So, What Is The Difference Between The Loess Line And The Regression Line?

Sun Ho Song

11/25/2017

Introduction

While taking one of University of California, Berkeley's upper division statistics courses, Stat 133 (Concepts in Computing with Data), I encountered the loess line and the regression line several times. Regardless of the type of graphing tools I used (baseplot function, ggplot2, and ggvis), I always had to draw the loess line and the regression line for the assignment.

So, what are these loess and regression lines? Well, such question was never answered during the lecture. I learned the concept of regression line both in high school and in college, but I have never heard of the word loess line before. So I decided to figure out the answer by myself. And I assume that many of other students in Stat 133 course are in a similar situation as I did. Many of them draw those lines for their assignments without knowing much about them. As a result, I hope everyone to have a better understanding in loess and regression lines after reading my post!

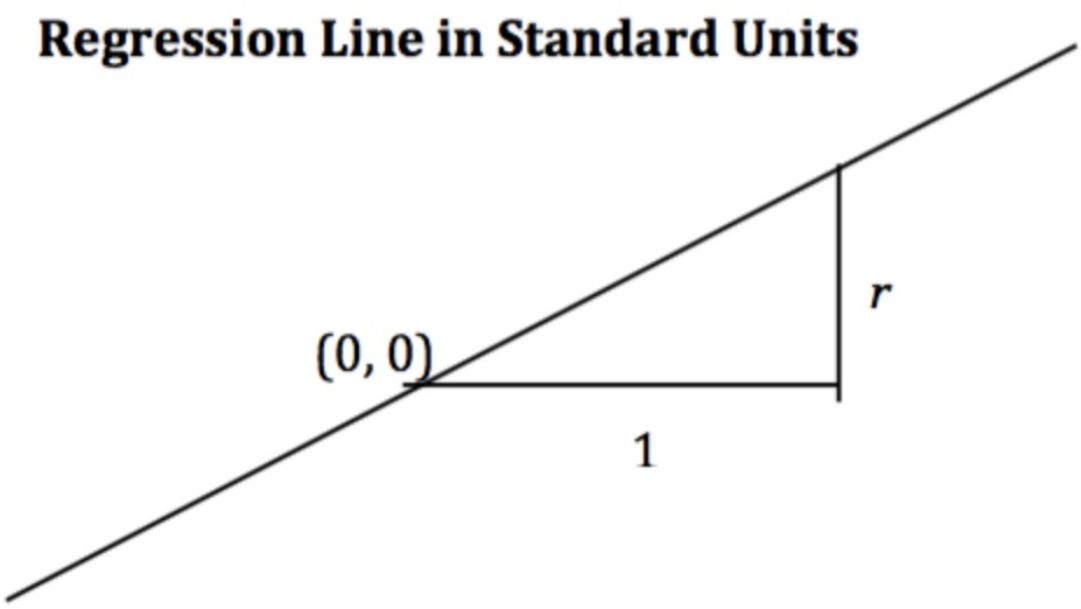
What is the regression line?

For those who have not heard of regression line before, I am pretty sure you have heard either least-squares regression or a line of best fit. All these three generally represent the same thing. In fact, least-squares method is the most common method for finding a regression line. The regression line is a straight line that minimizes the vertical distance between the line and the data points that are given. However, since the distance is composed of both positive and negative numbers, the sum of the vertical distance will give us 0. To prevent this meaningless result, we first square each vertical distance, then add them together so that no cancellation occurs. As a result of this addition, we get what is called sum of the squared errors. We divide sum of the squared errors by the number of data sets to get a mean squared error. Some might ask why not just transform all the distances into an absolute value. Though that is a good point, we have hard time differentiating the absolute values. Thus, squaring is a better method.
This notion of vertical distance indicates that if all the points lie on the regression line, it means that the vertical deviation from each point is 0, which also means that the mean squared error is 0.
The whole purpose of finding the regression line is to predict the value within the range (we do not use regression line to extrapolate to predict a very big or a very small number that is far from our data range)

