Homework Assignment 1: Solutions

DPI 610

Assignment Details

This homework assignment is due 2/19/2021 before midnight. Everyone must complete their own code and assignment, though you may discuss the homework with classmates.

To access the code and data for the assignment, either go to the FAS OnDemand cluster or download the zip file posted under HW1 on Canvas. To turn in the assignment, upload two files to Canvas: (1) your code file (.Rmd), and (2) a PDF version (click Knit to create PDF).

Technical Notes: First, for any questions involving CCES survey data, *ignore the survey weights*. We will learn about these later in the course. Second, part of this assignment is designed to give you some experience figuring out how to use key Base R functions. To look up help on a built-in R function, type `?` before the name of the function. For example, to look up help for `aggregate()`, type `?aggregate` in the console. Third, for training models in this homework assignment, note that I have split the North Carolina Voter data into a training set and a test set. The vector `train_id` contains the `id` numbers for people in the training set, and the vector `test_id` contains `id` numbers for people in the test set.

Setup

The following code loads the data set and packages necessary to complete the problem set. Please make sure to run the following code before starting the problem set.

```r
# Load Data
dat <- readRDS("nc_data.rds")
nc         <- dat$nc
nc_cces    <- dat$nc_cces
dictionary <- dat$dictionary

# Load packages
library(tidyverse)
library(pROC)
library(glmnet)

# Split NC voter data into training and test set
set.seed(02138)
n_train  <- floor(nrow(nc) * 0.75)
n_test   <- nrow(nc) - n_train
nc_split <- sample(c(rep(TRUE, n_train), rep(FALSE, n_test)))

train_id <- nc[nc_split == TRUE, "id"]
test_id  <- nc[nc_split == FALSE,"id"]
```

Question 1: Obama/Trump Voters

The CCES is a survey conducted before and after presidential and election years. There are slightly more than 2000 respondents for North Carolina. The survey data is an important tool for understanding the political views of voters. This question asks you to use some R functions to summarize some key insights from the data.

Question 1(a)

How many voters in the survey data voted for Barack Obama in 2012 and Donald Trump in 2016? What share of the full sample is this? (Hint: Look at the variables `voted_pres_12` and `voted_pres_16`.)

Answer 1(a)

```r
count <- nc_cces %>%
  filter(voted_pres_12 == "Barack Obama",
         voted_pres_16 == "Donald Trump") %>%
  count()

share <- round(count/nrow(nc_cces)*100)
```

48 voters in the survey data voted for Barack Obama in 2012 and Donald Trump in 2016, making for about 2% of the full sample.

Question 1(b)

Define a new indicator variable in `nc_cces` called `obama_trump` that takes `1` if a respondent voted for Barack Obama in 2012 and switched to Donald Trump in 2016, and takes `0` otherwise.

Answer 1(b)

```r
# Insert Answer Here
```

Question 1(c)

How do "Obama-Trump" voters in the CCES sample compare to the other survey respondents in terms of age, family income, and having a college education? Compare the two group in terms of the mean of age (`age`), the mean of family income (`familyincome`), and the proportion of college graduates (`college_grad_prob`).

Answer 1 (c)

```r
# Insert Answer Here
```

Question 1(d)

How do "Obama-Trump" voters in this sample compare to other survey respondents in terms of their stated partisan identification? (Hint: Look at variable `pid3`.) Report your answer as a share of the voters in each group (i.e., Democrats comprise XX% of all Obama-Trump voters, Republicans comprise XX% of all Obama-Trump voters, and so on). (Hint: You might want to use the `table()` and `prop.table()` functions.)

Answer 1(d)

```
# Insert Answer Here
```

Question 2: Turnout Model with Logistic Regression

This question uses North Carolina voter data to explore models of voter turnout.

Suppose we examined a simple model of turnout using logistic regression that took the following form:

```
## formula
fm <- voted2014 ~ voted2008 + voted2010 + voted2012 +
        age + I(age^2) + factor(party) + log(familyincome+1)+
        factor(density) + homeowner_score + married_prob + factor(socioecon_bg)

## estimate model (logistic regression)
model <- glm(fm, data = nc, subset= id %in% train_id,
             family = "binomial", na.action = na.exclude)
```

Question 2(a)

What is the theoretical basis for including past voter behavior in a model of voter turnout?

