


National University of Computer and Emerging Sciences, Lahore Campus

	Course Name:	Data Science	Course Code:	CS481
	Program:	BS (Computer Science)	Semester:	Spring 2017
	Duration:	180 Minutes	Total Marks:	50
	Paper Date:	18-05-2017	Weight	50 %
	Section:	ALL	Page(s):	12
	Exam Type:	Final		

Student : Name: _____ Roll No. _____ Section: _____

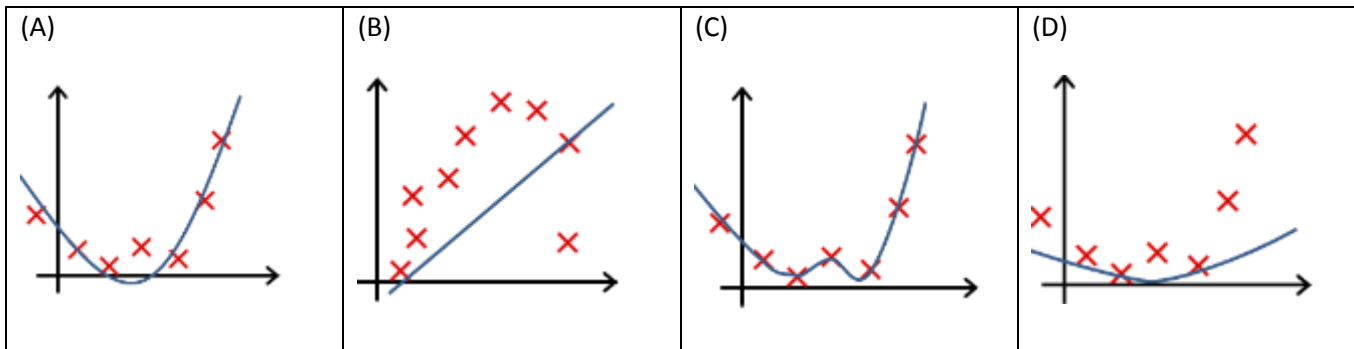
Instruction/Notes: Attempt the examination on the question paper and write concise answers. You can use extra sheet for rough work. Do not attach extra sheets used for rough with the question paper. Don't fill the table titled Questions/Marks.

Question	Objective	1	2	3	4	5	Total
Marks	/ 9	/ 14	/ 7	/ 7	/6	/7	/ 50

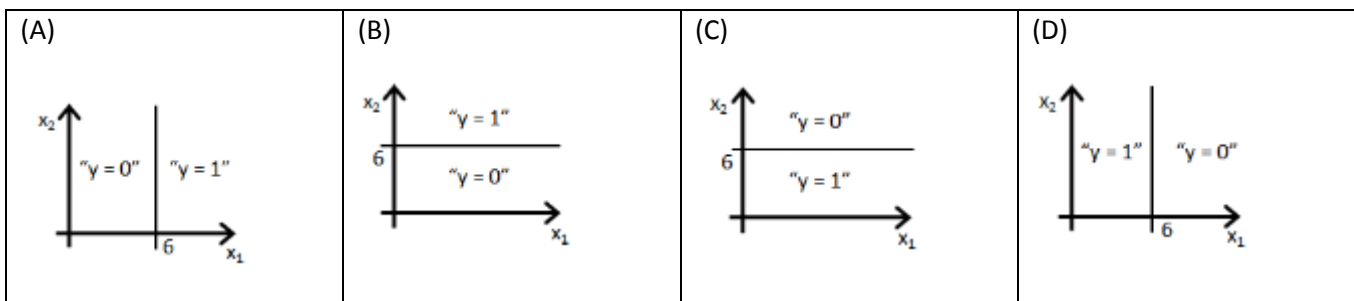
Section 1 (Objective part) [points 9]

Clearly circle the correct options.

Q1. In which of the following figure do you think the hypothesis is over-fitting the training set?



Q2. Suppose you train a logistic classifier $h_{\theta}(\mathbf{x}) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$. Suppose $\theta_0 = -6$, $\theta_1 = 1$, $\theta_2 = 0$. Which of following figures represents the decision boundary found by your classifier?



