**Linear Regression Subjective Questions and Answers**

**1. Explain the linear regression algorithm in detail.**

Linear regression algorithm is used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable(s). Here, we establish relationship between independent and dependent variables by fitting a best line. This best fit line is known as regression line and represented by a linear equation **y= a\*X + b**.

The best way to understand linear regression is to relive this experience of childhood. Let us say, you ask a child in fifth grade to arrange people in his class by increasing order of weight, without asking them their weights! What do you think the child will do? He / she would likely look (visually analyze) at the height and build of people and arrange them using a combination of these visible parameters. This is linear regression in real life! The child has actually figured out that height and build would be correlated to the weight by a relationship, which looks like the equation above.

In this equation:

* Y – Dependent Variable
* a – Slope
* X – Independent variable
* b – Intercept

These coefficients ‘a’ and ‘b’ are derived based on minimizing the sum of squared difference of distance between data points and regression line.

**For** **example,** here we have identified the best fit line having linear equation **y=0.2811x+13.9**. Now using this equation, we can find the weight, knowing the height of a person.

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Linear_Regression.png)

**Linear Regression is mainly of two types**: Simple Linear Regression and Multiple Linear Regression. **Simple Linear Regression** is characterized by one independent variable. And, **Multiple Linear Regression** (as the name suggests) is characterized by multiple (more than 1) independent variables. While finding the best fit line, you can fit a polynomial or curvilinear regression. And these are known as polynomial or curvilinear regression.

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

**For example**, consider following Data Set which is used to explain the assumptions of linear regression:

The data set which is used is the Advertising data set. This data set contains information about money spent on advertisement and their generated Sales. Money was spent on TV, radio and newspaper ads. It has 3 features namely TV, radio and newspaper and 1 target Sales.

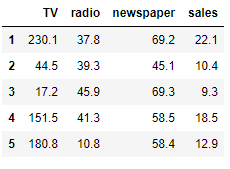
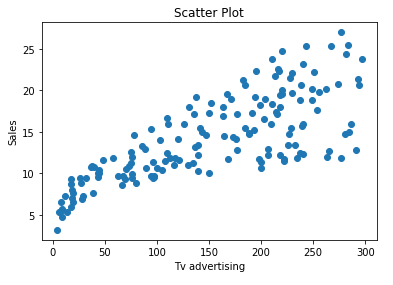


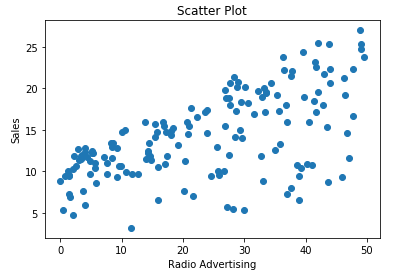
Fig: First 5 rows of the data set

**Assumptions of Linear Regression**

There are 5 basic assumptions of Linear Regression Algorithm:

**1. Linear Relationship between the features and target:** According to this assumption there is linear relationship between the features and target. Linear regression captures only linear relationship. This can be validated by plotting a scatter plot between the features and the target.





The first scatter plot of the feature TV vs Sales tells us that as the money invested on Tv advertisement increases the sales also increases linearly and the second scatter plot which is the feature Radio Vs Sales also shows a partial linear relationship between them, although not completely linear.

**2. Little or no Multicollinearity between the features:** Multicollinearity is a state of very high inter-correlations or inter-associations among the independent variables. It is therefore a type of disturbance in the data if present weakens the statistical power of the regression model. Pair plots and heatmaps (correlation matrix) can be used for identifying highly correlated features.

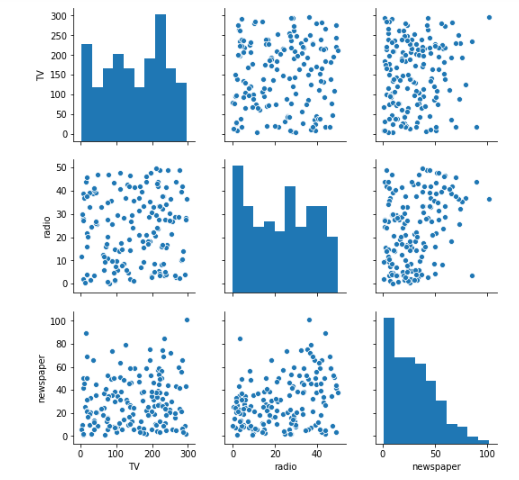


Fig: Pair plots of the features

The above pair plot shows no significant relationship between the features.

