Stat 462/862 Assignment 1

(Due on 11:59pm on Sept 27, 2023, on Crowdmark)

1. Load the data *Auto* after installing and loading the library *ISLR2*. and complete the following parts.

   (a) Look up the help for *Auto*.

   (b) Compute the dimension size of *Auto* and display its dimension names.

   (c) Write an expression that returns the summary statistics, as given by *summary()*, for each of the first eight columns in this matrix.

   (d) Apply the *pairs()* function to the matrix to create scatter plots of the first eight columns against one another.

   (e) Create a correlation matrix of the first seven columns

   (f) Select the cars that have < 16 *mpg*, but having eight cylinders.

   (g) Select the cars that have greater than average horsepower, but less than average weight.

   (h) Create a design matrix, X, that contains all 1's in the first column and the *hoursepower* in the second column. This will serve as our design matrix in part (m).

   (i) Assume that there is an approximate linear relationship between *horsepower* and *mpg*. Given the design matrix defined above, we can define a simple linear model as $y = X\beta + \epsilon$ where $y(mpg)$ is the dependant variable, X is the design matrix of independent variables, β is the vector of parameters and ϵ is the error term. The least squares estimate of β is: $\hat{\beta} = (X^T X)^{-1} X^T y$, where X^T is the transpose of X. Write an expression to calculate $\hat{\beta}$ in R.

   (j) In a single plot, draw the following subplots (1) a scatter plot of *mpg* (y) versus *horsepower* (x); (2) a histogram of *mpg*; (3) a Quantile-Quantiale plot of *mpg*;(4) a boxplot of *mpg* and *horsepower*.

2. Calculate the probability for each of the following events: (a) A standard normally distributed variable is larger than 3. (b) A normally distributed variable with mean 30 and standard deviation 4 is larger than 35. (c) Getting 10 out 10 successes in a

binomial distribution with probability of 0.83. (d) $X < 0.9$ when X has a standard uniform distribution. (e) $X > 6.5$ in a χ^2 distribution with 2 degrees of freedom.

3. This is a question to help understand *bias-variance trade-off*. Consider one input variable x and the response variable y. Suppose the true relationship between x and y is $y = 1.5 + 3.1x + 1.2x^2 + 4.5x^3 + \epsilon$ where $E(\epsilon) = 0$ and $Var(\epsilon) = 1$.

   (a) Generate 30 points as training data from the true function with x coming from a set of random values on (0,1), and make a plot of y versus x.

   (b) Fit data in (a) with each of the following models

   - Model 1: $y = \beta_0 + \beta_1 x + \epsilon$
   - Model 2: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$
   - Model 3: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$
   - Model 4: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \epsilon$
   - Model 5: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5 + \epsilon$
   - Model 6: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4 + \beta_5 x^5 + \beta_6 x^6 + \epsilon$

   and compute bias square, variance and MSE. To computer bias, variance and MSE, generate equally-spaced 100 points $(x_1, \ldots, x_{100})$ on (0,1). Compute the bias square $= \sum_{i=}^{100} (\hat{y}_i - (1.5 + 3.1x_i + 1.2x_i^2 + 4.5x_i^3))^2/100$, with $\hat{y}_i$ being the fitted value using one of the models above, compute variance $= \sum_{i=1}^{100} var(\hat{y}_i)/100$ with $var(\hat{y}_i)$ being the predictive variance at x_i, and computer MSE = bias square + variance. Note that both the fitted values and predictive variance can be obtained using the predict function.

   (c) Make a plot that displays three curves, one is for bias square versus the order of the linear model, one is for variance versus the order of the linear model, and one is for MSE versus the order of the linear model.

   (d) Repeat parts (a) – (c) several times, and report your finding.

4. To provide clients with quantitative information upon which to make rental decisions, a commercial real estate company evaluate vacancy rates, square footage, rental rates, and operating expenses for commercial properties in a large metropolitan are. See the data file *CommercialProperties.txt*. It lists the 81 suburban commercial properties

with the variables the age X_1, operating expense and taxes X_2, vacancy rates X_3, total square footage X_4 and the rental rates Y.

(a). In one figure, plot the four subfigures each of which is the boxplot for each predictor variable. What information do these plots provide?

(b). Obtain the scatter plot matrix and the correlation matrix. Interpret these and state your principal findings.

(c). Fit a regression model for the four predictor variables to the data. State the estimated regression function.

(d). Obtain the residuals and prepare a box plot of the residuals. Does the distribution appear to be fairly symmetrical?

(e). Obtain the prediction of mean response, its associated prediction error and 95% confidence interval based on the fitted model for the new input $X_1 = 5, X_2 = 8.25, X_3 = 0, X_4 = 250000$.

5. Consider the simple logistic regression model which assumes

$$P(Y = 1) = \pi(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

(a) Plot the logistic mean response π_i when $\beta_0 = -10$ and $\beta_1 = 0.4$.

(b) What is the value of X such that the logistic mean response $\pi(X)$ is equal to 0.5?

(c) Find the odds when $X = 25$ and $X = 26$.

(d) Let X_1 and X_2 be two inputs of the predictor X. The odds ratio of X_1 to X_2 is defined as $(\pi(X_1)/(1 - \pi(X_1)))/(\pi(X_2)/(1 - \pi(X_2)))$. Compute the odds ratio of $X = 26$ to $X = 25$. Is it equal to e^{β_1} as it should be?

6. Consider the dataset in banknote.txt. The description and background of the dataset can be seen at `https://archive.ics.uci.edu/dataset/267/banknote+authentication`. You can use R command "banknote= read.table("banknote.txt", sep = ",", header = FALSE)" to import the data. The variable names from left to right are variance, skewness, curtosis, entropy, and class. Suppose $Y = 1$ corresponds to the banknote is forgery and otherwise $Y = 0$ corresponds to the banknote is authentic. The predictors considered are the variance of Wavelet Transformed image (X_1), the skewness of Wavelet Transformed image (X_2), the curtosis of Wavelet Transformed image (X_3) and

the entropy of image (X_4). Answer the following questions using the logistic regression model with the four predictors in their first-order terms.

(a) Find the maximum likelihood estimates of $\beta_0, \beta_1, \beta_2$, β_3 and β_4. State the fitted mean response function $\hat{\pi}$, where $\pi = P(Y = 1)$.

(b) Let $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ and $\hat{\beta}_4$ be the maximum likelihood estimate of $\beta_1, \beta_2, \beta_3, \beta_4$, respectively. Obtain $e^{\hat{\beta}_1}$, $e^{\hat{\beta}_2}$ and $e^{\hat{\beta}_3}$. Interpret these numbers from the odds ratio point of view (see part (d) of question 5).

(c) What is the estimated probability that the banknote with $X_1 = 01, X_2 = -0.5, X_3 = 1, X_4 = 0.5$ is authentic?

7. (Graduate students only) Read Section 3.5 of the ISLR textbook. Import the dataset using the command

```
prostate = read.csv("http://www-stat.stanford.edu/~tibs/ElemStatLearn/
datasets/prostate.data",row.names=1,sep="\t")
```

Note that the last column indicates whether or not an entry is a training data point. Treat the data entries that are not training data as test data. Treat *lpsa* as the response variable and other columns (except the last columns) as predictors.

(a) Fit multiple linear regression models. Provide the model fits and compute the mean square prediction errors using the test data.

(b) Install and load the R package FNN, and use the *knn.reg* function to fit the k-nearest neighbourhood model, and compute the mean square prediction errors. Provide the mean square prediction errors for $k = 1, 3, 5, 10$.