Regression

Dr Muhammad Atif Tahir
Professor
NUCES Fast

Regression versus Classification

- Classification: the output variable takes class labels
- Regression: the output variable takes continuous values

Examples

- Predicting House Value
  - Actual Price: £100,000
  - Predicted 1: £99,950 (Very Good Prediction)
  - Predicted 1: £50,000 (Very Bad Prediction)

- Predicting Car Premium
  - Using Location, Age, History etc

Regression Techniques

- Linear Regression
- Ridge Regression
- Logistic Regression
- Lasso Regression
- And many more

Linear Regression

- Theoretically well motivated algorithm: developed from Statistical Learning Theory
- Empirically good performance: successful applications in many fields (stock prices, insurance etc)

Given examples $(x_i, y_i)_{i=1\ldots n}$

Predict y_{n+1} given a new point x_{n+1}

Formula

$$Y = a + bX$$

where

$$b = r\,\frac{SDy}{SDx}$$

$$a = \overline{Y} - b\overline{X}$$

Another formula for Slope:

Slope = $(N\Sigma XY - (\Sigma X)(\Sigma Y)) / (N\Sigma X^2 - (\Sigma X)^2)$

Where,

b = The slope of the regression line

a = The intercept point of the regression line and the y axis.

$\overline{X}$ = Mean of x values

$\overline{Y}$ = Mean of y values

SD_x = Standard Deviation of x

SD_y = Standard Deviation of y

Example

X Values	Y Values
60	3.1
61	3.6
62	3.8
63	4
65	4.1

Find Y if X = 64

To Find,

Least Square Regression Line Equation

Solution :

Step 1 :

Count the number of given x values.
N = 5

Step 2 :

Find XY, X^2 for the given values.
See the below table

X Value	Y Value	X*Y	X*X
60	3.1	60 * 3.1 =186	60 * 60 = 3600
61	3.6	61 * 3.6 = 219.6	61 * 61 = 3721
62	3.8	62 * 3.8 = 235.6	62 * 62 = 3844
63	4	63 * 4 = 252	63 * 63 = 3969
65	4.1	65 * 4.1 = 266.5	65 * 65 = 4225

Step 3 :

Now, Find ΣX, ΣY, ΣXY, ΣX^2 for the values

$\Sigma X = 311$

$\Sigma Y = 18.6$

$\Sigma XY = 1159.7$

$\Sigma X^2 = 19359$

Step 4

Substitute the values in the above slope formula given.

$$\text{Slope}(b) = (N\Sigma XY - (\Sigma X)(\Sigma Y)) / (N\Sigma X^2 - (\Sigma X)^2)$$

$$= ((5)*(1159.7)-(311)*(18.6))/((5)*(19359)-(311)^2)$$

$$= (5798.5 - 5784.6)/(96795 - 96721)$$

$$= 0.18783783783783292$$

Step 5 :

Now, again substitute in the above intercept formula given.
Intercept(a) = $(\Sigma Y - b(\Sigma X)) / N$
= (18.6 - 0.18783783783783292(311))/5
= -7.964

Step 6 :

