Subjective Questions – Linear Regression

1) Explain the linear regression algorithm in detail.

A simple linear regression model attempts to explain the relationship between a dependent variable and an independent one using a straight line.

Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable. This method is mostly used for forecasting and finding out cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.

In other words we try to find a best fit lines among given datapoints, this line also describes the relationship among dependent and independent variables.

Based on given datapoints we iterate to plot a best fit line for the model. The equation depicting linear regression model is:

$Y= b0 + b1*x + e$

Where,

Y = dependent variable

X = independent variables

b0 = intercept (constant)

b1 = slope of the line

e = sum of squared errors

We need to perform following steps to perform the algorithm:

1. EDA
2. Data Preparation (dummy variable, label encoding etc)
3. Splitting dataset into test-train
4. Rescaling
5. Building Model
6. Residual Analysis
7. Making prediction
8. Model evaluation

2) What are the assumptions of linear regression regarding residuals?

The assumption of the linear regression analysis regarding the residuals is homoscedasticity. The scatter plot is good way to check whether the data are homoscedastic (meaning the residuals are equal across the regression line), i.e. error terms should have constant variance and it should not follow any pattern. It should be independent of each other.

The assumption of normality is made, as it has been observed that the error terms generally follow a normal distribution with mean equal to zero in most cases.

3) What is the coefficient of correlation and the coefficient of determination?

coefficient of correlation:

Denoted by R, this is used in to measure a relationship is between two variables. The value may be between -1 and 1, where:

1 indicates a strong positive relationship.

-1 indicates a strong negative relationship.

zero indicates no relationship at all.

coefficient of determination:

Also known as R-squared, explains how much variance of data has been explained by the linear model.

More specifically, R-squared gives you the percentage variation in y explained by x-variables. The range is 0 to 1 (i.e. 0% to 100% of the variation in y can be explained by the x-variables.

It gives you an idea of how many data points fall within the results of the line formed by the regression equation. The higher the coefficient, the higher percentage of points the line passes through when the data points and line are plotted.

4) Explain the Anscombe's quartet in detail.

This can be best summarized as:

First- and second-order summary statistics don't say everything you might want to know about your data, so remember to plot it. The real story behind data points comes alive after plotting it.

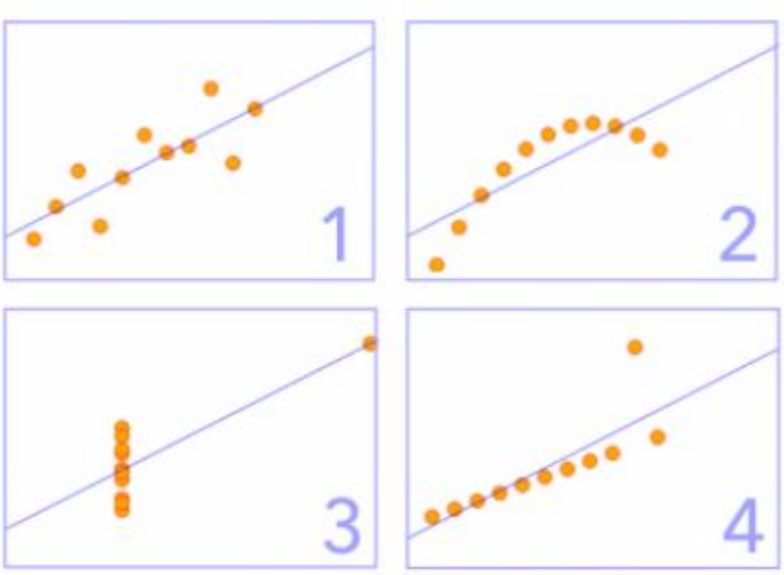
Summary statistics allow us to describe a vast, complex dataset using just a few key numbers. This gives us something easy to optimize against and use as a barometer for our business.

However, relying only on summary statistics and ignoring the overall distribution can be misleading. Calculating summary statistics, while useful, should only be one piece of your data analysis pipeline.