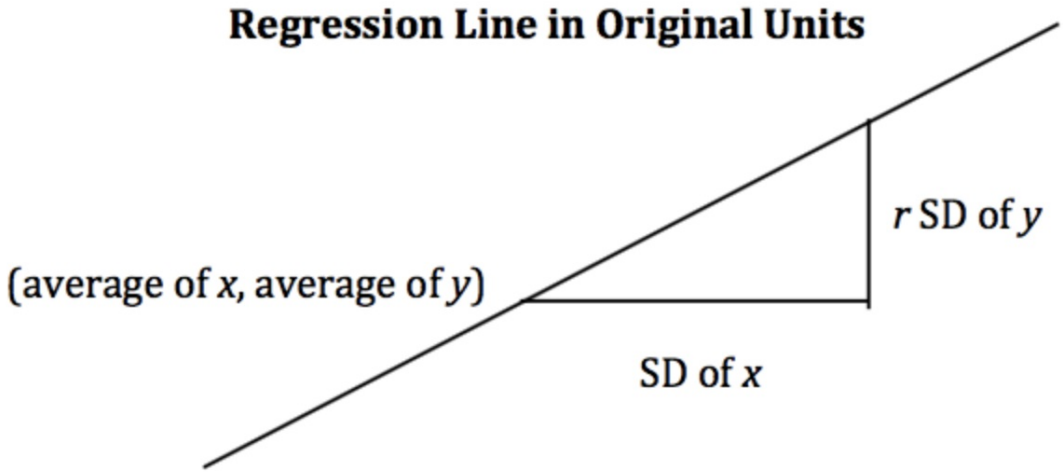
** Caution: Regression line is a straight line, which means it is only useful when the data points have a linear correlation.

Equation of the regression line

How the computation of a regression line differs depending on the unit:



If the unit is in the standard units, the y-intercept is at (0,0) and the slope is the value of r itself.



If the unit is in the original units, we have a totally different y-intercept and slope.

Equation of the regression line in original units:

The slope and intercept of the regression line in original units can be derived from the diagram above.

$$\textbf{slope of the regression line} = r \cdot \frac{\text{SD of } y}{\text{SD of } x}$$

$$\textbf{intercept of the regression line} = \text{average of } y - \text{slope} \cdot \text{average of } x$$

(These images are from Data 8 course textbook)

What is the loess line?

Just like the regression line, loess line also has many names. Loess line, lowess line (locally weighted scatterplot smoothing), and local polynomial regression represent the same thing. Loess line is a non-parametric regression method (meaning that it does not have a predetermined shape like regression line does) that is a mixture of several other regression models. The strength of the loess line is that it functions just like a regression line in a way that it minimizes the errors with the flexibility of nonlinear regression. However, it is almost impossible to find the loess line with hand because of its complicated computation (which is the reason why high school students and college students rarely confront this method in the theoretical statistics courses).
What is interesting about this loess line is that it has a smoothing parameter that changes the shape of the line. The smoothing parameter is a value between 0 and 1, but including 1, that represents the proportion of observations used for drawing the regression. As the value of the smoothing parameter increases, the smoother the line becomes.

Since the equation involves too much computations and a concept of weighting, I will not cover the equation in this post.
The advantage of using the loess line is first, it can be used for the data points with nonlinear pattern. Adding to that, rather than drawing a line that represents the whole data sets, loess line draws a prediction line for each section of the data sets that is found by dividing data sets into different segments, which is why it is not linear line. Overall, loess line draws a fitting line with a higher accuracy that aids predicting the data points when the data sets do not have a linear correlation.

What Is The Difference Between The Loess Line And The Regression Line?

Enough with the descriptions of the loess line and the regression line. Going back to the question of mine, we use the loess line and the regression line in different circumstances. We use loess line when the data points have nonlinear correlation and we use regression line when the data points have the linear correlation.

