Subjective Questions on Linear Regression

1. Explain the linear regression algorithm in detail.

Linear regression is a simple approach to supervised learning. It assumes that the dependence of Y on X1, X2, . . . Xp is linear.

Types of Linear Regression

- Simple Linear Regression (SLR)
- Multiple Linear Regression (MLR)

Least Square Method – Finding the best fit line

Least squares is a statistical method used to determine the best fit line or the regression line by minimizing the sum of squares created by a mathematical function. The “square” here refers to squaring the distance between a data point and the regression line. The line with the minimum value of the sum of square is the best-fit regression line. Regression Line, y = mx+c where,

y = Dependent Variable

x= Independent Variable ; c = y-Intercept

$$m = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

Examples where Linear Regression can be applied

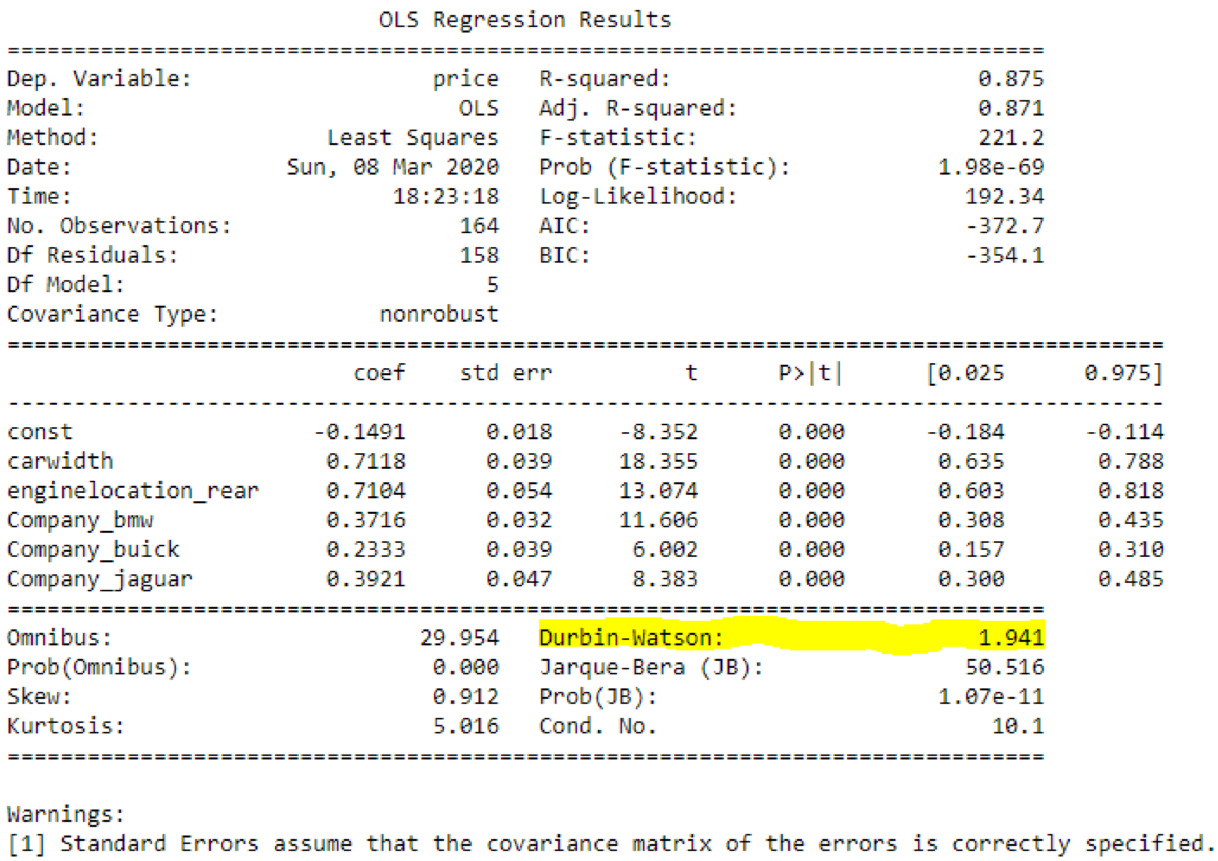
- Given a dataset of BMI (body mass index) and the fat percentage of the customers of a fitness center. the fitness center wants to predict the fat percentage of a new customer. This is a typical example of Linear Regression where the dependent variable to be predicted is numeric in nature.

