Homework 3 - Modern Applied Statistics II

James Young

Packages: ISLR, ggplot2, GGally Collaborators: Resources: stackoverflow.com

Please do the following problems from the text book ISLR or written otherwise.

1. Question 4.7.1 pg 168

Using a little bit of algebra, prove that (4.2) is equivalent to (4.3). In other words, the logistic function representation and logit representation for the logistic regression model are equivalent.

If we say $v = e^{B_0+B_1x}$

$p(X) = \frac{v}{1+v}$

$\frac{1}{p(x)} = \frac{1+v}{v} = 1 + \frac{1}{v}$

$v = \frac{1}{\frac{1}{p(x)}-1} = \frac{1}{\frac{1-p(x)}{p(x)}} = \frac{p(x)}{1-p(x)}$

2. Question 4.7.10(a-d) pg 171

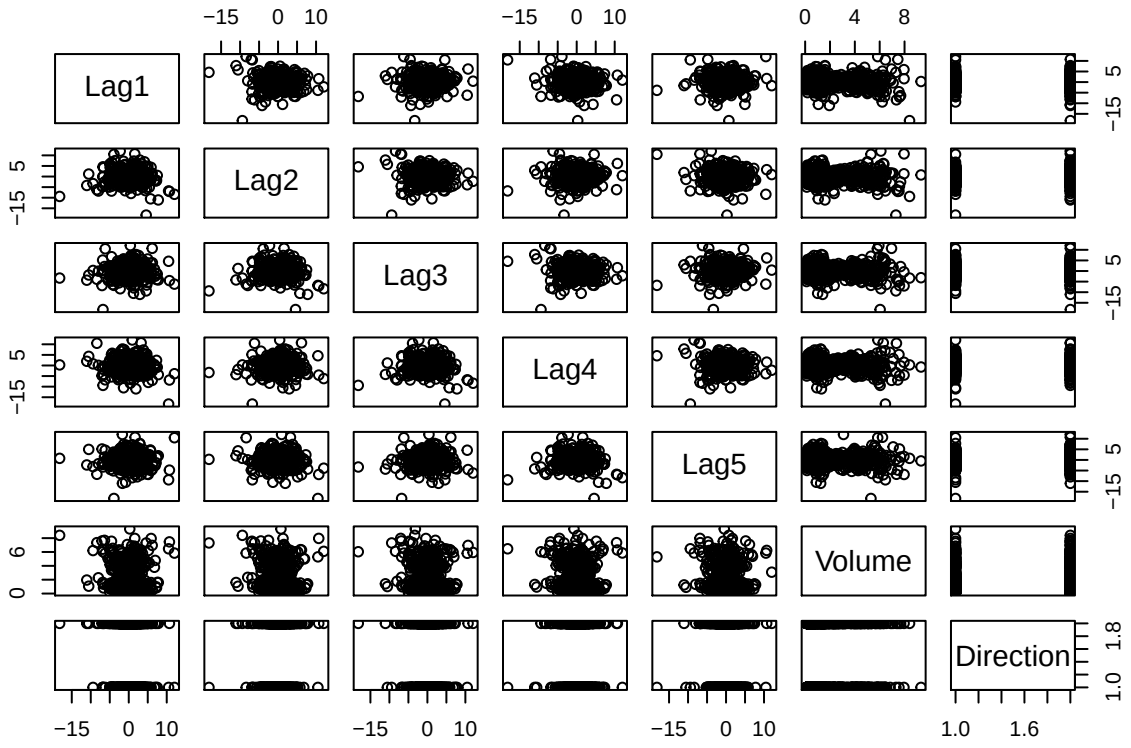
This question should be answered using the Weekly data set, which is part of the ISLR package. This data is similar in nature to the Smarket data from this chapter's lab, except that it contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

