Asssignment4

Ziwei Li

3/7/2020

This question should be answered using the Weekly data set, which is part of the ISLR package. This data is similar in nature to the Smarket data from this chapter's lab, except that it contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010. Write a data analysis report addressing the following problems.

```
library(ISLR)
```

```
names(Weekly)
```

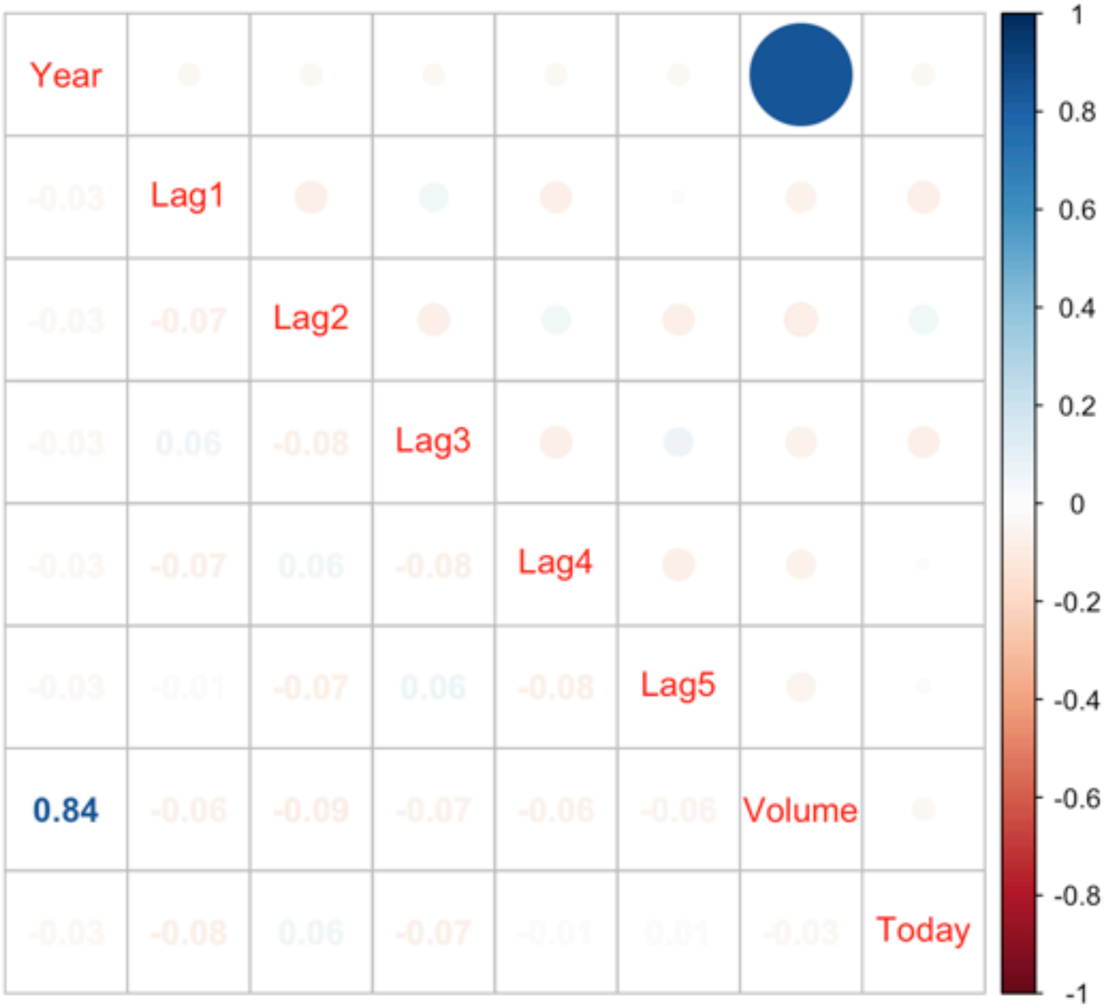
```
## [1] "Year"      "Lag1"      "Lag2"      "Lag3"      "Lag4"      "Lag5"
## [7] "Volume"    "Today"     "Direction"
```

