Assignment-based Subjective Questions

Akshay Ginodia

BHOOMISH ATHA

2. Table of Contents

1. Assignment-based Subjective Questions

A. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

- The demad of bike is less in the month of spring when compared with other seasons
- The demand bike increased in the year 2019 when compared with year 2018.
- Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.
- Bike demand is less in holidays in comparison to not being holiday.
- The demand of bike is almost similar throughout the weekdays.
- There is no significant change in bike demand with working day and non working day.
- The bike demand is high when weather is clear and Few clouds however demand is less in case of Lightsnow and light rainfall. We do not have any that for Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + For eg , so we can not derive any conclusion. May be the company is not operating on those days or there is no demand of bike.

B. Why is it important to use **drop_first=True** during dummy variable creation(2 mark)

drop_first=True drops the first column during dummy variable creation. Suppose, you have a column for gender that contains 4 variables- "Male", "Female", "Other", and "Unknown". So a person is either "Mal or "Female", or "Other". If they are not either of these 3, their gender is "Unknown".

We do NOT need another column for "Unknown".

It can be necessary for some situations, while not applicable for others. The goal is to reduce the number of columns by dropping the column that is not necessary. However, it is not always true. For some situations, we need to keep the first column.

Example

Suppose, we have 5 unique values in a column called "Fav_genre"- "Rock", "Hip hop", "Pop", "Metal", "Country" This column contains value While dummy variable creation, we usually generate 5 columns. In this case, drop_first=True is not applicable. A person may have more than one favorite genre. So dropping any of the columns would not be right. Hence, drop_first=False is the default parameter.

C. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

By looking at the pair plot temp variable has the highest (0.63) correlation with target variable 'cnt'.

D. How did you validate the assumptions of Linear Regression after building the model on the training set (3 marks)

Linear regression relies on five main assumptions. Being able to verify and act upon them is

especially important. Linear regression is probably the most important model in Data Science.

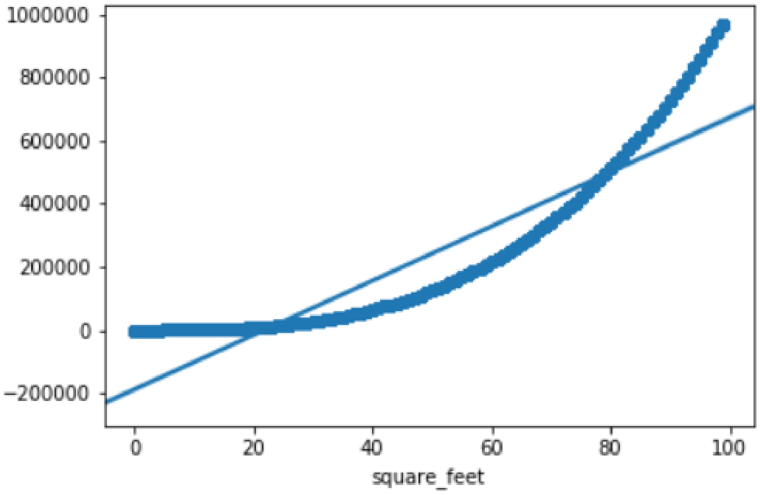
Despite its apparent simplicity, Linear regression is probably the most important model in Data Science it relies however on a few key assumptions (linearity, homoscedasticity, absence of multicollinearity, independence and normality of errors). Good knowledge of these is crucial to create and improve your model.

Linear Relationship

As obvious as this may seem, linear regression assumes that there exists a linear relationship between the dependent variable and the predictors.

How can it be verified?

Pair-wise scatterplots may be helpful in validating the linearity assumption as it is easy to visualize a linear relationship on a plot.



In the example above, the relationship between both variables is clearly not linear

In addition and similarly, a partial residual plot that represents the relationship between a predictor and the dependent variable while taking into account all the other variables may help visualize the “true nature of the relationship” between variables.

$$PartialResidual = Residual + \hat{\beta}_i X_i$$

What could it mean for the model if it is not respected?

If linearity is not respected, the regression will underfit and will not accurately model the relationship between the dependent and the independent variables.

