IEE 598: DATA SCIENCE FOR SYSTEM INFORMATICS: ASSIGNMENT 1

Hetul Varaiya
1211306106

Problem 1: Weighted Linear Regression

In the class, we have derived the Maximum Likelihood function for the following linear regression model

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I)$$

In this problems, we would like to extend this algorithm for the case of correlated noise. Suppose $\epsilon \sim N(0, \Sigma)$, where Σ is the covariance matrix of the noise, which is assumed to be given.

Hint: The pdf of the multivariate normal distribution of distribution with mean μ and covariance matrix Σ is given by $\det(2\pi\Sigma)^{-1/2}\exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$.

1. Please derive the likelihood function of the weighted linear regression.
2. Please derive the analytical solution.
3. Please analyze the time and space complexity of the analytical solution given X is an $n \times p$ matrix.

Solution on the next Page:

HETUL VARAIYA
121130606

Page No.
Date: / /

PROBLEM 1: WEIGHTED LINEAR REGRESSION:

As described in the class ~~MLE function for the following linear regression model~~ $y = X\beta + \varepsilon, \varepsilon \sim N(0, \sigma^2)$ has the pdf of the multivariate distribution,

$$(2\pi\Sigma)^{-1/2} e^{(-1/2(x-\mu)^T \Sigma^{-1}(x-\mu))}$$

1) Likelihood function of the weighted linear Regression

$$L = \prod_{i=1}^{N} f_i = \prod_{i=1}^{N} (2\pi\Sigma_i)^{-1/2} \exp\left\{-\left(\tfrac{1}{2\sigma}\right)(y_i - x_i\beta)^2\right\}$$

$$~~Let~~\ l = \log L = \sum_{i=1}^{N} \log((2\pi\Sigma)^{-1/2}) + \frac{1}{2}\frac{1}{2}\sum_{i=1}^{N} w_i(y_i - x_i\beta)^2 \quad w_i = \frac{1}{\Sigma_i}$$

$$\frac{dL}{d\beta} = 0 \Rightarrow \frac{dl}{d\beta} = 0 + \frac{2}{2}\sum_{i=1}^{N} w_i x_i(y_i - \beta x_i) = 0$$

$$\therefore \hat{\beta} = \frac{\sum_{i=1}^{N} w_i x_i y_i}{\sum_{i=1}^{N} w_i x_i^2}$$

$$\hat{\beta} = \left(\sum_{i=1}^{N} w_i x_i^2\right)^{-1}\left(\sum_{i=1}^{N} w_i x_i y_i\right) = \hat{\beta}_{WLS}$$

2. Derive the analytical solution.

$$\hat{\beta} = (X^T W X)^{-1}(X^T W y)$$

(3) $X^T = p \times n$

$X = n \times p$

$w : n \times n$

$y : n \times 1$

Date: / /

Time complexity:

$X^T w : O(pn^2)$

$X^T w X : O(p^2 n)$

$(X^T w X^T) = O(p^3)$

$X^T w : O(pn^2)$

$X^T w_y : O(pn)$

$(X^T W_x)^{-1} (X^T w_y) = O(p^2)$

$T = O(pn^2) + O(p^2(n)) + O(p^3) + O(pn^2) + O(pn) + O(p^2)$

Space Complexity:

$X^T : p_n$ $\quad X : n \times p$ $\quad w : n \times n$ $\quad y : n \times 1$

As the space required for w is $n \times n$
Assumption for this model $n > p$, we can say that, according to the notes from the lecture we can say that the space complexity of the algorithm will be $O(n^2)$

Problem 2: Prediction of the Housing price

In this problem, we will load the housing price data provided by 'houseprice.csv'. Here are a brief characteristics of the data set

Input variables

- crim: per capita crime rate by town
- zn: proportion of residential land zoned for lots over 25,000 sq.ft.
- indus: proportion of non-retail business acres per town
- chas: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- nox: nitric oxides concentration (parts per 10 million)
- rm: average number of rooms per dwelling
- age: proportion of owner-occupied units built prior to 1940
- dis: weighted distances to five Boston employment centres
- rad: index of accessibility to radial highways
- tax: full-value property-tax rate per $10,000
- ptratio: pupil-teacher ratio by town
- b: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
- lstat: % lower status of the population

Output variables

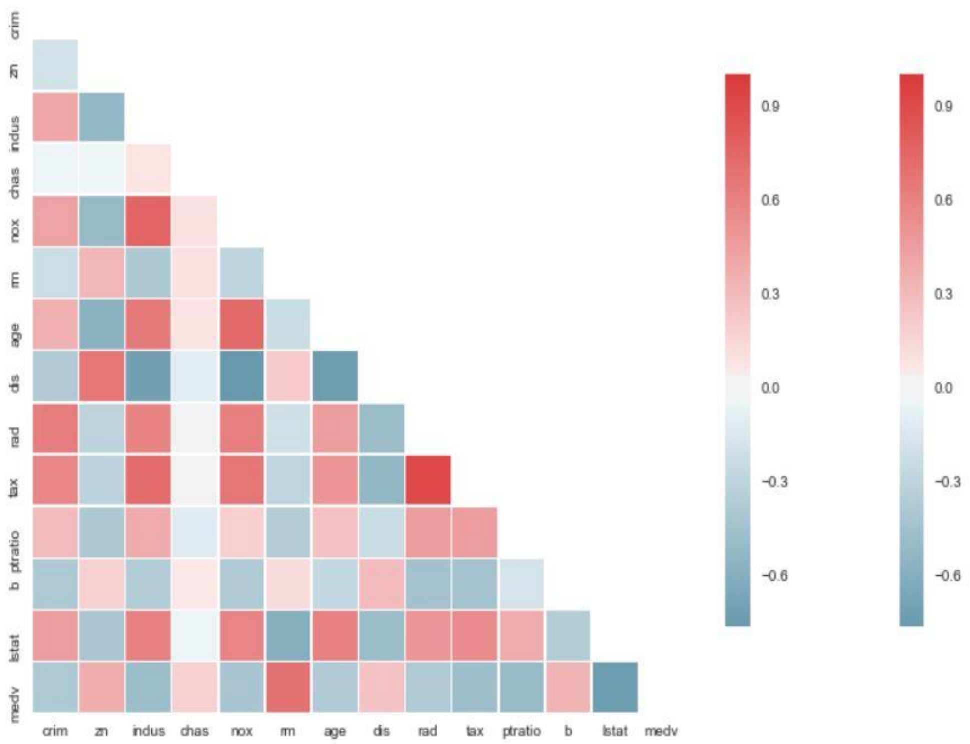
- medv: Median value of owner-occupied homes in $1000's, target variable

Please answer the following questions for this dataset.

