MATH 189 HW6

Zijian Su
Zelong Zhou
Xiangyi Lin

Last Updated: February 24, 2023

Concrete contributions

All problems were done by Zijian Su, Zelong Zhou, Xiangyi Lin. All contributing equally to this assignment. Everyone put in enough effort.

Overview

Packages

```
#install.packages("rmarkdown")
#install.packages("ggplot2")

#tinytex::install_tinytex()

#install.packages("scatterplot3d")
```

Question 1

Suppose we collect data for a group of students in a statistics class with variables X1 = hours studied, X2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficients:

$\hat{}B0 = -6$, $\hat{}B1 = 0.05$, $\hat{}B2 = 1$.

(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.

Answer:

From the question we can get: Y = -6 +0.05*X1 + 1* X2, I have X1 = 40, X2 = 3.5.
For probability we have this formula:
$P(X) = e^{(B0 + B1*X1 + B2*X2)} / 1 + e^{(B0 + B1*X1 + B2*X2)}$ Putting in X1,X2, B0, B1, B2, we can get:
$P(X) = e^{(-6 + 0.05*40 + 1*3.5)} / 1 + e^{(-6 + 0.05*40 + 1*3.5)} \sim= 0.37754$

Therefore, the probability of a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class is about 37.754%.

(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:

From the question, we know P(X) = 0.5. So, we can have:
$0.5 = e^{bs} / 1 + e^{bs}$

$1 + e^{bs} = 2e^{bs}$

$e^{bs} = 1$

$bs = \ln(1)$

$bs = 0$

bs = $\hat{}B0 + \hat{}B1*X1 + \hat{}B2*X2$, Putting in X2, B0, B1, B2, we can get:
0 = -6 + 0.05* X1 + 1* 3.5
0.05*X1 = 6 - 3.5
X1 = 2.5/0.05
X1 = 50

Therefore, a student in part (a) need to study 50 hours so this student can have a 50% chance of getting an A in this class.

Question 2

Consider the Weekly data set, which is part of the ISLR package. This data set consists of 1089 weekly percentage returns for the S&P 500 stock index over 21 years, from the beginning of 1990 to the end of 2010. It contains the following 9 variables.

Year: The year that the observation was recorded. **Lag1:** Percentage return for previous week. **Lag2:** Percentage return for 2 weeks previous. **Lag3:** Percentage return for 3 weeks previous. **Lag4:** Percentage return for 4 weeks previous. **Lag5:** Percentage return for 5 weeks previous. **Volume:** Volume of shares traded (average number of daily shares traded in billions).
Today: Percentage return for this week. **Direction:** A factor with levels Down and Up indicating whether the market had a positive or negative return on a given week.

(a)

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

Answer:

```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 4.1.3
```

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.1.3
```

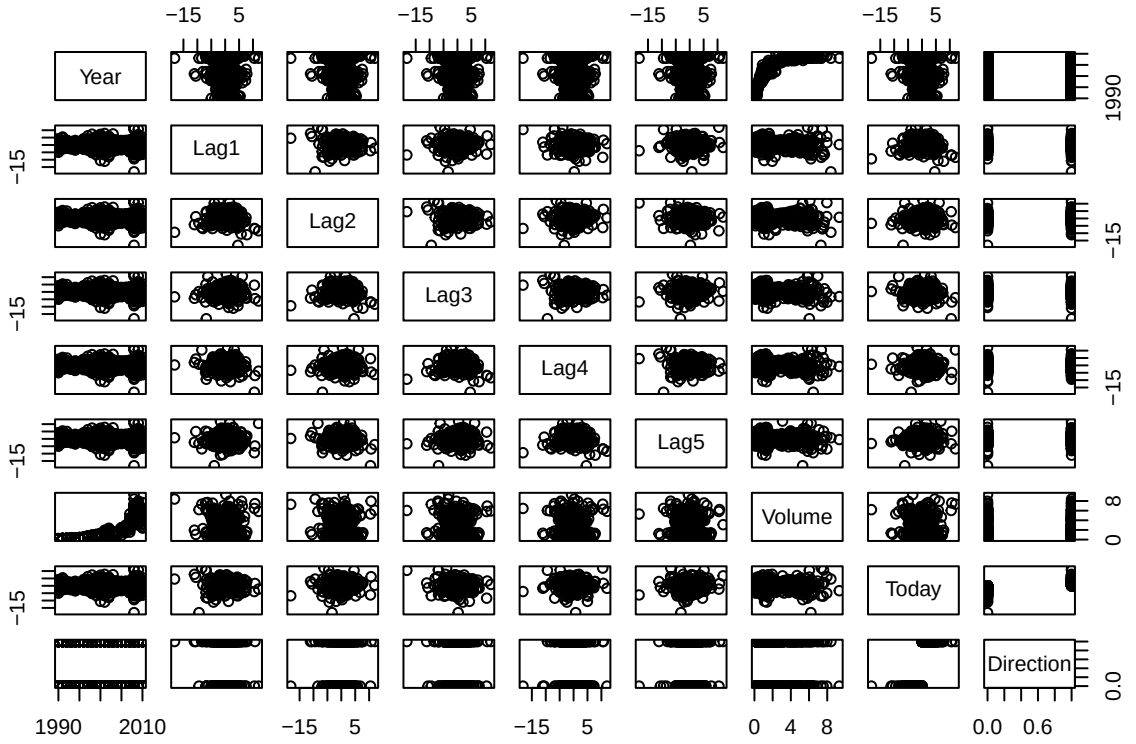
```
data <- Weekly
knitr::kable(summary(data))
```

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

