Q1 List down at least three main assumptions of linear regression and explain them in your own words. To explain an assumption, take an example or a specific use case to show why the assumption makes sense.

Answer 1

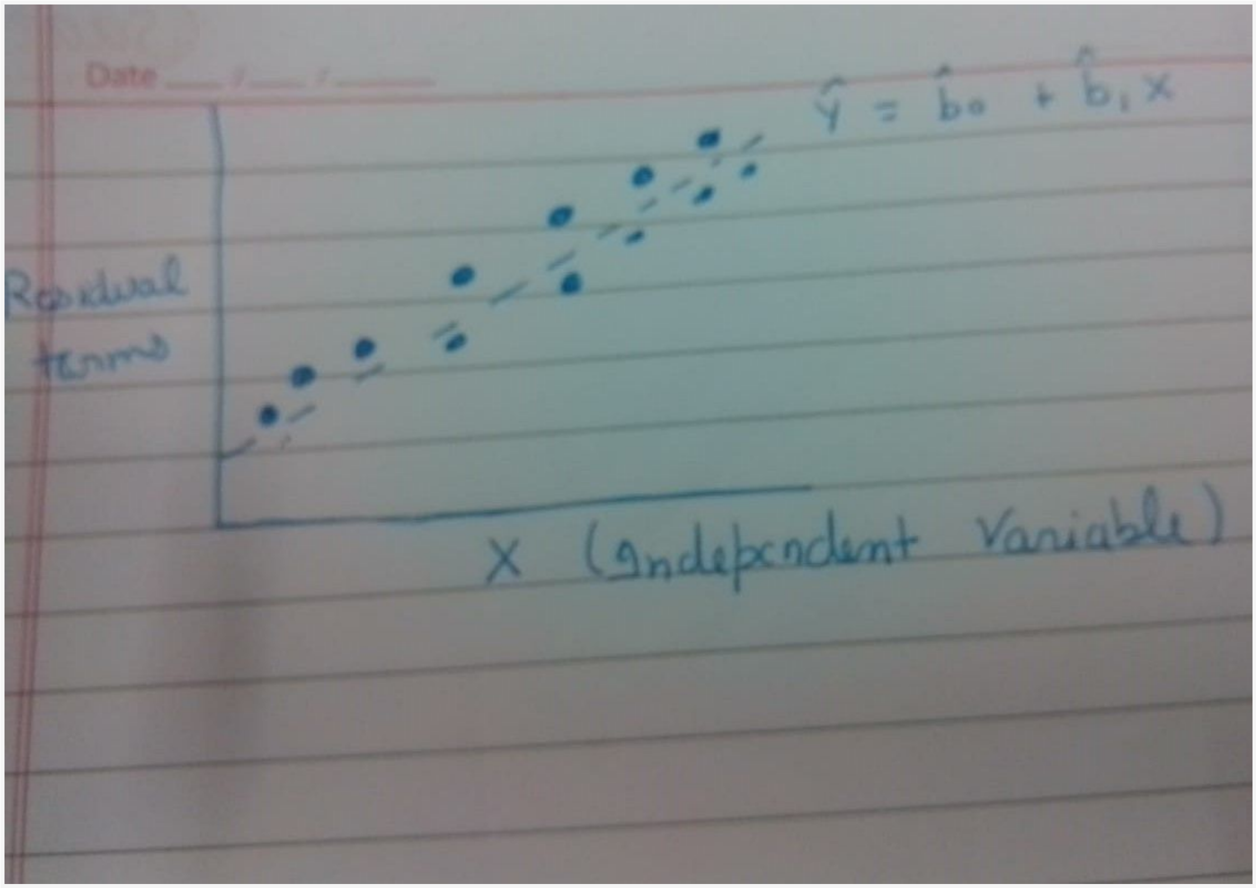
Assumptions:

1) **Linearity ⇒ According to Linear Regression concept relationship between dependent variable and independent variable should be linear if it is not linear then your model will not give you the correct prediction output. To Analyse the Relationship between dependent variable and independent variable we use Scatter plot. In Multiple regression one variable is not linear we will try to get something which is linear.**
2) **Heteroskedasticity ⇒ When the constant variance in residual term is not present then Heteroskedasticity comes in to action. Heteroskedasticity are of two types 1) Conditional 2) Unconditional**

   Unconditional Heteroskedasticity ⇒ is present when the variance of residual terms are not related to the values of independent variable. So it is not an issue for Linear Regression.