1. Please plot the correlation between all the input variables and output variables pairs. Please identify the first three pairs with strongest correlation (either positive or negative).
2. Please conduct the simple linear regression of the response 'medv' with each of the 13 variables 'crim', 'zn', 'indus', 'chas', 'nox', 'rm', 'age', 'dis', 'rad', 'tax', 'ptratio', 'b', and 'lstat'. Please split the data into training and testing and evaluate the testing accuracy for each model (13 in total). Please generate the plot for each model and report the Residual Sum of Square for each model.
3. Please use all the other input variables for multiple linear regression to predict the response 'medv' with all the 13 input variables and evaluate the testing accuracy. Please compare the accuracy with the simple linear regression with 13 variables.

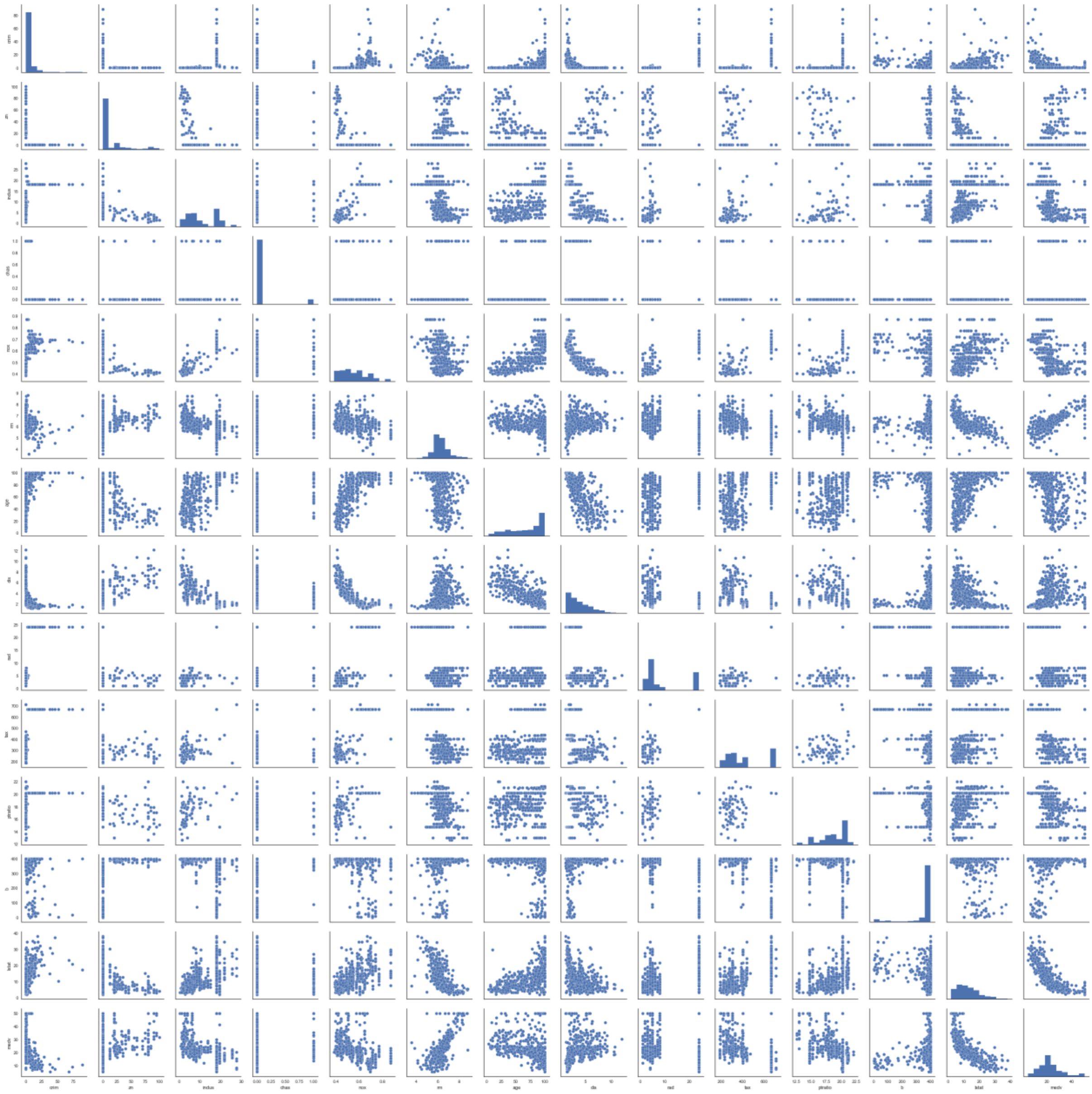
Important Conclusions from the question:

Part 1. Correlation Plot:



	crim	zn	indus	chas	nox	rm	age
crim	1.000000	-0.200469	0.406583	-0.055892	0.420972	-0.219247	0.352734
zn	-0.200469	1.000000	-0.533828	-0.042697	-0.516604	0.311991	-0.569537
indus	0.406583	-0.533828	1.000000	0.062938	0.763651	-0.391676	0.644779
chas	-0.055892	-0.042697	0.062938	1.000000	0.091203	0.091251	0.086518
nox	0.420972	-0.516604	0.763651	0.091203	1.000000	-0.302188	0.731470
rm	-0.219247	0.311991	-0.391676	0.091251	-0.302188	1.000000	-0.240265
age	0.352734	-0.569537	0.644779	0.086518	0.731470	-0.240265	1.000000
dis	-0.379670	0.664408	-0.708027	-0.099176	-0.769230	0.205246	-0.747881
rad	0.625505	-0.311948	0.595129	-0.007368	0.611441	-0.209847	0.456022
tax	0.582764	-0.314563	0.720760	-0.035587	0.668023	-0.292048	0.506456
ptratio	0.289946	-0.391679	0.383248	-0.121515	0.188933	-0.355501	0.261515
b	-0.385064	0.175520	-0.356977	0.048788	-0.380051	0.128069	-0.273534
lstat	0.455621	-0.412995	0.603800	-0.053929	0.590879	-0.613808	0.602339
medv	-0.388305	0.360445	-0.483725	0.175260	-0.427321	0.695360	-0.376955

