Linear Regression on Boston Housing-Prices Dataset

April 15, 2021

[1]:

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.datasets import load_boston
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
def linear_regression_all_features(X, y, plot, x_label="", y_label=""):
    # Step 1 split the dataset into training and test sets(80,20)
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
    lm = LinearRegression()
    # Step 2
    #fit the model
    lm.fit(X_train, y_train)
    predictions = lm.predict(X_test)

    # The coefficient(s).
    coef = lm.coef_

    # The mean square error.
    MSE = np.mean(((predictions - y_test) ** 2))

    # Explained variance score (1 is perfect prediction).
    vs = lm.score(X_test, y_test)

    if plot:
        plt.figure(figsize=(4, 3))
        ax = plt.axes()
        ax.scatter(X_test, y_test, color='turquoise')
        ax.scatter(X_train, y_train, color='pink')
        ax.plot(X_test, predictions, color='black', linewidth=3)

        ax.set_xlabel(x_label)
        ax.set_ylabel(y_label)

        plt.show()

    return coef, MSE, vs
```

1 Report the coefficients, mean squared error and variance score for the model on the test set

[2]:

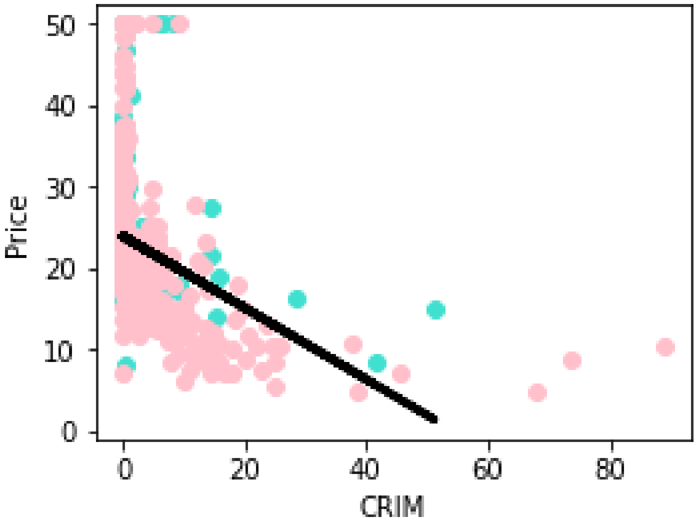
```
# Step 1 and Step 2 Output
data = load_boston()
X, y = data.data, data.target
coef, MSE, vs = linear_regression_all_features(X, y, False)
print("Coefficient:", coef)
print ("Mean squared error: %.2f" % MSE)
print ("Variance score: %.2f" % vs)
```

```
Coefficient: [-1.12241160e-01  6.18851654e-02  1.69236228e-02  3.31057053e+00
 -2.17012092e+01  3.74437883e+00 -5.79573689e-03 -1.70702714e+00
  2.89202004e-01 -1.17006172e-02 -9.48065547e-01  8.53879327e-03
 -4.90078381e-01]
Mean squared error: 22.63
Variance score: 0.70
```

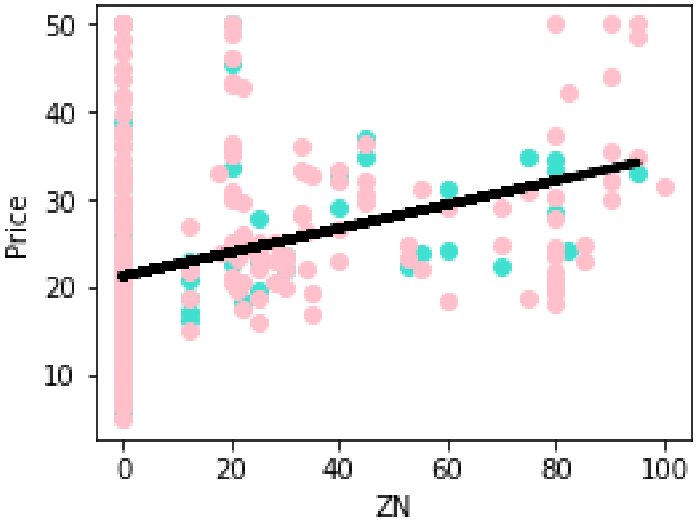
2 Report the coefficient, mean squared error and variance score for the model on the test set and 13 plots of the linear regression models generated on each feature

[3]:

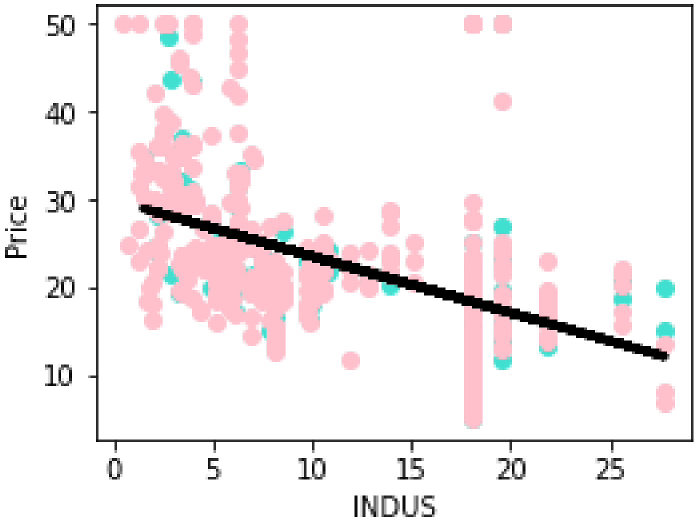
```
# Step 3
def linear_regression_each_feature(df,plot):
    names = data.feature_names
    features = list(df.columns)
    coefs = []
    MSEs = []
    vss = []
    for col in df[features]:
        coef, MSE, vs = linear_regression_all_features(np.reshape(df[col].
 ↪values, (-1, 1)), y, plot, names[col], 'Price' )
        coefs.append(coef)
        MSEs.append(MSE)
        vss.append(vs)
        print(names[col])
        print("Coefficient:", coef)
        print("Mean squared error: %.2f" % MSE)
        print("Variance score: %.2f" % vs)
        print("\n")
    return coefs, MSEs, vss

