Homework 3

Alex Soupir

February 8, 2020

Packages: ISLR, GGally

Collaborators:

Answer all questions specified on the problem and include a discussion on how your results answered/addressed the question.

Submit your **.rmd** file with the knitted **PDF** (or knitted Word Document saved as a PDF). If you are having trouble with .rmd, let us know and we will help you, but both the .rmd and the PDF are required.

This file can be used as a skeleton document for your code/write up. Please follow the instructions found under Content for Formatting and Guidelines. No code should be in your PDF write-up unless stated otherwise.

For any question asking for plots/graphs, please do as the question asks as well as do the same but using the respective commands in the GGPLOT2 library. (So if the question asks for one plot, your results should have two plots. One produced using the given R-function and one produced from the GGPLOT2 equivalent). This doesn't apply to questions that don't specifically ask for a plot, however I still would encourage you to produce both.

You do not need to include the above statements.

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

Question 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.

Since this can get crazy relatively quickly, creating a dummy variable for $e^{\beta_0+\beta_1 X}$ will make it less messy.

$K = e^{\beta_0+\beta_1 X}$

$\frac{p(X)}{1-p(X)} = K$

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

Flip

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

Subtract

$\frac{1}{p(X)} - 1 = \frac{1}{K}$

Flip

$K = \frac{1}{\frac{1}{p(X)}-1}$

$\frac{1}{\frac{1}{p(X)}-1} = \frac{1}{\frac{1}{p(X)}-\frac{p(X)}{p(X)}}$

$\frac{1}{\frac{1}{p(X)}-\frac{p(X)}{p(X)}} = \frac{1}{\frac{1-p(X)}{p(X)}}$

Flip

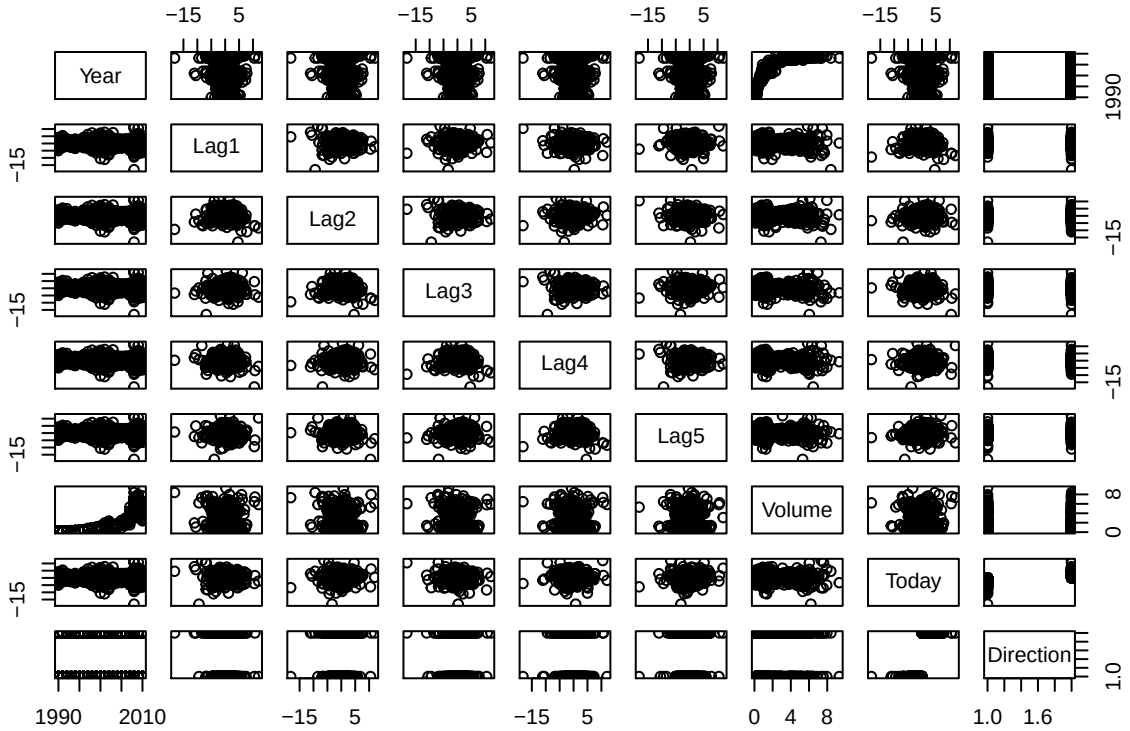
$\frac{1}{\frac{1-p(X)}{p(X)}} = \frac{p(X)}{1-p(X)}$