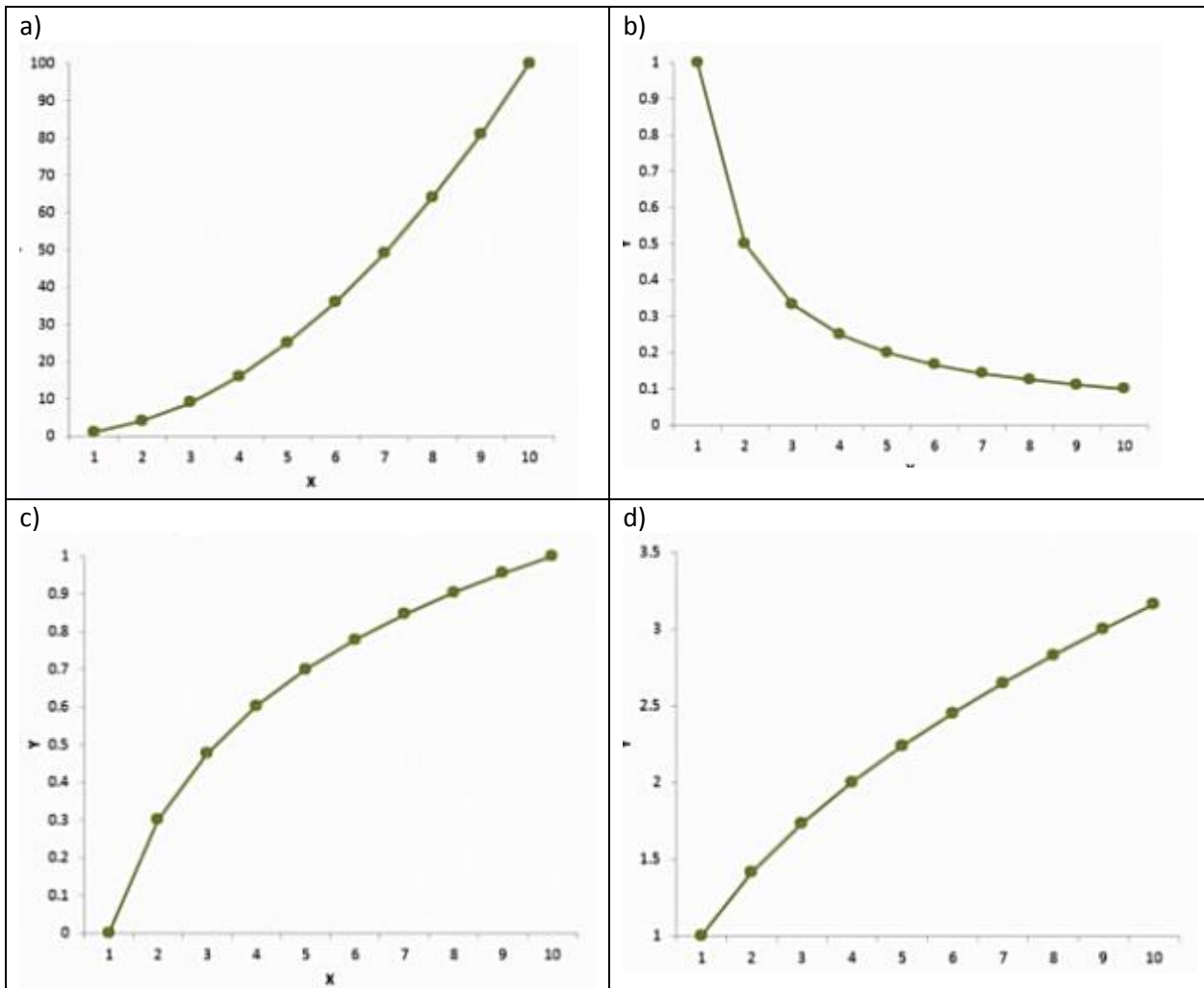
Q3. [True or False] Dimensionality reduction algorithms are one of the possible ways to reduce the computation time required to train a model.

