(1) derive formulas to solve liner regression problems.

$$E(w_1, w_0 | \mathcal{X}) = \frac{1}{N}\sum_{t=1}^{N}[r^t-(w_1x^t+w_0)]^2$$

平均均方誤差

推導 $w_1 = \frac{\sum_{t=1}^{N} x^t r^t - \bar{x}-\bar{r}N}{\sum_{t=1}^{N}(x^t)^2 - N\bar{x}^2} = \frac{\overline{xr}-\bar{x}\bar{r}}{\overline{x^2}-\bar{x}^2}$ and $w_0 = \bar{r} - w_1\bar{x}$

⇓ 證明 1

設 $a = r^t$, $b = (w_1x^t+w_0)$

$$g = \frac{1}{N}\sum(a-b)^2$$

$$\Rightarrow \frac{1}{N}\sum a - 2ab + b^2$$

↙ 代入

$$\Rightarrow \frac{1}{N}\sum r^{2t} - (2(r^t)(w_1x^t+w_0)) + (w_1x^t+w_0)^2$$

$$\Rightarrow \frac{1}{N}\sum r^{2t} - 2r^tw_1x^t - 2r^tw_0 + w_1^2x^{2t} + 2w_1x^tw_0 + w_0^2 \quad \cdots \quad (1)$$

$$\Rightarrow \frac{g}{dw_0} = 0 \Rightarrow \frac{g}{dw_0} = \frac{1}{N}\sum 0 - 0 - 2r^t + 0 + 2w_1x^t + 2w_0$$

$$\Rightarrow \frac{g}{dw_0} = \frac{1}{N}\sum -2r^t + 2w_1x^t + 2w_0$$

設 $\frac{g}{dw_0}$ 的極值 $= 0$

$$\frac{g}{dw_0} = \frac{1}{N}\sum \cancel{2}(-r^t + w_1x^t + w_0) = 0$$

$$\Rightarrow \frac{1}{N}\sum -r^t + w_1x^t + w_0 = 0$$

$$\Rightarrow \frac{1}{N}\sum -r^t + w_1x^t + \frac{1}{N}\cdot \underbrace{\boxed{Nw_0}}_{\sum_{t=1}^{N} w_0} = 0$$

$$\Rightarrow w_0 = \frac{1}{N}\sum -(-r^t + w_1x^t)$$

$$\Rightarrow w_0 = \frac{1}{N}\sum r^t - w_1x^t \Rightarrow \underline{w_0 = \bar{r} - w_1\bar{x}} \ \#$$

再對(1)做 $\frac{g}{dw_1}$ ⇐ 證明2

$$\Rightarrow \frac{g}{dw_1} = 0 - 2r^t x^t - 0 + 2x^{2t} w_1 + 2w_0 x^t + 0$$

$$\Rightarrow -2r^t x^t + 2x^{2t} w_1 + 2w_0 x^t$$

設極值 $=0$

$$\frac{g}{dw_1} = \frac{1}{N}\sum 2(-r^t x^t + x^{2t} w_1 + w_0 x^t) = 0$$

$$\Rightarrow \frac{1}{N}\sum w_0 x^t = \frac{1}{N}\sum r^t x^t - x^{2t} w_1 \quad = 0$$

$$\Rightarrow w_0 \cdot \frac{1}{N}\sum x^t = \frac{1}{N}\sum r^t x^t + \frac{1}{N}\sum -x^{2t} w_1$$

$$\Rightarrow w_0 \cdot \overline{x} = \overline{rx} + w_1 \cdot \frac{1}{N}\sum -x^{2t}$$

$$\Rightarrow w_0 \cdot \overline{x} = \overline{rx} - w_1 \cdot \overline{x^2}$$

又 $w_0 = \overline{r} - w_1 \overline{x}$ 代入

$$\Rightarrow (\overline{r} - w_1 \overline{x})\overline{x} = \overline{rx} - w_1 \overline{x^2}$$

$$\Rightarrow \overline{x}\,\overline{r} - w_1 (\overline{x})^2 = \overline{rx} - w_1 \overline{x^2}$$

$$\Rightarrow w_1 \overline{x^2} - w_1 (\overline{x})^2 = \overline{rx} - \overline{x}\,\overline{r}$$

$$\Rightarrow w_1 (\overline{x^2} - (\overline{x})^2) = \overline{rx} - \overline{x}\,\overline{r}$$

$$\Rightarrow w_1 = \frac{\overline{rx} - \overline{x}\,\overline{r}}{\overline{x^2} - (\overline{x})^2}$$

