Linear Regression

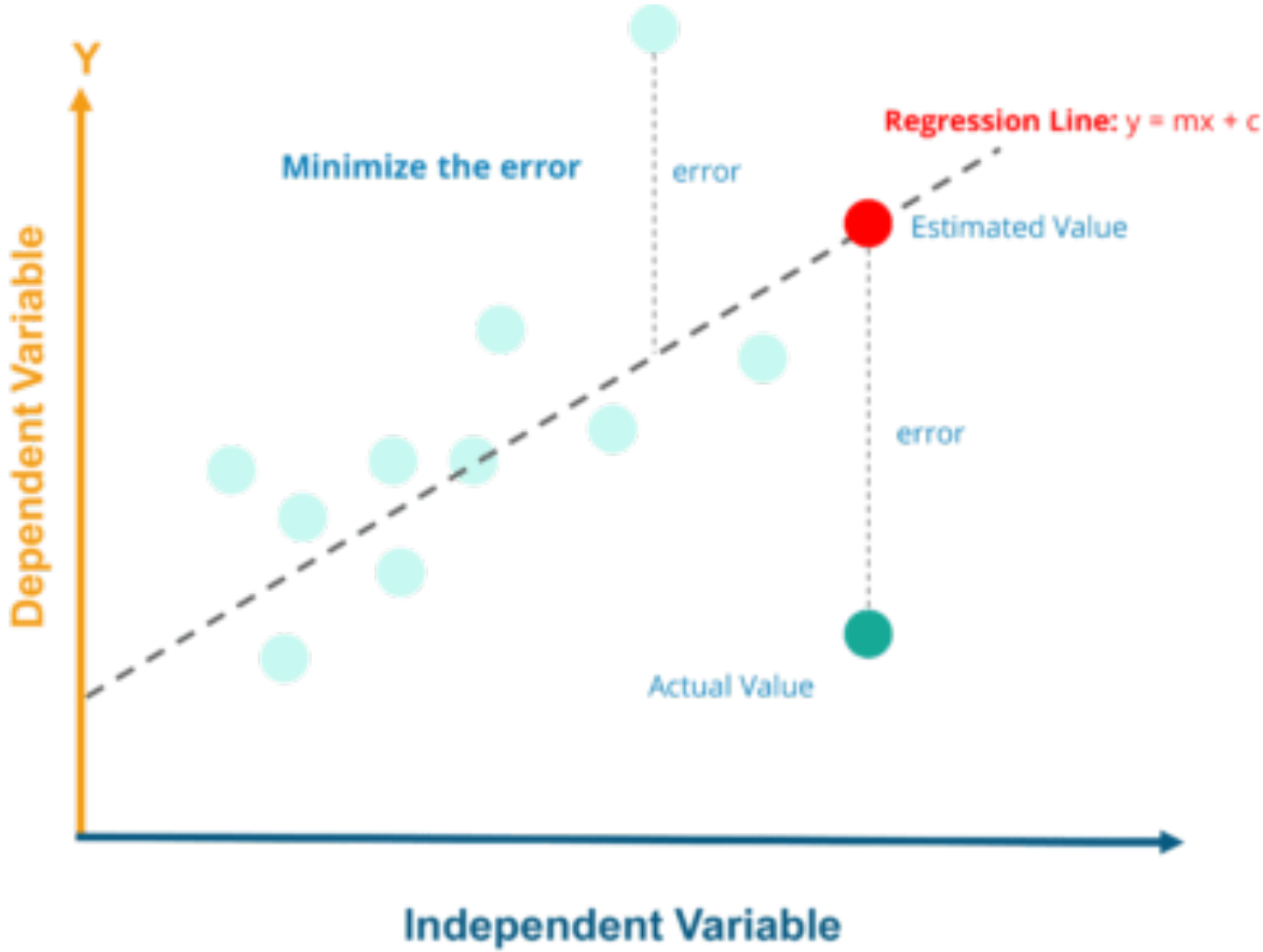
Introduction

- Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is an explanatory variable, and the other is a dependent variable.
- For our data set the independent/explanatory variable is the time and the dependent variable is the house price.
- A linear regression line has an equation of the form $Y = c + mX$, where X is the explanatory variable and Y is the dependent variable. The slope of the line is m, and c is the intercept (the value of y when $x = 0$).

