Linear Regression

as Supervised Machine Learning

Regression

Statistical method used to determine the strength and the character of the relationship between

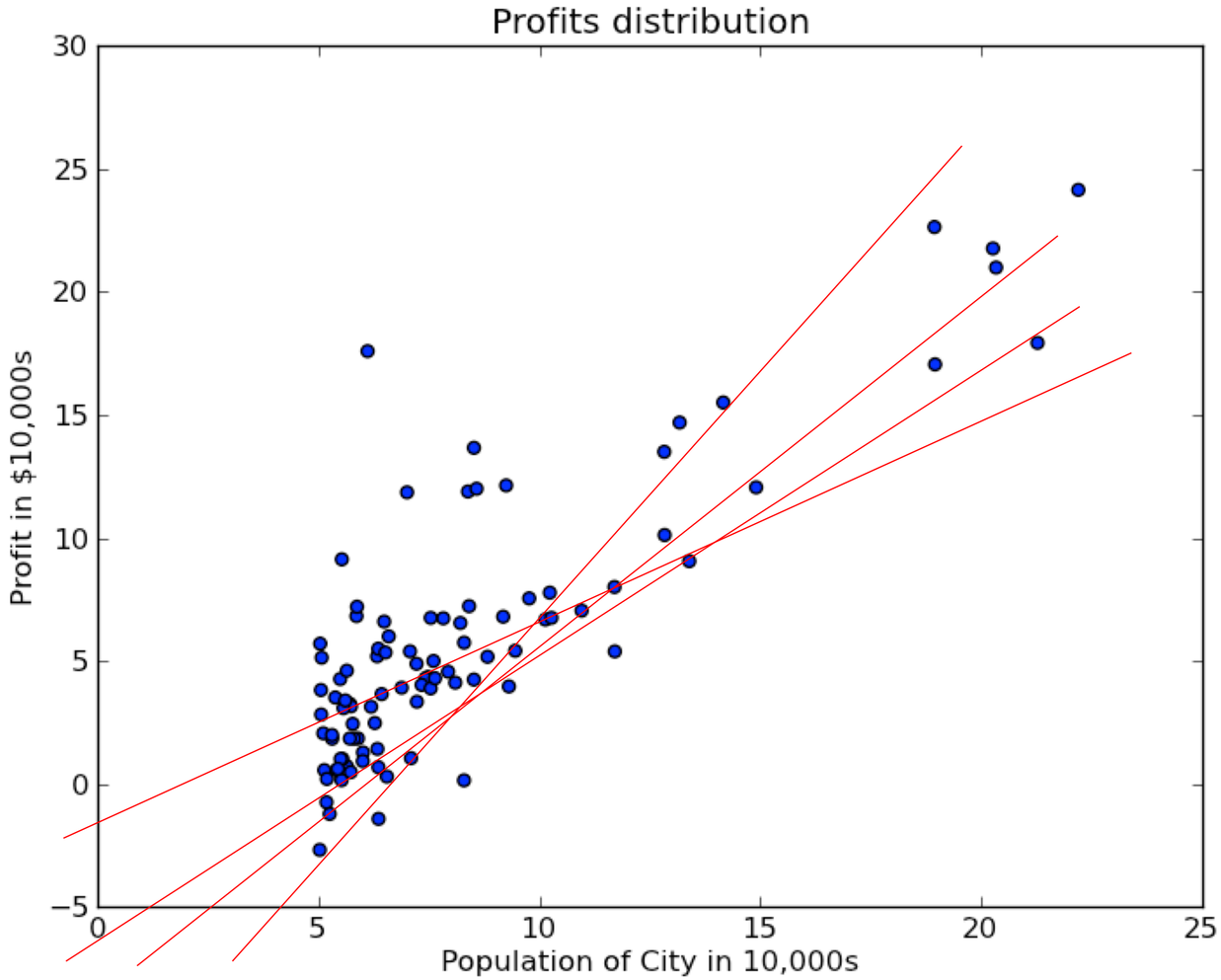
one dependent variable , Y, and

one or a series of other variables, known as independent variables, usually denoted by {X}

- Can be implemented for prediction of the output for a known input

Which line fits best?

