Name: Madhusudhan Anand
Email: maddymaster@gmail.com

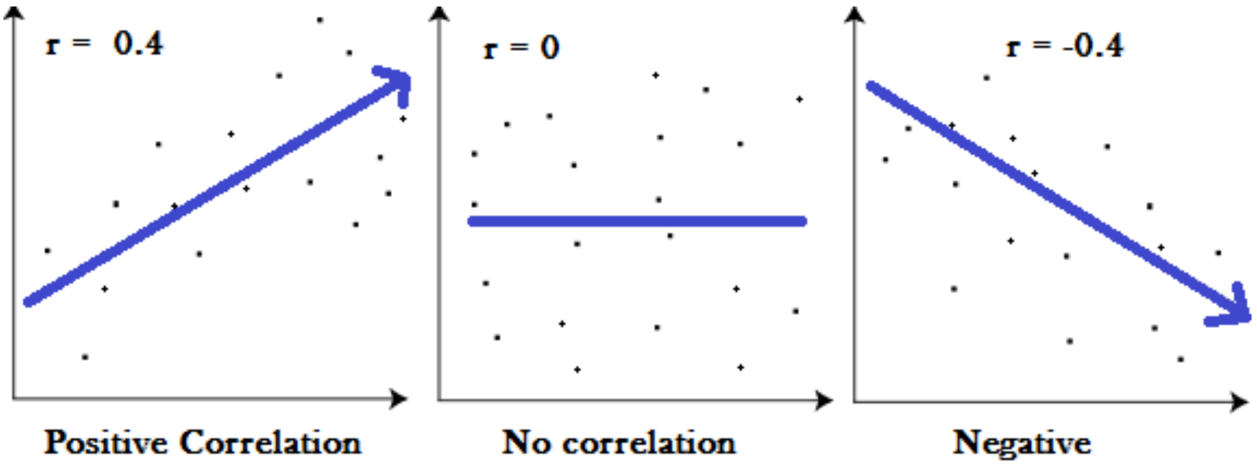
1. **What are the assumptions of linear regression regarding residuals?**

Answer: 1. Linear Regression assumes that the mean of the residuals is zero or very close to zero.
2. Then there will be same type scatter, that means The points higher on the x-axis have a larger variance than smaller values and the points have the same distance from the line.
3. The X variables do not find a correlation with the residuals
4. There is also an assumption that there won't be a perfect multicollinearity
5. The residuals are normally distributed

2. **What is the coefficient of correlation and the coefficient of determination?**

Answer: Coefficient of correlation is the R value, it defines how strong a relationship is between two variables. orrelation coefficient formulas are used to find how strong a relationship is between data. The formulae returns a value between -1 and 1, and its interpreted as

- 1 means there is a strong positive relationship.
- -1 should be interpreted as a strong negative relationship.
- A result of zero means that there is no relationship at all between the variables.



- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.
- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

3. Explain the Anscombe's quartet in detail.

Anscombes quartet has 4 data sets that have similar simple descriptive statistics, however, their distributions are very different and can be noted when their are visualized in a graph. constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties, they have eleven (x,y) points.