What could be done?

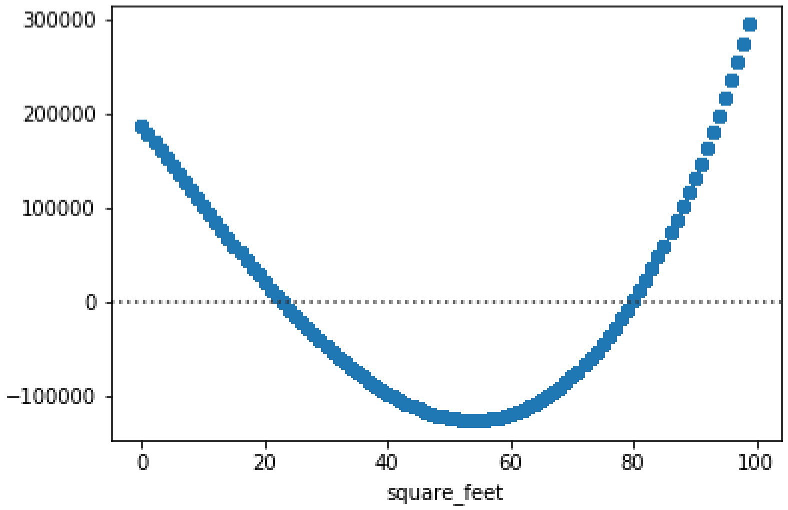
Independent variables and the dependent variables could be transformed so that the relationship between them is linear. For instance, you could find that the relationship is linear between the *log* of the dependent variables and some of the independent variables *squared* (c.f. *Polynomial Regression* and *Generalized Additive Models* (GAM) for an interesting generalization of this).

1. 2. Homoscedasticity

Homoscedasticity means that the residuals have constant variance no matter the level of the dependent variable.

How can it be verified?

To verify homoscedasticity, one may look at the residual plot and verify that the variance of the error terms is constant across the values of the dependent variable.



In the example above, there is heteroscedasticity as the variance of the residual is not constant

What could it mean for the model if it is not respected?

In the case of heteroscedasticity, the model will not fit all parts of the model equally and it will lead to biased predictions. It also often means that confounding variables, important predictors, have been omitted (it could also be due to the fact that the linearity assumption is not respected). While for the predictive context of data science, this may not be of utmost importance, heteroscedasticity is relatively more important in the context of inference, regarding the interpretability of the coefficients.

What could be done?

As heteroscedasticity generally reflects the absence of confounding variables, it can be tackled by reviewing the predictors and providing additional independent variables (and maybe even check that the linearity assumption is respected as well).

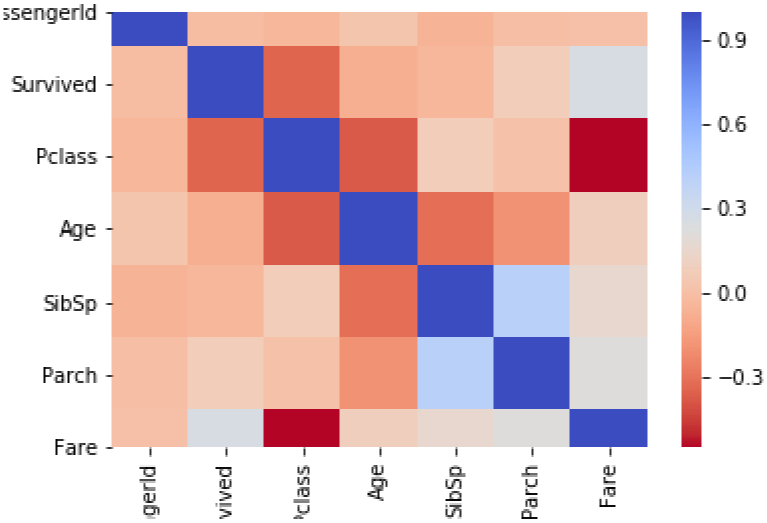
2. 3. Absence of Multicollinearity

Multicollinearity refers to the fact that two or more independent variables are highly correlated (or even redundant in the extreme case). While it may not be important for non-parametric methods, it is primordial for parametric models such as linear regression.