- a) True b) False

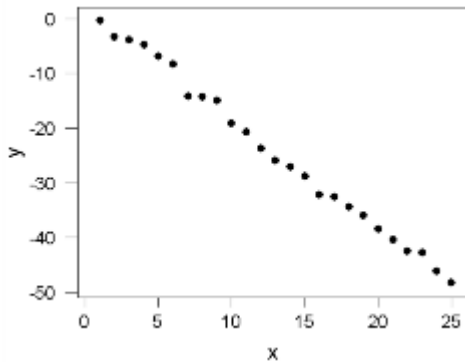
Q4. The most popularly used dimensionality reduction algorithm is Principal Component Analysis (PCA). Which of the following is/are true about PCA?

- a) PCA is an unsupervised method
- b) It searches for the directions that data have the largest variance
- c) Maximum number of principal components > number of features
- d) All principal components are not orthogonal to each other

Q5. Which one of the given graphs below represents Square Root Function? Explain your selection with the reason.



Q6. Shown below is a scatterplot of Y versus X.



Which choice is most likely to be the approximate value of R^2 ?

- A) -99.5%
- B) 2.0%
- C) 50.0%
- D) 99.5%

Q7. In the context of regression analysis, which of the following statements are true?

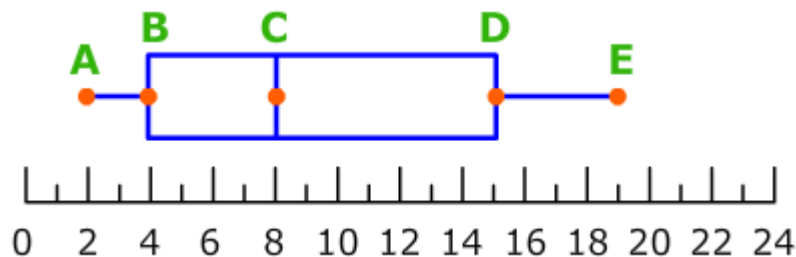
- I. When the sum of the residuals is greater than zero, the data set is nonlinear.
- II. A random pattern of residuals supports a linear model.
- III. A random pattern of residuals supports a non-linear model.

(A) I only (B) II only (C) III only (D) I and II (E) I and III

Q8. Imagine, you are solving a classification problems with highly imbalanced class. The majority class is observed 99% of times in the training data. Your model has 99% accuracy after taking the predictions on test data. Which of the following is true in such a case?

- A) Accuracy metric is not a good idea for imbalanced class problems.
- B) Accuracy metric is a good idea for imbalanced class problems.
- C) Precision and recall metrics are good for imbalanced class problems.
- D) Precision and recall metrics aren't good for imbalanced class problems.

Q9. What does the point labeled C represent on the box plot?



- a) Mean
- b) Median
- c) Mode
- d) Range

Section 2 (Subjective part) (marks 41)

Q1. [14 Marks] Short Questions:

A) [3 marks] Suppose a company X wants to outsource its 3 projects and for each project, there are 4 potential subcontractors, each having different time and cost proposal. Using Random Hypercube Design, generate a design matrix with 8 Design points.

B) [2 marks] You are a reviewer for the International conference on Learning Algorithms, and you read papers with the following experimental setups. Would you accept or reject each paper? Provide a one sentence justification. (This conference has short reviews.)

i). **accept/reject** “My algorithm is better than yours. Look at the training error rates!”

ii). **accept/reject** “My algorithm is better than yours. Look at the test error rates. Suppose we have Choosing λ based on the test data.”

C) [2 marks] Logistic regression is named after the log-odds of success (the logit of the probability) defined as

$$\ln \left(\frac{\mathbb{P}[Y = 1 \mid X = x]}{\mathbb{P}[Y = 0 \mid X = x]} \right)$$

Show that log-odds of success is a linear function of x .