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

```
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       
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747  
##  1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202  
##  Median :  0.2380   Median :  0.2340   Median :1.00268  
##  Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462  
##  3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373  
##  Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821  
##      Today          Direction 
##  Min.   :-18.1950   Down:484  
##  1st Qu.: -1.1540   Up  :605  
##  Median :  0.2410             
##  Mean   :  0.1499             
##  3rd Qu.:  1.4050             
##  Max.   : 12.0260             
```

```
corrplot::corrplot.mixed(cor(Weekly[,-9]))
```




```
# According to the corrplot, the relationship between year and volume seems to be str
ong. And there are no other patterns seem to be obvious.
```

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

```
mod1 = glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume,data=Weekly,family=binomial)
summary(mod1)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = Weekly)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6949  -1.2565   0.9913   1.0849   1.4579
##
## 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
```

```
# Lag2 seems to be significant, because the p-value of Lag2 is less than 0.05.
```

(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.

```
mod.prob = predict(mod1,type='response')
mod.pred = rep("Down",nrow(Weekly))
mod.pred[mod.prob > 0.5] = 'Up'
```

```
table(mod.pred,Weekly$Direction)
```

```
##
## mod.pred Down  Up
##     Down   54  48
##     Up    430 557
```

```r
(54+557)/nrow(Weekly)
```

```
## [1] 0.5610652
```

```r
# False positive is the type I error.
# False negative is the type II error.
```

(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).

```r
train = (Weekly$Year < 2009)
test = Weekly[!train,]
```

```r
mod2 = glm(Direction~Lag2,data=Weekly[train,],family=binomial)
mod2.probs = predict(mod2,test,type='response')
```

```r
mod2.pred = rep("Down",nrow(test))
mod2.pred[mod2.probs>0.5] = 'Up'
table(mod2.pred,test$Direction)
```

```
##
## mod2.pred Down Up
##      Down    9  5
##      Up     34 56
```

```r
(9+56)/nrow(test)
```

```
## [1] 0.625
```

(e) Repeat (d) using LDA.

```r
library(MASS)
```

```r
mod.lda = lda(Direction~Lag2,data=Weekly[train,])
mod.lda.pred = predict(mod.lda,test)
table(mod.lda.pred$class,test$Direction)
```

```
##
##         Down Up
##   Down     9  5
##   Up      34 56
```

```
(9+56)/nrow(test)
```

```
## [1] 0.625
```

Repeat (d) using QDA.

```
mod.qda = qda(Direction~Lag2,data=Weekly[train,])
mod.qda.pred = predict(mod.qda,test)
table(mod.qda.pred$class,test$Direction)
```

```
##
##         Down Up
##   Down     0  0
##   Up      43 61
```

```
(61+0)/nrow(test)
```

```
## [1] 0.5865385
```

(g) Which of these methods appears to provide the best results on this data?

```
# The LDA and logistic regression both have the same result(0.625), which is bigger t
han the QDA's test error(0.59)
```

(h) Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held out data.

```
# As for logistic regression, we may consdier the interaction between lag2 and lag4.
mod12 = glm(Direction~Lag2+Lag4+Lag2*Lag4,data=Weekly[train,],family=binomial)
mod12.prob = predict(mod12,test,type='response')
mod12.pred = rep('Down',length(mod12.prob))
mod12.pred[mod12.prob>0.5] = 'Up'
table(mod12.pred,test$Direction)
```

```
##
## mod12.pred Down Up
##       Down    4  4
##       Up     39 57
```

```
(4+39)/nrow(test)
```

```
## [1] 0.4134615
```

```
# As for LDA, we still consider the interaction between lag2 and lag4
mod.lda2 = lda(Direction~Lag2+Lag4+Lag2*Lag4,data=Weekly[train,])
mod.lda.pred2 = predict(mod.lda2,test)
table(mod.lda.pred2$class,test$Direction)
```

```
##
##        Down Up
##   Down    4  4
##   Up     39 57
```

```
(4+39)/nrow(test)
```

```
## [1] 0.4134615
```

```
# As for QDA, we still consider the interaction beween Lag2 and Lag4
mod.qda2 = qda(Direction~Lag2+Lag4+Lag2*Lag4,data=Weekly[train,])
mod.qda.pred2 = predict(mod.qda2,test)
table(mod.qda.pred2$class,test$Direction)
```

```
##
##        Down Up
##   Down   10 14
##   Up     33 47
```

```
(33+14)/nrow(test)
```

```
## [1] 0.4519231
```

```
# To sum up, if we use Lag2 and Lag4 and their interactions as predictiors, the LDA a
nd Logistic regression's test error is the same and also lower than QDA's test error.
Therefore, LDA and Logistic perform better than QDA.
```