Answer 2(a)

Insert answer here.

Question 2(b)

You can explore the estimates from the model by typing `summary(model)` in the console. The coefficient on age is 0.062. What is the interpretation of this coefficient? Given this estimate along with the estimate for age-squared, how do we think about the relationship between age and turnout?

Answer 2(b)

Insert answer here.

Question 2(c)

Use the model to produce a predicted probability of turnout for each person in both the test and training data sets. Call it `yhat`. Create a histogram of the predicted probabilities.

Answer 2(c)

```r
#Insert answer here
```

Question 2(d)

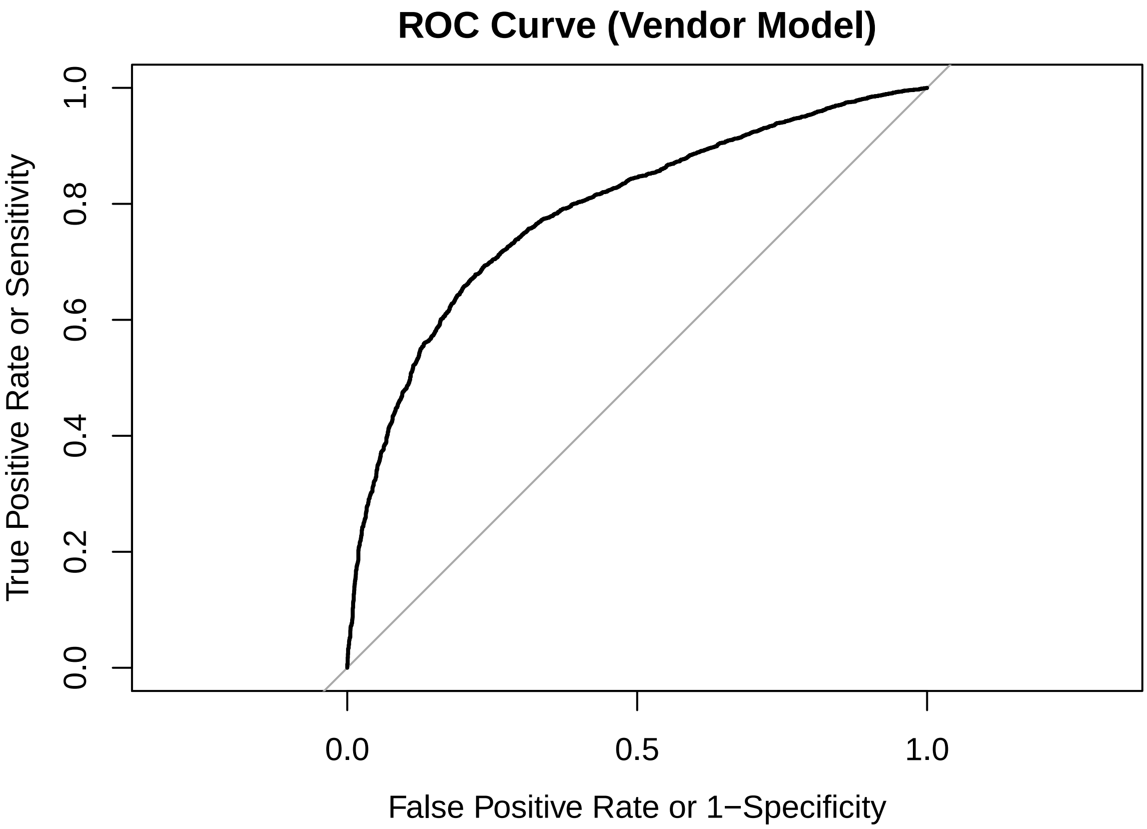
ROC Curve

Previously we have evaluated predictions by comparing them to observed outcomes and examining the accuracy, recall and precision. We are now going to introduce another diagnostic tool, the Receiver Operating Characteristic (ROC) Curve. As we change the threshold for identifying a "positive" prediction based on our model (i.e., predicting a voter will turnout), the relationship between precision and recall also changes. The ROC curve focuses on positive identifications and plots the False Positive Rate (x-axis) against the True Positive Rate (y-axis).

