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

1. **Normality assumption:** It is assumed that the error terms, ε(i), are normally distributed.
2. **Zero mean assumption:** It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.
3. **Constant variance assumption:** It is assumed that the residual terms have the same (but unknown) variance, σ^2. This assumption is also known as the assumption of homogeneity or homoscedasticity.
4. **Independent error assumption:** It is assumed that the residual terms are independent of each other, i.e. their pair-wise covariance is zero.

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

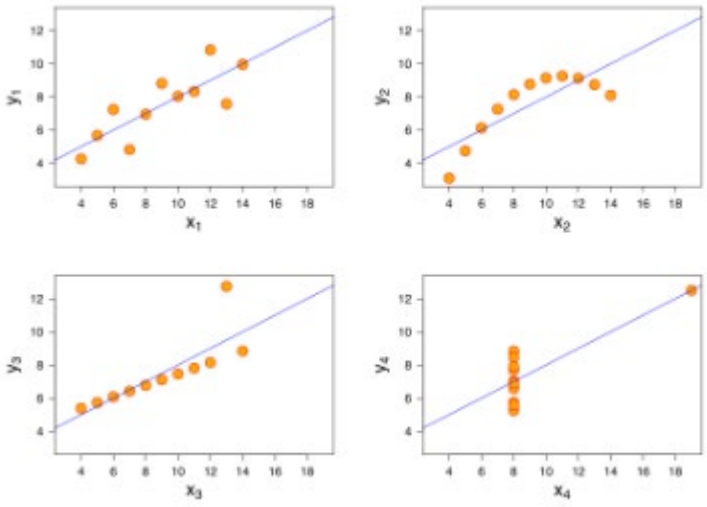
Coefficient of correlation gives the value of correlation between two variables indicating that when one of the variables is changing, how much it is impacting the other variable. Coefficient of determination, on the other hand, is just the coefficient of correlation squared. For simple linear regression, it is simply the R-squared value.

3. **Explain the Anscombe's quartet in detail.**

You should never just run a regression without having a good look at your data because simple linear regression has quite a few shortcomings:

1. It is sensitive to outliers
2. It models the linear relationships only
3. A few assumptions are required to make the inference

These phenomena can be best explained by the Anscombe's Quartet, shown below:



As we can see, all the four linear regression are exactly the same. But there are some peculiarities in the datasets which have fooled the regression line. While the first one seems to be doing a decent job, the second one clearly shows that linear regression can only model linear relationships and is incapable of handling any other kind of data. The third and fourth images showcase the linear regression model's sensitivity to outliers. Had the outlier not been present, we could have gotten a great line fitted through the data points. So, we should never ever run a regression without having a good look at our data.

4. **What is Pearson's R?**

If two variables are correlated, it is very much possible that they have some other sort of relationship and not just a linear one.

But the important point to note here is that there are two correlation coefficients that are widely used in regression. One is the Pearson's R correlation coefficient which is the correlation coefficient you've studied in the linear regression model. This correlation coefficient is designed for linear relationships and it might not be a good measure for if the relationship between the variables is non-linear. The other correlation coefficient is Spearman's R which is used to determine the correlation if the relationship between the variables is not linear. So even though, Pearson's R might give a correlation coefficient for non-linear relationships, it might not be reliable. For example, the correlation coefficients as given by both the techniques for the relationship $y = X^3$ for 100 equally separated values between 1 and 100 were found out to be:

$$\text{Pearson's R} \approx 0.91$$

$$\text{Spearman's R} \approx 1$$

And as we keep on increasing the power, the Pearson's R value consistently drop whereas the Spearman's R remains robust ar 1. For example, for the relationship y=X10 for the same data points, the coefficients were:

$$\text{Pearson's R} \approx 0.66$$

$$\text{Spearman's R} \approx 1$$

So, the takeaway here is that if you have some sense of the relationship being non-linear, you should look at Spearman's R instead or Pearson's R. It might happen that even for a non-linear relationship, the Pearson's R value might be high, but it is simply not reliable.

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

Scaling is the process of bringing all the variables to a same scale. Scaling is performed mostly during model building processes to bring everything to the same scale.

In normalized scaling, we use the maximum and the minimum values of a particular column to perform the scaling. For any datapoint 'X' in a column 'C', this scaling is performed using the formula: X – min(C)/max(C) – min(C).

Standardized scaling, on the other hand, brings all the data points in a normal distribution with mean zero and standard deviation one. It is performed using the formula: X – mean(C)/SD(C)

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

Recall that the formula for VIF is given as:

$VIF = 1/(1-R^2)$

Now, when you're calculating the VIF for one independent variable using all the other independent variables, if the R^2 value comes out to be 1, the VIF will become infinite. This is quite possible when one of the independent variables is strongly correlated with many of the other independent variables.