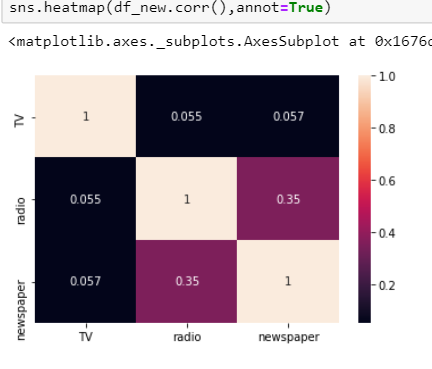


Fig: Heat Map (Correlation Matrix)

This heatmap gives us the correlation coefficients of each feature with respect to one another which are in turn less than 0.4. Thus the features aren’t highly correlated with each other.

**Removing highly correlated features is important:** The interpretation of a regression coefficient is that it represents the mean change in the target for each unit change in a feature when you hold all of the other features constant. However, when features are correlated, changes in one feature in turn shifts another feature/features. The stronger the correlation, the more difficult it is to change one feature without changing another. It becomes difficult for the model to estimate the relationship between each feature and the target independently because the features tend to change in unison.

**Multicollinearity should be treated:** If we have 2 features which are highly correlated we can drop one feature or combine the 2 features to form a new feature, which can further be used for prediction.

**3. Homoscedasticity Assumption:** Homoscedasticity describes a situation in which the error term (that is, the “noise” or random disturbance in the relationship between the features and the target) is the same across all values of the independent variables. A scatter plot of residual values vs predicted values is a good way to check for homoscedasticity. There should be no clear pattern in the distribution and if there is a specific pattern, the data is heteroscedastic.

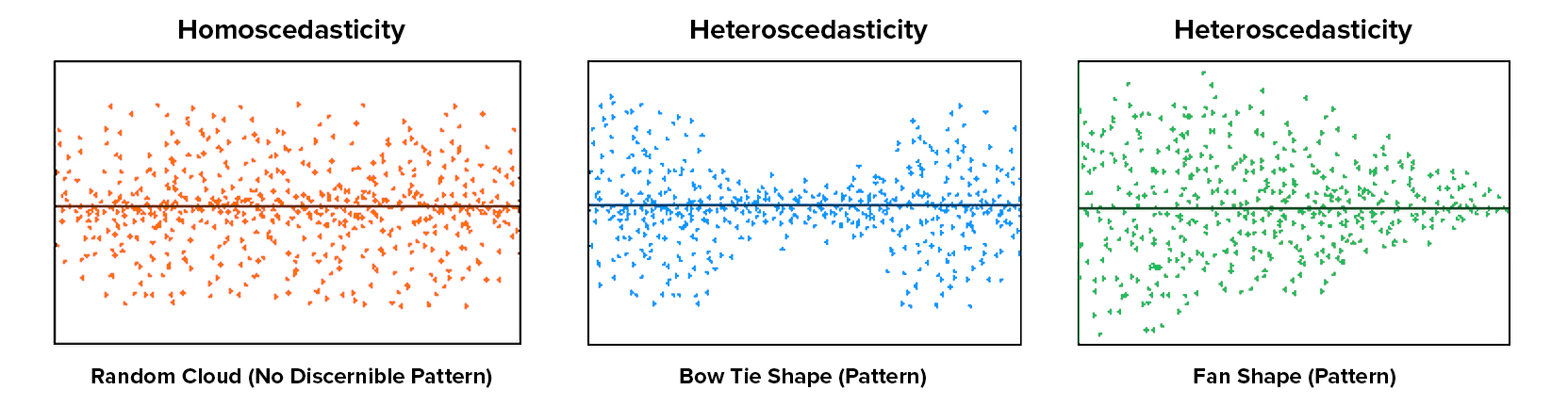


Fig: Homoscedasticity vs Heteroscedasticity

The leftmost graph shows no definite pattern i.e. constant variance among the residuals, the middle graph shows a specific pattern where the error increases and then decreases with the predicted values violating the constant variance rule and the rightmost graph also exhibits a specific pattern where the error decreases with the predicted values depicting heteroscedasticity

Python code for residual plot for the given data set:

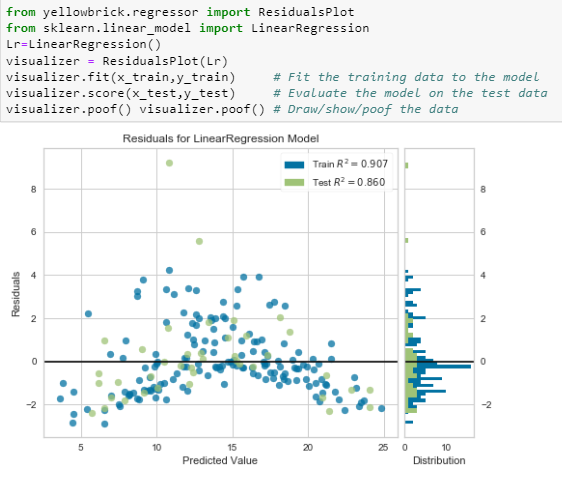


Fig: Error(residuals) vs Predicted values

**4. Normal distribution of error terms:** The fourth assumption is that the error(residuals) follow a normal distribution. However, a less widely known fact is that, as sample sizes increase, the normality assumption for the residuals is not needed. More precisely, if we consider repeated sampling from our population, for large sample sizes, the distribution (across repeated samples) of the ordinary least squares estimates of the regression coefficients follow a normal distribution. As a consequence, for moderate to large sample sizes, non-normality of residuals should not adversely affect the usual inferential procedures. This result is a consequence of an extremely important result in statistics, known as the central limit theorem.

Normal distribution of the residuals can be validated by plotting a q-q plot.

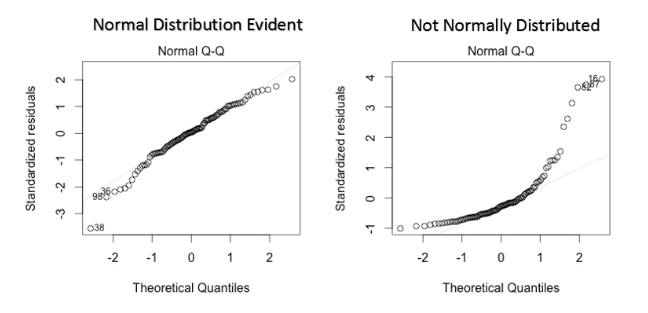


Fig: Q-Q Plots

Using the q-q plot we can infer if the data comes from a normal distribution. If yes, the plot would show fairly straight line. Absence of normality in the errors can be seen with deviation in the straight line.

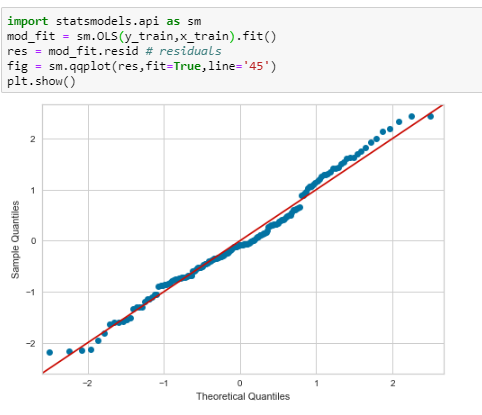
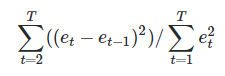


Fig: Q-Q Plot for the advertising data set

The q-q plot of the advertising data set shows that the errors(residuals) are fairly normally distributed. The histogram plot in the “Error(residuals) vs Predicted values” in assumption no.3 also shows that the errors are normally distributed with mean close to 0.

**5. Little or No autocorrelation in the residuals:** Autocorrelation occurs when the residual errors are dependent on each other. The presence of correlation in error terms drastically reduces model’s accuracy. This usually occurs in time series models where the next instant is dependent on previous instant.

Autocorrelation can be tested with the help of Durbin-Watson test. The null hypothesis of the test is that there is no serial correlation. The Durbin-Watson test statistics is defined as:



The test statistic is approximately equal to 2\*(1-r) where r is the sample autocorrelation of the residuals. Thus, for r == 0, indicating no serial correlation, the test statistic equals 2. This statistic will always be between 0 and 4. The closer to 0 the statistic, the more evidence for positive serial correlation. The closer to 4, the more evidence for negative serial correlation.

