Assignment 4

Q1. Implement the normal equation (closed form) regression for the Boston housing dataset. The dataset description can be found here. The target feature is variable no. 14, ’MEDV’, and the input variables are the remaining 13 variables.

Answer:

I implemented the closed form solution for calculating parameters

$$W = (X^T X)^{-1} X^T Y$$

```
def linear_regressor_closed_form(X,Y):
    return (np.linalg.pinv(np.matmul(X.T,X))@X.T@Y)
```

Q1a.Divide the dataset into training and testing using an 80:20 split ratio.

Answer:

```
shape of split data : train data and labels of training dataset:  (404, 13) (404,)
shape of test data and labels (102, 13) (102,)
```

Q1b.Perform Linear regression for all features and compute the RMSE for training as well as the testing set. (Note: There is no need to perform k-fold cross-validation for this part.)

Answer:

```
error on 80 percent dataset in case 1 and on entire dataset for question2  4.534460410890986
```

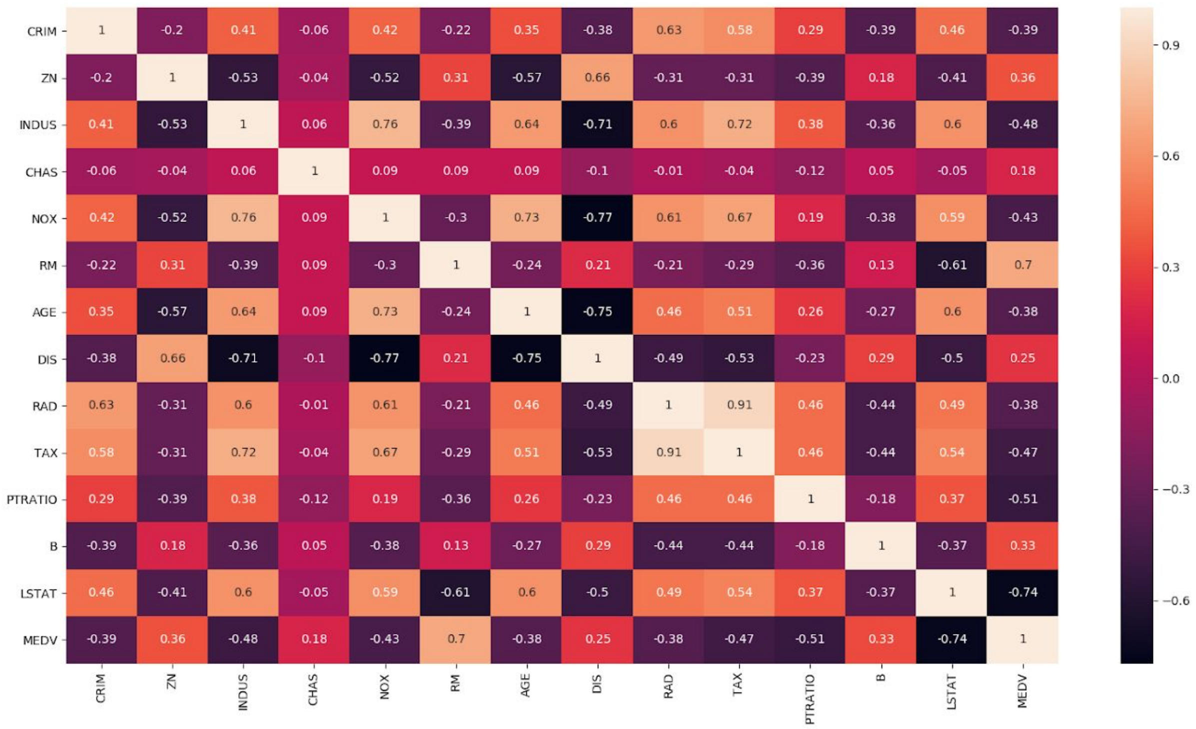
Common message as I have implemented a reusable function. Above is rmse for degree1 using all features.

Test set RMSE:

```
test set error when using all 13 features is 5.376194557068185
```

Q.1c Select the feature named ‘LSTAT’ for polynomial regression.

Answer: From the correlation matrix it is evident that LSTAT and RM are highly correlated with target variable MEDV. RM is positively correlated and LSTAT is negatively correlated.



However RM and LSTAT are also correlated with each other. We choose lstat here by picking the corresponding column from dataframe.

Q1d. K-fold-cross validation:

```
def k_fold_cross_validation(data, k, labels):
    divided_data = np.array_split(data, k)
    labels = np.array_split(labels, k)
    train_errors = []
    val_errors = []
    for i in range(k):
        #copy to avoid mutation of original data
        data_for_fold = divided_data.copy()
        validation_data = divided_data[i]
        validation_labels = labels[i]
        train_labels = labels.copy()
        del data_for_fold[i]
        del train_labels[i]
        print((np.concatenate(train_labels).shape))
        train_data_for_fold = np.concatenate( data_for_fold, axis=0 )
        weights_estimated = linear_regressor_closed_form(train_data_for_fold, np.concatenate(train_labels))
        predictions_train = predict(train_data_for_fold, weights_estimated)
        train_error = RMSE_error(np.concatenate(train_labels), predictions_train)
        print("train_error in 5 _fold  cross validation", train_error)
        train_errors.append(train_error)
        predictions_val = predict(validation_data, weights_estimated)
        val_error = RMSE_error(validation_labels, predictions_val)
        val_errors.append(val_error)
        print("val_error in 5 fold cross validation", val_error)
    return train_errors, val_errors
```

Q1e

Performing k-fold cross validation for different degrees of polynomial:

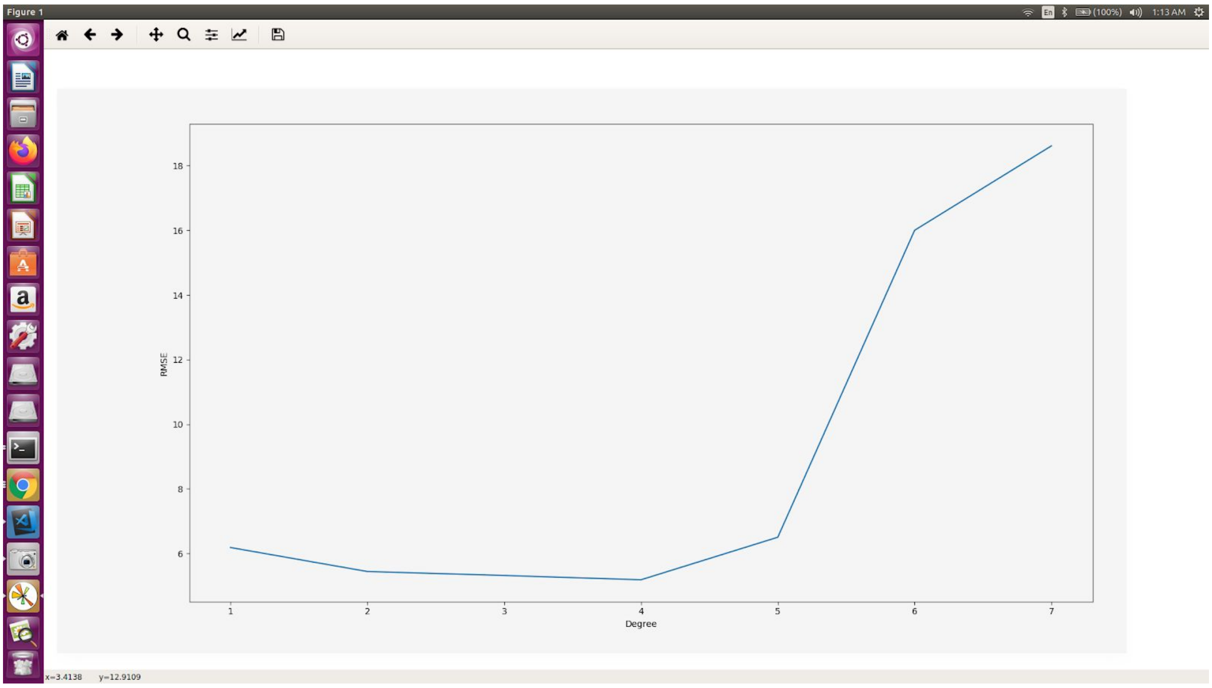
Sample output: (not exhaustive).

```
rain_error in 5 _fold  cross validation 6.451338929759388
al_error in 5 fold cross validation 5.112272612932732
323,)
rain_error in 5 _fold  cross validation 6.261610013233084
al_error in 5 fold cross validation 5.991334914304872
323,)
rain_error in 5 _fold  cross validation 5.848379651051819
al_error in 5 fold cross validation 7.495842834216626
323,)
rain_error in 5 _fold  cross validation 6.186131323262918
al_error in 5 fold cross validation 6.244847326434905
324,)
rain_error in 5 _fold  cross validation 6.174428441534089
al_error in 5 fold cross validation 6.295075516649748
_fold_cross_validation mean train error for degree 1 and validations errors are:  6.18437767176826 6.2278746409077765
_with_bias_terms [ 1.      6.72    45.1584]
alue of weights using closed form linear regressor is (404, 3) (3,)
rror on 80 percent dataset in case 1 and on entire dataset for question2  5.45258754044224
323,)
rain_error in 5 _fold  cross validation 5.666989526242768
al_error in 5 fold cross validation 4.5736543063317265
323,)
rain_error in 5 _fold  cross validation 5.487193802493503
al_error in 5 fold cross validation 5.383633484671677
323,)
rain_error in 5 _fold  cross validation 5.195286012243591
al_error in 5 fold cross validation 6.4442187234786195
323,)
rain_error in 5 _fold  cross validation 5.406448096805383
al_error in 5 fold cross validation 5.644579219721294
324,)
rain_error in 5 _fold  cross validation 5.452651222386622
al_error in 5 fold cross validation 5.463420454845786
_fold_cross_validation mean train error for degree 2 and validations errors are:  5.4417137320343745 5.50190123780982
```

