STAT 380 – Logistic Regression Part 2 (Lecture 13)

RECALL: At the end of Lecture 12, we used a Titanic passenger's age and a logistic regression model to produce a predicted probability of surviving the sinking of the ship. The predicted probabilities were compared to a threshold (chosen to be 0.38) in order to classify each passenger as someone that we believed would survive the sinking of the Titanic or not. The results for the first 5 passengers are summarized below:

Row Num	Survived	Age	Predicted Prob. of Surviving	Predicted Value of Surviving	Type
1	No	22	0.4007780	Yes	Incorrect; False Positive
2	Yes	38	0.3675797	No	Incorrect; False Negative
3	Yes	26	0.3923793	Yes	Correct; True Positive
4	Yes	35	0.3737199	No	Incorrect; False Negative
5	No	35	0.3737199	No	Correct; True Negative

IDEA: We can summarize the results through the use of a tool known as a confusion matrix.

EXAMPLE 1: Using L12_titanic3.csv, build a logistic regression model for predicting whether a person survived the sinking of the Titanic using the person's age. Using a threshold of 0.38, predict each person's survival status (pred_surv) and create a confusion matrix summarizing the results of the classifier.

a. Create the confusion matrix.

```
#Create the confusion matrix - Basic
table(pred_surv, titanic$Survived)
```

```
##
## pred_surv  No Yes
##       No  206 135
##       Yes 339 207
```

b. Based on our confusion matrix, find the following;

i. sensitivity?

ii. recall?

iii. specificity?

iv. true negative rate?

v. precision?

vi. accuracy?

NOTE: There's a multitude of technical terms related to probabilities associated with the 2x2 confusion matrix. See:

https://en.wikipedia.org/wiki/Sensitivity_and_specificity

If you are looking for an academic source:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2636062/

Definition: Sensitivity is the ability of the test to correctly classify an individual as 'having the condition.' Since this terminology originated in the field of biostatistics, we often think about the confusion matrix in terms of diagnosing a condition. The sensitivity is also called recall, hit rate, or the true positive rate.

$$Sensitivity = Recall = Hit\ Rate = \ True\ Positive\ Rate = P(Predict +\ |\ Cond +) = \frac{TP}{TP + FN}$$

Definition: Specificity is the ability of a test to correctly classify an individual as 'lacking the condition.' The specificity is also called the selectivity or true negative rate.

$$Specificity = Selectivity = \ True\ Negative\ Rate = P(Predict -|\ Cond -) = \frac{TN}{TN + FP}$$

NOTE: We are also interested in the probability: $1 - Specificity$, which is called the false positive rate.

Definition: Precision is the percentage of patients with a positive test who actually have the condition.

$$Precision = Positive\ Predictive\ value = P(Cond +|\ Pred +) = \frac{TP}{TP + FP}$$

Definition: The accuracy is the percentage of correct classifications.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

EXAMPLE 2: A useful function from the "caret" package is the "confusionMatrix" function. In addition to producing the confusion matrix, it also calculates some probabilities for you.

a. Explain the reason that the results shown below are incorrect.

```
#use confusionMatrix from caret library
confusionMatrix(data = as.factor(pred_surv),
                      reference = as.factor(titanic$Survived))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  206 135
##        Yes 339 207
##
##                Accuracy : 0.4656
##                  95% CI : (0.4324, 0.4991)
##     No Information Rate : 0.6144
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : -0.0151
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.3780
##             Specificity : 0.6053
##          Pos Pred Value : 0.6041
##          Neg Pred Value : 0.3791
##              Prevalence : 0.6144
##          Detection Rate : 0.2322
##    Detection Prevalence : 0.3844
##       Balanced Accuracy : 0.4916
##
##        'Positive' Class : No
##
```

b. Look at the help file and fix the code.

IDEA: Ideally, we want a high value for sensitivity (true positive rate) and a low value for 1-specificity (false positive rate).

Definition: A receiver operating characteristic (ROC) curve is a usual summary of a binary classifier. ROC curves plot the sensitivity (y-axis) vs. 1-specificity (x-axis) for a variety of threshold values.

NOTE: The area under the curve (AUC) is a commonly used metric for assessing the quality of the model. Classifiers that perform no better than random chance would have an AUC of 0.50, while the closer the AUC is to 1, the better the classifier has worked.