df = pd.DataFrame(X)
_,_,_ = linear_regression_each_feature(df, True)
```



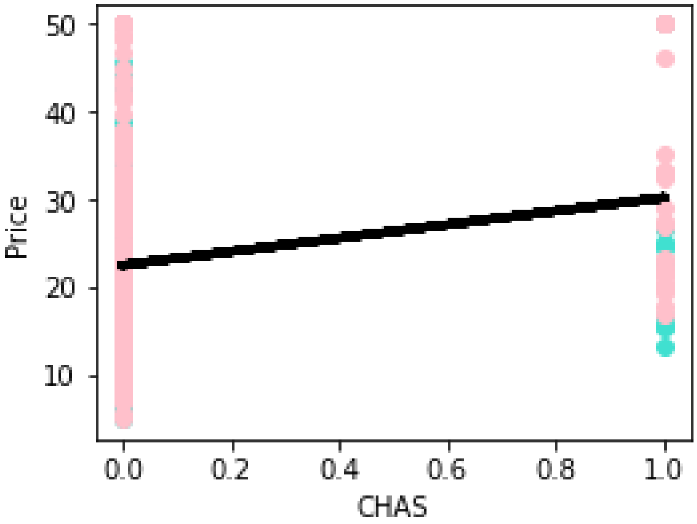
```
CRIM
Coefficient: [-0.4390441]
Mean squared error: 80.80
Variance score: 0.03
```



```
ZN
Coefficient: [0.13588277]
Mean squared error: 49.27
Variance score: 0.22
```

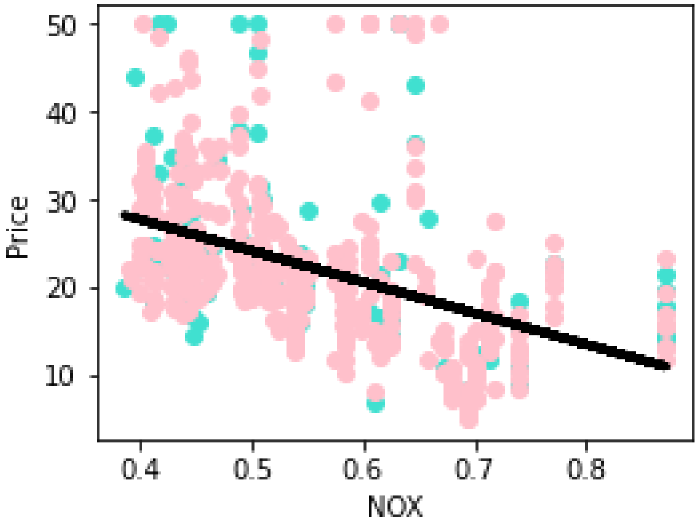


```
INDUS
Coefficient: [-0.64038778]
Mean squared error: 47.54
Variance score: 0.27
```

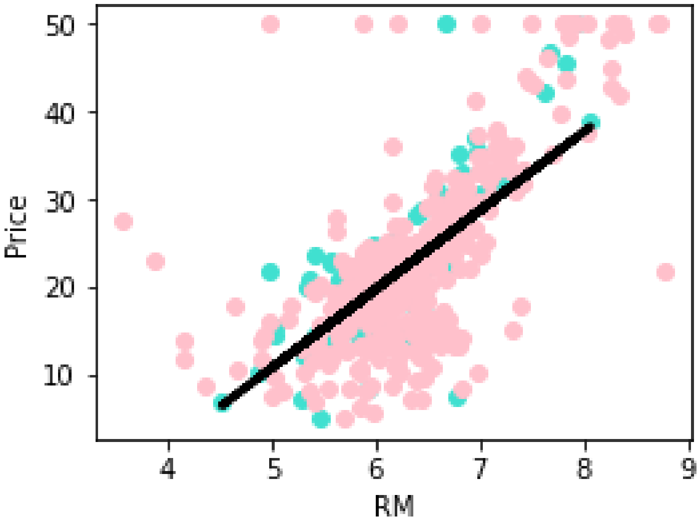


