```
> ########################################################
> # Homework Coding Questions 2                          #
> ########################################################
>
> ########################################################
> # Exercise 6.7 Question 3 - Iris Dataset               #
> ########################################################
>
> #installing the rpart package to access datasets
> install.packages("rpart")
>
> #Calling up the Iris dataset
> iris
> #Saving the data in a new variable iris_df
> iris_df <- iris
>
> #Creating a variable to store 0/1 labels for species
> #Converting species to a factor and numeric in order to use gsub to
create labels
> iris_df$Group <- as.numeric(as.factor(iris_df$Species))
>
> #Using gsub to replace the factor levels with labels 0 and 1
> iris_df$Group <- gsub(1, 0, iris_df$Group) #Replace 1 with 0 in the
iris_df, group column
> iris_df$Group <- gsub(2, 0, iris_df$Group) #Replace 2 with 0 in the
iris_df, group column
> iris_df$Group <- gsub(3, 1, iris_df$Group) #Replace 3 with 1 in the
iris_df, group column
>
> #converting entire variable into a numeric type
> iris_df$Group <- as.numeric(iris_df$Group)
>
> #3b - Building a regression model to predict the observation being
Virginica
> iris_reg <- glm(Group ~
Sepal.Length+Sepal.Width+Petal.Length+Petal.Width,
+                  data = iris_df, family= "binomial")
>
> exp(-2.465)-1 # getting change in odds for the first coefficient
[1] -0.9149912
> exp(-6.681)-1 # getting change in odds for the second coefficient
[1] -0.9987455
> exp(9.429)-1 # getting change in odds for the third coefficient
[1] 12443.08
> exp(18.286)-1 # getting change in odds for the fourth coefficient
[1] 87399489
>
> #Printing the summary for the regression model
> summary(iris_reg)

Call:
glm(formula = Group ~ Sepal.Length + Sepal.Width + Petal.Length +
    Petal.Width, family = "binomial", data = iris_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 summary shows that only Petal.Length has a p-value that is less
than 0.05 (0.0465), which makes it the only statistically significant
coefficient in the model.All other coefficients can still be used to
explain the model and calculation even though they are statistically
insignificant.
> #Predicting the probability of Virginica using all the data in the
dataframe
> predict(iris_reg, iris_df , type="response")
>
> #3c - Calculating the probability of a new plant being virginica using
the given parameters
> #Creating a function that calculates the probabilty, holding the
intercepts constant
> viriginica_success <- function (sepal_length, sepal_width,
petal_length, petal_width){
+   calc <- -42.638-2.645*(sepal_length)-6.681*(sepal_width)+
+     9.429*(petal_length)+18.286*(petal_width)
+   return(calc) #this returns the calculated value based on the user
inputs
+ }#Closing the function
>
> #Calling the function to calculate the probability of Virginica where:
> #sepal_length = 9
> #sepal_width = 5
> #petal_length = 10
> #petal_width = 7
>
> viriginica_success(9,5,10,7)
[1] 122.444
>
> #Calculating the probabilty of success, 1, for Virginica
> #prob_succ <- 1/(1+1/exp(viriginica_success(9,5,10,7)))
>
> exp(viriginica_success(9,5,10,7))/(1+exp(viriginica_success(9,5,10,7)))
[1] 1
>
> #This code has been successfully tested by user and results
> # probability value of 1
>
> ########################################################
```

```
> # Exercise 6.7 Question 4 - Kyphosis Dataset          #
> ######################################################
> #Calling the rpart library
> library(rpart)
>
> #Saving the kyphosis data set as a dataset
> kyphosis_df <- kyphosis
> #Inspecting the data
> View(kyphosis_df)
>
> #Using gsub function to replace absent and present with the new labels,
0/1
> #Replace "absent" with 0 in the kyphosis_df, kyphosis column
> kyphosis_df$Kyphosis <- gsub("absent",0, kyphosis_df$Kyphosis)
> #Replace "present" with 1 in the kyphosis_df, kyphosis column
> kyphosis_df$Kyphosis <- gsub("present",1, kyphosis_df$Kyphosis)
>
> #Converting the kyphosis variable to a numberic
> kyphosis_df$Kyphosis <- as.numeric(kyphosis_df$Kyphosis)
>
>
> #Building a regression model
> present_prob <- glm(Kyphosis ~ Age+Number+Start,
+                     data = kyphosis_df, family = "binomial")
>
> exp(0.010930)-1 # getting change in odds for the first coefficient
[1] 0.01098995
> #10.9% change in odds for success can be expected for every one unit
change in kyphosis
> exp(0.410601)-1 # getting change in odds for the second coefficient
[1] 0.5077237
> #50.7% change in odds for success can be expected for every one unit
change in kyphosis
> exp(-2.06510)-1 # getting change in odds for the third coefficient
[1] -0.8731944
> #87.3% change in odds for success can be expected for every one unit
change in kyphosis
>
>
> #Printing the summary of the regression model
> summary(present_prob)

Call:
glm(formula = Kyphosis ~ Age + Number + Start, family = "binomial",
    data = kyphosis_df)

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
>
> # Based on the summary, only Start has a p-value that is less than 0.05
(0.00229), this makes it the only statistically significant coefficient
in the model.All other coefficients can still be used to explain the
model for and calculation even though they are statistically
insignificant and should be equal to zero.
