1

(a)
(1) 先移除未用到的參數，並做 one hot encoding。

(2) 根據欲獲取的 training data 比例求出 training data 筆數，並隨機從總筆數抽出做為 training data 的 index，其餘 index 就做為 test set。

(3) 分別將 training set, test set 以 Training set 的 mean, std 做標準化即得 training set & test set。

(c) $J(w) = \frac{1}{m}\underbrace{(Xw-Y)^T(Xw-Y)}_{MSE\text{-}train} + \frac{1}{2}\lambda w^T w$

objective function:
$(Xw-Y)^T(Xw-Y) + \frac{\alpha}{2}w^T w$
$= \frac{m}{m}(Xw-Y)^T(Xw-Y) + \frac{m\lambda}{2}w^T w$

Normalize → $w = (X^TX + \alpha I)^{-1}X^TY = (X^TX + m\lambda I)^{-1}X^TY$

(f)
(1) pseudo-inverse 及 regularization without bias figure 的結果較偏離正確值，regularization with bias 及 Bayesian Linear Regression 所得結果接近正確值，又以 Bayesian Linear Regression 表現最佳。

(2) 因 model(d)(e) 皆有 bias 項，多了一個參數，$MSE = Bias(\hat{\theta}) + Var(\hat{\theta})$
當多了一個參數時即可降低 $Bias(\hat{\theta})$ 而使 MSE ↓

2

(1) 影響薪水是否>50K最主要兩個parameter為：age - martial - status
而工作相關的：workclass, hours-per-week. occupation 也有一定影響力。財務相關(capital-gain, capital-loss)也有相關，而與個人相關的以native-country最相關，其他如sex, race的影響力相對普通。整體而言，影響薪水的因素主要還是以年齡(年資)、工作性質與內容、工時為主，而一些個人天生條件，如性別、種族其實非主要影響因素。

(2) test data各parameter中包含了一些train data未出現的類別，因此會影響預測的表現，RMSE表現較training時差。