```
CHAS
```

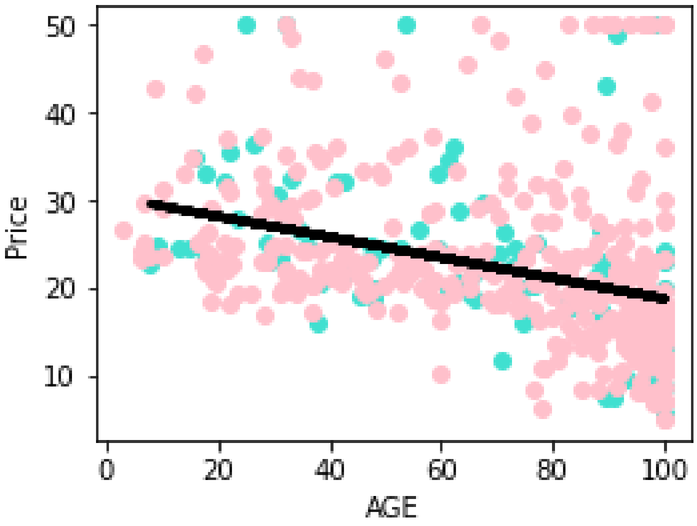
```
Coefficient: [7.69023199]
Mean squared error: 85.72
Variance score: -0.10
```



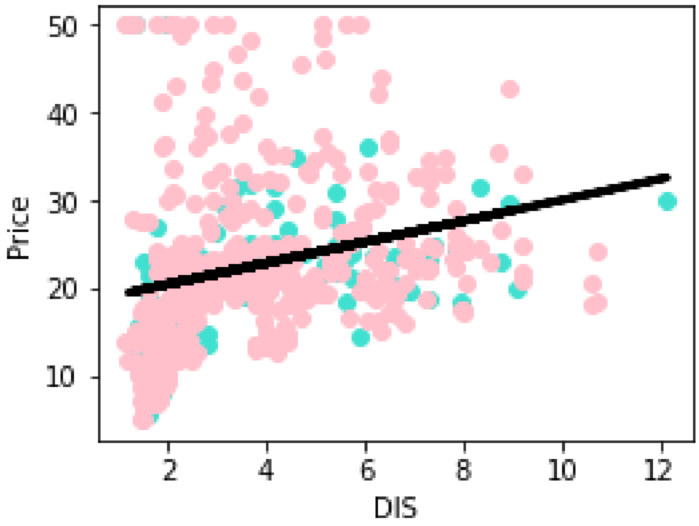
```
NOX
Coefficient: [-35.4375942]
Mean squared error: 88.82
Variance score: 0.10
```



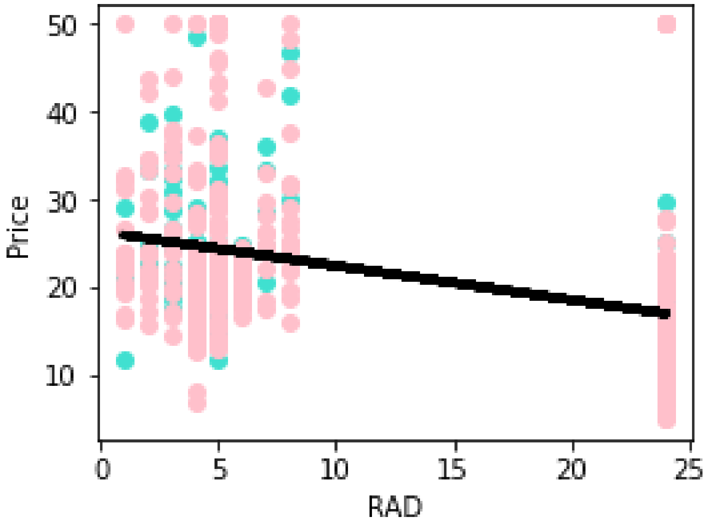
```
RM
Coefficient: [8.93689179]
Mean squared error: 30.42
Variance score: 0.58
```



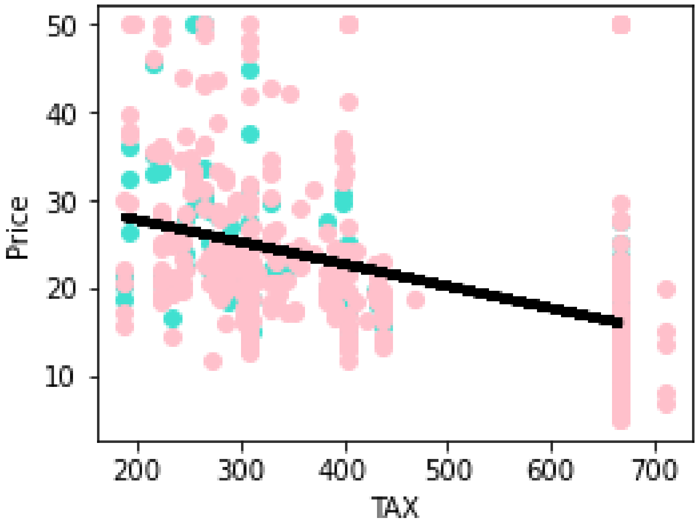
```
AGE
Coefficient: [-0.1169018]
Mean squared error: 66.39
Variance score: 0.19
```



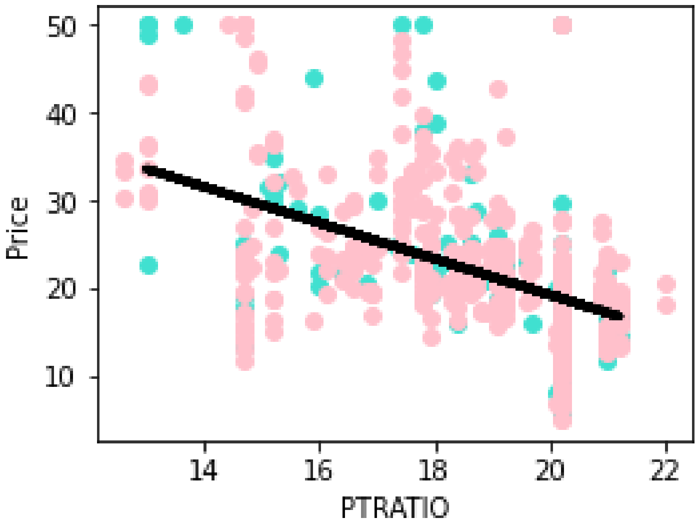
```
DIS
Coefficient: [1.19452222]
Mean squared error: 52.49
Variance score: 0.03
```



