M12Guided

Alanna Hazlett

2024-04-20

Data were collected from 3154 males aged 39 to 59 in the San Francisco area in 1960. They all did not have coronary heart disease at the beginning of the study.

Variables of interest: chd: whether the person developed coronary heart disease during annual follow ups in the study, with a '1' indicating the person developed coronary heart disease, and a '0' indicating the person did not develop coronary heart 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

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.

```
Data<-wcgs
#set.seed used for reproducability
set.seed(6021)
sample<-sample.int(nrow(Data), floor(.50*nrow(Data)), replace = F)
train<-Data[sample, ]
test<-Data[-sample, ]
#Creating logistic regression
result<-glm(chd~age+sdp+cigs+dibep,family=binomial,data=train)
```

Problem 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.

```
summary(result)
```

```
##
## Call:
## glm(formula = chd ~ age + sdp + cigs + dibep, family = binomial,
##     data = train)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.065578   1.036178  -7.784 7.03e-15 ***
```

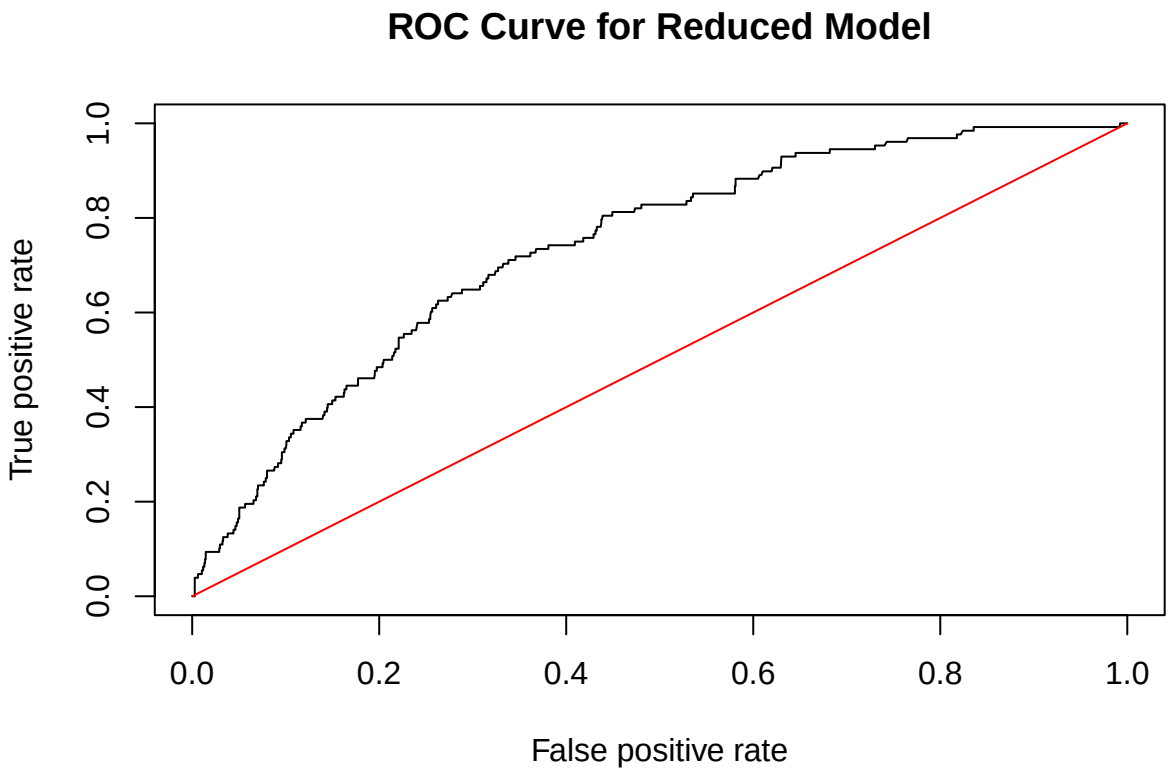
```
## age          0.060880   0.016560   3.676 0.000237 ***
## sdp          0.020757   0.005595   3.710 0.000207 ***
## cigs         0.020642   0.006035   3.421 0.000625 ***
## dibepB      -0.531792   0.198281  -2.682 0.007318 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 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
```

Log(odds) for person having chd increases with age, sdp, and cigs. The log(odds) decreases for a person with personality type B compared to person with personality type A.

Problem 2

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

```
##predicted probs for test data
preds<-predict(result,newdata=test, type="response")
##produce the numbers associated with classification table
rates<-ROCR::prediction(preds, test$chd)
##store the true positive and false postive rates
roc_result<-ROCR::performance(rates,measure="tpr", x.measure="fpr")
##plot ROC curve and overlay the diagonal line for random guessing
plot(roc_result, main="ROC Curve for Reduced Model")
lines(x = c(0,1), y = c(0,1), col="red")
```



Our model does a better job classifying the observation than random guessing.

Problem 3

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

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

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

AUC confirms that our model does better than random guessing, as it is larger than 0.5.

Problem 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
```

$$accuracy = \frac{TP + TN}{n} = \frac{0 + 1449}{1577} = 0.9188$$
$$TPR = \frac{TP}{FN + TP} = \frac{0}{128 + 0} = 0$$
$$FPR = \frac{FP}{TN + FP} = \frac{0}{1449 + 0} = 0$$

Problem 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.
Yes, looking at the confusion matrix we immediately can see that our model is only as good as random guessing, because the true column is missing. The values in the true columns are what make up the numerators for TPR and FPR, and with them both being zero that means that they are equal to eachother. When TPR = FPR the model is classifying based on random guessing.

Problem 6

Discuss if the threshold should be adjusted. Will it be better to raise or lower the threshold? Briefly explain. We would need to discuss with a subject matter expert to determine which error we should minimize: * Increasing FPR/Decreases FNR If we change the threshold it would be to make it lower. Lowering the threshold makes it easier for an observation to be classified as 1 by the model. When we lower the threshold TPR increases and FPR increases.

Problem 7

Based on your answer in part 6, adjust the threshold accordingly, and create the corresponding confusion matrix. Report the accuracy, TPR, and FPR for this threshold.

```
print(table(test$chd, preds>0.4))
```

```
##
##       FALSE TRUE
##   no   1445    4
##   yes   128    0
```

```
print(table(test$chd, preds>0.3))
```

```
##
##       FALSE TRUE
##   no   1440    9
##   yes   123    5
```

```
print(table(test$chd, preds>0.1))
```

```
##
##       FALSE TRUE
##   no   1114  335
##   yes    57   71
```

$$accuracy = \frac{TP + TN}{n} = \frac{5 + 1440}{1577} = 0.9163$$
$$TPR = \frac{TP}{FN + TP} = \frac{5}{123 + 5} = 0.0391$$
$$FPR = \frac{FP}{TN + FP} = \frac{9}{1440 + 9} = 0.0062$$

Problem 8

Comment on the results from the confusions matrices in parts 4 and 7. What do you think is happening? Adjusting our threshold results in the model classifying more observations as 1, meaning that more students will be classified as having driven drunk: * We will have an increase in the false positive cases, where students truly have not driven drunk and are classified as having driven drunk. * We will have an increase in the true positive cases, where students who have truly driven drunk are classified as having driven drunk.