Homework 3

Yuan Li, N19728558, NetID: yl6606

Part I: Written Exercises

1. (a) Answer:

Set the variable for the cancer volume as x_1, the variable for the patient's age as x_2. Set x_3 which is assigned the value 1 if the cancer is Type I, and x_4 which is assigned the value 1 if the cancer is Type II. So models 1 & 2 should be:

$$\hat{y}_1 = m_1 * x_1 + d_1$$
$$\hat{y}_2 = m_2 * x_1 + n_2 * x_2 + d_2$$

Where m_i, n_i, di are parameters, i for the number of model i.

1. (b) Answer:

Given the variable which we mentioned in 1.(a), we can get the model 3 as:

$$\hat{y}_3 = (m_3 + p_3 * x_3 + q_3 * x_4) * x_1 + n_3 * x_2 + d_3$$

1. (c) Answer:

The number of parameters in model 1: 2.
The number of parameters in model 2: 3.
Model 3 is the most complex. Because there are 4 variables and 5 parameters.

1. (d) Answer:

The first three rows of the matrix X for model 1:

$$\begin{bmatrix} 0.7 \\ 1.3 \\ 1.6 \end{bmatrix}$$

The first three rows of the matrix X for model 2:

$$\begin{bmatrix} 0.7 & 55 \\ 1.3 & 65 \\ 1.6 & 70 \end{bmatrix}$$

The first three rows of the matrix X for model 3:

$$\begin{bmatrix} 0.7 & 55 & 1 & 0 \\ 1.3 & 65 & 0 & 1 \\ 1.6 & 70 & 0 & 1 \end{bmatrix}$$

1. (e) Answer:

We should choose model 2.

Because the validation MSE of model 2 is smallest, which means it can get the best accuracy. While the training MSE of model 3 is smallest, but its validation MSE is bigger than model 3, so it shouldn't be selected.

2. Answer:

We can assume that the crop yields are a linear model which depends on the amount of rainfall, the amount of fertilizer, the average temperature, and the number of sunny days. Set crop yields as y, and other 4 variables as x_1, x_2, x_3, x_4, so the model can be:

$$\begin{aligned}\hat{y} &= w_1 * x_1 + w_2 * x_2 + w_3 * x_3 + w_4 * x_4 + w_5 \\ &= W * X\end{aligned}$$

And it should be a regression problem.

3. (a) Answer:

the matrix X should be:

$$X = \begin{bmatrix} 28540 \\ 40133 \\ 39900 \\ 0 \\ 0 \\ 42050 \\ 43220 \\ 39565 \\ 40400 \\ 54605 \end{bmatrix}$$

3. (b) Answer:

Based on the closed-form formula given in class, we can get the $E_{in}(w)$ as:

$$\begin{aligned}E_{in}(w) &= \frac{1}{10}\sum_{i=1}^{10}(\hat{y}_i + y_i)^2 \\ &= \frac{1}{N}||Xw - y||_2^2 \\ \nabla E_{in}(w) &= \begin{bmatrix} \frac{\partial E_{in}(w)}{\partial w_0} \\ \frac{\partial E_{in}(w)}{\partial w_1} \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ 0 \end{bmatrix}\end{aligned}$$

Use python can easily get the result. The code is as follow:

```
import numpy as np
x = np.array([137, 135, 127, 122, 120, 118, 118, 117, 117, 114])
y = np.array([28540, 40133, 39900, 0, 0, 42050, 43220, 39565, 40400, 54506])

A = np.vstack([x, np.ones(len(x))]).T
m, c = np.linalg.lstsq(A, y, rcond=None)[0]
```

Python code

So the equation is $g(x) = -338.5366 * x_1 + 74302.1369$

3. (c) Answer:

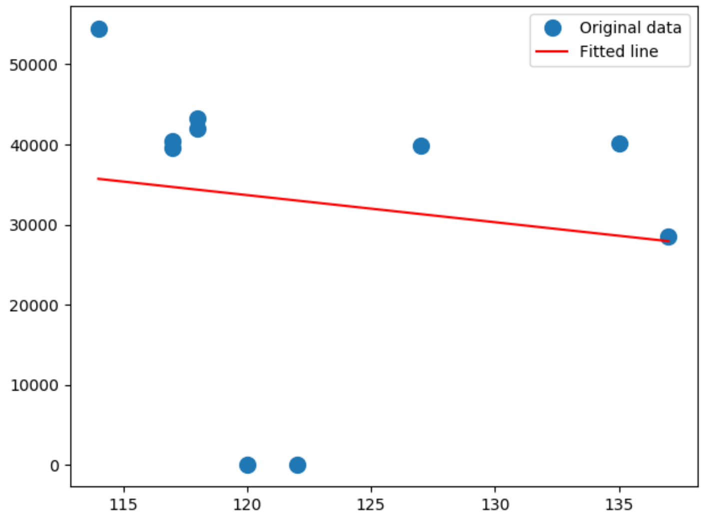
We can use Python code to get the picture:

```
import matplotlib.pyplot as plt
_ = plt.plot(x, y, 'o', label='Original data', markersize=10)
_ = plt.plot(x, m*x + c, 'r', label='Fitted line')
```

```
_ = plt.legend()
plt.show()
```

