1-a X_1: cancer volume

X_2: patient's age

X_3: cancer type

Model 1: $\hat{y_1} = w_1 X_1 + w_0 = [w_0\ w_1]\begin{bmatrix} 1 \\ X_1 \end{bmatrix}$

Model 2: $\hat{y_2} = w_2 X_2 + w_1 X_1 + w_0 = [w_0\ w_1\ w_2]\begin{bmatrix} 1 \\ X_1 \\ X_2 \end{bmatrix}$

1-b $\hat{y_3} = w_2 X_2 + w_1 X_1 \cdot X_3 + w_1' X_1 (1 - X_3) + w_0$

$= w_2 X_2 + w_1 (X_1 X_3) + w_1' X_1 - w_1' (X_1 X_3) + w_0$

$\cancel{= w_2 X_2 + (w_1 - w_1') X_1 X_3 + w_1' X_1 + w_0}$

$= w_2 X_2 + w_1 (X_1 X_3) + w_1' (X_1 - X_1 X_3) + w_0$

$= [w_0\ w_1'\ w_1\ w_2]\begin{bmatrix} 1 \\ X_1 - X_1 X_3 \\ X_1 X_3 \\ X_2 \end{bmatrix}$

1-c model 1 has 2 parameters

model 2 has 3 parameters

model 3 has 4 parameters $\rightarrow$ the most complex model

1-d: model 1:

$$\begin{bmatrix} 1 & 0.7 \\ 1 & 1.3 \\ 1 & 1.6 \end{bmatrix}$$

model 2

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

model 3

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

1-e: model 2, for 2 reasons:

(1) model 2 has less MSE than model 1

(2) model 3's ~~has~~ training error is much lower than its validation error, which may implies overfitting.

---

2. feature selection:

x_1 — rainfall

x_2 — fertilizer

x_3 — average temperature

x_4 — number of sunny days

result

y — crop yields

linear model: $y = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4$

3-a

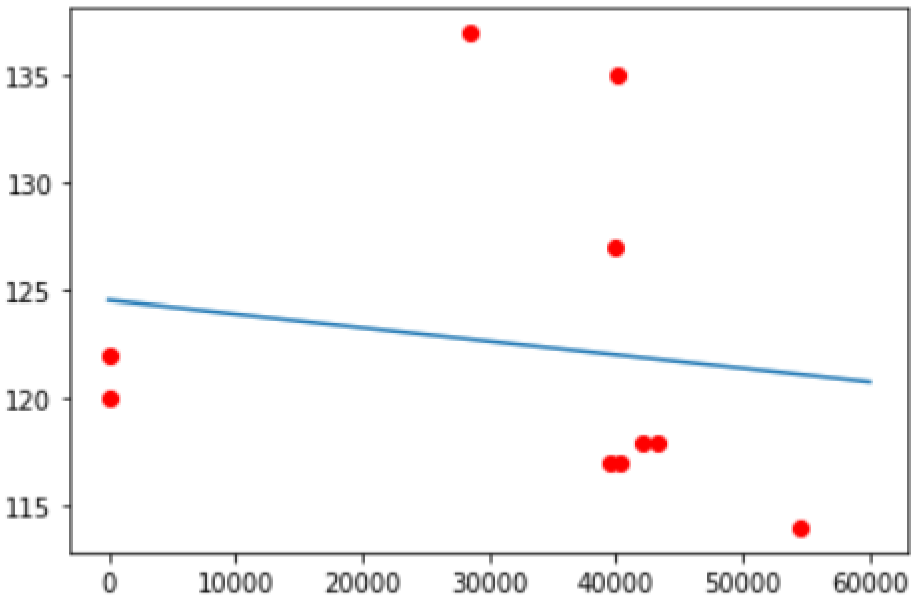
$$X = \begin{bmatrix} 1 & 28540 \\ 1 & 40133 \\ 1 & 39900 \\ 1 & 0 \\ 1 & 0 \\ 1 & 42050 \\ 1 & 43220 \\ 1 & 39565 \\ 1 & 40400 \\ 1 & 54506 \end{bmatrix}$$

3-b:

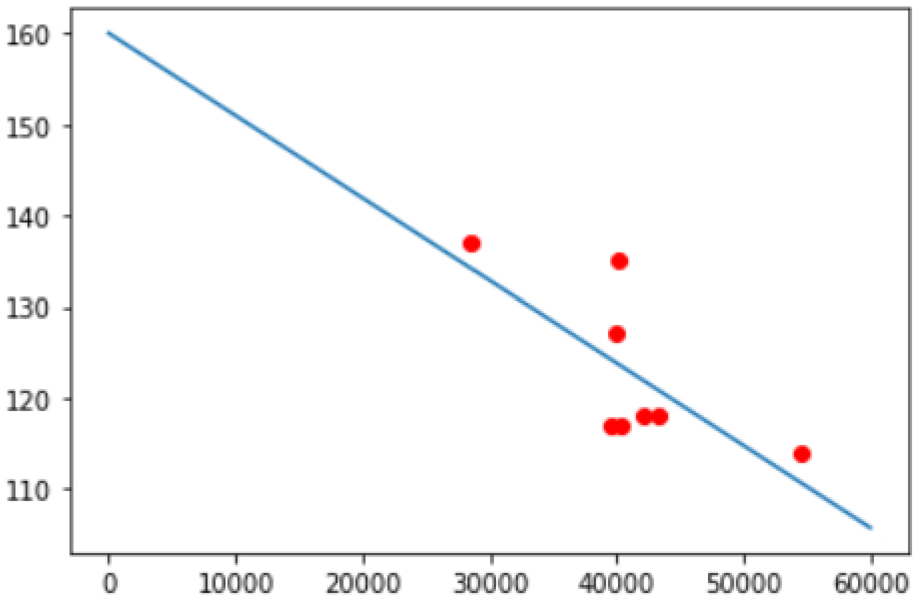
$y = [w_0 \; w_1] \cdot X$

$w_0 \approx 124$

$w_1 \approx -6.3 \times 10^{-5}$



3-c before removing the noise



3-f-c after removing the noise

3-d $R^2 = \frac{ESS}{TSS} = \frac{12.087}{566.5} = 0.02$

3-e $X = 40000$

$y = w_0 + w_1 X = 122.05$

3-f-b: $\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 160 \\ -9\times10^{-4} \end{bmatrix}$

$y = 160 - (9\times10^{-4}) X$

3-f-d: $R^2 = \cancel{0.51} \frac{284.89}{558.875} = 0.51$

3-f-e: 123.815

4-a $RSS(w) = \sum_{i=1}^{N} (\hat{y_i} - y_i)^2$

$= \sum_{i=1}^{N} (wx_i - y_i)^2$

$= \| Xw - y \|_2^2$

$= (Xw-y)^T (Xw-y)$

$= \boxed{w^T X^T X w - 2 w X^T y + y^T y}$

4-b $\nabla_w = 2X^TXw - 2X^Ty = 0$

$X^TXw = X^Ty$ $\boxed{w = \frac{X^Ty}{X^TX}}$

4-b

X is N*1, y is N*1, therefore X.T.dot(X) = a value, X.T.dot(y) is also a value.

5-a $p(y|w,x) = p(\varepsilon)$

$\varepsilon^{(i)} = y^{(i)} - w^T x^{(i)}$

$W_{ML} = \underset{w}{\arg\max} \prod_{i=1}^{N} p(\varepsilon^{(i)})$

$L(w) = \prod_{i=1}^{N} p(\varepsilon^{(i)})$

$\log(L(w)) = \sum_{i=1}^{N} \log \frac{1}{2b} e^{\frac{-|\varepsilon^{(i)}|}{b}}$

