Analyzing the Statistical Effectiveness of Diagnostic Tests for COVID-19 Survival Using the Receiver Operating Characteristic (ROC) Curve

Mengle Hu, Simeng (Hannah) Li, and Ruiyi Liu, and Dr. Paul Kvam
Department of Mathematics, *University of Richmond, VA 23173*

Introduction

Diagnostic tests that use markers to determine whether a patient is diseased or healthy are standard tools in medical screening. For the diagnosis of many modern diseases, the difference in marker measurements used to screen healthy patients from diseased patients can be subtle, and statistical researchers work to develop the most effective tool to discern this difference.

Misclassification costs are often asymmetric; that is, the cost of misclassifying a healthy patient into the diseased group (a false-positive result) is often less than the cost of misclassifying a diseased patient into the healthy group. We will focus on a tool that has been especially useful in recent decades called the receiver operating characteristic (ROC) curve.

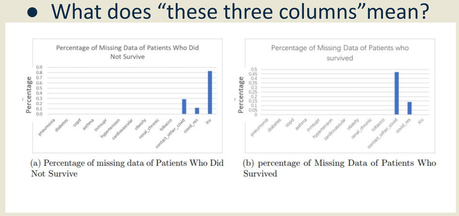
Dataset

Introduction of Data

- Released by the Mexican government
- Covid 19 survived outcomes over the past two years
- 500,000 patients dataset that contains columns of id, sex, patient type, entry date, date symptoms, date died, intubed, pneumonia, diabetes, copd, asthma, inmsupr, hypertension, cardiovascular, obesity, renal chronic, tobacco, contact other Covid, Covid res, and icu.

Cleaning of Data

- Many observations were plagued by missing values.
- Misleading if we delete all data with missing values.
- Study which variables were most frequently missing.
- What does “these three columns”mean?



(a) Percentage of missing data of Patients Who Did Not Survive

(b) percentage of Missing Data of Patients Who Survived

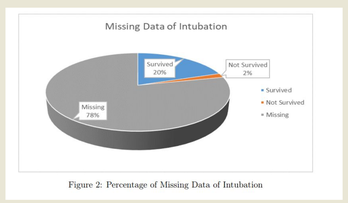


Figure 2: Percentage of Missing Data of Intubation

- Deleted these three columns, and then deleted the patients with missing values.
- Used survival status as response and not intubation status.

Logistic Regression

What is Logistic Regression

- A way to predict a binary response with numerical predictors.
- Describe the data and explain the relationship between one dependent binary variable and one or more independent variables.

How to use Logistic Regression

For binary response, with every observation, we recorded $(y_i, x_{1i}, x_{2i}, \cdots, x_{ki})$ but also the unrecorded $0 < p_i < 1$. We model the unobserved p:

$$p_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon.$$

Because the relationship between p_i and the explanatory variables is only rarely a linear one, we transformed p and assigned this logit transformation to ρ:

$$\rho = logit(\hat{p}_i) = \log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon.$$

Therefore, given any input values $x_1, \cdots, x_k$, the model will generate a prediction $logit(\hat{p})$, which we can map back to get p. We used **age** and **risk factors** as the predictor variables and **survival results** (survive or not) as the response variable.

```
Coefficients:
             Estimate Std. Error z value             Pr(>|z|)
(Intercept) -6.148415   0.022448    -274 <0.0000000000000002 ***
age          0.053665   0.000389     138 <0.0000000000000002 ***
nriskfactors 0.550312   0.004425     124 <0.0000000000000002 ***
```

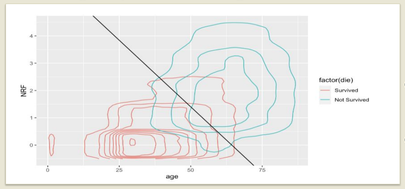
- The results of the logistic regression model.

Construct a Classifier through the Logistic Regression

We use above table draws a (straight) line across the plot for classification, indicating observations on one side of the line would be classified as 0 (survive) and the other side as 1 (not survive). The line would be solved as:

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -6.148415 + 0.053665 x_{ai} + 0.550312 x_{ri}$$

Because of the density of observations, the plot graph fails to effectively show how the two groups differ in age and the number of risk factors. Therefore, we considered a contour plot, a graphic technique that represents a 3-D surface in a 2-D format.



- admitted patients in the bottom left corner have higher survival rates
- admitted patients in the upper right corner tend not to survive.

Error Rate

Logistic regression has a high rate of misclassification (proportion of people who are predicted to be alive but not survived) when used as a classifier. In this case, we are trying to figure out the error rate of the model. If we use the threshold of $\hat{\rho}$ = 0.0633 to predict whether an admitted patient would be survive based on their age and number of risk factors, we can test our classifier as a method of model checking.

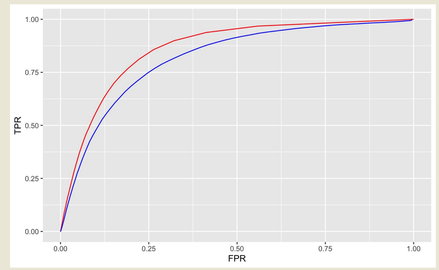
Receiver Operating Characteristic (ROC) Curve

What is the ROC Curve

A receiver operating characteristic curve, or a ROC curve, is a graphical plot that shows the ability of diagnosis of a binary classifier with a varied discrimination threshold. The ROC curve is made by plotting the true positive rate (TPR) against the false positive rate (FPR) at different threshold settings.

TPR: how many correct positive results occur among all positive samples available during the test.
FPR: how many incorrect positive results occur among all negative samples available during the test.

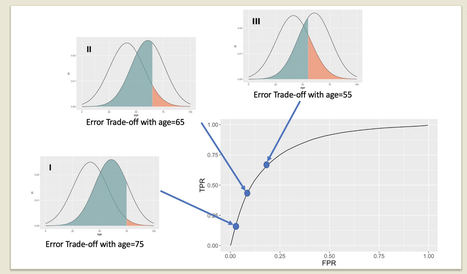
ROC Curve Using Different Threshold



- The red ROC curve uses predicted values from a logistic regression using age, number of risk factors and number of days before hospital as variables.
- The blue ROC curve uses age as discrimination threshold.
- We can clearly see that the AUC (area under curve) for red ROC is larger and larger AUC usually represents a better model.

ROC Curve Using Age as Threshold and the Error Trade-off

- Plots I,II, and III put two subgroups (Survived & Not Survived) together, using age 55, 65, and 75 as the thresholds, where the left curve indicates survived patients and the right curve refers to the patients who did not survive. We can see a trend that patients survived are generally younger as expected.



- The colored parts show the error trade-off between the two subgroups, where the orange part shows the patients who are predicted to not survive (but actually survived), and the green part shows the patients who are predicted to survive (but actually did not).