- Predict the sales of a retail store based on its size, given the dataset of sales of retail stores and their sizes. This is yet another example of linear regression where the sales is a numeric in nature

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

- There should be no correlation between the residual (error) terms.
- Absence of this phenomenon is known as **Autocorrelation**.
- 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. If the error terms are correlated, the estimated standard errors tend to underestimate the true standard error.
- Also, lower standard errors would cause the associated p-values to be lower than actual. This will make us incorrectly conclude a parameter to be statistically significant.
- How to check: Look for **Durbin – Watson (DW) statistic**. It must lie between 0 and 4. If DW = 2, implies no autocorrelation, 0 < DW < 2 implies positive autocorrelation while 2 < DW < 4 indicates negative autocorrelation. Also, you can see residual vs time plot and look for the seasonal or correlated pattern in residual values.

  In our car assignment, the D-W values hover around 1.95, which indicates we don't have Autocorrelation



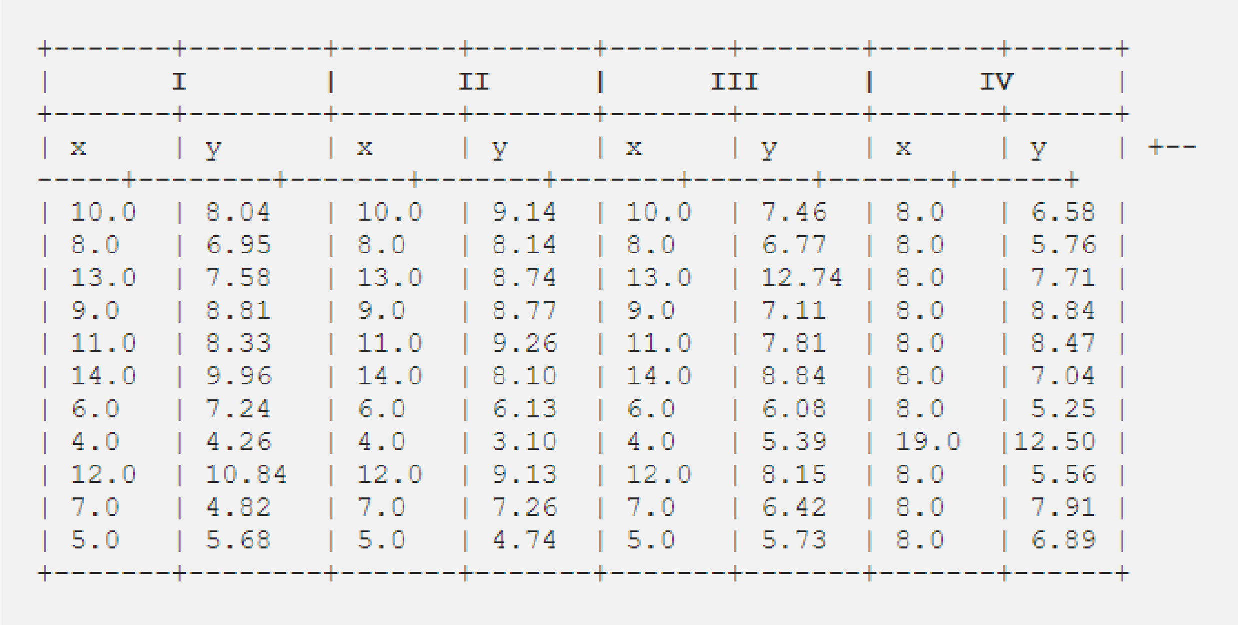
```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                  price   R-squared:                       0.875
Model:                            OLS   Adj. R-squared:                  0.871
Method:                 Least Squares   F-statistic:                     221.2
Date:                Sun, 08 Mar 2020   Prob (F-statistic):           1.98e-69
Time:                        18:23:18   Log-Likelihood:                 192.34
No. Observations:                 164   AIC:                            -372.7
Df Residuals:                     158   BIC:                            -354.1
Df Model:                           5                                         
Covariance Type:            nonrobust                                         
