SL HW 7

Joshua Ingram

11/10/2019

Problem 1

```r
head(Weekly)
```

```
##   Year   Lag1   Lag2   Lag3   Lag4   Lag5    Volume  Today Direction
## 1 1990  0.816  1.572 -3.936 -0.229 -3.484 0.1549760 -0.270      Down
## 2 1990 -0.270  0.816  1.572 -3.936 -0.229 0.1485740 -2.576      Down
## 3 1990 -2.576 -0.270  0.816  1.572 -3.936 0.1598375  3.514        Up
## 4 1990  3.514 -2.576 -0.270  0.816  1.572 0.1616300  0.712        Up
## 5 1990  0.712  3.514 -2.576 -0.270  0.816 0.1537280  1.178        Up
## 6 1990  1.178  0.712  3.514 -2.576 -0.270 0.1544440 -1.372      Down
```

a.

```r
round(cor(Weekly[,-9]), 2)
```

```
##         Year  Lag1  Lag2  Lag3  Lag4  Lag5 Volume Today
## Year    1.00 -0.03 -0.03 -0.03 -0.03 -0.03   0.84 -0.03
## Lag1   -0.03  1.00 -0.07  0.06 -0.07 -0.01  -0.06 -0.08
## Lag2   -0.03 -0.07  1.00 -0.08  0.06 -0.07  -0.09  0.06
## Lag3   -0.03  0.06 -0.08  1.00 -0.08  0.06  -0.07 -0.07
## Lag4   -0.03 -0.07  0.06 -0.08  1.00 -0.08  -0.06 -0.01
## Lag5   -0.03 -0.01 -0.07  0.06 -0.08  1.00  -0.06  0.01
## Volume  0.84 -0.06 -0.09 -0.07 -0.06 -0.06   1.00 -0.03
## Today  -0.03 -0.08  0.06 -0.07 -0.01  0.01  -0.03  1.00
```

There is a strong relationship between Year and Volume (correlation of 0.84)

b.

```r
glm.obj <- glm(Direction ~ .-Today, family = "binomial", data = Weekly)
vif(glm.obj)
```

```
##     Year     Lag1     Lag2     Lag3     Lag4     Lag5   Volume
## 3.490471 1.019615 1.031364 1.021486 1.029219 1.015926 3.558933
```

```r
summary(glm.obj)
```

```
##
## Call:
## glm(formula = Direction ~ . - Today, family = "binomial", data = Weekly)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7071  -1.2578   0.9941   1.0873   1.4665
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 17.225822  37.890522   0.455   0.6494
## Year        -0.008500   0.018991  -0.448   0.6545
## Lag1        -0.040688   0.026447  -1.538   0.1239
## Lag2         0.059449   0.026970   2.204   0.0275 *
## Lag3        -0.015478   0.026703  -0.580   0.5622
## Lag4        -0.027316   0.026485  -1.031   0.3024
## Lag5        -0.014022   0.026409  -0.531   0.5955
## Volume       0.003256   0.068836   0.047   0.9623
## ---
## 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: 1486.2  on 1081  degrees of freedom
## AIC: 1502.2
##
## Number of Fisher Scoring iterations: 4
```

no variabels had a VIF greater than 5, so our logistic regression model includes all variables to predict direction, except today.

c.

(based on summary() from part b) only lag2 is signifcant with $\alpha = 0.05$ and lag1 with $\alpha = 0.20$

Interpretations:

lag1: holding all other variables constant, if we for every one percentage increase of lag1 there will be decrease of 0.041 in the $logit(\hat{\pi})$

lag2: holding all other variables constant, if we for every one percentage increase of lag2 there will be an increase of 0.059 in the $logit(\hat{\pi})$

d.

```r
glm.probs <- predict(glm.obj, type="response")
glm.pred <- ifelse(glm.probs > 0.50, "Up", "Down")

conf.mat <- table(glm.pred, Weekly$Direction)
conf.mat
```

```
##
## glm.pred Down  Up
##     Down   56  47
##     Up    428 558
```

```r
mean(glm.pred == Weekly$Direction)
```

```
## [1] 0.56382
```