We need a criteria for choosing the best

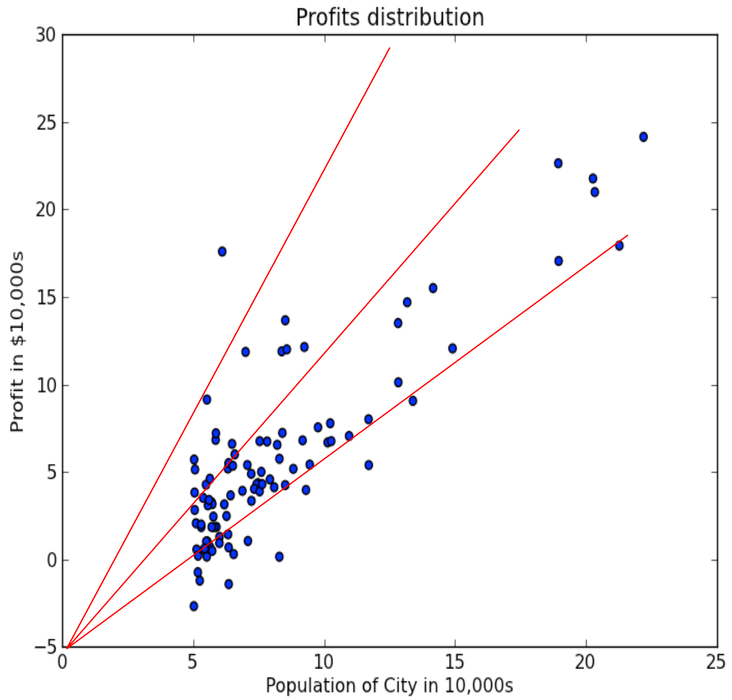


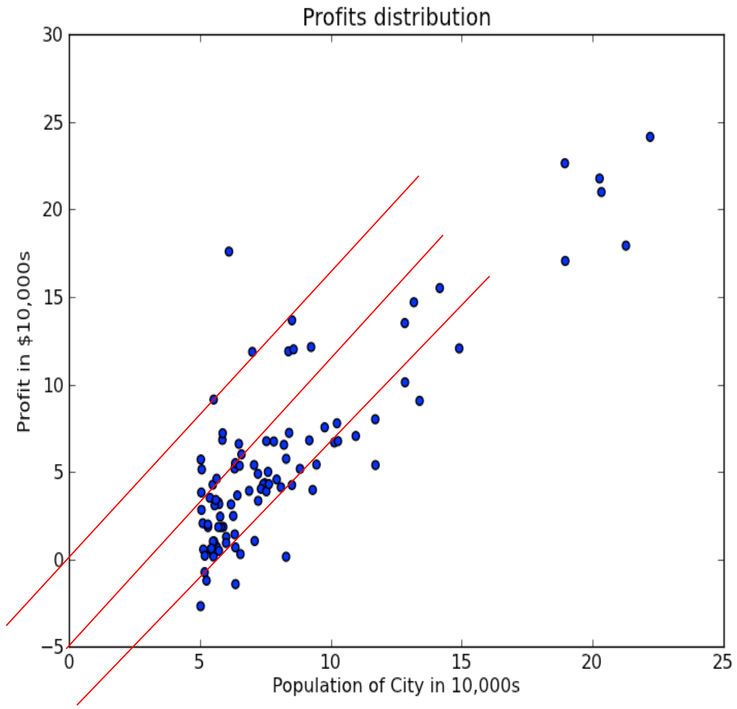
Intercept and Scope

Lines differ in

- angle towards the X axis
- crossing point location on Y axis

The line equation contains both





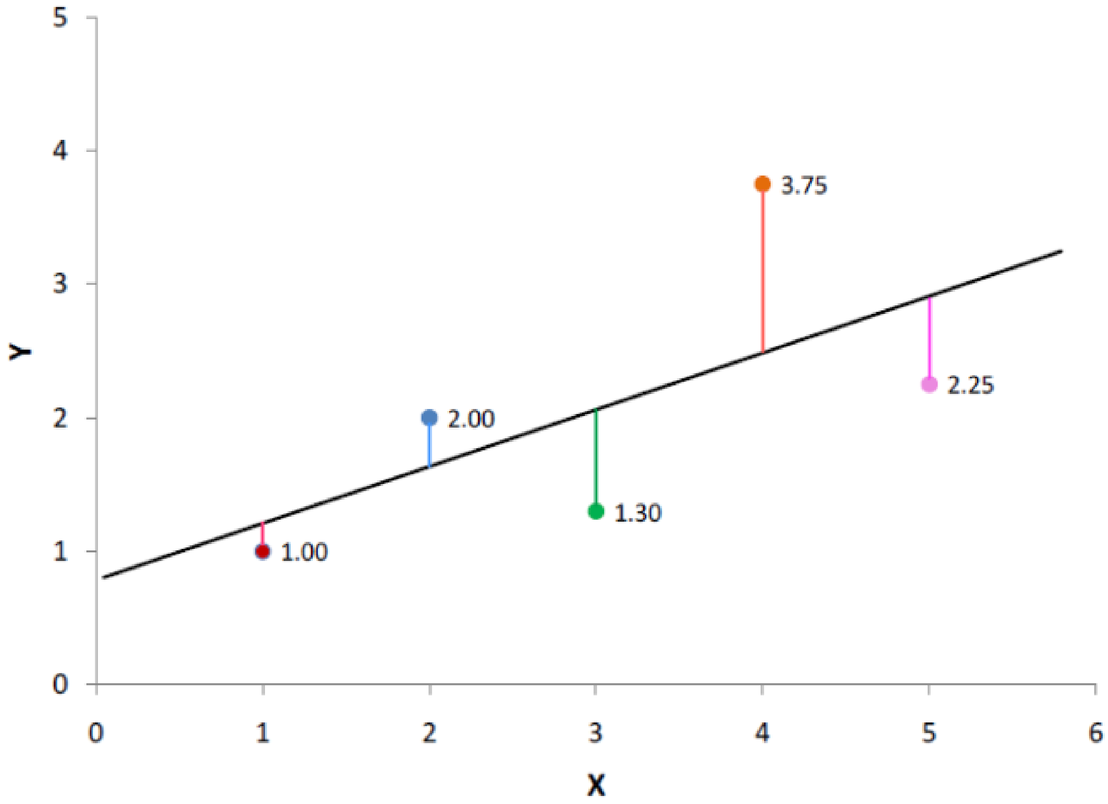
Best Fitting Line

The coefficients a and b would be such that produce the minimal or no error of calculation of Y out of X