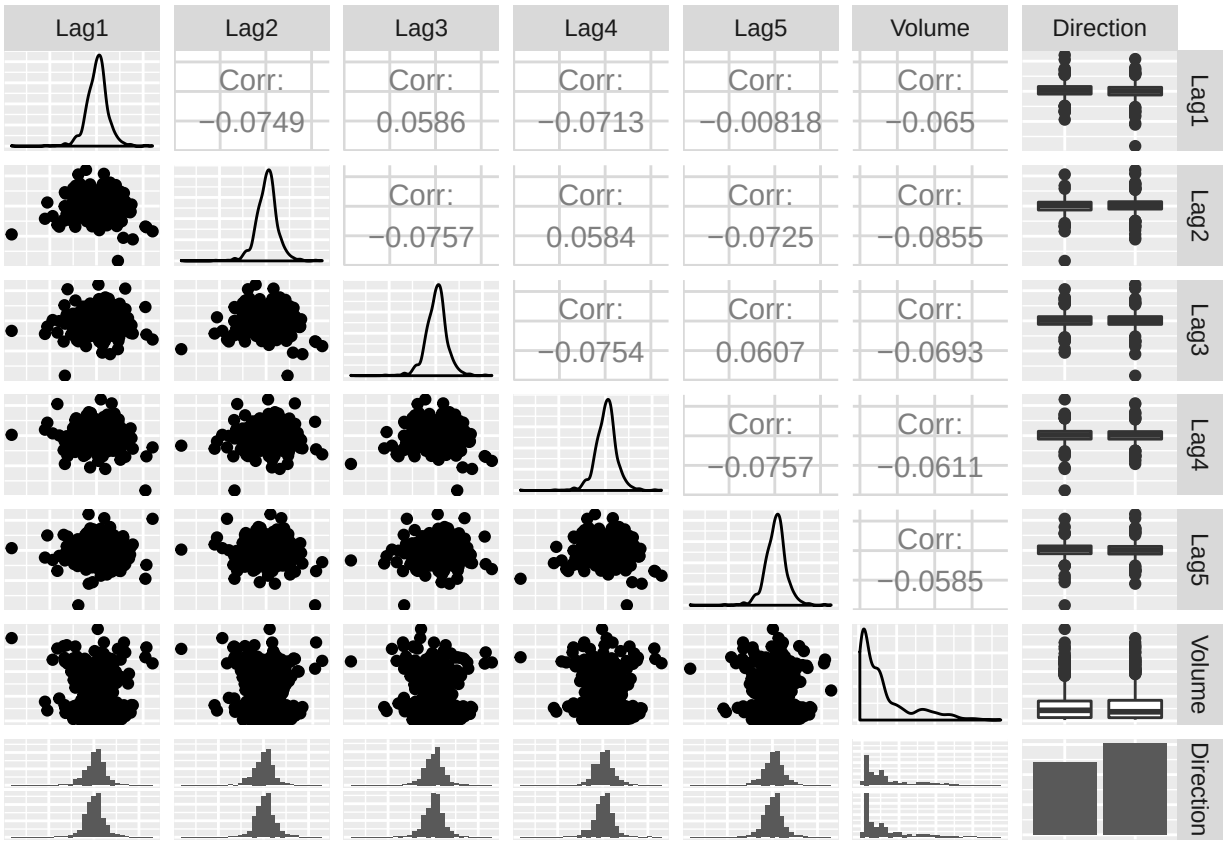
a) Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

I removed variables from the visualization that won't be used in the regression in order to make the remaining variables more visible. Based on the numerical summary and visualizations there doesn't seem to be any strong patterns. On the bright side we can also see there really is not colinearity, either.

```
##       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
##
##
##
##
```





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?

The only predictor variable that seems to be statistically significant at alpha = 0.05 is "Lag2".

```
##
## Call:
## glm(formula = Direction ~ ., 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
```

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.

The confusion matrix below shows True Postive (top left), False Positive (top right), True Negative (bottom right), and False Positive (bottom left). True positive and true negative show correct predictions while false positive and false negative show type I and type II errors, respectively. This is to say, you can be wrong in different ways and depending on the problem, it may be important which way you are wrong.

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

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

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).