In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the Auto data set. Write a data analysis report addressing the following problems.

```
summary(Auto)
```

```
##       mpg          cylinders      displacement     horsepower
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0
##
##      weight      acceleration        year           origin
##  Min.   :1613   Min.   : 8.00   Min.   :70.00   Min.   :1.000
##  1st Qu.:2225   1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000
##  Median :2804   Median :15.50   Median :76.00   Median :1.000
##  Mean   :2978   Mean   :15.54   Mean   :75.98   Mean   :1.577
##  3rd Qu.:3615   3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000
##  Max.   :5140   Max.   :24.80   Max.   :82.00   Max.   :3.000
##
##                  name
##  amc matador       :  5
##  ford pinto        :  5
##  toyota corolla    :  5
##  amc gremlin       :  4
##  amc hornet        :  4
##  chevrolet chevette:  4
##  (Other)           :365
```

(a) Create a binary variable, mpg01, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median.

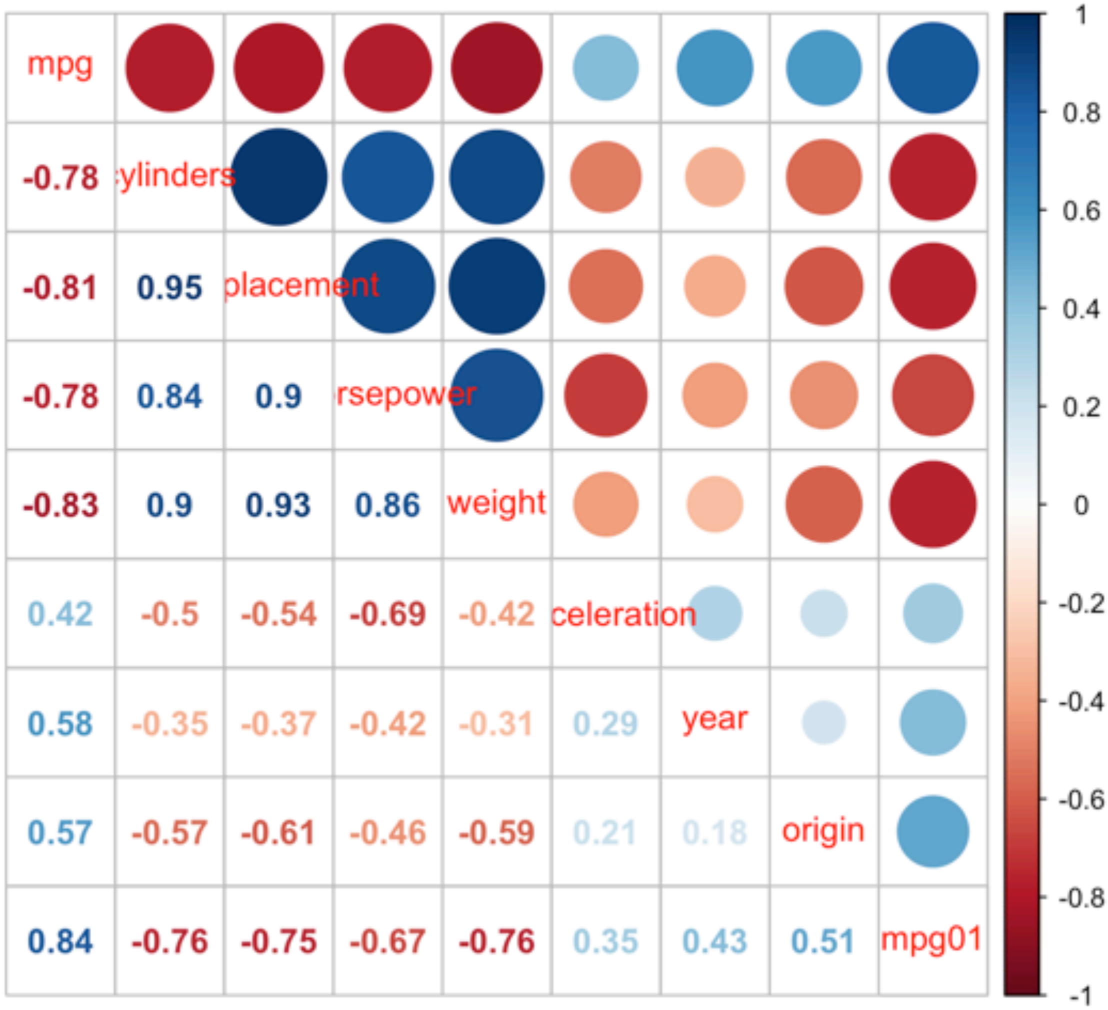
```
Auto2 = Auto
Auto2$mpg01 = ifelse(Auto2$mpg > median(Auto2$mpg),1,0)
```

```
summary(Auto2)
```

```
##       mpg          cylinders      displacement     horsepower
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0
##
##      weight      acceleration        year           origin
##  Min.   :1613   Min.   : 8.00   Min.   :70.00   Min.   :1.000
##  1st Qu.:2225   1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000
##  Median :2804   Median :15.50   Median :76.00   Median :1.000
##  Mean   :2978   Mean   :15.54   Mean   :75.98   Mean   :1.577
##  3rd Qu.:3615   3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000
##  Max.   :5140   Max.   :24.80   Max.   :82.00   Max.   :3.000
##
##                 name         mpg01
##  amc matador       :  5   Min.   :0.0
##  ford pinto        :  5   1st Qu.:0.0
##  toyota corolla    :  5   Median :0.5
##  amc gremlin       :  4   Mean   :0.5
##  amc hornet        :  4   3rd Qu.:1.0
##  chevrolet chevette:  4   Max.   :1.0
##  (Other)           :365
```