#

(2-1) 利用 linear regression 預測攝影機拍攝之灰階值對應之實際灰度值

預測 $x = 38162,\ f(x) = ?$
$x = 21537,\ f(x) = ?$
$x = 50000,\ f(x) = ?$

資料點編號	座標（383,255）的灰階值	BM-7A 所量測輝度值
1	17703	1009
2	19079	1102
3	20620	1202
4	22181	1310
5	23632	1399
6	24911	1497
7	26371	1598
8	27986	1707
9	29467	1806
10	30960	1906
11	32226	2001
12	33672	2103
13	35146	2209
14	36572	2309
15	37929	2402
16	39274	2500
17	40610	2601
18	42063	2703
19	43332	2803
20	44696	2902
21	46154	3008

$f(x) = W_1 x + W_0$，由證明 1 and 2 得知

$$W_0 = \bar{r} - W_1 \bar{x}, \quad W_1 = \frac{\overline{rx} - \bar{x}\,\bar{r}}{\overline{x^2} - (\bar{x})^2}$$

令 x = 拍攝之灰階值，r = 實際之灰度值

$\therefore\ x = \{17703, 19079 \ldots\},\ r = \{1009, 1102 \ldots\}$

計算 $\bar{r} = \frac{1}{N}\sum_{t=1}^{N} r \doteqdot 2003.66$

$\bar{x} = \frac{1}{N}\sum_{t=1}^{N} x \doteqdot 32123.04$

可得 $W_1 \doteqdot 0.07$

$W_0 \doteqdot -253.22$

$$\therefore\ \boxed{\begin{aligned} f(38162) &= 2427.95 \\ f(21537) &= 1259.92 \\ f(50000) &= 3259.66 \end{aligned}}$$

(取到小數點第二位)

#

對應 python code

```python
import numpy as np

def get_w1(x, r):  2 usages
    x_bar = np.mean(x)
    r_bar = np.mean(r)
    return (np.mean(x * r) - x_bar * r_bar) / (np.mean(x ** 2) - x_bar ** 2)
```

→ 各自計算 $\bar{x}, \bar{r}$ (利用 np.mean 算平均)

↳ 對應到 $w_1 = \dfrac{\overline{rx} - \bar{x}\bar{r}}{\overline{x^2} - (\bar{x})^2}$

```python
def get_w0(x, r):  1 usage
    return np.mean(r) - get_w1(x, r) * np.mean(x)
```

↳ 對應到 $W_0 = \bar{r} - W_1\bar{x}$

計算 $f(x) = w_1x + w_0$

```python
def f_of_x(x, w1, w0):  1 usage
    return w1 * x + w0

if __name__ == "__main__":
    x = np.array([
        17703, 19079, 20620, 22181, 23632, 24911, 26371, 27986, 29467, 30960,
        32226, 33672, 35146, 36572, 37929,39274, 40610, 42063, 43332, 44696,
        46154
    ])
```

⇒ 灰階值 (拍攝)

```python
    r = np.array([
        1009, 1102, 1202, 1310, 1399, 1497, 1598, 1707, 1806, 1906,
        2001, 2103, 2209, 2309, 2402,2500, 2601, 2703, 2803, 2902, 3008
    ])
```

→ 灰度值 (實際)

```python
    w1 = get_w1(x, r)
    w0 = get_w0(x, r)

    print(f"w1 = {w1:.2f}")
    print(f"w0 = {w0:.2f}")
```

→ show calculated W_0 and W_1 (取小數2位)

```python
    for x in [38162, 21537, 50000]:
        print(f"f({x}) = {f_of_x(x, w1, w0):.2f}")
```

$[38162, 21537, 50000] \to x$

→ 取小數 2 位

```
w1 = 0.07
w0 = -253.22
f(38162) = 2427.95
f(21537) = 1259.92
f(50000) = 3259.66
```