```
RAD
Coefficient: [-0.38762479]
Mean squared error: 56.26
Variance score: 0.20
```

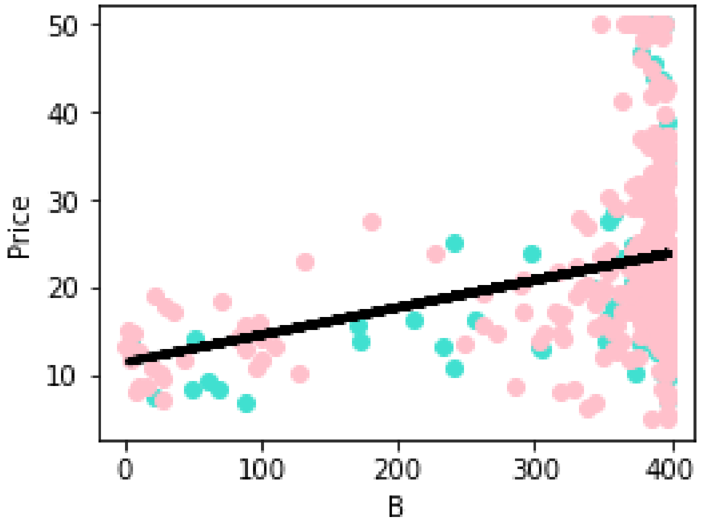


```
TAX
Coefficient: [-0.02501992]
Mean squared error: 45.16
Variance score: 0.33
```

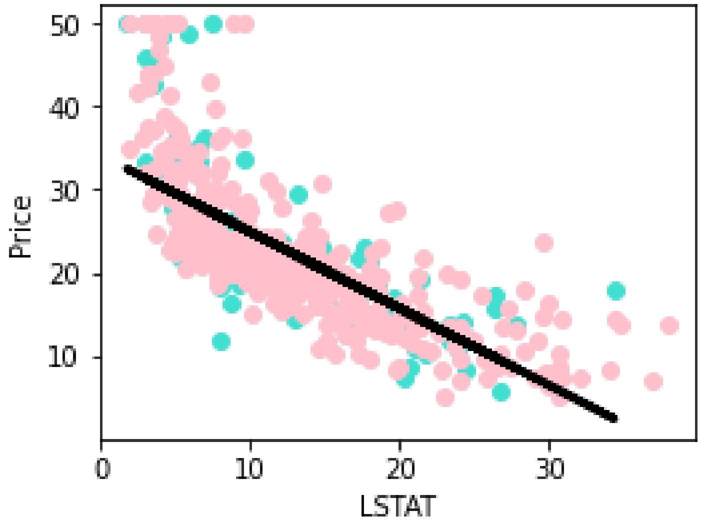


```
PTRATIO
```

```
Coefficient: [-2.04464556]
Mean squared error: 70.84
Variance score: 0.31
```



```
B
Coefficient: [0.03112865]
Mean squared error: 92.22
Variance score: 0.14
```



```
LSTAT
Coefficient: [-0.92225893]
Mean squared error: 48.73
Variance score: 0.54
```

[4]:

```
# Step 4 Perform 10 iterations of Step 1, Step 2, and Step 3
sum_coefs = 0
sum_MSEs = 0
sum_vss = 0
for i in range(0, 10):
    coef, MSE, vs = linear_regression_all_features(X, y, False)
    sum_coefs += coef
    sum_MSEs += MSE
    sum_vss += vs

avg_coefs_all = sum_coefs / 10.0
avg_MSEs_all = sum_MSEs / 10.0
avg_vss_all = sum_vss / 10.0

print("average coefficient: " + str(avg_coefs_all))
print("average Mean square error: " + str(avg_MSEs_all))
print("average Variance score: " + str(avg_vss_all))
```

