27.5: Cross-validation (Section 26.6.1)

Cross-validation is a powerful technique that allows us to estimate how well our results will generalize to a new dataset. Here we will build our own crossvalidation code to see how it works, continuing the logistic regression example from the previous section.

In cross-validation, we want to split the data into several subsets and then iteratively train the model while leaving out each subset (which we usually call *folds*) and then test the model on that held-out fold Let's write our own code to do this splitting; one relatively easy way to this is to create a vector that contains the fold numbers, and then randomly shuffle it to create the fold assigments for each data point.

```
nfolds <- 4 # number of folds

# we use the kronecker() function to repeat the folds
fold <-  kronecker(seq(nfolds),rep(1,npatients/nfolds))
# randomly shuffle using the sample() function
fold <- sample(fold)

# add variable to store CV predictions
disease_df <- disease_df %>%
  mutate(CVpred=NA)

# now loop through folds and separate training and test data
for (f in seq(nfolds)){
  # get training and test data
  train_df <- disease_df[fold!=f,]
  test_df <- disease_df[fold==f,]
  # fit model to training data
  glm_result_cv <- glm(heartattack ~ biomarker, data=train_df,
                 family=binomial())
  # get probability of heart attack on test data
  pred <- predict(glm_result_cv,newdata = test_df)
  # convert to prediction and put into data frame
0.5  disease_df$CVpred[fold==f] = (pred>)

}
```

Now let's look at the performance of the model:

```
# create table comparing predicted to actual outcomes
CrossTable(disease_df$CVpred,
           disease_df$heartattack,
           prop.t=FALSE,
           prop.r=FALSE,
           prop.chisq=FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Col Total |
## |-------------------------|
##
##
## Total Observations in Table:  1000
##
##
##                    | disease_df$heartattack
## disease_df$CVpred |     FALSE |      TRUE | Row Total |
## ------------------|-----------|-----------|-----------|
##             FALSE |       416 |       269 |       685 |
##                   |     0.832 |     0.538 |           |
## ------------------|-----------|-----------|-----------|
##              TRUE |        84 |       231 |       315 |
##                   |     0.168 |     0.462 |           |
## ------------------|-----------|-----------|-----------|
##      Column Total |       500 |       500 |      1000 |
##                   |     0.500 |     0.500 |           |
## ------------------|-----------|-----------|-----------|
##
##
```

Now we see that the model only accurately predicts less than half of the heart attacks that occurred when it is predicting to a new sample. This tells us that this is the level of prediction that we could expect if were to apply the model to a new sample of patients from the same population.
