Regression and Loess Line

Yeon Mi Hwang

11/28/2017

Regression and Loess Line

[Introduction]

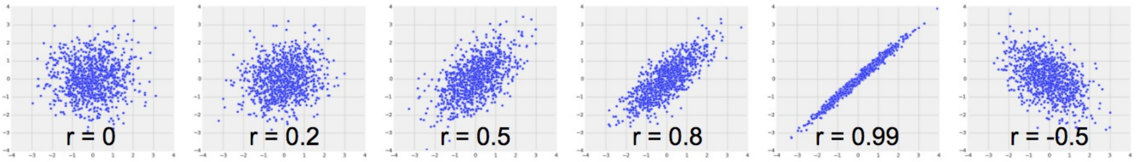
Do you know what exactly regression and loess line is?   I mean, we have used both lines pretty frequently throughout the stat133 course, but really, do we really know what they are? There may be some readers who completely understand about it, but for me, whose first stat course is stat133, I absolutely had no idea what loess line was. Regression line was not an exception. Although I have seen several regression lines often times in other courses, I did not understand the mechanism of regression line. I just used it. That is why I chose `Regression line` and `loess line` as the topic of this blog post.
I will first introduce you to slightly familiar `Regression line` and then we will dig into `loess line` afterwards.
Hope you enjoy it!

[Regression Line]

1) the correlation coefficent r

- tools to measure linear association
- -1 <= r <= 1
- `r=1` : scatter is perfect straight line up (every positive increase of 1 in one variable, there is a positive increase of 1 in the other)
- `r=-1` : scatter is perfect straight line down (every postive increase of 1 in one variable, there is a negative decrease of 1 in the other)
- `r=0` : no linear association; uncorrelated

Following image displays the general trend of scatter plot and the value of correlation coefficent r



- as you can see in the above image, as the absolute value of `r` approaches 1, we can see more obvious trend in the scatter plot.

2) definition of r

- *correlation coefficient(r)* = average of product of `x in standard units` and `y in standard units`.
- `x in standard unit` = (given x - average of x)/SD of x
- `y in standard unit` = (estimate of y - average of y)/SD of y

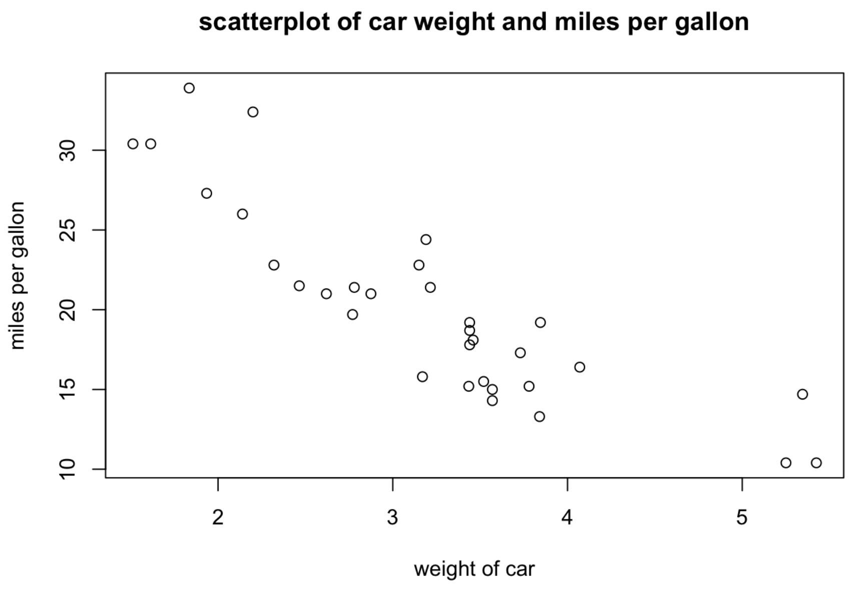
3) Regression line equation

- `y in standard unit` = `slope` * `x in standard unit` + `intercept`
- **slope of the regression line** = `r` * (`SD of y` / `SD of x`)
- **intercept of the regression line** = `average of y` - `slope` * `average of x`