We will use the function `roc()` from the `pROC` R package (loaded in the setup code chunk) to plot a ROC curve. Note we perform this operation within the training data, not the test data.

Let's start by plotting the ROC curve for the predictions already made by the *vendor* in the variable `vote2014_prob`.

```r
# compute roc curve
nc.roc <- roc(
  response  = nc$voted2014[nc$id %in% train_id],
  predictor = nc$vote2014_prob[nc$id %in% train_id]/100
)

# plot the ROC curve
plot(nc.roc,
     xlab = "False Positive Rate or 1-Specificity",
     ylab = "True Positive Rate or Sensitivity",
     main = "ROC Curve (Vendor Model)",
     legacy.axes = TRUE)
```



Create an ROC curve for the predictions from *our* model of voter turnout. Assign it to an object called `nc.roc2` and then use the `plot()` function to visualize the ROC curve.

Answer 2(d)

```
#Insert answer here
```

Question 2(e)

A variety of methods can be used to determine the threshold for predicting someone voted versus not voted. One is to place relative weights on the cost of a false positive and a false negative. Suppose we decide the cost of a false negative is 1/2 the cost of a false positive. Then we can use the `coords()` function to determine the optimal threshold.

```
# calculate the best threshold
threshold <- coords(roc = nc.roc2, x = "best", best.weights = c(.5, .6), transpose = TRUE,
                    ret = c("threshold", "accuracy", "precision", "recall"))
```

Apply the optimal threshold determined with the above code to the out-of-sample predictions made for the test set. Create a 2x2 table showing the number of false positives, false negatives, true positives and true negatives.

Answer 2(e)

```
# Insert answer here.
```

Question 2(f)

In the context of predicting voter turnout, is it more important for a campaign to avoid False Positives (predicting people will vote but they do not) or False Negatives (predicting someone will not vote but they do)?

Answer 2(f)

Insert answer here.

Question 3: Turnout Model with Penalized Logistic Regression

In the previous question, we used our knowledge of the political science literature, and past studies of voter turnout, to choose which variables should be included in our model. In some cases, we may not have the background or information to know which variables to include in a model. Nonetheless, we rarely want to include every conceivable variable as a predictor in a logistic regression.

Question 3(a)

What are the trade-offs between including more variables (and even interactions between variables) versus fewer variables when trying to make out-of-sample predictions?

Answer 3(a)

Insert answer here.

Question 3(b)

One way to help with variable selection, so that we can consider but not necessarily include a wide range of covariates, is to implement a penalized regression, known as a LASSO model. The lasso model forces the sum of the absolute value of the coefficients in the model to be below a fixed value (called `lambda`), which leads to setting some of the coefficients to zero. The effect is to omit some variables from the model when making predictions, leading to a sparser model.

But in order to do this, we need to be able to pick a good value for `lambda`. We use a method called cross-validation. We split the data into, say, 10 different pieces and then try running our model with different values of lambda on them. We can then evaluate what value for `labmda` allows us to minimize the errors in our predictions.