← 執行結果

(2-2) 找出 MEDV 與其餘 13個特徵的各自 linear regressio 模型

各特徵值實際意義

特徵 (Feature)	中文描述 (Chinese Description)
CRIM	鄉鎮人均犯罪率
ZN	規劃為超過 25,000 平方英尺地塊的住宅用地比例
INDUS	每個鄉鎮非零售商業用地的比例
CHAS	查爾斯河虛擬變數（若地塊毗鄰河流則為 1；否則為 0）
NOX	一氧化氮濃度（百萬分之幾）
RM	每棟住宅的平均房間數
AGE	1940 年以前建造的自住單位比例
DIS	到波士頓五個就業中心的加權距離
RAD	輻射狀公路可達性指數
TAX	每 $10,000 全值房產稅率
PTRATIO	鄉鎮師生比例
B	$1000(\mathrm{Bk}-0.63)^2$（其中 Bk 為該鎮黑人比例）
LSTAT	人口中地位較低的百分比
MEDV	自住房屋的中位價值（以千美元計）

令 $x = \text{MEDV}$, $y = [\text{CRIM} \sim \text{B}]$ ⇒共 13組

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0.00632	18.00	2.310	0	0.5380	6.5750	65.20	4.0900	1	296.0	15.30	396.90	4.98	24.00
0.02731	0.00	7.070	0	0.4690	6.4210	78.90	4.9671	2	242.0	17.80	396.90	9.14	21.60
0.02729	0.00	7.070	0	0.4690	7.1850	61.10	4.9671	2	242.0	17.80	392.83	4.03	34.70
0.03237	0.00	2.180	0	0.4580	6.9980	45.80	6.0622	3	222.0	18.70	394.63	2.94	33.40
0.06905	0.00	2.180	0	0.4580	7.1470	54.20	6.0622	3	222.0	18.70	396.90	5.33	36.20
0.02985	0.00	2.180	0	0.4580	6.4300	58.70	6.0622	3	222.0	18.70	394.12	5.21	28.70

資料共 506 筆 由證明 1 and 2 可計算出 13組各自

linear regression model $f(x) = W_1 x + W_0$

利用程式計算出 13 組模行

繼續利用 (1-1) 所定義的 functions

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

def get_w1(x, r):  2 usages   new *
    x_bar = np.mean(x)
    r_bar = np.mean(r)
    return (np.mean(x * r) - x_bar * r_bar) / (np.mean(x ** 2) - x_bar ** 2)

def get_w0(x, r):  1 usage   new *
    return np.mean(r) - get_w1(x, r) * np.mean(x)

def f_of_x(x, w1, w0):  1 usage   new *
    return w1 * x + w0
```

← 14個特徵

```python
if __name__ == "__main__":
    features = [
        "per capita crime rate by town(CRIM)",
        "proportion of residential land zoned for lots over 25,000 sq.ft(ZN)",
        "proportion of non-retail business acres per town(INDUS)",
        "Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)[CHAS]",
        "nitric oxides concentration (parts per 10 million)[NOX]",
        "average number of rooms per dwelling(RM)",
        "proportion of owner-occupied units built prior to 1940(AGE)",
        "weighted distances to five Boston employment centres(DIS)",
        "index of accessibility to radial highways(RAD)",
        "full-value property-tax rate per $10,000(TAX)",
        "pupil-teacher ratio by town(PTRATIO)",
        "1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town[B]",
        "% lower status of the population(LSTAT)",
        "Median value of owner-occupied homes in $1000's(MEDV)"
    ]
    df = pd.read_csv( filepath_or_buffer: "housing.data", sep='\s+', names=features)
```

add features to title

↳以空格分割 data

→排除 MEDV

