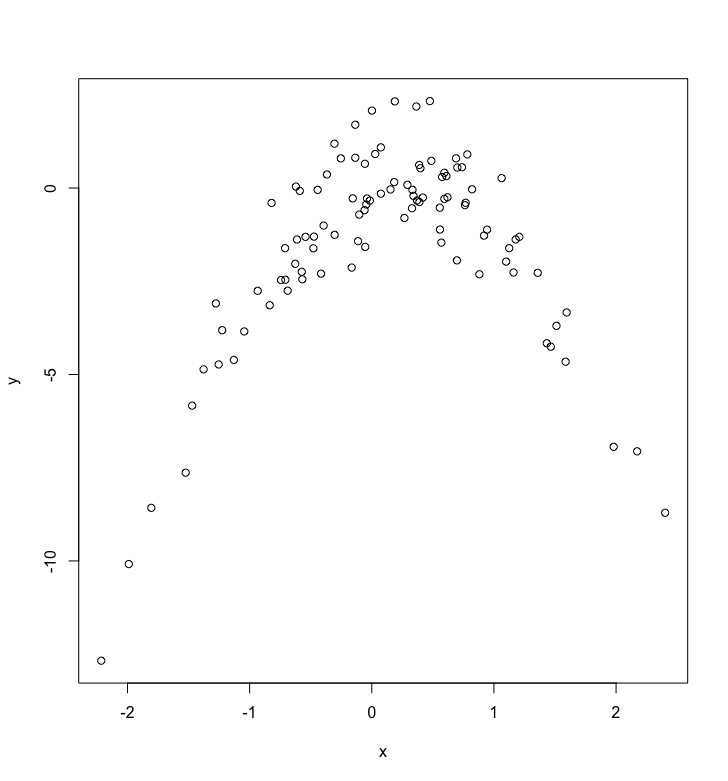
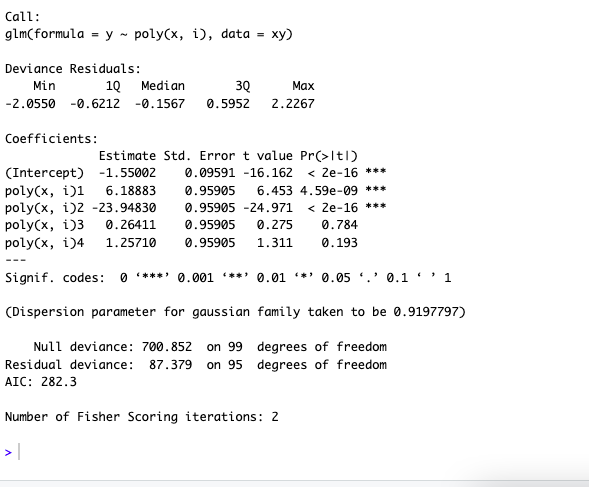
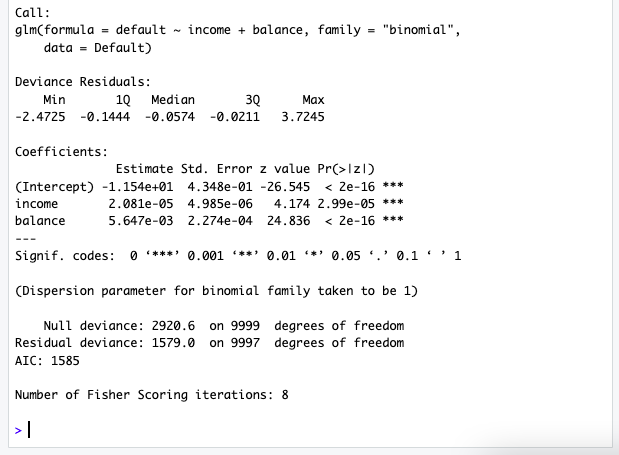
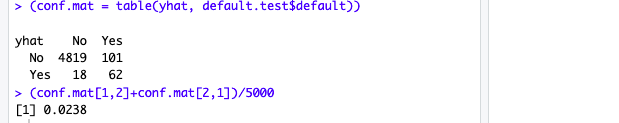
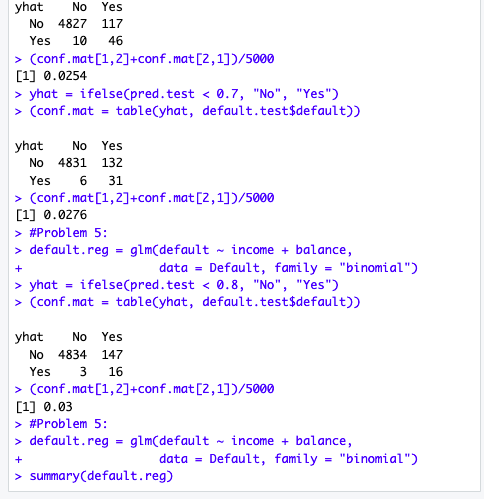
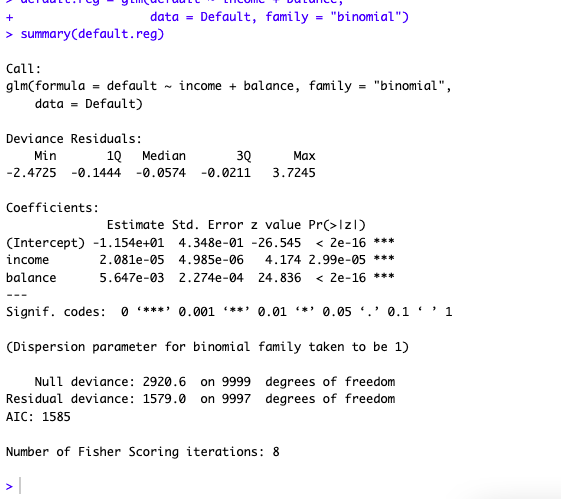
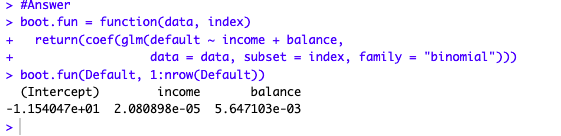
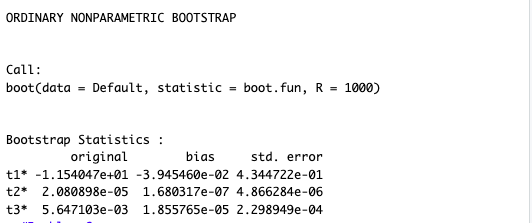
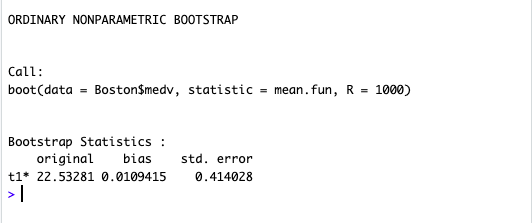
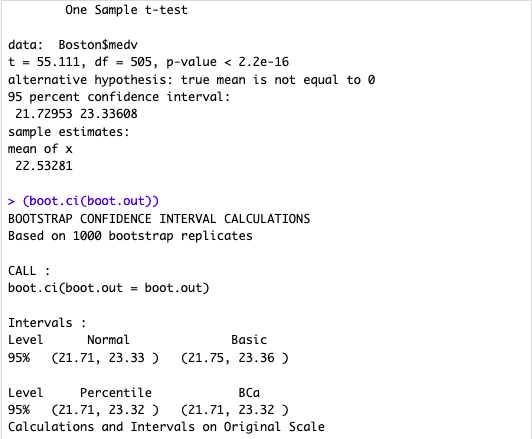
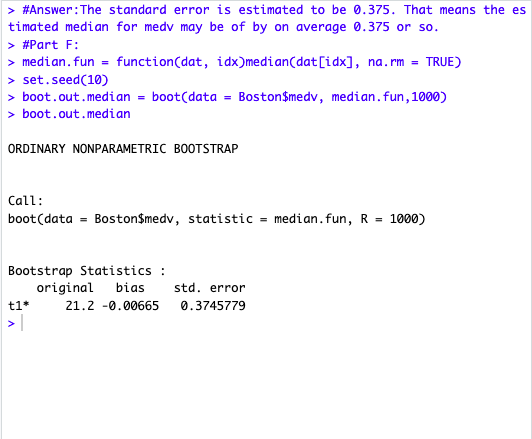
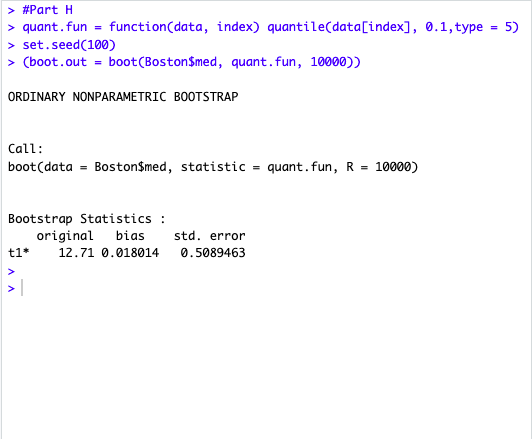
* **Problem 1:**
  + **(A)explain how k-fold cross-validation is implemented.**
  + **(B)What are the advantages and disadvantages of k-fold cross-validation relative to:**
    - The validation set approach?
    - LOOVC?
  + **Answer:**
    - **(A)** This approach involves randomly k-fold CV dividing the set of observations into k groups, or folds, of approximately equal size. The frst fold is treated as a validation set, and the method is fitted on the remaining k − 1 folds. The mean squared error, MSE1, is then computed on the observations in