Midterm Assignemnt, ME314 2019

Summer School 2019 midsession examination

ME314 Introduction to Data Science and Machine Learning

Suitable for all candidates

Instructions to candidates

- Complete the assignment by adding your answers directly to the RMarkdown document, knitting the document, and submitting the HTML file to Moodle.
- Time allowed: due 19:00 on Wednesday, 7th August 2019.
- Submit the assignment via Moodle (https://shortcourses.lse.ac.uk/course/view.php?id=158).

Question 1:

This question should be answered using the `Carseats` data set, which is part of the **ISLR** package. This data contains simulated data set containing sales of child car seats at 400 different stores.

```
data("Carseats", package = "ISLR")
```

1. Fit a regression model predicting Sales using Advertising and Price as predictors. Interpret the coefficients, the R^2, and the Residual standard error from the regression (by explaining each in a few statements).

```
lm.fit = lm(formula = Sales ~ Advertising + Price, data = Carseats)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = Sales ~ Advertising + Price, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9011 -1.5470 -0.0223  1.5361  6.3748
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.003427   0.606850  21.428  < 2e-16 ***
## Advertising  0.123107   0.018079   6.809 3.64e-11 ***
## Price       -0.054613   0.005078 -10.755  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.399 on 397 degrees of freedom
## Multiple R-squared:  0.2819, Adjusted R-squared:  0.2782
## F-statistic: 77.91 on 2 and 397 DF,  p-value: < 2.2e-16
```

(Answer 1): “One unit (in thousands dollars) increase in Advertising, i.e., local advertising budget, results in 0.1231 unit (in thousands) increase in sales, while holding everything else constant.”

(Answer 2): R-squared: 0.2819 “The proportion of the variance for the Sales variable that is explained by the Advertising and Sales variables is 0.2819.”

(Answer 3): Residual Standard Error: 2.399 “The standard deviation of the residual values, or the difference between a set of observed and predicted values for Sales is 2.399. It shows how well a set of data points fit with the actual model.”

2. Fit a second model by adding Urban as an interactive variable with Advertising. Interpret the two new coefficients produced by adding this interaction to the Advertising variable that was already present from the first question, in a few statements.

```
lm.fit2 = lm(formula = Sales ~ Advertising + Price + Urban + Adverti
sing*Urban, data = Carseats)
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = Sales ~ Advertising + Price + Urban + Advertising *
##     Urban, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8696 -1.5640 -0.0284  1.5323  6.3818
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          12.988703   0.666534  19.487  < 2e-16 ***
## Advertising           0.128015   0.034404   3.721 0.000227 ***
## Price                -0.054519   0.005110 -10.670  < 2e-16 ***
## UrbanYes              0.004602   0.369040   0.012 0.990057
## Advertising:UrbanYes -0.006666   0.040564  -0.164 0.869559
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.405 on 395 degrees of freedom
## Multiple R-squared:  0.2819, Adjusted R-squared:  0.2747
## F-statistic: 38.77 on 4 and 395 DF,  p-value: < 2.2e-16
```

(Answer 1): The coefficient of Urban (0.0046) tells us the starting difference in sales between the stores in an urban and rural location. The coeefficent of the intercation variable (-0.006666) tells us the additional effect of Advertising has on Sales for the stores in an urban location.

3. Which of these two models is preferable, and why?

(Answer 1): I would prefer the first model, as the two added coefficent are not statistically significant in the second model. That is, the different effect of the locations of stores has no statistical significance on Sales. The R-squared value remains the same and the adjusted R-squared value has fallen. Thus, the added two variables does not provide any additional explaination to the variantion for Sales.

Question 2:

You will need to load the core library for the course textbook and any other libraries you find suitable to answer the question:

```
data("Weekly", package = "ISLR")
library("MASS")
library("class")
```