```python
    for feature in features[:-1]:
        print(f"Features: {feature}")
        x = df[feature].values
        r = df["Median value of owner-occupied homes in $1000's(MEDV)"].values
        w1 = get_w1(x, r)
        w0 = get_w0(x, r)
        print(f"w1: {w1:.2f}, w0: {w0:.2f}")
        print(f"f(x)={w1:.2f}x + {w0:.2f}")
        fig = plt.figure()
        plt.scatter(x, r, color='lightpink', edgecolors='darkolivegreen')
        plt.plot( *args: x, f_of_x(x, w1, w0), 'r-')
        plt.xlabel(feature)
        plt.ylabel("Median value of owner-occupied homes in $1000's(MEDV)")
        plt.savefig(f"img/MEDV-{feature}.png")
        print("---")
```

→[CRIM ~ LSTA] 特徵 value ⇒ x

⇒ MEDV ⇓ r

→計算 W_0 and W_1

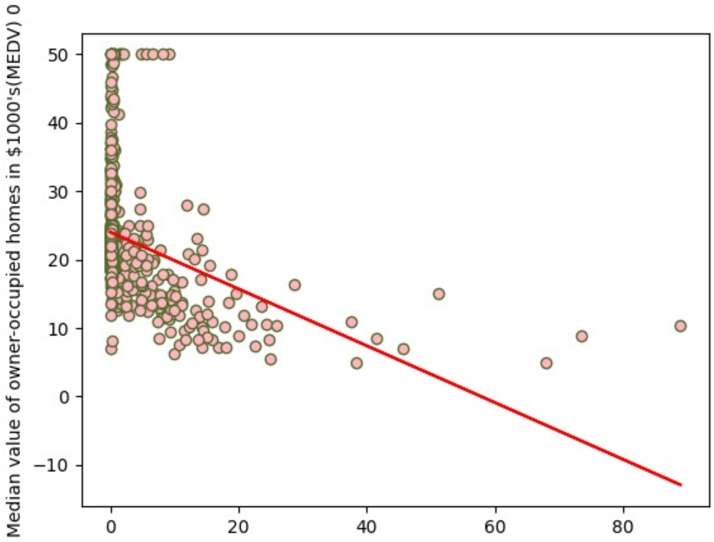
