Homework#7

Neshma Simon

11/17/2020

Group Members: Fareha & Hertz

```
data_use1$earn_lastyr <- as.factor(data_use1$ERNYR_P)
levels(data_use1$earn_lastyr) <- c("0","$01-$4999","$5000-$9999","$10000-$14999","$15000-$19999","$2000

# We're trying to predict the health insurance of asian women with highschool and college education.
model_logit1 <- glm(NOTCOV ~ AGE_P + I(AGE_P^2) + female + Asian + educ_hs + educ_smcoll + educ_as + edu

d_region <- data.frame(model.matrix(~ data_use1$REGION))
d_region_born <- data.frame(model.matrix(~ factor(data_use1$region_born)))  # snips any with zero in the
dat_for_analysis_sub <- data.frame(
  data_use1$NOTCOV,
  data_use1$AGE_P,
  data_use1$female,
  data_use1$AfAm,
  data_use1$Asian,
  data_use1$RaceOther,
  data_use1$Hispanic,
  data_use1$educ_hs,
  data_use1$educ_smcoll,
  data_use1$educ_as,
  data_use1$educ_bach,
  data_use1$educ_adv,
  data_use1$married,
  data_use1$widowed,
  data_use1$divorc_sep,
  d_region[,2:4],
  d_region_born[,2:12])
names(dat_for_analysis_sub) <- c("NOTCOV",
                                 "Age",
                                 "female",
                                 "AfAm",
                                 "Asian",
                                 "RaceOther",
                                 "Hispanic",
                                 "educ_hs",
                                 "educ_smcoll",
                                 "educ_as",
                                 "educ_bach",
                                 "educ_adv",
                                 "married",
                                 "widowed",
```

```
                                   "divorc_sep",
                                   "Region.Midwest",
                                   "Region.South",
                                   "Region.West",
                                   "born.Mex.CentAm.Carib",
                                   "born.S.Am",
                                   "born.Eur",
                                   "born.f.USSR",
                                   "born.Africa",
                                   "born.MidE",
                                   "born.India.subc",
                                   "born.Asia",
                                   "born.SE.Asia",
                                   "born.elsewhere",
                                   "born.unknown")

require("standardize")
set.seed(654321)
NN <- length(dat_for_analysis_sub$NOTCOV)
restrict_1 <- as.logical(round(runif(NN,min=0,max=0.6))) # use fraction as training data
restrict_1 <- (runif(NN) < 0.1) # use 10% as training data
summary(restrict_1)
   Mode   FALSE    TRUE
logical  100833   11220
dat_train <- subset(dat_for_analysis_sub, restrict_1)
dat_test <- subset(dat_for_analysis_sub, !restrict_1)
sobj <- standardize(NOTCOV ~ Age + female + AfAm + Asian + RaceOther + Hispanic +
                      educ_hs + educ_smcoll + educ_as + educ_bach + educ_adv +
                      married + widowed + divorc_sep +
                      Region.Midwest + Region.South + Region.West +
                      born.Mex.CentAm.Carib + born.S.Am + born.Eur + born.f.USSR +
                      born.Africa + born.MidE + born.India.subc + born.Asia +
                      born.SE.Asia + born.elsewhere + born.unknown, dat_train, family = binomial)

# We use this code to predict using the test sets we created and use summary to look at the effect of di

s_dat_test <- predict(sobj, dat_test)
model_lpm1 <- lm(sobj$formula, data = sobj$data)
summary(model_lpm1)
Call:
lm(formula = sobj$formula, data = sobj$data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.61297 -0.12838 -0.08277 -0.02818  1.04196

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             0.727911   0.074442   9.778  < 2e-16 ***
Age                    -0.017009   0.004587  -3.708 0.000210 ***
female1                -0.007195   0.002942  -2.445 0.014496 *
AfAm1                  -0.011912   0.004393  -2.712 0.006704 **
Asian1                 -0.011946   0.008631  -1.384 0.166354
RaceOther1              0.062762   0.010463   5.999 2.05e-09 ***
Hispanic1               0.014111   0.004589   3.075 0.002111 **
```

```
educ_hs1              0.041919   0.004495   9.325  < 2e-16 ***
educ_smcoll1          0.025201   0.004867   5.178 2.28e-07 ***
educ_as1              0.025662   0.006105   4.204 2.65e-05 ***
educ_bach1           -0.004483   0.005306  -0.845 0.398202
educ_adv1            -0.015007   0.006708  -2.237 0.025298 *
married1             -0.019758   0.004348  -4.544 5.57e-06 ***
widowed1             -0.030269   0.008770  -3.451 0.000560 ***
divorc_sep1           0.008816   0.006337   1.391 0.164174
Region.Midwest1       0.014089   0.004897   2.877 0.004019 **
Region.South1         0.038268   0.004434   8.631  < 2e-16 ***
Region.West1          0.010803   0.004683   2.307 0.021099 *
born.Mex.CentAm.Carib1  0.150176   0.006269  23.956  < 2e-16 ***
born.S.Am1            0.086226   0.016372   5.267 1.42e-07 ***
born.Eur1             0.023163   0.013233   1.750 0.080074 .
born.f.USSR1          0.017727   0.032907   0.539 0.590116
born.Africa1          0.059734   0.017441   3.425 0.000617 ***
born.MidE1            0.063055   0.028634   2.202 0.027676 *
born.India.subc1      0.039849   0.016065   2.481 0.013132 *
born.Asia1            0.034574   0.015979   2.164 0.030501 *
born.SE.Asia1         0.038399   0.013319   2.883 0.003948 **
born.elsewhere1       0.004298   0.023403   0.184 0.854274
born.unknown1         0.026249   0.029657   0.885 0.376120
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3072 on 11154 degrees of freedom
Multiple R-squared:  0.1108,    Adjusted R-squared:  0.1086
F-statistic: 49.64 on 28 and 11154 DF,  p-value: < 2.2e-16
```

```
pred_vals_lpm <- predict(model_lpm1, s_dat_test)
pred_model_lpm1 <- (pred_vals_lpm > 0.5)
table(pred = pred_model_lpm1, true = dat_test$NOTCOV)
       true
pred        0     1
  FALSE 88187 12157
  TRUE    261   265
```

```
model_logit1 <- glm(sobj$formula, family = binomial, data = sobj$data)
summary(model_logit1)
Call:
glm(formula = sobj$formula, family = binomial, data = sobj$data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7804  -0.5018  -0.3986  -0.2632   2.9370

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             2.69782    0.87347   3.089 0.002011 **
Age                    -0.19411    0.05107  -3.801 0.000144 ***
female1                -0.07271    0.03133  -2.321 0.020308 *
AfAm1                  -0.09339    0.04789  -1.950 0.051191 .
Asian1                 -0.13093    0.10570  -1.239 0.215465
RaceOther1              0.47154    0.08467   5.569 2.56e-08 ***
Hispanic1               0.13459    0.04606   2.922 0.003475 **
```

```
educ_hs1               0.37334    0.04334   8.615  < 2e-16 ***
educ_smcoll1           0.23167    0.04953   4.677 2.91e-06 ***
educ_as1               0.23206    0.06380   3.637 0.000275 ***
educ_bach1            -0.20639    0.06920  -2.982 0.002861 **
educ_adv1             -0.58966    0.12690  -4.647 3.38e-06 ***
married1              -0.18068    0.04531  -3.988 6.67e-05 ***
widowed1              -0.42442    0.12056  -3.521 0.000431 ***
divorc_sep1            0.11064    0.06220   1.779 0.075258 .