Explanations, Examples, and Discussions

To make my blog post more reproducible, I will be using the data set that is built in R: Motor Trend Car Road Tests. Also, for my purpose, I will be using ggplot2 package to demonstrate the difference between the loess line and the regression line visually. The version of ggplot2 package that I am using is 2.2.1; the version of RStudio that I am using is 1.1.383; the version of R that I am using is 3.4.2.

To start with, I will make a code chunk to download the ggplot2 package:

```
# Installing ggplot2 package if you do not have one already
# install.packages("ggplot2")
```

Then, I will call the ggplot2 package from the library:

```
# Calling ggplot2 package from the library
library(ggplot2)
```

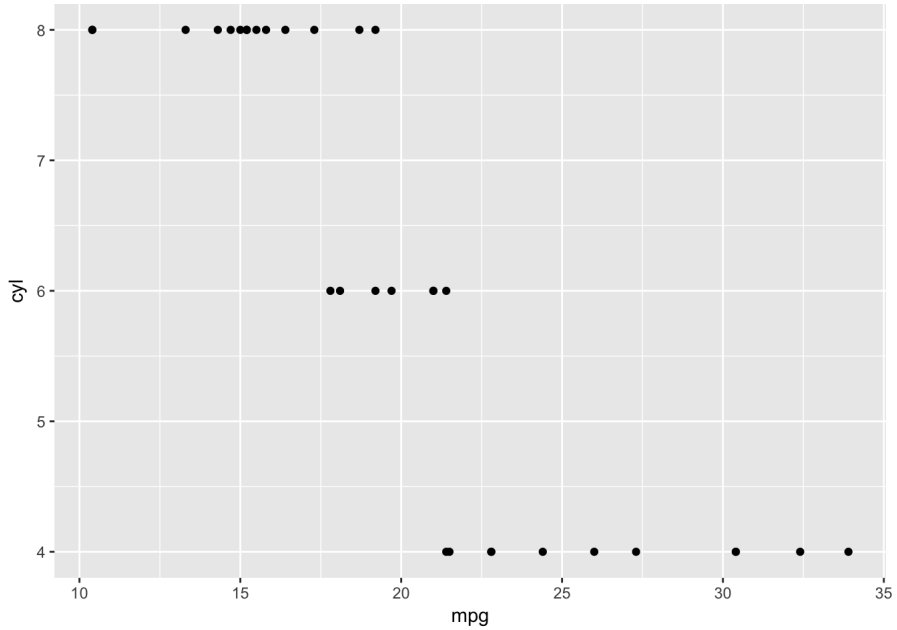
As I mentioned above, I will be using the Motor Trend Car Road Tests data sets that is already built in R.

```
# Showing first 10 rows of the mtcars data set
head(mtcars, 10)
```

```
##                   mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
```

From this data set, I will be looking at the correlation between the Miles/gallon (mpg column) and the Number of cylinders (cyl column). Let's start with drawing the scatterplot between the mpg (x-axis) and cyl (y-axis) columns.

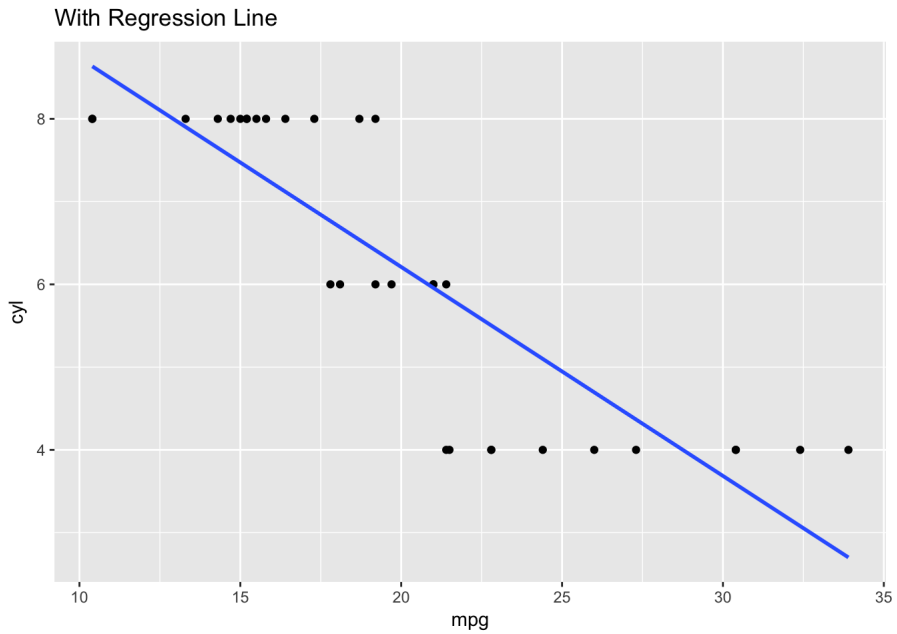
```
#Drawing mpg on the x-axis and cyl on the y-axis
ggplot(mtcars, aes(mpg, cyl)) + geom_point()
```