   Conditional Heteroskedasticity ⇒ In conditional residuals are systematically related to the independent variable

   Heteroskedasticity can be detected by scatter plot and BP Chi square Test



3) **Outliers ⇒ Outliers have no specific definition for having 50 degree Celsius in switzerland is an outlier for that country . But same Temperature is not an outlier in Africa. But In regression outliers Can change the prediction output. In case of large number of outliers in Dataset we prepare two Models one with outliers and one without Outliers and choose the one model which suits better.**

Question 2 By now you have seen multiple model evaluation metrics used for regression models, such as r-squared, adjusted r-squared, RMSE, the residual plot etc.

In this question, you are required to explain at least three regression model evaluation metrics in your own words.

1. For the final model that you have built, explain each evaluation metric with its intuition (i.e. what and how it measures) and relate the intuition to its mathematical formula. You may use figures or examples to explain if needed. Limit your answer to 1000 words for this part.
2. Compare the advantages and disadvantages of any three evaluation metrics. If you do not think there's any advantage or disadvantage of a certain metric, mention that. Limit your answer to 1000 words for this part.

Answer 2. 1) RSS(Residual sum Square) ⇒ It is the sum of the squares of the difference between the predicted value and actual value . Lesser the value of RSS better is the model. Because lesser the value of RSS lesser is the error between predicted and actual value and better is the model

RSS == (y1_actual -y1_pred)^2 +(y2_actual - y2_pred)^2+........n

For every model

2) R_squared ⇒ RSS is a absolute quantity it will be different for different units so for that we need a standardise which will give us the result through which we can decide which model is best. So for that we calculate TSS (Total sum of square)

Y_avg = (y1 +y2+y3+y4+y5+y6+y7……yn)/n

TSS= (y1 - y_avg)^2+…………..(yn-y_avg)^2

By this we can also infer that any model we built should be better than model

B0 = Y_avg

R_squared = 1 - RSS/TSS

R_squared = 0.878means that by our model we can able to explain about

87.8% variation of the data

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.878
Model:                            OLS   Adj. R-squared:                  0.873
Method:                 Least Squares   F-statistic:                     163.7
Date:                Sun, 19 Aug 2018   Prob (F-statistic):           1.11e-59
Time:                        16:25:28   Log-Likelihood:                 182.89
No. Observations:                 143   AIC:                            -351.8
Df Residuals:                     136   BIC:                            -331.0
Df Model:                           6
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           -0.0740      0.011     -6.574      0.000      -0.096      -0.052
enginesize       0.9861      0.042     23.741      0.000       0.904       1.068
bmw              0.2098      0.030      7.004      0.000       0.151       0.269
porsche          0.2178      0.042      5.239      0.000       0.136       0.300
volvo            0.1073      0.029      3.675      0.000       0.050       0.165
rotor            0.2274      0.036      6.333      0.000       0.156       0.298
fivecylinder     0.1401      0.024      5.745      0.000       0.092       0.188
==============================================================================
Omnibus:                        9.030   Durbin-Watson:                   2.147
Prob(Omnibus):                  0.011   Jarque-Bera (JB):                9.636
Skew:                           0.478   Prob(JB):                      0.00808
Kurtosis:                       3.839   Cond. No.                         8.73
==============================================================================
```

3) VIF(Variance Inflation Factor) ⇒ It calculates how well you can predict particular variable using all other variables. If two variables are same you can

	Var	Vif
0	enginesize	5.53
1	stroke	4.44
5	volvo	1.48
4	porsche	1.41
6	minusTwo	1.38
7	dohcv	1.25
10	twelvecylinder	1.24
2	bmw	1.14
9	fivecylinder	1.12
3	peugeot	1.08
8	rotor	1.06

39]:

	Var	Vif
0	enginesize	1.37
1	bmw	1.10
5	fivecylinder	1.10
2	porsche	1.09
3	volvo	1.08
4	rotor	1.00

predict one variable from the other so that you can delete one of the variable. Such a situation is called multicollinearity situation

Example ⇒ x1, x2,x3,x4..........xn

VIF = 1/1-R_squared

Answer 2.2 Advantages of Multicollinearity

1) It help reduce the number of variables in model
2) It gives the relation between two independent variables
3) By using this property you can predict one independent variable from other
4) It reduces the complexity of model

5) It helps to reduce the difference between R_squared and adjusted R_squared

Disadvantages of Multicollinearity

1) Multicollinearity reduces the precision of the estimate coefficients, which weakens the statistical power of your regression model.
2) The coefficient estimates can swing wildly based on which other independent variables are in the model. Cofficents are very sensitive to small change in the model as shown below in the figure