Region.Midwest1        0.18917    0.06143   3.080 0.002073 **
Region.South1          0.43674    0.05330   8.194 2.53e-16 ***
Region.West1           0.16924    0.05698   2.970 0.002975 **
born.Mex.CentAm.Carib1  0.94692    0.05253  18.025  < 2e-16 ***
born.S.Am1             0.70720    0.12982   5.447 5.11e-08 ***
born.Eur1              0.27642    0.14314   1.931 0.053475 .
born.f.USSR1           0.14494    0.52524   0.276 0.782591
born.Africa1           0.57577    0.15597   3.692 0.000223 ***
born.MidE1             0.67806    0.26081   2.600 0.009327 **
born.India.subc1       0.48919    0.18127   2.699 0.006963 **
born.Asia1             0.41842    0.18432   2.270 0.023201 *
born.SE.Asia1          0.43475    0.15039   2.891 0.003841 **
born.elsewhere1        0.02772    0.27777   0.100 0.920511
born.unknown1          0.29552    0.31308   0.944 0.345211
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8222.7  on 11182  degrees of freedom
Residual deviance: 7160.6  on 11154  degrees of freedom
AIC: 7218.6

Number of Fisher Scoring iterations: 6
```

```
pred_vals <- predict(model_logit1, s_dat_test, type = "response")
pred_model_logit1 <- (pred_vals > 0.5)
table(pred = pred_model_logit1, true = dat_test$NOTCOV)
       true
pred        0     1
  FALSE 87299 11181
  TRUE   1149  1241
```

```
# We create a data object that is standardized and also divide it into trainding and test sets. Standard
```

```
require(stargazer)
stargazer(model_logit1, type = "text")
```

```
==================================================
                         Dependent variable:
                     ---------------------------
                                sobj
--------------------------------------------------
Age                           -0.194***
                               (0.051)

female1                       -0.073**
```

	(0.031)
AfAm1	-0.093*
	(0.048)
Asian1	-0.131
	(0.106)
RaceOther1	0.472***
	(0.085)
Hispanic1	0.135***
	(0.046)
educ_hs1	0.373***
	(0.043)
educ_smcoll1	0.232***
	(0.050)
educ_as1	0.232***
	(0.064)
educ_bach1	-0.206***
	(0.069)
educ_adv1	-0.590***
	(0.127)
married1	-0.181***
	(0.045)
widowed1	-0.424***
	(0.121)
divorc_sep1	0.111*
	(0.062)
Region.Midwest1	0.189***
	(0.061)
Region.South1	0.437***
	(0.053)
Region.West1	0.169***
	(0.057)
born.Mex.CentAm.Carib1	0.947***
	(0.053)
born.S.Am1	0.707***
	(0.130)
born.Eur1	0.276*

```
                                (0.143)

born.f.USSR1                     0.145
                                (0.525)

born.Africa1                    0.576***
                                (0.156)

born.MidE1                      0.678***
                                (0.261)

born.India.subc1                0.489***
                                (0.181)

born.Asia1                      0.418**
                                (0.184)

born.SE.Asia1                   0.435***
                                (0.150)

born.elsewhere1                  0.028
                                (0.278)

born.unknown1                    0.296
                                (0.313)

Constant                        2.698***
                                (0.873)

--------------------------------------------------
Observations                    11,183
Log Likelihood                -3,580.288
Akaike Inf. Crit.              7,218.577
==================================================
Note:                  *p<0.1; **p<0.05; ***p<0.01
```

```
require(e1071)
svm.model <- svm(as.factor(NOTCOV) ~ ., data = sobj$data, cost = 10, gamma = 0.1)