(b) Explore the data graphically in order to investigate the association between mgp01 and the other features. Which of the other features seem most likely to be useful in predicting mpg01? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

```
corrplot::corrplot.mixed(cor(Auto2[,c(-9)]))
```




```
# It might be useful to use cylinders, displacement, weight, and horsepower to predic
t mpg01 because their correlations are higher compared to the rest of the predcitors.
```

(c) Split the data into a training set and a test set.

```
library(caTools)
sample = sample.split(Auto2,SplitRatio = 0.75)
```

```
train = subset(Auto2,sample ==TRUE)
test = subset(Auto2,sample == FALSE)
```

(d) Perform LDA on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?

```
library(MASS)
mod.lda2 = lda(mpg01~cylinders+displacement+weight+horsepower,data=train)

lda.pred = predict(mod.lda2,test,type='response')
table(lda.pred$class,test$mpg01)
```

```
##
##      0  1
##    0 49  3
##    1 11 54
```

```
(8+2)/nrow(test)
```

```
## [1] 0.08547009
```

```
# Test error is 0.08547009
```

(e) Perform QDA on the training data in order to predict mpg01 using the variables that seemed most

associated with mpg01 in (b). What is the test error of the model obtained?

```
mod.qda2 = qda(mpg01~cylinders+displacement+weight+horsepower,data=train)

qda.pred = predict(mod.lda2,test,type='response')
table(qda.pred$class,test$mpg01)
```

```
##
##      0  1
##    0 49  3
##    1 11 54
```

```
(8+2)/nrow(test)
```

```
## [1] 0.08547009
```

f) Perform logistic regression on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?

```
mod.log = glm(mpg01~cylinders+displacement+weight+horsepower,data=train,family=binomi
al)
log.pred = predict(mod.log,test,type='response')
```

```
mod.log.pred = rep(0,nrow(test))
mod.log.pred[log.pred>0.5] = 1
table(mod.log.pred,test$mpg01)
```

```
##
## mod.log.pred  0  1
##            0 51  5
##            1  9 52
```

```
(6+4)/nrow(test)
```

```
## [1] 0.08547009
```

```
# The teset error is 0.08547009
```

1. (PCR and PLS, *10 pt* **Bonus**) Are the following sentences about principal component regression (PCR) and partial least square (PLS) True or False? Briefly justify your answer.