How can it be verified?

Often, a tell-tale sign of multicollinearity is the fact that some of the estimated coefficients have the “wrong” sign (i.e. the coefficient related to the size of a being negative in a model attempting to predict house prices).

Pairwise correlations could be the first step to identify potential relationships between various independent variables.



A correlation heatmaps may allow to quickly notice pair-wise correlations

A more thorough method, however, would be to look at the Variance Inflation Factors (VIF). It is calculated by regressing each independent variable on all the others and calculating a score as follows:

$$VIF = \frac{1}{1 - R^2}$$

Formula to calculate the VIF of an independent variable

Hence, if there exists a linear relationship between an independent variable and the others, it will imply a large *R-squared* for the regression and thus a larger VIF. As a rule of thumb, VIFs scores above 5 are generally indicators of multicollinearity (above 10 it can definitely be considered an issue).

What could it mean for the model if it is not respected?

The model may be producing inaccurate coefficient estimates that could thus not be interpreted. It may thus hurt inference power and possibly predictive performance.

In the presence of multicollinearity, the regression’s results may also become unstable and vary tremendously depending on the training data.

What could be done?

Multicollinearity can be fixed by performing feature selection: deleting one or more independent variables.

A common approach is to use backward subset-regression: start by building a regression with all the potential independent variables and iteratively remove variables with high VIF and using domain-specific knowledge.

Another method could be to isolate and keep only the interaction effects between multiple independent variables (using intuition or regularization generally).

As multicollinearity is reduced, the model will become more stable and the coefficients' interpretability will be improved.

3. 4. Independence of residuals (absence of auto-correlation)

Autocorrelation refers to the fact that observations' errors are correlated.

How can it be verified?

To verify that the observations are not auto-correlated, we can use the **Durbin-Watson test**. The test will output values between 0 and 4. The closer it is to 2, the less auto-correlation there is between the various variables (0–2: positive auto-correlation, 2–4: negative auto-correlation).

What could it mean for the model if it is not respected?

Auto-correlation could mean that the linearity of the relationship is not respected or that variables may have been omitted.

Auto-correlation would lead to spurious relationships between the independent variables and the dependent variable.

What could be done?

For time-series, one could add a lag variable. Another potential way to tackle this is to modify the variables from absolute value to relative change (i.e. instead of a stock price, it could be the change percentage from one period to the next).

More generally, variables should be further fine-tuned and added to the model.

4. 5. Normality of Errors

If the residuals are not normally distributed, Ordinary Least Squares (OLS), and thus the regression, may become biased.

How can it be verified?

To verify the normality of error, an easy way is to draw the distribution of residuals against levels of the dependent variable. One can use a QQ-plot and measure the divergence of the residuals from a normal distribution. If the resulting curve is not normal (i.e. is skewed), it may highlight a problem.

What could it mean for the model if it is not respected?

If it is not respected, it may highlight the presence of large outliers or highlight other assumptions being violated (i.e. linearity, homoscedasticity). As a result, calculating t-statistics and confidence intervals with the standard methodologies will become biased.

What could be done?

In the case where errors are not normally distributed, one could verify that the other assumptions are respected (i.e. homoscedasticity, linearity), as it may often be a tell-tale sign of such a violation, and fine-tune the model accordingly.

Otherwise, one should also attempt to treat the large outliers in the data and check if the data could not be separate subsets using different models.

In addition to the numerous assumptions enumerated above, it may be very relevant to verify that your linear regression is not extrapolating outside of the range of the training data and that there are no outliers or single records that may skew/have too much leverage on the regression (cf. Cook's distance and rule of thumbs for detecting outliers)

As highlighted throughout this post, despite its apparent simplicity, linear regression relies on numerous assumptions. When building a model, it is important to verify that they are being respected and to tackle potential violations in case they arise. I hope this guide will help you understand better the various assumptions behind the linear regression and give you the tools needed to tackle potential problems you may face as you use it.

E. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand for the shared bikes? (2 marks)

Temperature (temp) - A coefficient value of ‘0.5636’ indicated that a unit increase in temp variable increases the bike hire numbers by 0.5636 units.
Weather Situation 3 (weathersit_3) - A coefficient value of ‘-0.3070’ indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.3070 units.
Year (yr) - A coefficient value of ‘0.2308’ indicated that a unit increase in yr variable increases the bike hire numbers by 0.2308 units.

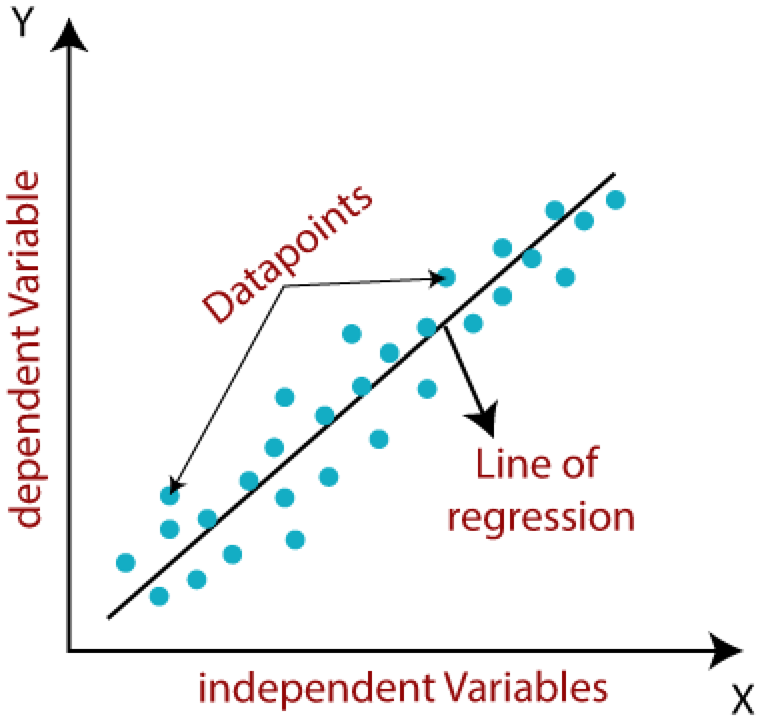
5. General Subjective Questions

A. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Mathematically, we can represent a linear regression as

$y= a_0+a_1x+ \varepsilon$

Here,

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

Types of Linear Regression

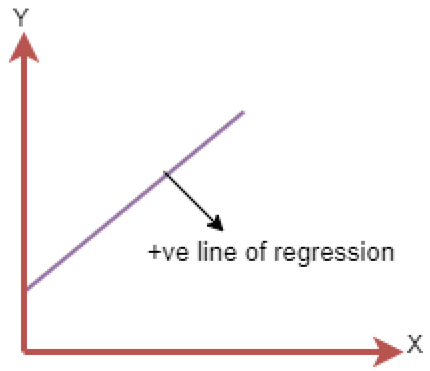
Linear regression can be further divided into two types of the algorithm:

Simple Linear Regression:
If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

Multiple Linear regression: *If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.*

Linear Regression Line

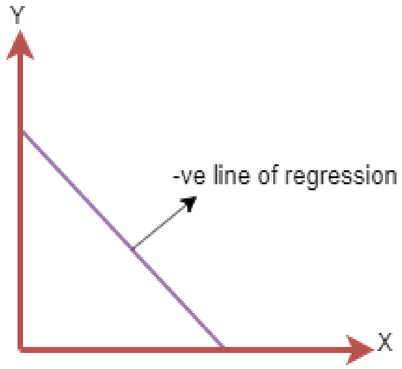
A linear line showing the relationship between the dependent and independent variables is called a **regression line**. A regression line can show two types of relationship:



The line equation will be: $\mathbf{Y= a_0+a_1x}$

Positive Linear Relationship:

If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



The line of equation will be: $\mathbf{Y= -a_0+a_1x}$

Negative Linear Relationship:

If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.

Finding the best fit line:

When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

The different values for weights or the coefficient of lines (a_0, a_1) gives a different line of regression, so we need to calculate the best values for a_0 and a_1 to find the best fit line, so to calculate this we use cost function.