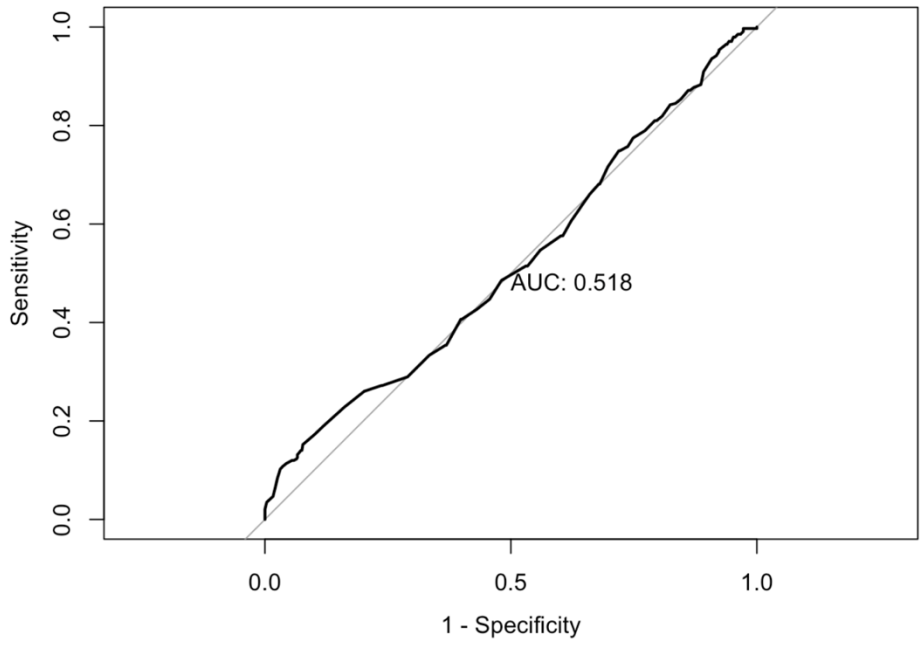
NOTE: A great tutorial for demonstrating a variety of ways to create ROC curves in R may be found at: https://rviews.rstudio.com/2019/03/01/some-r-packages-for-roc-curves/

EXAMPLE 3: For the logistic regression model created in Example 1, create the ROC curve and find the area under the curve. Comment on the effectiveness of our model for making predictions.

```
#Using roc from pROC library
test_roc = roc(response = titanic$Survived,
               predictor = pred_prob,
               plot = TRUE, print.auc = TRUE,
               legacy.axes=TRUE)
```

```
## Setting levels: control = No, case = Yes
```

```
## Setting direction: controls < cases
```




```
#View AUC
as.numeric(test_roc$auc)
```

```
## [1] 0.5179436
```

NOTE: The ROC curve and AUC are often used for the Test set in order to evaluate the performance of the classifier. (This metric has an advantage over using the accuracy since the accuracy requires a single threshold to create the confusion matrix. Unfortunately, the ROC curve and AUC are usually only used when the response is binary.)

EXAMPLE 4: Perform an 80/20 training/testing split of the Titanic data using a seed of 123. Let us see if we can improve upon the performance of our classifier by including some other variables.

a. Backward elimination is a variable selection procedure in which we begin by placing all variables in the model. Variables are removed one at a time according to some criterion until it is no longer beneficial to remove anything else. The code below implements backward elimination and the output shows the results. NOTE: We used the step function from the stats package. Since stats is part of base R, no library command is needed. Our decision-making is based on the Akaike Information Criterion (AIC), which is a common metric for comparing models. Smaller values are more desirable.

```
#Train/Test split
set.seed(123)
train_ind <- sample(1:nrow(Titanic), floor(0.8*nrow(Titanic)))
set.seed(NULL)

Train <- Titanic[train_ind, ]
Test <- Titanic[-train_ind, ]

#Build Intercept Only Model. NOTE: ~ 1 tells R that you only want an intercept
int_only_model <- glm(SurvivedNum ~ 1, family = binomial, data = Train)

#Build model with all potential regressors.
#In code below, SurvivedNum ~ . tells R to use all columns in dataset to predi
ct SurvivedNum
#SurvivedNum ~ . -Survived tells R to use all columns except Survived to predi
ct SurvivedNum
full_model <- glm(SurvivedNum ~ . -Survived, family = binomial, data = Train)

#Perform backward elimination
#Have R do it all
stats::step(object = full_model,
            scope = list(lower = int_only_model, upper = full_model),
            data = Train,
            direction = "backward")
```

```
Start:  AIC=631.16
SurvivedNum ~ (Survived + Pclass + Sex + Age + Siblings + Parch +
    Fare) - Survived

           Df Deviance    AIC
- Parch     1   617.21 629.21
<none>          617.16 631.16
- Fare      1   620.13 632.13
- Siblings  1   629.95 641.95
- Age       1   643.24 655.24
- Pclass    1   677.16 689.16
- Sex       1   792.43 804.43
```

```
Step:  AIC=629.21
SurvivedNum ~ Pclass + Sex + Age + Siblings + Fare

            Df Deviance    AIC
<none>           617.21 629.21
- Fare       1   620.16 630.16
- Siblings   1   631.62 641.62
- Age        1   643.36 653.36
- Pclass     1   679.40 689.40
- Sex        1   797.68 807.68
```

b. Build the model selected in Part a. for predicting the odds (or probability) of survival using PClass, age, sex, fare, and siblings using the train. Find the overall accuracy on the test set based on a threshold of 0.5. (0.80337)

c. Find the AUC associated with the test set.