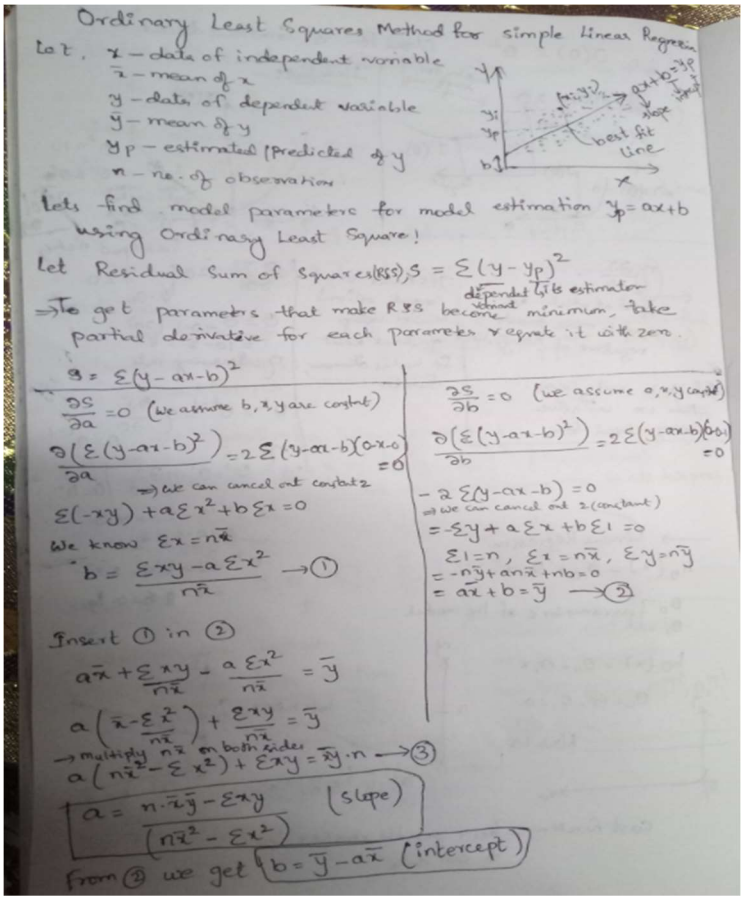
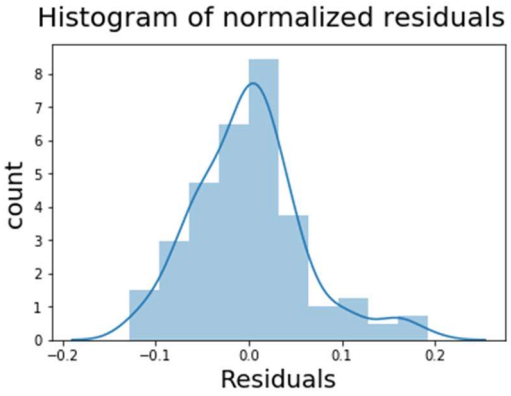
1. Explain the linear regression algorithm in detail.

    a. A linear regression model attempts to explain the linear relationship between a dependent variable and independent variables(1 independent variable incase of Simple Linear Regression and more than 1 independent variables incase of Multiple Linear Regression.)
    b. Linear Regression algorithm for a given training data set, it will try to fit a best straight line in case of simple linear regression or a hyperplane for multiple regression using a metric called Residual Sum of Squares (RSS). The residuals are the difference between actual value and predicted values. Cost function is to minimize RSS using Ordinary Least Square
    c. We create a model using linear regression algorithm with different independent variables each time until we get a model which has significant variables which can explain target variable or dependent variable.
    d. yp=ax+b where b is the intercept and x is the independent variable and yp is the predicted value for the dependent variable. This is the best fit line we got for our data.
    e. To understand how we got a and b parameters that make RSS become minimum we take partial derivative for each parameter and equate it with zero
    f. Please check the image for obtaining parameters using Ordinary Least Square method. Its nice to compute the values.



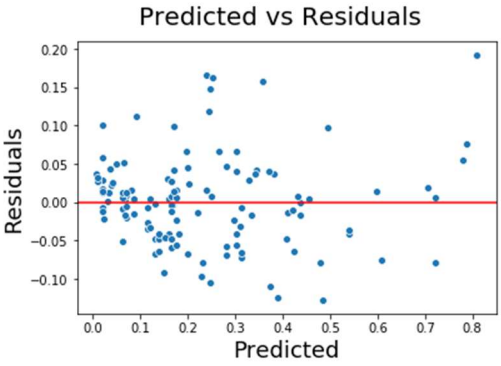
2. What are the assumptions of linear regression regarding residuals?
    a. **Residuals are normally distributed**, this is important to make inferences on the model that we have built, we need a notion of distribution of residuals. The residual/error terms generally follow a normal distribution with mean equal to zero. We can use histogram or q-q plot to plot the residuals.

