Homework Coding 2

KhoaLe

02/07/2022

Chapter 6

Question 6.7.3

(a)

```r
# Using the isis dataset to the data frame
mydf_iris <- iris

# Grouping label - Coverting to binary
mydf_iris$binary <- gsub("setosa", "0", mydf_iris$Species)
mydf_iris$binary <- gsub("versicolor", "0", mydf_iris$binary)
mydf_iris$binary <- gsub("virginica", "1", mydf_iris$binary)


# Convert to numeric
mydf_iris$binary <- as.numeric(mydf_iris$binary)
```

(b)

```r
# Reading library
library(splitstackshape)

# Sampling into training and testing
training_testing_iris <- stratified(as.data.frame(mydf_iris), group = 6,
                                    size = 0.8, bothSets = T)
# Deleting Species
df <- subset(mydf_iris, select = -c(Species))

# Building a logistic regression with using all data
my_logit <- glm(binary ~ Sepal.Length+Sepal.Width+Petal.Length+Petal.Width,
                data = df, family = "binomial")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
# Summary my_logit
summary(my_logit)
```

```
##
## Call:
## glm(formula = binary ~ Sepal.Length + Sepal.Width + Petal.Length +
##     Petal.Width, family = "binomial", data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.01105  -0.00065   0.00000   0.00048   1.78065
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -42.638     25.708  -1.659   0.0972 .
## Sepal.Length   -2.465      2.394  -1.030   0.3032
## Sepal.Width    -6.681      4.480  -1.491   0.1359
## Petal.Length    9.429      4.737   1.990   0.0465 *
## Petal.Width    18.286      9.743   1.877   0.0605 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 190.954  on 149  degrees of freedom
## Residual deviance:  11.899  on 145  degrees of freedom
## AIC: 21.899
##
## Number of Fisher Scoring iterations: 12
```

The result of summary my_logit **(c)**

```
# Calculate the probability
my_prob <- 1/(1+exp(-(-42.638 +-2.465*9 +-6.681*5 + 9.429*10 + 18.286*7)))

# Print result
print(my_prob)
```

```
## [1] 1
```

Base on the logistic regression model, there will be 100% chance of a new plant being a Virginca.

Question 6.7.4

(a)

```
# Reading library
library(rpart)

# Using the kyphosis dataset to the data frame
mydf_kyphosis <- kyphosis

# Converting variable to numeric
mydf_kyphosis$binary <- gsub("present", "1", mydf_kyphosis$Kyphosis)
```

```
mydf_kyphosis$binary <- gsub("absent", "0", mydf_kyphosis$binary)

# Converting to numeric
mydf_kyphosis$binary <- as.numeric(mydf_kyphosis$binary)
```

(b)

```
# Building a logistic regression with using all data
my_logit1 <- glm(binary ~ Age+Number+Start,
                data = mydf_kyphosis, family = "binomial")
# Summary my_logit
summary(my_logit1)
```

```
##
## Call:
## glm(formula = binary ~ Age + Number + Start, family = "binomial",
##     data = mydf_kyphosis)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3124  -0.5484  -0.3632  -0.1659   2.1613
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.036934   1.449575  -1.405  0.15996
## Age          0.010930   0.006446   1.696  0.08996 .
## Number       0.410601   0.224861   1.826  0.06785 .
## Start       -0.206510   0.067699  -3.050  0.00229 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.234  on 80  degrees of freedom
## Residual deviance: 61.380  on 77  degrees of freedom
## AIC: 69.38
##
## Number of Fisher Scoring iterations: 5
```

Base on the summary, we have:

- The variables of Age and Number are both insignificant
- Start variable has a great impact on our regression with p-value = 0.00229

(c)

```
# Calculate the probability
my_prob1 <- 1/(1+exp(-(-2.036934 + 0.010930*50 + 0.410601*5 + -0.206510*10)))

# Print result
print(my_prob1)
```

```
## [1] 0.1820486
```

Base on the logistic regression model, there will be 18.2% chance of being "present" for Age, Start, and Number