From

$$Y_i = a*X_i + b + e_i$$

$$e_i = Y_i - a*X_i - b$$



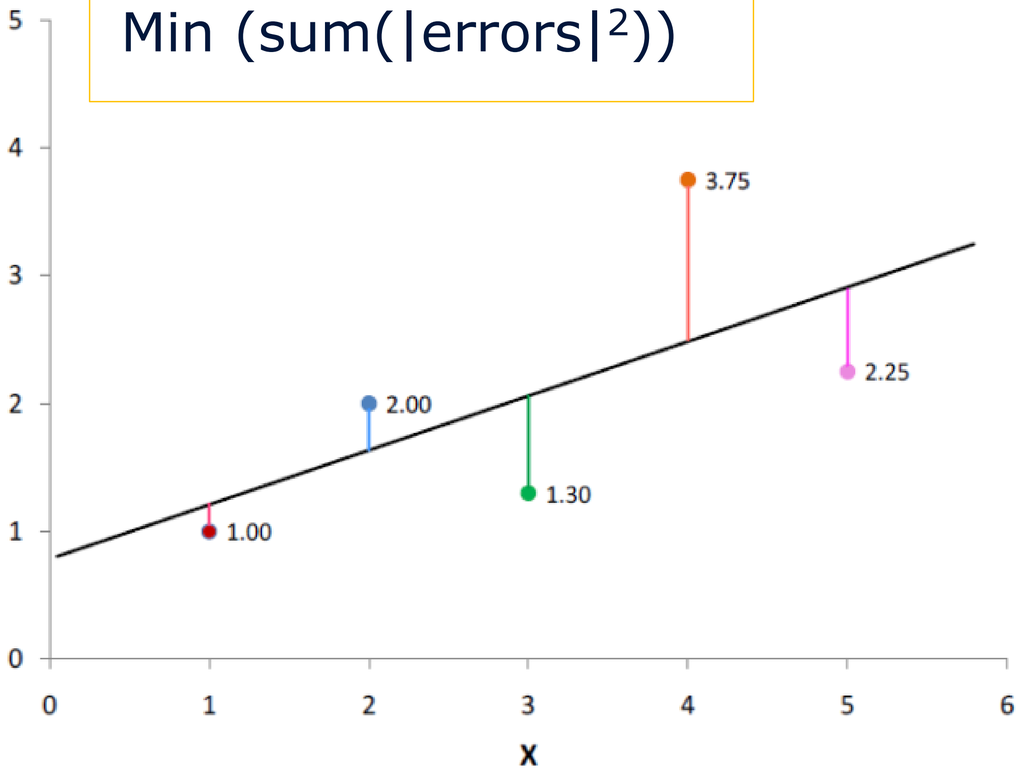
Best Fitting Line

The best-fitting line minimizes the distances of our data points to the line

Min (sum(errors))

Min (sum(|errors|))

Min (sum($|errors|^2$))



Mean Absolute Error (MAE) is the mean of the absolute value of the errors:

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Mean Squared Error (MSE) is the mean of the squared errors:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Root Mean Squared Error (RMSE) is the square root of the mean of the squared errors:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Quiz:

What are the best b and a?

x	y
3	0
6	-3
4	-1
5	-2

$$y = a^* x + b$$

Quiz:

What are the best fitting line between x and y?