See results below. 62.5% of observations were correctly predicted. False positive and false negative seem fairly balanced.

```
##
## fit2.pred Down Up
##      Down    9  5
##      Up     34 56
```

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

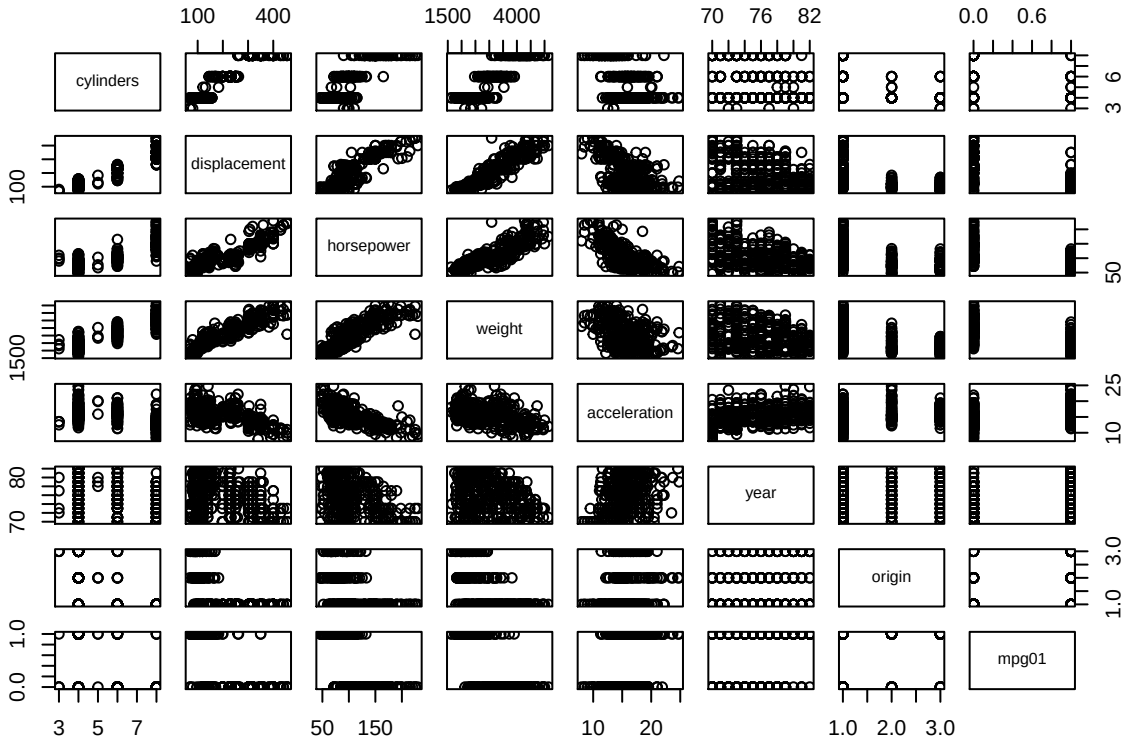
3. Question 4.7.11(a,b,c,f) pg 172 In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the Auto data set.

