1. Read the paper “Diagnosis of multiple cancer types by shrunken centroids of gene expression” by Tibshirani et al. (2001). Summarize your understanding in three sentences.

2. (Exercise 2.6 in the textbook) Consider a regression problem with inputs $x_i$ and outputs $y_i$, and a parameterized model $f_\theta(x)$ to be fit by least squares. Show that if there are observations with tied or identical values of $x$, then the fit can be obtained from a reduced weighted least squares problem.

3. Exercise 2.7(a)(b).

4. Exercise 2.8.

5. (k-Nearest Neighbor for Classification)
   You may use the functions knn or knn1 in the R library “class” for this problem. Submit your codes along with your results.
   
   (a) Fit k-nearest neighbor classifier with a range of values $k$ for the training data in Scenario 1, $k = \{1, 4, 7, 10, 13, 16, 30, 45, 60, 80, 100, 150, 200\}$. Report both training and testing errors for each k-NN classifier. Plot two curves: the training error vs the degree of freedom $n/k$, and the testing error vs $n/k$, in one same figure (Similar to Figure 2.4 in the textbook).

   (b) Repeat (a) for Scenario 2.

   (c) Based on the plots obtained in (a) and (b), describe the different patterns between two curves. How should you choose the best $k$ and recommend your best $k$ for each scenario.

6. Classify the 1’s, 2’s, 3’s for the zip code data.
   
   (a) Use the $k$-nearest neighbor classification with $k = 1, 3, 5, 7, 15$. Show both the training and test error for each choice.

   (b) Implement the LDA method and report its training and testing errors. Note: Before carrying out the LDA analysis, you are suggested to delete variable 16 first from the data, since the variable takes a constant value and it can cause the singularity of the covariance matrix. In general, a constant variable does not have a discriminating power to separate two classes.