Homework 6

PubH 7406: Biostatistical Inference II – Jared D. Huling

Due: Monday, May 1, 2023 at 2:30pm CT

HW 6 - 30 points

Instructions

Turn in the homework in the form of a PDF. It is fine to use existing functions to answer questions. However, note that simply providing output from statistical software is not sufficient and will not receive full points. Any output/results must be interpreted in the context of the problem and accompanying explanations of the results are necessary. Use clearly-defined and explained statistical notation to accompany results. For example, any statistical tests should be accompanied by a formal statement of the null and alternative hypotheses with additional context explaining what the hypotheses mean in the context of the problem. Please follow the instructions on homeworks in the syllabus in order to receive full credit.

If you have any questions, please ask in the course Q&A on Canvas so that others can see any responses.

Goals

We learned about model evaluation primarily in the context of regression models for continuous outcomes. The concepts we learned about for estimating out-of-sample prediction error (cross validation) can be used for other loss functions and other types of outcomes. Here we will explore using cross validation for models for binary outcomes. We will use the area under the ROC curve to evaluate the predictive performance of our model.

Data: Evaluating heart failure prediction models

Cardiovascular diseases kill approximately 17 million people globally every year, and they mainly exhibit as myocardial infarctions and heart failures. Heart failure (HF) occurs when the heart cannot pump enough blood to meet the needs of the body. A study focused on heart failure patients who were admitted to a cardiology center in Pakistan during April-December of 2015. All the patients were aged 40 years or above, having left ventricular systolic dysfunction, belonging to NYHA class III and IV. Available electronic medical records of patients quantify symptoms, body features, and clinical laboratory test values can potentially be used to predict mortality due to heart failure. Data in this study include age, ejection fraction, serum creatinine, serum sodium, anemia, platelets, creatinine phosphokinase, blood pressure, gender, diabetes and smoking status.

In this exercise we will develop a model predicting the risk of cardiovascular mortality and will aim to evaluate the generalizability and calibration. First we will develop the risk model, then use cross validation to assess predictive performance, and then validate on an external dataset.

	Variable	Description
1	age	Age
2	anaemia	Condition in which one lacks enough healthy red blood cells to carry adequate oxygen to body's tissues and anemia is frequent comorbidity of hear failure and is associated with the poor outcomes
3	creatinine_phosphokinase	An enzyme in the body, when total cpk level is high, it most often means there has been injury or stress to muscle tissue, the heart or the brain
4	diabetes	Metabolic disease that causes high blood sugar
5	ejection_fraction	Measurement of the percentage of blood leaving your heart each time it contracts
6	high_blood_pressure	Indicates if the blood pressure is high or not
7	platelets	Count of platelets in the blood
8	serum_creatinine	Measures the level of creatinine in the blood and provides an estimate of how well one's kidneys work
9	serum_sodium	serum sodium
10	sex	Male or Female
11	smoking	Yes or No
12	DEATH_EVENT	If the patient deceased during the follow up period. This is the outcome

The training data can be read in as follows:

```
fpath <- "https://jaredhuling.org/data/pubh7406/heart_failure_dataset.csv"
heartf <- read.csv(fpath)
dplyr::glimpse(heartf)
```

```
## Rows: 150
## Columns: 12
## $ age                       <int> 75, 80, 75, 45, 45, 65, 65, 68, 53, 75, 58,
## $ anaemia                   <int> 0, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0
## $ creatinine_phosphokinase <int> 582, 123, 81, 981, 582, 52, 128, 220, 63, 5
## $ diabetes                  <int> 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1
## $ ejection_fraction         <int> 20, 35, 38, 30, 14, 25, 30, 35, 60, 30, 38,
## $ high_blood_pressure       <int> 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0
## $ platelets                 <dbl> 265000, 388000, 368000, 136000, 166000, 276
## $ serum_creatinine          <dbl> 1.90, 9.40, 4.00, 1.10, 0.80, 1.30, 1.60, 0
## $ serum_sodium              <int> 130, 133, 131, 137, 127, 137, 136, 140, 135
## $ sex                       <int> 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0, 0, 1
## $ smoking                   <int> 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0
## $ DEATH_EVENT               <int> 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1
```

Develop a model for mortality risk

```
glmfit <- glm(DEATH_EVENT ~ <develop a model!>, family = binomial(),
              data = heartf)
```

Evaluate model on validation data

The *validation* data can be read in as follows. This dataset is only to be used for validation of your model in parts 4-5.

```
fpath_val <- "https://jaredhuling.org/data/pubh7406/heart_failure_dataset_new_sample.csv"
heartf_val <- read.csv(fpath_val)
```

Exercises for heart failure data

1. (5 points) For your model, evaluate the ROC curve and the area under the ROC curve on the in-sample data (the data you used to fit your model). Recall the `pROC` package has functionality for plotting ROC curves and computing the area under them.

   ```
   library(pROC)
   roc_in_sample <- pROC::roc(heartf$DEATH_EVENT,
         predict(glmfit, newdata = heartf, type = "response"))
   roc_in_sample$auc
   plot(roc_in_sample)
   ```

2. (10 points) Use K-fold cross validation to estimate the out-of-sample area under the ROC curve and compare this with the in-sample AUC. Why does LOOCV not work for AUC?

3. (5 points) For your model, investigate the impact of changing the value of K in K-fold cross validation? What are the tradeoffs for larger versus smaller values? How variable are your results if you use different random seeds?

4. (5 points) Evaluate the area under the ROC curve on the validation data. How does this compare with your estimate from K-fold cross validation?

5. (5 points) How well calibrated is your model? Use an appropriate method to assess model calibration. Use the validation dataset for calibration.