學號：B03902101　系級： 資工四　姓名：楊力權

請實做以下兩種不同 feature 的模型，回答第 (1) ~ (3) 題：

1. 抽全部 9 小時內的污染源 feature 的一次項(加 bias)
2. 抽全部 9 小時內 pm2.5 的一次項當作 feature(加 bias)

備註 ：

a. NR 請皆設為 0，其他的數值不要做任何更動

b. 所有 advanced 的 gradient descent 技術(如: adam, adagrad 等) 都是可以用的

1. (2%)記錄誤差值 (RMSE)(根據 kaggle public+private 分數)，討論兩種 feature 的影響

	public	private
全部	8.82919	7.36600
pm2.5	8.28692	5.70832

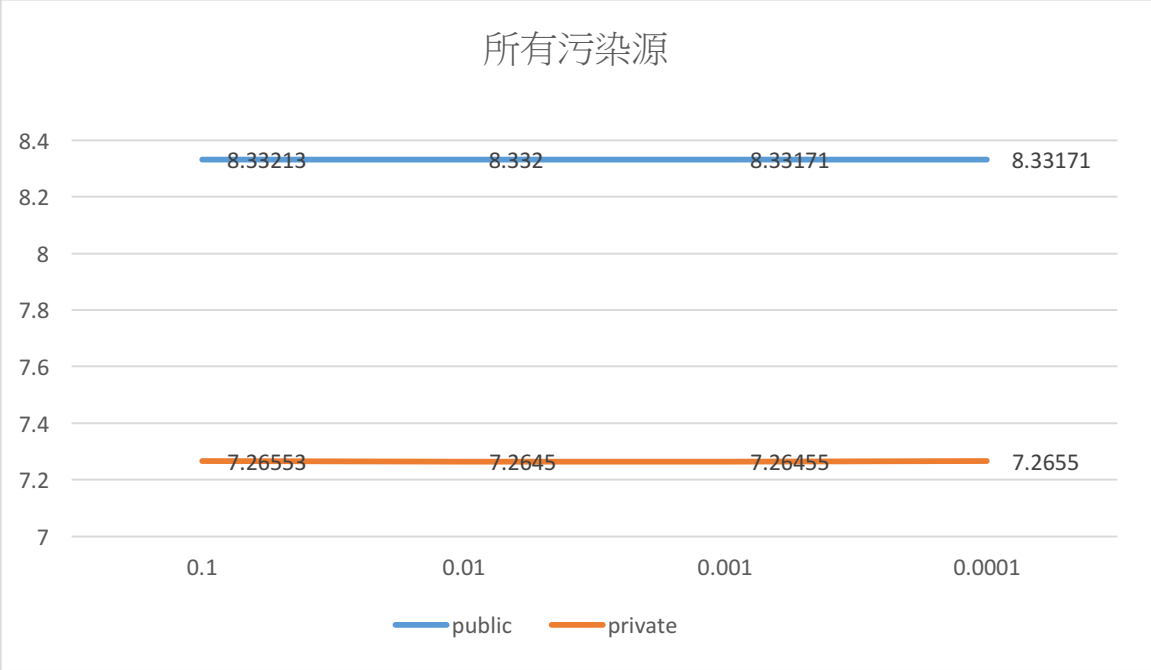
只做 pm2.5 的結果比拿所有污染源的結果還要好，是因為 18 項 feature 中有太多不相關的雜訊存在，因此訓練這些 feature 互相影響後的模型不比只訓練與 pm2.5 最有直接關係的前 9 小時 pm2.5 的模型好。