	dis	rad	tax	ptratio	b	lstat	medv
crim	-0.379670	0.625505	0.582764	0.289946	-0.385064	0.455621	-0.388305
zn	0.664408	-0.311948	-0.314563	-0.391679	0.175520	-0.412995	0.360445
indus	-0.708027	0.595129	0.720760	0.383248	-0.356977	0.603800	-0.483725
chas	-0.099176	-0.007368	-0.035587	-0.121515	0.048788	-0.053929	0.175260
nox	-0.769230	0.611441	0.668023	0.188933	-0.380051	0.590879	-0.427321
rm	0.205246	-0.209847	-0.292048	-0.355501	0.128069	-0.613808	0.695360
age	-0.747881	0.456022	0.506456	0.261515	-0.273534	0.602339	-0.376955
dis	1.000000	-0.494588	-0.534432	-0.232471	0.291512	-0.496996	0.249929
rad	-0.494588	1.000000	0.910228	0.464741	-0.444413	0.488676	-0.381626
tax	-0.534432	0.910228	1.000000	0.460853	-0.441808	0.543993	-0.468536
ptratio	-0.232471	0.464741	0.460853	1.000000	-0.177383	0.374044	-0.507787
b	0.291512	-0.444413	-0.441808	-0.177383	1.000000	-0.366087	0.333461
lstat	-0.496996	0.488676	0.543993	0.374044	-0.366087	1.000000	-0.737663
medv	0.249929	-0.381626	-0.468536	-0.507787	0.333461	-0.737663	1.000000



From the correlation plot it can be concluded that the correlation between

1. Full-value property Tax(TAX) and the index of accessibility to the radial highways(RAD) is maximum(positively) with 91.0228% correlation.

2. Nitric Oxide Concentration(NOX) and proportion of non-retail business acres per town(INDUS) are positively correlated (76.36%).

3. Nitric Oxide Concentration(NOX) and weighted distances to five Boston employment centres(DIS) are negatively correlated (76.9)

Correlation between the output Variable and the Regressor RM has the highest positive correlation and the correlation between the output variable MEDV and the regressor LSTAT has the highest negative correlation.

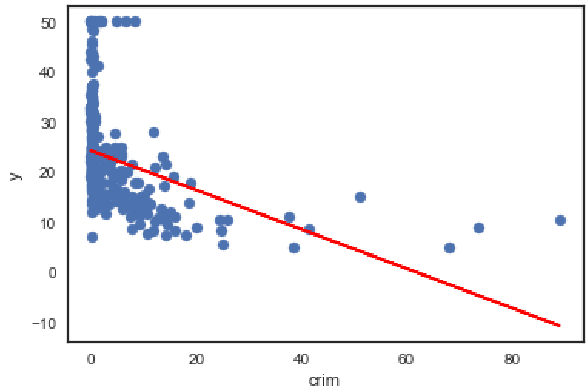
Part 2 and Part 3:

Shown below are the regression plots for the Simple Linear Regression between the Output Variable MEDV and the corresponding input variables.