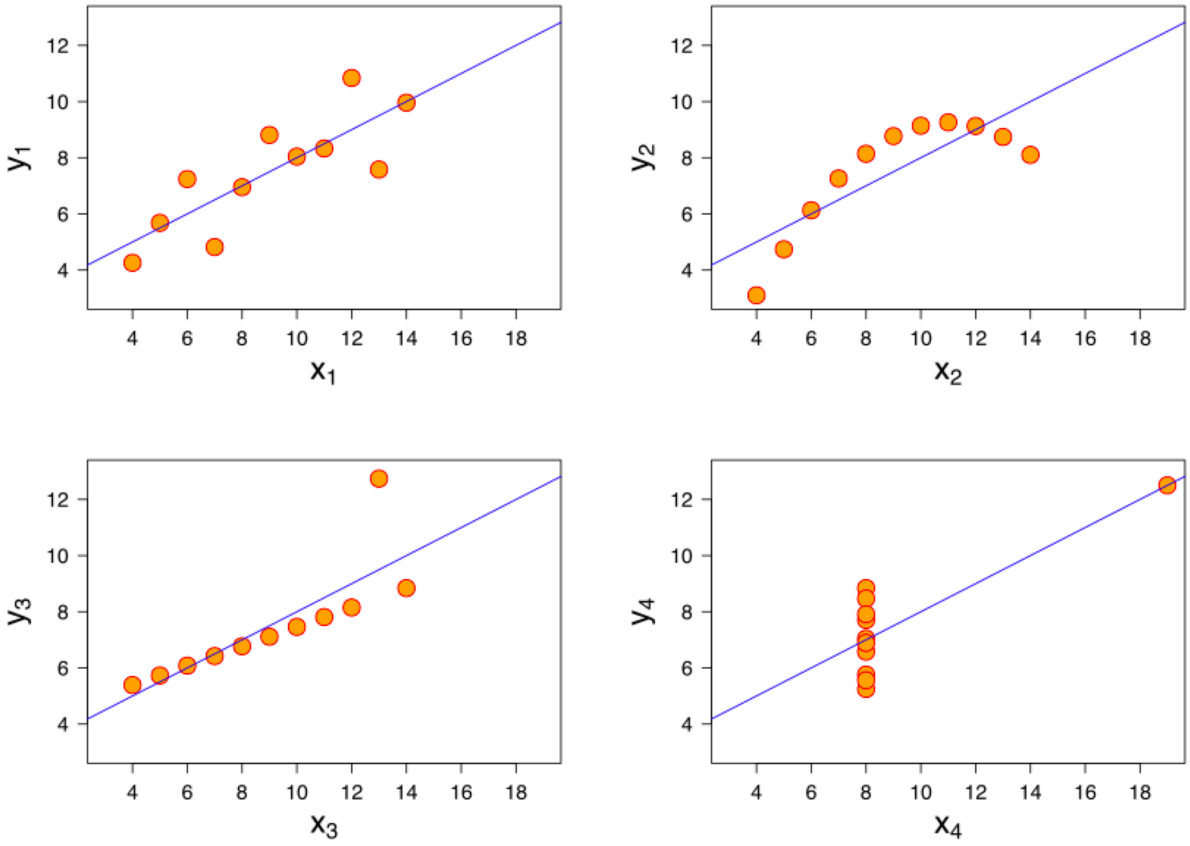
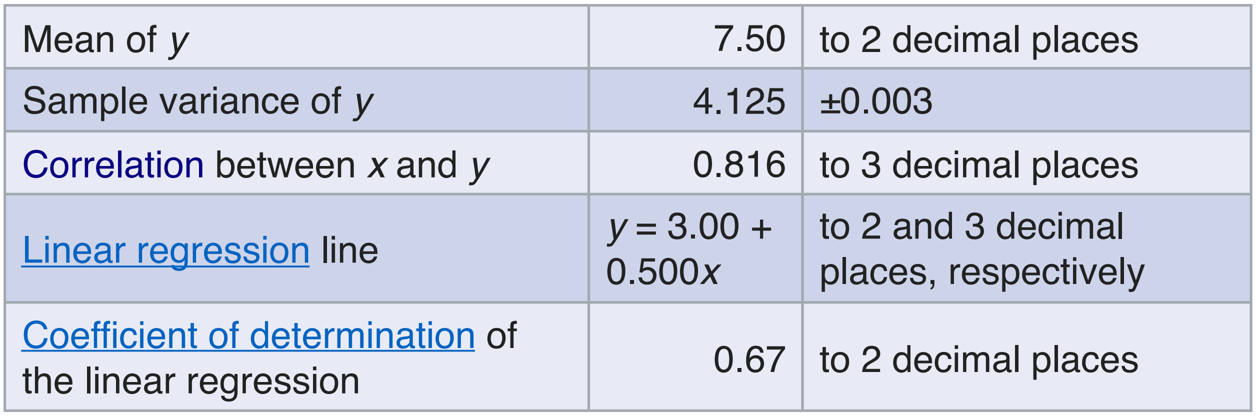


Image source: Wikipedia.