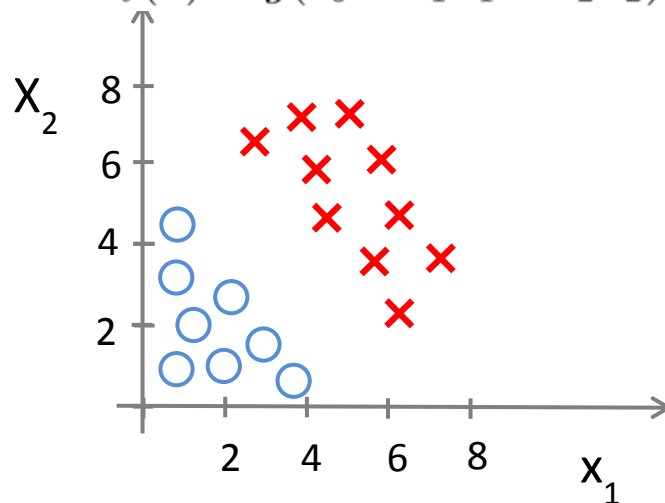
- D) [3 points]** Suppose you have implemented regularized linear regression to predict housing prices. However, when you test your hypothesis in a new set of houses, you find that it makes unacceptably large errors (high bias or high variance) in its prediction. You can try some of the options given in the first column in order to fix the problem. Mark 'y' in the second or third column for all 6 options.

If you try	Fixes high bias	Fixes High variance
Get more training examples		
Try smaller sets of features		
Try decreasing λ		
Try increasing λ		
Try getting additional features		
Try adding polynomial features		

E) [2 points] Logistic Regression – Decision Boundary:

We consider the following model of logistic regression for binary classification with a sigmoid function

Model:
$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1^2 + \theta_2 x_2)$$



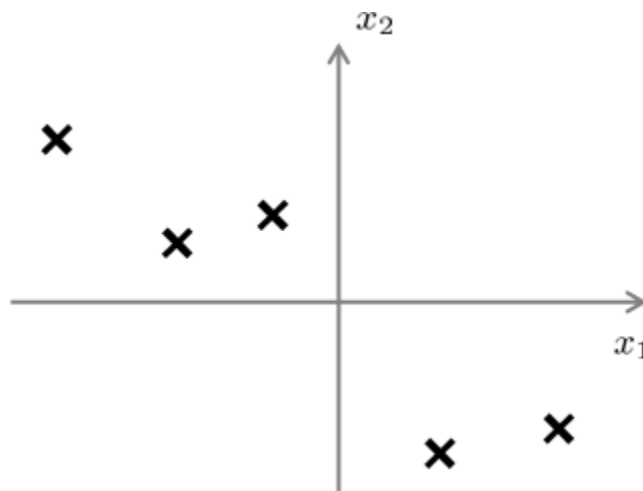
Suppose the trained parameter values are $\theta_0 = -8$, $\theta_1 = 2$ AND $\theta_2 = 2$

Draw the decision boundary and Show your working.

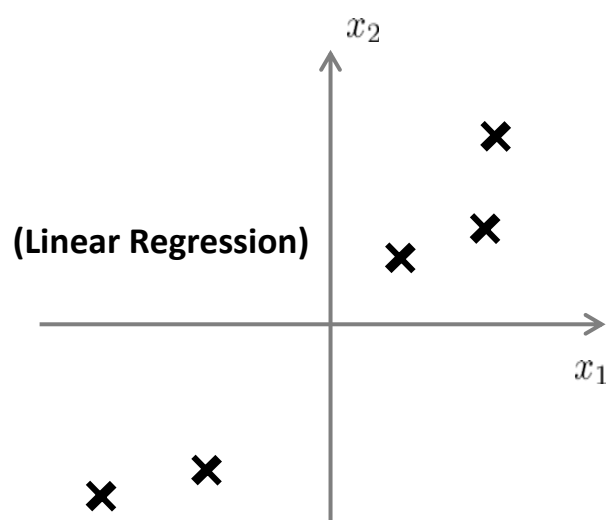
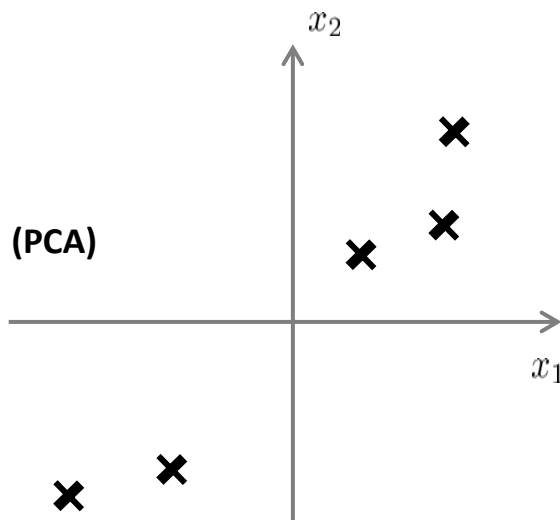
- F) [2 points] What is the **overfitting problem** and what can be the possible cause for this problem? Write down all possible options for addressing the overfitting problem.