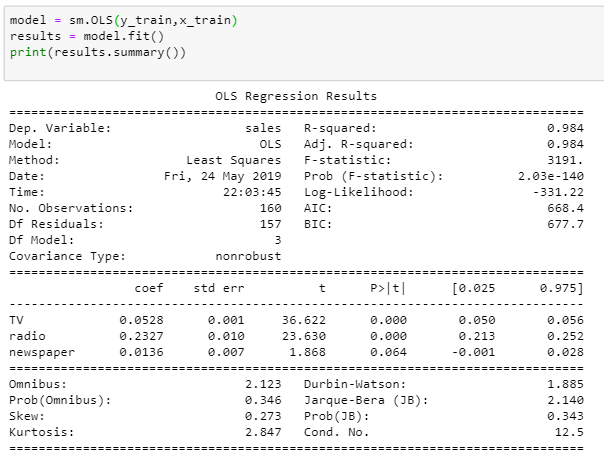


Fig: Summary of the fitted Linear Model

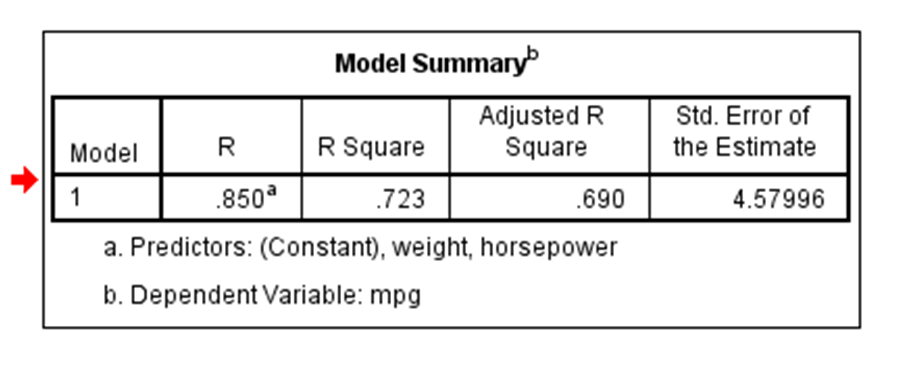
From the above summary note that the value of Durbin-Watson test is 1.885 quite close to 2 as said before when the value of Durbin-Watson is equal to 2, ‘r’ takes the value 0 from the equation 2\*(1-r), which in turn tells us that the residuals are not correlated.

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

**Coefficient of correlation** is “R” value which is given in the summary table in the Regression output. **Coefficient of determination** is also called as “R-square”. Multiply R times R to get the R-square value. In other words, Coefficient of Determination is the square of Coefficient of Correlation.

R-square or coefficient of determination shows percentage variation in y which is explained by all the x variables together. Higher the better. It is always between 0 and 1. It can never be negative – since it is a squared value.

It is easy to explain the R-square in terms of regression. It is not so easy to explain the R in terms of regression.

[](http://blog.uwgb.edu/bansalg/files/2013/11/Regression-model-summary-table.png)

Coefficient of Correlation is the R value i.e. .850 (or 85%). Coefficient of Determination is the R-square value i.e. .723 (or 72.3%). R-square is simply square of R i.e. R times R.

**Coefficient of Correlation:** is the degree of relationship between two variables say x and y. It can go between -1 and 1.  1 indicates that the two variables are moving in unison. They rise and fall together and have perfect correlation. -1 means that the two variables are in perfect opposites. One goes up and other goes down, in perfect negative way. Any two variables in this universe can be argued to have a correlation value. If they are not correlated, then the correlation value can still be computed which would be 0. The correlation value always lies between -1 and 1 (going thru 0 – which means no correlation at all – perfectly not related).

Correlation can be rightfully explained for simple linear regression – because you only have one x and one y variable. For multiple linear regression R is computed, but then it is difficult to explain because we have multiple variables involved here. That’s why R-square is a better term. You can explain R-square for both simple linear regressions and for multiple linear regressions.

**4. Explain the Anscombe’s quartet in detail.**

Anscombe’s Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize **COMPLETELY,**when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