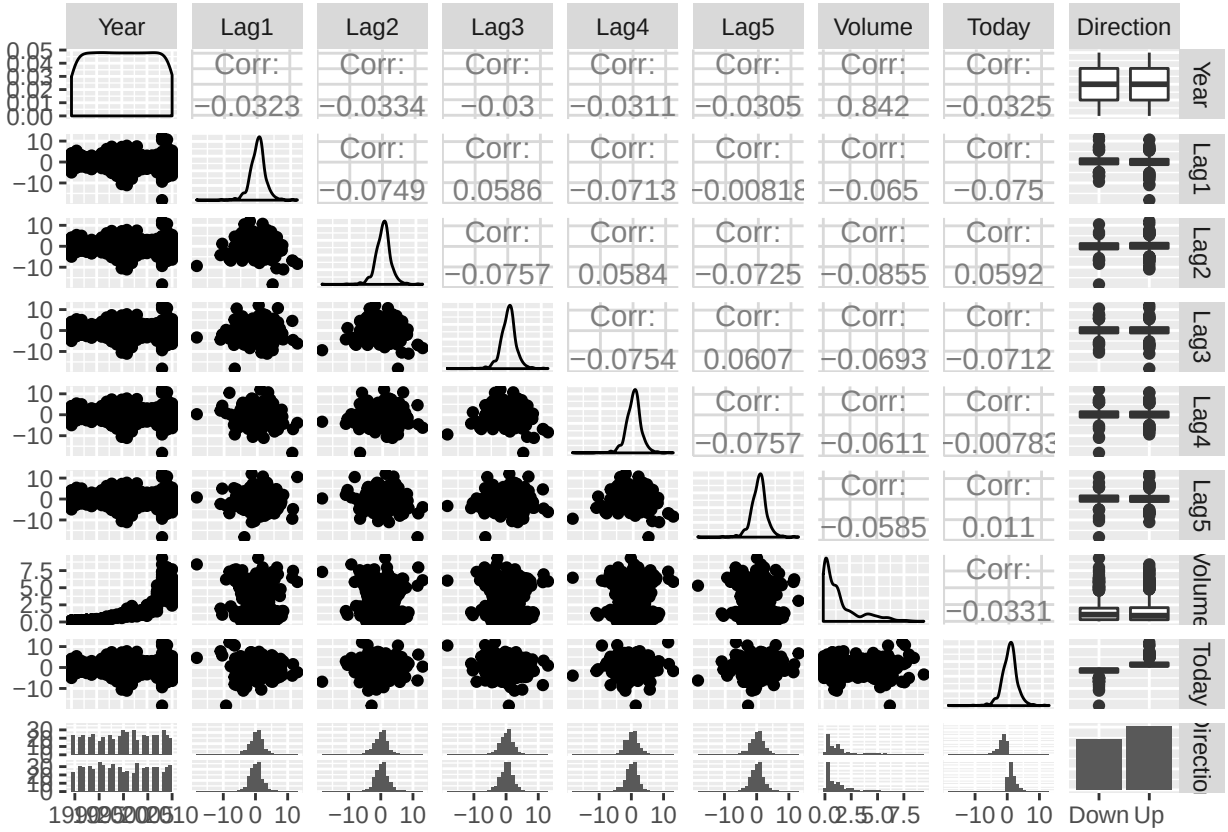
Question 2. Question 4.7.10(a-d) pg 171

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 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?