All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

For all four datasets:

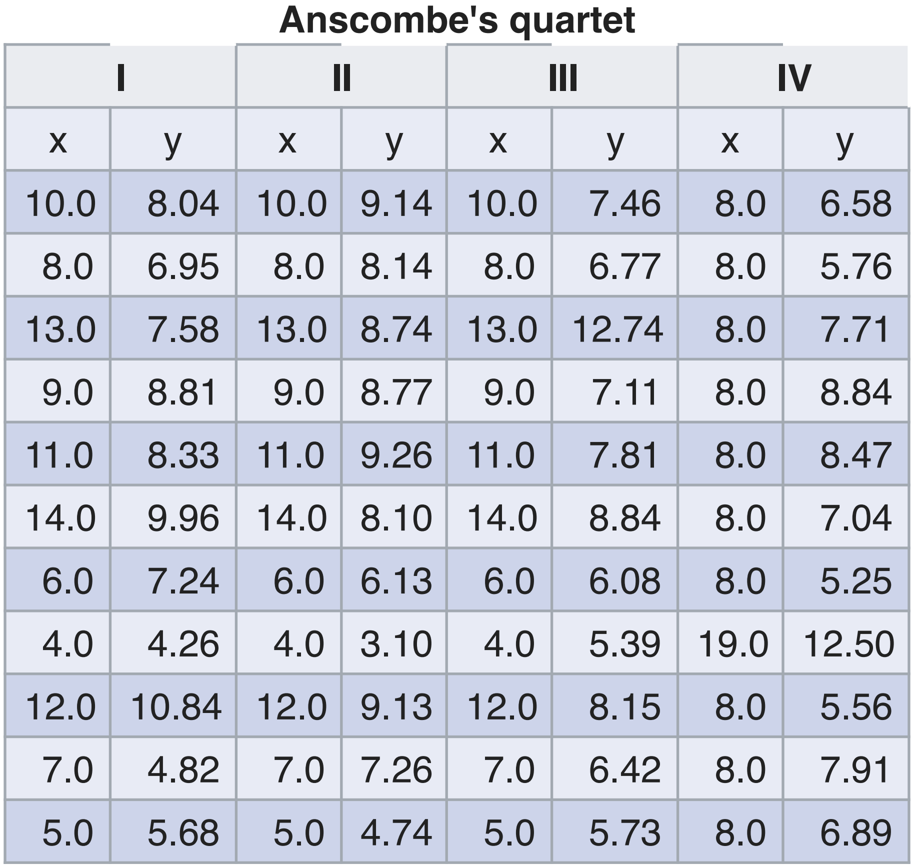
Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x*	11	exact



Mean of *y*	7.50	to 2 decimal places
Sample variance of *y*	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	*y* = 3.00 + 0.500*x*	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression	0.67	to 2 decimal places

x

[1]



Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Now you can see that the summary statistics show that the means and the variances were identical for x and y across the groups :

- Firstly notice that Mean of x is straight 9 and the mean of y is clearly 7.50 for each of the dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset

- And there is a strong relationship between x and y as the correlation coefficient is 0.816 for each dataset
- Now, the Dataset I appears to have a clean and is well-fitting the linear models.
- However, Dataset II is not at all distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

4. What is Pearson's R?

Answer: As we know that the strength of two variables' relationship with each other is gathered by R, is the **Pearson** product-moment **correlation** coefficient.
A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.

- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.
- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

5. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

scaling is a method used to normalize **the** range of independent variables or features of data. Now for example, the units can be of different range, there can be a certain value of the same attribute in thousands and there could be something in decimals and in the same column or so there could be something in millions (for ex population in a class versus a country or measure of weight in grams versus kilograms). Now when they are used without data pre-processing the variables with very large or very low values can stand out as outliers while they are just not, they are just on a different scale of measurement when the data was given. Hence as a important step to prepare data for any sort of machine learning, linear regression etc., we need to ensure that the data is normalized and there is a scale.

The result of **standardization** (or **Z-score normalization**) is that the features will be rescaled so that they'll have the properties of a standard normal distribution. μ=0

μ = 0 and σ=1

So that is how Standardized scaling is taken into account.

An alternative approach to Z-score normalization (or standardization) is the so-called **Min-Max scaling**(often also simply called "normalization" - a common cause for ambiguities).

In this approach, the data is scaled to a fixed range - usually 0 to 1.
The cost of having this bounded range - in contrast to standardization - is that we will end up with smaller standard deviations, which can suppress the effect of outliers.

6. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

ViF stands for Variance inflation factor and is used to help detect multicollinearity in regression analysis. Multicollinearity is when there's correlation between predictors (i.e. independent variables) in a model; it's presence can adversely affect your **regression** results.

In other words, ViF shows the degree to which a regression coefficient will be affected. Now, because of the variable's redundancy with other independent variables, the value of ViF can go higher. As the squared multiple correlation of any predictor variable with the other predictors approaches unity, the corresponding VIF becomes infinite indicating multicollinearity that will have an exponential coorrelation