       For our car-predictions, the following histogram is generated which shows residuals are normalized.



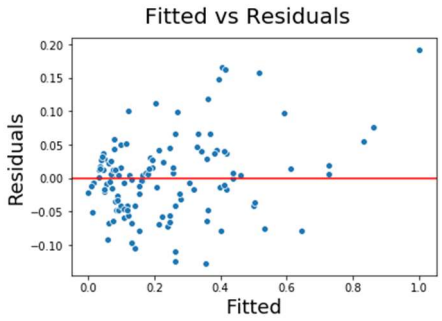
    b. **Residuals are independent of each other**, we should not see any visible patterns with the predicted and residuals.

       For our car-predictions, the following graph is generated which shows residuals are independent.

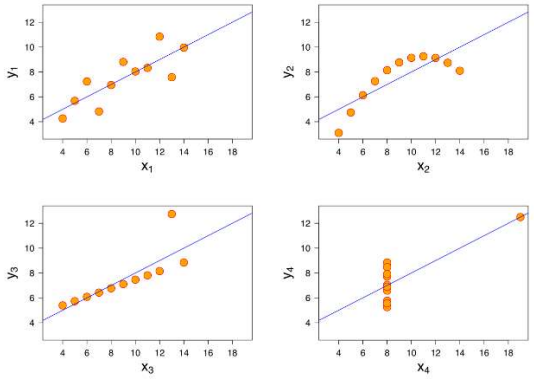


    c. **Residuals have constant variance (Homoscedasticity)** , the variance should not increase or decrease as error values changes. We will plot residuals vs actual y values.

       For our car-predictions, the following graph is generated which shows most of the residuals are having constant variance.



3. What is the coefficient of correlation and the coefficient of determination?
    a. Coefficient of correlation is "R" value measures the strength and direction of linear relationship between two variables. The value of *r* is such that $-1 \leq r \leq +1$. The + and – signs are used for positive linear correlations and negative linear correlations, respectively.
    b. Coefficient of determination is "r^2 " gives the proportion of the variance(fluctuation) of 1 variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in making predicitions from a certain model. It is the ratio of explained variation to total variation. The coefficient of determination is such that $0 < r^2 < 1$, and denotes the strength of the linear association between x and y.
4. Explain the Anscombe's quartet in detail.
    a. 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(ex: SUM, AVERAGE, STDDEV, MEAN, VARIANCE, CORRELATION COEFFICIENT,..).
    b. But things change completely when they are graphed. Each graph tells a different story irrespective of their similar summary statistics. Below are the graphs for the 4 data sets.



        i. Dataset I ,appear to have clean and well-fitting linear models.
        ii. Dataset II ,is not distributed normally.
        iii. In Dataset III, the distribution is linear, but the calculated regression is thrown off by an outlier.
        iv. Dataset IV ,shows that one outlier is enough to produce a high correlation coefficient.
    c. 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.