By observing this scatterplot, what do you see? Well, it seems like cyl and mpg have negative correlation: as the value of mpg increases, the

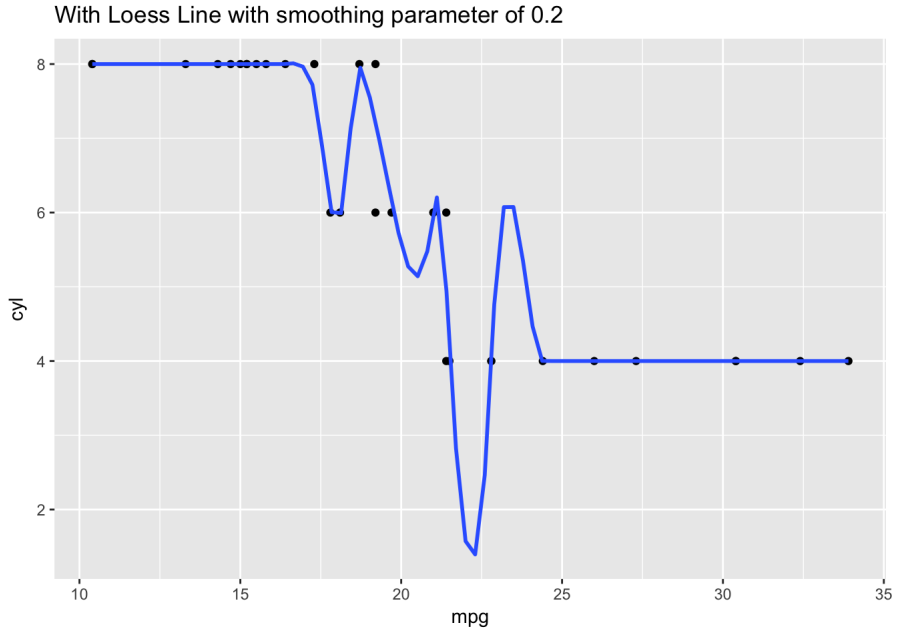
number of cyl decreases. Here is another question. Is the negative correlation linear or nonlinear? This negative correlation is closer to nonlinear correlation because as the value of mpg increases, the value of cyl decreases dramatically like a stair.
Now, let's actually draw the regression line and the loess line on the scatterplot.

```
# Drawing regression line on the scatterplot
ggplot(mtcars, aes(mpg, cyl)) + geom_point() + geom_smooth(method = "lm", se = FALSE) + ggtitle("With Regression L
ine")
```