Cost function-

- The different values for weights or coefficient of lines (a_0, a_1) gives the different line of regression, and the cost function is used to estimate the values of the coefficient for the best fit line.

- Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing.
- We can use the cost function to find the accuracy of the **mapping function**, which maps the input variable to the output variable. This mapping function is also known as **Hypothesis function**.

For Linear Regression, we use the **Mean Squared Error (MSE)** cost function, which is the average of squared error occurred between the predicted values and actual values. It can be written as:

For the above linear equation, MSE can be calculated as:

$$\text{MSE}= 1\frac{1}{N}\sum_{i=1}^{n}(y_i - (a_1x_i + a_0))^2$$

Where,

N=Total number of observation
Yi = Actual value
($a1x_i+a_0$)= Predicted value.

Residuals: The distance between the actual value and predicted values is called residual. If the observed points are far from the regression line, then the residual will be high, and so cost function will high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function.

Gradient Descent:

- Gradient descent is used to minimize the MSE by calculating the gradient of the cost function.
- A regression model uses gradient descent to update the coefficients of the line by reducing the cost function.
- It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

Model Performance:

The Goodness of fit determines how the line of regression fits the set of observations. The process of finding the best model out of various models is called **optimization**. It can be achieved by below method:

1. R-squared method:

- R-squared is a statistical method that determines the goodness of fit.
- It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%.
- The high value of R-square determines the less difference between the predicted values and actual values and hence represents a good model.
- It is also called a **coefficient of determination,** or **coefficient of multiple determination** for multiple regression.
- It can be calculated from the below formula:

$$\text{R-squared}=\frac{\textbf{Explained variation}}{\textbf{Total Variation}}$$

Assumptions of Linear Regression

Below are some important assumptions of Linear Regression. These are some formal checks while building a Linear Regression model, which ensures to get the best possible result from the given dataset.

- **Linear relationship between the features and target:** Linear regression assumes the linear relationship between the dependent and independent variables.
- **Small or no multicollinearity between the features:** Multicollinearity means high-correlation between the independent variables. Due to multicollinearity, it may difficult to find the true relationship between the predictors and target variables. Or we can say, it is difficult to determine which predictor variable is affecting the target variable and which is not. So, the model assumes either little or no multicollinearity between the features or independent variables.

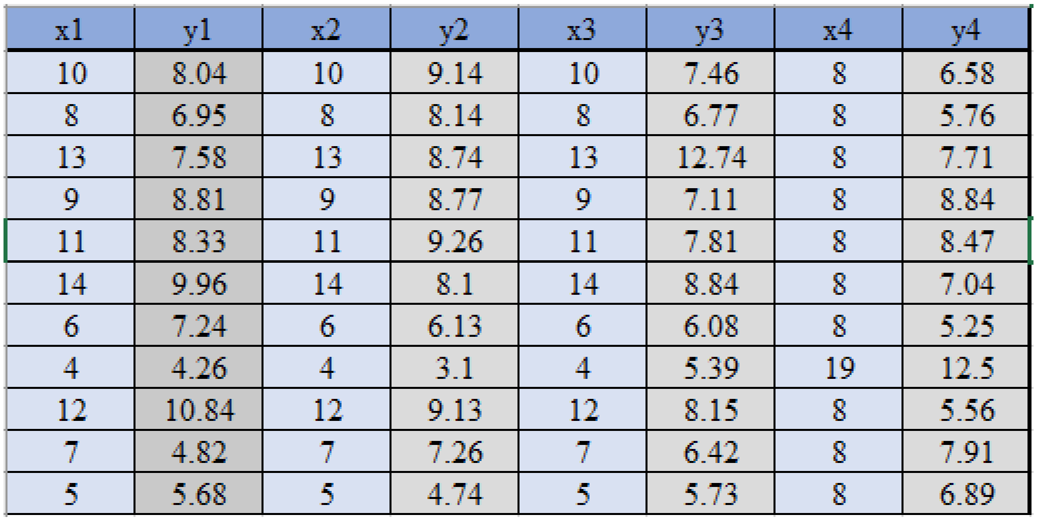
- Homoscedasticity Assumption:
  Homoscedasticity is a situation when the error term is the same for all the values of independent variables. With homoscedasticity, there should be no clear pattern distribution of data in the scatter plot.
- Normal distribution of error terms:
  Linear regression assumes that the error term should follow the normal distribution pattern. If error terms are not normally distributed, then confidence intervals will become either too wide or too narrow, which may cause difficulties in finding coefficients.
  It can be checked using the **q-q plot**. If the plot shows a straight line without any deviation, which means the error is normally distributed.

