Least squares in linear regression

Linear regression is a traditional entry point into the extremely useful, exciting and trending area of computer science called machine learning. Which for our basic purpose means that a service or program can "learn" from data without a programmer explicitly intervening to change things. Some of people favorite computer features such as targeted advertising and predictive text use machine learning algorithms. Or maybe for an example people actually like a service like Spotify. 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 variable (more than one are possible). Some of its best use cases involve time ordered data, or just continuous variable data in general that you want a numerical answer for. 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. (Programs, 2018)

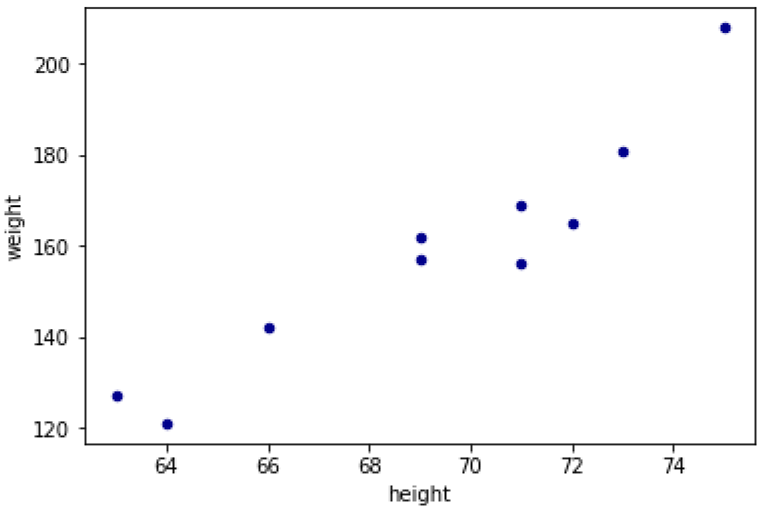
To actually use linear regression correctly there are a few key assumptions that need to be met or your regression model is going to have flaws some of which are: Linear relationship between dependent and independent vars, Multivariate normality (normal distribution). No multicollinearity which happens when independent variables are correlated with each other. No auto-correlation which is where y(x) is not independent from other values such as y(x-1), y(x+1), etc. Homoscedasticity which is where residuals are equals across the regression line, residuals being the difference between the predicted value and observed value. (Solutions, 2019)

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 from (Programs, 2018).

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

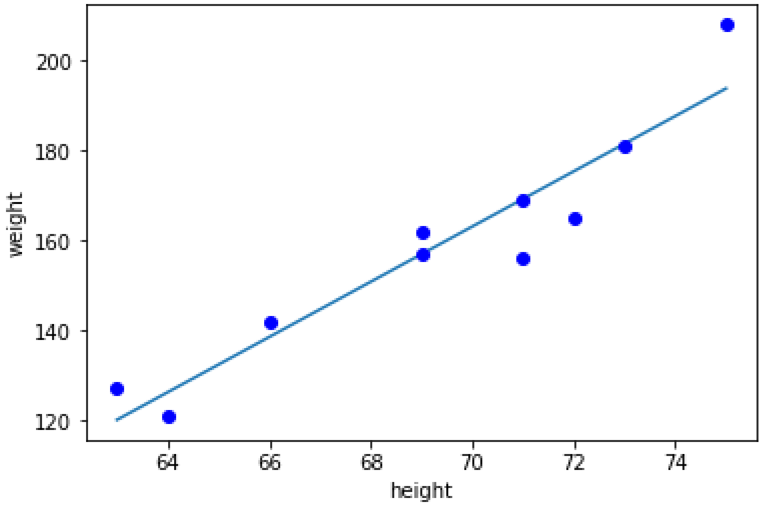


In order to find the best fit line one way is using the method of least squares. To do so you take the sum of squared prediction errors which the sum of squares of $e_i = y_i - \hat{y}_i$. Where e_i is the error of the prediction, y_i is the actual value and $\hat{y}_i$ is the prediction value. You take the square because otherwise when we sum the values up the positive and negative values would cancel each other out. You do this because we want to minimize the error, or distance from the line to the true values. 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 $\mathrm{b}_1 =$ $\frac{\sum_{i=1}^{n}(x_i - x)(y_i - y)}{\sum_{i=1}^{n}(x_i - x)^2}$. This will give the coefficients that minimize the error between the predicted y and y actual. Plugging in the equation for $\hat{y}_i$ in $Q = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ you get $Q = \sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2$.