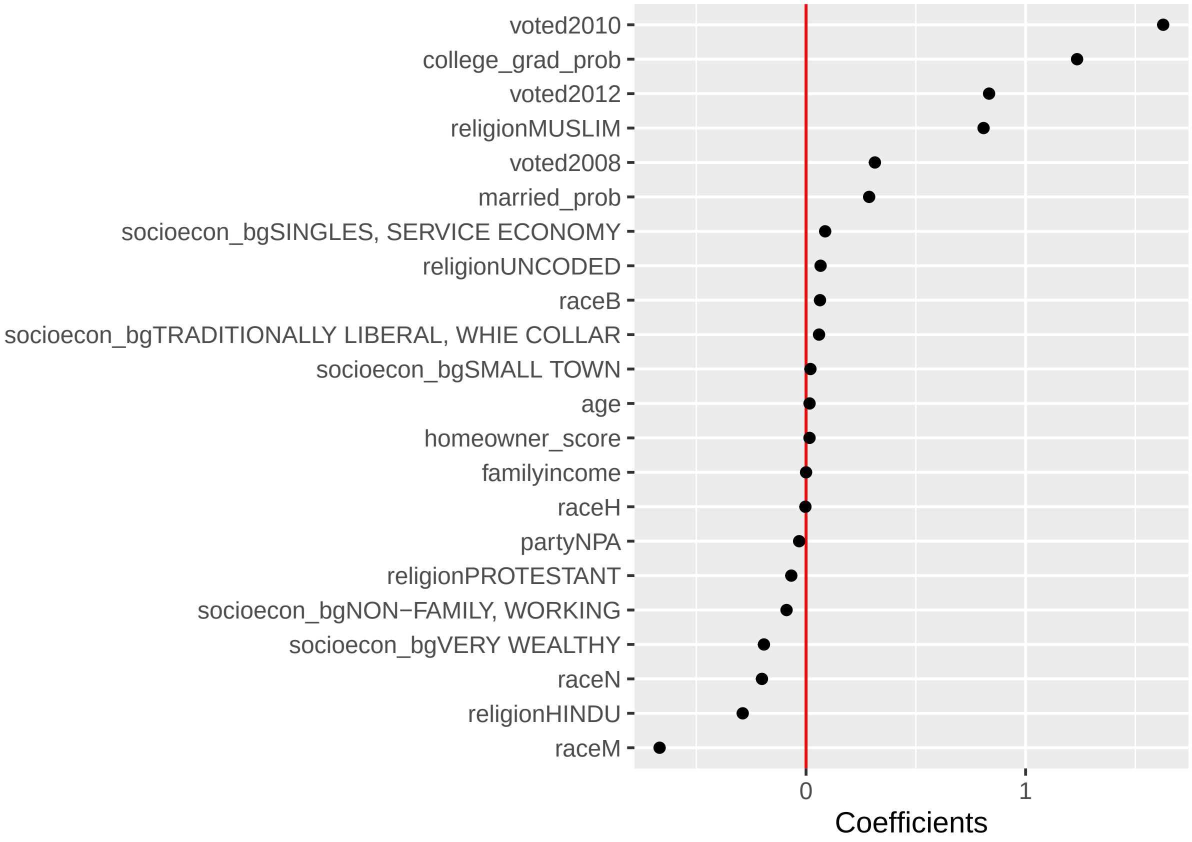
The `glmnet` package in R helps us to fit a lasso model, and it even performs cross-validation for us. The code below walks through an example for fitting a penalized logistic regression to our NC voter data.

Setup

```r
# Input names of all the variables in our data that we would like to use
turnout_model_variables <- c(
  "voted2014","voted2008","voted2010", "voted2012",           # turnout history & outcome
  "age","gender", "race","socioecon_bg", "familyincome",      # demographics 1
  "density", "ever_donor","college_grad_prob",                # demographics 2
  "hunter_prob","married_prob","religion", "homeowner_score", # demographics 3
  "party", "ideology_score","dem_score"                       # political variables
)

# clean NC dataset
nc_clean <- na.omit(nc[, c("id", turnout_model_variables)])

# Identify the rows in the NC data set in the training set and in the test set
row_train <- which(nc_clean$id %in% train_id)
row_test  <- which(nc_clean$id %in% test_id)

# To input the covariates into the glmnet function, we must convert
#  the data into a matrix form
# The function model.matrix() does this for us.
X <- model.matrix(voted2014 ~., data = nc_clean)[,-c(1,2)]

# create training and test matrices/vectors
## training set
X_train <- X[row_train,]
y_train <- nc_clean$voted2014[row_train]

## testing set
X_test <- X[row_test,]
y_test <- nc_clean$voted2014[row_test]
```