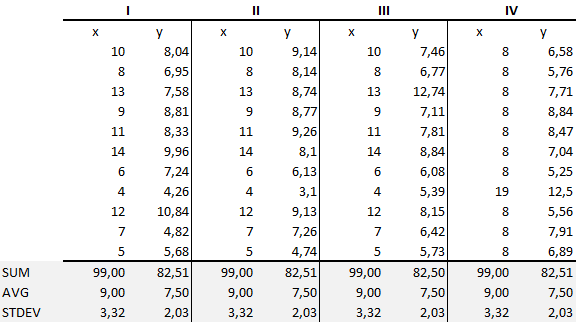
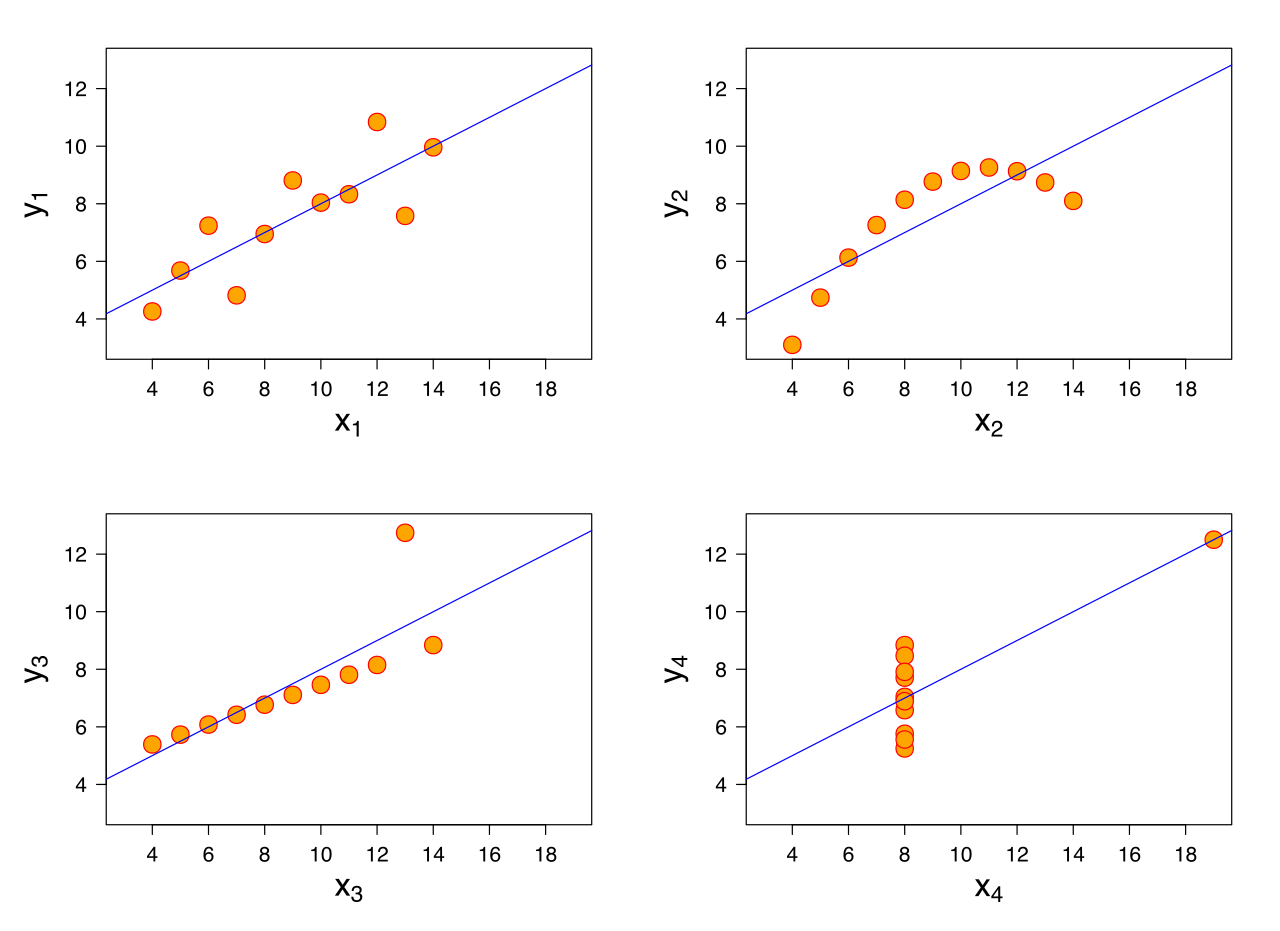


Fig: Quartet’s Summary Stats

The summary statistics show that the means and the variances were identical for x and y across the groups:

* Mean of x is 9 and mean of y is 7.50 for each dataset.
* Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
* The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



* Dataset I appears to have clean and well-fitting linear models.
* Dataset II is not 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.

**5. What is Pearson’s R?**

Pearson's correlation coefficient (r) for continuous (interval level) data ranges from -1 to +1:

|  |  |  |
| --- | --- | --- |
| r = -1 | data lie on a perfect straight line with a negative slope | Data lie on a perfect straight line with a negative slope. |
| r = 0 | no linear relationship between the variables | No linear relationship between the variables. |
| r = +1 | data lie on a perfect straight line with a positive slope | Data lie on a perfect straight line with a positive slope. |

Positive correlation indicates that both variables increase or decrease together, whereas negative correlation indicates that as one variable increases, so the other decreases, and vice versa.

**Significance**

The t-test is used to establish if the correlation coefficient is significantly different from zero, and, hence that there is evidence of an association between the two variables. There is then the underlying assumption that the data is from a normal distribution sampled randomly. If this is not true, the conclusions may well be invalidated. If this is the case, then it is better to use Spearman's coefficient of rank correlation (for non-parametric variables).

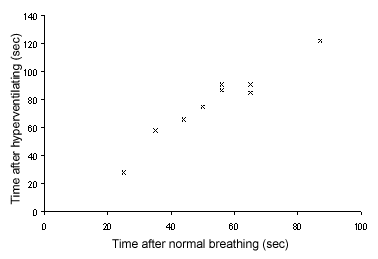
It is interesting to note that with larger samples, a low strength of correlation, **for example** r = 0.3, can be highly statistically significant (i.e. p < 0.01). However, is this an indication of a meaningful strength of association?

