Homework #7

Collin Rafferty

11/10/2021

```
load("/cloud/project/Household_Pulse_data.RData")

Household_Pulse_data$vaxx <- (Household_Pulse_data$RECVDVACC == "yes got
vaxx")
is.na(Household_Pulse_data$vaxx) <- which(Household_Pulse_data$RECVDVACC ==
"NA")

# Like last week, I decided to set the NAs equal to zero.

require(gtsummary)

## Loading required package: gtsummary

## #BlackLivesMatter

tb1 <- data.frame(Col = Household_Pulse_data$vaxx,
 Row = Household_Pulse_data$EEDUC)
tbl_cross(tb1, row=Row, col=Col, percent="row")

## Table printed with {flextable}, not {gt}. Learn why at
## http://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in the code chunk
header.
```

	Col			
Characteristic	FALSE	TRUE	Unknown	Total
Row				
less than hs	115 (28%)	290 (71%)	6 (1.5%)	411 (100%)
some hs	269 (29%)	652 (70%)	15 (1.6%)	936 (100%)
HS diploma	1,647 (21%)	6,097 (78%)	113 (1.4%)	7,857 (100%)
some coll	2,396 (16%)	12,022 (82%)	178 (1.2%)	14,596 (100%)
assoc deg	1,132 (15%)	6,266 (83%)	110 (1.5%)	7,508 (100%)
bach deg	1,565 (7.8%)	18,272 (91%)	238 (1.2%)	20,075 (100%)
adv deg	813 (4.6%)	16,727 (94%)	191 (1.1%)	17,731 (100%)

	Col			
Characteristic	FALSE	TRUE	Unknown	Total
Total	7,937 (11%)	60,326 (87%)	851 (1.2%)	69,114 (100%)

```
tb2 <- data.frame(Col = Household_Pulse_data$vaxx,
 Row = Household_Pulse_data$RRACE)
tbl_cross(tb2, row=Row, col=Col, percent="row")

## Table printed with {flextable}, not {gt}. Learn why at
## http://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in the code chunk
header.
```

	Col			
Characteristic	FALSE	TRUE	Unknown	Total
Row				
White	6,349 (11%)	49,875 (88%)	714 (1.3%)	56,938 (100%)
Black	834 (15%)	4,510 (83%)	68 (1.3%)	5,412 (100%)
Asian	134 (3.8%)	3,401 (96%)	26 (0.7%)	3,561 (100%)
Other	620 (19%)	2,540 (79%)	43 (1.3%)	3,203 (100%)
Total	7,937 (11%)	60,326 (87%)	851 (1.2%)	69,114 (100%)

```
tb3 <- data.frame(Col = Household_Pulse_data$vaxx,
 Row = Household_Pulse_data$REGION)
tbl_cross(tb3, row=Row, col=Col, percent="row")

## Table printed with {flextable}, not {gt}. Learn why at
## http://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in the code chunk
header.
```

	Col			
Characteristic	FALSE	TRUE	Unknown	Total
Row				
Northeast	828 (7.9%)	9,532 (91%)	118 (1.1%)	10,478 (100%)
South	2,913 (13%)	19,499 (86%)	268 (1.2%)	22,680 (100%)
Midwest	1,729 (13%)	11,714 (86%)	208 (1.5%)	13,651 (100%)
West	2,467 (11%)	19,581 (88%)	257 (1.2%)	22,305 (100%)
Total	7,937 (11%)	60,326 (87%)	851 (1.2%)	69,114 (100%)

```
pick_use1 <- (Household_Pulse_data$REGION == "South" &
Household_Pulse_data$GENID_DESCRIBE== "male" & Household_Pulse_data$RRACE==
"White")
dat_use1 <- subset(Household_Pulse_data, pick_use1)
```