crim

RSS is 10059.847880803769

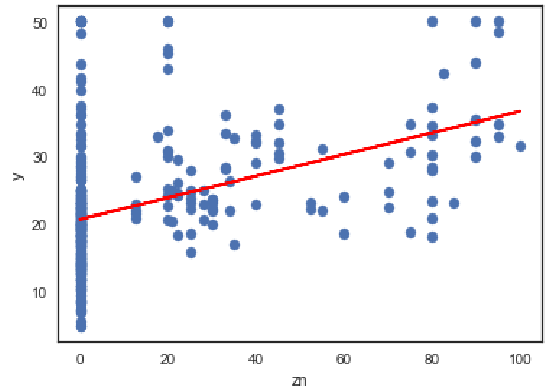
MSE is 66.18320974213006



zn

RSS is 11625.335502239657

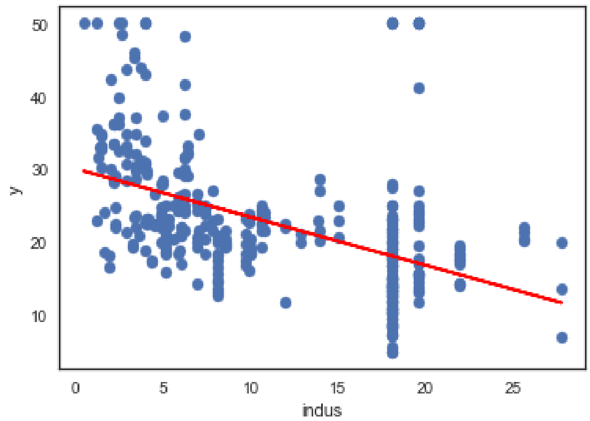
MSE is 76.48247040947143



indus

RSS is 9152.250060521186

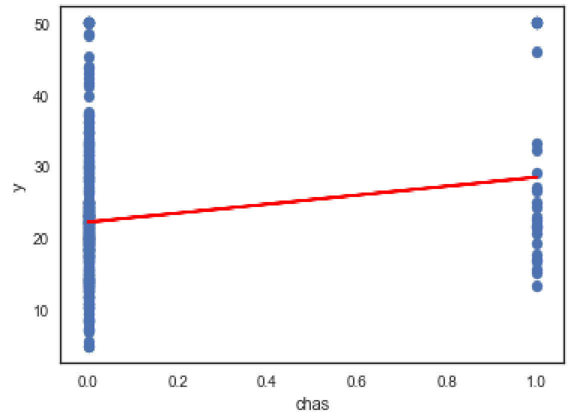
MSE is 60.21217145079728



chas

RSS is 11337.26915588018

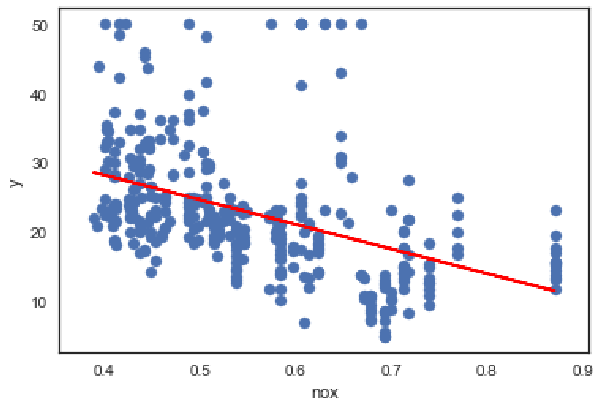
MSE is 74.58729707815908



nox

RSS is 10365.974176116722

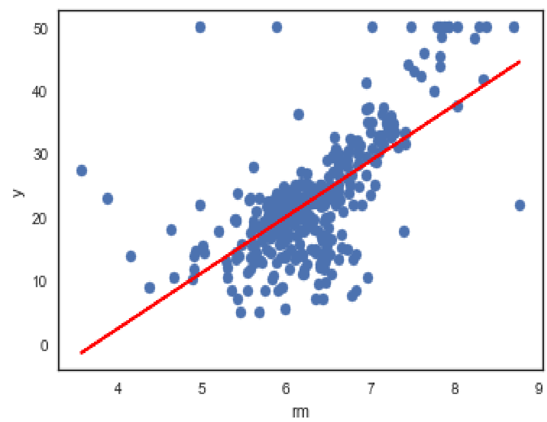
MSE is 68.1971985270837



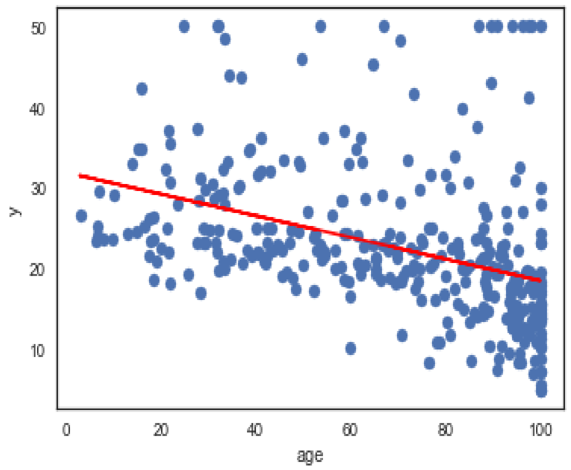
rm

RSS is 4962.945204383877

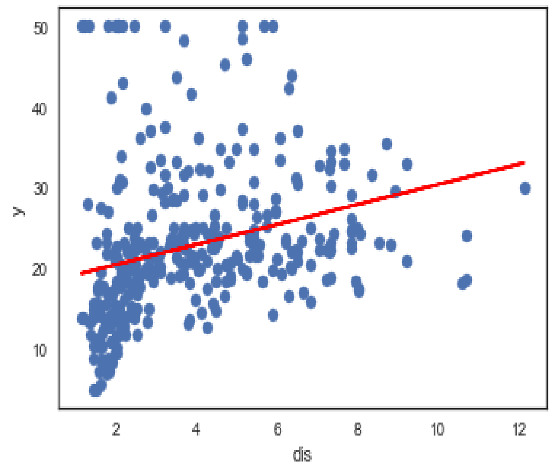
MSE is 32.650955291999196



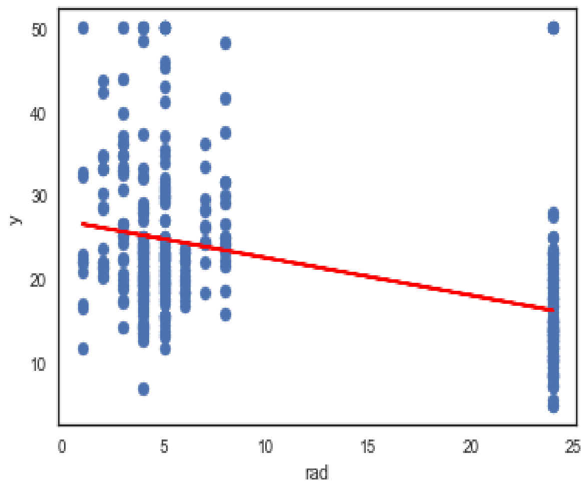
age
RSS is 10546.103863896924
MSE is 69.38226226247976



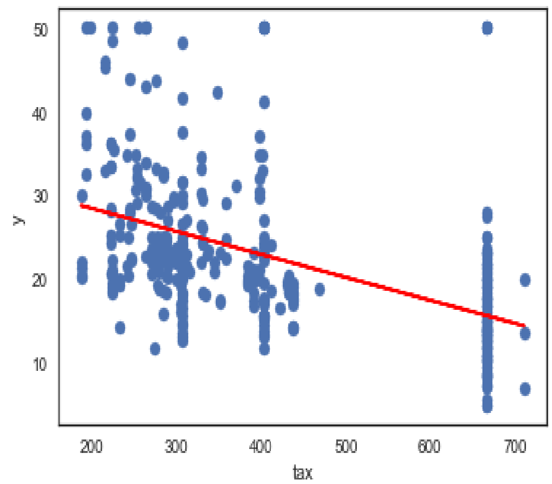
dis
RSS is 11545.721707816529
MSE is 75.95869544616137



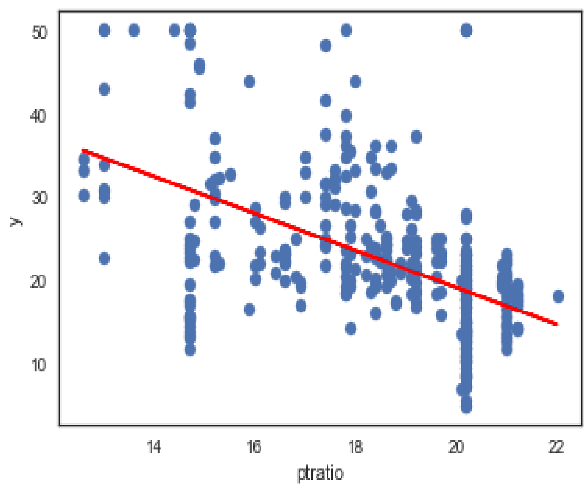
rad
RSS is 10966.868808746465
MSE is 72.15045268912148