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                  -0.1491      0.018     -8.352      0.000      -0.184      -0.114
carwidth                0.7118      0.039     18.355      0.000       0.635       0.788
enginelocation_rear     0.7104      0.054     13.074      0.000       0.603       0.818
Company_bmw             0.3716      0.032     11.606      0.000       0.308       0.435
Company_buick           0.2333      0.039      6.002      0.000       0.157       0.310
Company_jaguar          0.3921      0.047      8.383      0.000       0.300       0.485
==============================================================================
Omnibus:                       29.954   Durbin-Watson:                   1.941
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               50.516
Skew:                           0.912   Prob(JB):                     1.07e-11
Kurtosis:                       5.016   Cond. No.                         10.1
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

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.
- R square is also called coefficient of determination. Multiply R times R to get the R square value. **In other words, Coefficient of Determination is the square of Coefficient of Correlation.**
- The strength of a linear regression model is mainly explained by R^2, where $R^2 = 1 - (RSS/TSS)$ where RSS: Residual sum of squares and TSS: Total sum of squares

4. Explain the Anscombe’s quartet in detail.

- The trend of ignoring the visualizations when we have the summary statistics with us can be horrible with respect to predictions.
- Anscombe, a great statistician identified this trend and in order to eliminate this dangerous trend, he created specific data points **(4 sets of 11 data points)** whose summary statistics such as mean, standard deviation, correlation are almost intact

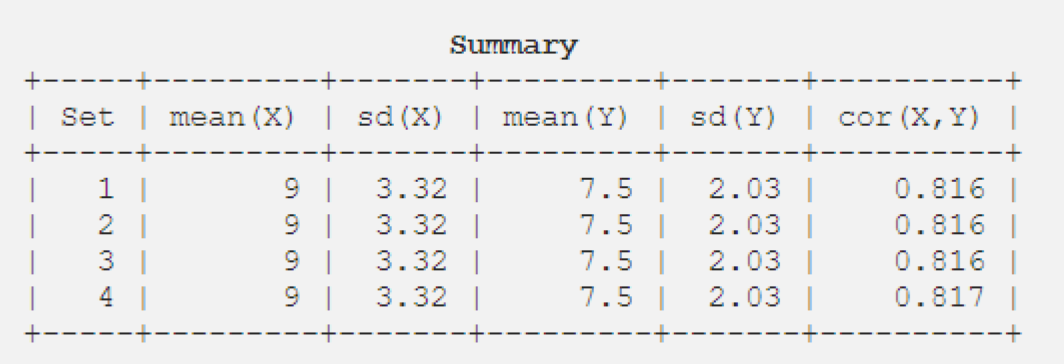