**Just because two variables are related, it does not necessarily mean that one directly causes the other.**

**For example**

Nine students held their breath, once after breathing normally and relaxing for one minute, and once after hyperventilating for one minute. The table indicates how long (in sec) they were able to hold their breath. Is there an association between the two variables?

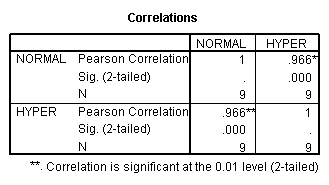
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Subject** | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** |
| **Normal** | 56 | 56 | 65 | 65 | 50 | 25 | 87 | 44 | 35 |
| **Hypervent** | 87 | 91 | 85 | 91 | 75 | 28 | 122 | 66 | 58 |



The chart shows the scatter plot (drawn in MS Excel) of the data, indicating the reasonableness of assuming a linear association between the variables.

Hyperventilating times are considered to be the dependent variable, so are plotted on the vertical axis.

**Output from SPSS and Minitab are shown below:**

**SPSS  
Select Analysis>Correlation>Bi-variate**  


**Minitab  
Correlations: Normal, Hypervent**

Pearson correlation of Normal and Hypervent = 0.966  
P-Value = 0.000

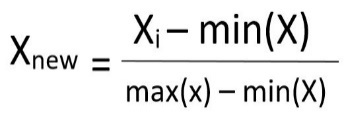
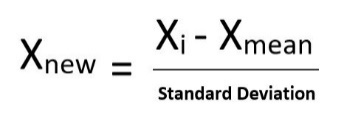
In conclusion, the printouts indicate that the strength of association between the variables is very high (r = 0.966), and that the correlation coefficient is very highly significantly different from zero (P < 0.001). Also, we can say that 93% (0.9662) of the variation in hyperventilating times is explained by normal breathing times.

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

**Feature Scaling** is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

**For example,** if an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that’s actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

**Techniques to perform Feature Scaling**  
Consider the two most important ones:

* **Min-Max Normalization:**This technique re-scales a feature or observation value with distribution value between 0 and 1.  
  
* **Standardization:**It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.  
  

**Consider following sample dataset:**

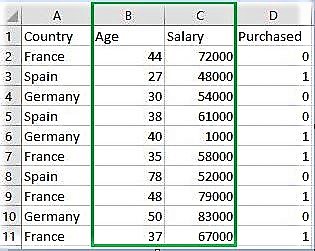


Fig: Sample data set

**Original data values:**

[[ 44 72000]

[ 27 48000]

[ 30 54000]

[ 38 61000]

[ 40 1000]

[ 35 58000]

[ 78 52000]

[ 48 79000]

[ 50 83000]

[ 37 67000]]

**After min max Scaling:**

[[ 0.33333333 0.86585366]

[ 0. 0.57317073]

[ 0.05882353 0.64634146]

[ 0.21568627 0.73170732]

[ 0.25490196 0.]

[ 0.15686275 0.69512195]

[ 1. 0.62195122]

[ 0.41176471 0.95121951]

[ 0.45098039 1.]

[ 0.19607843 0.80487805]]

**After Standardization:**

[[ 0.09536935 0.66527061]

[-1.15176827 -0.43586695]

[-0.93168516 -0.16058256]

[-0.34479687 0.16058256]

[-0.1980748 -2.59226136]

[-0.56487998 0.02294037]

[ 2.58964459 -0.25234403]

[ 0.38881349 0.98643574]

[ 0.53553557 1.16995867]

[-0.41815791 0.43586695]]

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

**VIF** is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. **For example**, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:  
  
X\_1=C+ α\_2 X\_2+α\_3 X\_3+⋯  
〖VIF〗\_1=1/(1-R\_1^2 )  
  
Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:  
  
X\_2=C+ α\_1 X\_1+α\_3 X\_3+⋯  
〖VIF〗\_2=1/(1-R\_2^2 )  
If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity.

**For example:**

VIF of X1 = 1/ (1 - R-squared of X1 on all other ‘X’s).

If you only have 1 X or that X is orthogonal with all the other ‘X’s; then

VIF = 1/ (1-0) = 1 - so no variance inflation

 If two ‘X’s are perfectly correlated

VIF = 1/ (1-1) = 1/0 = infinity that is the estimate is as imprecise as it can be.