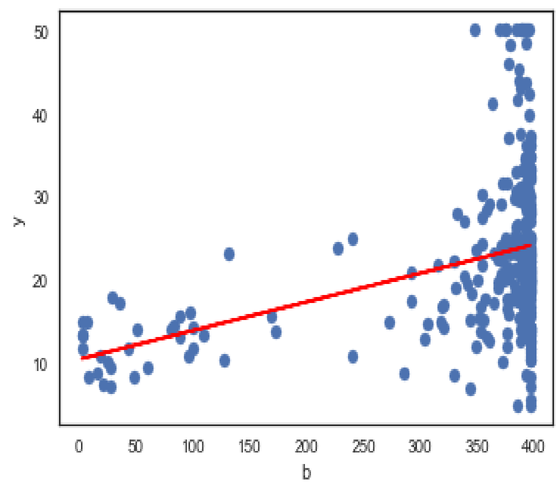
tax
RSS is 9971.22337606783
MSE is 65.60015378991993

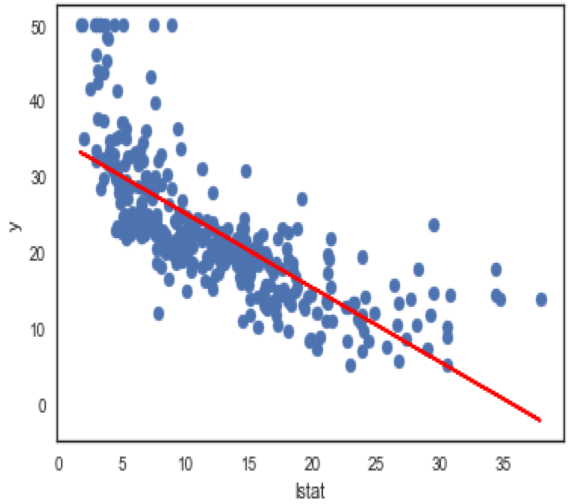


ptratio
RSS is 9166.516821882715
MSE is 60.3060317229126



b
RSS is 10799.728204022871
MSE is 71.05084344751889



lstat RSS is 5263.044242835275 MSE is 34.62529107128471 	#Multiple Linear Regression: The **RSS** for the Multiple Linear Regression is **3270.651523** The **MSE** for the Multiple Linear Regression is **21.517444**

Using the library sklearn.model_selection and importing the train_test_split I was able to separate the data into two different sets of training and testing data in a 70 – 30 way(70% training data and 30% testing data) for both the input and output variables.

After training the model and predicting the output variable the testing accuracy turned out to be the highest for the model between RM and MEDV with the model Mean Square Error to be the lowest. The testing accuracy for the model with multiple input variables using the Multiple Linear Regression turned out to be the least as the mean Square error is 21.52 which is less than the simple linear regression for all the models displayed separately.

Clearly it is seen that model with more number of parameters works better specifically in this case. This result cannot be generalized but we can say that for this model it turns out that it is better if we choose to include more number of variables to get a better testing accuracy.

Another important observation from the above findings is that the Residual Sum of Squares values(RSS) for the model is the least for LSTAT which means it has most of the points near to the regression fit giving the RSS value to be least among all the other input variables.

Problem 3: Polynomial Regression Model

Suppose the true function $f(x) = \sin(x)$, we sample 50 points from 0 to 10 as our training samples, the goal is to use the polynomial regression to fit this sin function.

Please follow the guide in HW1Prob3.ipynb to answer the questions. Please complete the code in the notebook in the following 4 parts.

1. Part 1: Visualize the data in the notebook
2. Part 2: generate the Design Matrix for Polynomial Regression
3. Part 3: Use training and testing split
4. Part 4: (Bonus and open question): for this question, since we are simulating the example and we do know the true function. Can you use simulation to compute the bias and variance for different order of polynomial function? If so, please use simulation to estimate the bias and variance for polynomial model from order 1 to 20 and comment on the result.

Please complete the code in the notebook and answer the following question: (You can also download the code as .py file and work outside the notebook)

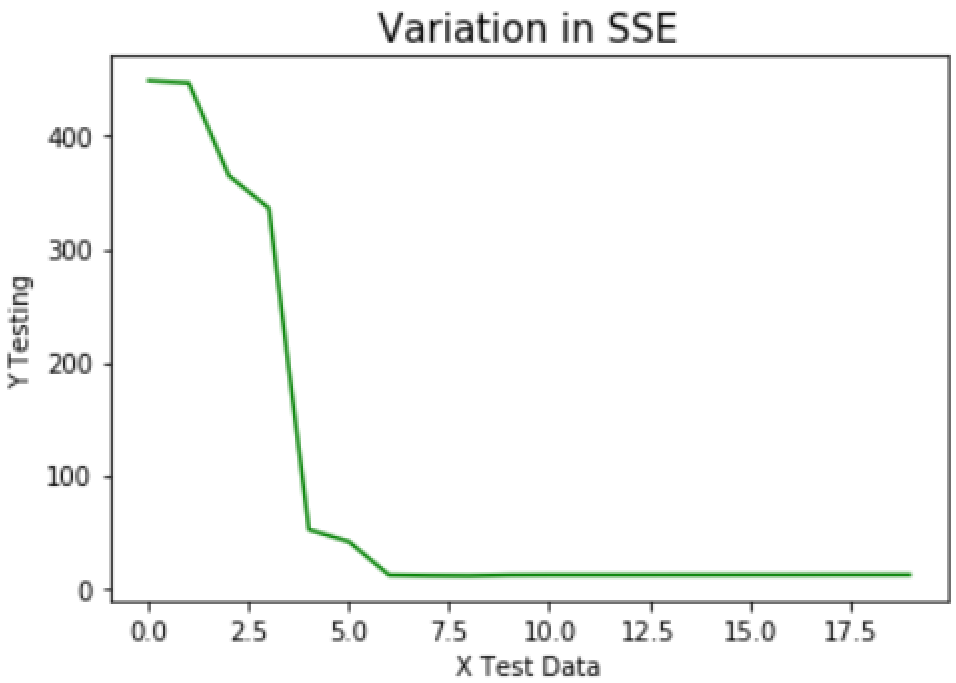
1. Which model you would like to use? (Polynomial order and coefficients) How and why do you choose this model?
2. Please provide a plot for the fitted models.