4) Example of Regression line

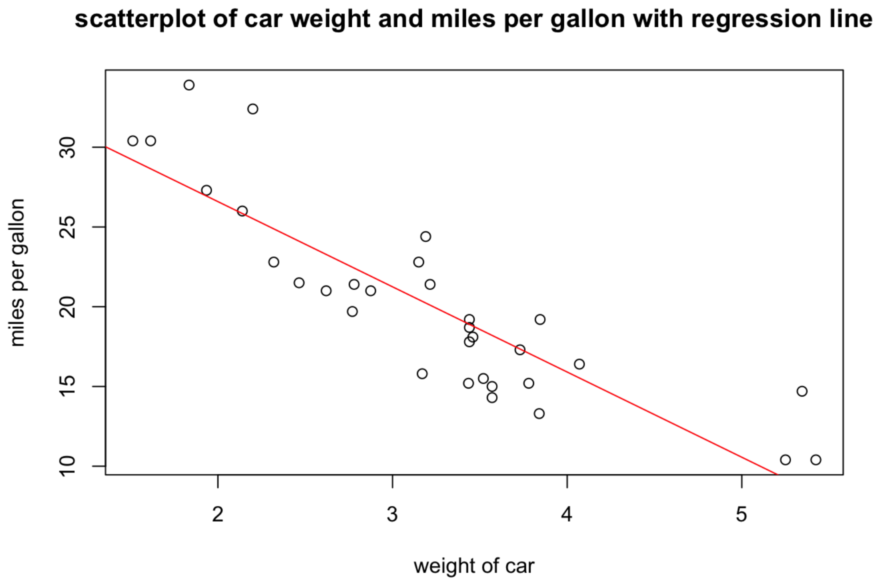
I will be using `mtcars` dataset which is comes together with Rstudio.
Following data displays scatterplot of car weight and miles per gallon.
+ we can obviously see the negative linear association between the car weight and miles per gallon.

```
#scatterplot between car weight and car miles per gallon
plot(mtcars$wt,mtcars$mpg,
     main = "scatterplot of car weight and miles per gallon",
     xlab="weight of car",
     ylab="miles per gallon")
```



I will add the regression line that matches above scatter plot.
regression line can be simply added by using function `abline(lm(y_axis~x_axis))`.
You just have to add that line to the code for the plot.

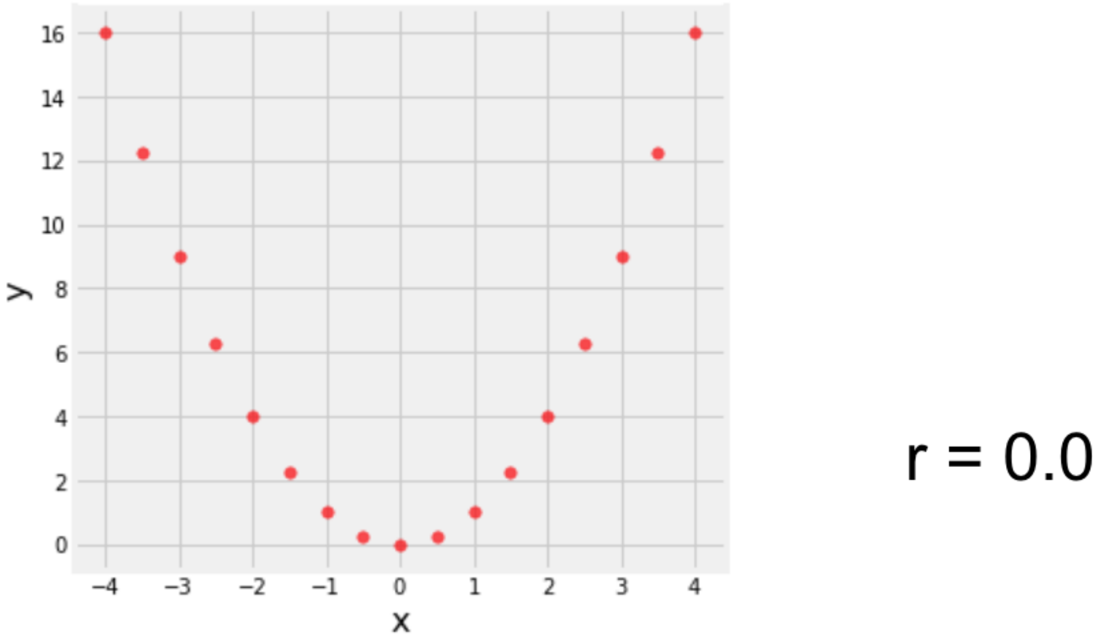
```
#scatterplot of car weight and mpg with regression line
plot(mtcars$wt,mtcars$mpg,
     main = "scatterplot of car weight and miles per gallon with regression line",
     xlab="weight of car",
     ylab="miles per gallon")
abline(lm(mtcars$mpg~mtcars$wt),col="red")
```