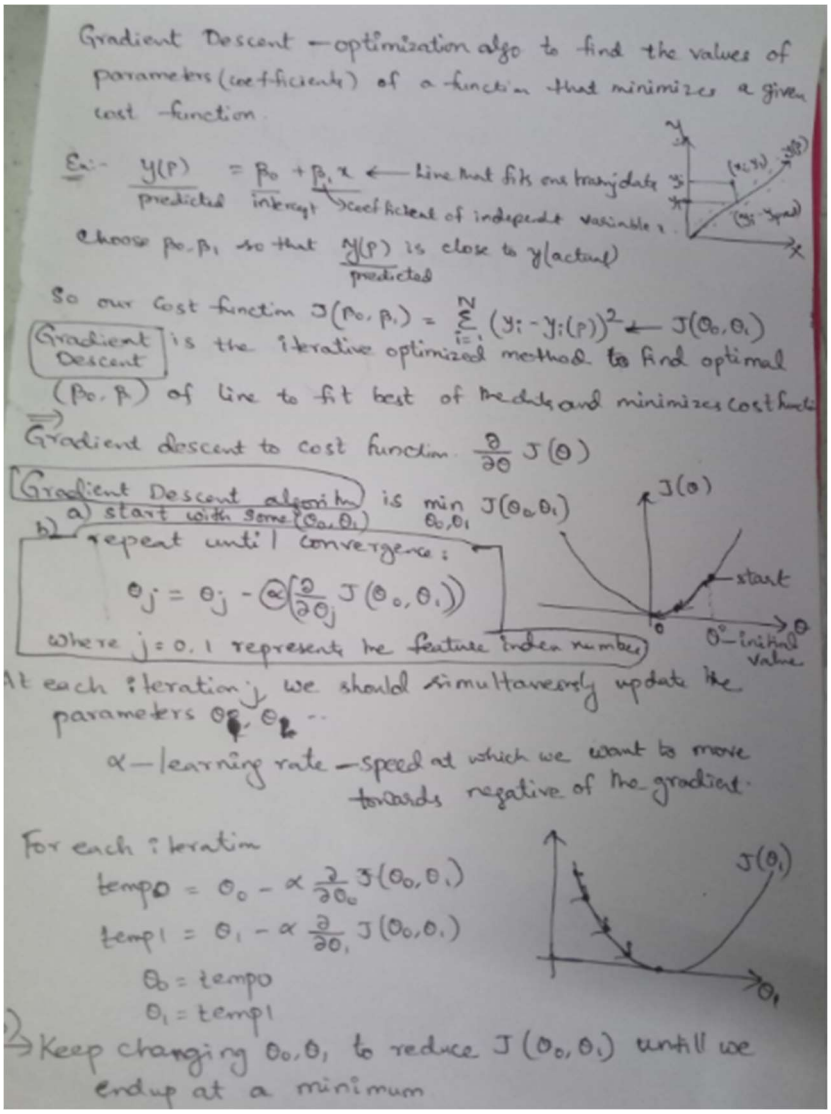
5. What is Pearson's R?
    a. Pearson's correlation coefficient is also referred as Pearson's R. It is a measure of strength and direction of linear relationship between two variables. The value of *r* is such that $-1 \leq r \leq +1$. The + and – signs are used for positive linear correlations and negative linear correlations, respectively.

6. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
    a. Feature scaling or scaling is essential step of data processing which is applied to features to normalize when the range of raw data varies widely.
        i. i.e, Scaling is important when we have features which vary in magnitudes, units and range and bring them to same level of magnitudes
        ii. Scaling improve or speedup gradient descent algorithm.
        iii. Scaling does not affect other parameters such as p-value and R-squared, it affects only the coefficients.
    b. Standardized scaling brings all the data into standard normal distribution with mean 0 and standard deviation 1. Whereas Normalized or MinMax scaling, brings all the data in range of 0-1. The formulae used as given below:
        i. Standaridization : X = (x-mean(x))/stddev(X)
        ii. Normalization (MinMax Scaling): X = (x-min(x))/(max(x)-min(x))

7. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
    a. VIF, calculates how well one independent variable is explained by all other independent variables combined. VIF(i) = 1/(1-R-squared(i)).
    b. So when R-squared(i) becomes 1 then VIF(i) becomes infinity. R-squared(i) will be 1 when 100% variance of one variable is explained by another i.e, when two variables are exactly same. So we should look at variables with exact collinearity using correlation matrix or heatmap.

8. What is the Gauss-Markov theorem?
    a. Gauss-Markov theorem states that ordinary least squares (OLS) estimator of the coefficients of a linear regression model is the best linear unbiased estimator (BLUE) if error terms satisfy below conditions.
        i. Expected value of error term is zero for all observations
        ii. Variance of the error term is constant
        iii. Error term is independently distributed and not correlated.

9. Explain the gradient descent algorithm in detail.
    a. Gradient descent mainly we start with some theta0, theta1 for simple linear regression with intercept(theta0) and one coefficient for independent variable(theta1)
    b. Keep changing thet0 and theta1 to reduce cost function J(theta0,theta1) until we end up at a minimum



10. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
    a. The Q-Q plot is a graphical tool to assess if two data sets come from populations with a common distribution.Q-Q plot of the quantiles of X versus the quantiles/ppf of a distribution
    b. Importance of Q-Q plot in linear regression we can plot residuals to check if they are normally distributed instead of using histogram alone to check.
    c. Compare X against dist. The default is scipy.stats.distributions.norm (a standard normal).
        i. import statsmodels.api as sm
        ii. res = lm5.resid #after training the model(lm5) we get residuals and plot q-q plot
        iii. fig = sm.qqplot(res, stats.norm, fit=True, line='45') # we can change stats.norm to stats.t or any distribution
        iv. plt.show()
    d. A perfectly normal distribution would exactly follow a line with slope = 1 and intercept = 0.

For our car-predictions, the following graph is generated which shows that residuals are normally distributed.

e.