The overall accuracy of our model is 56.382%. We had 47 false positives and 428 false positives.

e.

```r
train <- (Weekly$Year<2009)
Weekly.Test <- Weekly[!train,]
dim(Weekly.Test)
```

```
## [1] 104   9
```

```r
Direction.Test <- Weekly$Direction[!train]

glm.train <- glm(Direction ~ .-Today, family="binomial", data = Weekly, subset = train)

glm.test.prob <- predict(glm.obj, type="response", newdata = Weekly.Test)
glm.test.pred <- ifelse(glm.test.prob > 0.50, "Up", "Down")

conf.mat <- table(glm.test.pred, Direction.Test)
conf.mat
```

```
##              Direction.Test
## glm.test.pred Down Up
##          Down   15 12
##          Up     28 49
```

```r
mean(glm.test.pred == Direction.Test)
```

```
## [1] 0.6153846
```

f.

It seems that the accuracy obtained in part (e) would be a more trustworthy value for our model forecasting performance. This is because in part (e) we have a training set and a testing set, so the 61.53% accuracy is with data that the model has not seen before, as opposed to the model in part (d) that used data to both train the model and predicted that same data.

Problem 2

```r
head(Caravan)
```

```
##   MOSTYPE MAANTHUI MGEMOMV MGEMLEEF MOSHOOFD MGODRK MGODPR MGODOV MGODGE
## 1      33        1       3        2        8      0      5      1      3
## 2      37        1       2        2        8      1      4      1      4
## 3      37        1       2        2        8      0      4      2      4
## 4       9        1       3        3        3      2      3      2      4
## 5      40        1       4        2       10      1      4      1      4
## 6      23        1       2        1        5      0      5      0      5
##   MRELGE MRELSA MRELOV MFALLEEN MFGEKIND MFWEKIND MOPLHOOG MOPLMIDD
## 1      7      0      2        1        2        6        1        2
## 2      6      2      2        0        4        5        0        5
## 3      3      2      4        4        4        2        0        5
## 4      5      2      2        2        3        4        3        4
## 5      7      1      2        2        4        4        5        4
## 6      0      6      3        3        5        2        0        5
##   MOPLLAAG MBERHOOG MBERZELF MBERBOER MBERMIDD MBERARBG MBERARBO MSKA
## 1        7        1        0        1        2        5        2    1
## 2        4        0        0        0        5        0        4    0
## 3        4        0        0        0        7        0        2    0
## 4        2        4        0        0        3        1        2    3
## 5        0        0        5        4        0        0        0    9
## 6        4        2        0        0        4        2        2    2
##   MSKB1 MSKB2 MSKC MSKD MHHUUR MHKOOP MAUT1 MAUT2 MAUT0 MZFONDS MZPART
## 1     1     2    6    1      1      8     8     0     1       8      1
## 2     2     3    5    0      2      7     7     1     2       6      3
## 3     5     0    4    0      7      2     7     0     2       9      0
## 4     2     1    4    0      5      4     9     0     0       7      2
## 5     0     0    0    0      4      5     6     2     1       5      4
## 6     2     2    4    2      9      0     5     3     3       9      0
##   MINKM30 MINK3045 MINK4575 MINK7512 MINK123M MINKGEM MKOOPKLA PWAPART
## 1       0        4        5        0        0       4        3       0
## 2       2        0        5        2        0       5        4       2
## 3       4        5        0        0        0       3        4       2
## 4       1        5        3        0        0       4        4       0
## 5       0        0        9        0        0       6        3       0
## 6       5        2        3        0        0       3        3       0
##   PWABEDR PWALAND PPERSAUT PBESAUT PMOTSCO PVRAAUT PAANHANG PTRACTOR
## 1       0       0        6       0       0       0        0        0
## 2       0       0        0       0       0       0        0        0
## 3       0       0        6       0       0       0        0        0
## 4       0       0        6       0       0       0        0        0
## 5       0       0        0       0       0       0        0        0
## 6       0       0        6       0       0       0        0        0
##   PWERKT PBROM PLEVEN PPERSONG PGEZONG PWAOREG PBRAND PZEILPL PPLEZIER
## 1      0     0      0        0       0       0      5       0        0
## 2      0     0      0        0       0       0      2       0        0
## 3      0     0      0        0       0       0      2       0        0
## 4      0     0      0        0       0       0      2       0        0
## 5      0     0      0        0       0       0      6       0        0
## 6      0     0      0        0       0       0      0       0        0
##   PFIETS PINBOED PBYSTAND AWAPART AWABEDR AWALAND APERSAUT ABESAUT AMOTSCO
## 1      0       0        0       0       0       0        1       0       0
```

```
## 2       0       0        0        2       0       0        0       0       0
## 3       0       0        0        1       0       0        1       0       0
## 4       0       0        0        0       0       0        1       0       0
## 5       0       0        0        0       0       0        0       0       0
## 6       0       0        0        0       0       0        1       0       0
##   AVRAAUT AAANHANG ATRACTOR AWERKT ABROM ALEVEN APERSONG AGEZONG AWAOREG
## 1       0        0        0      0     0      0        0       0       0
## 2       0        0        0      0     0      0        0       0       0
## 3       0        0        0      0     0      0        0       0       0
## 4       0        0        0      0     0      0        0       0       0
## 5       0        0        0      0     0      0        0       0       0
## 6       0        0        0      0     0      0        0       0       0
##   ABRAND AZEILPL APLEZIER AFIETS AINBOED ABYSTAND Purchase
## 1      1       0        0      0       0        0       No
## 2      1       0        0      0       0        0       No
## 3      1       0        0      0       0        0       No
## 4      1       0        0      0       0        0       No
## 5      1       0        0      0       0        0       No
## 6      0       0        0      0       0        0       No
```