$= N\log\frac{1}{2b} + \sum_{i=1}^{N} \frac{1}{-b} |\varepsilon^{(i)}|$

same for any w

$\log(L(w)) \Rightarrow g(w) = \frac{1}{-b} \sum_{i=1}^{N} |\varepsilon^{(i)}|$

$= -\frac{1}{b} \sum_{i=1}^{N} |y^{(i)} - w^T x^{(i)}|$ #

~~5-b~~ 6 $g(w) = (y - Xw)^T \Omega (y - Xw)$

$= (y^T - w^T X^T)\Omega(y - Xw)$

$= (y^T\Omega - w^T X^T \Omega)(y - Xw)$

$x^T\Omega y = y^T \Omega x$

$= y^T\Omega y \boxed{- w^T x^T \Omega y - y^T \Omega x w} + w^T x^T \Omega x w$

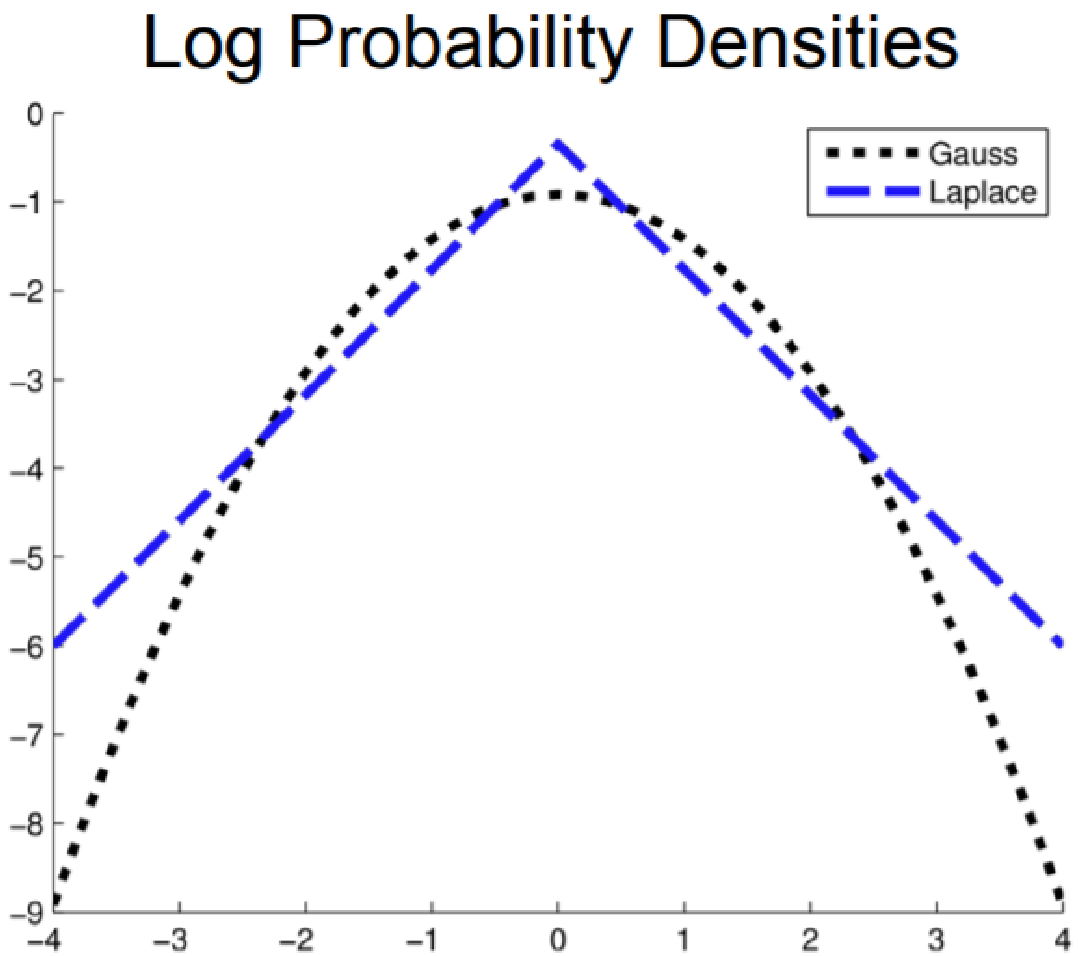
$= w^T(x^T\Omega x)w - 2(y^T\Omega^T x)w + y^T \Omega y$

$\frac{\partial g(w)}{\partial w} = 2x^T\Omega x w - 2y^T\Omega^T x$

$x^T\Omega x w = y^T \Omega^T x$

$w = (x^T\Omega x)^{-1} y^T \Omega^T x$

5-b



The probability of noise data appearing in Laplace Distribution is higher than Gaussian Distribution, therefore the Laplace model would be more robust than the Gaussian.