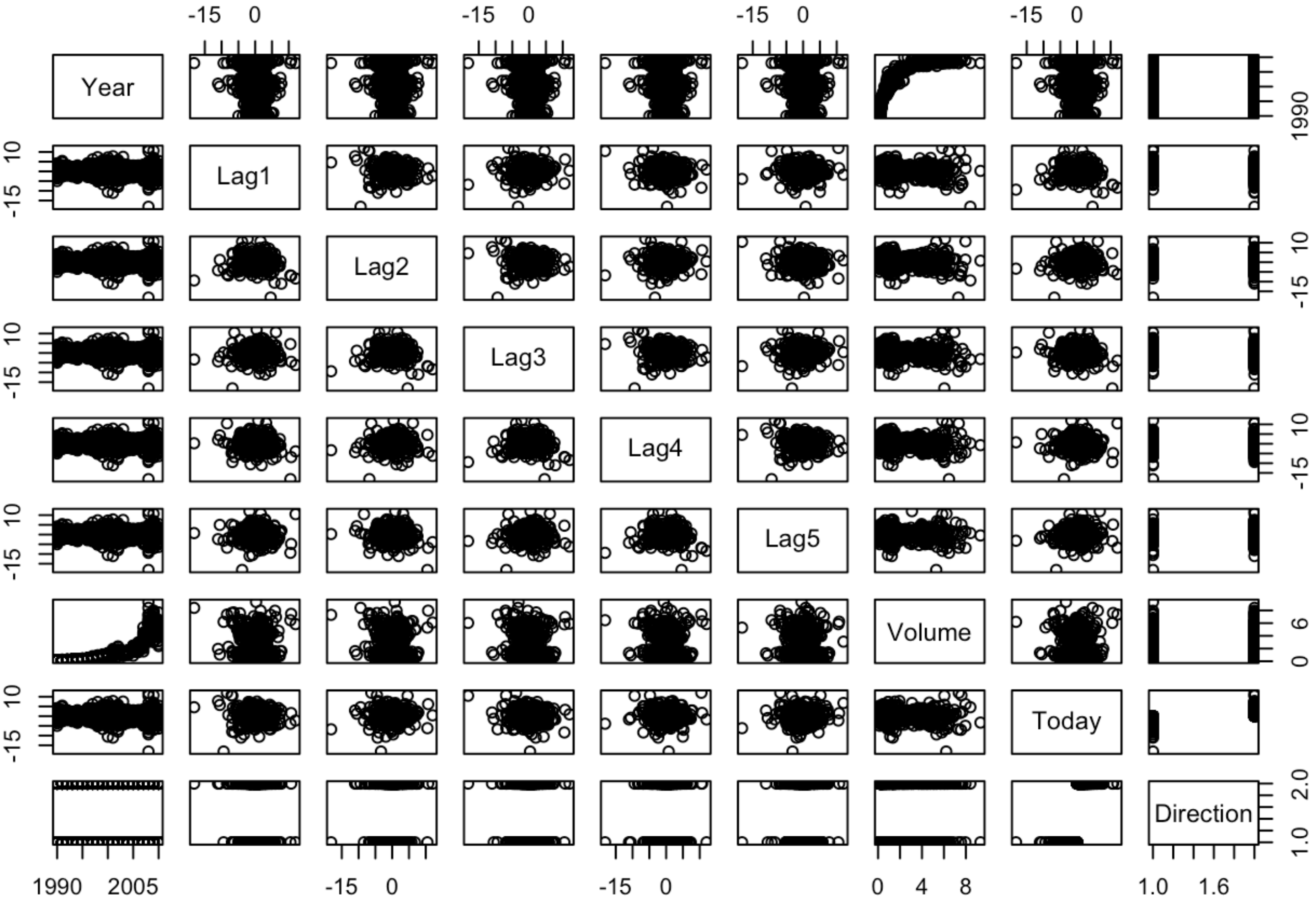
This question should be answered using the `Weekly` data set, which is part of the **ISLR** package. This data contains 1,089 weekly stock returns for 21 years, from the beginning of 1990 to the end of 2010.

1. Perform exploratory data analysis of the `Weekly` data (produce some numerical and graphical summaries). Discuss any patterns that emerge.

```
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
```