a.

```r
glm.obj2 <- glm(Purchase ~ ., family="binomial", data = Caravan)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
summary(glm.obj2)
```

```
##
## Call:
## glm(formula = Purchase ~ ., family = "binomial", data = Caravan)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7047  -0.3711  -0.2450  -0.1588   3.2916
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.542e+02  1.116e+04   0.023  0.98183
## MOSTYPE      6.580e-02  4.624e-02   1.423  0.15468
## MAANTHUI    -1.832e-01  1.927e-01  -0.951  0.34157
## MGEMOMV     -2.696e-02  1.399e-01  -0.193  0.84723
## MGEMLEEF     2.096e-01  1.016e-01   2.063  0.03911 *
## MOSHOOFD    -2.767e-01  2.076e-01  -1.333  0.18247
## MGODRK      -1.142e-01  1.069e-01  -1.068  0.28535
## MGODPR      -1.910e-02  1.177e-01  -0.162  0.87112
## MGODOV      -1.618e-02  1.055e-01  -0.153  0.87818
## MGODGE      -6.817e-02  1.113e-01  -0.612  0.54024
## MRELGE       2.310e-01  1.566e-01   1.475  0.14031
## MRELSA       8.509e-02  1.466e-01   0.580  0.56169
## MRELOV       1.467e-01  1.562e-01   0.939  0.34759
```

```
## MFALLEEN    -8.291e-02  1.311e-01  -0.633  0.52702
## MFGEKIND    -1.154e-01  1.337e-01  -0.863  0.38813
## MFWEKIND    -8.140e-02  1.417e-01  -0.575  0.56561
## MOPLHOOG     9.717e-04  1.311e-01   0.007  0.99408
## MOPLMIDD    -9.077e-02  1.365e-01  -0.665  0.50605
## MOPLLAAG    -1.994e-01  1.376e-01  -1.449  0.14740
## MBERHOOG     8.883e-02  9.349e-02   0.950  0.34204
## MBERZELF     3.918e-02  9.897e-02   0.396  0.69219
## MBERBOER    -1.169e-01  1.104e-01  -1.059  0.28951
## MBERMIDD     1.353e-01  9.191e-02   1.472  0.14106
## MBERARBG     3.976e-02  9.067e-02   0.438  0.66104
## MBERARBO     9.954e-02  9.143e-02   1.089  0.27628
## MSKA         2.690e-02  1.035e-01   0.260  0.79502
## MSKB1       -8.801e-03  1.011e-01  -0.087  0.93064
## MSKB2        1.200e-02  9.081e-02   0.132  0.89485
## MSKC         9.016e-02  9.958e-02   0.905  0.36527
## MSKD        -2.468e-02  9.724e-02  -0.254  0.79967
## MHHUUR      -1.472e+01  8.140e+02  -0.018  0.98557
## MHKOOP      -1.469e+01  8.140e+02  -0.018  0.98561
## MAUT1        1.819e-01  1.514e-01   1.202  0.22953
## MAUT2        1.507e-01  1.371e-01   1.099  0.27162
## MAUT0        9.325e-02  1.436e-01   0.649  0.51603
## MZFONDS     -1.445e+01  9.359e+02  -0.015  0.98768
## MZPART      -1.451e+01  9.359e+02  -0.016  0.98763
## MINKM30      1.181e-01  1.006e-01   1.174  0.24039
## MINK3045     1.366e-01  9.650e-02   1.415  0.15694
## MINK4575     1.009e-01  9.667e-02   1.043  0.29678
## MINK7512     1.144e-01  1.027e-01   1.114  0.26513
## MINK123M    -1.607e-01  1.449e-01  -1.109  0.26738
## MINKGEM      9.214e-02  9.945e-02   0.927  0.35417
## MKOOPKLA     6.856e-02  4.642e-02   1.477  0.13966
## PWAPART      5.954e-01  3.901e-01   1.526  0.12693
## PWABEDR     -2.757e-01  4.635e-01  -0.595  0.55196
## PWALAND     -4.405e-01  1.035e+00  -0.425  0.67052
## PPERSAUT     2.306e-01  4.199e-02   5.491 4.01e-08 ***
## PBESAUT      1.215e+01  4.029e+02   0.030  0.97595
## PMOTSCO     -8.101e-02  1.147e-01  -0.706  0.48006
## PVRAAUT     -2.106e+00  2.557e+03  -0.001  0.99934
## PAANHANG     1.014e+00  9.371e-01   1.082  0.27917
## PTRACTOR     7.229e-01  4.278e-01   1.690  0.09107 .
## PWERKT      -5.525e+00  4.805e+03  -0.001  0.99908
## PBROM        2.170e-01  4.865e-01   0.446  0.65559
## PLEVEN      -2.382e-01  1.170e-01  -2.036  0.04173 *
## PPERSONG    -4.523e-01  2.094e+00  -0.216  0.82901
## PGEZONG      1.444e+00  1.029e+00   1.404  0.16033
## PWAOREG      8.239e-01  5.943e-01   1.386  0.16565
## PBRAND       2.401e-01  7.714e-02   3.113  0.00185 **
## PZEILPL     -8.658e+00  3.261e+03  -0.003  0.99788
## PPLEZIER    -1.886e-01  3.259e-01  -0.579  0.56289
## PFIETS       3.664e-01  8.325e-01   0.440  0.65985
## PINBOED     -1.068e+00  8.764e-01  -1.219  0.22301
## PBYSTAND    -1.676e-01  3.321e-01  -0.505  0.61373
## AWAPART     -9.293e-01  7.802e-01  -1.191  0.23364
## AWABEDR      4.197e-01  1.082e+00   0.388  0.69824
```

```
## AWALAND      2.762e-01  3.528e+00   0.078  0.93758
## APERSAUT    -3.902e-02  1.772e-01  -0.220  0.82566
## ABESAUT     -7.298e+01  2.417e+03  -0.030  0.97591
## AMOTSCO      2.418e-01  3.772e-01   0.641  0.52142
## AVRAAUT     -4.490e+00  1.078e+04   0.000  0.99967
## AAANHANG    -1.351e+00  1.687e+00  -0.801  0.42322
## ATRACTOR    -2.376e+00  1.524e+00  -1.559  0.11899
## AWERKT      -8.749e-01  9.682e+03   0.000  0.99993
## ABROM       -1.060e+00  1.549e+00  -0.684  0.49367
## ALEVEN       4.789e-01  2.245e-01   2.133  0.03291 *
## APERSONG     3.997e-01  4.329e+00   0.092  0.92644
## AGEZONG     -3.163e+00  2.706e+00  -1.169  0.24247
## AWAOREG     -3.212e+00  3.433e+00  -0.936  0.34939
## ABRAND      -4.118e-01  2.787e-01  -1.477  0.13956
## AZEILPL      1.047e+01  3.261e+03   0.003  0.99744
## APLEZIER     2.516e+00  1.010e+00   2.490  0.01276 *
## AFIETS       2.318e-01  5.699e-01   0.407  0.68420
## AINBOED      1.947e+00  1.412e+00   1.378  0.16812
## ABYSTAND     1.078e+00  1.103e+00   0.977  0.32870
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2635.5  on 5821  degrees of freedom
## Residual deviance: 2243.5  on 5736  degrees of freedom
## AIC: 2415.5
##
## Number of Fisher Scoring iterations: 17
```