→作圖

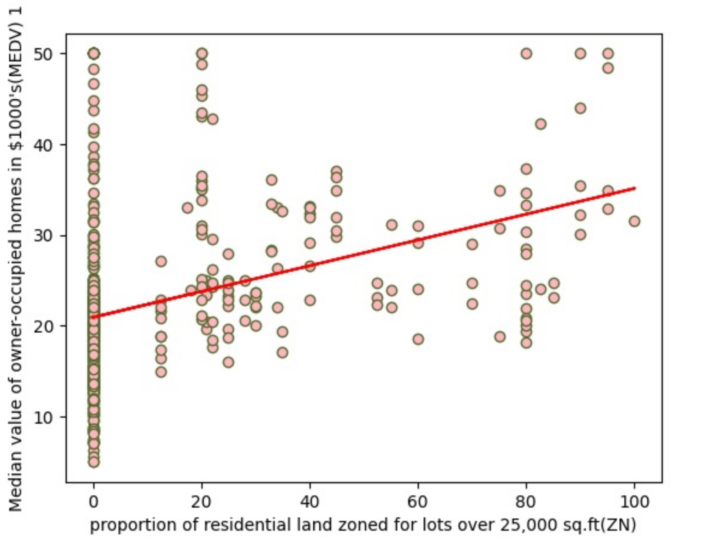
↳由 linear regression model 得出 predict 結果 f(x)=y

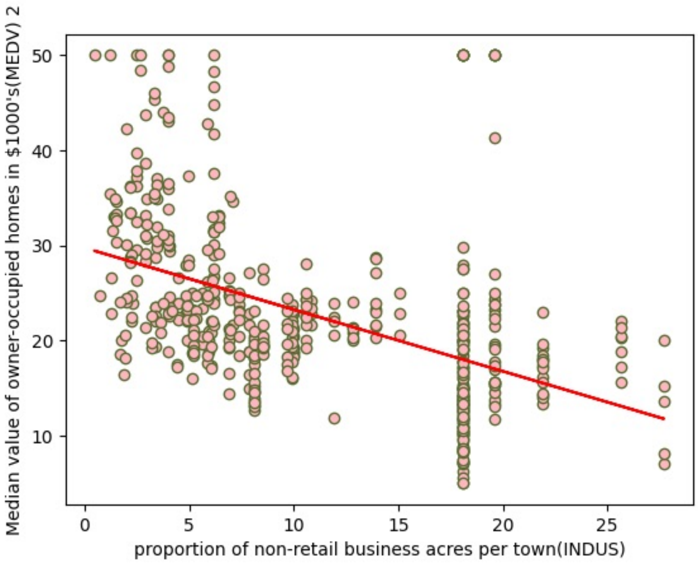
以下為13組特徵值所建立出的 linear regression model (展示取至小數2位) #

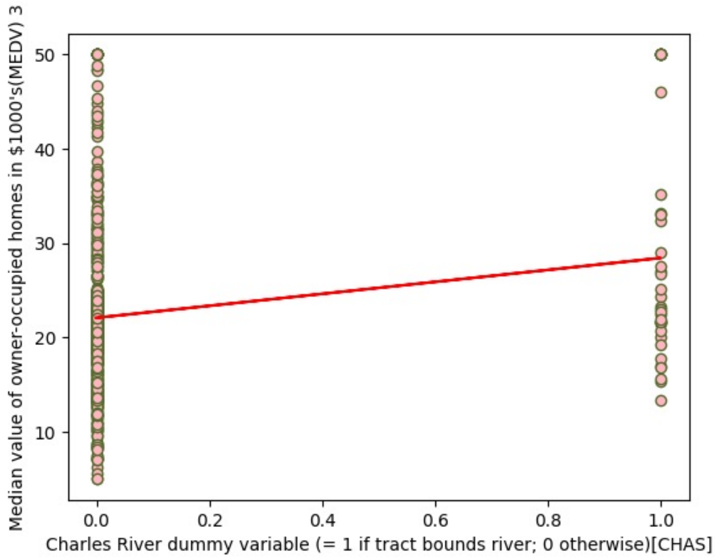
```
Features: per capita crime rate by town(CRIM)
w1: -0.42, w0: 24.03
f(x)=-0.42x + 24.03
---
Features: proportion of residential land zoned for lots over 25,000 sq.ft(ZN)
w1: 0.14, w0: 20.92
f(x)=0.14x + 20.92
---
Features: proportion of non-retail business acres per town(INDUS)
w1: -0.65, w0: 29.75
f(x)=-0.65x + 29.75
---
Features: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)[CHAS]
w1: 6.35, w0: 22.09
f(x)=6.35x + 22.09
---
Features: nitric oxides concentration (parts per 10 million)[NOX]
w1: -33.92, w0: 41.35
