Extra Credit Assignment:
Cross-Validation for Logistic Regression
Due: Thursday, May 7

Instructions: This extra credit assignment is worth 3 points. The R code is provided, so you just need to run the code and interpret the output. Review the lecture 19 slides before working on this assignment.

For this assignment, you will again use the 2012/16 election data set for US counties discussed in HW 7. You will estimate a logistic regression model to predict whether or not Trump wins a county using some demographic variables from the US Census. Cross-validation will be used to evaluate the performance of logistic regression model.

```
# load data set
county_votes16 <- readRDS(url("https://ericwfox.github.io/data/county_votes16.rds"))
```

(a) Run the code below to randomly split the data into a 70% training and 30% test set. Then estimate a logistic regression model for `trump_win` on the training set, using the 8 demographic variables as predictors. Use `summary()` to print the results.

```
set.seed(999) # set seed for reproducibility
n <- nrow(county_votes16)
floor(0.7*n)
train <- sample(1:n, 2178)
glm_train <- glm(trump_win ~ pct_pop65 + pct_black + pct_white + pct_hispanic
                 + pct_asian + highschool + bachelors + income,
                 data = county_votes16, subset = train, family = binomial)
```

(b) Some of the predictors in the logistic regression model fit in part (a) are not significant. Run the code below to perform backwards stepwise variable selection. Use `summary()` to print the results.

```
glm_train2 <- step(glm_train)
```

(c) Run the code below to make a confusion matrix between the actual and predicted values on the test set. A 0.5 probability threshold is used to classify each point (county) in the test set as a Trump win or a Trump loss. Use the confusion matrix to calculate the accuracy (percent correctly classified), sensitivity (percent of Trump wins (1) correctly classified), and specificity (percent of Trump losses (0) correctly classified).

```
county_votes16_test <- county_votes16[-train, ]
probs_test <- predict(glm_train2, newdata = county_votes16_test, type = "response")
length(probs_test)
```

```
preds_test <- rep(0, 934)
preds_test[probs_test > 0.5] <- 1
tb <- table(prediction = preds_test,
            actual = county_votes16_test$trump_win)
addmargins(tb)
```

(d) Run the following code to plot the ROC curve and compute the AUC. How does the model perform on the test set according to these metrics?

```
library(pROC)
roc_obj <- roc(county_votes16_test$trump_win, probs_test)
plot(1 - roc_obj$specificities, roc_obj$sensitivities, type="l",
     xlab = "1 - Specificity", ylab = "Sensitivity")
abline(0, 1, lty=2)
auc(roc_obj)
```

(e) In terms of the cross-validation results how does the multiple logistic regression model, which uses the demographic variables as predictors, compare with the simple logistic model from lecture 19, which uses `obama_pctvotes` as a predictor? Note that the same seed, `set.seed(999)`, was used when making the random 70% training and 30% test set split, so the comparison is valid.