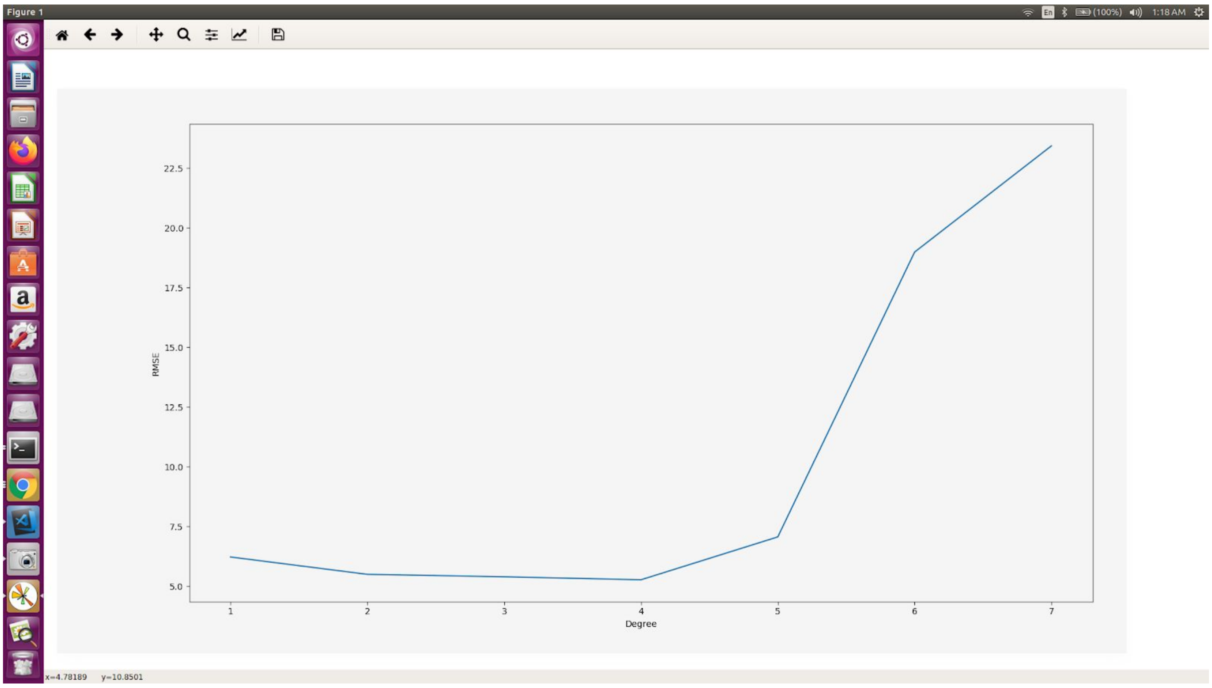
Q1f.

Computing RMSE and mean of RMSE for different degrees(1,2,3,4,5,6,7 etc.) and plotting it:

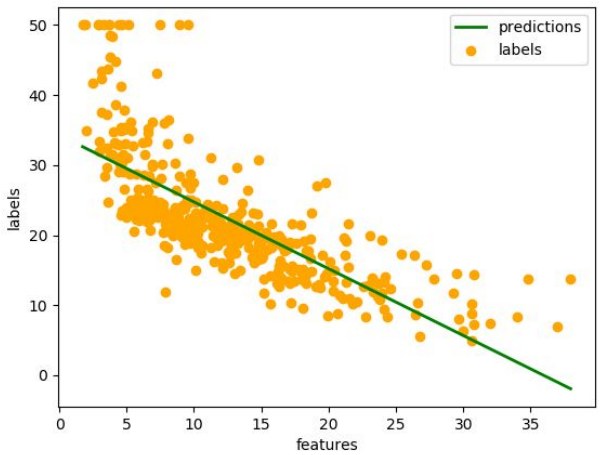
Train mean RMSE plot:



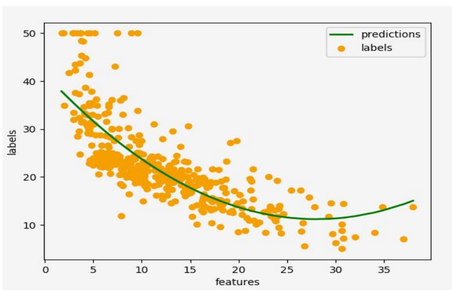
Validation Mean RMSE plot:



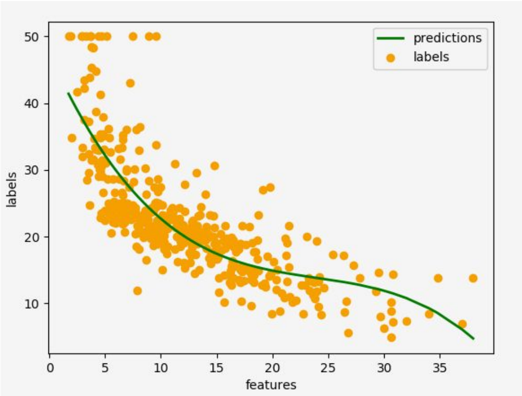
Fitted line plots for degree 1:



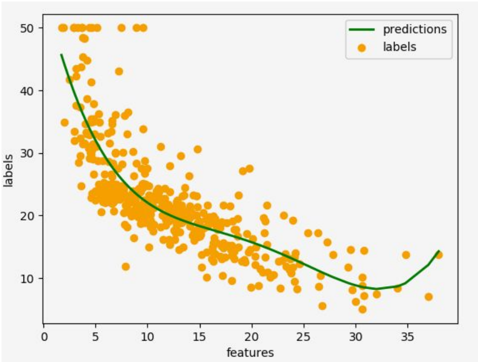
For degree 2:



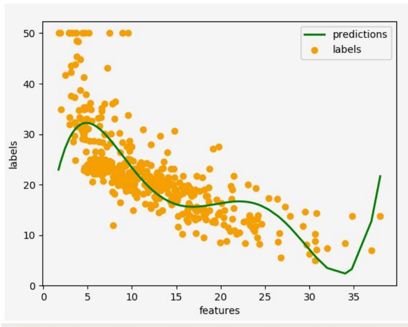
For degree 3:



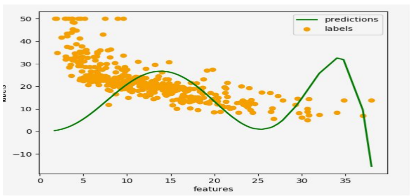
For degree 4:



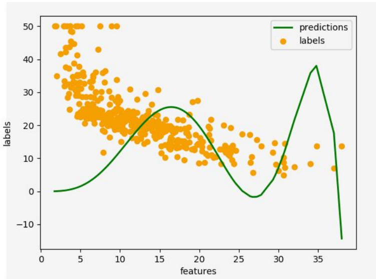
For degree 5:



For degree 6:



For degree 7:



Q1g.
Choosing degree of polynomial with lowest mean validation RMSE, and performing regression on training set and test set :

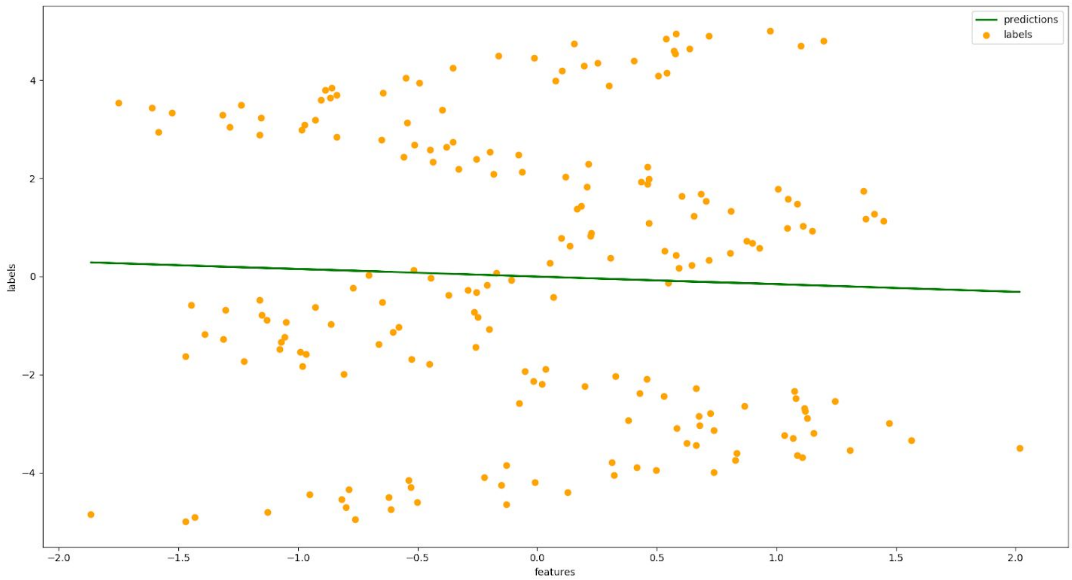
```
min_degree 4
best degree is 4 and  validation error is  5.274042824968437


RMSE on 80 percent train dataset for best degree is  5.197111228704988


RMSE on 20 percent test dataset for best degree is  5.513721075111767
```

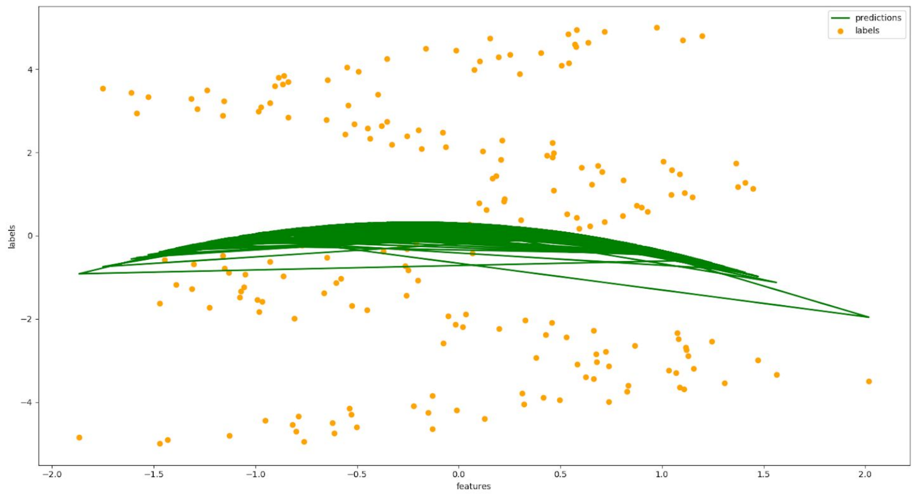
Q2
A and B

.
Fitted line for degree1:

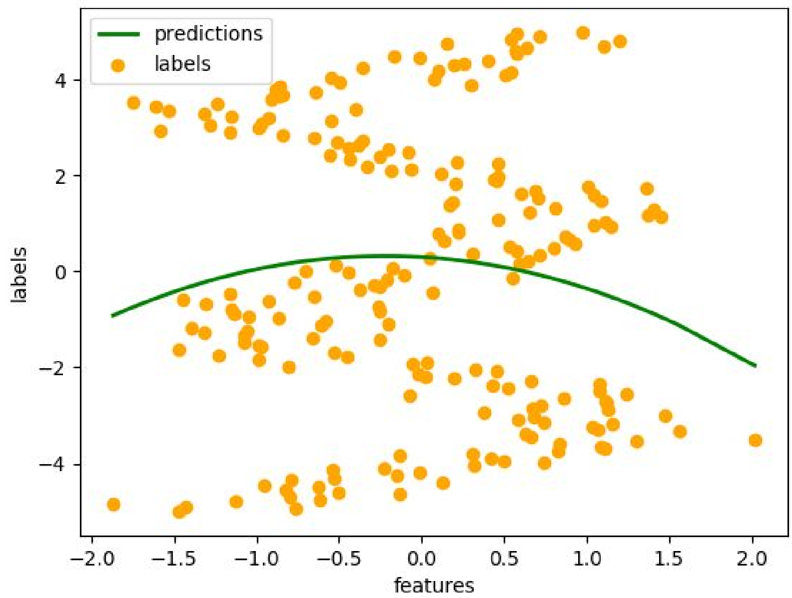


RMSE is 2.8984450960530683

For degree 2:

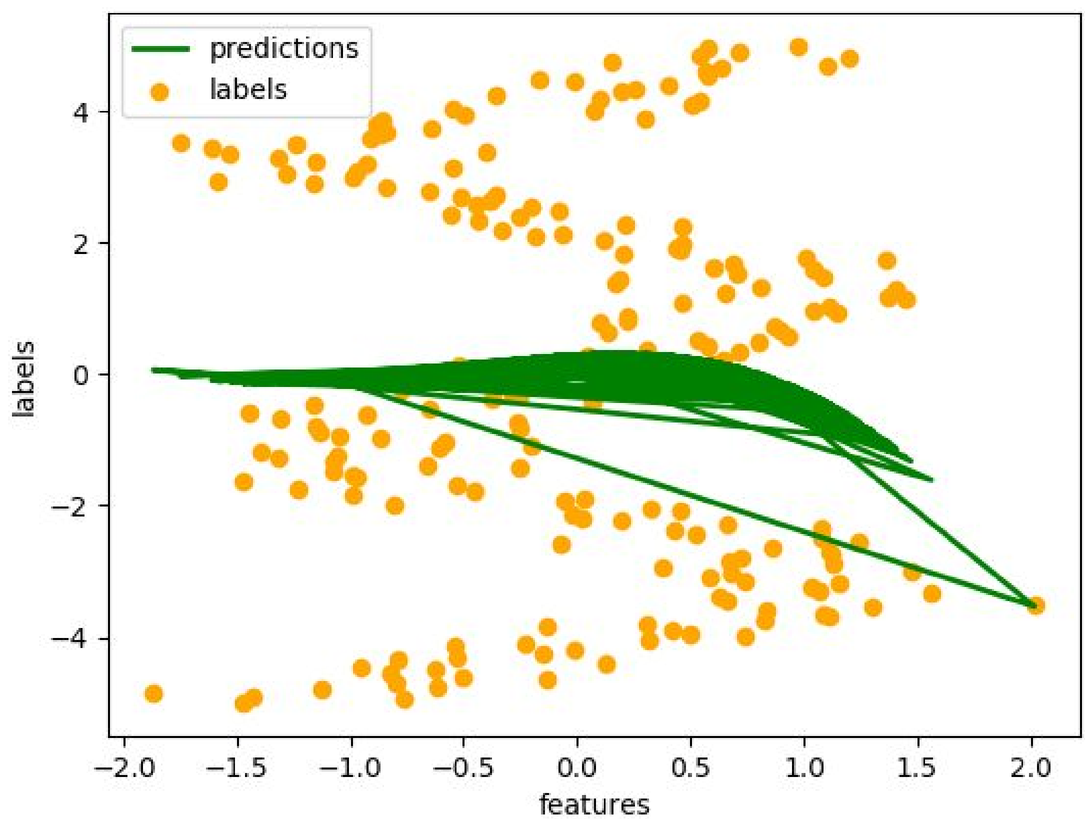


I could take the average line for a more clear plot but wanted to illustrate the default plot and fitted model.Or i could sort x values before line plot
After sorting X values the plot we get is

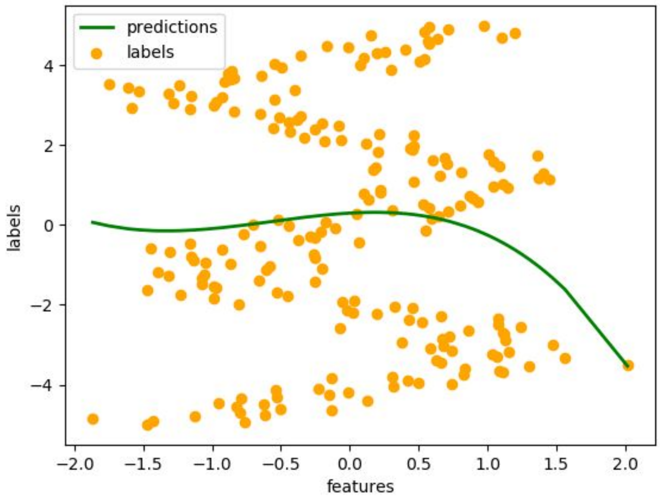


RMSE for degree2: 2.88000505198886

For degree 4: Before sorting for line plot :



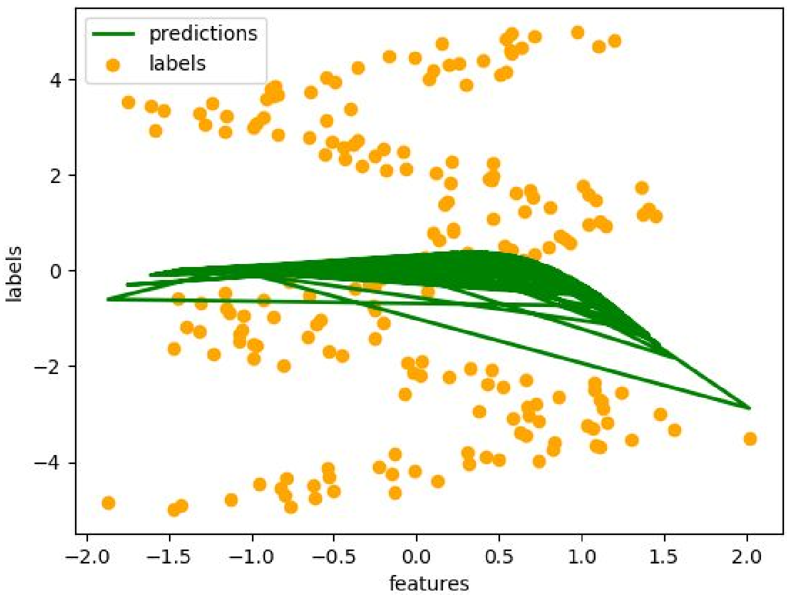
After sorting for line plot:



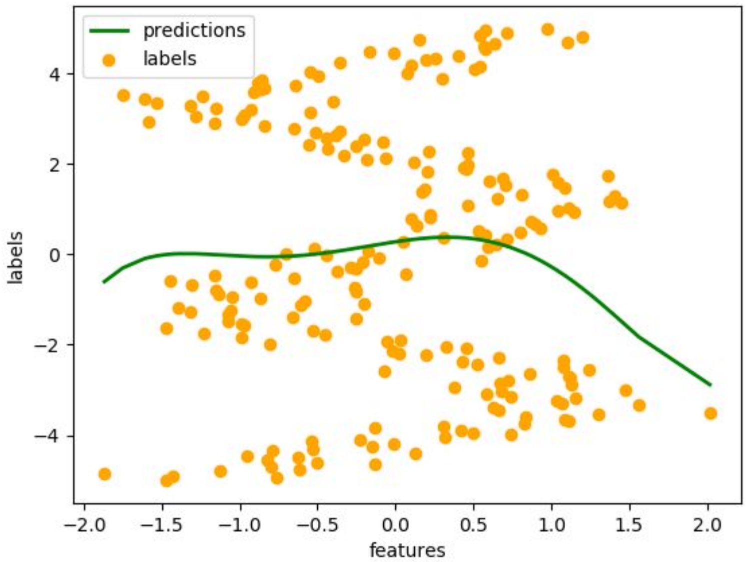
RMSE is:

```
2.8726896919751668
```

For degree 5: Before sorting :

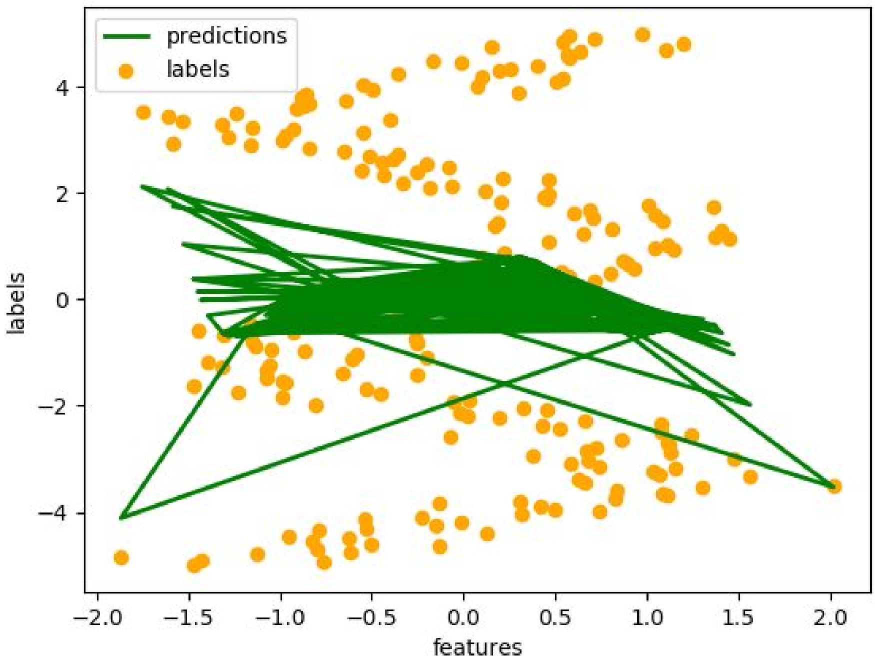


After sorting for line plot:

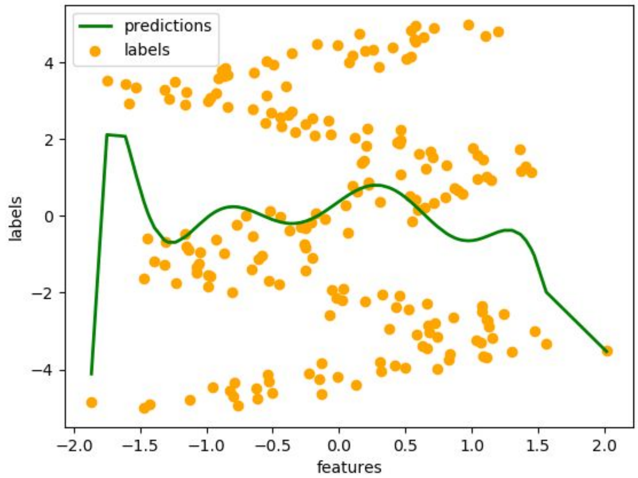


RMSE is: 2.870418058719128

For degree 10:Before sorting

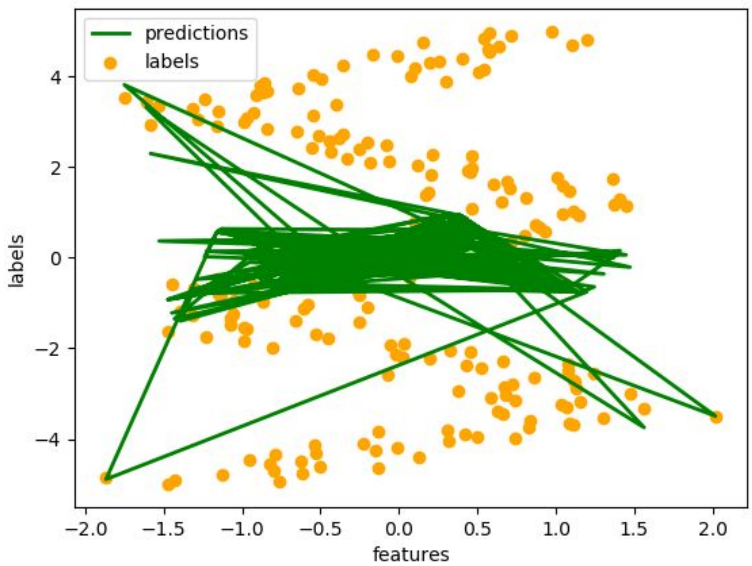


After sorting for line plot:

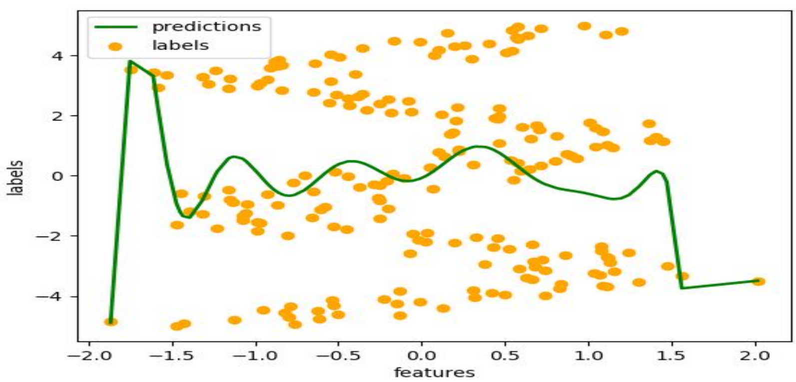


RMSE is 2.8293330859285635

For degree 15: Before sorting:

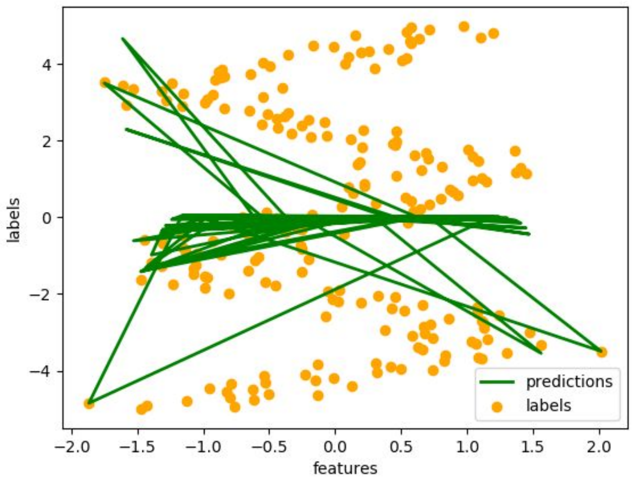


After sorting line plot becomes:

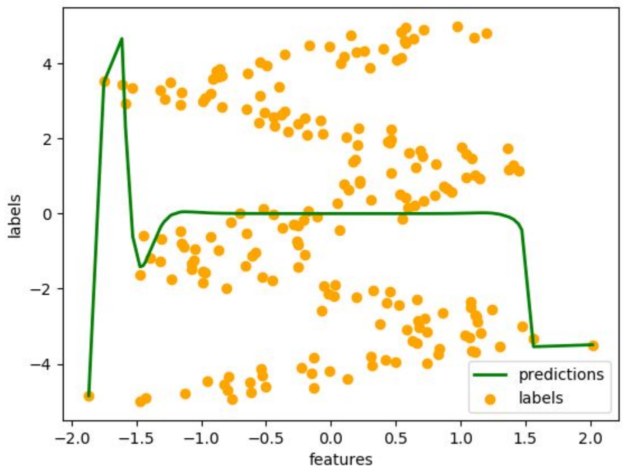


RMSE: 2.78364761632952

For degree 30: Before sorting

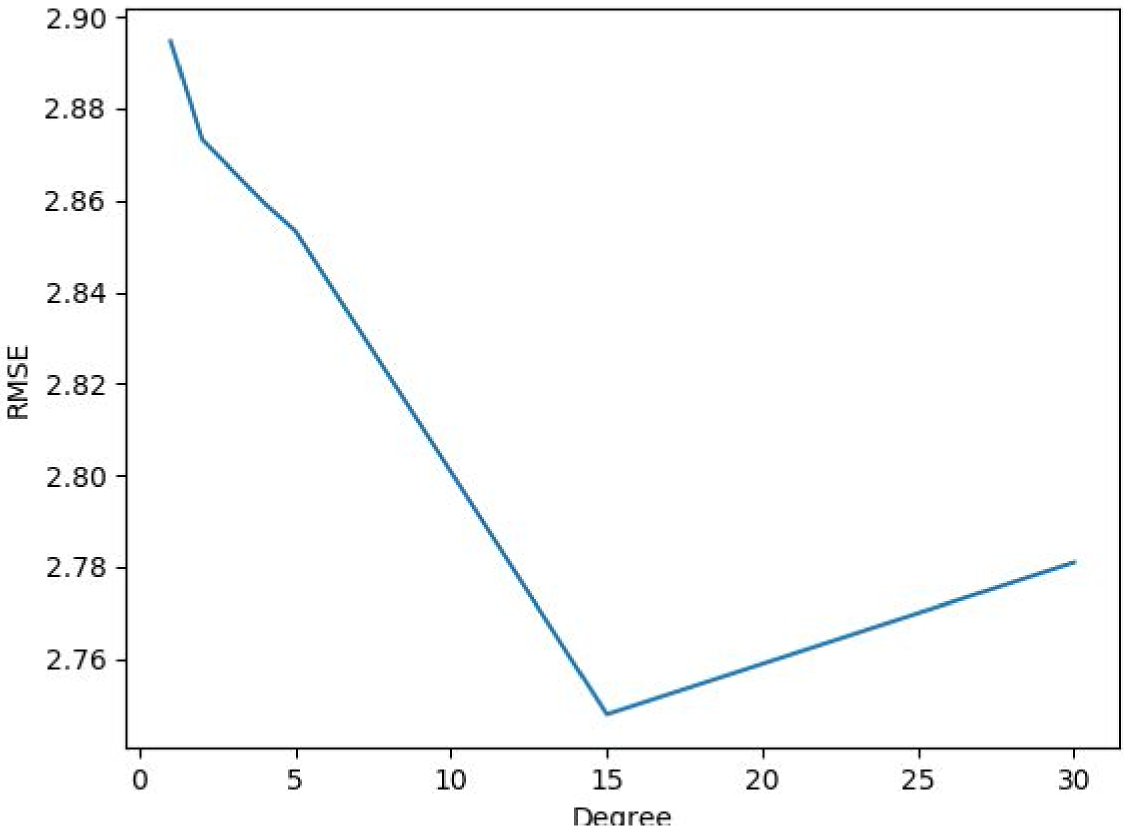


After sorting:

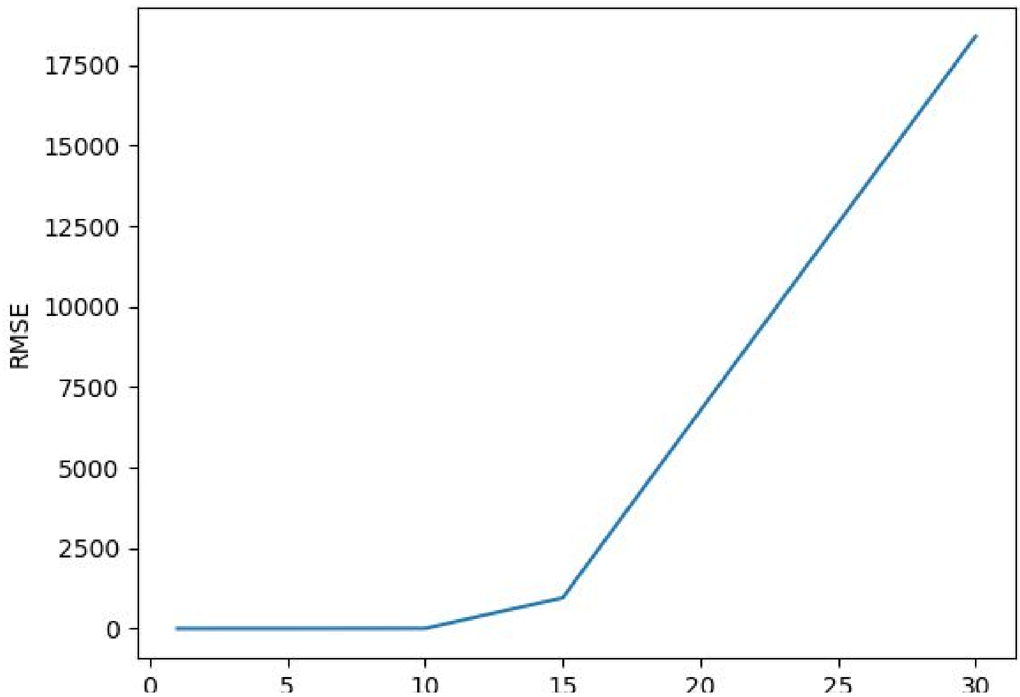


RMSE: 2.8162022462288094

RMSE plot for different degrees



Mean RMSE plot of val set when performing k-fold cross validation is



Sample RMSE during cross validations (only for degree 5 and 10 screenshot is given for brevity but was performed for all degrees as seen in above plot)

```
train_error in 5 _fold  cross validation 2.8539701384613547
val_error in 5 fold cross validation 3.2084837653531113
(160,)
train_error in 5 _fold  cross validation 2.854672892325671
val_error in 5 fold cross validation 3.0605412047228477
(160,)
train_error in 5 _fold  cross validation 2.8179448437444927
val_error in 5 fold cross validation 3.144316122516692
(160,)
train_error in 5 _fold  cross validation 2.8532052412981654
val_error in 5 fold cross validation 2.9721621906710993
(160,)
train_error in 5 _fold  cross validation 2.8869203186951444
val_error in 5 fold cross validation 2.8255074481833757
k_fold_cross_validation mean train error for degree 5 and validations errors are:  2.853342686904966 3.042202146289425
X_with_bias_terms [ 1.         -0.61215961  0.37473939 -0.22940032  0.14042961 -0.08596533
  0.0526245  -0.0322146   0.01972047 -0.01207208  0.00739004]


value of weights using closed form linear regressor is (200, 11) (11,)


error on 80 percent dataset in case 1 and on entire dataset for question2  2.8293330859285635




(160,)
train_error in 5 _fold  cross validation 2.824170666866898
val_error in 5 fold cross validation 20.533673608036114
(160,)
train_error in 5 _fold  cross validation 2.7956939932109126
val_error in 5 fold cross validation 3.1047145635508877
(160,)
train_error in 5 _fold  cross validation 2.7710049293444357
val_error in 5 fold cross validation 3.2927424713257167
(160,)
train_error in 5 _fold  cross validation 2.791367689645203
val_error in 5 fold cross validation 3.0248301908819033
(160,)
train_error in 5 _fold  cross validation 2.821695812592324
val_error in 5 fold cross validation 2.91205111292586
k_fold_cross_validation mean train error for degree 10 and validations errors are:  2.8007866183319545 6.573602389344098
```