svm.pred <- predict(svm.model, s_dat_test)
table(pred = svm.pred, true = dat_test$NOTCOV)
    true
pred     0     1
   0 86565 10432
   1  1883  1990
```

```
require('randomForest')
set.seed(54321)
model_randFor <- randomForest(as.factor(NOTCOV) ~ ., data = sobj$data, importance=TRUE, proximity=TRUE)
print(model_randFor)

Call:
  randomForest(formula = as.factor(NOTCOV) ~ ., data = sobj$data,      importance = TRUE, proximity = T
Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 5
```

```
OOB estimate of  error rate: 11.57%
Confusion matrix:
  0   1 class.error
0 9778  59 0.005997764
1 1235 111 0.917533432
```

```
round(importance(model_randFor),2)
0      1 MeanDecreaseAccuracy MeanDecreaseGini
Age                  27.75   7.19                30.47           224.63
female                3.05  -3.52                 0.57            22.64
AfAm                 -5.05   9.69                 1.49            15.69
Asian                13.47  -9.70                10.37             9.56
RaceOther             1.52  10.07                 6.28            14.58
Hispanic             -6.50  19.02                16.71            52.34
educ_hs              16.68  -9.48                13.62            21.47
educ_smcoll          15.21  -5.48                13.98            14.54
educ_as              11.91  -1.71                11.12            11.81
educ_bach            14.89  11.59                17.21            19.00
educ_adv             13.73  19.76                21.33            13.58
married              20.92 -16.08                20.64            27.93
widowed               2.67  -1.08                 2.41             6.98
divorc_sep            9.47  -2.83                 9.07            13.07
Region.Midwest        1.84  -0.95                 1.30            11.19
Region.South          7.04   8.54                12.13            23.13
Region.West           7.81  -2.98                 7.57            13.72
born.Mex.CentAm.Carib  3.55  43.51               29.50           123.52
born.S.Am            -7.66  14.29                -1.11             7.84
born.Eur              4.79   0.40                 4.79             6.38
born.f.USSR          -5.68  -0.97                -5.70             0.84
born.Africa          -1.21   9.93                 3.99             5.55
born.MidE             7.20  -0.10                 6.85             4.08
born.India.subc       5.73  -3.47                 4.99             4.67
born.Asia             0.76   1.40                 1.29             4.45
born.SE.Asia          0.90   4.56                 2.34             5.84
born.elsewhere        3.73  -4.00                 2.60             3.43
born.unknown         -5.86  -2.86                -6.30             1.94
```

```
varImpPlot(model_randFor)
pred_model1 <- predict(model_randFor,  s_dat_test)
table(pred = pred_model1, true = dat_test$NOTCOV)
true
pred     0     1
0 87845 11178
1   603  1244
```

```
# Elastic Net
require(glmnet)
model1_elasticnet <-  glmnet(as.matrix(sobj$data[,-1]),sobj$data$NOTCOV)
# default is alpha = 1, lasso
-lasso only selects variables that are important for the predcitions.
par(mar=c(4.5,4.5,1,4))
plot(model1_elasticnet)
#See PLOT1
vnat=coef(model1_elasticnet)
```

```
vnat=vnat[-1,ncol(vnat)] # remove the intercept, and get the coefficients at the end of the path
axis(4, at=vnat,line=-.5,label=names(sobj$data[,-1]),las=1,tick=FALSE,cex.axis=0.5)
#See PLOT2
```

```
plot(model1_elasticnet, xvar = "lambda")