Part 1: Visualize the data

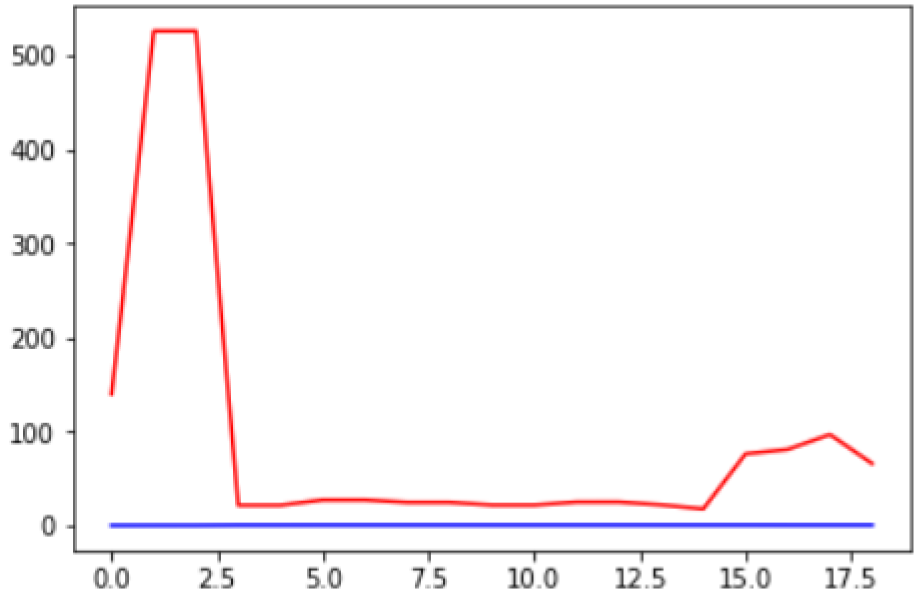
Part 2: design matrix polynomial regression

```
Array ([[ 1    1    1],
        [ 2    4    8.],
        [ 3    9    27.]])
```

Part 3:



Part 4:

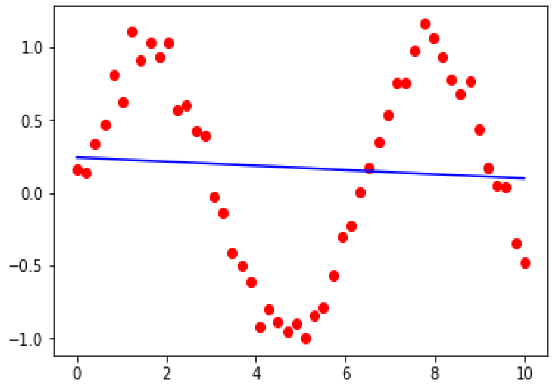


Part 5.

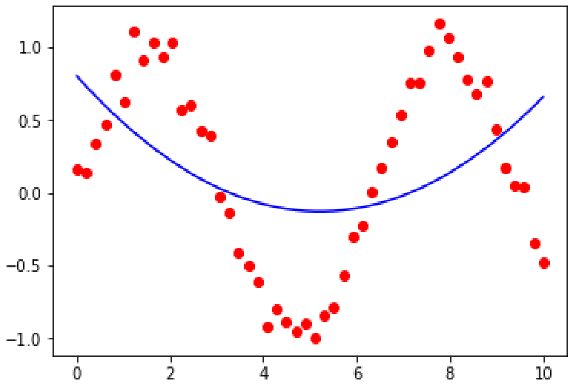
The preferable model that can be used to get the best model would be with the degree 8 because it has the least variance and bias shift according to the scatter points of the data points. The variance gradually increases as the number of regressors in the model increases.