```
+-------+--------+-------+-------+-------+-------+-------+-------+------+
|       I        |       II      |      III      |       IV      |
+-------+--------+-------+-------+-------+-------+-------+-------+------+
| x     | y      | x     | y     | x     | y     | x     | y     |  +--
-----+--------+-------+-------+-------+-------+-------+------+
| 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  |
+-------+--------+-------+-------+-------+-------+-------+-------+------+
```

Anscombe's Quartet 1

The above figure indicates the 4 sets he created with 11 data points. This is called as **Anscombe’s quartet**

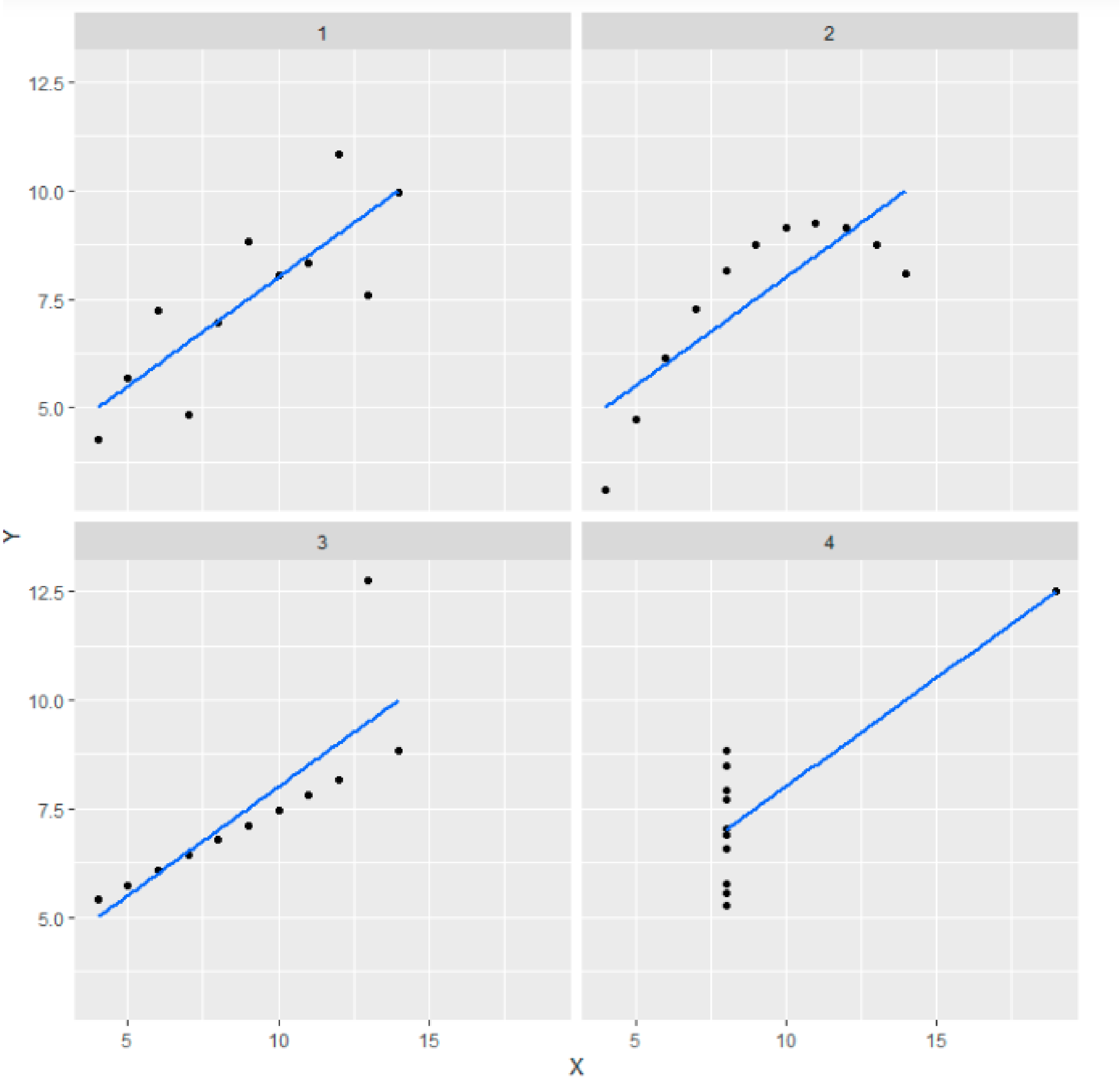
The summary statistics of the data points gave the following