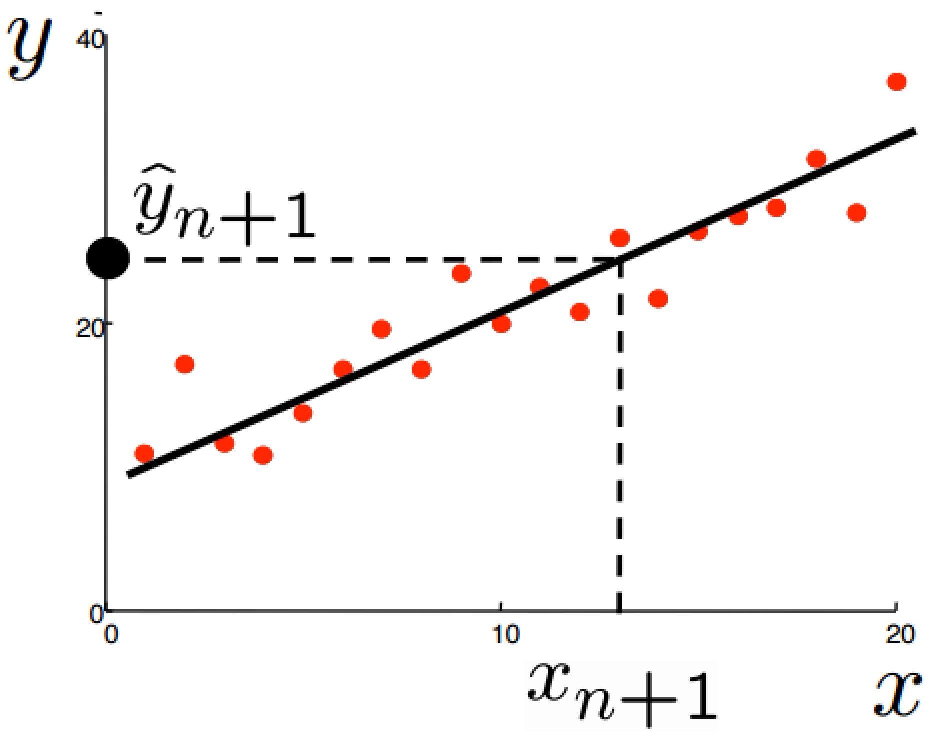
Then substitute these values in regression equation formula
Regression Equation(y) = a + bx
= -7.964 + 0.188x
Suppose if we want to calculate the approximate y value for the variable x = 64 then, we can substitute the value in the above equation
Regression Equation(y) = a + bx
= -7.964 + 0.188(64)
= 4.068

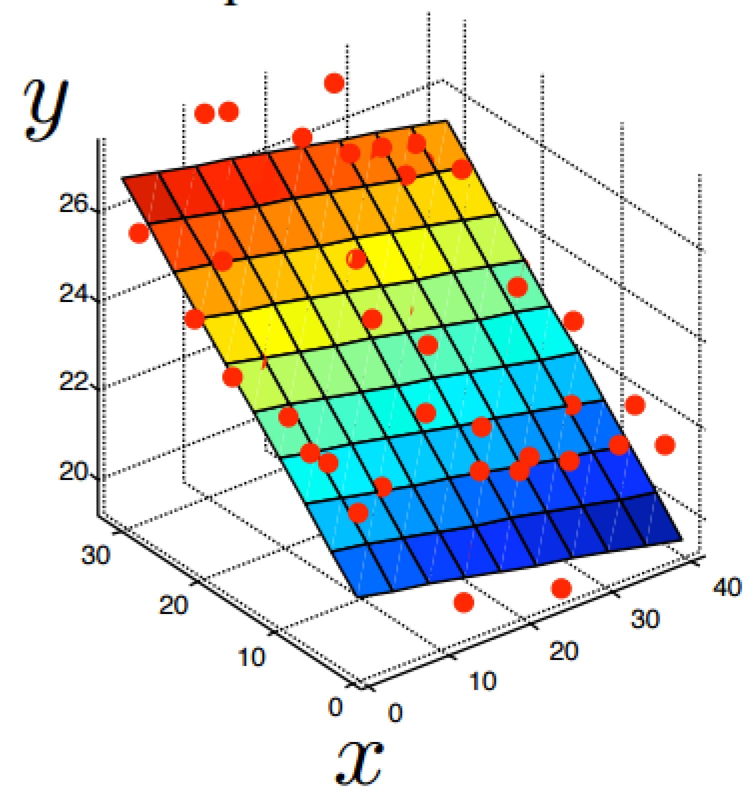
Linear regression

We wish to estimate $\hat{y}$ by a linear function of our data x :

$$
\begin{aligned}
\hat{y}_{n+1} &= w_0 + w_1 x_{n+1,1} + w_2 x_{n+1,2} \\
&= w^\top x_{n+1}
\end{aligned}
$$