x	y
3	0
6	-3
4	-1
5	-2

$$y = a * x + b$$

-1 : slope

+3 : intercept

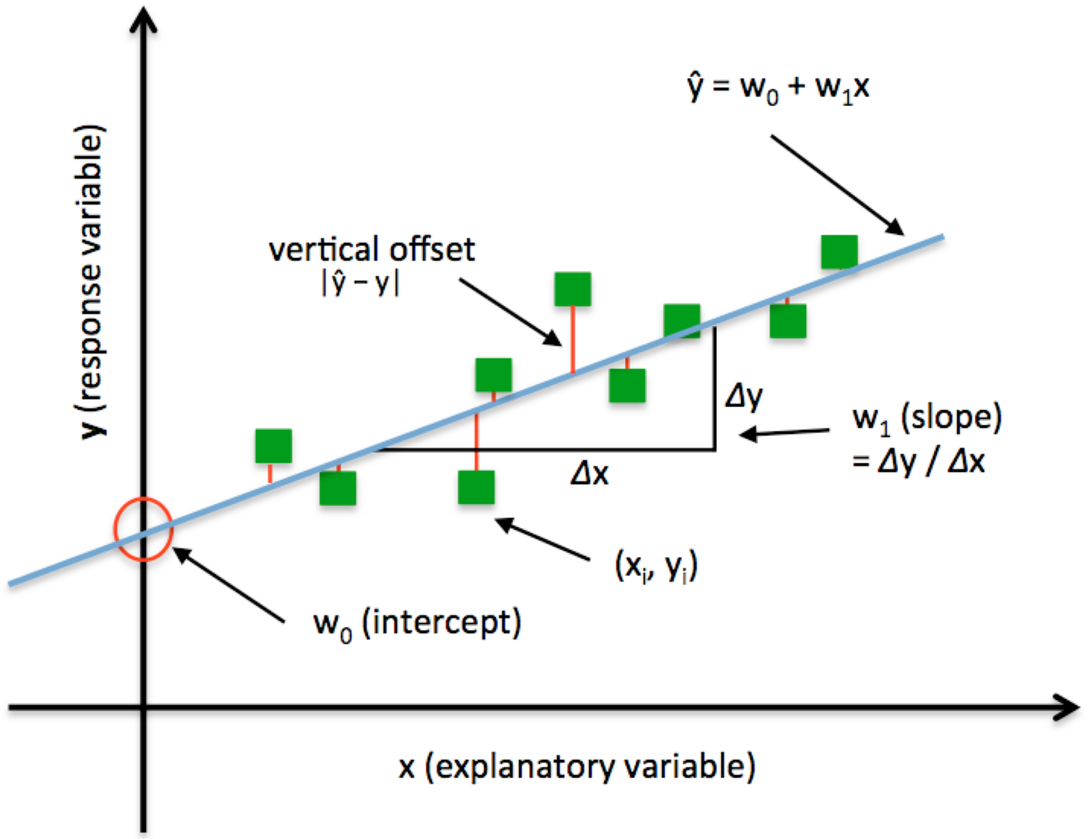
Regression Line

The equation of a line

$$Y = a*X + b$$

Regression uses the best-fitting straight line through the points on a scatter plot to predict how X causes Y to change

X is the input, it is known
if we also know a and b, we can calculate the output Y
the machine learning analysis needs to discover a and b



w0 – intercept (b)
w1 – slope (a)
Δy – errors, *residuals*

Regression Example

https://www.youtube.com/watch?v=IpGxLWOIZy4

Example

Can I sell or buy this house?

How to estimate the right price for it?

Can base my estimation on the available information!

Hypothesis

bigger the house – more expensive it it
- X – size
- Y - price

This is a linear function!

$$Y_i = f(X_i, \beta) + e_i$$

Y_i = dependent variable

f = function

X_i = independent variable

β = unknown parameters

e_i = error terms

I need a couple of observations on the market

- sizes vs prices

Example

- vector {X} of sizes
- vector {Y} of prices

$$Y_i = f(X_i, \beta) + e_i$$

Y_i = dependent variable

f = function

X_i = independent variable

β = unknown parameters

e_i = error terms

More specific:

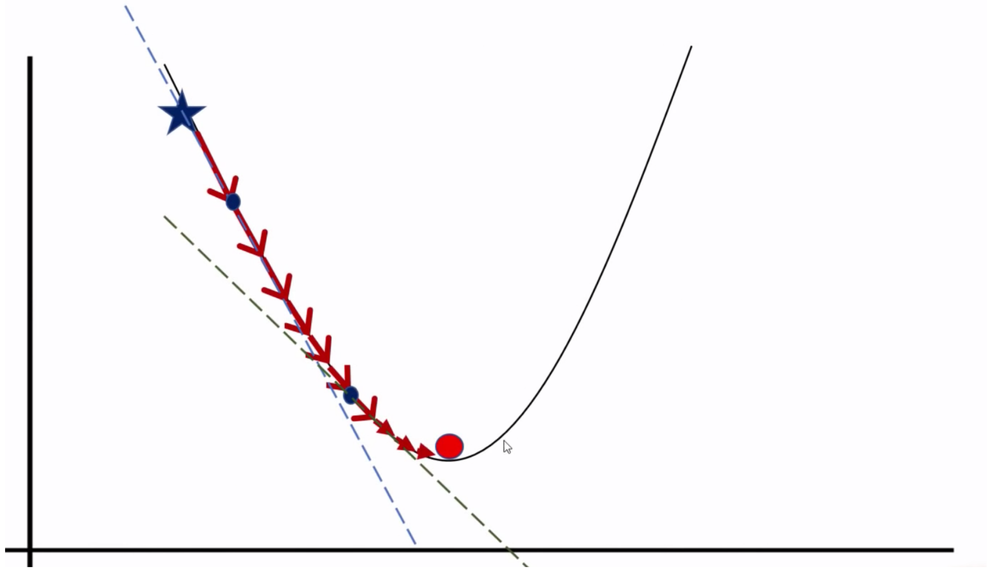
$$Y_i = f(X_i, a, b) + e_i$$

$$Y_i = a*X_i + b + e_i$$

How To Validate the Model?

Model Training

- Gradient Descent
  - iterative optimization algorithm
  - start with random values, predict and adjust