```
                          Summary
+-----+---------+-------+---------+-------+----------+
| Set | mean(X) | sd(X) | mean(Y) | sd(Y) | cor(X,Y) |
+-----+---------+-------+---------+-------+----------+
|   1 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   2 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   3 |       9 |  3.32 |     7.5 |  2.03 |    0.816 |
|   4 |       9 |  3.32 |     7.5 |  2.03 |    0.817 |
+-----+---------+-------+---------+-------+----------+
```

As seen in the above Summary Statistics image, the mean, standard deviation and correlation between two variables are almost identical.

However, on plotting the actual data points, it revealed something different



Anscombe's quartet — Plot

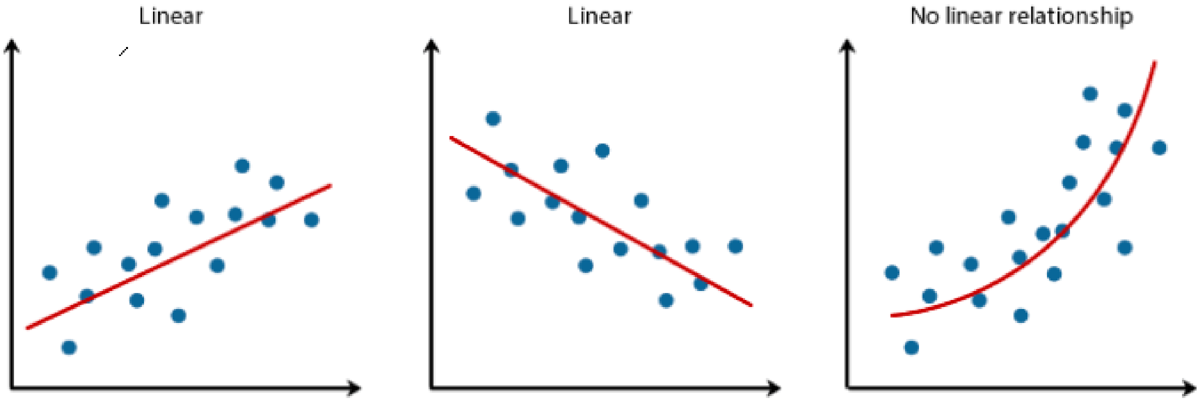
The graphs are completely different whereas the statistics are identical.

Quote of the question: To have analysis that you can trust,
Plot your data, plot you must.

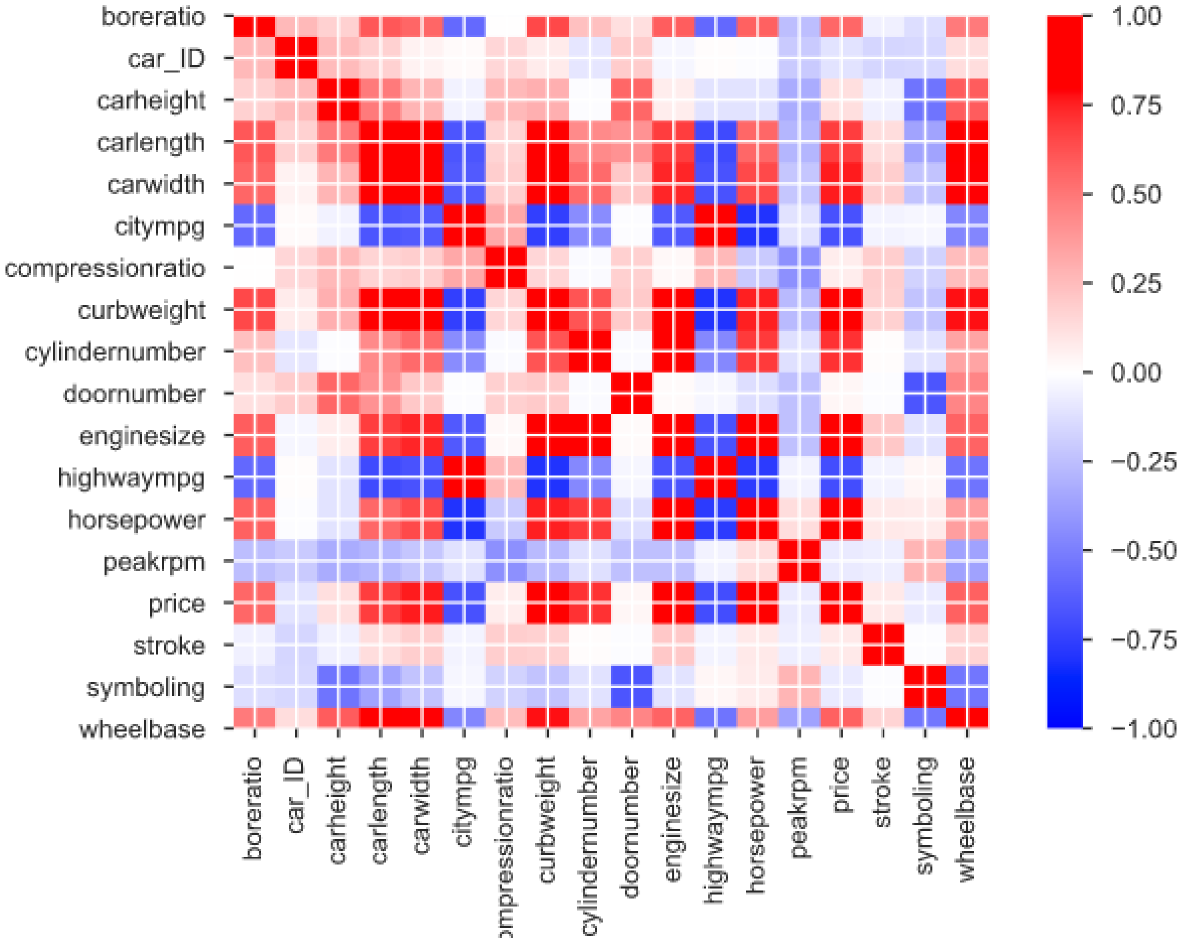
Reference Used: https://towardsdatascience.com/fables-of-data-science-anscombes-quartet-2c2e1a07fbe6

5. What is Pearson's R?

- Correlation is an analysis that indicates the strength of association between two variables and the direction of the relationship.
- In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and -1, where +1 indicates a perfectly positive correlation between the variables and -1 indicates a perfectly negative correlation between variables
- Perason's R or Perason's correlation is a type of correlation which mainly deals with **linear association** between two variables
- Pearson's R assumes that both the variables are normally distributed (that is follows the Gaussian curve with mean as 0.0)
- The two variables in question should have a linear relationship. That is, changes in one variable should linearly affect the other variable, as seen in first two plots (left to right) of below image



- The variables should be in Homoscedascity, i.e: equal variance



Pearson's R in Car Assignment 1

The above heatmap reveals the Pearson’s R in the car assignment that we performed. Based on this we inferred that

1. Price (Dependent variable) is highly positively correlated with Horse power, car width, curbweight and Engine Size