>
> #Using the predict function to caculate the probabilty
> predicted_kyphosis <- predict(present_prob, kyphosis_df, type =
"response")
>
> #calculating the probability of success for present in the dataset
> present_success <- function(Age, Number, Start){
+   success <- -2.036934+0.010930*(Age)+0.410601*(Number)-
0.206510*(Start)
+   return(success)
+ }#closing User Defined Function
>
> #Calling the function to calculate the probability  where:
> #Age = 50
> #Number = 5
> #Start = 10
>
> present_success(50,5,10)
[1] -1.502529
> exp(present_success(50,5,10))/(1+exp(present_success(50,5,10)))
[1] 0.1820486
> present_succ <- 1/(1+1/exp(present_success(50,5,10)))
>
> #The probability of sucess of kyphosis present is 0.18
>
> ############################################################
> # Question 5 - Homoscedactic/Heteroscedastic Test          #
> ############################################################
> #Installing package to fit linear regression
> install.packages("lmtest")
> #loading the lmtest to run the linear regression
> library(lmtest)
>
> #Calculating the linear regression of each variable pair
> iris_sepal_1 <- lm(Sepal.Length ~ Sepal.Width, iris_df)
> summary(iris_sepal_1)

Call:
lm(formula = Sepal.Length ~ Sepal.Width, data = iris_df)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5561 -0.6333 -0.1120  0.5579  2.2226

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.5262     0.4789   13.63   <2e-16 ***
Sepal.Width  -0.2234     0.1551   -1.44    0.152
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8251 on 148 degrees of freedom
Multiple R-squared:  0.01382,     Adjusted R-squared:  0.007159
F-statistic: 2.074 on 1 and 148 DF,  p-value: 0.1519

>
> #This model is not statistically significant as the p-value is higher
than 0.05
>
> #Pair 2
> iris_sepal_2 <- lm(Sepal.Length ~ Petal.Length, iris_df)
> summary(iris_sepal_2)

Call:
lm(formula = Sepal.Length ~ Petal.Length, data = iris_df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.24675 -0.29657 -0.01515  0.27676  1.00269

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.30660    0.07839   54.94   <2e-16 ***
Petal.Length  0.40892    0.01889   21.65   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4071 on 148 degrees of freedom
Multiple R-squared:   0.76,  Adjusted R-squared:  0.7583
F-statistic: 468.6 on 1 and 148 DF,  p-value: < 2.2e-16

>
> #This model is statistically significant as the p-value is lower than
0.05
>
> #Pair 3
> iris_sepal_3 <- lm(Sepal.Length ~ Petal.Width, iris_df)
> summary(iris_sepal_3)

Call:
lm(formula = Sepal.Length ~ Petal.Width, data = iris_df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.38822 -0.29358 -0.04393  0.26429  1.34521

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.77763    0.07293   65.51   <2e-16 ***
Petal.Width  0.88858    0.05137   17.30   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.478 on 148 degrees of freedom
Multiple R-squared:  0.669,  Adjusted R-squared:  0.6668
F-statistic: 299.2 on 1 and 148 DF,  p-value: < 2.2e-16

>
> #This model is also statistically significant as the p-value is lower
than 0.05
>
>
> #Calculating the bptest of each variable pair to test for
homoscedasticity or heteroscedasticcity
> bptest(iris_sepal_1)

      studentized Breusch-Pagan test

data:  iris_sepal_1
BP = 0.78243, df = 1, p-value = 0.3764

>
> bptest(iris_sepal_2)

      studentized Breusch-Pagan test

data:  iris_sepal_2
BP = 2.7561, df = 1, p-value = 0.09688

>
> bptest(iris_sepal_3)

      studentized Breusch-Pagan test

data:  iris_sepal_3
BP = 12.357, df = 1, p-value = 0.0004393

> #Plotting scatter plots
> plot(x= iris_df$Sepal.Width, y= iris$Sepal.Length, type= "p")
> # For this pair, we can conclude that heteroscedasticity is present
because the p-value of the test
> # is lower than 0.05.
> plot(x= iris_df$Petal.Length, y= iris$Sepal.Length, type= "p")
> # For this pair, we can conclude that homoscedasticity is present. the
error variances are equal.
> plot(x= iris_df$Petal.Width, y= iris$Sepal.Length, type= "p")
> # For this pair, we can conclude that heteroscedasticity is present
because the p-value of the test
> # is lower than 0.05.
> #the script has minimal errors
> #The following warning message is displayed whenever the iris_reg is
run
> #Warning message:
> #glm.fit: fitted probabilities numerically 0 or 1 occurred
```