P – Degree of the polynomials

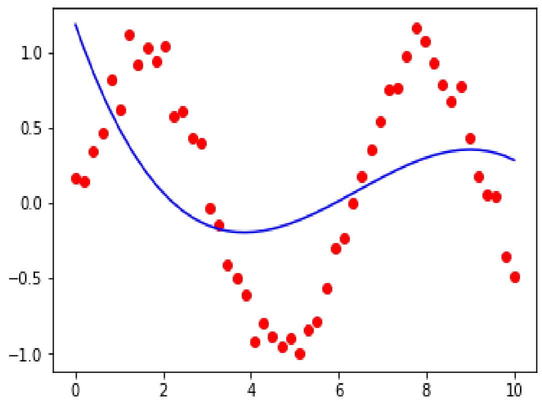
P = 1



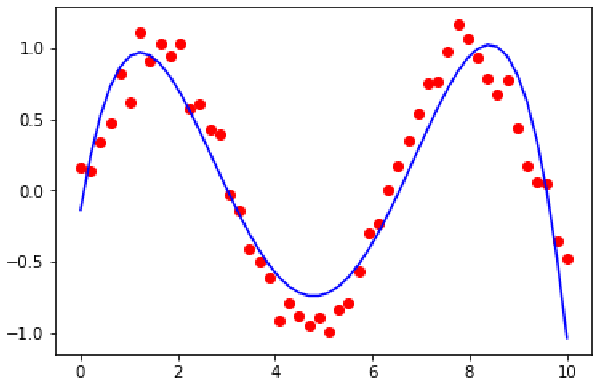
P = 2



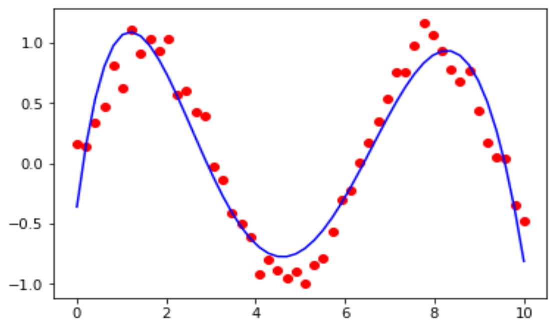
P = 3



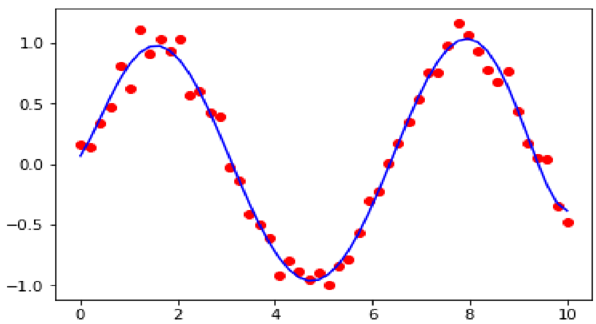
P = 4



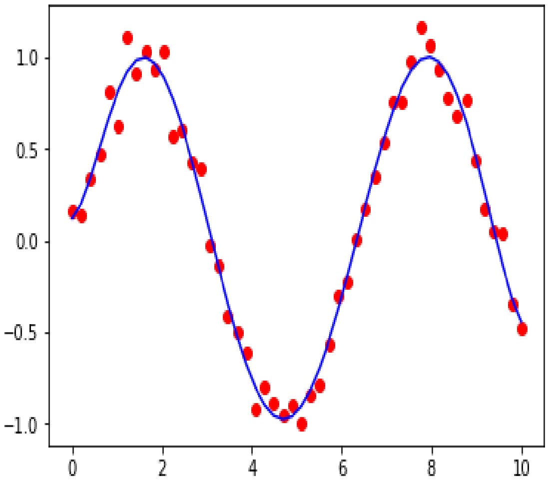
P = 5



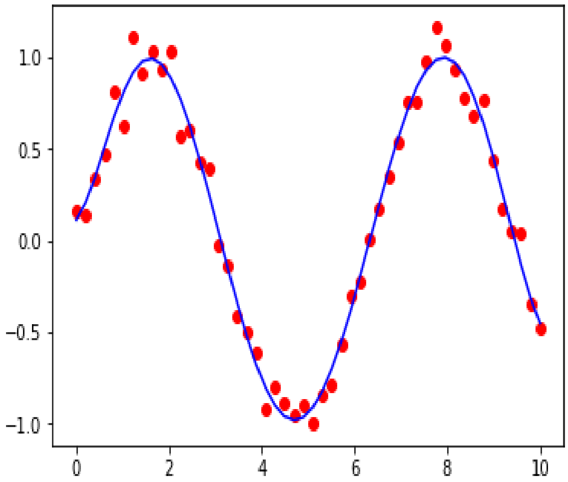
P = 6



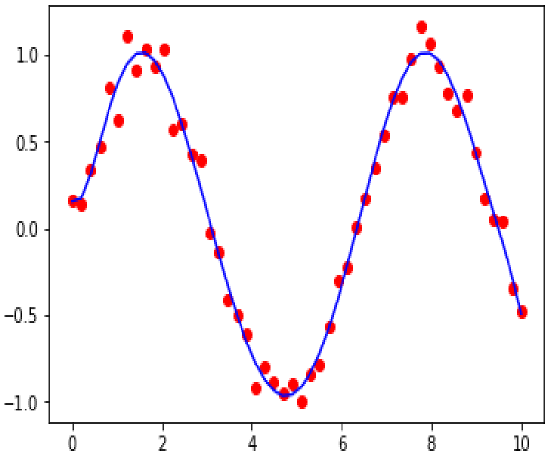
P = 7



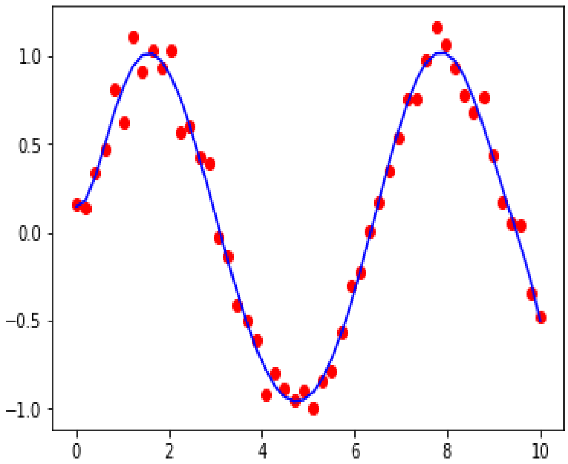
P = 8 (Best Model)