**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.

**Gauss Markov Assumptions:**

There are five Gauss Markov assumptions (also called conditions):

1. **Linearity**: the parameters we are estimating using the OLS method must be themselves linear.
2. **Random**: our data must have been randomly sampled from the population.
3. **Non-Collinearity**: the regressors being calculated aren’t perfectly correlated with each other.
4. **Exogeneity**: the regressors aren’t correlated with the error term.
5. **Homoscedasticity**: no matter what the values of our regressors might be, the error of the variance is constant.

**Purpose of the Assumptions:**

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

Checking how well our data matches these assumptions is an important part of estimating regression coefficients. When you know where these conditions are violated, you may be able to plan ways to change your experiment setup to help your situation fit the ideal Gauss Markov situation more closely.

In practice, the Gauss Markov assumptions are **rarely all met perfectly**, but they are still useful as a benchmark, and because they show us what ‘ideal’ conditions would be. They also allow us to pinpoint problem areas that might cause our estimated regression coefficients to be inaccurate or even unusable.

**The Gauss-Markov Assumptions in Algebra:**

We can summarize the Gauss-Markov Assumptions succinctly in algebra, by saying that a linear regression model represented by

yi = xi‘ β + εi

and generated by the ordinary least squares estimate is the best linear unbiased estimate (BLUE) possible if

* E{εi} = 0, i = 1, . . . , N
* {ε1……εn} and {x1…..,xN} are independent
* cov{εi, εj} = 0, i, j = 1,…., N I ≠ j.
* V{ε1 = σ2, i= 1, ….N

**These assumptions can be read as:**

* The expected value of the error term is zero.
* The second assumption is collinearity.
* The third is exogeneity.
* The fourth is homoscedasticity.

**9. Explain the gradient descent algorithm in detail.**

**Gradient Descent** is an optimization algorithm used for minimizing the cost function in various machine learning algorithms. It is basically used for updating the parameters of the learning model.

**Types of gradient Descent:**

1. **Batch Gradient Descent:**This is a type of gradient descent which processes all the training examples for each iteration of gradient descent. But if the number of training examples is large, then batch gradient descent is computationally very expensive. Hence if the number of training examples is large, then batch gradient descent is not preferred. Instead, we prefer to use stochastic gradient descent or mini-batch gradient descent.
2. **Stochastic Gradient Descent:** This is a type of gradient descent which processes 1 training example per iteration. Hence, the parameters are being updated even after one iteration in which only a single example has been processed. Hence this is quite faster than batch gradient descent. But again, when the number of training examples is large, even then it processes only one example which can be additional overhead for the system as the number of iterations will be quite large.
3. **Mini Batch gradient descent:** This is a type of gradient descent which works faster than both batch gradient descent and stochastic gradient descent. Here b examples where b<m are processed per iteration. So even if the number of training examples is large, it is processed in batches of b training examples in one go. Thus, it works for larger training examples and that too with lesser number of iterations.

**For example, variables used:**  
Let ‘m’ be the number of training examples.  
Let ‘n’ be the number of features.

**Note:** if b == m, then mini batch gradient descent will behave similarly to batch gradient descent.

**Algorithm for batch gradient descent:**  
Let hθ(x) be the hypothesis for linear regression. Then, the cost function is given by:  
Let Σ represents the sum of all training examples from i=1 to m.

Jtrain(θ) = (1/2m) Σ( hθ(x(i)) - y(i))2

Repeat {

θj = θj – (learning rate/m) \* Σ( hθ(x(i)) - y(i))xj(i)

For every j =0 …n

}

Where xj(i) Represents the jth feature of the ith training example. So if m is very large, then the derivative term fails to converge at the global minimum.

**Algorithm for stochastic gradient descent:**

1) Randomly shuffle the data set so that the parameters can be trained evenly for each type of data.  
2) As mentioned above, it takes into consideration one example per iteration.

Hence,

Let (x(i),y(i)) be the training example

Cost(θ, (x(i),y(i))) = (1/2) Σ( hθ(x(i)) - y(i))2

Jtrain(θ) = (1/m) Σ Cost(θ, (x(i),y(i)))

Repeat {

For i=1 to m{

θj = θj – (learning rate) \* Σ( hθ(x(i)) - y(i))xj(i)

For every j =0 …n

} }

**Algorithm for mini batch gradient descent:**

Say ‘b’ be the no of examples in one batch, where b < m. Assume b = 10, m = 100; **Note:** However, we can adjust the batch size. It is generally kept as power of 2. The reason behind it is because some hardware such as GPUs achieve better run time with common batch sizes such as power of 2.

Repeat {

For i=1,11, 21,…..,91

Let Σ be the summation from i to i+9 represented by k.