#See PLOT4
plot(model1_elasticnet, xvar = "dev", label = TRUE)
#See PLOT4
print(model1_elasticnet)
Call:  glmnet(x = as.matrix(sobj$data[, -1]), y = sobj$data$NOTCOV)

   Df  %Dev   Lambda
1   0  0.00 0.088570
2   1  1.26 0.080700
3   1  2.30 0.073530
4   1  3.17 0.067000
5   1  3.89 0.061050
6   1  4.49 0.055630
7   1  4.98 0.050680
8   1  5.40 0.046180
9   1  5.74 0.042080
10  1  6.02 0.038340
11  1  6.26 0.034930
12  2  6.48 0.031830
13  2  6.70 0.029000
14  2  6.87 0.026430
15  3  7.06 0.024080
16  4  7.35 0.021940
17  5  7.63 0.019990
18  8  7.94 0.018210
19  8  8.27 0.016600
20  9  8.55 0.015120
21 10  8.80 0.013780
22 11  9.04 0.012550
23 11  9.26 0.011440
24 11  9.45 0.010420
25 11  9.60 0.009497
26 12  9.73 0.008654
27 13  9.86 0.007885
28 14  9.98 0.007184
29 16 10.09 0.006546
30 16 10.20 0.005965
31 19 10.31 0.005435
32 19 10.42 0.004952
33 19 10.50 0.004512
34 20 10.57 0.004111
35 22 10.64 0.003746
36 22 10.70 0.003413
37 22 10.75 0.003110
38 23 10.79 0.002834
39 23 10.83 0.002582
40 23 10.86 0.002353
41 24 10.89 0.002144
42 25 10.92 0.001953
```

```
43 25 10.94 0.001780
44 25 10.96 0.001622
45 25 10.98 0.001477
46 25 11.00 0.001346
47 25 11.01 0.001227
48 27 11.02 0.001118
49 27 11.03 0.001018
50 27 11.04 0.000928
51 27 11.04 0.000846
52 27 11.05 0.000770
53 27 11.06 0.000702
54 27 11.06 0.000640
55 27 11.06 0.000583
56 27 11.07 0.000531
57 27 11.07 0.000484
58 27 11.07 0.000441
59 28 11.07 0.000402
60 28 11.07 0.000366
61 28 11.08 0.000333
62 28 11.08 0.000304
63 28 11.08 0.000277
64 28 11.08 0.000252
65 28 11.08 0.000230
66 28 11.08 0.000209
67 28 11.08 0.000191
68 28 11.08 0.000174
69 28 11.08 0.000158
70 28 11.08 0.000144
71 28 11.08 0.000131
72 28 11.08 0.000120
73 28 11.08 0.000109
74 28 11.08 0.000099
75 28 11.08 0.000091
```

```
cvmodel1_elasticnet = cv.glmnet(data.matrix(sobj$data[,-1]),data.matrix(sobj$data$NOTCOV))
> cvmodel1_elasticnet$lambda.min
[1] 0.0001443494

log(cvmodel1_elasticnet$lambda.min)
[1] -8.843274

coef(cvmodel1_elasticnet, s = "lambda.min")
29 x 1 sparse Matrix of class "dgCMatrix"
                                 1
(Intercept)            2.702811502
Age                   -0.016789554
female                 0.014075640
AfAm                   0.023175664
Asian                  0.020474503
RaceOther             -0.124977273
Hispanic              -0.028413577
educ_hs               -0.082913192
educ_smcoll           -0.049456270
educ_as               -0.050130906
educ_bach              0.009229021
```

```
educ_adv              0.030080994
married               0.038868601
widowed               0.059846243
divorc_sep           -0.017459103
Region.Midwest       -0.026492937
Region.South         -0.074880389
Region.West          -0.020042744
born.Mex.CentAm.Carib -0.299403382
born.S.Am            -0.169856235
born.Eur             -0.044589247
born.f.USSR          -0.031071025
born.Africa          -0.116969502
born.MidE            -0.122811448
born.India.subc      -0.074507298
born.Asia            -0.064225767
born.SE.Asia         -0.072313410
born.elsewhere       -0.005630858
born.unknown         -0.048605718
```

```
pred1_elasnet <- predict(model1_elasticnet, newx = data.matrix(s_dat_test), s = cvmodel1_elasticnet$laml
pred_model1_elasnet <- (pred1_elasnet < mean(pred1_elasnet))
table(pred = pred_model1_elasnet, true = dat_test$NOTCOV)
       true
pred        0     1
  FALSE 60142  4362
  TRUE  28306  8060
```

```
model2_elasticnet <-  glmnet(as.matrix(sobj$data[,-1]),sobj$data$NOTCOV, alpha = 0)
```