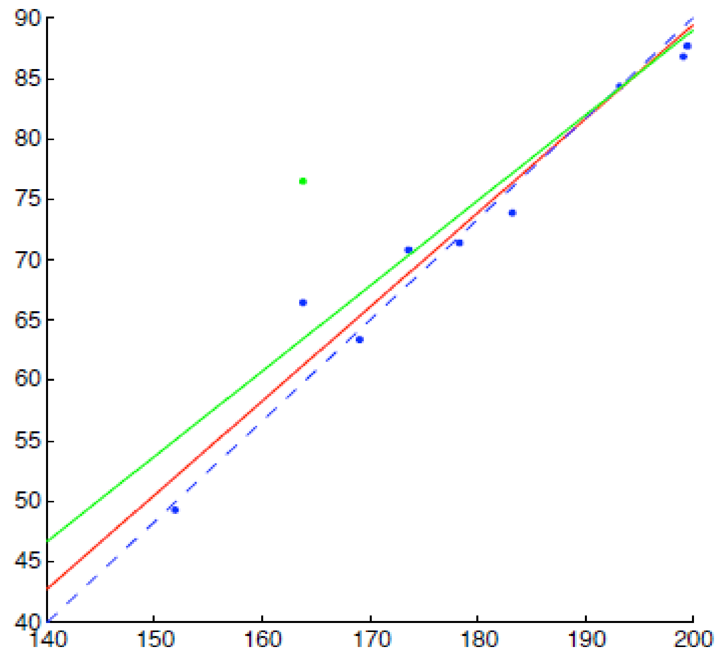
$$= 317.85$$

The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,850

Outliers

- After a regression line has been computed for a group of data, a point which lies far from the line (and thus has a large residual value) is known as an outlier.
- Such points may represent erroneous data, or may indicate a poorly fitting regression line.
- These points have may have a significant impact on the slope of the regression line.
- Depending on their location may have a major impact on the regression line.

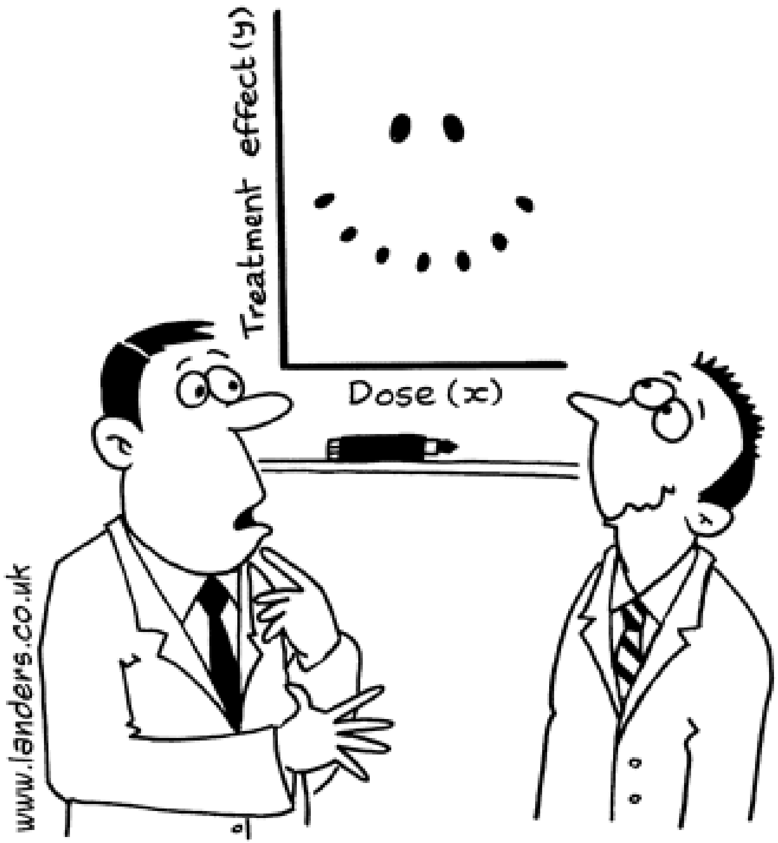
Outliers



One of the blue points got moved up 10 units to the green point, changing the red regression line to the green line.

Summary

- ...



"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."