f(x)=-33.92x + 41.35
---
Features: average number of rooms per dwelling(RM)
w1: 9.10, w0: -34.67
f(x)=9.10x + -34.67
---
Features: proportion of owner-occupied units built prior to 1940(AGE)
w1: -0.12, w0: 30.98
f(x)=-0.12x + 30.98
---
Features: weighted distances to five Boston employment centres(DIS)
w1: 1.09, w0: 18.39
f(x)=1.09x + 18.39
Features: index of accessibility to radial highways(RAD)
w1: -0.40, w0: 26.38
f(x)=-0.40x + 26.38
---
Features: full-value property-tax rate per $10,000(TAX)
w1: -0.03, w0: 32.97
f(x)=-0.03x + 32.97
---
Features: pupil-teacher ratio by town(PTRATIO)
w1: -2.16, w0: 62.34
f(x)=-2.16x + 62.34
---
Features: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town[B]
w1: 0.03, w0: 10.55
f(x)=0.03x + 10.55
---
Features: % lower status of the population(LSTAT)
w1: -0.95, w0: 34.55
f(x)=-0.95x + 34.55
---
```

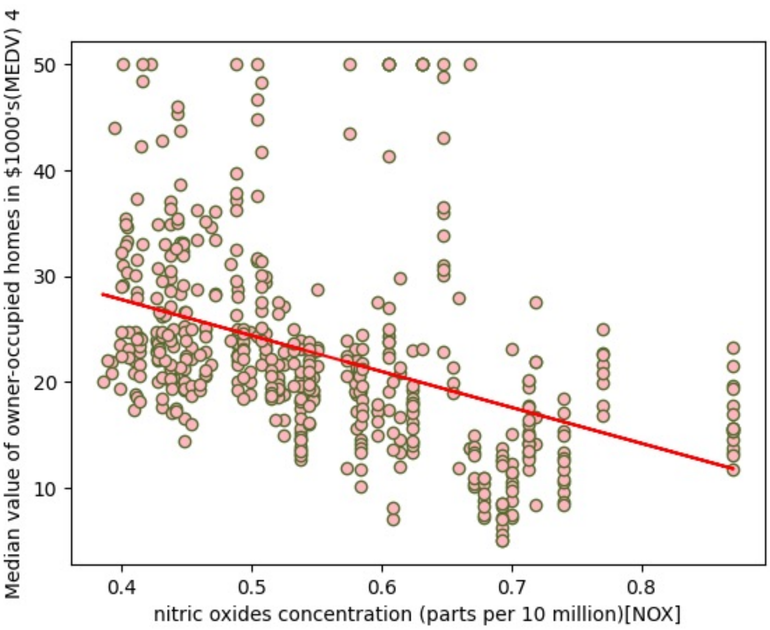
以下為各 MEDV 對各13個特徵值的 linear regression model圖 #

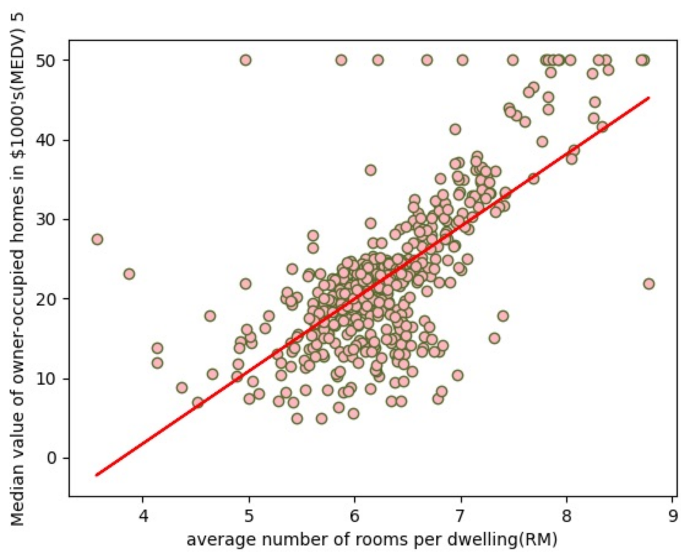


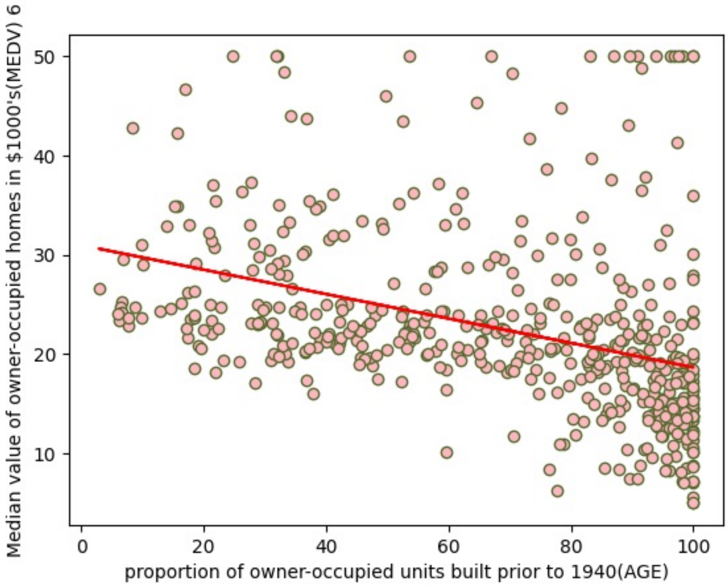


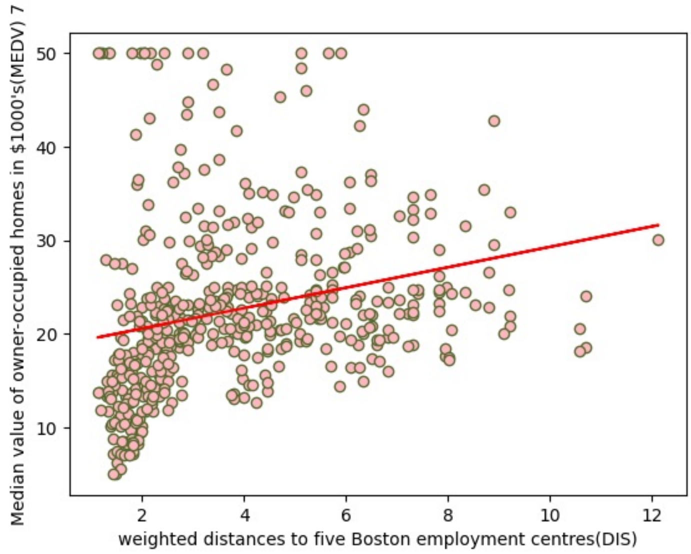


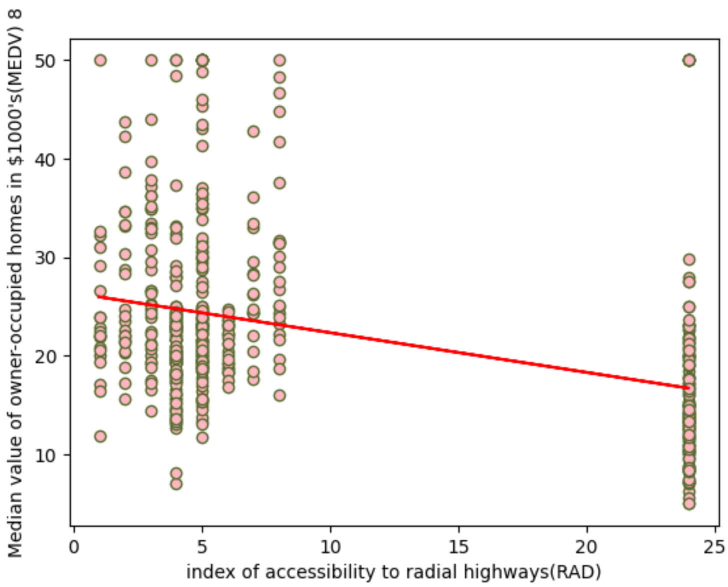


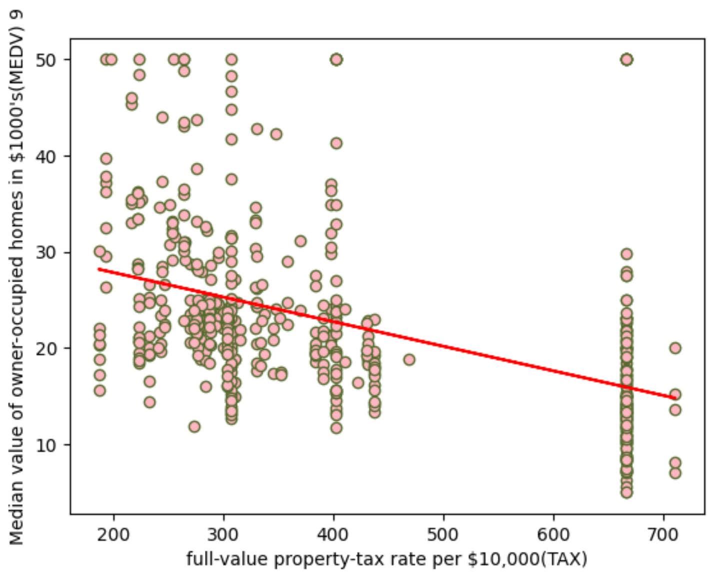


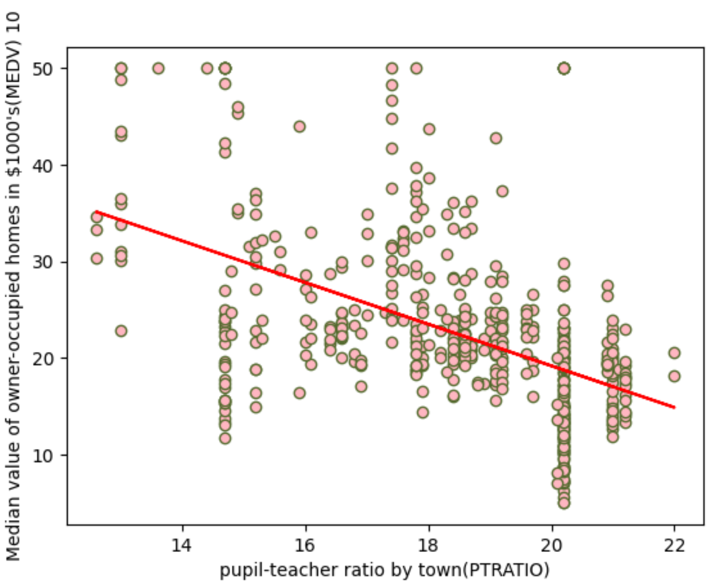


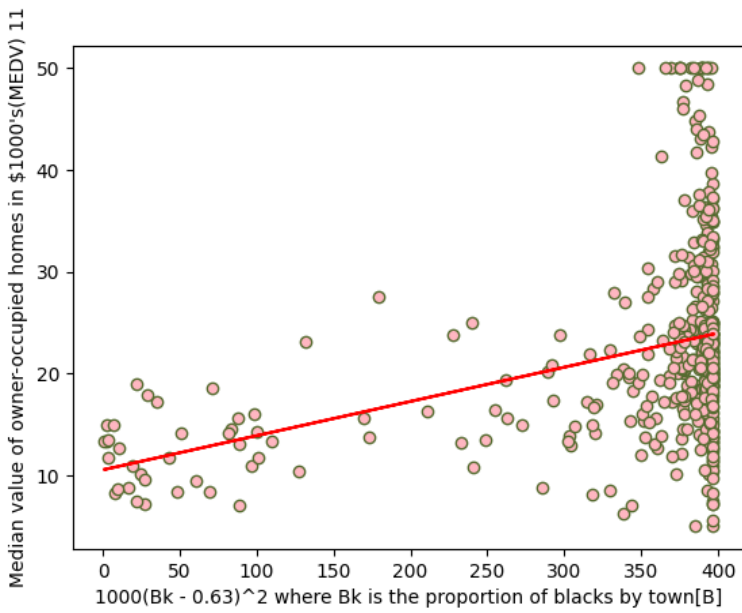

Median value of owner-occupied homes in $1000's(MEDV) 6
10 20 30 40 50
0 20 40 60 80 100
proportion of owner-occupied units built prior to 1940(AGE)

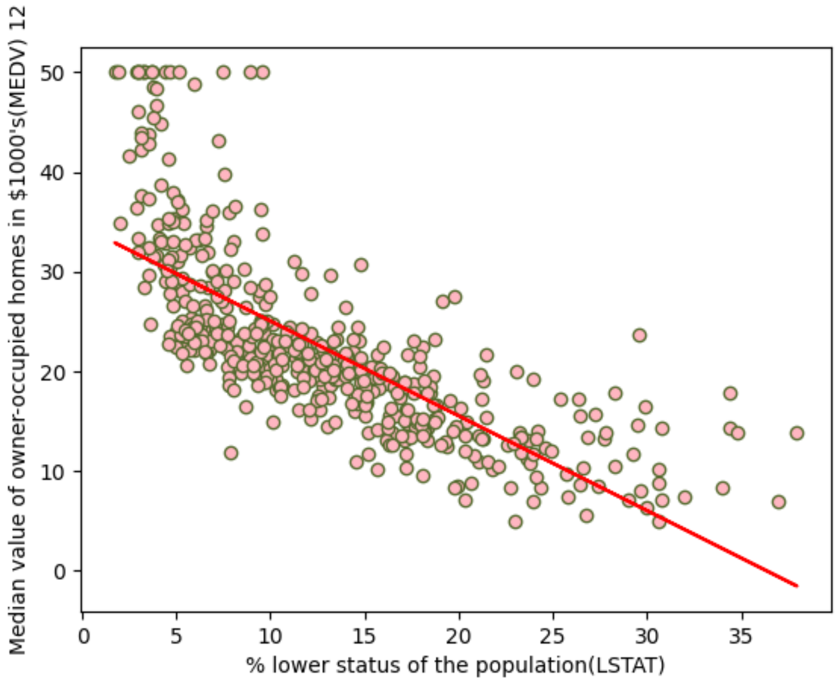

Median value of owner-occupied homes in $1000's(MEDV) 7
10 20 30 40 50
2 4 6 8 10 12
weighted distances to five Boston employment centres(DIS)


Median value of owner-occupied homes in $1000's(MEDV) 8
10 20 30 40 50
0 5 10 15 20 25
index of accessibility to radial highways(RAD)


Median value of owner-occupied homes in $1000's(MEDV) 9
10 20 30 40 50
200 300 400 500 600 700
full-value property-tax rate per $10,000(TAX)


Median value of owner-occupied homes in $1000's(MEDV) 10
10 20 30 40 50
14 16 18 20 22
pupil-teacher ratio by town(PTRATIO)


Median value of owner-occupied homes in $1000's(MEDV) 11
10 20 30 40 50
0 50 100 150 200 250 300 350 400
1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town[B]


Median value of owner-occupied homes in $1000's(MEDV) 12
% lower status of the population(LSTAT)
0
5
10
15
20
25
30
35
40
50