```
data$Direction = ifelse(data$Direction == "Up", 1, 0)
cor(data)
```

```
##                  Year         Lag1        Lag2        Lag3         Lag4
## Year       1.00000000 -0.032289274 -0.03339001 -0.03000649 -0.031127923
## Lag1      -0.03228927  1.000000000 -0.07485305  0.05863568 -0.071273876
## Lag2      -0.03339001 -0.074853051  1.00000000 -0.07572091  0.058381535
## Lag3      -0.03000649  0.058635682 -0.07572091  1.00000000 -0.075395865
## Lag4      -0.03112792 -0.071273876  0.05838153 -0.07539587  1.000000000
## Lag5      -0.03051910 -0.008183096 -0.07249948  0.06065717 -0.075675027
## Volume     0.84194162 -0.064951313 -0.08551314 -0.06928771 -0.061074617
## Today     -0.03245989 -0.075031842  0.05916672 -0.07124364 -0.007825873
## Direction -0.02220025 -0.050003804  0.07269634 -0.02291281 -0.020549456
##                   Lag5      Volume        Today   Direction
## Year      -0.030519101  0.84194162 -0.032459894 -0.02220025
## Lag1      -0.008183096 -0.06495131 -0.075031842 -0.05000380
## Lag2      -0.072499482 -0.08551314  0.059166717  0.07269634
## Lag3       0.060657175 -0.06928771 -0.071243639 -0.02291281
## Lag4      -0.075675027 -0.06107462 -0.007825873 -0.02054946
## Lag5       1.000000000 -0.05851741  0.011012698 -0.01816827
## Volume    -0.058517414  1.00000000 -0.033077783 -0.01799521
## Today      0.011012698 -0.03307778  1.000000000  0.72002470
## Direction -0.018168272 -0.01799521  0.720024704  1.00000000
```

```
pairs(data)
```



From the above data, the year and volume should be related, and the direction and today can also be considered related. The rest can be considered irrelevant.

(b)

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

Answer:

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

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

We can see that only Lag2's p-value < 0.05, Lag2 reject the null hypothesis($B = 0$). So Lag2 is statistically significant .

The remaining P values are all greater than 0.05, supporting null hypothesis($B = 0$).

(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:

```
data<- Weekly
Probs = predict(data_fits, type='response')
Pred = ifelse(Probs>0.5, "Up", "Down")
table(Pred, data$Direction)
```

```
##
## Pred   Down  Up
##   Down   54  48
##   Up    430 557
```

overall fraction of correct predictions: (54+557) / (54+48+430+557) ~= 0.56106

We can know that the prediction accuracy of this model is only about 56%.Also, the model correctly predicts most of the ups (48 false positives(type I errors)), but mispredicts most of the downs(430 false negatives(type II errors)).

(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_set = data[data$Year <= 2008,]
test_set = data[data$Year > 2008,]

data_fits2 = glm(Direction ~ Lag2, data = train_set, family = binomial)

Probs_2 = predict(data_fits2, test_set, type = "response")
Pred_2 = ifelse(Probs_2>0.5, "Up", "Down")

table(Pred_2, test_set$Direction)
```

```
##
## Pred_2 Down Up
##   Down    9  5
##   Up     34 56
```

overall fraction of correct predictions: (9+56) / (9+5+34+56) = 0.625

The accuracy is 62.5%, and we have 5 type I errors and 34 type II errors.