The mean of x is 69.3, and the mean of y is 158.8.

i	x_i Height inches	y_i Weight actual lbs	$\hat{y}_i$ Estimated Weight lbs	$e_i = y_i - \hat{y}_i$	$e_i = (y_i - \hat{y}_i)$^2	b_0	$\sum_{i=1}^{n}(x_i - x)(y_i - y)$	$\sum_{i=1}^{n}(x_i - x)^2$
1	63	127	120.13	6.8715	47.22	-259.668	200.34	39.69
2	64	121	126.27	-5.266	27.73	-271.805	200.34	28.09
3	66	142	138.54	3.459	11.97	-263.08	55.44	10.89
4	69	157	156.95	0.0465	0.002	-266.493	0.54	0.09
5	69	162	156.95	5.0465	25.47	-261.493	-0.96	0.09
6	71	156	169.23	-13.2285	174.99	-279.768	-4.76	2.89
7	71	169	169.23	-0.2285	0.052	-266.768	17.34	2.89
8	72	165	175.37	-10.366	107.45	-276.906	16.74	7.29
9	73	181	181.50	-0.5035	0.25	-267.043	82.14	13.69
10	75	208	193.78	14.2215	202.25	-252.319	280.44	32.49
				Sum approx. = 0	Sum = 597.34= error	Sum = -266.534	Sum = 847.6	Sum = 138.1
						b_0= -266.534	$b_1 = \frac{847.6}{138.1} = 6.1375$	

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



With a best fit line of $\hat{y}_{i(weight)} = -266.534 + 6.1375x_{i(height)}$.

With Linear Algebra

That is but one way to find the best fit line. There are a few strategies that can be employed to find a best fit line and an answer to an otherwise unsolvable system of equations. Within linear algebra is one such equation the "normal equation" $A^TA\hat{x} = A^Tb$. If we look at the same problem using matrices. If we put all of our points in matrices where Ax=b,

$$A=\begin{bmatrix}1 & 63\\1 & 64\\1 & 66\\1 & 69\\1 & 69\\1 & 71\\1 & 71\\1 & 72\\1 & 73\\1 & 75\end{bmatrix} \qquad x=\begin{matrix}b_0\\b_1\end{matrix} \qquad \text{and} \qquad B=\begin{bmatrix}127\\121\\142\\157\\162\\156\\169\\165\\181\\208\end{bmatrix}$$

We already know this system is inconsistent so we cannot solve it traditionally. This is where the normal equation $A^TA\hat{x} = A^Tb$ comes in.

$$A^TA\hat{x}\ produces\ the\ matrix\ \begin{matrix}10 & 693\\693 & 48163\end{matrix} \qquad A^Tb\ \text{produces}\ \begin{matrix}1588\\110896\end{matrix}$$

Then with A^TA being invertible we can solve for $\hat{x}$ by taking $(A^TA)^{-1}$. B which is:

$$\begin{bmatrix}\frac{48163}{1381} & -\frac{693}{1381}\\-\frac{693}{1381} & \frac{10}{1381}\end{bmatrix} \cdot \begin{bmatrix}1588\\110896\end{bmatrix} = \begin{bmatrix}-\frac{368084}{1381}\\\frac{8476}{1381}\end{bmatrix}$$

Checking against the previous results b_0= -266.534 and b_1 = 6.1375 which is perfect. Now for why this works has to do with A having independent columns. Because A has independent columns then Ax = 0 and the only solution is x = 0 so A^TA is invertible which makes it so we can solve $A^TA\hat{x} = A^Tb$ for the best solution using the inverse matrix of A^TA. (Lay, 2012)

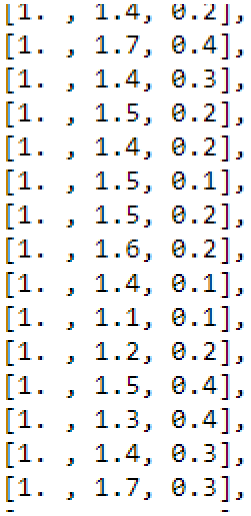
Multivariate Regression

Regression is possible with 1 …. n variables and as you may suspect it looks something like equation

$$\hat{y}_i = b_0 + b_1x_i + b_2x_2 + \cdots + \ b_nx_n$$

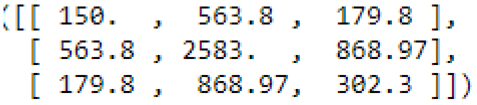
The techniques used previously apply here.

One such example going back to the iris data from (Raschka, 2015) we can use a regression to estimate what a irises sepal length is given the petal length and width. Why do this? Because taking measurements is hard. So, like before we form up the matrix A (only a small portion of the matrix)=



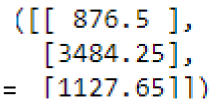
```
[1. , 1.4, 0.2],
[1. , 1.7, 0.4],
[1. , 1.4, 0.3],
[1. , 1.5, 0.2],
[1. , 1.4, 0.2],
[1. , 1.5, 0.1],
[1. , 1.5, 0.2],
[1. , 1.6, 0.2],
[1. , 1.4, 0.1],
[1. , 1.1, 0.1],
[1. , 1.2, 0.2],
[1. , 1.5, 0.4],
[1. , 1.3, 0.4],
[1. , 1.4, 0.3],
[1. , 1.7, 0.3],
```