Question 6.7.5

```
# Install package
#install.packages("lmtest")

#Loading library
library("lmtest")
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

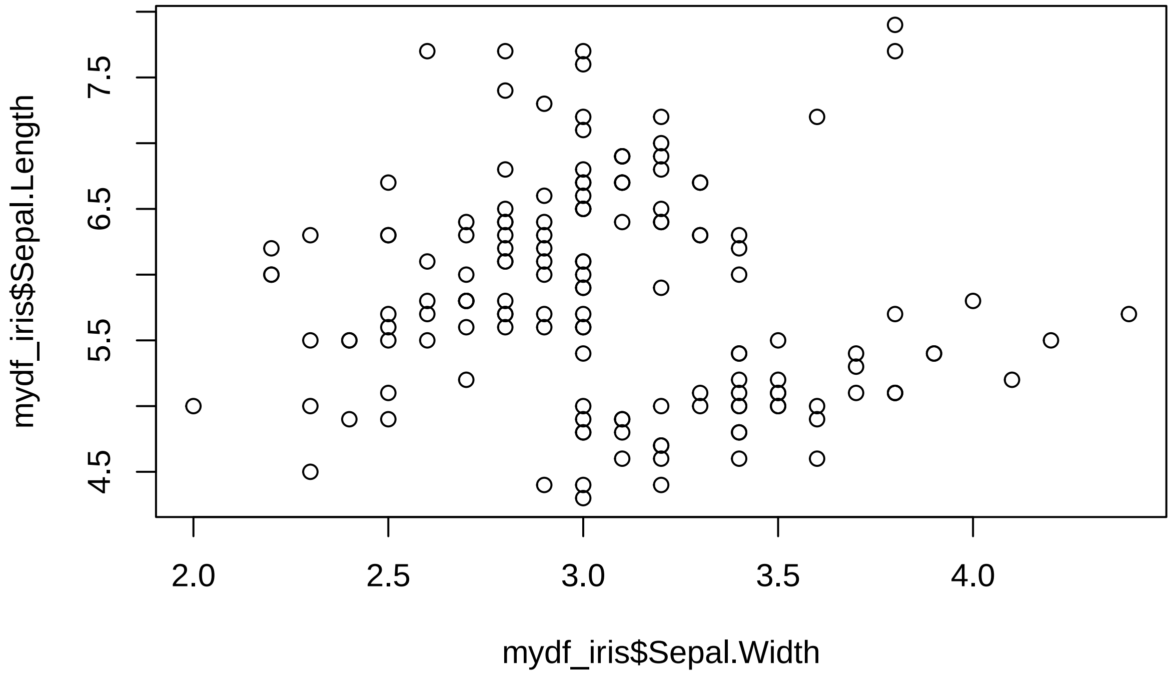
```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

Definition of homoscedastic and heterscedasticity (According to Frost, n.d): - Heteroscedasticity is a change in the spread of residuals over a range of measured values that is systematic. Because ordinary least squares (OLS) regression implies that all residuals are obtained from a population with a fixed variance, heteroscedasticity is a concern (homoscedasticity).

Note from the book (Kurnicki, n.d):

- If p-value is lower than 0.5 or 1, we must reject the null hypothesis and come to the conclusion that heteroscedasticity exists.
- The data is homoscedastic, according to the null hypothesis (error variances are all equal). This is why, in order to conclude homoscedasticity, we need larger p-values.

```
# Set x = Sepal.Width, y = Sepal.Length
plot(x= mydf_iris$Sepal.Width, y=mydf_iris$Sepal.Length, type = "p")
```



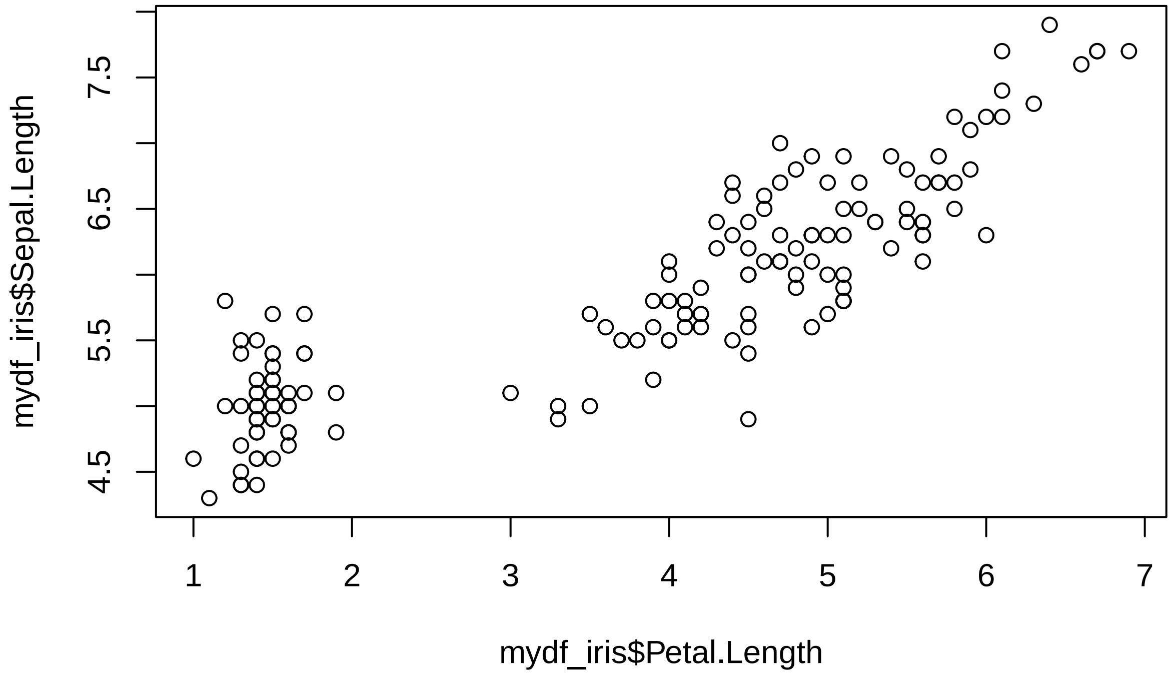
As a result, this is homoscedastic

```
# Testing for heteroscendasticity by using linear
# x = Sepal.Width, y = Sepal.Length
swidth_slength <- lm(Sepal.Length~Sepal.Width, data = mydf_iris)
bptest(swidth_slength)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  swidth_slength
## BP = 0.78243, df = 1, p-value = 0.3764
```

