LogisticInference_start

Dima Mikhaylov

12/1/2021

```
library(faraway)
data <- wcgs
n <- dim(data)[1]
names(data)
```

```
##  [1] "age"     "height"  "weight"  "sdp"     "dbp"     "chol"    "behave"
##  [8] "cigs"    "dibep"   "chd"     "typechd" "timechd" "arcus"
```

The variables of interest are: * chd: '1' indicating the person disease, and a '0' indicating the person did not develop disease. * age: age in years, *sdp: systolic blood pressure in mm Hg,* dbp: diastolic blood pressure in mm Hg, * cigs: number of cigarettes smoked per day, * dibep: behavior type, labeled A and B for aggressive and passive respectively.

```
#data$chd <- factor(data$chd)
#levels(data$chd) <-c(1, 0) # 1 for No and 0 for Yes
#levels(data$chd)
```

Recall that we split the data into a training set and a test set (50-50 split) using set.seed(6021). Be sure to do this split, fit the logistic regression with the training data.

```
set.seed(6021) ##for reproducibility to get the same split
sample<-sample.int(nrow(data), floor(.50*nrow(data)), replace = F)
train<-data[sample, ] ##training data frame
test<-data[-sample, ] ##test data frame
dim(train)[1]
```

```
## [1] 1577
```

From the previous guided question set, we went with a logistic regression model with age, sdp, cigs, and dibep as the predictors, dropping dbp from the model. We will now evaluate how our model performs in classifying the test data.

```
result <- glm(chd ~ age + sdp + cigs + dibep, family='binomial', data = train)
summary(result)
```

```
##
## Call:
## glm(formula = chd ~ age + sdp + cigs + dibep, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2095  -0.4515  -0.3488  -0.2748   2.6961
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.597370   1.025411  -8.384  < 2e-16 ***
## age          0.060880   0.016560   3.676 0.000237 ***
## sdp          0.020757   0.005595   3.710 0.000207 ***
## cigs         0.020642   0.006035   3.421 0.000625 ***
## dibepB       0.531792   0.198281   2.682 0.007318 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 893.04  on 1576  degrees of freedom
## Residual deviance: 838.25  on 1572  degrees of freedom
## AIC: 848.25
##
## Number of Fisher Scoring iterations: 5
```

1. Based on the estimated coefficients of your logistic regression, briefly comment on the relationship between the predictors and the (log) odds of developing heart disease.

Positive slopes thus positive relationship meaning all the predictors increase log oadds of developing the disease.

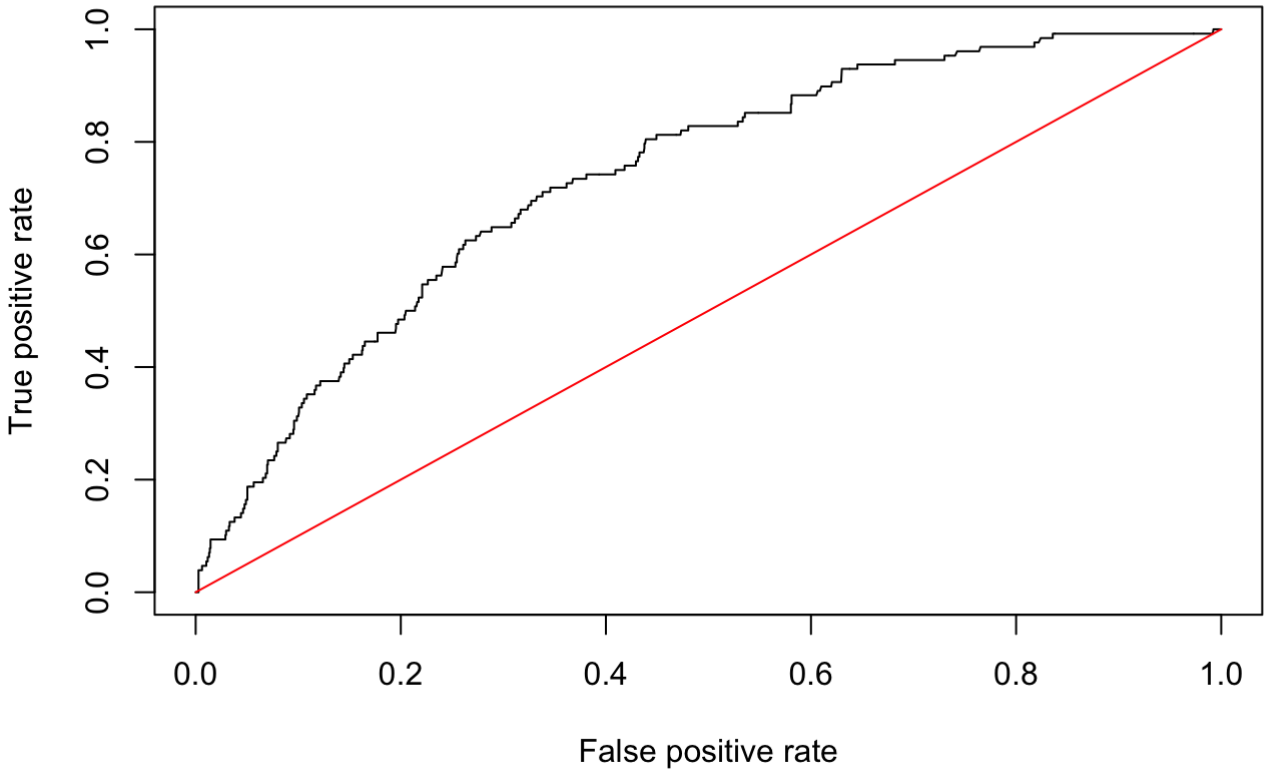
2. Validate your logistic regression model using an ROC curve. What does your ROC curve tell you?

```r
#install.packages("ROCR")
library(ROCR)

preds <- predict(result, newdata=test, type='response')
rates <- prediction(preds, test$chd)
roc_result <- performance(rates, measure = 'tpr', x.measure = 'fpr')
roc_result
```

```
## A performance instance
##   'False positive rate' vs. 'True positive rate' (alpha: 'Cutoff')
##   with 1220 data points
```

```r
plot(roc_result, title="The ROC curve")+
lines(x = c(0,1), y = c(0, 1), col="red")
```




```
## integer(0)
```

3. Find the AUC associated with your ROC curve. What does your AUC tell you?

```
auc <- performance(rates, measure="auc")
auc@y.values
```

```
## [[1]]
## [1] 0.7371679
```

4. Create a confusion matrix using a cutoff of 0.5. Report the accuracy, true positive rate (TPR), and false positive rate (FPR) at this cutoff.

```
table(test$chd, preds>0.5)
```

```
##
##       FALSE
##   no   1449
##   yes   128
```

5. Based on the confusion matrix in part 4, a classmate says the logistic regression at this cutoff is as good as random guessing. Do you agree with your classmate's statement? Briefly explain.
6. Discuss if the threshold should be adjusted. Will it be better to raise or lower the threshold? Briefly explain.
7. Based on your answer in part 6, adjust the threshold accordingly, and create the cor- responding confusion matrix. Report the accuracy, TPR, and FPR for this threshold.

```
table(test$chd, preds>0.7)
```

```
##
##       FALSE
##   no   1449
##   yes   128
```

8. Comment on the results from the confusions matrices in parts 4 and 7. What do you think is happening?