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


Year
Lag1
Lag2
Lag3
Lag4
Lag5
Volume
Today
Direction


Year
Lag1
Lag2
Lag3
Lag4
Lag5
Volume
Today
Direction
Corr: −0.0323
Corr: −0.0334
Corr: −0.03
Corr: −0.0311
Corr: −0.0305
Corr: 0.842
Corr: −0.0325
Corr: −0.0749
Corr: 0.0586
Corr: −0.0713
Corr: −0.00818
Corr: −0.065
Corr: −0.075
Corr: −0.0757
Corr: 0.0584
Corr: −0.0725
Corr: −0.0855
Corr: 0.0592
Corr: −0.0754
Corr: 0.0607
Corr: −0.0693
Corr: −0.0712
Corr: −0.0757
Corr: −0.0611
Corr: −0.00783
Corr: −0.0585
Corr: 0.011
Corr: −0.0331
Down Up

`Volume` and `Year` are highly correlated to each other, 0.842. Another interesting pattern that can be seen from the second correlation plot matrix is the lag correlations. `Lag1` to `Lag2`, `Lag2` to `Lag3`, etc. has really similar correlations of about -0.076. Although low, this is interesting that each lagged day shares a similar relationship.

(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 printthe results. Do any of the predictors appear to be statistically significant? If so, which ones?

```
##
## 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
```

Of the lag variables and the `Volume` variable, the only one that was significant was the `Lag2` variable. Since the coefficient is positive, as `Lag2` increases, so does the probability of the market having a positive return for that week.

(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.

```
## Confusion Matrix:

##
## preds2.b Down  Up
##     Down   54  48
##     Up    430 557

## Overall fraction of correct predictions:

## [1] 0.5610652

## Incorrect Up predictions, should be down:

## [1] 0.4356636
```

```
## Incorrect Down predictions, should be up:

## [1] 0.4705882
```

The confusion matrix is telling us that the model predicts the right market direction movement 56.1% of the time, while predicting Up when it should be predicting Down 43.5% of the time and predicting Down when it should be predicting Up 47.1% of the time. It appears that the model is favoring the prediction of Up, which might come from the data having more Up data observations than down observations (484 down and 605 up in the whole data set).

(d) Now fit the logistic regression model using the 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, that data from 2009 and 2010).

```
##
## Call:
## glm(formula = Direction ~ Lag2, family = binomial, data = train.2)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.536  -1.264   1.021   1.091   1.368
##
## 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

## Confusion Matrix of data from 2009 and 2010:

##
## preds2.d Down Up
##     Down    9  5
##     Up     34 56

## Overall fraction of correct predictions:

## [1] 0.625

## Incorrect Up predictions, should be down:

## [1] 0.3777778

## Incorrect Down predictions, should be up:

## [1] 0.3571429
```

This model performed better as far as predicting the correct market movement, having the correct market direction prediction 62.5% of weeks in 2009 and 2010 after being trained on the previous 18 years with only `Lag2` which was the only significant predictor in the previous model. The predictions of Up when it should be Down decreased to 37.8% and the predictions

of Down when it should have been up decreased to 35.7% compared to the whole data trained model.