No autocorrelations:

The linear regression model assumes no autocorrelation in error terms. If there will be any correlation in the error term, then it will drastically reduce the accuracy of the model. Autocorrelation usually occurs if there is a dependency between residual errors.

B. Explain the Anscombe’s quartet in detail.(3 marks)

C. *Anscombe’s Quartet* is the modal example to demonstrate the importance of data visualization which was developed by the statistician *Francis Anscombe* in 1973 to signify both the importance of plotting data before analyzing it with statistical properties. It comprises of four data-set and each data-set consists of eleven (x,y) points. The basic thing to analyze about these data-sets is that they all share the same descriptive statistics(mean, variance, standard deviation etc) but different graphical representation. Each graph plot shows the different behavior irrespective of statistical analysis.



x1	y1	x2	y2	x3	y3	x4	y4
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

Four Data-sets

Apply the statistical formula on the above data-set,

Average Value of x = 9

Average Value of y = 7.50

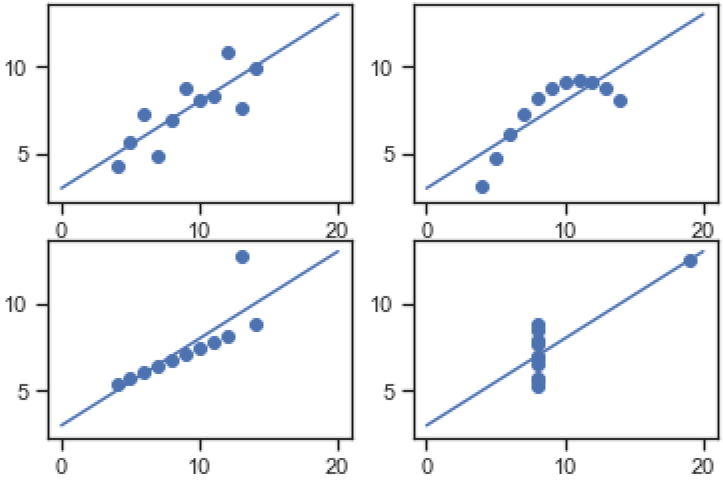
Variance of x = 11

Variance of y =4.12

Correlation Coefficient = 0.816

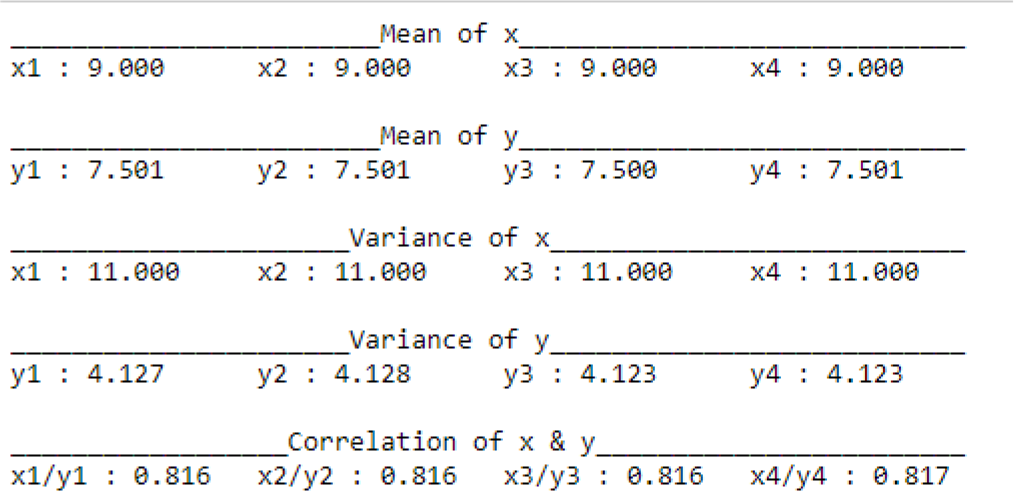
Linear Regression Equation : y = 0.5 x + 3

However, the statistical analysis of these four data-sets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represent the different behavior.