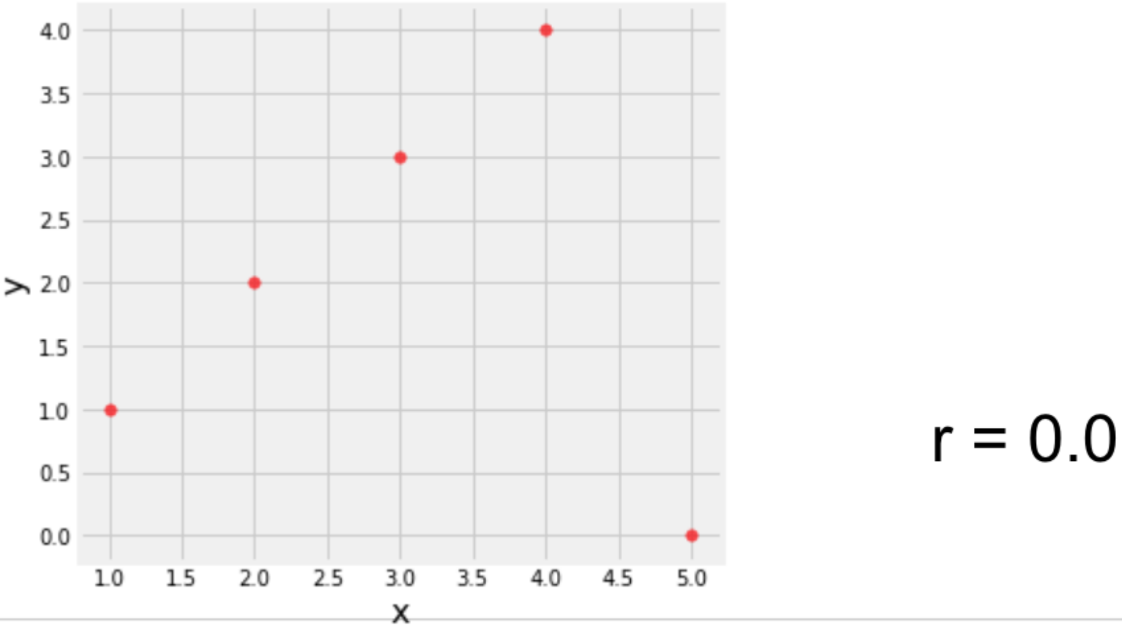
As you can see in the graph above, the red line matches the general trend of the scatter plot. It shows linear trend(positive association) of the data

5) disadvanatge of correlation coefficient r

- If the data does not show linear trend, r is not an effective tool to measure the association. In other words, `r` can only describe the data that has a simple association which is positive or negative.
  if you see the picture below, we can see there is a trend between x variable and y variable. However the trend is not linear, so the correlation coefficent turns out to be 0. `r` is not a good tool to measure the degree of nonlinear association



- In addition to nonlinear association, we also have to watch out for outliers. Few outlier of data can make the data to have little association.
  In the following image, we can obviously see there is a linear trend(let's ignore the outlier for a second)
  Regardless of the obvious linear trend, the outlier makes the data set to have no linear association.
  In such cases, we may have to reorganize the data set(for example, remove the outlier), so the regression line can better reflect the linear association of the data set.