```
average coefficient: [-1.07711963e-01  4.64055441e-02  1.29488521e-02
3.09031782e+00
 -1.77612987e+01  3.62484899e+00  2.23022585e-03 -1.49096663e+00
  2.99264924e-01 -1.07942717e-02 -9.29543279e-01  9.87637055e-03
 -5.55499366e-01]
average Mean square error: 20.328327791570242
average Variance score: 0.7535063998516651
```

[5]:

```
avg_coefs = [0.0] * 13
avg_MSEs = [0.0] * 13
avg_vss = [0.0] * 13
names = data.feature_names
for i in range(0, 10):
    print("\nIteration " + str(i))
    df = pd.DataFrame(X)
    coef, MSE, vs = linear_regression_each_feature(df, False)
    for j in range(0, 13):
        avg_coefs[j] += coef[j] / 13.0
        avg_MSEs[j] += MSE[j] / 13.0
        avg_vss[j] += vs[j] / 13.0
```

```python
print("\nAverages: ")
for k in range(0, 13):
    print("\n" + names[k])
    print("average coefficient: " + str(avg_coefs[k]))
    print("average Mean square error: " + str(avg_MSEs[k]))
    print("average Variance score: " + str(avg_vss[k]))
```

```
Iteration 0
CRIM
Coefficient: [-0.36891375]
Mean squared error: 75.99
Variance score: 0.21


ZN
Coefficient: [0.13921563]
Mean squared error: 69.53
Variance score: 0.18


INDUS
Coefficient: [-0.65757171]
Mean squared error: 73.62
Variance score: 0.21


CHAS
Coefficient: [7.60254559]
Mean squared error: 68.18
Variance score: -0.05


NOX
Coefficient: [-34.62602971]
Mean squared error: 67.73
Variance score: 0.16


RM
Coefficient: [9.44399216]
Mean squared error: 49.72
Variance score: 0.36


AGE
Coefficient: [-0.120893]
```

```
Mean squared error: 72.92
Variance score: 0.18


DIS
Coefficient: [1.16280275]
Mean squared error: 73.75
Variance score: 0.03


RAD
Coefficient: [-0.39635705]
Mean squared error: 80.55
Variance score: 0.14


TAX
Coefficient: [-0.02438432]
Mean squared error: 68.90
Variance score: 0.27


PTRATIO
Coefficient: [-2.17122752]
Mean squared error: 68.23
Variance score: 0.24


B
Coefficient: [0.0356249]
Mean squared error: 89.42
Variance score: 0.07


LSTAT
Coefficient: [-0.96150729]
Mean squared error: 28.27
Variance score: 0.61



Iteration 1
CRIM
Coefficient: [-0.39455029]
Mean squared error: 65.76
Variance score: 0.14
```

```
ZN
Coefficient: [0.13196774]
Mean squared error: 83.44
Variance score: 0.13


INDUS
Coefficient: [-0.66104169]
Mean squared error: 57.94
Variance score: 0.24


CHAS
Coefficient: [6.84509425]
Mean squared error: 78.38
Variance score: -0.00


NOX
Coefficient: [-33.44615384]
Mean squared error: 73.97
Variance score: 0.20


RM
Coefficient: [9.52806613]
Mean squared error: 53.02
Variance score: 0.21


AGE
Coefficient: [-0.12520957]
Mean squared error: 57.40
Variance score: 0.13


DIS
Coefficient: [1.12813101]
Mean squared error: 81.79
Variance score: 0.02


RAD
Coefficient: [-0.4161014]
Mean squared error: 51.10
Variance score: 0.12
```

```
TAX
Coefficient: [-0.0263536]
Mean squared error: 47.93
Variance score: 0.14


PTRATIO
Coefficient: [-2.14971847]
Mean squared error: 72.48
Variance score: 0.20


B
Coefficient: [0.03197662]
Mean squared error: 66.46
Variance score: 0.17


LSTAT
Coefficient: [-0.95683029]
Mean squared error: 33.41
Variance score: 0.59



Iteration 2
CRIM
Coefficient: [-0.50532326]
Mean squared error: 119.24
Variance score: 0.04


ZN
Coefficient: [0.13991379]
Mean squared error: 64.67
Variance score: 0.18


INDUS
Coefficient: [-0.64268651]
Mean squared error: 56.33
Variance score: 0.30


CHAS
Coefficient: [7.54293404]
Mean squared error: 86.67
Variance score: -0.02
```

```
NOX
Coefficient: [-33.99111663]
Mean squared error: 68.29
Variance score: 0.18


RM
Coefficient: [9.09120013]
Mean squared error: 39.41
Variance score: 0.51


AGE
Coefficient: [-0.11614956]
Mean squared error: 63.54
Variance score: 0.21


DIS
Coefficient: [1.1369547]
Mean squared error: 89.30
Variance score: 0.02


RAD
Coefficient: [-0.39334197]
Mean squared error: 49.64
Variance score: 0.19


TAX
Coefficient: [-0.02554722]
Mean squared error: 55.00
Variance score: 0.25


PTRATIO
Coefficient: [-1.96624815]
Mean squared error: 82.19
Variance score: 0.27


