# Homework 3 Solutions - MATH 4322

### Instructions

1. Due date: February 26, 2024
2. Scan or Type your answers and submit only one file. (If you submit several files only the recent one uploaded will be graded).
3. Preferably save your file as PDF before uploading. Submit in Canvas under Homework 3.
4. These questions are from *An Introduction to Statistical Learning with Applications in R* by James, et. al., chapter 4.

### Problem 1

Suppose we collect data for a group of students in a statistics class with variables $X_1$ = hours studied, $X_2$ =undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$.

(a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.

(b) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?

**Answer**

(a) The model is:

$$p(\hat{X}) = \frac{exp(-6 + 0.05 \times \text{hours} + \text{GPA})}{1 + exp(-6 + 0.05 \times \text{hours} + \text{GPA})}$$

Thus $p(\hat{X}) = 0.3775$.

(b) Use this as the model:

$$log\left(\frac{p(X)}{1 - p(X)}\right) = -6 + 0.05h + 3.5$$

$$\log(1) = -2.5 + 0.05h$$
$$2.5 = 0.05h$$
$$h = 50$$

## Problem 2

This question should be answered using the `Weekly` data set, which is part of the `ISLR2` package. This data set consists of percentage returns for the S&P 500 stock index over 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010. For each week, we have recorded the percentage returns for each of the five previous trading weeks, `Lag1` through `Lag5`. We have also recorded `Volume` (the average number of shares traded on the previous week, in billions), `Today` (the percentage return for the week in question) and `Direction` (whether the market was Up or Down on this week). Our goal is to predict `Direction` (a qualitative response) using the other features.

(a) Produce some numerical and graphical summaries of the `Weekly` data. Do there appear to be any patterns?

**Answer**

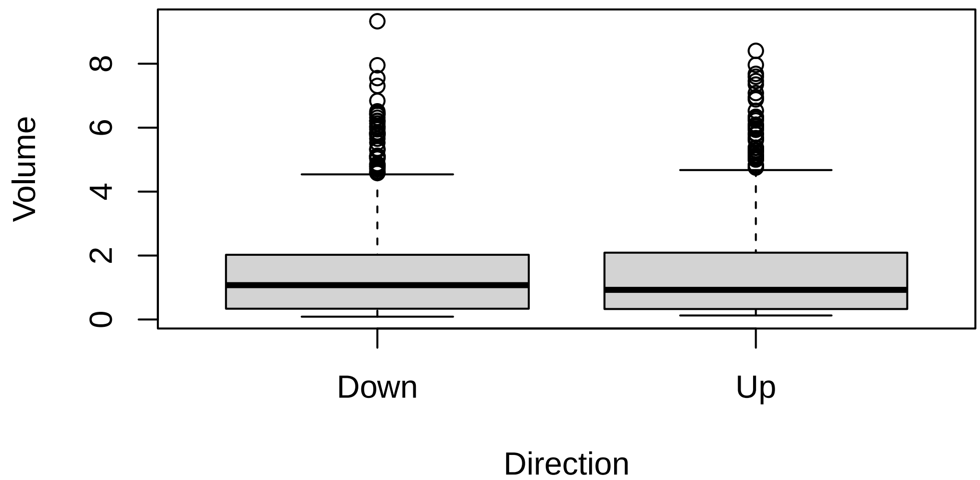
```r
library(ISLR2)
summary(Weekly)
```

```
      Year           Lag1               Lag2               Lag3         
 Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950  
 1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580  
 Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.2410  
 Mean   :2000   Mean   :  0.1506   Mean   :  0.1511   Mean   :  0.1472  
 3rd Qu.:2005   3rd Qu.:  1.4050   3rd Qu.:  1.4090   3rd Qu.:  1.4090  
 Max.   :2010   Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260  
      Lag4               Lag5              Volume            Today         
 Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747   Min.   :-18.1950  
 1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202   1st Qu.: -1.1540  
 Median :  0.2380   Median :  0.2340   Median :1.00268   Median :  0.2410  
 Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462   Mean   :  0.1499  
 3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373   3rd Qu.:  1.4050  
 Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821   Max.   : 12.0260  
 Direction 
 Down:484  
 Up  :605  
```

```r
cor(Weekly[,c(2,3,4,5,6,7)])
```

```
              Lag1        Lag2        Lag3        Lag4         Lag5
Lag1    1.000000000 -0.07485305  0.05863568 -0.07127388 -0.008183096
Lag2   -0.074853051  1.00000000 -0.07572091  0.05838153 -0.072499482
Lag3    0.058635682 -0.07572091  1.00000000 -0.07539587  0.060657175
Lag4   -0.071273876  0.05838153 -0.07539587  1.00000000 -0.075675027
Lag5   -0.008183096 -0.07249948  0.06065717 -0.07567503  1.000000000
Volume -0.064951313 -0.08551314 -0.06928771 -0.06107462 -0.058517414
             Volume
Lag1   -0.06495131
Lag2   -0.08551314
Lag3   -0.06928771
Lag4   -0.06107462
Lag5   -0.05851741
Volume  1.00000000
```

```r
boxplot(Volume ~ Direction, data = Weekly)
```