Estimate the penalized regression with `cv.glmnet()` and `glmnet()`

```r
# Run the cross validation portion of the model
# This step allows us to pick the value for
#  lambda that will minimize mean cross-validated error
set.seed(02139)
cv_lasso <- cv.glmnet(X_train, y_train, alpha = 1, family = "binomial")

# Refit the model using the best value for lambda
model2 <- glmnet(X_train, y_train, alpha = 1, family = "binomial",
                 lambda = cv_lasso$lambda.min)


# see which variables were actually included in model
est_coefs <- as.matrix(coef(model2))
# knitr::kable(est_coefs[est_coefs[,1] != 0, ], digits = 3,
#              col.names = c("Coefficients"))
```



Briefly interpret the estimated coefficients. What kind of pattern do you observe in the estimate?

Answer 3(b)

Insert answer here.

Question 3(c)

Does this approach yield improvements in out-of-sample prediction, compared to the approach taken in Question Two? We investigate the performance of this penalized logistic regression by computing the ROC curve and classification accuracy.

Setup

Let's compute the predicted turnout based on the result from the penalized logistic regression.

```
# Predict in sample and find threshold
yhat_lasso_train <- predict(model2, newx = X_train, type = "response")
nc.roc4 <- roc(response  = y_train,
               predictor = as.numeric(yhat_lasso_train))
thres_lasso <- coords(nc.roc4, x = "best", transpose = TRUE,
                      ret = c("threshold","accuracy","precision","recall"))
```

```
# Predict out-of-sample observations
yhat_lasso_test <- predict(model2, newx = X_test, type = "response")
```

Compute the ROC curve based on `yhat_lasso_test`, and compute the classification accuracy based on the threshold computed in `thres_lasso`.

Answer 3(c)

```
#Insert Answer Here
```

Question 3(d) Optional Challenge

Building on the example of penalized logistic regression above, create a model (estimated in the training set) predicting whether someone is registered with the Democratic Party (see the `party` variable for this information). Use the variables listed below in your model.

```
# outcome variable
nc$dem <- as.numeric(nc$party=="DEM")

# all variables used in this question
party_model_variables <- c(
  "dem","voted2008","voted2010", "voted2012",                  # turnout history & outcome
  "age","gender", "race","socioecon_bg", "familyincome",      # demographics 1
  "density", "ever_donor","college_grad_prob",                 # demographics 2
  "hunter_prob","married_prob","religion", "homeowner_score", # demographics 3
  "ideology_score"                                             # political variables
)
```

Answer 3(d)

```
#Insert answer here
```

Question 3(e) Optional Challenge

In the training data set, determine a threshold for classifying someone as a Democrat using an ROC curve and the `coords` function. Explain why you chose this threshold.

```
#Insert answer here
```

Question 3(f) Optional Challenge

Make the out-of-sample prediction of whether someone is a Democrat in the test data. What is the share of false positives, false negatives, true positives, and true negatives based on your prediction?

Answer 3(f)

```
#Insert answer here
```

Question 4: Optional Challenge Question

Suppose you are once again working for Richard Burr (R-NC), a senator from North Carolina. You have come to believe that your political future depends on successful messaging and fundraising efforts directed at voters who supported Donald Trump in 2016 and who have a past history of making donations. This is challenging since some of these voters may not be registered Republicans.

Question 4(a)

Use the `voted_pres16` variable in the CCES survey as your outcome, and create a model assessing the relationship between a vote for Trump and age, gender, party, race, family income, college graduate status, and marital status. Report estimated coefficients, and briefly interpret the estimates.

Answer 4(a)

```
#Insert answer here
```

Question 4(b)

Use the estimates from the above model to then predict the probability of having voted for Trump *in the NC Voter File Data* (i.e., object named `nc`). Use the data on past donations to create a vector of voters who you believe (1) voted for Trump in 2016 and (2) have donated money in the past. How many of these voters are there in the NC Voter file data? What is the gender breakdown of these voters?

Answer 4(b)

```
#Insert answer here
```