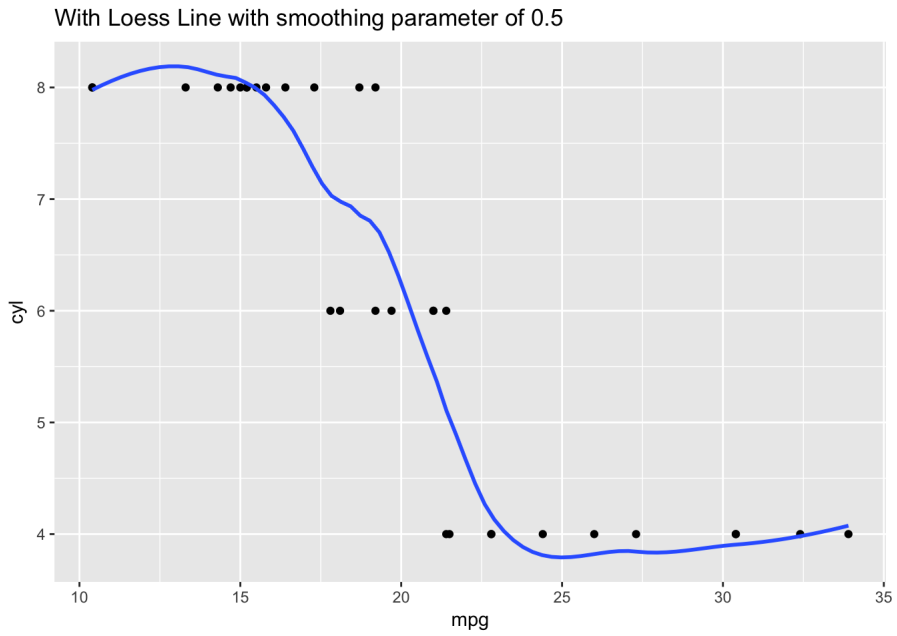
Well, although the regression line goes through the midpoint of all the points, it is hard to say that regression line best represents the correlation between mpg and cyl. Above, we said the pattern of data points is like a shape of stairs. We want to show a line that best represent that pattern. Going back to the regression line, regression line predicts that if a car has a mpg of 32.5, then the car would likely to have 3 cylinders. Is this true? According to our data sets, no. The prediction is totally off because we tried to predict a value for the data sets using the linear regression method when the data set really does not have a linear correlation.

```
# Drawing loess line with smoothing parameter of 0.2 on the scatterplot
ggplot(mtcars, aes(mpg, cyl)) + geom_point() + geom_smooth(method = 'loess', span = 0.2, se = FALSE) + ggtitle("Wi
th Loess Line with smoothing parameter of 0.2")
```



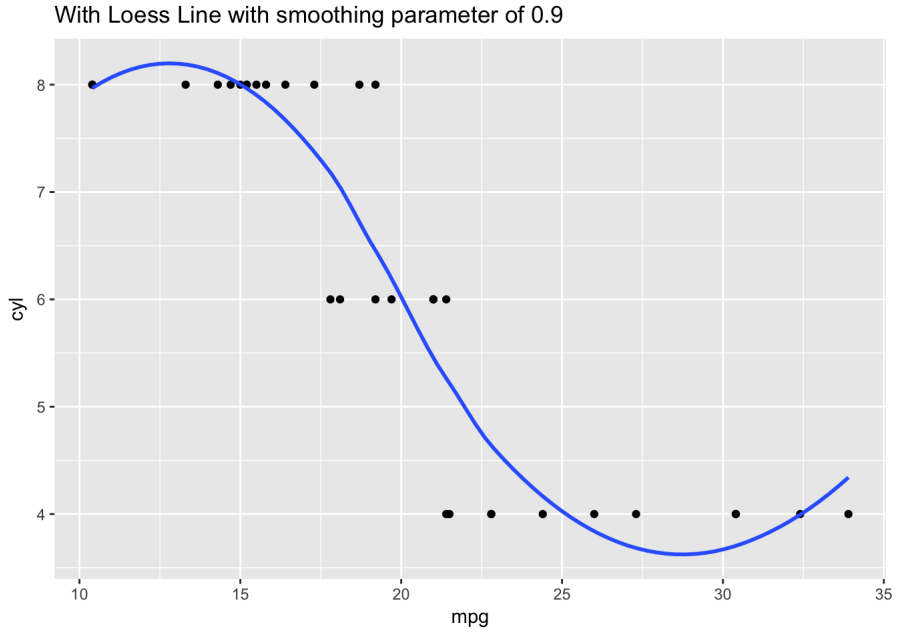
This is where the loess line comes in. Since the data sets have nonlinear correlation, we need to use a loess line. And by using the loess line with very small smoothing parameter, the loess line goes through almost every single points on the graph, increasing accuracy. Let us ask the same question that we did for the regression line. Loess line predicts that if a car has a mpg of 32.5, then the car would likely to have 4 cylinders. Is this true? Yes! The loess line got it perfectly. This shows that loess line is much better in predicting when the data sets do not have a linear correlation. However, since this loess line with the smoothing parameter of 0.2 does not seem to represent the nonlinear correlation pattern that we are looking for, let us try with a higher smoothing parameter.

```
# Drawing loess line with smoothing parameter of 0.5 on the scatterplot
ggplot(mtcars, aes(mpg, cyl)) + geom_point() + geom_smooth(method = 'loess', span = 0.5, se = FALSE) + ggtitle("Wi
th Loess Line with smoothing parameter of 0.5")
```