   (i) Both PCR and PLS come up with orthogonal features.

   (ii) Let $\boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)}, \cdots, \boldsymbol{Z}^{(p)}$ be the features obtained by PLS. For an intermediate $k < p$, we fit a regression model $\boldsymbol{Y}$ on $\boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)}, \cdots, \boldsymbol{Z}^{(k)}$, and obtain a predicted response $\hat{\boldsymbol{Y}}^{(k)}$ on the training set. Then $\hat{\boldsymbol{Y}}^{(k)}$ is orthogonal to the subsequent features $\boldsymbol{Z}^{(k+1)}, \boldsymbol{Z}^{(k+2)}, \cdots, \boldsymbol{Z}^{(p)}$. Therefore, to compute $\hat{\boldsymbol{Y}}^{(k+1)}$ (the predicted response based on $\boldsymbol{Z}^{(1)}, \boldsymbol{Z}^{(2)}, \cdots, \boldsymbol{Z}^{(k)}, \boldsymbol{Z}^{(k+1)}$), we can first regress $\boldsymbol{Y}$ on $\boldsymbol{Z}^{(k+1)}$ only and obtain $\hat{\boldsymbol{Y}}(\boldsymbol{Z}^{(k+1)})$ (the predicted response based on $\boldsymbol{Z}^{(k+1)}$ only), and then let
$$\hat{\boldsymbol{Y}}^{(k+1)} \leftarrow \hat{\boldsymbol{Y}}^{(k)} + \hat{\boldsymbol{Y}}(\boldsymbol{Z}^{(k+1)}).$$

   (iii) The first feature from PLS is more predictive towards the response in the training set than that from PCR.

   (iv) In the procedures of constructing features from PCR and PLS, the earlier a feature is included in the regression model, the faster the training R^2 increases.

   (v) Using the features from PCR and PLS has lower training errors and test errors than those of the original linear model.

   Hint. Read more in Section 6.3 from the Textbook before getting started.

(i)
False.
Two principal components Z1 and Z2 of the Principal Component Analysis have the zero-correlation condition. In other words, the second principal component direction must be perpendicular, or perpendicular orthogonal, to the first principal component direction.
Partial Least Squares, use the annotated label to maximize inter-class variance. Principal components are pairwise orthogonal. Principal components are focus on maximize correlation.

2. (Logistic Regression, *10 pt*) Suppose we have observation pairs $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ where $\boldsymbol{X}_i \in \mathbb{R}^p$, $Y_i \in \{0,1\}$.

   (i) Suppose $Y_i \sim \textbf{Binomial}(p_i)$ where $p_i \in [0,1]$ is a parameter. Write down the probability mass function (PMF) of Y_i. What's the expectation of Y_i?

   (ii) In the terminology of generalized linear model (GLM), there is a link function $g : \mathbb{R} \to \mathbb{R}$ that establishes the relationship between the expectations of Y_i's and the linear combinations of $\boldsymbol{X}_i$'s

$$g(\mathbb{E}Y_i) = \boldsymbol{X}_i^T \boldsymbol{\beta} \quad (\forall 1 \leqslant i \leqslant n)$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unknown coefficient parameter for the linear part.

Now for the Logistic regression problem, g is the **log-odds/logit** link

$$g(\mu) := \log\left(\frac{\mu}{1-\mu}\right). \quad (\mu \in [0,1])$$

First write down the log-likelihood function of Y_i in terms of parameter $\boldsymbol{\beta}$ given the observation $(\boldsymbol{X}_i, Y_i)$. Then write down the joint log-likelihood.