[Loess line]

1) Definition of Loess

- LOWESS or LOESS is actually an acronym. LOWESS means Locally Weighted Scatterplot Smoothing and LOESS means Locally Weighted Smoothing.(LOWESS and LOESS is basically the same thing.)
- It may be understood as "LOcal regrESSion"
- non-parametric approach that fits multiple regression in local neighborhood

I bet this definition is quite difficult to understand. So let's begin with the word 'parametric'.

The word "parametric" means that researcher and analyst assume in advance that data fits some type of distribution.
If some type of distribution is assumed in advance, the smooth curve can often misinterpret the data.
Therefore, in such case (when the data is misinterpreted) non-parametric smoothers is a better choice. (and loess line is non-parametric smoother!)
I know that what I just said is super confusing. To make it simple, non-parametric approach means that LOESS tries to find the curve without assuming any type of distribution shape.
As mentioned earlier, loess line can be the acronym of "LOcal regrESSion".
That being said, multiple regression line is drawn in each infinitely small region of entire range(local neighborhood).
And those infinitely many regression lines are connected to create a loess line!

2) Characteristics of Loess line

- It builds upon “classical method” such as linear and non-linear least square regressions.
- LOESS line address the issue that classical procedures do not perform well.
- It combines the **simplicity** of linear least square regression and also the **flexibility** of nonlinear regression.

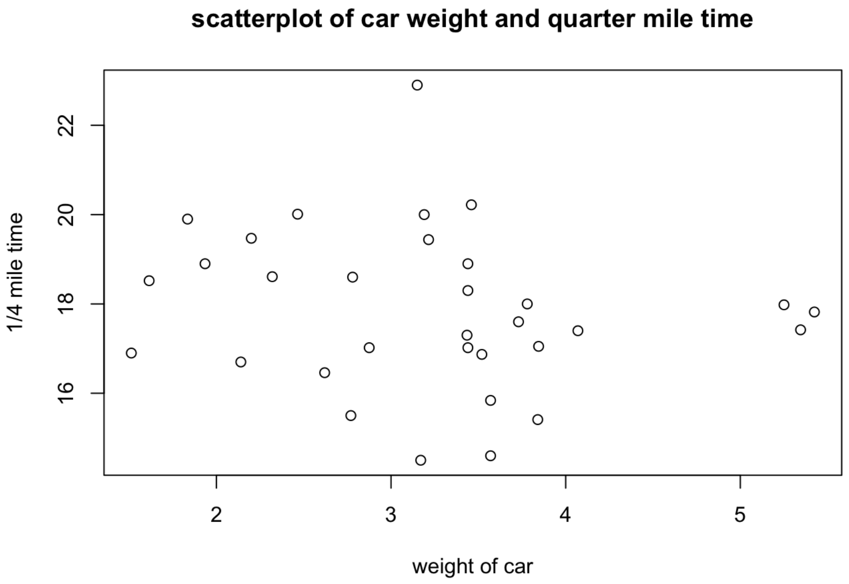
3) usage of Loess line

- scatterplot with noisy data value & sparse data points & weak interrelationships
- linear regression in which least squares fitting does not create a line of best fit
- social science data exploration! (this kind of data set has lots of noisy data value)