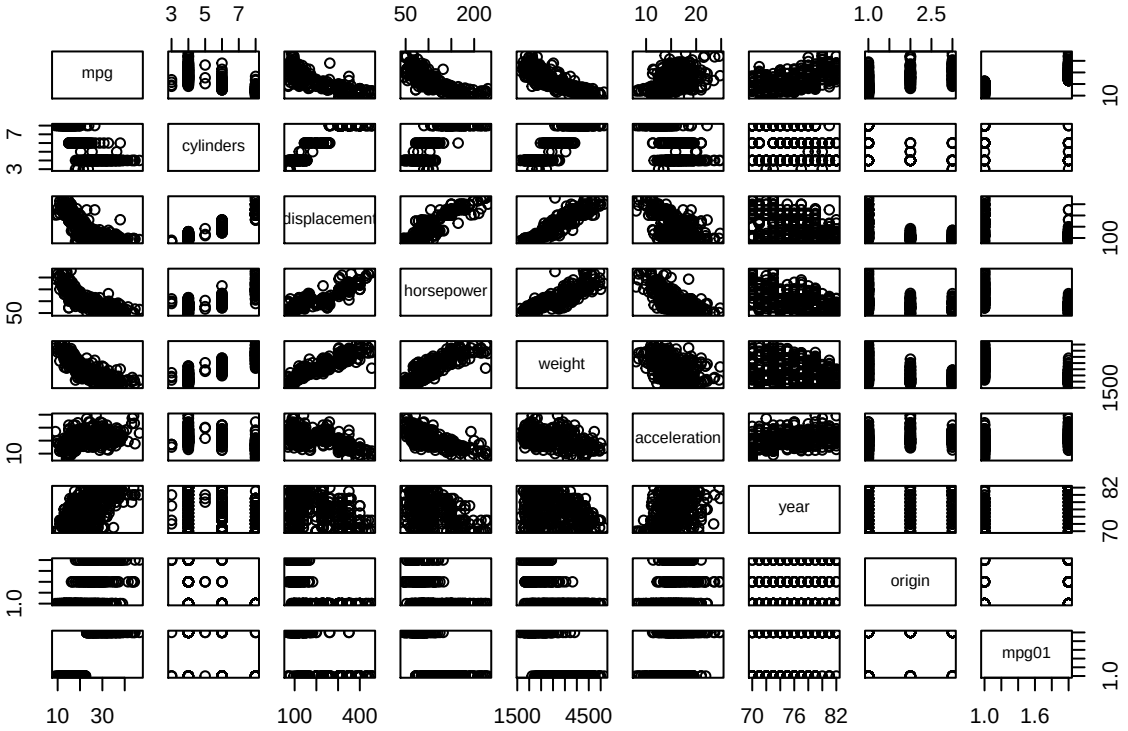
Question 3. Question 4.7.11(a,b,c,f) pg 172

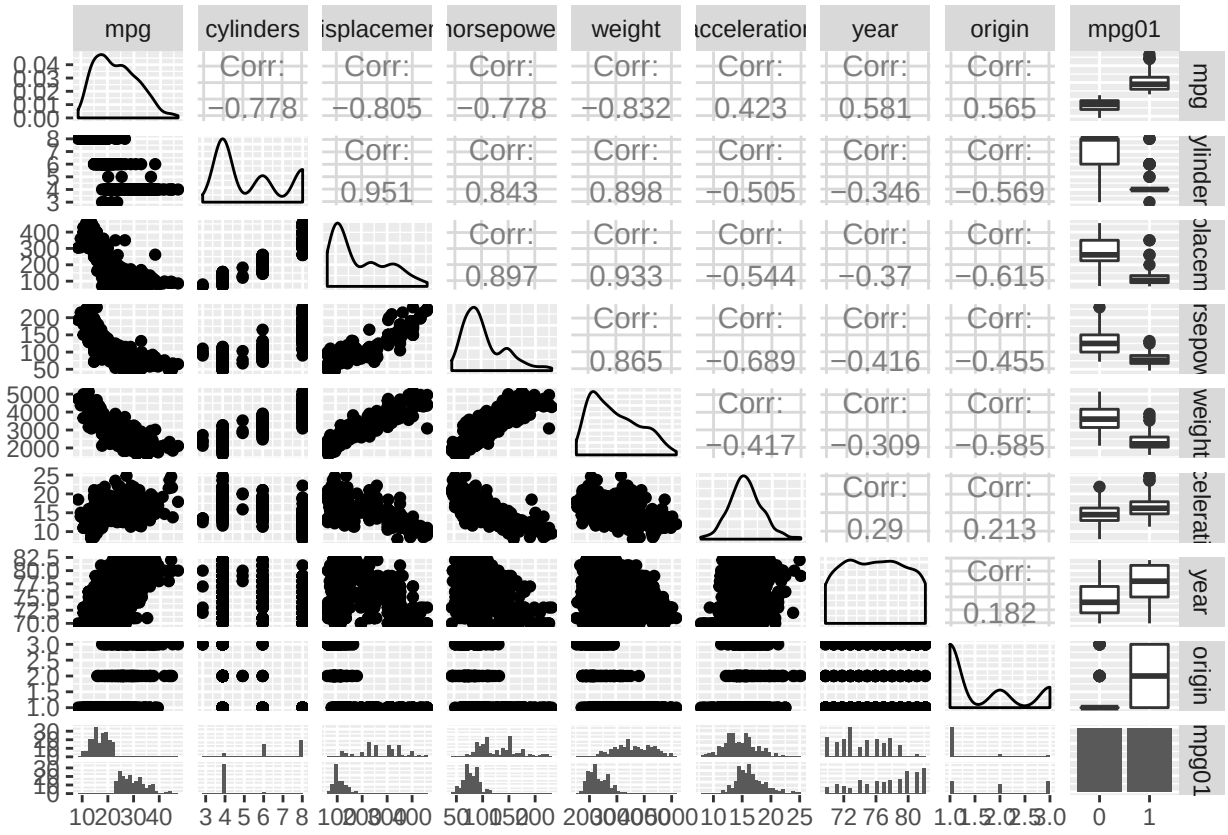
In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the `Auto` data set.