Does this look much better? Yes! R has found some patterns that we were looking for. Although the line does not go through most of the points as it did for the smoothing parameter of 0.2, as a result of smoothing, we got a pattern that better represents the data.

```
# Drawing loess line with smoothing parameter of 0.9 on the scatterplot
ggplot(mtcars, aes(mpg, cyl)) + geom_point() + geom_smooth(method = 'loess', span = 0.9, se = FALSE) + ggtitle("Wi
th Loess Line with smoothing parameter of 0.9")
```



Does this mean that if we have a higher smoothing parameter, the higher accuracy we will have? According to the graph above, that is not quite true. Because the line is too smoothed, we have lost some accuracy when mpg is between 10 and 15, and between 25 and 30. However, our representation of the loess line was successful in a way that we showed that loess line works better for the data sets with nonlinear correlation than the regression line does.

Conclusions (with take home message)

I wrote this post to go through the difference between the regression line and the loess line. Though I have been using these two lines for my assignments multiple times, I did not have a solid understanding in the loess line because it was my first time accessing it. I thought other students, too, may not have a very good understanding in loess line, or even in the regression line. So I tried to show them how they are different and how they are differently used. After the research and the demonstration on the actual data sets, I realized that the regression line is used when the data sets have a linear correlation (regardless positive or negative) and the loess line is used when the data sets do not have a linear correlation (regardless of positive or negative). Overall, we can use these lines to predict other values within the range of the data sets and we learned that loess line does a much better job when the data sets have a nonlinear correlation.

Take home message :
The reason why we have never heard of loess line during the theoretical statistics courses is that it is nearly impossible to compute the line manually. Only with the help of technology, this line can be found. Repeating myself, regression line is used when the data sets have a linear correlation and the loess line is used when the data sets do not have a linear correlation.
However, in most of times we do not know whether the data sets that we are dealing with have a linear correlation or not because most likely we will be dealing with the data sets that we have never seen before or that we are not familiar with. To prevent from drawing a wrong prediction line, it is better to draw both regression line and loess line to see which line fits better, just like students in Stat 133 did for their multiple

assignments.

References

- https://www.inferentialthinking.com/chapters/13/3/method-of-least-squares.html
- http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm
- https://en.wikipedia.org/wiki/Local_regression
- http://www.statisticshowto.com/lowess-smoothing
- https://www.statsdirect.com/help/nonparametric_methods/loess.htm
- http://geog.uoregon.edu/bartlein/old_courses/geog414f03/lectures/lec05.htm
- https://blogs.sas.com/content/iml/2016/10/17/what-is-loess-regression.html
- http://sites.stat.psu.edu/~fxc11/Stat462_STABLE/Lect12_lowess.pdf
- http://blog.naver.com/meta_com/220723151711
- https://stackoverflow.com/questions/15633714/adding-a-regression-line-on-a-ggplot