4) Example of Loess line

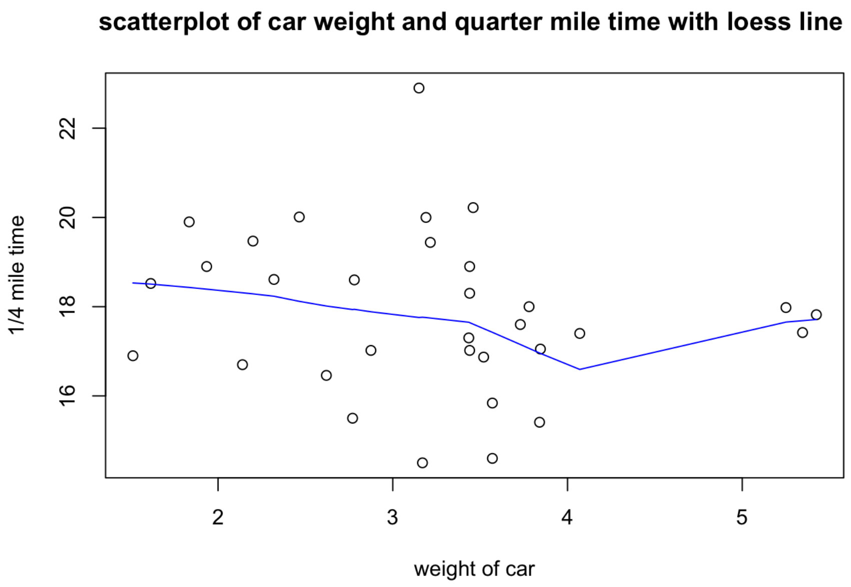
The graph below displays the scatterplot of car weight and quarter mile time. In contrast to previous scatterplot, it is quite difficult to observe the linear association with naked eye. It seems the weight of car is randomly distrbuted across the quarter mile time.

```
#scatterplot of car weight and quarter mile time
plot(mtcars$wt,mtcars$qsec,
     main = "scatterplot of car weight and quarter mile time",
     xlab="weight of car",
     ylab="1/4 mile time")
```