the held-out fold. This procedure is repeated k times; each time, a diferent group of observations is treated as a validation set. This process results in k estimates of the test error, MSE1, MSE2, . . . , MSEk.
    - The k-fold CV estimate is computed by averaging these values,
      * CV(k) = 1/k^k \_i=1 MSE\_i
    - **(B)** Advantages
      * Less computationally intensive
      * Doesn’t lose in estimation quality
      * The variability in the estimates are negligible
* **Problem 2:**
  + **Suppose that we use some statistical learning method to make a prediction for the response Y for a particular value of the predictor X. Carefully describe how we might estimate the standard deviation of our prediction.**
    - **Answer:**
      * If we suppose using some statistical learning method to make a prediction for the response Y for a particular value of the predictor X we might estimate the standard deviation of our prediction by using the bootstrap approach. The bootstrap approach works by repeatedly sampling observations (with replacement) from the original data set B times, for some large value of B, each time fitting a new model and subsequently obtaining the RMSE of the estimates for all B models.
* **Problem 3:**
  + **Generate a simulated data set as follows:**
  + **A.) In this data set, what is n and what is p? Write out the model used to generate the data in equation form.**
    - **Answer**: Y = 2X^2 + epsilon, n = 100, p = 2
  + **B.)Create a scatter plot of X against Y . Comment on what you find.**
* **Answer: **
  + - Set a random seed, and then compute the LOOCV errors that result from fitting the following four
    - models using least squares:
      * Y = β0 + β1X + ε
      * Y = β0 + β1X + β2X2 + ε
      * Y = β0 + β1X + β2X2 + β3X3 + ε
      * Y = β0 + β1X + β2X2 + β3X3 + β4X4 + ε
    - **Answers:** 
      * [1] 7.2881616 0.9374236 0.9566218 0.9539049
      * [1] 7.2881616 0.9374236 0.9566218 0.9539049

****

* **Note:** you might find it helpful to use the data.frame() function to create a single data set containing both X and Y .