I thought it would be interesting to look at a subset of only white males living in the south because, in lab 6, I found them to have the lowest probability of being vaccinated out of the subset I was looking at.

```
d_marstat <- data.frame(model.matrix(~ dat_use1$MS))
d_pubhlth<- data.frame(model.matrix(~ dat_use1$PUBHLTH))
d_x <- data.frame(model.matrix(~ dat_use1$SEXUAL_ORIENTATION))
d_income <- data.frame(model.matrix(~ dat_use1$INCOME))
d_educ <- data.frame(model.matrix(~ dat_use1$EEDUC))
d_vaxx <- data.frame(model.matrix(~ dat_use1$vaxx))

dat_for_analysis_sub <- data.frame(
  d_vaxx[,2],
  d_educ[!is.na(dat_use1$vaxx),2:7],
  d_marstat[!is.na(dat_use1$vaxx),2:6],
  d_income[!is.na(dat_use1$vaxx),2:9],
  d_pubhlth[!is.na(dat_use1$vaxx),2:3],
  d_x[!is.na(dat_use1$vaxx),2:6])

names(dat_for_analysis_sub) <-
sub("dat_use1.","",names(dat_for_analysis_sub))
names(dat_for_analysis_sub)[1] <- "vaxx"

summary(d_vaxx)

##   X.Intercept. dat_use1.vaxxTRUE
##  Min.   :1     Min.   :0.0000
##  1st Qu.:1     1st Qu.:1.0000
##  Median :1     Median :1.0000
##  Mean   :1     Mean   :0.8899
##  3rd Qu.:1     3rd Qu.:1.0000
##  Max.   :1     Max.   :1.0000

summary(d_educ)

##   X.Intercept. dat_use1.EEDUCsome.hs dat_use1.EEDUCHS.diploma
##  Min.   :1     Min.   :0.00000       Min.   :0.0000
##  1st Qu.:1     1st Qu.:0.00000       1st Qu.:0.0000
##  Median :1     Median :0.00000       Median :0.0000
##  Mean   :1     Mean   :0.01314       Mean   :0.1025
##  3rd Qu.:1     3rd Qu.:0.00000       3rd Qu.:0.0000
##  Max.   :1     Max.   :1.00000       Max.   :1.0000
##  dat_use1.EEDUCsome.coll dat_use1.EEDUCassoc.deg dat_use1.EEDUCbach.deg
##  Min.   :0.0000          Min.   :0.00000         Min.   :0.0000
##  1st Qu.:0.0000          1st Qu.:0.00000         1st Qu.:0.0000
```

```
##  Median :0.0000          Median :0.00000         Median :0.0000
##  Mean   :0.1971          Mean   :0.08772         Mean   :0.3119
##  3rd Qu.:0.0000          3rd Qu.:0.00000         3rd Qu.:1.0000
##  Max.   :1.0000          Max.   :1.00000         Max.   :1.0000
##  dat_use1.EEDUCadv.deg
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2814
##  3rd Qu.:1.0000
##  Max.   :1.0000
```

```
require("standardize")
```

```
## Loading required package: standardize
```

```
##
##  ************************************************************
##           Loading standardize package version 0.2.2
##      Call standardize.news() to see new features/changes
##  ************************************************************
```

```
set.seed(12345)
NN <- length(dat_for_analysis_sub$vaxx)
restrict_1 <- (runif(NN) < 0.3)
summary(restrict_1)
```

```
##    Mode   FALSE    TRUE
## logical    4793    2039
```

```
dat_train <- subset(dat_for_analysis_sub, restrict_1)
dat_test <- subset(dat_for_analysis_sub, !restrict_1)
```

```
summary(dat_test$INCOMEHH.income..200k..)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0       0       0       0
```

```
sobj <- standardize(vaxx ~  EEDUCsome.hs + EEDUCHS.diploma + EEDUCsome.coll +
EEDUCassoc.deg + EEDUCbach.deg + EEDUCadv.deg + MSmarried + MSwidowed +
MSdivorced + MSseparated +PUBHLTHno.public.health.ins+
MSnever+INCOMEHH.income.less.than..25k+INCOMEHH.income..25k....34.9k+INCOMEHH
.income..35k...49.9+
INCOMEHH.income..50k...74.9+INCOMEHH.income..75...99.9+INCOMEHH.income..100k.
..149+INCOMEHH.income..150...199+ SEXUAL_ORIENTATIONbisexual+
SEXUAL_ORIENTATIONdont.know+SEXUAL_ORIENTATIONgay.or.lesbian+SEXUAL_ORIENTATI
ONsomething.else+SEXUAL_ORIENTATIONstraight, dat_train, family = binomial)
```

```
summary(sobj$data)
```

```
##       vaxx        EEDUCsome.hs EEDUCHS.diploma EEDUCsome.coll
EEDUCassoc.deg
```

```
##  Min.   :0.0000   1:  33       1: 203          1: 425         1: 192
##  1st Qu.:1.0000   0:2006       0:1836          0:1614         0:1847
##  Median :1.0000
##  Mean   :0.8813
##  3rd Qu.:1.0000
##  Max.   :1.0000
##  EEDUCbach.deg EEDUCadv.deg MSmarried MSwidowed MSdivorced MSseparated
##  1: 600        1: 568       1:1354    1:  79    1: 229     1:  28
##  0:1439        0:1471       0: 685    0:1960    0:1810     0:2011
##
##
##
##
##  PUBHLTHno.public.health.ins MSnever  INCOMEHH.income.less.than..25k
##  1: 933                      1: 341   1: 116
##  0:1106                      0:1698   0:1923
##
##
##
##
##  INCOMEHH.income..25k....34.9k INCOMEHH.income..35k...49.9
##  1: 158                        1: 279
##  0:1881                        0:1760
##
##
##
##
##  INCOMEHH.income..50k...74.9 INCOMEHH.income..75...99.9
##  1: 226                      1: 332
##  0:1813                      0:1707
##
##
##
##
##  INCOMEHH.income..100k...149 INCOMEHH.income..150...199
##  1: 162                      1: 245
##  0:1877                      0:1794
##
##
##
##
##  SEXUAL_ORIENTATIONbisexual SEXUAL_ORIENTATIONdont.know
##  1:  31                     1:  21
##  0:2008                     0:2018
##
##
##
##
##  SEXUAL_ORIENTATIONgay.or.lesbian SEXUAL_ORIENTATIONsomething.else
##  1: 119                           1:   9
```

```
##  0:1920                            0:2030
##
##
##
##
##  SEXUAL_ORIENTATIONstraight
##  1:1842
##  0: 197
##
##
##
##
```

```
s_dat_test <- predict(sobj, dat_test)
```