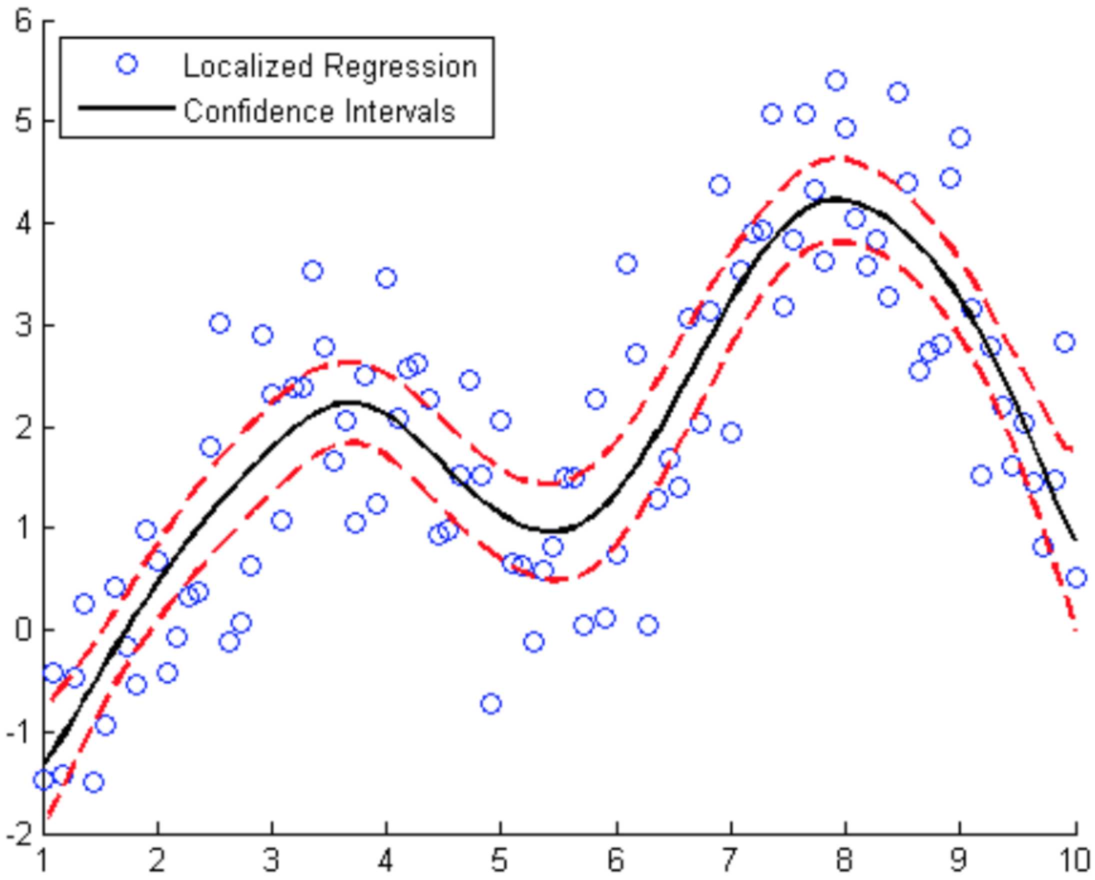


The graph below displays the scatterplot of car weight and quarter mile time and loess line added to it.
Similar to the way we added regression line on the plot, we have to add function `lines(lowess(x axis,y axis))` to the code in order to add a loess line.

```
#scatter plot of car weight and quarter mile time with loess line
plot(mtcars$wt,mtcars$qsec,
     main = "scatterplot of car weight and quarter mile time with loess line ",
     xlab="weight of car",
     ylab="1/4 mile time")
lines(lowess(mtcars$wt,mtcars$qsec),col="blue")
```



The above graph shows scatterplot with the loess line added to it.
The shape of the good fit line looks a lot different from the regression line which was straight.
The dataset I had does not have enough data point, so this may not be a good example to show the example of loess line.   Often, loess line shows a smooth curve to interpret the trend.
Following picture is a bit more common example of graph with lowess line.
To be honest, I do not know what this data is about but I thought it is a good graph of LOWESS line.



5) Benefit of Loess line

- It provides a flexible approach to represent data(since it is non-parametric!! type of distribution is not assumed in advance)
- easy to use!!

6) Disadvantage of Loess line

- requires fairly large, densely sampled data set in order to produce good models !!
- does not produce a regression function that can be easily represented by a mathemathical formula.
- Although I gave you the equation to get the linear regression line equation in the regression line section, I could not provide mathemathical model of loess line because it cannot easily be represented.
- it is computationally intensive

- similar to regression line, it is prone to the effect of outlier in the data set

[Conclusion/Take Home Message]

Through this blog post, I introduced you to the very concept of linear regression line and loess line! Along with the concept, I introduced you to the **characteristics**,**example, advantage** and **disadvantage** of each type of regression line.
Although we have used both types of regression line in order to interpret our data set, most of us did not really have an understanding about these regression lines.   While I was doing hw4 shiny part, curiosity about that two regression lines arose so I decided to write a blog post about it!   I hope you now gained some new insight about the loess and regression line, and I hope whenever you encoutner these regression lines in other classes you at least know what these lines are for!

Reference

image of loess line : (https://blogs.mathworks.com/loren/2011/01/13/data-driven-fitting/)
image of linear correlation example : (https://www.inferentialthinking.com/chapters/13/2/regression-line.html)
tutorial : (https://www.statmethods.net/graphs/scatterplot.html)
tutorial : (http://onlinestatbook.com/2/regression/intro.html)
tutorial : (http://www.dummies.com/education/math/statistics/how-to-calculate-a-regression-line/)
tutorial : (https://www.rdocumentation.org/packages/DescTools/versions/0.99.19/topics/lines.loess)
tutorial : (http://www.statisticshowto.com/what-is-a-regression-equation/)