Least squares method

Least squares fitting (also called least squares estimation) is a way to find the best fit curve or line for a set of points. In this technique, the sum of the squares of the offsets (residuals) are used to estimate the best fit curve or line instead of the absolute values of the offsets. The resulting equation gives you a y-value for any x-value, not just those x and y values plotted with points.

$$\mathrm{m} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$


Y
Dependent Variable
Minimize the error
error
Regression Line: y = mx + c
Estimated Value
error
Actual Value
Independent Variable

Advantages of least squares fitting

- Least squares allows the residuals to be treated as a continuous quantity where derivatives (measures of how much a function's output changes when an input changes) can be found. This is invaluable, as the point of finding an equation in the first place is to be able to predict where other points on the line (even points that are way beyond the original points) might lie.

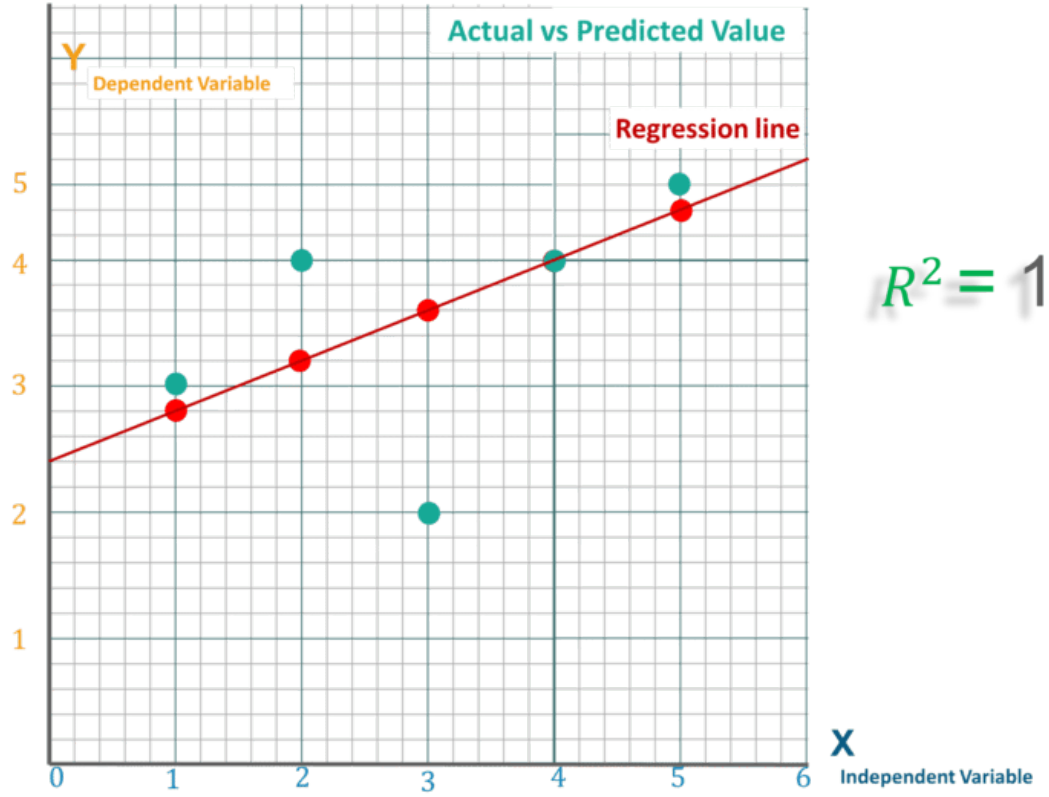
- Chi-square fitting takes categorical values which allows a test to be made of whether the variance of the population has a pre-determined value. The value of the population is not continuous as it takes discrete values.

Disadvantages of Least Squares Fitting

Outliers can have a disproportionate effect if you use the least squares fitting method of finding an equation for a curve. This is because the squares of the offsets are used instead of the absolute value of the offsets; outliers naturally have larger offsets and will affect the line more than points closer to the line. These disproportionate values may be beneficial in some cases.

R Square Method – Goodness of Fit

R–squared value is the statistical measure to show how close the data are to the fitted regression line



$$R^2 = 1 - \frac{\sum (y_p - \bar{y})^2}{\sum (y - \bar{y})^2}$$

R Square Method(cont.)

R-squared does not indicate whether a regression model is adequate. You can have a low R-squared value for a good model, or a high R-squared value for a model that does not fit the data

References

- https://www.edureka.co/blog/linear-regression-in-python/
- https://towardsdatascience.com/linear-regression-python-implementation-ae0d95348ac4
- https://machinelearningmastery.com/linear-regression-for-machine-learning/
- https://ml-cheatsheet.readthedocs.io/en/latest/gradient_descent.html
- https://www.statisticshowto.datasciencecentral.com/least-squares-regression-line/