2. Price (Dependent variable) is highly negatively correlated with highwaympg and citympg

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.
- Different Types of Scaling are MinMax Scaling (Normalized) and Standard Scaling

MinMaxScaling

- This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

Standard Scaling

- 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.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

In our car assignment we have used MinMax Scaling to scale the train and test data.

It is important to note that, on train data, we fit_transform the data whereas on the test data, we transform the data to scaled values

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

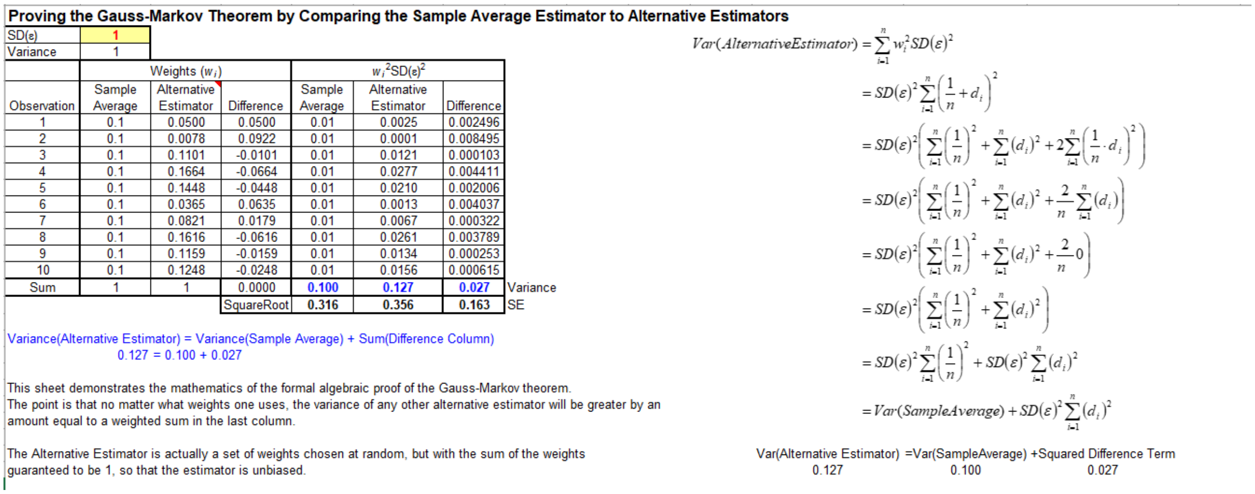
- Variance Inflation Factor is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity.
- In order to determine VIF, we fit a regression model between the independent variables.
- If there is no correlation, then VIF is 1

- **If there is perfect correlation, then VIF = infinity.**
- A large value of VIF indicates that there is a correlation between the variables.
- If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.
- Generally, a VIF until 5 is acceptable, whereas VIF greater than 10 indicates that model has serious multicollinearity issues.

8. What is the Gauss-Markov theorem?

- The Gauss Markov theorem says that, under certain conditions, the ordinary least squares (OLS) estimator of the coefficients of a linear regression model is the best linear unbiased estimator that is, the estimator that has the smallest variance among those that are unbiased and linear in the observed output variables.
- The **mathematical expectation** of each error is assumed to be zero, and all of them have the same (unknown) variance.
- "OLS is BLUE" (best linear unbiased estimator) and is known as the Gauss–Markov theorem
- The Gauss–Markov theorem also works in reverse: when the data generating process does not follow the classical econometric model, ordinary least squares is typically no longer the preferred estimator.

The below image shows the Proof for Gauss-Markow theorem in UniVariate Analysis



Proving the Gauss-Markov Theorem by Comparing the Sample Average Estimator to Alternative Estimators

SD(ε)	1
Variance	1