Q2. [2 + 2 + 3 points]: Using Principal Component Analysis (PCA):

- A) Considering the dataset given below in the figure, find all principal components. You don't need any calculation, just draw and show in which direction of the vectors?



- B) Differentiate between PCA and Linear Regression. Explain your answer using Figures.



- c) **PCA**: Find the Eigenvalues for the matrix as given below and compute the first principal component.

$$\begin{bmatrix} 5 & 8 & 16 \\ 4 & 1 & 8 \\ -4 & -4 & -11 \end{bmatrix}$$

Q3: [7 points] Exploratory Data Analysis: Data Transformation

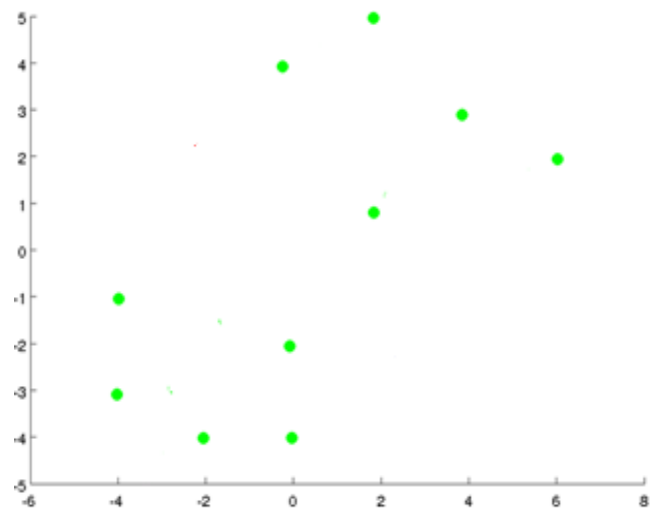
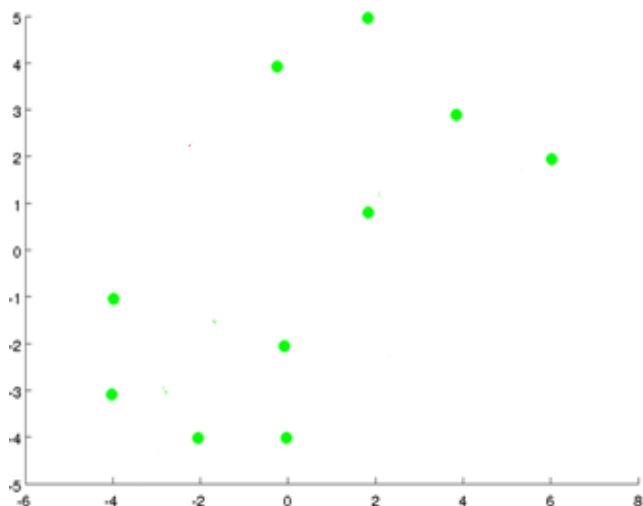
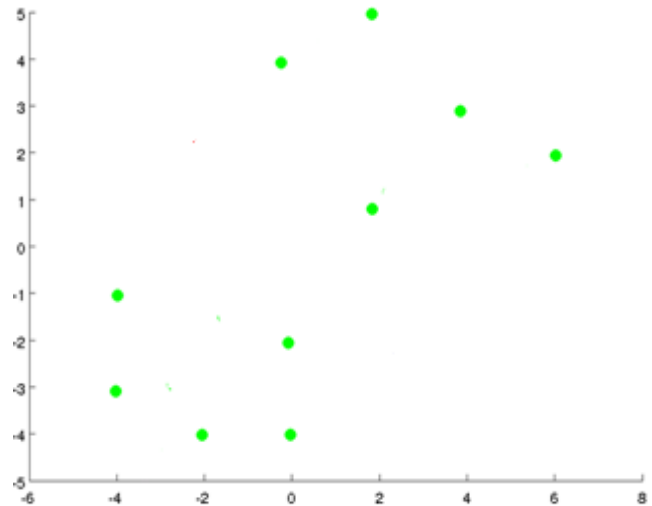
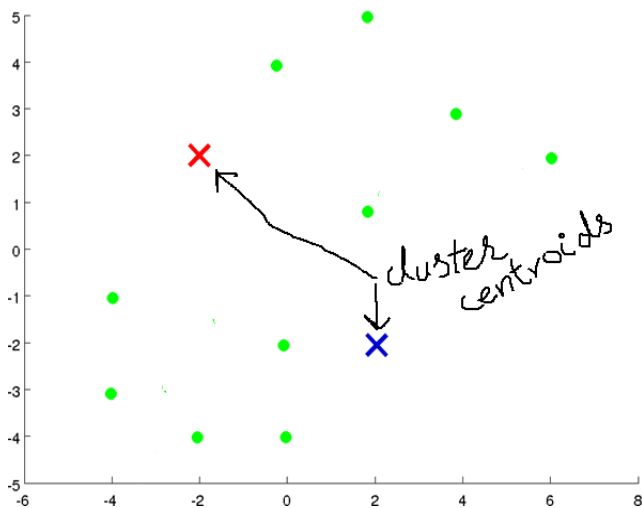
Suppose the relationship between X (independent variable) and Y (dependent variable) is represented by a power function $Y = 3X^2$. The data for X is given in the table below. For the given data, we cannot best fit a linear Line.

