Stat 6021: Guided Question Set 12

Tom Lever

11/28/22

For this question set, we will continue using the Western Collaborative Study Group (WCGS) data set, which is from a study regarding heart disease. Data were collected from 3,154 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. The data set comes from the `faraway` package and is called `wcgs`. The variables of interest are:

- `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 predictors, dropping `dbp` from the model. We will now evaluate how our model performs in classifying the test data.

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

```
library(faraway)
data_set <- wcgs
set.seed(6021)
number_of_observations <- nrow(data_set)
indices_of_observations <- sample.int(number_of_observations, floor(0.5 * number_of_observations), repla
training_data_set <- data_set[indices_of_observations, ]
testing_data_set <- data_set[-indices_of_observations, ]
head(training_data_set, n = 3)
```

```
##       age height weight sdp dbp chol behave cigs dibep chd typechd timechd
## 2275   53     71    142 150  78  218     A2   40     B  no    none    3127
## 21132  46     71    180 110  80  260     B3    0     A  no    none    2887
## 11520  39     71    180 114  78  234     B3    0     A  no    none    2985
##         arcus
## 2275  present
## 21132  absent
## 11520 present
```

```
generalized_linear_model <- glm(chd ~ age + sdp + dbp + cigs + dibep, family = "binomial", data = traini
generalized_linear_model
```

```
## 
## Call:  glm(formula = chd ~ age + sdp + dbp + cigs + dibep, family = "binomial",
##     data = training_data_set)
```

```
##
## Coefficients:
## (Intercept)          age          sdp          dbp         cigs       dibepB
##    -8.83568      0.06021      0.01512      0.01203      0.02137      0.52691
##
## Degrees of Freedom: 1576 Total (i.e. Null);  1571 Residual
## Null Deviance:       893
## Residual Deviance: 837.5     AIC: 849.5
```

The logistic regression equation is

$$ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -8.836 + 0.06\ age + 0.015\ sdp + 0.012\ dbp + 0.021\ cigs + 0.527\ I_1$$

where $I_1 = 1$ for behavior type B, and 0 for behavior type A.

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

The regression coefficient for `age` is 0.06.

For an additional year on average, the estimated log odds of developing coronary heart disease increases by 0.06, while controlling for the other predictors (age, systolic blood pressure, diastolic blood pressure, and behavior type).

For an additional year on average, the estimated odds of developing coronary heart disease gets multiplied by a factor of $exp(0.06)$, while controlling for the other predictors.

The regression coefficient for `sdp` is 0.015.

For an additional point in systolic blood pressure, the estimated log odds of developing coronary heart disease increases by 0.015, while controlling for the other predictors.

For an additional point in systolic blood pressure, the estimated odds of developing coronary heart disease gets multiplied by a factor of $exp(0.015)$, while controlling for other predictors.

The regression coefficient for `dbp` is 0.012.

For an additional point in diastolic blood pressure, the estimated log odds of developing coronary heart disease increases by 0.012, while controlling for the other predictors.

For an additional point in diastolic blood pressure, the estimated odds of developing coronary heart disease gets multiplied by a factor of $exp(0.012)$, while controlling for other predictors.

The regression coefficient for `cigs` is 0.021.

For an additional cigarette smoked per day on average, the estimated log odds of developing coronary heart disease increases by 0.021, while controlling for the other predictors.

For an additional cigarette smoked per day on average, the estimated odds of developing coronary heart disease gets multiplied by a factor of $exp(0.021) = 1.021$, while controlling for the other predictors.