	Weights (w_i)			w_i^2SD(ε)^2			
Observation	Sample Average	Alternative Estimator	Difference	Sample Average	Alternative Estimator	Difference	
1	0.1	0.0500	0.0500	0.01	0.0025	0.002496	
2	0.1	0.0078	0.0922	0.01	0.0001	0.008495	
3	0.1	0.1101	-0.0101	0.01	0.0121	0.000103	
4	0.1	0.1664	-0.0664	0.01	0.0277	0.004411	
5	0.1	0.1448	-0.0448	0.01	0.0210	0.002006	
6	0.1	0.0365	0.0635	0.01	0.0013	0.004037	
7	0.1	0.0821	0.0179	0.01	0.0067	0.000322	
8	0.1	0.1616	-0.0616	0.01	0.0261	0.003789	
9	0.1	0.1159	-0.0159	0.01	0.0134	0.000253	
10	0.1	0.1248	-0.0248	0.01	0.0156	0.000615	
Sum	1	1	0.0000	**0.100**	**0.127**	**0.027**	Variance
			SquareRoot	**0.316**	**0.356**	**0.163**	SE

Variance(Alternative Estimator) = Variance(Sample Average) + Sum(Difference Column)
0.127 = 0.100 + 0.027

This sheet demonstrates the mathematics of the formal algebraic proof of the Gauss-Markov theorem.
The point is that no matter what weights one uses, the variance of any other alternative estimator will be greater by an amount equal to a weighted sum in the last column.

The Alternative Estimator is actually a set of weights chosen at random, but with the sum of the weights guaranteed to be 1, so that the estimator is unbiased.

$$
\begin{aligned}
Var(AlternativeEstimator) &= \sum_{i=1}^{n} w_i^2 SD(\varepsilon)^2 \\
&= SD(\varepsilon)^2 \sum_{i=1}^{n}\left(\frac{1}{n}+d_i\right)^2 \\
&= SD(\varepsilon)^2\left(\sum_{i=1}^{n}\left(\frac{1}{n}\right)^2+\sum_{i=1}^{n}(d_i)^2+2\sum_{i=1}^{n}\left(\frac{1}{n}\cdot d_i\right)^2\right) \\
&= SD(\varepsilon)^2\left(\sum_{i=1}^{n}\left(\frac{1}{n}\right)^2+\sum_{i=1}^{n}(d_i)^2+\frac{2}{n}\sum_{i=1}^{n}(d_i)\right) \\
&= SD(\varepsilon)^2\left(\sum_{i=1}^{n}\left(\frac{1}{n}\right)^2+\sum_{i=1}^{n}(d_i)^2+\frac{2}{n}0\right) \\
&= SD(\varepsilon)^2\left(\sum_{i=1}^{n}\left(\frac{1}{n}\right)^2+\sum_{i=1}^{n}(d_i)^2\right) \\
&= SD(\varepsilon)^2\sum_{i=1}^{n}\left(\frac{1}{n}\right)^2+SD(\varepsilon)^2\sum_{i=1}^{n}(d_i)^2 \\
&= Var(SampleAverage)+SD(\varepsilon)^2\sum_{i=1}^{n}(d_i)^2
\end{aligned}
$$

Var(Alternative Estimator) =Var(SampleAverage) +Squared Difference Term
0.127 0.100 0.027

References: Various educational sites

9. Explain the gradient descent algorithm in detail.