Let us consider the table below:

Set	I		II		III		IV	
Summary	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Mean	9	9	9	9	9	9	9	9
Sd	3.32	3.32	3.32	3.32	3.32	3.32	3.32	3.32
corr(x,y)	0.816	0.816	0.816	0.816	0.816	0.816	0.817	0.817

All the vitals of the four sets seem to be similar, however the real story comes alive after plotting these sets:



5) What is Pearson's R?

Denoted by R, this is used in to measure a relationship is between two variables. Pearson's r can range from -1 to 1. An r of -1 indicates a perfect negative linear relationship between variables, an r of 0 indicates no linear relationship between variables, and an r of 1 indicates a perfect positive linear relationship between variables. Figure 1 shows a scatter plot for which r = 11 indicates a strong positive relationship. The two variables here can have different units.

1 indicates a strong positive relationship.

-1 indicates a strong negative relationship.

zero indicates no relationship at all.

6) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

As the name suggests it is a technique used in case we need to compare variables of a contrasting units e.g. in case of our "Geely Auto" assignment we had wheelbase vs curbweight vs price all these had contrasting values, if we try to plot them on the same graph wheelbase would be very marginal and might not be noted, for such cases we need to bring them to one scale where these are comparable.

We need to perform this during regression because most of the machine learning algorithms use Eucledian distance between two data points in their computations. If not scaled properly it would cause a problem.

Normalization rescales the values into a range of [0,1]. This might be useful in some cases where all parameters need to have the same positive scale. However, the outliers from the data set are lost. The dataset hence is compressed between 0 and 1.

$X_{changed} = (X - Xmin)/(Xmax - Xmin)$

Standardization rescales data to have a mean (μ) of 0 and standard deviation (σ) of 1 (unit variance). The data spread is hence spread across mean value of zero.

$X_{changed} = (X - \mu)/\sigma$

For most applications standardization is recommended.

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

VIF is abbreviation to Variance Inflation Factor, let's try to understand given condition via it's mathematical formula:

VIF = 1/(1-r^2)

Where, r = coefficient of correlation

and r^2 = coefficient of determination

hence, an infinite VIF implies that r=r^2=1. This would mean that model is able to explain all the variance and it is a perfect fit.

In VIF, each feature is regression against all other features. If R2 is more which means this feature is correlated with other features. When R2 reaches 1, VIF reaches infinity. It means the variables are perfectly correlated with each-other. That would imply that set of variables in consideration are completely redundant and we might need to treat collinearity.

8) What is the Gauss-Markov theorem?

The Gauss Markov theorem tells us that if a certain set of assumptions are met, the ordinary least squares estimate for regression coefficients gives you the best linear unbiased estimate (BLUE) possible.

There are five Gauss Markov assumptions:

i) Linearity: the parameters we are estimating using the OLS method must be themselves linear.
ii) Random: our data must have been randomly sampled from the population.
iii) Non-Collinearity: the regressors being calculated aren't perfectly correlated with each other.
iv) Exogeneity: the regressors aren't correlated with the error term.
v) Homoscedasticity: no matter what the values of our regressors might be, the error of the variance is constant.

The Gauss Markov assumptions guarantee the validity of ordinary least squares for estimating regression coefficients.

9) Explain the gradient descent algorithm in detail.

Gradient descent is an optimization algorithm used to minimize cost function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient. In machine learning, we use gradient descent to update the parameters of our model. Parameters refer to coefficients in Linear Regression and weights in neural networks. This has to be done in a way to achieve most cost effective way.

Types:

i) Full Batch Gradient Descent Algorithm

ii) Stochastic Gradient Descent Algorithm

In full batch gradient descent algorithms, you use whole data at once to compute the gradient, whereas in stochastic you take a sample while computing the gradient.

10) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

As the name suggests the Q-Q plot, or quantile-quantile plot are plots of two quantiles against each other. It is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. The purpose of Q Q plots is to find out if two sets of data come from the same distribution.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

The normal Q Q plot is one way to assess normality. However, you don't have to use the normal distribution as a comparison for your data; you can use any continuous distribution as a comparison.

If the datasets in consideration near similar distribution the Q–Q plot will approximately lie on the line y = x.

In linear regression it can be used to assess if residuals are normally distributed. A 45^0 angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line. We need to observe if there is any patterns that deviate from this - particularly anything that looks curvilinear (bending at either end) or s shaped.