Least squares in linear regression

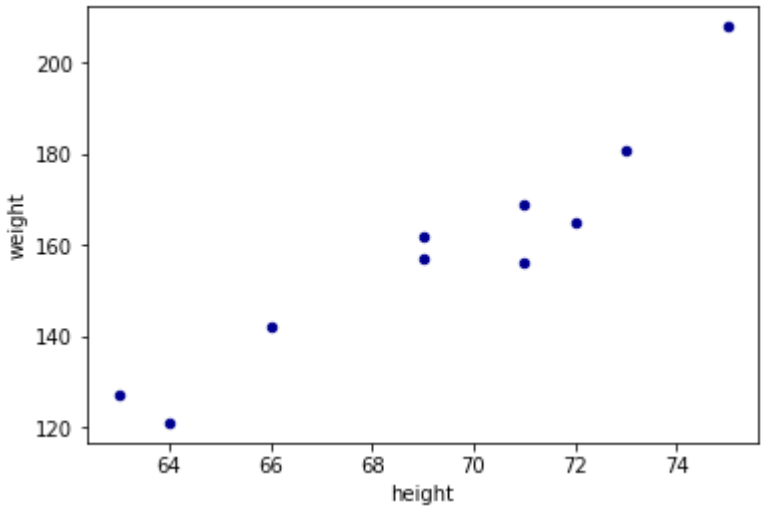
Linear regression is a traditional entry point into the extremely useful, exciting and trending area of computer science. At its core linear regression algorithms are a predictive analysis of data. In its basic form it determines the relationship between a dependent variable and an independent (more than one are possible). Some of its best use cases involve time ordered data, or just continuous variable data in general. In its most basic form, we have the equation $\hat{y}_i = b_0 + b_1 x_i$ and if you think it looks like an equation for a line you would be correct. What you are looking at is the equation for the best fit line where b_0 is some constant, b_1 is the regression coefficient, x_i is the independent variable, and $\hat{y}_i$ is the predicted value.

The data and equations

In order to do some calculations and show how all of this works we need some data, so below is 10 height/weight data points from people.

ht	wt
63	127
64	121
66	142
69	157
69	162
71	156
71	169
72	165
73	181
75	208

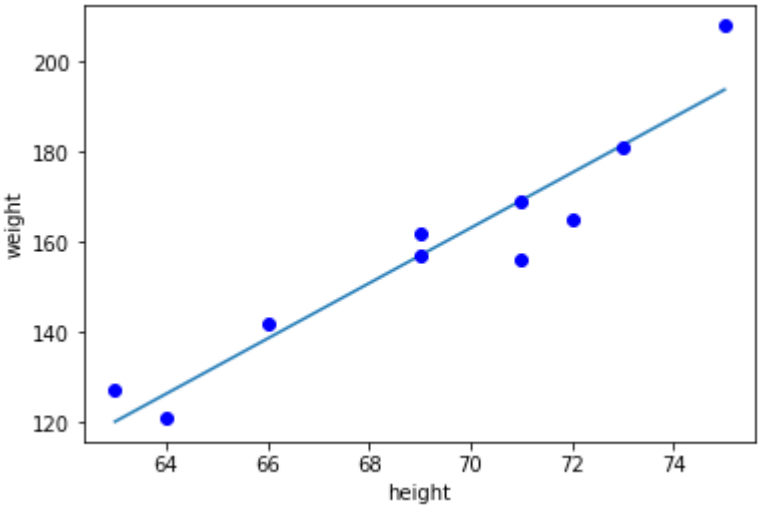
Here is what it looks like as a scatter plot



In order to find the best fit line one way is using least squares. Take the sum of squared prediction errors which the sum of Squares of $e_i = y_i - \hat{y}_i$. Where e_i is the error of the prediction, y_i is the actual value and $\hat{y}_i$ is the prediction value. We take the square because otherwise when we sum the values up the positive and negative values would cancel each other out. We do this because we want to minimize the error which we do using the equation: $Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$. Now with that equation we can find the sum of the squared error. So the next step is to find the least square estimate for b_1 and b_0 . In order to do this we take derivates for b_0 and b_1 which produces the two equations: $b_0 = y - b_1 x$, and

$$b1 = \frac{\sum_{i=1}^{n} (x_i - x)(y_i - y)}{\sum_{i=1}^{n} (x_i - x)^2} .$$

After carrying out the math and graphing both the points and best fit line we have:



Conclusion

Now for the big reason we have done all of this. After you gotten your sample data and found your best fit line you can now predict What a persons weight will be based on their height. If finding someone's predicted weight based on their height linear regression may not be that useful however the techniques used to predict what someone's weight will be based on their height can be used in all kinds of scenarios. Like predicting what a stocks price may be, how many purchases to expect in spring and so on.