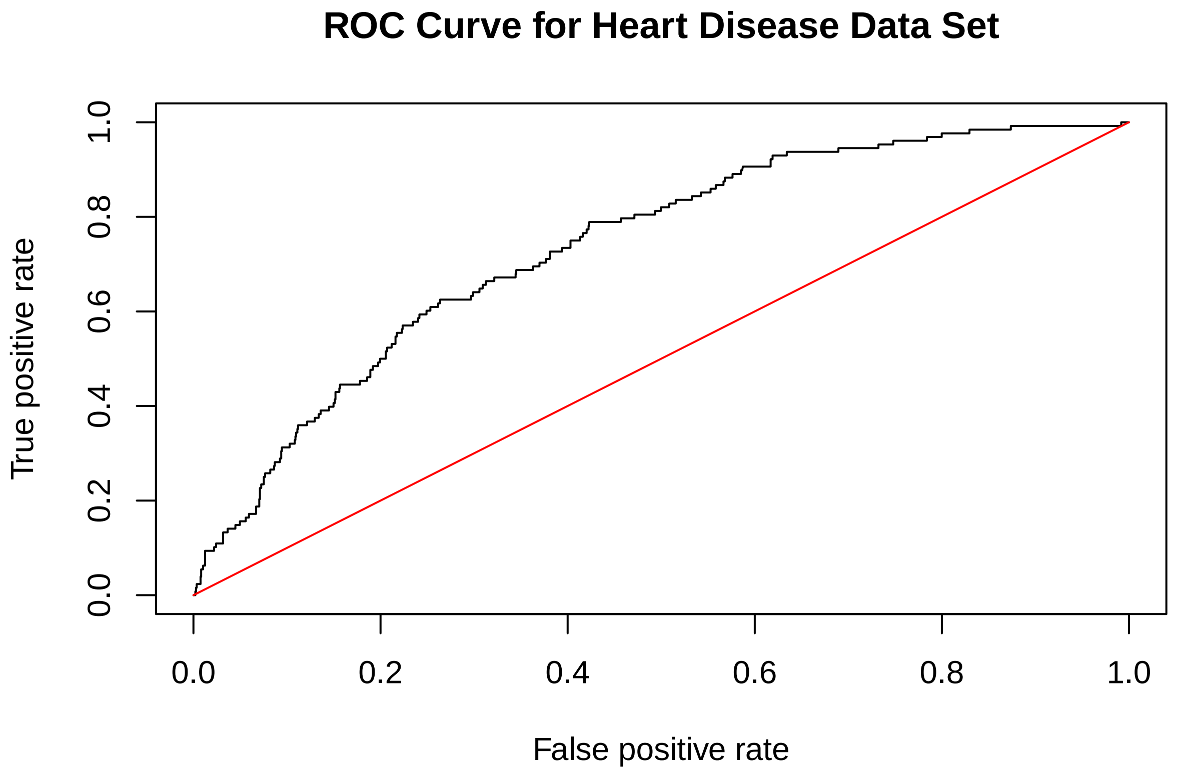
The regression coefficient for `dibep` is 0.527.

The estimated log odds of developing heart disease for males with type B (passive) behavior is 0.527 higher than for males with type A (aggressive) behavior, while controlling for the other predictors.

The estimated odds of developing heart disease for males with type B behavior is $exp(0.69) = 1.694$ times the odds for males with type A behaviors, while controlling for the other predictors.

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

```
library(ROCR)
predicted_indicators_of_coronary_heart_disease <- predict(generalized_linear_model, newdata = testing_da
prediction_instance <- prediction(predicted_indicators_of_coronary_heart_disease, testing_data_set$chd)
roc_curve <- performance(prediction_instance, measure = "tpr", x.measure = "fpr")
plot(roc_curve, main = "ROC Curve for Heart Disease Data Set")
lines(x = c(0, 1), y = c(0, 1), col = "red")
```



Our ROC curve tells us that our logistic regression model is a little less than halfway from a logistic regression model that predicts indicators of coronary heart disease randomly and without relying on any predictors (symbolized by the diagonal line) to a logistic regression model that predicts indicators of coronary heart disease perfectly [symbolized by the point (0, 1)].

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

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

```
## [[1]]
## [1] 0.7359844
```

The Area Under the Curve of the ROC curve tells us that our logistic regression model is a little less than halfway from a logistic regression model that predicts indicators of coronary heart disease randomly and without relying on any predictors (symbolized by the area under the diagonal line) to a logistic regression model that predicts indicators of coronary heart disease perfectly (symbolized by the entire plot area).

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(testing_data_set$chd, predicted_indicators_of_coronary_heart_disease > 0.5)
```

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

$$a = \frac{FP}{TP} = \frac{0}{0} = UNDEFINED$$
$$TPR = \frac{TP}{FN + TP} = \frac{0}{128 + 0} = \frac{0}{128}$$
$$FPR = \frac{0}{1449 + 0} = \frac{0}{1449}$$

5. Based on the confusion matrix in part 4, a classmate says the logistic regression model is as good as random guessing. Do you agree with your classmate's statement?

I agree. The true and false positive rates are zero for both our logistic regression model and a random model.

6. Discuss if the threshold should be adjusted. Will it be better to raise or lower the threshold? Briefly explain.

If we lower the threshold, we decrease the number of false negatives, increase the number of false positives, and produce a logistic regression model that is worse than a random model in terms of its ability to correctly predict the indicator of coronary heart disease for a middle-aged male.

If we raise the threshold, we decrease the number of false positives, increase the number of false negatives, and produce a logistic regression model that is better than a random model in terms of its ability to correctly predict the indicator of coronary heart disease for a middle-aged male.

We raise the threshold.

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

```
table(testing_data_set$chd, predicted_indicators_of_coronary_heart_disease > 0.75)
```

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

$$a = \frac{FP}{TP} = \frac{0}{0} = UNDEFINED$$
$$TPR = \frac{TP}{FN + TP} = \frac{0}{128 + 0} = \frac{0}{128}$$
$$FPR = \frac{0}{1449 + 0} = \frac{0}{1449}$$

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

The confusion matrices in parts 4 and 7 are identical. For thresholds of 0.5 and 0.75, our logistic regression model is consistently predicting that all middle-age males do not develop coronary heart disease. Our logistic regression model needs to be adjusted.