W23 STAT 362 R for Data Science

Assignment 5

Due: 24 Mar 11:59pm.

Q1: Install and load the package `ISLR2`.

Q1: Consider the dataset `Hitters` in `ISLR2`. Remove the rows with NA values using the following code.

```
library(ISLR2)
Hitters <- na.omit(Hitters)
```

Use `?Hitters` and read the description of the dataset. Which variables are categorical?

Q2 (a): Create a scatterplot to visualize the relationship between players' salaries and their years of experience in Major League Baseball. Use `ggplot()`.

Q2 (b): Create a scatterplot to visualize the relationship between players' salaries and their years of experience in Major League Baseball. Use different colors according to the player's division. Use `ggplot()`.

Q3 (a): Create a histogram to visualize the distribution of player salaries in the `Hitters` dataset using `geom_histogram` with 20 bins. Describe one feature that you see from the histogram.

Q3 (b): Create a kernel density estimate to visualize the distribution of player salaries in the `Hitters` dataset. Use `geom_density`.

Q3 (c): Compare the histogram and the kernel density estimate in (a) and (b), and describe one difference between them.

Q4 (a): Fit a simple linear regression with response `Salary` and predictor `CHits` using the `Hitters` dataset

Q4 (b): Describe the interpretation of the estimated regression coefficient of `CHits`.

Q5 (a): Use the following code to find the 20%, 40%, 60%, 80% quantiles of `CHits`

```
CHits_quan <- quantile(Hitters$CHits, c(0.2, 0.4, 0.6, 0.8))
```

Then, create a new categorical variable called `cat_CHits` with 5 groups for the categorized variable of `CHits` according to the quantiles above. Use the following code.

```
cat_CHits <- rep(0, nrow(Hitters))
cat_CHits[Hitters$CHits <= CHits_quan[1]] <- 1
cat_CHits[Hitters$CHits > CHits_quan[1] & Hitters$CHits <= CHits_quan[2]] <- 2
cat_CHits[Hitters$CHits > CHits_quan[2] & Hitters$CHits <= CHits_quan[3]] <- 3
cat_CHits[Hitters$CHits > CHits_quan[3] & Hitters$CHits <= CHits_quan[4]] <- 4
cat_CHits[Hitters$CHits > CHits_quan[4]] <- 5
cat_CHits <- factor(cat_CHits)
```

Fit a simple linear regression with response `Hitters$Salary` and predictor `cat_CHits`.

Q5 (b): Describe the interpretation of the estimated regression coefficients of `cat_CHits`.

Q6: Split the `Hitters` dataset into `train` and `test` using the following code.

```
set.seed(1)
index <- sample(nrow(Hitters), nrow(Hitters) * 0.5)
train <- Hitters[index, ]
test <- Hitters[-index, ]
```

Q6 (a): Perform a multiple linear regression with response `Salary` and the remaining variables as covariates. Use the dataset `train` only.

Q6 (b): Which variables are statistically significant at significance level 0.05?

Q6 (c): Describe the interpretation of the estimated regression coefficient of `CHits`.

Q7 (a): Compute the training error in terms of mean squared error for the model fitted in Q6.

Q7 (b): Compute the testing error in terms of mean squared error for the model fitted in Q6.

Download the two datasets `iris_train.csv` and `iris_test.csv` from onQ (onQ -> Content -> Datasets). Import them to R using `read.csv()`.

Q8: Create an additional response called `versicolor` to indicate if the iris species is `versicolor` using the following code.

```
iris_train$versicolor <- as.numeric(iris_train$Species == "versicolor")
iris_test$versicolor <- as.numeric(iris_test$Species == "versicolor")
```

Fit a logistic regression with response `versicolor` and features `Sepal.Length`, `Sepal.Width`, `Petal.Length` and `Petal.Width`. Find the estimates of the regression coefficients and their corresponding *p*-values using `summary()`.

Q9: Based on the results in Q8, which variable(s) should be important for classifying if the iris is versicolor.

Q10: Find the classification accuracy using `iris_test` for your model in Q8.