- A) Transform the data in such a way that we can fit a Linear Line using Linear Regression. Plot the transformed data on a graph and fit a linear line.
- B) Find out the approximated intercept term and slope of the line.

Observation# 1	1	2	3	4	5	6	7
X	1	2	2.5	3	3.5	4	5

Q4. [6 points]: Clustering Algorithm (K-means)

Given the training set ($m = 10$) in the table 1 (which are represented by circles in the Figure, group the data into 2 cohesive clusters ($K=2$). **Write values for $c^{(i)}$ and μ_k in each iteration of the K-means algorithm** (where $c^{(i)}$ = index of cluster (1,2,...K) to which example $X^{(i)}$ is currently assigned, and μ_k is cluster centroid k). Stop when converge or after 3 steps whichever comes first. (Note: distance calculation should be done on the question paper).



<

Solution:

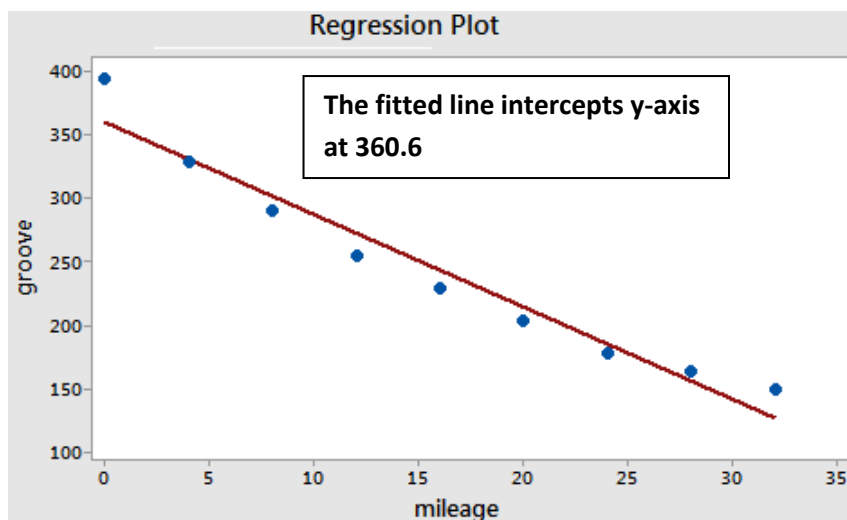
Q5. [7 points] A scientific laboratory conducted an experiment in order to answer the following research question: "Is tire tread wear linearly related to mileage?"