θj = θj – (learning rate/size of (b) ) \* Σ( hθ(x(k)) - y(k))xj(k)

For every j =0 …n

}

**Convergence trends in different variants of Gradient Descents:**

**In case of Batch Gradient Descent,** the algorithm follows a straight path towards the minimum. If the cost function is convex, then it converges to a global minimum and if the cost function is not convex, then it converges to a local minimum. Here the learning rate is typically held constant.

**In case of stochastic gradient Descent and mini-batch gradient descent,** the algorithm does not converge but keeps on fluctuating around the global minimum. Therefore, in order to make it converge, we have to slowly change the learning rate. However, the convergence of Stochastic gradient descent is much noisier as in one iteration, it processes only one training example.

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

A **Quantile-Quantile** plot (also known as a QQ-plot) is another way you can determine whether a dataset matches a specified probability distribution. QQ-plots are often used to determine whether a dataset is normally distributed. Graphically, the QQ-plot is very different from a histogram. As the name suggests, the horizontal and vertical axes of a QQ-plot are used to show quantiles.

Quartiles divide a dataset into four equal groups, each consisting of 25 percent of the data. But there is nothing particularly special about the number four. You can choose any number of groups you please.

Another popular type of quantile is the percentile, which divides a dataset into 100 equal groups. **For example**, the 30th percentile is the boundary between the smallest 30 percent of the data and the largest 70 percent of the data. The median of a dataset is the 50th percentile of the dataset. The 25th percentile is the first quartile, and the 75th percentile the third quartile.

With a QQ-plot, the quantiles of the sample data are on the vertical axis, and the quantiles of a specified probability distribution are on the horizontal axis. The plot consists of a series of points that show the relationship between the actual data and the specified probability distribution. If the elements of a dataset perfectly match the specified probability distribution, the points on the graph will form a 45-degree line.

**For example**, this figure shows a normal QQ-plot for the price of Apple stock from January 1, 2013 to December 31, 2013.

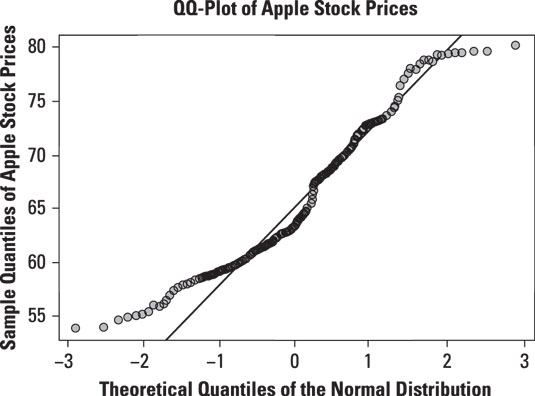


Fig: Normal QQ-plot of daily prices for Apple stock.

The QQ-plot shows that the prices of Apple stock do not conform very well to the normal distribution. In particular, the deviation between Apple stock prices and the normal distribution seems to be greatest in the lower left-hand corner of the graph, which corresponds to the left tail of the normal distribution. The discrepancy is also noticeable in the upper right-hand corner of the graph, which corresponds to the right tail of the normal distribution.

The graph shows that the smallest prices of Apple stock are not small enough to be consistent with the normal distribution; similarly, the largest prices of Apple stock are not large enough to be consistent with the normal distribution. This shows that the tails of the Apple stock price distribution are too “**thin**” or “**skinny**” compared with the normal distribution. The conclusion to be drawn from this is that the Apple stock prices are not normally distributed.

This figure shows a normal QQ-plot for the daily returns to Apple stock from January 1, 2013 to December 31, 2013:

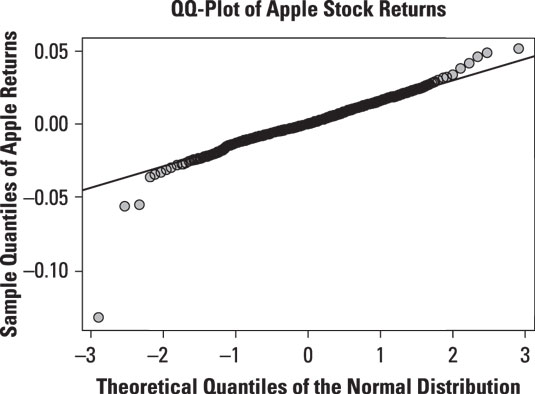


Fig: Normal QQ-plot of daily returns to Apple stock.

The QQ-plot shows that the returns to Apple stock do not conform to the normal distribution, either. In this case, the smallest returns to Apple stock are too small to be consistent with the normal distribution. Similarly, the largest returns to Apple stock are too large to be consistent with the normal distribution. This shows that the tails of the Apple return distribution are too “**thick**” or “**fat**” compared with the normal distribution. Therefore, Apple returns are not normally distributed.