Graphical Representation of Anscombe's Quartet

- Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
- Data-set II — shows a curve shape but doesn't show a linear relationship (might be quadratic?).
- Data-set III — looks like a tight linear relationship between *x* and *y*, except for one large outlier.
- Data-set IV — looks like the value of *x* remains constant, except for one outlier as well.



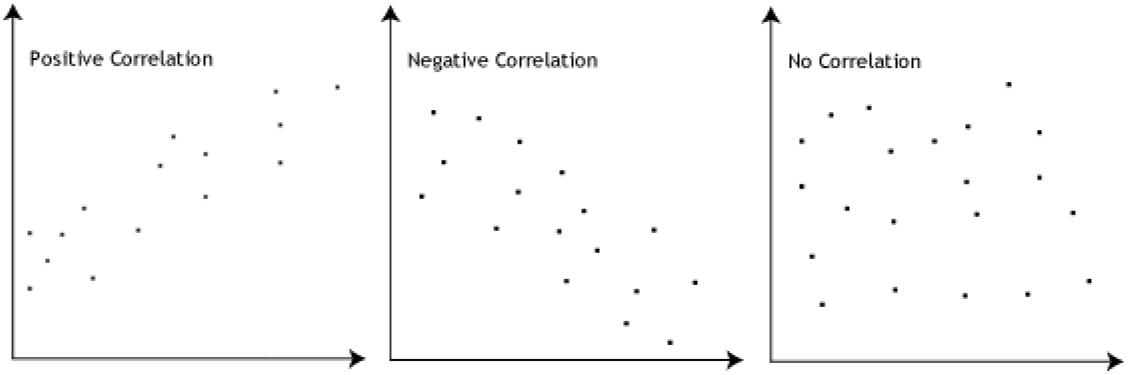
```
________________________Mean of x_____________________________
x1 : 9.000      x2 : 9.000      x3 : 9.000      x4 : 9.000

________________________Mean of y_____________________________
y1 : 7.501      y2 : 7.501      y3 : 7.500      y4 : 7.501

______________________Variance of x___________________________
x1 : 11.000     x2 : 11.000     x3 : 11.000     x4 : 11.000

______________________Variance of y___________________________
y1 : 4.127      y2 : 4.128      y3 : 4.123      y4 : 4.123

__________________Correlation of x & y________________________
x1/y1 : 0.816   x2/y2 : 0.816   x3/y3 : 0.816   x4/y4 : 0.817
```

D. What is Pearson’s R? (3 marks)

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association



E. Pearson r Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

- =correlation coefficient
- =values of the x-variable in a sample
- =mean of the values of the x-variable
- =values of the y-variable in a sample
- =mean of the values of the y-variable

F. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?(3 marks)

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

S.NO.	Normalisation	Standardisation
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.

S.NO.	Normalisation	Standardisation
7.	It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

G. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

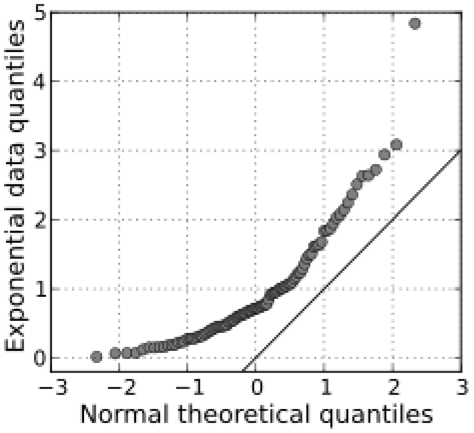
If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

H. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree 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.

A Q Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.