- Gradient descent is an optimisation algorithm that optimises the objective function (cost function for linear regression) to reach the optimal solution
- Given that the equation for the line that's fit the data as: **y(p)=β0+β1x** where β0 is the intercept of the fitted line and β1 is the coefficient for the independent variable x, the Gradient Descent algorithm finds the optimal betas and thetas which reduce the overall cost function of the mode.
- Gradient descent is an iterative form solution of order one. So to compute optimal thetas, we need to apply Gradient Descent to the Cost function, which is given as follows: ∂∂θJ(θ)

Process of Computing Gradient Descent in Python

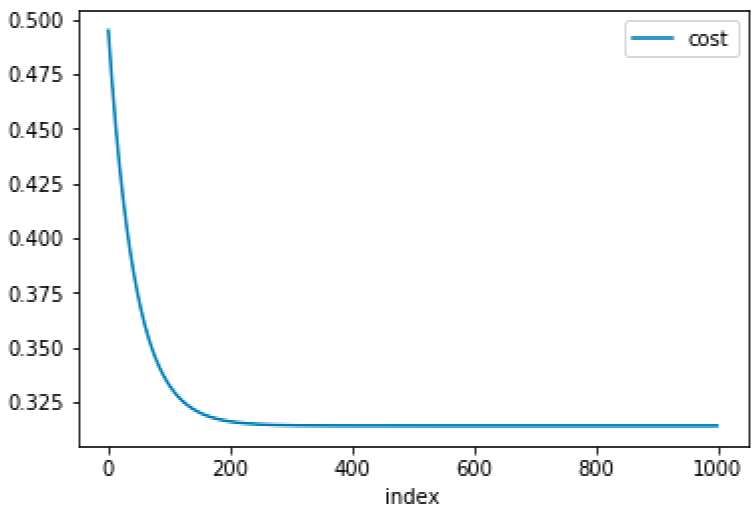
- Import numpy and pandas
- Assign feature variable X and response variable y
- Add a columns of 1s as an intercept to X.
- Convert X and y to arrays
- Define theta (vector representing coefficients (intercept, other features))
- Define cost function

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_0^{(i)}$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_1^{(i)}$$

$$\theta_2 := \theta_2 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_2^{(i)}$$

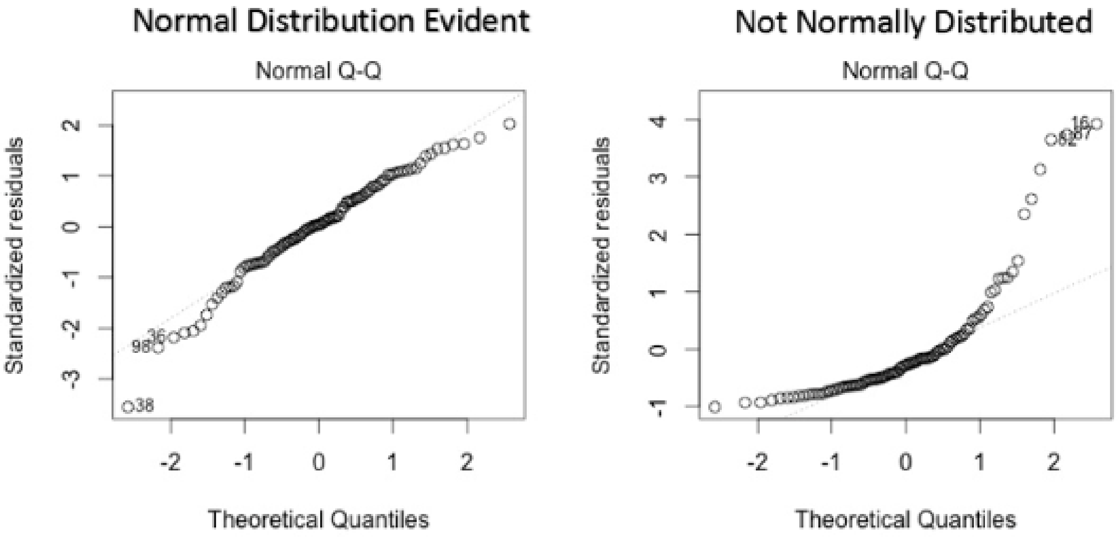
- Define gradient descent function whose arguments are X, y, learning rate alpha, theta and iterations
- The function returns the cost as a data frame
- Print costs with various values of coefficients b0, b1, b2
- Plot the costs for each occurrence



As seen in the above image, the cost gradually reduce to zero almost as the indexes increase.

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

- If the error terms are non- normally distributed, confidence intervals may become too wide or narrow.
- Once confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on minimization of least squares.
- Presence of non – normal distribution suggests that there are a few unusual data points which must be studied closely to make a better model.
- The presence of non- normal distribution can be analyzed using a Q-Q plot



Normal Distribution Evident

Not Normally Distributed

Image Source: Wikipedia

- This q-q or quantile-quantile is a scatter plot which helps us validate the assumption of normal distribution in a data set.
- Using this 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.