After running a summary on incomes over 200K, I found its Min=Max=0, so I dropped it from the model. I didn't want it to cause an error in the model.

```
require(stargazer)
```

```
## Loading required package: stargazer
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary
Statistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
model_lpm1 <- lm(sobj$formula, data = sobj$data)
stargazer::stargazer(model_lpm1, type="text", title = "Model_lmp1")
```

```
##
## Model_lmp1
## ===============================================================
##                                       Dependent variable:
##                                   ---------------------------
##                                              sobj
## ---------------------------------------------------------------
## EEDUCsome.hs1                               -0.134***
##                                              (0.046)
##
## EEDUCHS.diploma1                             -0.061
##                                              (0.039)
##
## EEDUCsome.coll1                              -0.050
##                                              (0.038)
##
## EEDUCassoc.deg1                              -0.005
##                                              (0.039)
##
```

```
## EEDUCbach.deg1                                0.015
##                                              (0.038)
##
## EEDUCadv.deg1                                 0.021
##                                              (0.038)
##
## MSmarried1                                   -0.031
##                                              (0.055)
##
## MSwidowed1                                   -0.012
##                                              (0.058)
##
## MSdivorced1                                  -0.027
##                                              (0.056)
##
## MSseparated1                                 -0.073
##                                              (0.063)
##
## PUBHLTHno.public.health.ins1                -0.038***
##                                              (0.008)
##
## MSnever1                                     -0.046
##                                              (0.056)
##
## INCOMEHH.income.less.than..25k1               0.031*
##                                              (0.016)
##
## INCOMEHH.income..25k....34.9k1                0.017
##                                              (0.014)
##
## INCOMEHH.income..35k...49.91                  0.013
##                                              (0.012)
##
## INCOMEHH.income..50k...74.91                  0.021
##                                              (0.013)
##
## INCOMEHH.income..75...99.91                   0.015
##                                              (0.012)
##
## INCOMEHH.income..100k...1491                 0.032**
##                                              (0.015)
##
## INCOMEHH.income..150...1991                  0.029**
##                                              (0.014)
##
## SEXUAL_ORIENTATIONbisexual1                  0.118**
##                                              (0.047)
##
## SEXUAL_ORIENTATIONdont.know1                 -0.016
##                                              (0.051)
```

```
##
## SEXUAL_ORIENTATIONgay.or.lesbian1           0.074*
##                                             (0.041)
##
## SEXUAL_ORIENTATIONsomething.else1            0.052
##                                             (0.065)
##
## SEXUAL_ORIENTATIONstraight1                  0.029
##                                             (0.038)
##
## Constant                                   0.876***
##                                             (0.269)
##
## ------------------------------------------------------------
## Observations                                 2,039
## R2                                           0.085
## Adjusted R2                                  0.074
## Residual Std. Error                  0.311 (df = 2014)
## F Statistic                       7.782*** (df = 24; 2014)
## ============================================================
## Note:                              *p<0.1; **p<0.05; ***p<0.01
```

```
pred_vals_lpm <- predict(model_lpm1, s_dat_test)

pred_model_lpm1 <- (pred_vals_lpm > mean(pred_vals_lpm))
table(pred = pred_model_lpm1, true = dat_test$vaxx)
```

```
##        true
## pred       0    1
##   FALSE  350 1762
##   TRUE   160 2521
```

```
# logit
model_logit1 <- glm(sobj$formula, family = binomial, data = sobj$data)
stargazer::stargazer(model_logit1, type="text", title = "Model_logit1")
```

```
##
## Model_logit1
## ============================================================
##                                      Dependent variable:
##                                  ------------------------
##                                             sobj
## ------------------------------------------------------------
## EEDUCsome.hs1                              -0.810*
##                                            (0.429)
##
## EEDUCHS.diploma1                           -0.466
##                                            (0.398)
##
## EEDUCsome.coll1                            -0.413
##                                            (0.393)
```

```
##
## EEDUCassoc.deg1                          -0.054
##                                          (0.406)
##
## EEDUCbach.deg1                            0.197
##                                          (0.399)
##
## EEDUCadv.deg1                             0.335
##                                          (0.404)
##
## MSmarried1                               -6.935
##                                         (408.216)
##
## MSwidowed1                               -6.726
##                                         (408.216)
##
## MSdivorced1                              -6.898
##                                         (408.216)
##
## MSseparated1                             -7.169
##                                         (408.216)
##
## PUBHLTHno.public.health.ins1            -0.410***
##                                          (0.080)
##
## MSnever1                                 -7.054
##                                         (408.216)
##
## INCOMEHH.income.less.than..25k1           0.303*
##                                          (0.161)
##
## INCOMEHH.income..25k....34.9k1            0.189
##                                          (0.141)
##
## INCOMEHH.income..35k...49.91              0.141
##                                          (0.116)
##
## INCOMEHH.income..50k...74.91              0.207
##                                          (0.135)
##
## INCOMEHH.income..75...99.91               0.165
##                                          (0.119)
##
## INCOMEHH.income..100k...1491             0.384**
##                                          (0.188)
##
## INCOMEHH.income..150...1991              0.330**
##                                          (0.163)
##
## SEXUAL_ORIENTATIONbisexual1               7.835
```

```
##                                              (201.099)
##
## SEXUAL_ORIENTATIONdont.know1                   0.015
##                                               (0.433)
##
## SEXUAL_ORIENTATIONgay.or.lesbian1             0.861**
##                                               (0.411)
##
## SEXUAL_ORIENTATIONsomething.else1              0.483
##                                               (0.655)
##
## SEXUAL_ORIENTATIONstraight1                    0.260
##                                               (0.350)
##
## Constant                                      -9.614
##                                             (1,241.050)
##
## ---------------------------------------------------------------
## Observations                                    2,039
## Log Likelihood                                -656.697
## Akaike Inf. Crit.                             1,363.394
## ===============================================================
## Note:                               *p<0.1; **p<0.05; ***p<0.01
```

```
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$vaxx)
```

```
##        true
## pred       0    1
##   FALSE    6    5
##   TRUE   504 4278
```