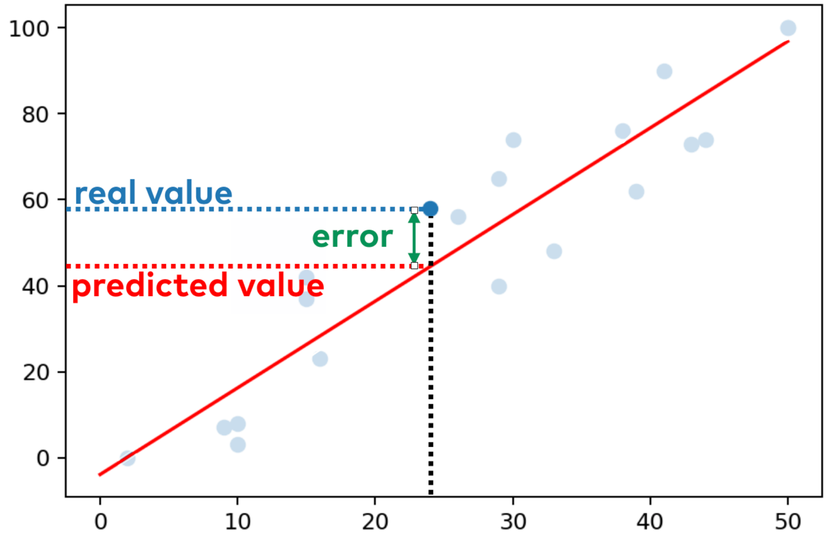


Model Validation

ERROR = ACTUAL – PREDICTED

Loss function

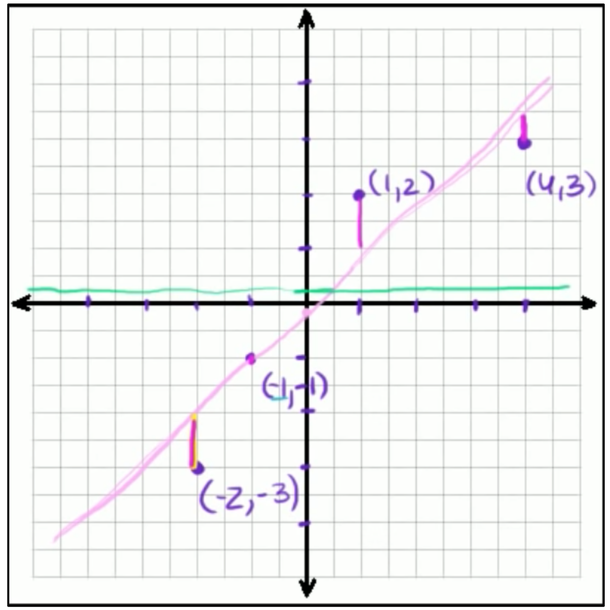
- calculate the errors
- minimize the errors



R-squared value, also known as the coefficient of determination

- a statistical measure of how close the data is to the regression line, or how well the model fits the observations
- if the data is perfectly on the line, the r-squared value would be 1, or 100%, meaning that the model fits perfectly
- R-squared is the percentage of the variation of the dependent variable that is explained by this linear model

How to Calculate R-squared?



The errors can have two sources:

1) lack of perfect fit of the model over Y values, the model can not explain the residuals
2) the total variance of Y set itself

1. can be assessed by the squared sum of the residuals (observed - expected)

$$\text{Rss} = \sum_i (y_i - \hat{y}_i)^2$$

2. can be assessed as variance of Y – squared sum of the distance to the mean value of Y (observed – mean)

$$\text{Tss} = \sum_i (y_i - \overline{y})^2$$

The proportion

$$\frac{Rss}{Tss} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2} \qquad [\%]$$

is this part of Y error, which CANNOT be explained by this model

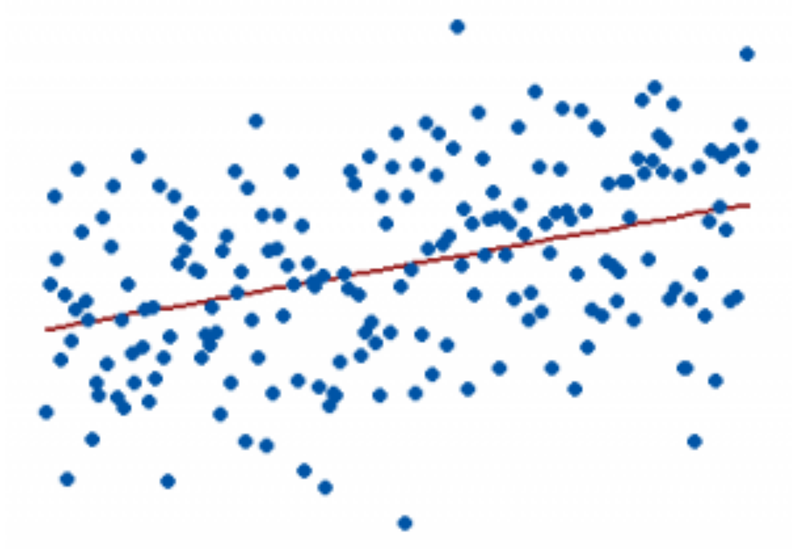
Therefore,

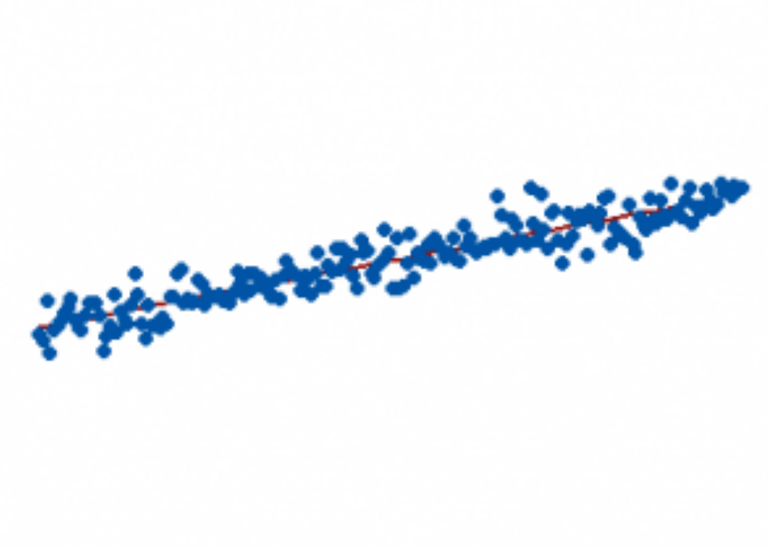
$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

is the regression error EXPLAINED by the model.