B
Coefficient: [0.0333894]
Mean squared error: 66.23
Variance score: 0.09
```

```
LSTAT
Coefficient: [-0.94174782]
Mean squared error: 34.85
Variance score: 0.57



Iteration 3
CRIM
Coefficient: [-0.4636581]
Mean squared error: 76.85
Variance score: 0.04


ZN
Coefficient: [0.14313979]
Mean squared error: 57.34
Variance score: 0.14


INDUS
Coefficient: [-0.63525018]
Mean squared error: 74.02
Variance score: 0.21


CHAS
Coefficient: [5.439218]
Mean squared error: 84.84
Variance score: 0.06


NOX
Coefficient: [-34.40974225]
Mean squared error: 68.07
Variance score: 0.18


RM
Coefficient: [9.0683577]
Mean squared error: 34.94
Variance score: 0.56


AGE
Coefficient: [-0.11850679]
```

```
Mean squared error: 62.39
Variance score: 0.16


DIS
Coefficient: [1.10016833]
Mean squared error: 82.63
Variance score: 0.06


RAD
Coefficient: [-0.41615423]
Mean squared error: 66.58
Variance score: 0.10


TAX
Coefficient: [-0.02671341]
Mean squared error: 63.72
Variance score: 0.17


PTRATIO
Coefficient: [-2.27168156]
Mean squared error: 71.99
Variance score: 0.14


B
Coefficient: [0.03262838]
Mean squared error: 33.04
Variance score: 0.29


LSTAT
Coefficient: [-0.95753707]
Mean squared error: 37.55
Variance score: 0.54



Iteration 4
CRIM
Coefficient: [-0.45494033]
Mean squared error: 77.09
Variance score: 0.08
```

```
ZN
Coefficient: [0.1385655]
Mean squared error: 85.40
Variance score: 0.14


INDUS
Coefficient: [-0.66002891]
Mean squared error: 50.84
Variance score: 0.24


CHAS
Coefficient: [6.72479839]
Mean squared error: 78.46
Variance score: -0.00


NOX
Coefficient: [-31.9321121]
Mean squared error: 55.38
Variance score: 0.27


RM
Coefficient: [9.20374165]
Mean squared error: 49.98
Variance score: 0.48


AGE
Coefficient: [-0.12850565]
Mean squared error: 57.09
Variance score: 0.04


DIS
Coefficient: [1.09738281]
Mean squared error: 92.98
Variance score: 0.04


RAD
Coefficient: [-0.4118916]
Mean squared error: 84.85
Variance score: 0.12
```

```
TAX
Coefficient: [-0.02727016]
Mean squared error: 63.57
Variance score: 0.07


PTRATIO
Coefficient: [-2.15849335]
Mean squared error: 78.81
Variance score: 0.16


B
Coefficient: [0.0336487]
Mean squared error: 61.02
Variance score: 0.11


LSTAT
Coefficient: [-0.90834105]
Mean squared error: 39.47
Variance score: 0.60



Iteration 5
CRIM
Coefficient: [-0.44185127]
Mean squared error: 75.24
Variance score: 0.13


ZN
Coefficient: [0.14362632]
Mean squared error: 83.72
Variance score: 0.10


INDUS
Coefficient: [-0.63048194]
Mean squared error: 69.60
Variance score: 0.22


CHAS
Coefficient: [5.01299735]
Mean squared error: 93.94
Variance score: 0.06
```

```
NOX
Coefficient: [-34.8916728]
Mean squared error: 65.72
Variance score: 0.13


RM
Coefficient: [8.80855934]
Mean squared error: 30.97
Variance score: 0.65


AGE
Coefficient: [-0.11951211]
Mean squared error: 70.29
Variance score: 0.17


DIS
Coefficient: [1.09153418]
Mean squared error: 72.02
Variance score: 0.04


RAD
Coefficient: [-0.39242397]
Mean squared error: 77.74
Variance score: 0.16


TAX
Coefficient: [-0.02560052]
Mean squared error: 54.52
Variance score: 0.25


PTRATIO
Coefficient: [-2.2185797]
Mean squared error: 42.75
Variance score: 0.24


B
Coefficient: [0.03381127]
Mean squared error: 71.13
Variance score: 0.14
```

```
LSTAT
Coefficient: [-0.94926716]
Mean squared error: 43.44
Variance score: 0.51



Iteration 6
CRIM
Coefficient: [-0.43278488]
Mean squared error: 91.53
Variance score: 0.08


ZN
Coefficient: [0.15949607]
Mean squared error: 65.49
Variance score: -0.04


INDUS
Coefficient: [-0.62619494]
Mean squared error: 58.44
Variance score: 0.32


CHAS
Coefficient: [8.45952381]
Mean squared error: 78.59
Variance score: -0.08


NOX
Coefficient: [-32.1707631]
Mean squared error: 74.84
Variance score: 0.24


RM
Coefficient: [8.75558223]
Mean squared error: 36.77
Variance score: 0.60


AGE
Coefficient: [-0.11641587]
```

```
Mean squared error: 82.69
Variance score: 0.16


DIS
Coefficient: [0.99781036]
Mean squared error: 106.75
Variance score: 0.06


RAD
Coefficient: [-0.36930968]
Mean squared error: 56.45
Variance score: 0.25


TAX
Coefficient: [-0.02622125]
Mean squared error: 72.71
Variance score: 0.17


PTRATIO
Coefficient: [-2.15289392]
Mean squared error: 58.01
Variance score: 0.31