There really is no correlation between the lag variables, and looking at the volume weather the direction is up or down is about the same.

(b) Use the full data set to perform a logistic regression with `Direction` as the response and the five lag variables plus `Volume` as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

```
week.fit = glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume,
               data = Weekly,family = "binomial")
summary(week.fit)
```

```
Call:
glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
    Volume, family = "binomial", data = Weekly)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.26686    0.08593   3.106   0.0019 **
Lag1        -0.04127    0.02641  -1.563   0.1181
Lag2         0.05844    0.02686   2.175   0.0296 *
Lag3        -0.01606    0.02666  -0.602   0.5469
```

```
Lag4        -0.02779    0.02646  -1.050   0.2937
Lag5        -0.01447    0.02638  -0.549   0.5833
Volume      -0.02274    0.03690  -0.616   0.5377
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1496.2  on 1088  degrees of freedom
Residual deviance: 1486.4  on 1082  degrees of freedom
AIC: 1500.4

Number of Fisher Scoring iterations: 4
```

It appears that only Lag2 is significant in predicting if the stock will go "Up".

(c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

**Answer**

```
percent.w = predict.glm(week.fit,type = "response")
predict.w = ifelse(percent.w < 0.5,"Down","Up")
table(predict.w,Weekly$Direction)
```

```
predict.w Down  Up
     Down   54  48
     Up    430 557
```

Fraction of correct predictions: $\frac{611}{1089} = 0.5611$
This is not a good model to predict if a stock will go up or down.

(d) Now fit the logistic regression model using a training data period from 1990 to 2008, with `Lag2` as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).

**Answer**

```
train.w = Weekly[Weekly$Year <2009,]
test.w = Weekly[Weekly$Year > 2008,]
week.fit2 = glm(Direction ~ Lag2, data = train.w, family = "binomial")
```

```r
summary(week.fit2)
```

```

Call:
glm(formula = Direction ~ Lag2, family = "binomial", data = train.w)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.20326    0.06428   3.162  0.00157 **
Lag2         0.05810    0.02870   2.024  0.04298 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1354.7  on 984  degrees of freedom
Residual deviance: 1350.5  on 983  degrees of freedom
AIC: 1354.5

Number of Fisher Scoring iterations: 4
```

```r
percent.w = predict.glm(week.fit2,newdata = test.w, type = "response")
predict.w = ifelse(percent.w < 0.5,"Down","Up")
table(predict.w,test.w$Direction)
```

```

predict.w Down Up
     Down    9  5
     Up     34 56
```

Fraction of correct predictions: $\frac{65}{104} = 0.625$

## Problem 3

This problem involves writing functions.

(a) Write a function, `Power()`, that prints out the result of raising 2 to the 3rd power. In other words, your function should compute $2^3$ and print out the results. *Hint: Recall that* $x^a$ raises $x$ to the power $a$. Use the `print()` function to output the result.