b.

```r
glm.prob <- predict(glm.obj2, type="response")
glm.pred <- ifelse(glm.prob > 0.50, "Yes", "No")

conf.mat <- table(glm.pred, Caravan$Purchase)
conf.mat
```

```
##
## glm.pred   No  Yes
##      No  5466  341
##      Yes    8    7
```

```r
mean(glm.pred == Caravan$Purchase)
```

```
## [1] 0.940055
```

```r
prop.table(conf.mat)
```

```
##
```

```
## glm.pred          No         Yes
##      No  0.938852628 0.058570938
##      Yes 0.001374098 0.001202336
```

```r
mean(Caravan$Purchase == "No")
```

```
## [1] 0.9402267
```

We have an overall accuracy of 94.0055% with this model. The false negative rate is 5.9% and the false positive rate is .12%.

If we were to just use a "model" of predicting that all custommers would purchase insurance, we would be have an accuracy of 94.02267%, which is actually more accurate than our model.

c.

```r
prop.table(conf.mat, 1)
```

```
##
## glm.pred         No        Yes
##      No  0.94127777 0.05872223
##      Yes 0.53333333 0.46666667
```

$P(\text{obesrveYes} \mid \text{predictYes}) = .46666667$

Problem 3

```r
head(Default)
```

```
##   default student   balance    income
## 1      No      No  729.5265 44361.625
## 2      No     Yes  817.1804 12106.135
## 3      No      No 1073.5492 31767.139
## 4      No      No  529.2506 35704.494
## 5      No      No  785.6559 38463.496
## 6      No     Yes  919.5885  7491.559
```

a.

```r
glm.obj3 <- glm(default~., family="binomial", data = Default)
summary(glm.obj3)
```

```
##
## Call:
## glm(formula = default ~ ., family = "binomial", data = Default)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4691  -0.1418  -0.0557  -0.0203   3.7383
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
## studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
## balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
## income       3.033e-06  8.203e-06   0.370  0.71152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1571.5  on 9996  degrees of freedom
## AIC: 1579.5
##
## Number of Fisher Scoring iterations: 8
```

All of the predictors in this model are significant except for income.

b.

```r
set.seed(1)
cv.errors <- cv.glm(Default, glm.obj3, K=10)$delta[1]
cv.errors
```

```
## [1] 0.02138616
```

test error: 0.02138616

If we did not use set.seed(1) we would get different test errors everytime we ran the code due to the random component in k-fold CV

c.

```r
set.seed(1)
glm.obj3.new <- glm(default~.-income, family = "binomial", data = Default)
cv.errors <- cv.glm(Default, glm.obj3.new, K=10)$delta[1]
cv.errors
```

```
## [1] 0.02136638
```

test error: 0.02136638

This test error is slightly lower with less variables compared to the model above with all the predictors included.

d.

```r
# using LOOCV
#cv.errors <- cv.glm(Default, glm.obj3)$delta[1]
#cv.errors
```

We did not use LOOCV and used 10-fold CV because it would take significantly longer to compute the test error using LOOCV. I used the code above to do LOOCV and my computer was performing the function for almost a minute until I halted the execution.