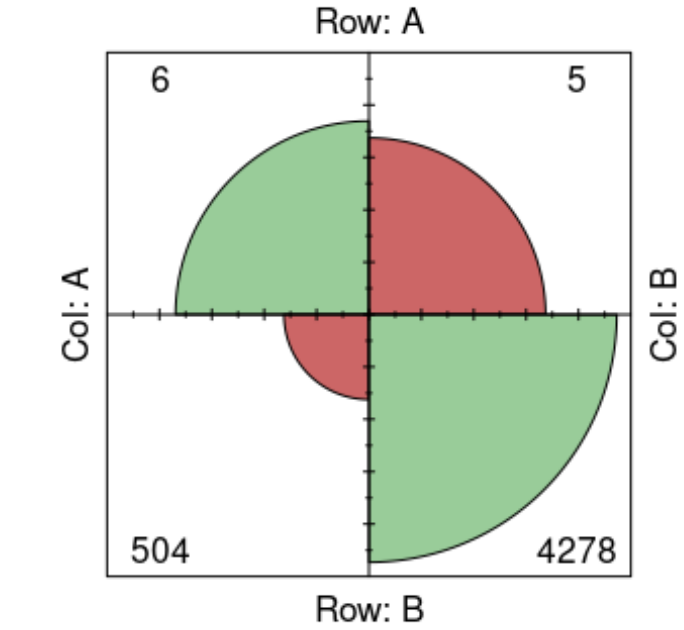
```
pred_model_logit2 <- (pred_vals > 0.3)
table(pred = pred_model_logit2, true = dat_test$vaxx)
```

```
##       true
## pred      0    1
##   TRUE  510 4283
```