(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.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0     0.5     0.5     1.0     1.0
```

(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.





Using scatter plots suggested in the problem, there are a couple variables that seem to be decent predictors of the `mpg01` binary response. The variables `cylinders`, `displacement`, `horsepower`, and `weight` are all decently negatively correlated with `mpg01`. As these predictors increase, on their own, `mpg01` is more likely to be *0* than *1*.

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

(d) 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?

```
##
## Call:
## glm(formula = mpg01 ~ cylinders + displacement + horsepower +
##     weight, family = binomial, data = train3.c)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4220  -0.2350   0.1411   0.4126   3.2135
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  10.7939760  1.7947878   6.014 1.81e-09 ***
## cylinders     0.0648083  0.3600721   0.180   0.8572
## displacement -0.0137065  0.0084145  -1.629   0.1033
## horsepower   -0.0361916  0.0148458  -2.438   0.0148 *
## weight       -0.0018330  0.0007464  -2.456   0.0141 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 432.32  on 311  degrees of freedom
## Residual deviance: 177.05  on 307  degrees of freedom
## AIC: 187.05
##
## Number of Fisher Scoring iterations: 7
```

```
## Confusion matrix of `Auto` testing split:
```

```
##
## preds3.f  0  1
##        0 39  0
##        1  5 36
```

```
##
## Model error rate of the `Auto` testing split:
```

```
## [1] 0.0625
```

Something that I thought was interesting was that even though `displacement` and `cylinders` was fairly correlated with the `mpg01` response variable alone, they weren't significant in the model with the other variables present. This matches up with what Cami was saying about predictors in different models may no longer be significant if they were in another. She mentioned that the variables may be highly correlated to each other, and in this case `displacement` and `cylinders` are highly correlated (0.951).

Question 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.*

```
mss.calc = function(cut_off, pred_probs, orig_resp){
  pred_resp = ifelse(pred_probs > cut_off, 1, 0)
  class_matrix = table(pred_resp, orig_resp)
  misclass = (class_matrix[2]+class_matrix[3])/sum(class_matrix)
  sensit = class_matrix[4]/(class_matrix[2]+class_matrix[4])
  specif = class_matrix[1]/(class_matrix[1]+class_matrix[3])

  cat("These values are based on `1` as a positive reponse\n")
  cat("and `0` as the negative response.\n\n")
  cat("Misclassification rate (rate of false positive and false negative):\n")
  print(misclass)
  cat("Sensitivity (rate of correct `1` or positives):\n")
  print(sensit)
  cat("Specificity (rate of correct `0` or negatives):\n")
  print(specif)
}
```

Testing the function on the results from Question 4.7.3.f

```
## These values are based on `1` as a positive reponse
## and `0` as the negative response.
##
## Misclassification rate (rate of false positive and false negative):
```

```
## [1] 0.0625
## Sensitivity (rate of correct `1` or positives):
## [1] 0.8780488
## Specificity (rate of correct `0` or negatives):
## [1] 1
```

Looking at how sensitivity and specificity are calculated, it appears that it depends on what is assigned as for a `1` and `0`. Since I have 'yes' of higher than the median as a `1`, and 'no' as `0`, I have to flip which rows are sensitivity and specificity?

References

- stackoverflow
- statinfer.com