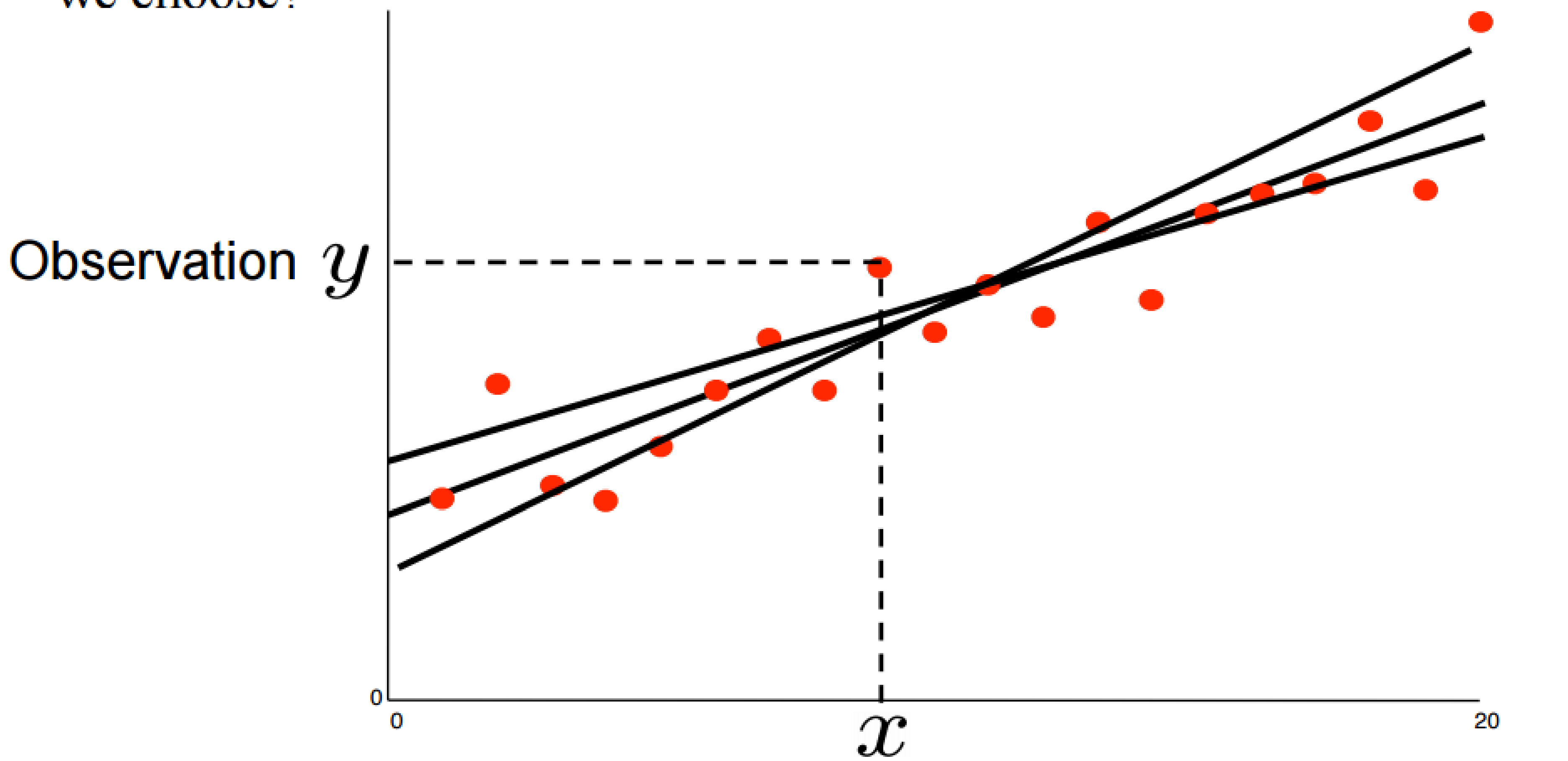
where w is a parameter to be estimated and we have used the standard convention of letting the first component of x be 1.





Choosing the regressor

Of the many regression fits that approximate the data, which should we choose?



Evaluation Measure

- Mean Squared Error

Actual (Y)	Predicted (Y')	Y'-Y	Square (Y'-Y)
41	43.6	2.6	6.76
45	44.4	-0.6	0.36
49	45.2	-3.8	14.44
47	46	-1	1
44	46.8	2.8	7.84

Sum of Error = 30.4 /5 = 6.08

Regression Techniques in Python

- Linear Least Square
- Ridge
- Lasso

http://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html

References

- https://people.eecs.berkeley.edu/~jordan/courses/294-fall09/lectures/regression/slides.pdf
- https://www.easycalculation.com/analytical/learn-least-square-regression.php

Questions!