1-a X_1: cancer volume

X_2: patient's age

X_3: cancer type

Model 1: $\hat{y}_1 = w_1X_1 + w_0 = [w_0\ w_1]\begin{bmatrix}1\\X_1\end{bmatrix}$

Model 2: $\hat{y}_2 = w_2X_2 + w_1X_1 + w_0 = [w_0\ w_1\ w_2]\begin{bmatrix}1\\X_1\\X_2\end{bmatrix}$

1-b $\hat{y}_3 = w_2X_2 + w_1X_1 \cdot X_3 + w_1'X_1(1-X_3) + w_0$

$= w_2X_2 + w_1(X_1X_3) + w_1'X_1 - w_1'(X_1X_3) + w_0$

$\cancel{= w_2X_2 + (w_1 - w_1')X_1X_3 + w_1'X_1 + w_0}$

$= w_2X_2 + w_1(X_1X_3) + w_1'(X_1 - X_1X_3) + w_0$

$= [w_0\ w_1'\ w_1\ w_2]\begin{bmatrix}1\\X_1 - X_1X_3\\X_1X_3\\X_2\end{bmatrix}$

1-c model 1 has 2 parameters

model 2 has 3 parameters

model 3 has 4 parameters → the most complex model

If we consider model 3, then it would be most complex model.

If we only consider model 1 and 2, then model 2 is more complex than model 1.

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

model 3 has 3 features: x1-x1*x3, x1*x3 and x2

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: $y = w_0 + w_1 X_1 + w_2 X_2 + w_3 X_3 + w_4 X_4$

3-a

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

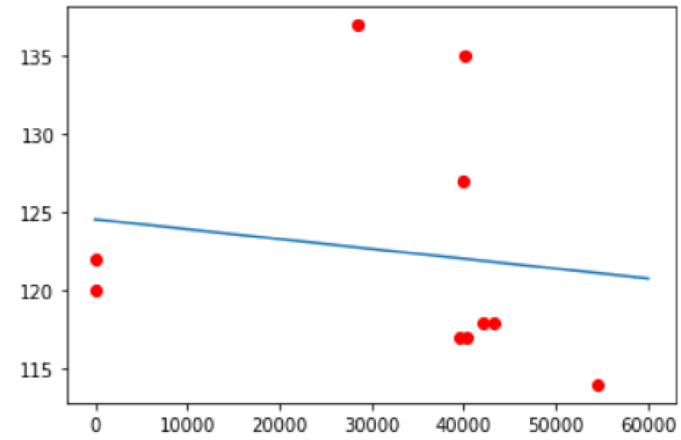
3-b:

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

$w_0 \approx 124$

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

y = (-6.3*10^(-5))*x + 124



3-c before removing the noise

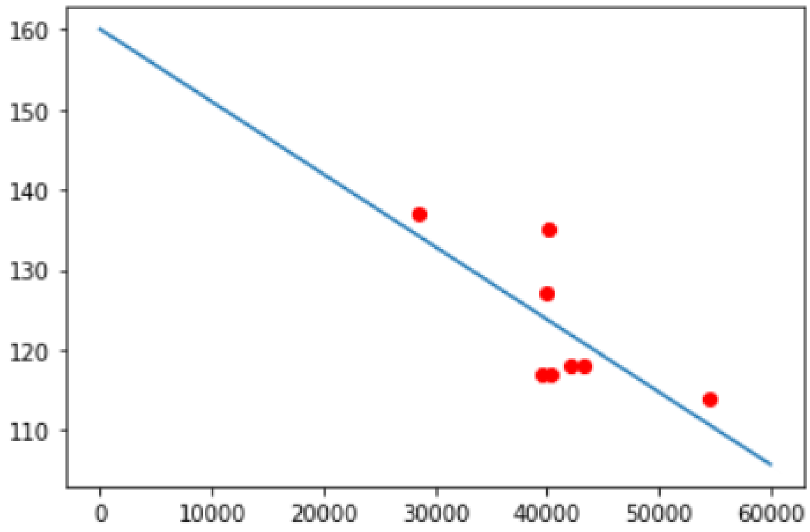
3-d $R^2 = \frac{ESS}{TSS} = \frac{12.087}{566.5} = 0.02$

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-c after removing the noise

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

3-f-e: 123.815

$$
\begin{aligned}
\text{4-a}\quad RSS(w) &= \sum_{i=1}^{N} (\hat{y_i} - y_i)^2 \\
&= \sum_{i=1}^{N} (w x_i - y_i)^2 \\
&= \| Xw - y \|_2^2 \\
&= (Xw - y)^T (Xw - y) \\
&= \boxed{w^T X^T X w - 2 w X^T y + y^T y}
\end{aligned}
$$

$$
\begin{aligned}
\text{4-b}\quad \nabla_w &= 2X^T X w - 2X^T y = 0 \\
X^T X w &= X^T y
\end{aligned}
\qquad \boxed{w = \frac{X^T y}{X^T X}}
$$

w is scalar, so I write in this way.

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

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

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

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

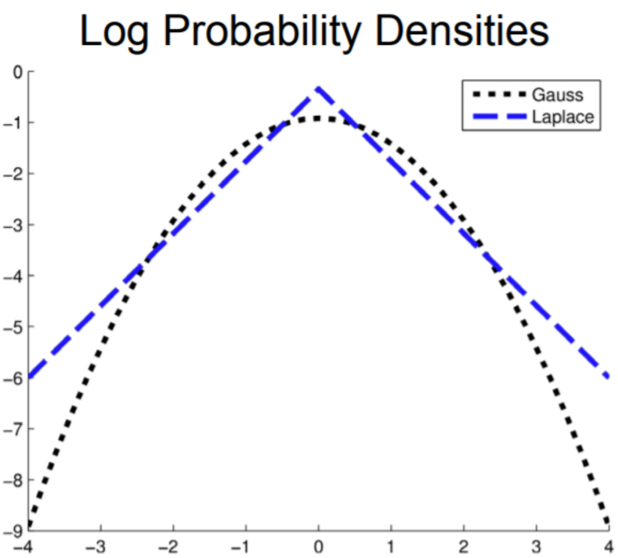
$$\log(L(w)) = \sum_{i=1}^{N} \log \frac{1}{2b} e^{\frac{-|\varepsilon^{(i)}|}{b}}$$

$$= \underbrace{N \log \frac{1}{2b}}_{\text{same for any } w} + \sum_{i=1}^{N} \frac{1}{-b} |\varepsilon^{(i)}|$$

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

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

5-b



The picture above shows that the probability of noise data appearing in Laplace Distribution is higher than Gaussian Distribution.

One reason is that the log of the Laplace's PDF is an absolute polynomial, while the Gaussian is quadratic. The probability of Laplace decreases slower than the probability of Gaussian. Therefore, the noise data in Laplace has a higher probability than Gaussian.

6 $\quad g(w) = (y - Xw)^T \Omega (y - Xw)$

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

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

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

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

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

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

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