學號：B03201031　系級：　數學四　姓名：王楷

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

	private	public	total
全部汙染源	5.36561	7.59287	12.95848
只有 pm2.5	5.62839	7.45285	13.08124

從此表格觀察，若觀察 private 的 RMSE，只有 pm2.5 feature 的 model 誤差略高於抽取全部汙染源的 model，而 public 的結果則相反，而將 public+private，只有 pm2.5 的略高，但兩者相差非常小，很難判斷哪個 model 比較好

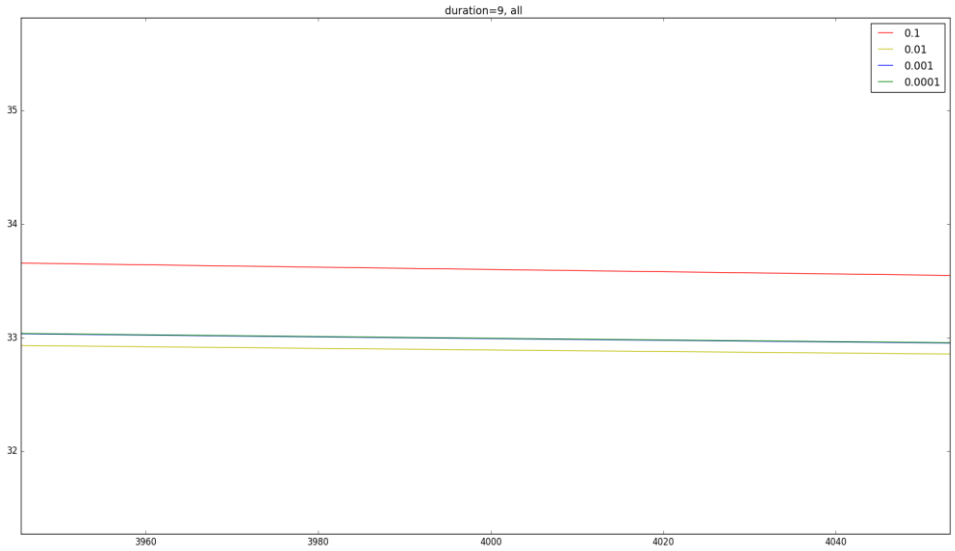
2. (1%)將 feature 從抽前 9 小時改成抽前 5 小時，討論其變化

	private	Public	total
全部汙染源	5.41397	7.64271	13.05668
只有 pm2.5	5.79853	7.57784	13.37637

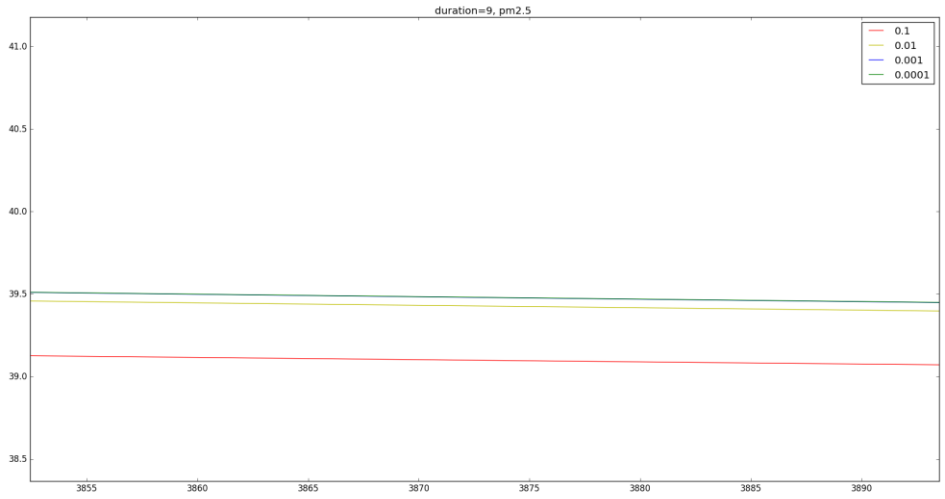
總體來說，兩個 model 在兩種結果(public, private)，的誤差都上升了，其中只有 pm2.5 feature 的 model 又更明顯，可以判斷此模型太 simple

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

   Take all as feature:



Take pm2.5 as feature



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.$(X^TX)^{-2}X^Ty$

The normal equations can be derived directly from a matrix representation of the problem as follows. The objective is to minimize

$$S(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}^{\mathrm{T}}\mathbf{y} - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}.$$

Note that : $(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y})^{\mathrm{T}} = \mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}$ has the dimension 1x1 (the number of columns of y), so it is a scalar and equal to its own transpose, hence $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}$ and the quantity to minimize becomes $S(\boldsymbol{\beta}) = \mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\beta}.$ Differentiating this with respect to $\boldsymbol{\beta}$ and equating to zero to satisfy the first-order conditions gives $-\mathbf{X}^{\mathrm{T}}\mathbf{y} + (\mathbf{X}^{\mathrm{T}}\mathbf{X})\boldsymbol{\beta} = 0,$ which is equivalent to the above-given normal equations. A sufficient condition for satisfaction of the second-order conditions for a minimum is that $\mathbf{X}$ have full column rank, in which case $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is positive definite

so we pick (c).