Python code

The picture should be:



3. (d) Answer:

$$\begin{aligned} R^2 &= 1 - \frac{RSS}{TSS} \\ &= 1 - \frac{\sum_{i=1}^{10}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{10}(y_i - \overline{y})^2} \\ &= 0.021337 \end{aligned}$$

3. (e) Answer:

Put all the parameters into equation $g(x)$:

$$40000 = -338.5366 * x + 74302.1369$$

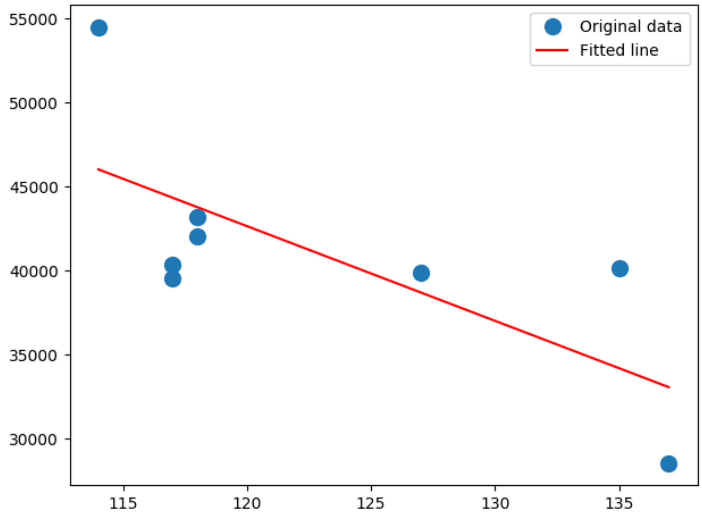
So the mid-career salary should be: $x \approx 101$

3. (f) Answer:

After removing all the outliers, the matrix X should be:

$$X = \begin{bmatrix} 28540 \\ 40133 \\ 39900 \\ 42050 \\ 43220 \\ 39565 \\ 40400 \\ 54605 \end{bmatrix}$$

So the $g(x)$ becomes $g(x) = -563.4511 * x + 110273.3075$. Picture becomes:



And $R^2 = 0.509764$, the mid-career salary is still: $x \approx 101$

4. (a) Answer:

Because we want to force the predicted value $\hat{y} = 0$ when $x = 0$, w should be $\begin{bmatrix} w_1 \end{bmatrix}$
So the cost function should be:

$$\begin{aligned} E_{in}(w) &= \sum(\hat{y}_i - y_i) \\ &= ||w_1 X - y||_2^2 \\ &= w_1^2 X^T X - 2w_1 X^T y + y^T y \end{aligned}$$

4. (b) Answer:

$$\nabla E_{in}(w) = 2w_1 X^T X - 2X^T y$$

In order to get the w that minimizes the RSS, so:

$$\begin{aligned} 2w_1 X^T X - 2X^T y = 0 \\ w_1 X^T X = X^T y \\ w = \begin{bmatrix} w_1 \end{bmatrix} = (X^T X)^{-1} X^T y \end{aligned}$$

5. Answer:

We can set Ω as the matrix for the weigh. So the problem becomes: how to find

$$\begin{aligned} \hat{w}_{WLS} &= argmin(y - Xw)^T \Omega (y - Xw) \\ \frac{\partial \hat{w}_{WLS}}{\partial w} &= \frac{\partial}{\partial w}(y - Xw)^T \Omega (y - Xw) \\ &= \frac{\partial}{\partial w}(y^T \Omega y - y^T \Omega X w - w^T X^T \Omega y + w^T X^T \Omega X w) \\ &= X^T \Omega X w - X^T \Omega y = 0 \end{aligned}$$

$$\begin{aligned} X^T \Omega X w &= X^T \Omega y \\ w &= (X^T \Omega X)^{-1} X^T \Omega y \end{aligned}$$

So the result should be $w = (X^T \Omega X)^{-1} X^T \Omega y$