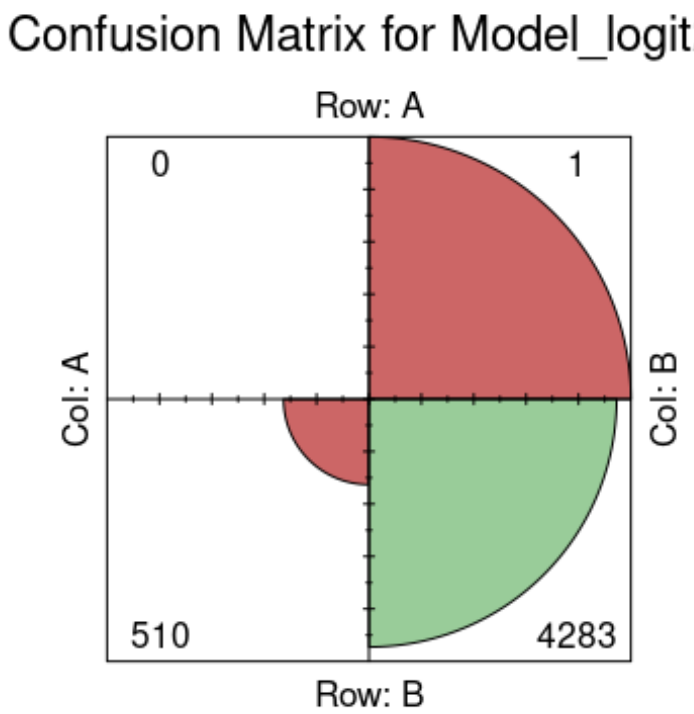
```
pred_model_logit3 <- (pred_vals > 0.8)
table(pred = pred_model_logit3, true = dat_test$vaxx)
```

```
##        true
## pred       0    1
##   FALSE  210  597
##   TRUE   300 3686
```