Higher the R square value, better is the predicted model!

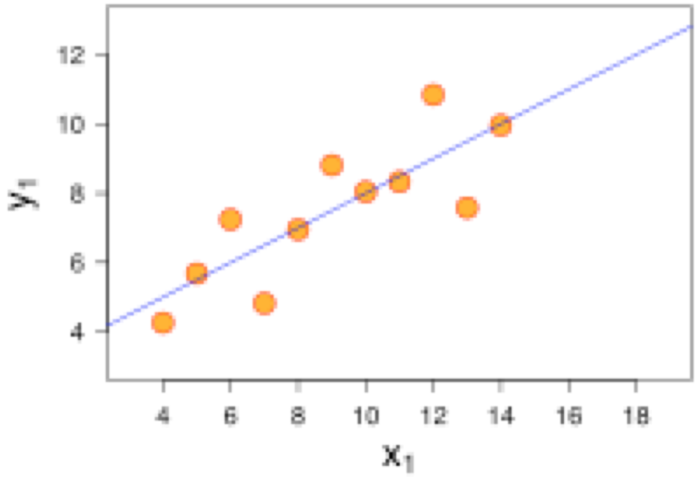
Which of these two regressions has higher R-squared coefficient?

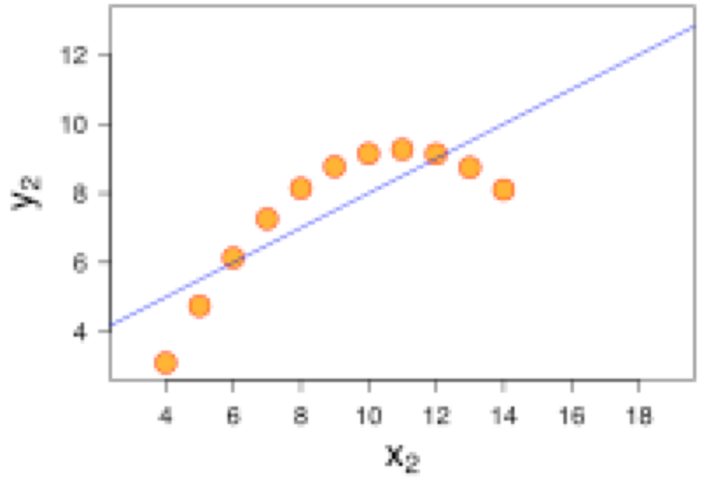


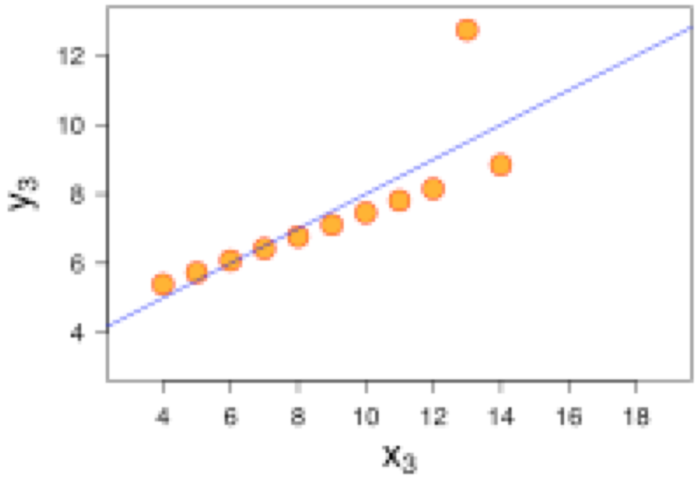


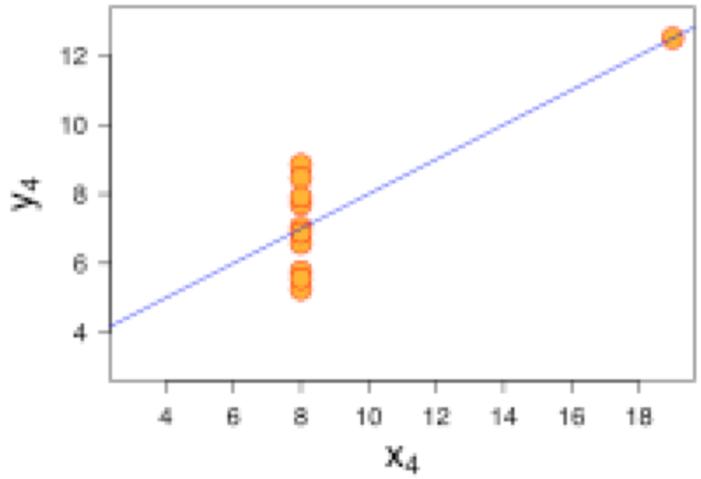
Other Types of Regression

Line is NOT Always the Best Fit





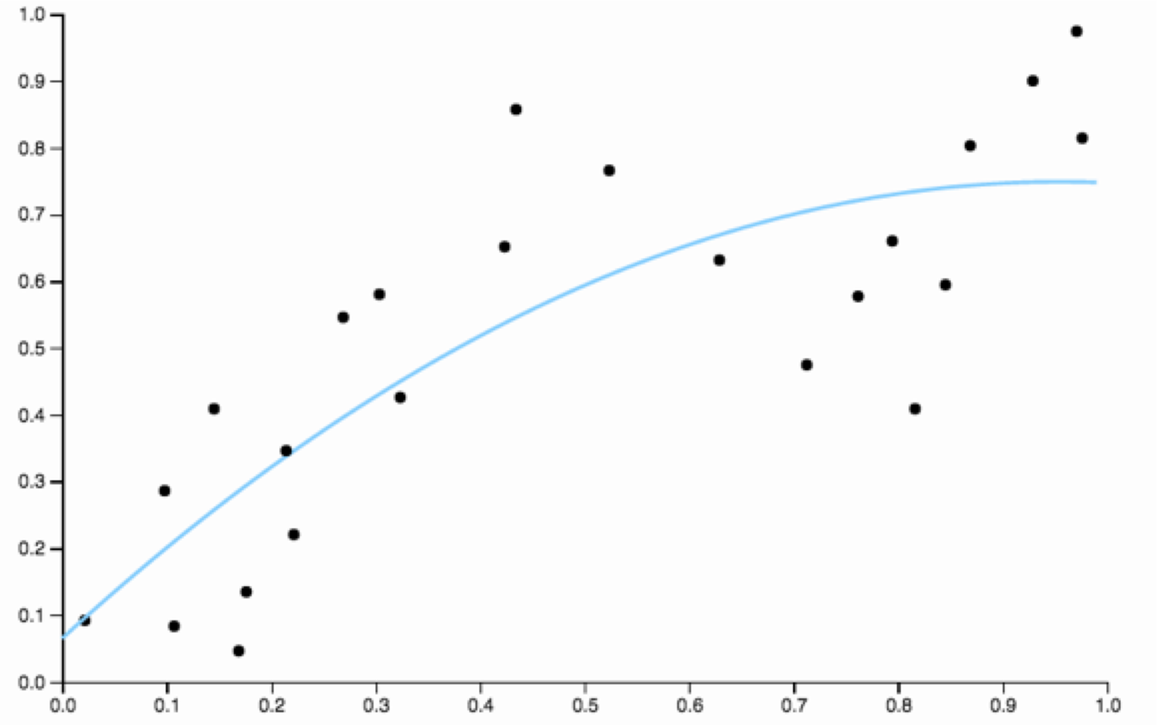




Multiple Linear Regression