Applying the normal equation, we are left with A^TA =



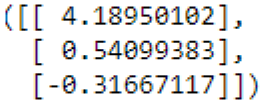
```
([[ 150.  ,  563.8 ,  179.8 ],
  [ 563.8 , 2583.  ,  868.97],
  [ 179.8 ,  868.97,  302.3 ]])
```

and A^Tb =



```
([[ 876.5 ],
  [3484.25],
  [1127.65]])
```

Solving with $(A^T A)^{-1}. A^T b$ gives the least squares solution

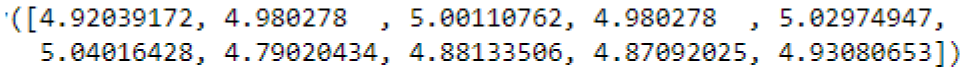


```
([[ 4.18950102],
  [ 0.54099383],
  [-0.31667117]])
```

And the equation for the sepal length from petal length and width is

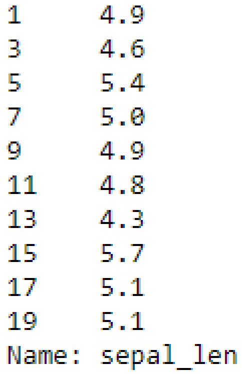
$\hat{y}_{(sL)} = 4.18950102 + 0.54099383 x_{1(P\ len)} - 0.31667117 x_{2(P\ wid)}$

Now to see how well we did let's look at some of the predicted values compared to the actual



```
([4.92039172, 4.980278  , 5.00110762, 4.980278  , 5.02974947,
  5.04016428, 4.79020434, 4.88133506, 4.87092025, 4.93080653])
```

Compare to the actual Y of data reserved for testing the first

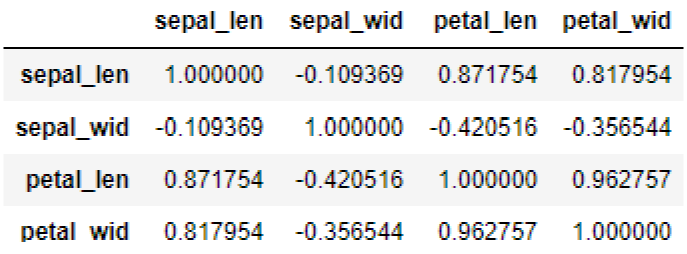


```
1     4.9
3     4.6
5     5.4
7     5.0
9     4.9
11    4.8
13    4.3
15    5.7
17    5.1
19    5.1
Name: sepal_len
```

With an R^2 value of 0.5748598957765443 which on the surface means it Is an alright fit.

Conclusion

Though I originally wanted to build a classification algorithm for the irises using linear regression and least squares I fell short. Linear regression can serve respectably well as a classification algorithm if the man assumptions are met. For one in the regression I did do petal length and width are quite correlated with each other which you do not want as seen by the correlation matrix



	sepal_len	sepal_wid	petal_len	petal_wid
sepal_len	1.000000	-0.109369	0.871754	0.817954
sepal_wid	-0.109369	1.000000	-0.420516	-0.356544
petal_len	0.871754	-0.420516	1.000000	0.962757
petal_wid	0.817954	-0.356544	0.962757	1.000000

When I was trying to construct the classification algorithm using linear regression to try and determine what variety of iris it was given a petal length and width, I ran into the problem of linear regression not being good for non-continuous variables. So as a whole my multivariate linear regression classification algorithm attempt was a series of lessons in how not to do things.

Linear regression is much better suited to something like heights and weights, or changes over time. After you gotten your sample data and found your best fit line you can now predict what a person's weight will be based on their height. If linear regression was only useful for something like finding someone's predicted weight based on their height, or the sepal length of an iris it would not be all that exciting. Especially when you consider that in this case it will only model a certain subset of the population well. For example, let's assume that height weight data was from adult males. If we tried predicting a female's weight using our best fit line, we may be quite a bit off the mark. Even worse if we looked at a child let's say 36 inches tall our best fit line would predict that the child weights -45.584 pounds. Which is bad news for that kid. These are just a few of the pitfalls you can fall into using linear regression. 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 stock's price may be in the future, how many purchases to expect in spring, a computer recognizing a face from a picture, predicting heart attack risk, and so on.

References

Lay, D. (2012). *Linear Algebra and its applications* . Boston: Addison-Wesley.

Programs, D. o. (2018). *Regression Methods*. Retrieved from PennState, Eberly College of Science: https://newonlinecourses.science.psu.edu/stat501/node/252/

Raschka, S. (2015). *Principal Component Analysis*. Retrieved from sebastianraschka: https://sebastianraschka.com/Articles/2015_pca_in_3_steps.html

Solutions, S. (2019). *CompleteDissertation*. Retrieved from Statistics Solutions : https://www.statisticssolutions.com/assumptions-of-linear-regression/