```
]:  const             -0.050776
    enginelocation     0.120385
    enginesize         1.052466
    stroke            -0.069856
    bmw                0.196690
    peugeot            0.043323
    porsche            0.147472
    volvo              0.147670
    minusTwo          -0.146717
    dohcv              0.026399
    rotor              0.240688
    fivecylinder       0.141346
    twelvecylinder    -0.212092
    dtype: float64
```

```
]:  1  print(ln_1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.896
Model:                            OLS   Adj. R-squared:                  0.886
Method:                 Least Squares   F-statistic:                     93.27
Date:                Sun, 19 Aug 2018   Prob (F-statistic):           9.39e-58
Time:                        14:58:44   Log-Likelihood:                 194.03
No. Observations:                 143   AIC:                            -362.1
Df Residuals:                     130   BIC:                            -323.5
Df Model:                          12
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const             -0.0508      0.022     -2.262      0.025      -0.095      -0.006
enginelocation     0.1204      0.093      1.295      0.198      -0.064       0.304
enginesize         1.0525      0.049     21.430      0.000       0.955       1.150
stroke            -0.0699      0.041     -1.685      0.094      -0.152       0.012
bmw                0.1967      0.029      6.711      0.000       0.139       0.255
peugeot            0.0433      0.024      1.783      0.077      -0.005       0.091
porsche            0.1475      0.066      2.230      0.027       0.017       0.278
volvo              0.1477      0.034      4.347      0.000       0.080       0.215
minusTwo          -0.1467      0.057     -2.592      0.011      -0.259      -0.035
dohcv              0.0264      0.093      0.284      0.777      -0.157       0.210
rotor              0.2407      0.034      7.022      0.000       0.173       0.309
fivecylinder       0.1413      0.024      6.013      0.000       0.095       0.188
twelvecylinder    -0.2121      0.079     -2.693      0.008      -0.368      -0.056
```

```
: const           -0.083661
  enginesize       1.036992
  bmw              0.198608
  porsche          0.204122
  volvo            0.149915
  minusTwo        -0.147332
  rotor            0.234826
  fivecylinder     0.132557
  twelvecylinder  -0.186686
  dtype: float64
```

```
: 1 print(ls_1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.889
Model:                            OLS   Adj. R-squared:                  0.882
Method:                 Least Squares   F-statistic:                     134.1
Date:                Sun, 19 Aug 2018   Prob (F-statistic):           4.29e-60
Time:                        15:37:06   Log-Likelihood:                 189.39
No. Observations:                 143   AIC:                            -360.8
Df Residuals:                     134   BIC:                            -334.1
Df Model:                           8
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const             -0.0837      0.012     -7.227      0.000      -0.107      -0.061
enginesize         1.0370      0.045     22.888      0.000       0.947       1.127
bmw                0.1986      0.029      6.800      0.000       0.141       0.256
porsche            0.2041      0.040      5.050      0.000       0.124       0.284
volvo              0.1499      0.034      4.386      0.000       0.082       0.218
minusTwo          -0.1473      0.058     -2.559      0.012      -0.261      -0.033
rotor              0.2348      0.035      6.766      0.000       0.166       0.303
fivecylinder       0.1326      0.024      5.598      0.000       0.086       0.179
twelvecylinder    -0.1867      0.076     -2.468      0.015      -0.336      -0.037
==============================================================================
Omnibus:                        8.505   Durbin-Watson:                   2.159
Prob(Omnibus):                  0.014   Jarque-Bera (JB):                9.929
Skew:                           0.405   Prob(JB):                      0.00698
Kurtosis:                       4.006   Cond. No.                         15.0
==============================================================================
```

Advantages of TSS(Total Sum of Square)

1) TSS= (y1 - y_avg)^2+..............(yn-y_avg)^2 when such a linear model where there is no independent variable then you built a model where you use intercept y= intercept. So it can help you to built a model without independent variable.
2) It can act whether the given model is good or not. Any other model built with independent variable should be better than than model Y= intercept where intercept = (y1+y2+....yn)/n
3) It helps to calculate R_squared

Advantages of R_Squared

1) It helps to avoid overfitting of the model
2) It helps to explain the how good the model is
3) It helps to explain the variation in the data

4) It helps to calculate the vif which determines the multicollinearity