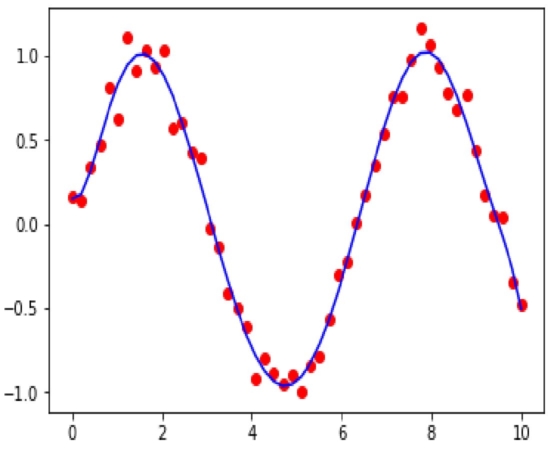
P = 9



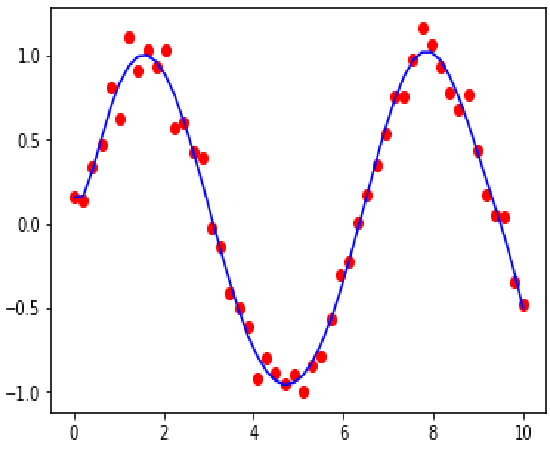
P = 10



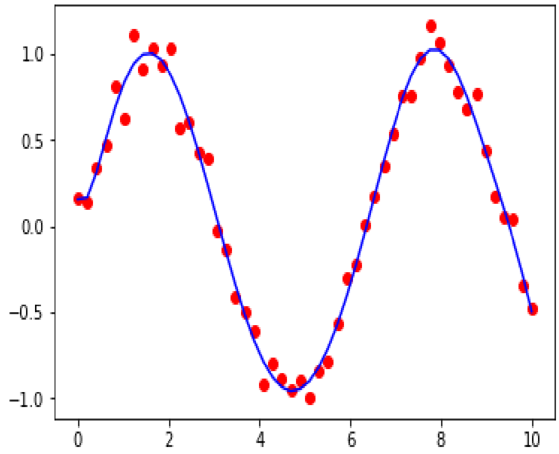
P = 11



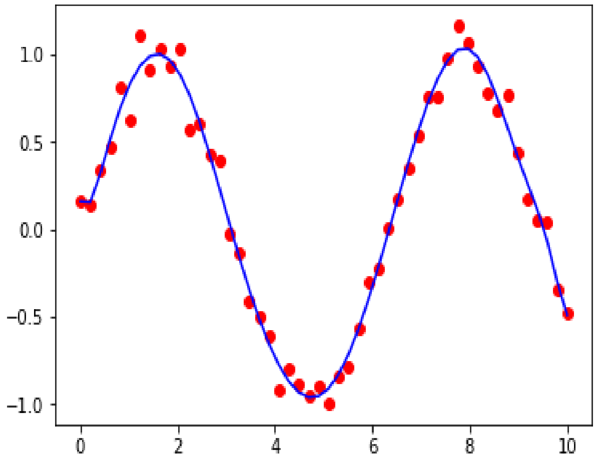
P = 12



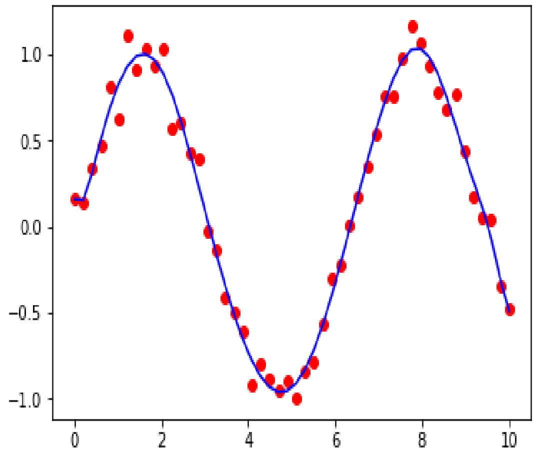
P = 13



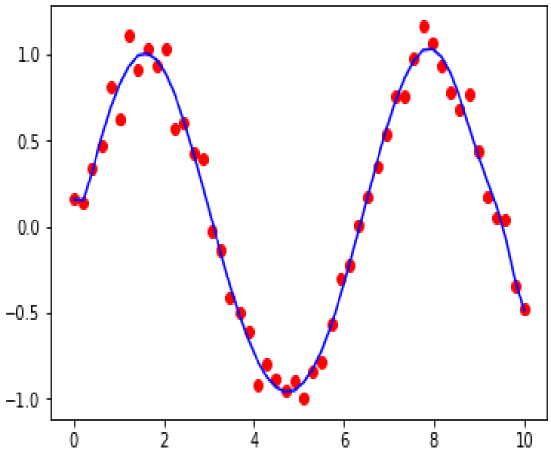
P = 14



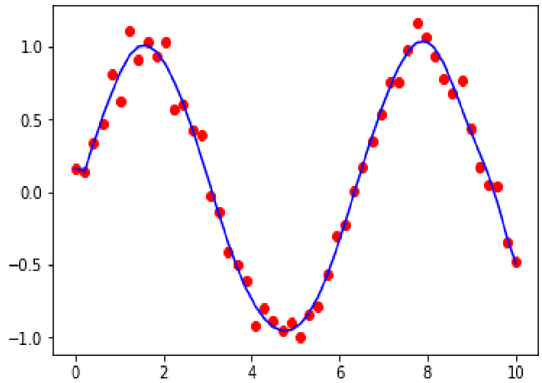
P = 15



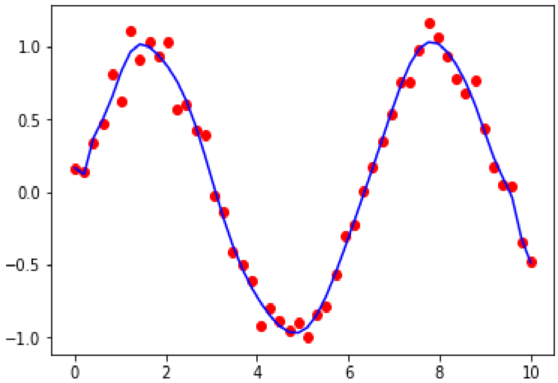
P = 16

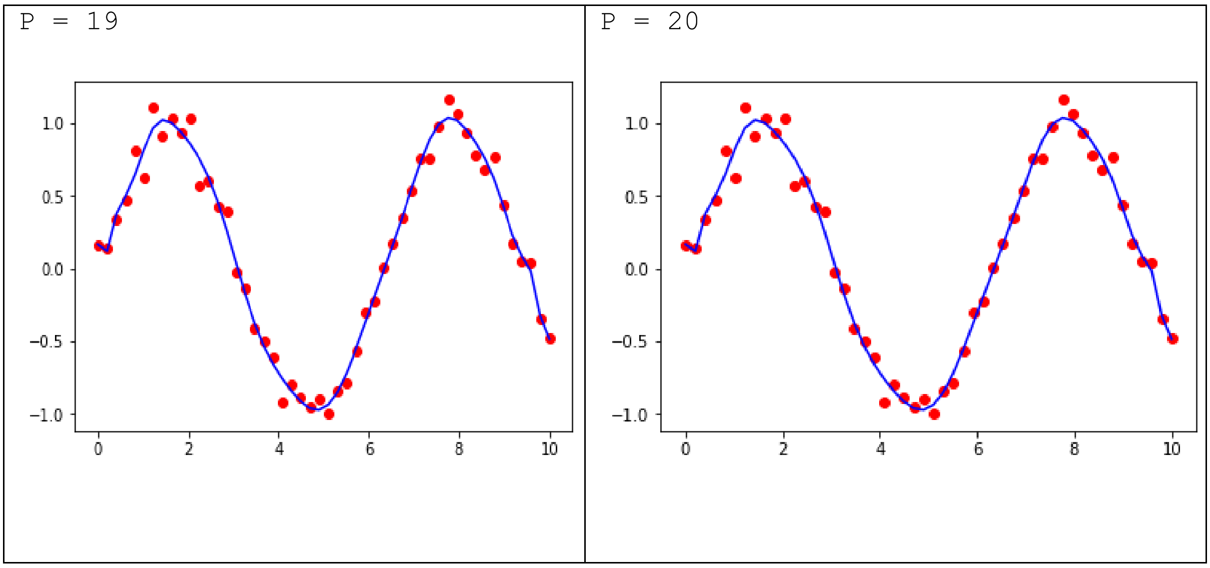


P = 17



P = 18





Combined Model Plot for all the 20 models to show the variation in different models.