As a result of the experiment, the researchers obtained a data set as shown in Table below, containing the mileage (x, in 1000 miles) driven and the depth of the remaining groove (y, in mils).

mileage	groove
0	394.33
4	329.50
8	291.00
12	255.17
16	229.33
20	204.83
24	179.00
28	163.83
32	150.33

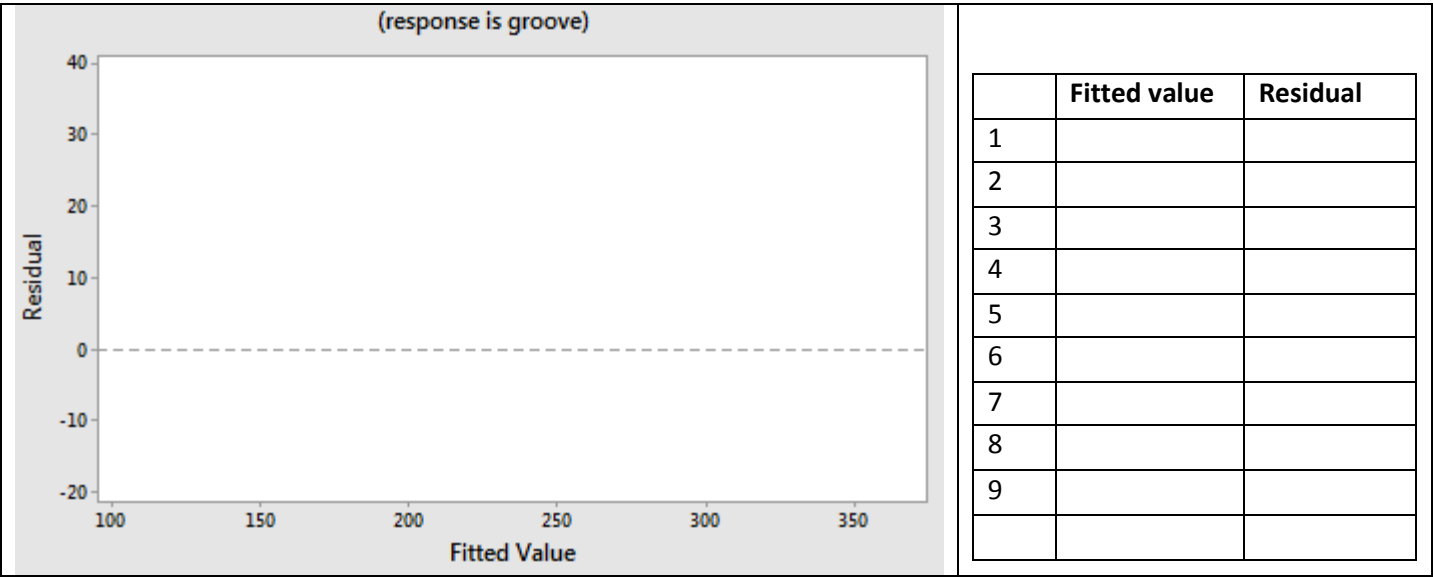


The fitted line plot of the resulting data suggests that there is a relationship between groove depth and mileage.



a) What will be the value of θ_0 and θ_1 for our simple linear hypothesis: $\text{groove} = \theta_0 + \theta_1 \text{mileage}$

b) Plot the residual of the simple linear regression model of the given data set against the fitted value (predicted values). Plot the residuals on the Figure below, and moreover, fill in the table with Fitted (Predicted) values and Residuals.



c) How does your Residual plot helps to determine if your regression function is linear or not-linear.