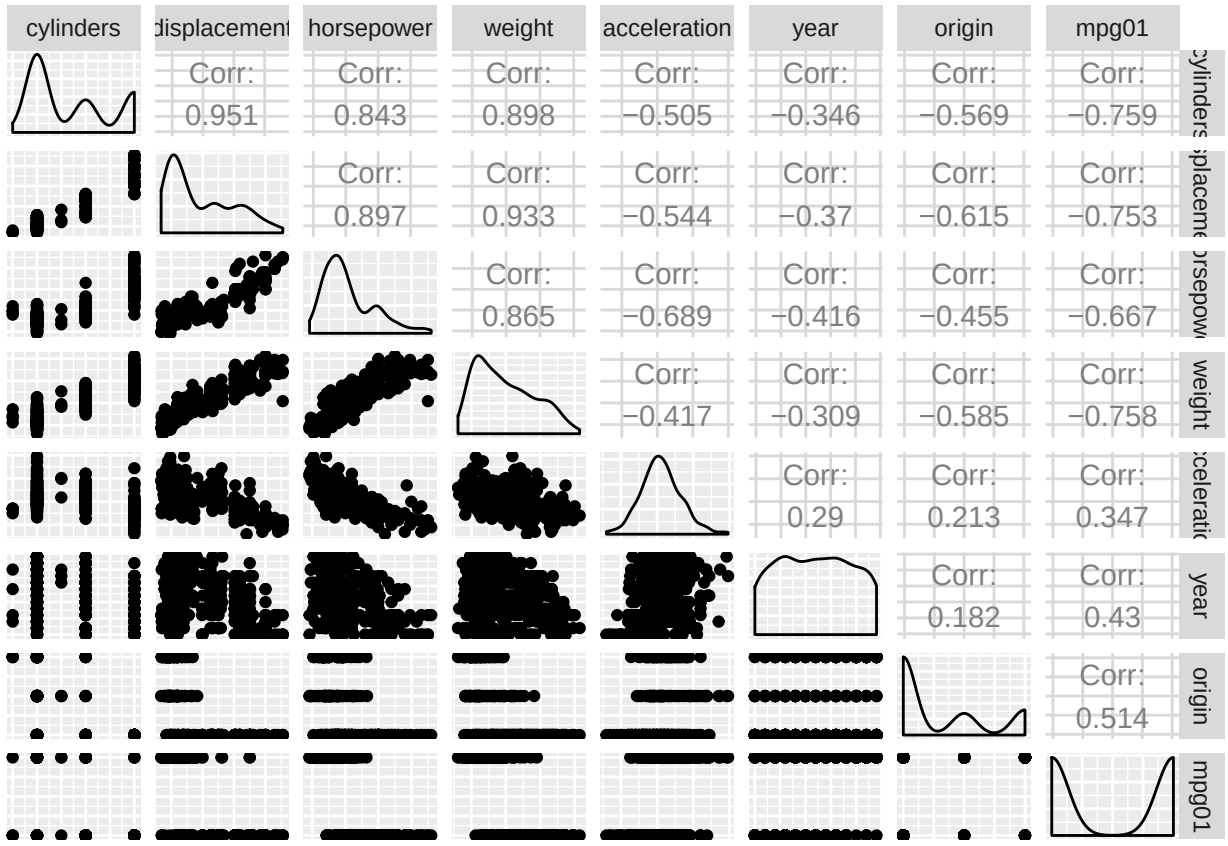
a) Create a binary variable, mpg01, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median. You can compute the median using the median() function. Note you may find it helpful to use the data.frame() function to create a single data set containing both mpg01 and the other Auto variables.

Completed.

b) Explore the data graphically in order to investigate the association between mpg01 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.

Based on the visualization, horsepower and weight predictor variables seem to provide the most discriminatory power for determining if mpg was above or below the median mpg.





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

Used a 60/40 split.

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?

Horsepower and weight seemed to discriminate mpg performance the best from a visual standpoint, so I used them as predictor variables. The test error is ~0.11.

```
##
## logit.pred  0  1
##          0 69  7
##          1 13 68
```

```
## [1] 0.1273885
```

4. Write a function in RMD that calculates the misclassification rate, sensitivity, and specificity. The inputs for this function are a cutoff point, predicted probabilities, and original binary response. Test your function using the model from 4.7.10 b. (Post any questions you might have regarding this on the discussion board, this needs to be an actual function, using the function() command, not just a chunk of code). This will be something you will want to use throughout the semester, since we will be calculating these a lot! *Show the function code you wrote in your final write-up.*

Function code is included directly below. See results further below.

```
BinaryResults = function(cut_off, pred, orig){
  pred = ifelse(pred > cut_off, 1, 0)
  class_matrix = table(pred, orig)
  misclass = (class_matrix[2]+class_matrix[3])/sum(class_matrix)
  sensit = (class_matrix[1]/(class_matrix[1]+class_matrix[2]))
  specif = (class_matrix[4]/(class_matrix[4]+class_matrix[3]))

  cat("These values are predicated on `Down` being the positive reponse\n")
  cat("and `Up` being the negative response.\n\n")
  cat("Misclassification rate:\n")
  print(misclass)
  cat("Sensitivity:\n")
  print(sensit)
  cat("Specificity:\n")
  print(specif)
}
```

```
## These values are predicated on `Down` being the positive reponse
## and `Up` being the negative response.
##
## Misclassification rate:
## [1] 0.4389348
## Sensitivity:
## [1] 0.1115702
## Specificity:
## [1] 0.9206612
```