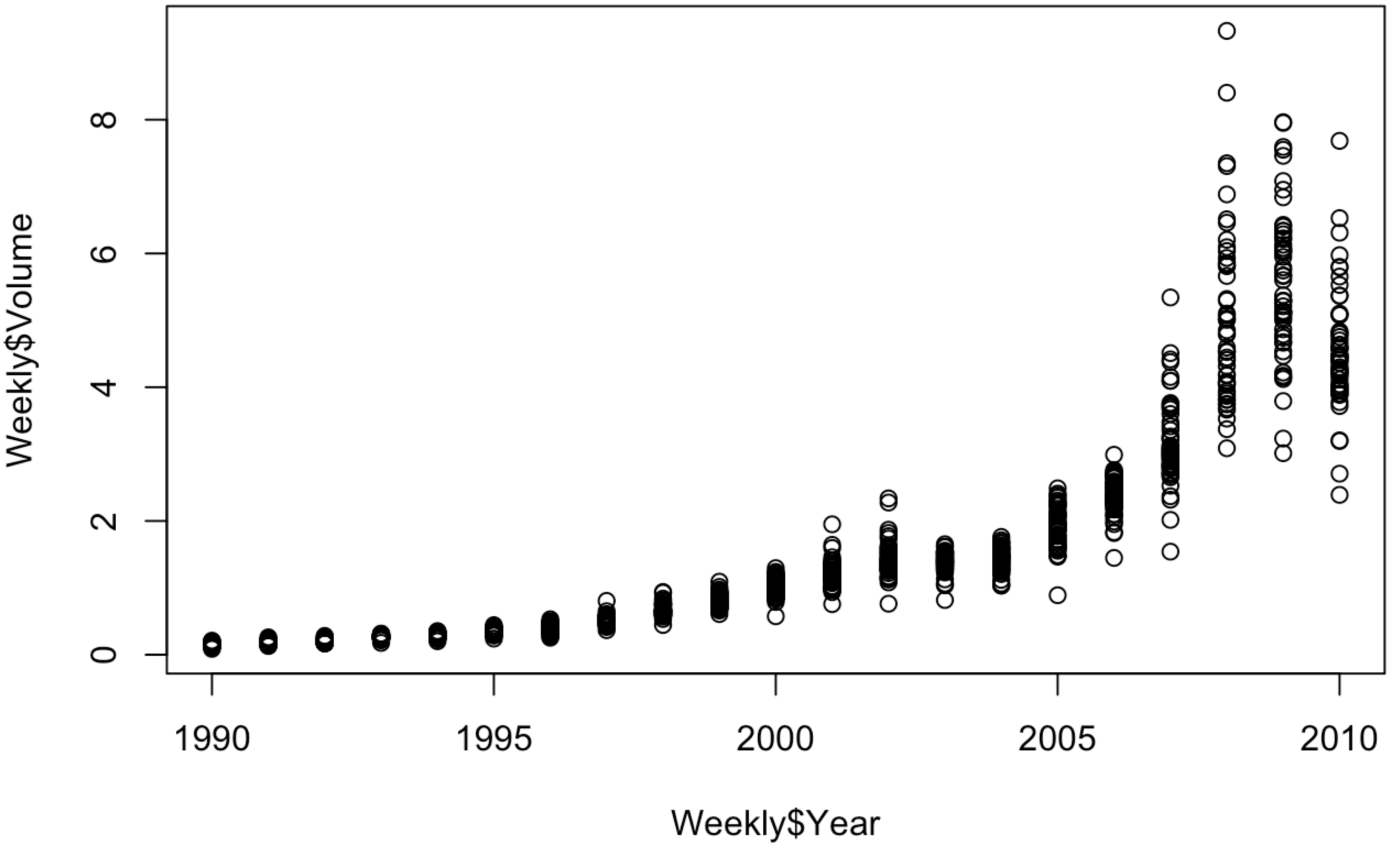
```
pairs(Weekly)
```


Year
Lag1
Lag2
Lag3
Lag4
Lag5
Volume
Today
Direction
-15
0
10
1990
2005
0
4
8
6
1.0
1.6
2.0

```
cor(Weekly[,-9])
```

```
##               Year         Lag1        Lag2        Lag3         L
ag4
## Year    1.00000000 -0.032289274 -0.03339001 -0.03000649 -0.031127
923
## Lag1   -0.03228927  1.000000000 -0.07485305  0.05863568 -0.071273
876
## Lag2   -0.03339001 -0.074853051  1.00000000 -0.07572091  0.058381
535
## Lag3   -0.03000649  0.058635682 -0.07572091  1.00000000 -0.075395
865
## Lag4   -0.03112792 -0.071273876  0.05838153 -0.07539587  1.000000
000
## Lag5   -0.03051910 -0.008183096 -0.07249948  0.06065717 -0.075675
027
## Volume  0.84194162 -0.064951313 -0.08551314 -0.06928771 -0.061074
617
## Today  -0.03245989 -0.075031842  0.05916672 -0.07124364 -0.007825
873
##                Lag5      Volume        Today
## Year   -0.030519101  0.84194162 -0.032459894
## Lag1   -0.008183096 -0.06495131 -0.075031842
## Lag2   -0.072499482 -0.08551314  0.059166717
## Lag3    0.060657175 -0.06928771 -0.071243639
## Lag4   -0.075675027 -0.06107462 -0.007825873
## Lag5    1.000000000 -0.05851741  0.011012698
## Volume -0.058517414  1.00000000 -0.033077783
## Today   0.011012698 -0.03307778  1.000000000
```

```
plot(Weekly$Year, Weekly$Volume)
```



```
Weekly[row.names(Weekly)[which.max(Weekly$Volume)],1]
```

```
## [1] 2008
```