Hint. In the PMF of Y_i, replace p_i by some quantities in terms of $\boldsymbol{X}_i^T\boldsymbol{\beta}$ through the link between $\mathbb{E}Y_i$ and $\boldsymbol{X}_i^T\boldsymbol{\beta}$.

(i)

$$P(x; p, n) = \binom{n}{x}(p)^x(1-p)^{n-x}$$
$$E(Y) = np$$

(ii)

(ii) $P(X_i) = Pr(y_i = 1 | X_i) = \pi(X_i)$

$Pr(y_i = 0 | X_i) = 1 - p(X_i)$

$$\log \frac{P(X_i)}{1-P(X_i)} = \sum_{j=0}^{p} \beta_j X_{i,j}$$

$$\log \frac{p(X_i;\beta)}{1-p(X_i;\beta)} = X_i\beta$$

$$p(X_i;\beta) = \frac{e^{X_i\beta}}{1+e^{X_i\beta}}$$

$$1 - p(X_i;\beta) = \frac{1}{1+e^{X_i\beta}}$$

log-likelihood

$$l(\beta|x_i) = (1-y_i)\log[1-p(X_i;\beta)] + y_i \log p(X_i;\beta)$$
$$= \begin{cases} \log p(X_i;\beta) & \text{if } y_i = 1 \\ \log[1-p(X_i;\beta)] & \text{if } y_i = 0 \end{cases}$$

Joint-log-likelihood

$$L(\beta|X_1 \ldots X_n) = \sum_{i=1}^n l_i(\beta|X_i)$$
$$= \sum_{i=1}^n (1-y_i)\log[1-p(X_i;\beta)] + y_i \log p(X_i;\beta)$$
$$= \sum_{i=1}^n y_i \log \frac{p(X_i;\beta)}{1-p(X_i;\beta)} + \log[1-p(X_i;\beta)]$$
$$= \sum_{i=1}^n y_i X_i \beta - \log(1+e^{X_i\beta})$$

3. (Logistic Regression, Textbook 4.6, *15 pt*) 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) Provide an interpretation of each coefficient in the model. Note that β_0 corresponds to an additional intercept in the model.

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

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

(a)
β_1 is the coefficient of 'hours studied'. Exp(0.05) = 1.051271, meaning odds of earning an A are multiplied by 1.051271, with each 1 hour increase in 'hours studied'
β_2 is the coefficient of 'undergrad GPA'. Exp(1) = 2.718282, meaning odds of earning an A are multiplied by 2.718282, with each 1-unit increase in GPA.
(b)

$$p(x1 = 40, x2 = 3.5) = \frac{e^{-6+0.05*40+1*3.5}}{1+e^{-6+0.05*40+1*3.5}} = 0.3775407$$

(c)

$$p(x1 = X1, x2 = 3.5) = \frac{e^{-6+0.05*X1+1*3.5}}{1+e^{-6+0.05*X1+1*3.5}} = 0.5$$

$$e^{-6+0.05*X1+1*3.5} = 1$$

$$e^{-6+0.05*X1+1*3.5} = 1$$
$$0.05 * X1 = 2.5$$
$$X1 = 50$$

He needs 50 hours study to reach the 50% chance of getting an A.

4. (LDA and QDA, Textbook 4.5, *20 pt*) We now examine the differences between linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA).

   (a) If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set?

   (b) If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set?

   (c) In general, as the sample size n increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?

   (d) True or False: Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer.

(a)
If the Bayes decision boundary is linear, we may expect that LDA performs better on the training set and test set. Because the QDA will suffer from high variance without a corresponding decrease in bias.
(b)
If the Bays decision boundary is non-linear, we may expect the QDA will outperform the LDA on both training set and test set. Because QDA is more flexible and does not share a common covariance matrix we may expect a decrease in bias.
(c)
As sample size n increase, we may expect that QDA may outperform the LDA. Because as n gets large, reducing variance of classifier is not a major concern, or if the assumption of a common covariance matrix for the K classifiers clearly untenable. Besides, QDA is more flexible than LDA. Then, QDA may perform better than LDA.
(d)
False.
When the number of observations is small but with more predictors, QDA may lead to overfitting due to the flexibility of QDA. Therefore, when considering reducing the variance, it is better to use LDA.