B
Coefficient: [0.03505458]
Mean squared error: 76.20
Variance score: 0.06


LSTAT
Coefficient: [-0.93012325]
Mean squared error: 42.81
Variance score: 0.55



Iteration 7
CRIM
Coefficient: [-0.43661787]
Mean squared error: 67.25
Variance score: 0.18
```

```
ZN
Coefficient: [0.14494489]
Mean squared error: 71.81
Variance score: 0.08


INDUS
Coefficient: [-0.67498653]
Mean squared error: 70.78
Variance score: 0.13


CHAS
Coefficient: [6.67622069]
Mean squared error: 73.93
Variance score: 0.01


NOX
Coefficient: [-34.52169101]
Mean squared error: 56.67
Variance score: 0.16


RM
Coefficient: [8.87751449]
Mean squared error: 49.37
Variance score: 0.49


AGE
Coefficient: [-0.12250417]
Mean squared error: 83.52
Variance score: 0.10


DIS
Coefficient: [1.03646342]
Mean squared error: 82.54
Variance score: 0.08


RAD
Coefficient: [-0.41134614]
Mean squared error: 66.76
Variance score: 0.15
```

```
TAX
Coefficient: [-0.02492925]
Mean squared error: 72.22
Variance score: 0.24


PTRATIO
Coefficient: [-2.26513473]
Mean squared error: 59.01
Variance score: 0.17


B
Coefficient: [0.03344326]
Mean squared error: 55.34
Variance score: 0.17


LSTAT
Coefficient: [-0.93735683]
Mean squared error: 40.27
Variance score: 0.57



Iteration 8
CRIM
Coefficient: [-0.401462]
Mean squared error: 69.40
Variance score: 0.11


ZN
Coefficient: [0.14509813]
Mean squared error: 66.49
Variance score: 0.08


INDUS
Coefficient: [-0.63083511]
Mean squared error: 64.84
Variance score: 0.27


CHAS
Coefficient: [5.8718845]
Mean squared error: 68.29
Variance score: 0.05
```

```
NOX
Coefficient: [-33.16568826]
Mean squared error: 84.09
Variance score: 0.17


RM
Coefficient: [8.71797568]
Mean squared error: 38.59
Variance score: 0.59


AGE
Coefficient: [-0.11423798]
Mean squared error: 78.96
Variance score: 0.20


DIS
Coefficient: [1.12049473]
Mean squared error: 76.08
Variance score: 0.05


RAD
Coefficient: [-0.4073681]
Mean squared error: 85.39
Variance score: 0.11


TAX
Coefficient: [-0.0256449]
Mean squared error: 46.58
Variance score: 0.26


PTRATIO
Coefficient: [-2.04032022]
Mean squared error: 76.89
Variance score: 0.28


B
Coefficient: [0.03469737]
Mean squared error: 58.63
Variance score: 0.06
```

```
LSTAT
Coefficient: [-0.93521581]
Mean squared error: 30.98
Variance score: 0.63



Iteration 9
CRIM
Coefficient: [-0.44771132]
Mean squared error: 72.80
Variance score: 0.20


ZN
Coefficient: [0.12890189]
Mean squared error: 59.25
Variance score: 0.24


INDUS
Coefficient: [-0.60215678]
Mean squared error: 66.89
Variance score: 0.31


CHAS
Coefficient: [7.61080739]
Mean squared error: 101.78
Variance score: -0.01


NOX
Coefficient: [-33.85150987]
Mean squared error: 53.25
Variance score: 0.26


RM
Coefficient: [8.66726668]
Mean squared error: 31.31
Variance score: 0.62


AGE
Coefficient: [-0.1211096]
```

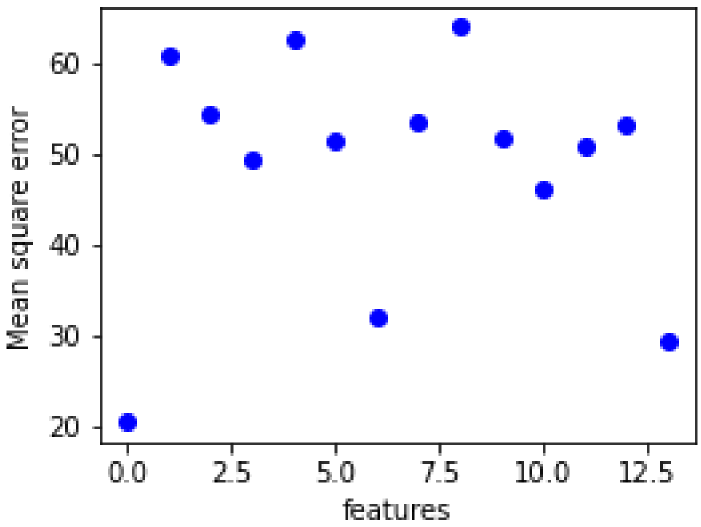
```
Mean squared error: 65.37
Variance score: 0.17