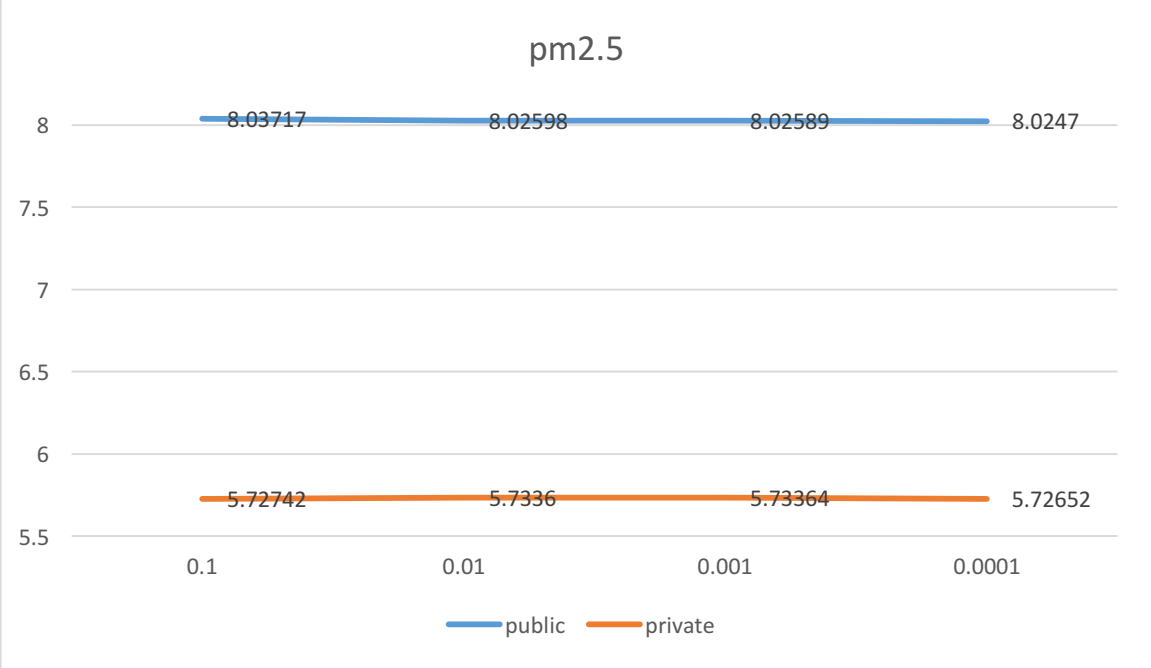
2. (1%)將 feature 從抽前 9 小時改成抽前 5 小時，討論其變化

	public	private
全部	8.48199	5.66906
pm2.5	8.06294	5.85941

很明顯能夠發現 5 小時的 feature 有著比較好的結果, 而原因可能是因為前 6~9 的污染源對於第 10 小時的 pm2.5 已經沒有影響了，因此減少不必要雜訊可以獲得更好的預測模型。

3. (1%)Regularization on all the weight with λ=0.1、0.01、0.001、0.0001，並作圖





4. (1%)在線性回歸問題中，假設有 N 筆訓練資料，每筆訓練資料的特徵 (feature) 為一向量 x^n，其標註(label)為一存量 y^n，模型參數為一向量 w (此處忽略偏權值 b)，則線性回歸的損失函數(loss function)為 n=1Nyn-xnw2 。若將所有訓練資料的特徵值以矩陣 $X = [x^1\ x^2 \dots x^N]^T$ 表示，所有訓練資料的標註以向量 $y = [y^1\ y^2 \dots y^N]^T$ 表示，請問如何以 X 和 y 表示可以最小化損失函數的向量 w ？請寫下算式並選出正確答案。(其中 X^TX 為 invertible)

a. $(X^TX)X^Ty$
b. $(X^TX)^{-0}X^Ty$
c. $(X^TX)^{-1}X^Ty$
d. $(X^TX)^{-2}X^Ty$

設有N筆資料，一筆資料有n個feature
W是一個n x 1的矩陣代表weight $W = [w_1\ w_2\ \dots\ w_n\ ]^T$
Y是一個N x 1的矩陣代表每筆資料推測出的正確數值 $Y = [y_1\ y_2\ \dots\ y_N]^T$
X是一個N x n的矩陣代表N筆資料每筆有n個feature $X = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & X_{Nn} \end{bmatrix}$
設Loss function為 $L = \sum_{i=0}^{N}(y_i - w_1x_1 - w_2x_2 - \cdots - w_nx_n)^2 = (Y - XW)^2$
對W做微分希望能最小化L $\frac{\delta}{\delta w}L = 2X^T(Y - XW) = 0$
因此$2X^T(Y - XW) = 0$
因為X^TX可逆 可得W $X^TY = X^TXW \implies W = (X^TX)^{-1}X^TY$

所以答案是 C.