  + Repeat (c) using another random seed, and report your results. Are your results the same as what you got in (c)? Why?
  + Which of the models in (c) had the smallest LOOCV error? Is this what you expected? Explain your answer.
  + Comment on the statistical significance of the coefficient estimates that results from fitting each of the models in (c) using least squares. Do these results agree with the conclusions drawn based on the cross-validation results?
* **Problem 4:**
  + **We will use a logistic regression to predict the probability of default using income and balance on the Default data set in the ISLR package. We will now estimate the test error of this logistic regression model using the validation set approach. Do not forget to set a random seed before beginning your analysis.**
    - (A) Fit a logistic regression model that uses income and balance to predict default.
  + ****
    - Using the validation set approach, estimate the test error of this model. In order to do this, you must perform the following steps:
      * Split the sample set into a training set and a validation set.
      * Fit a multiple logistic regression model using only the training observations.
      * Obtain a prediction of default status for each individual in the validation set by computing the posterior probability of default for that individual, and classifying the individual to the default category if the posterior probability is greater than 0.5.
      * Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.
      * **Answer:** 
        + ****
    - Repeat the process in (b) three times, using three different splits of the observations into a training set and a validation set. Comment on the results obtained.
    - 
      * They all stayed between 2.5% and 3%
* **Problem 5:**
  + **We continue to consider the use of a logistic regression model to predict the probability of default using income and balance on the Default data set. In particular, we will now compute estimates for the standard errors of the income and balance logistic regression coefficients in two different ways: (1) using the bootstrap, and (2) using the standard formula for computing the standard errors in the glm() function. Do not forget to set a random seed before beginning your analysis.**
    - Using the summary() and glm() functions, determine the estimated standard errors for the coefficients associated with income and balance in a multiple logistic regression model that uses both predictors.
      * ****
      * Standard error for βˆ2 = 0.000005 and βˆ3 = 0.00023
    - Write a function, boot.fn(), that takes as input the Default data set as well as an index of the observations, and that outputs the coefficient estimates for income and balance in the multiple logistic regression model.
    - Use the boot() function together with your boot.fn() function to estimate the standard errors of the logistic regression coefficients for income and balance.
      * 
    - Comment on the estimated standard errors obtained using the glm() function and using your bootstrap function.
      * The standard error using both methods is very close
* **Problem 6:**
  + **We will now consider the Boston housing data set, from the MASS library.**
    - Based on this data set, provide an estimate for the population mean of medv. Call this estimate ˆμ.
      * 22.53281
    - Provide an estimate of the standard error of ˆμ. Interpret this result. Hint: We can compute the standard error of the sample mean by dividing the sample standard deviation by the square root of the number of observations.
      * 0.4088611
    - Now estimate the standard error of ˆμ using the bootstrap. How does this compare to your answer from



* + - * Both errors are very close to 0.41
    - Based on your bootstrap estimate from (c), provide a 95% confidence interval for the mean of medv. Compare it to the results obtained using t.test(Boston$medv). Hint: You can approximate a 95% confidence interval using the formula [ˆμ −2 ×SE(ˆμ), ˆμ + 2 ×SE(ˆμ)].
      * 
    - Based on this data set, provide an estimate, ˆμmed, for the median value of medv in the population.
      * 21.2
    - We now would like to estimate the standard error of ˆμmed. Unfortunately, there is no simple formula for computing the standard error of the median. Instead, estimate the standard error of the median using the bootstrap. Comment on your findings.
      * Answer:The standard error is estimated to be 0.375. That means the estimated median for medv may be of by on average 0.375 or so.
    - Based on this data set, provide an estimate for the tenth percentile of medv in Boston suburbs. Call this quantity ˆμ0.1. (You can use the quantile() function.)
      * 10%
      * 12.71
    - Use the bootstrap to estimate the standard error of ˆμ0.1. Comment on your findings.
      * 
      * The estimate may be inaccurate by give or take 0.51