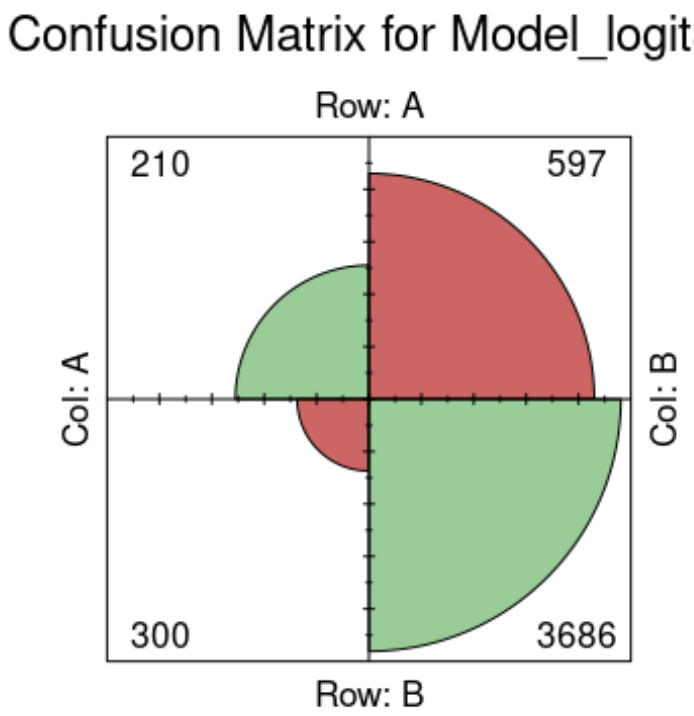
```
cm1 <- as.table(matrix(c(6, 5, 504, 4278), nrow = 2, byrow = TRUE))
fourfoldplot(cm1, color = c("#CC6666", "#99CC99"),
conf.level = 0, margin = 1, main = "Confusion Matrix for Model_logit1 ")
```




```
cm2 <- as.table(matrix(c(0, 1, 510, 4283), nrow = 2, byrow = TRUE))
fourfoldplot(cm2, color = c("#CC6666", "#99CC99"),
conf.level = 0, margin = 1, main = "Confusion Matrix for Model_logit2 ")
```



```
cm3 <- as.table(matrix(c(210, 597, 300, 3686), nrow = 2, byrow = TRUE))
fourfoldplot(cm3, color = c("#CC6666", "#99CC99"),
conf.level = 0, margin = 1, main = "Confusion Matrix for Model_logit3 ")
```



The tables and graphs show as the model's cutoff value is reduced, the model's predictive accuracy increases to a point. For example, when the cutoff value was set to pred_vals > .8, the model made 4283 true positive predictions, but when the cutoff value was set to pred_vals > .5, the model only predicted 3686 true positives. Both false negative and false positive also increase substantially from model one to model three. However, there is a certain point when the cutoff value is too low, and the model's predictive accuracy declines. For example, looking at model two, where the cutoff was set to pred_vals >.3, it predicted 4278 true positives. This is five fewer true positives than model one, in which the cutoff value was set to pred_vals >.5.