The p-value is 0.3764

```
# Set x = Petal.Length, y = Sepal.Length
plot(x= mydf_iris$Petal.Length, y=mydf_iris$Sepal.Length, type = "p")
```



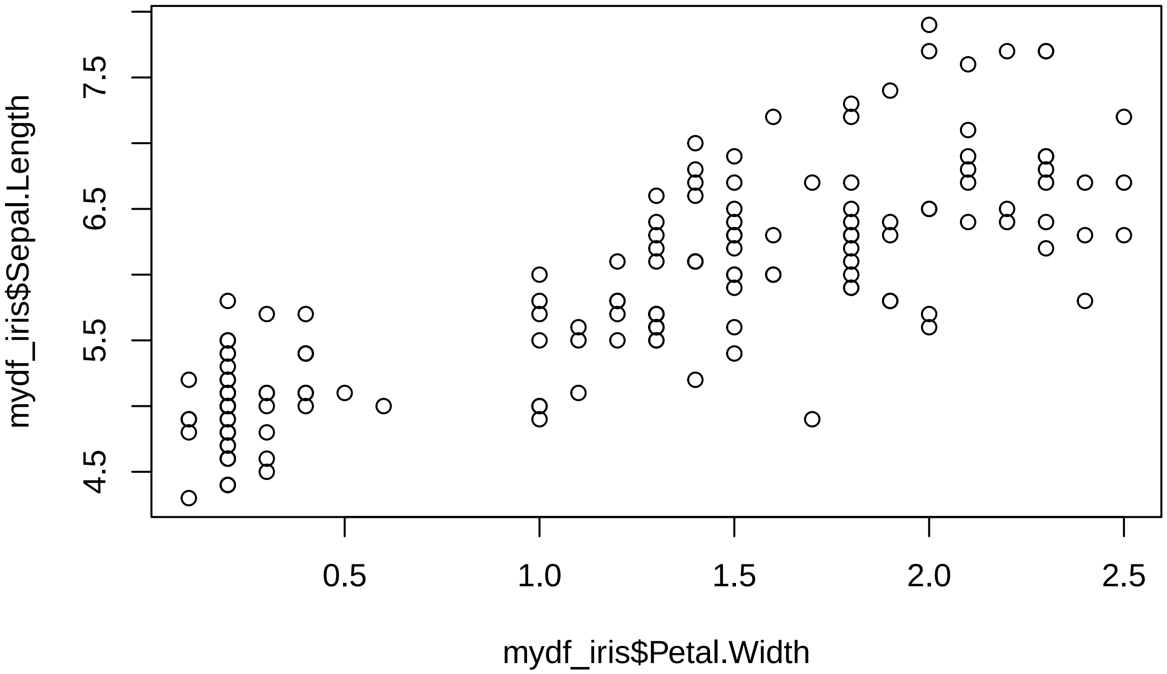
As a result, this is heteroscedastic

```
# Testing for heteroscendasticity by using linear
# x = Petal.Length, y = Sepal.Length
plength_slength <- lm(Petal.Length~Sepal.Length, data = mydf_iris)
bptest(plength_slength)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  plength_slength
## BP = 3.1239, df = 1, p-value = 0.07715
```

The p-value is 0.07715

```
# Set x = Petal.Width, y = Sepal.Length
plot(x= mydf_iris$Petal.Width, y=mydf_iris$Sepal.Length, type = "p")
```



As a result, this is heteroscedastic

```r
# Testing for heteroscendasticity by using linear
# x = Petal.Width, y = Sepal.Length
plength_slength <- lm(Petal.Width~Sepal.Length, data = mydf_iris)
bptest(plength_slength)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  plength_slength
## BP = 0.65185, df = 1, p-value = 0.4195
```

The p-value is 0.4195

Reference:

Frost, J.(n.d). Heteroscedasticity in Regression Analysis. Retrieved on February 08, 2022. Available at: https://statisticsbyjim.com/regression/heteroscedasticity-regression/.

Kurnicki, T.(n.d). Learn R. By coding.