DIS
Coefficient: [1.12915282]
Mean squared error: 74.72
Variance score: 0.05


RAD
Coefficient: [-0.4048638]
Mean squared error: 53.13
Variance score: 0.18


TAX
Coefficient: [-0.02565223]
Mean squared error: 53.94
Variance score: 0.25


PTRATIO
Coefficient: [-2.1775272]
Mean squared error: 50.50
Variance score: 0.30


B
Coefficient: [0.03304081]
Mean squared error: 112.50
Variance score: 0.03


LSTAT
Coefficient: [-0.92973013]
Mean squared error: 49.90
Variance score: 0.45



Averages:

CRIM
average coefficient: [-0.33444716]
average Mean square error: 60.857729678185734
average Variance score: 0.09279574320584574
```

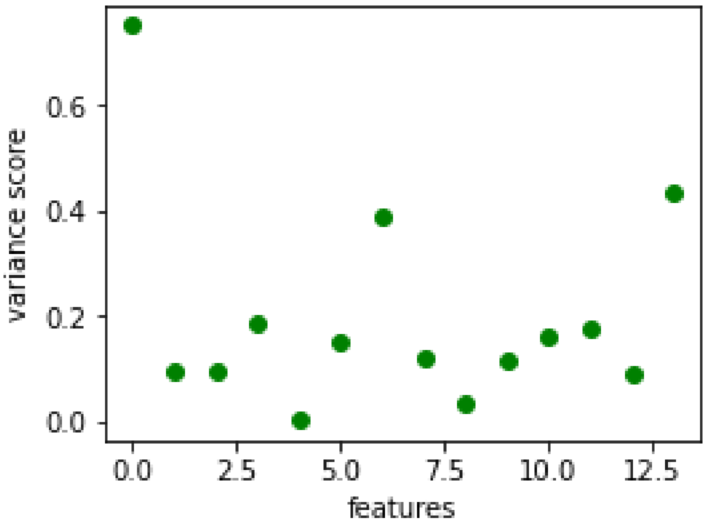
```
ZN
average coefficient: [0.10883614]
average Mean square error: 54.39611906810181
average Variance score: 0.09465221113358974

INDUS
average coefficient: [-0.4939411]
average Mean square error: 49.48512094094062
average Variance score: 0.18682032412742147

CHAS
average coefficient: [5.21430954]
average Mean square error: 62.543405363108356
average Variance score: 0.0009537141618401858

NOX
average coefficient: [-25.92357535]
average Mean square error: 51.385750301216845
average Variance score: 0.14928451580257093

RM
average coefficient: [6.93555817]
average Mean square error: 31.851796349944685
average Variance score: 0.3888173939733907

AGE
average coefficient: [-0.09254187]
average Mean square error: 53.39766273352915
average Variance score: 0.11734320570289072

DIS
average coefficient: [0.8462227]
average Mean square error: 64.04345845708131
average Variance score: 0.035066801510610535

RAD
average coefficient: [-0.309166]
average Mean square error: 51.70791777901924
average Variance score: 0.11696395198032873

TAX
average coefficient: [-0.01987053]
average Mean square error: 46.082576075551806
average Variance score: 0.16074109899182298

PTRATIO
average coefficient: [-1.65937114]
average Mean square error: 50.8349736127547
average Variance score: 0.1780231260298484

B
average coefficient: [0.02594733]
average Mean square error: 53.07488819891271
average Variance score: 0.09174987640537803

LSTAT
average coefficient: [-0.7236659]
average Mean square error: 29.303034065881924
average Variance score: 0.4321188357514304
```

3 To compare the model performance, provide 1. mean square error vs features and 2. variance score vs features

[6]:

```python
# provide the following plots: 1. mean square error vs features 2. variance score vs features
avg_coefs.insert(0, avg_coefs_all)
avg_MSEs.insert(0, avg_MSEs_all)
avg_vss.insert(0, avg_vss_all)

features = list(range(14))
plt.figure(figsize=(4, 3))
ax = plt.axes()
ax.scatter(features, avg_MSEs, color='blue')
ax.set_xlabel("features")
ax.set_ylabel("Mean square error")
plt.show()

plt.figure(figsize=(4, 3))
ax = plt.axes()
ax.scatter(features, avg_vss, color='green')
ax.set_xlabel("features")
ax.set_ylabel("variance score")
plt.show()
```





4 Analysis

1. Based upon the linear models you generated, which feature appears to be most predictive for the target feature?

The LSTAT (% lower status of the population) appears to be most predictive for the target feature.

2. Suppose you need to select two features for a linear regression model to predict the target feature. Which two features would you select? Why?

I would select LSTAT (% lower status of the population) and RM (average number of rooms per dwelling) because they have the lowest mean square errors and highest variance scores.

3. Examine all the plots and numbers you have, do you have any comments on them? Do you find any surprising trends? Do you have any idea about what might be causing this surprising trend in the data? This is a descriptive question meant to encourage you to interpret your results and express yourself.

I find that the weighted distance to 5 Boston employment centres is negatively correlated to the housing price is surprising. The causing might be the employment centre has the most weight is in the area with low housing price.