(Answer 1): Recent years correlate with higher volume of shares traded. Year 2008 has the highest volume of shares traded.

2. Fit a logistic regression with `Direction` as the response and different combinations of lag variables plus `Volume` as predictors. Use the period from 1990 to 2008 as your training set and 2009-2010 as your test set. Produce a summary of results.

```
train=subset(Weekly, Year<'2009')
# Fit a model
model1 <- glm(Direction ~ Lag1+Lag2+Lag3+Lag4+Lag5+Volume, data=train, family=binomial)
model2 <- glm(Direction ~ Lag1+Lag3+Lag4+Lag5+Volume, data=train, family=binomial)
model3 <- glm(Direction ~ Lag3+Lag4+Lag5+Volume, data=train, family=binomial)
summary(model1)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7186  -1.2498   0.9823   1.0841   1.4911
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.33258    0.09421   3.530 0.000415 ***
## Lag1        -0.06231    0.02935  -2.123 0.033762 *
## Lag2         0.04468    0.02982   1.499 0.134002
## Lag3        -0.01546    0.02948  -0.524 0.599933
## Lag4        -0.03111    0.02924  -1.064 0.287241
## Lag5        -0.03775    0.02924  -1.291 0.196774
## Volume      -0.08972    0.05410  -1.658 0.097240 .
## ---
## 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: 1342.3  on 978  degrees of freedom
## AIC: 1356.3
##
## Number of Fisher Scoring iterations: 4
```

```
test = subset(Weekly, Year > '2008')
results <- predict(model1, test, type = "response")
```

```
Do any of the predictors appear to be statistically significant in y
our training set? If so, which ones?
```

(Answer 1): In model 1, Lag1 appears to be statistically significant at 5% significance level. In the same model, Volume appears to be statistically significant at 10% significance level. The rest of the predictors do not appear to be statistically significant at 10% significance level.

3. From your test set, compute the confusion matrix, and calculate accuracy, precision, recall and F1.

```
predict = rep("Down", 1089)
predict[results > 0.5] = "Up"
confusion_matrix <- table(predict, Weekly$Direction)
confusion_matrix
```

```
##
## predict Down  Up
##    Down  343 443
##    Up    141 162
```

```
Accuracy <- (84+531)/1089
Accuracy
```

```
## [1] 0.5647383
```

```
Precision <- 531/(531+74)
Precision
```

```
## [1] 0.877686
```

```
Recall <- 531/(531+400)
Recall
```

```
## [1] 0.5703545
```

```
F1 <- 2 * ((Precision * Recall) /(Precision + Recall))
F1
```

```
## [1] 0.6914062
```

```
Explain what the confusion matrix is telling you about the types of
mistakes made by logistic regression, and what can you learn from ad
ditional measures of fit like accuracy, precision, recall, and F1.
```

(Answer 1): The confusion matrix tells us the following types of mistakes: False Negative: how many times our classifier predicts “no” (Down), but the actual is “yes” (Up). False Positive: how many times our classifier predicts “yes” (Up), but the actual is “no” (Down).

(Accuracy): the ratio of correctly predicted observation to the total observations.

(Precision): the ratio of correctly predicted “positive” observations to the total predicted “positive” observations.

(Recall): the ratio of correctly predicted “positive” observations to the all observations in actual class, i.e., “up”.

(F1): the weighted average of Precision and Recall.

4. (Extra credit) Experiment with alternative classification methods.

   Present the results of your experiments reporting method, associated confusion matrix, and measures of fit on the test set like accuracy, precision, recall, and F1.