**Answer**

```
Power = function(x) {
  print(x^3)
}
Power(2)
```

```
[1] 8
```

(b) Create a new function, `Power2()`, that allows you to pass any two numbers, $x$ and $a$, and prints out the value of $x^a$. You can do this by beginning your function with the line

```
Power2 <- function(x, a) {
```

You should be able to call your function by entering, for instance,

```
Power2(3, 8)
```

on the command line. This should output the value of $3^8$, namely, 6,561.

**Answer**

```
Power2 = function(x,a) {
  print(x^a)
}
Power2(3,8)
```

```
[1] 6561
```

**Answer**

```
Power2(10,3)
```

```
[1] 1000
```

```
Power2(8,17)
```

```
[1] 2.2518e+15
```

```
Power2(131,3)
```

```
[1] 2248091
```

(d) Now create a new function, `Power3()`, that actually returns the result $x^a$ as an `R` object, rather than simply printing it to the screen. That is, if you store the value $x^a$ in an object called `result` within your function, then you can simply `return()` this result, using the following line:

```
return(result)
```

The line above should be the last line in your function, before the } symbol.

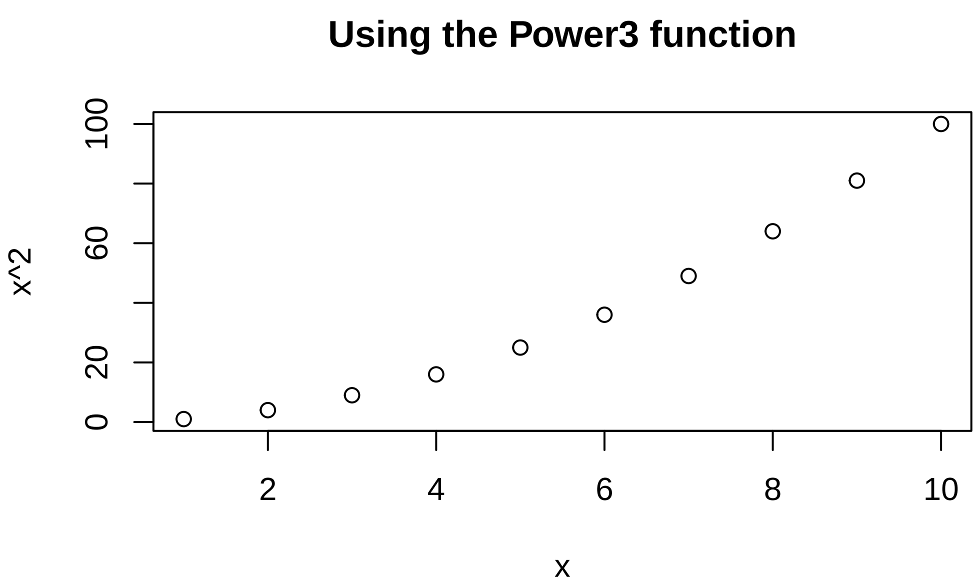
**Answer**

```
Power3 = function(x,a) {
  p3 = x^a
  return(p3)
}
```

(e) Now using the `Power3()` function, create a plot of $f(x) = x^2$. The $x$-axis should display a range of integers from 1 to 10, and the $y$-axis should display $x^2$. Label the axes appropriately, and use an appropriate title for the figure.

**Answer**

```
plot(1:10,Power3(1:10,2),xlab = "x",ylab = "x^2",main = "Using the Power3 function")
```



(f) Create a function, `PlotPower()`, that allows you to create a plot of $x$ against $x^a$ for a fixed $a$ and for a range of values of $x$. For instance, if you call

```
PlotPower(1:10, 3)
```

then a plot should be created with an $x$-axis taking on values $1, 2, \ldots, 10$, and a $y$-axis taking on values $1^3, 2^3, \ldots, 10^3$.

**Answer**

```
PlotPower = function(x,a) {
  plot(x,Power3(x,a),xlab = "x",ylab = "x^a",main = "")
}
PlotPower(1:10,3)
```