Linear Regression with Multiple Independent Variables

$$Y = a_0 + a_1*X_1 + a_2*X_2 + \ldots + a_n*X_n$$



Non-Linear Regression

Non-linear functions can have elements like exponentials, logarithms, fractions, and others.

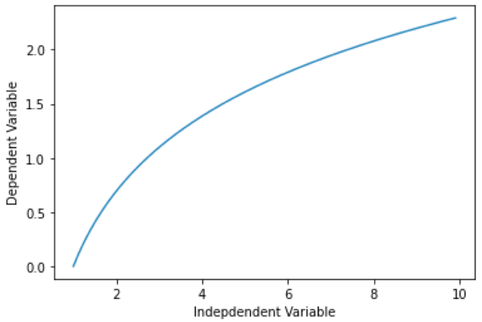
Examples:

$y=\log(x)$

$y=\log(ax3+bx2+cx+d)$

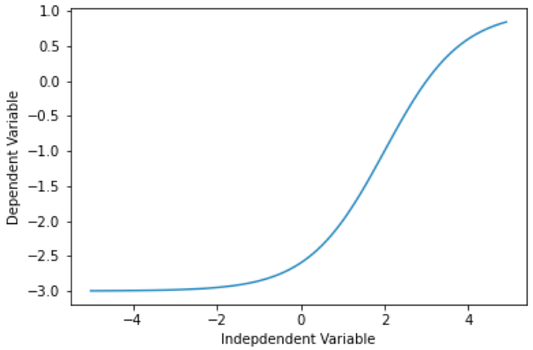
Logarithmic

y = np.log(x)



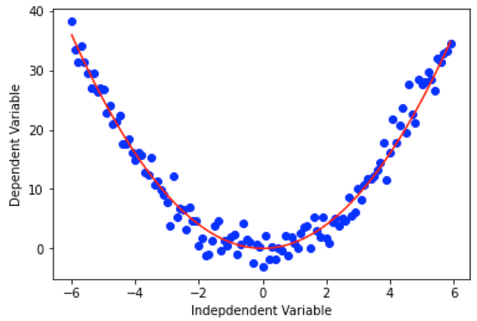
Sigmoidal

y = 1-4/(1+np.power(3, X-2))



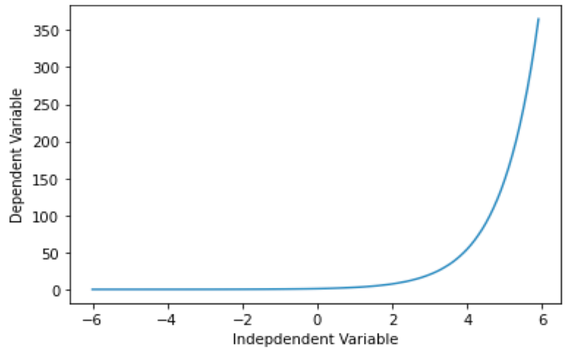
Quadratic

y = np.pow(x, 2)



Exponential

y = np.exp(x)



https://www.kaggle.com/ibrahimbahbah/non-linear-regression-tutorial/notebook

Polynomial Regression

Polynomial Regression

$$Y = 4*X^4 - 3*X^3 - X^2 - 3*X + 3$$

A form of regression in which the relationship between the observation and the outcome is modelled as an n^{th} degree polynomial

Method: we create new variables

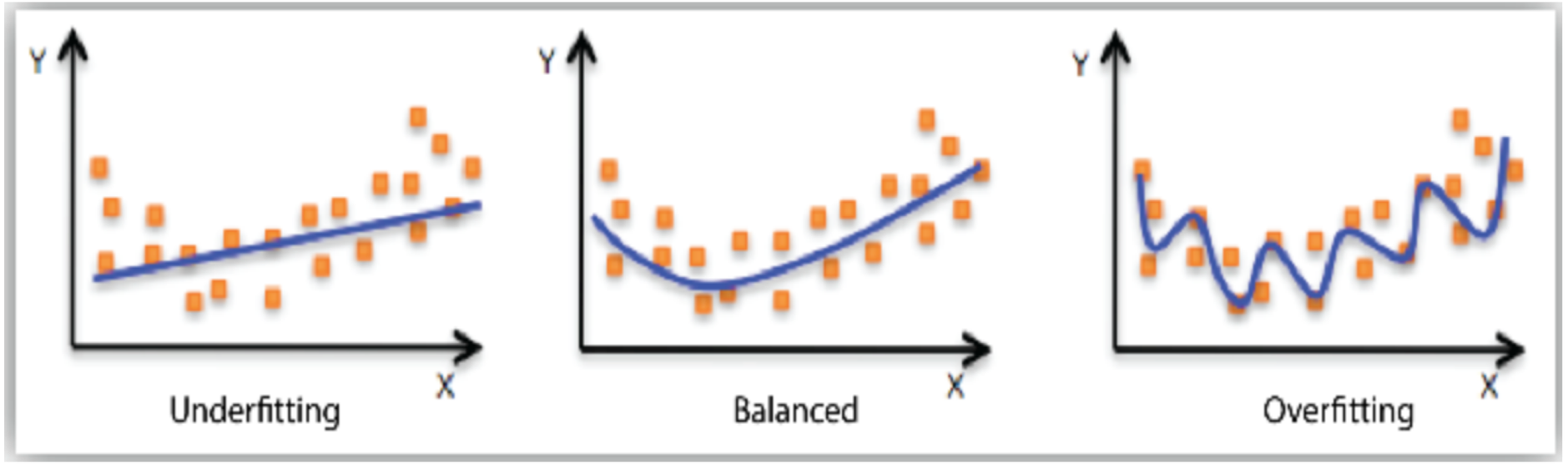
$$Z_1 = X^4$$
$$Z_2 = X^3$$
$$Z_3 = X^2$$

and replace them in the polynom

$$Y = 4*Z_1 - 3*Z_2 - Z_3 - 3*Z + 3$$

Now we can implement multiple linear regression like above.

Overfitting
vs
Underfitting



Advantages of Regression

One of the most widely used methods

- Simple and easy to understand
- Requires minimum resources
- Training of a model is much faster than other similar models

- For being meaningful, regression analysis must be applied to:
  - quantitative variables
  - without outliers
- Not useful beyond the range

Why is it called Regression?

- Sir Francis Galton, in his 1885 Presidential address before the anthropology section of the British Association for the Advancement of Science (Stigler, 1986), described a study he had made that compared the heights of children with the heights of their parents. He examined the heights of parents and their grown children, perhaps to gain some insight into what degree height is an inherited characteristic.
- He thought he had made a discovery when he found that the heights of the children tended to be more moderate than the heights of their parents. For example, if parents were very tall the children tended to be tall but shorter than their parents. If parents were very short the children tended to be short but taller than their parents were. This discovery he called "regression to the mean," with the word "regression" meaning to come back to.
- He published his results in a paper, "Regression Towards Mediocrity In Hereditary Stature," (Galton, F. (1886)).

Reference

- https://www.scribbr.com/statistics/standard-deviation/
- https://www.ritchieng.com/machine-learning-evaluate-linear-regression-model/#15.-Model-Evaluation-Metrics-for-Regression
- https://blogs.oracle.com/datascience/types-of-machine-learning-and-top-10-algorithms-everyone-should-know-v2
- https://medium.com/@erika.dauria/looking-at-r-squared-721252709098
- https://www.ritchieng.com/machine-learning-evaluate-linear-regression-model/#15.-Model-Evaluation-Metrics-for-Regression