

P8130 Final Project: Predicting Breast Cancer Survival

Guadalupe Antonio Lopez, Gustavo Garcia-Franceschini, Derek Lamb

Abstract

Introduction

Background info/research on breast cancer patients:

Breast cancer occurs due to abnormal cell growths in breast tissue. Although it is most often found in females, 1 out of every 100 diagnosed patients in the US is a male. Other breast cancer risk factors include, increase in age, family history or personal history of breast cancer, radiation exposure, obesity, alcohol use, among many more. Research suggests that postmenopausal hormone therapy is a risk factor due the combination of estrogen and progesterone used to treat signs and symptoms of menopause.

Additionally, the patient's breast cancer stage is important to consider when determining the severity of the cancer and how to treat it. The American Joint Committee on Cancer (AJCC) TNM system is the most common, and contains clinical and pathologic systems. The pathologic stage is determined by examining the tissue removed during surgery, while the clinical stage is based on results of a physical exam, biopsy, and imaging tests. Nevertheless, both systems are composed of the size of the tumor, the spread to nearby lymph nodes and/or to distant sites, their estrogen and/or progesterone receptor status, the grade of the cancer, and if the cancer makes too much of HER2 protein.

Although, most recently, breast cancer survival rates have increase and number of deaths decreased, it would be important to explore the risk factors of breast cancer. For this project, we will investigate the odds of breast cancer survival given most of the risk factors previously mentioned.

Methods

Data source We obtained a deidentified set containing data on 4024 breast cancer patients. this dataset contains both demographic information, such as patient age, race, and marital status; clinical information

such as tumor stage, tumor size, hormone therapies (progesterone and estrogen), regional node positive, and regional node examined; and outcome information: the number of months the patient had survived prior to study conclusion, and their alive/dead status at the end of the study.

Data cleaning We combined the regional node positive and regional node examined variables into a “regional node proportion positive” variable. This variable, but neither the node positive nor node examined variables were in the model. Further, we decided to discard the T stage and N stage variables, as they captured information already contained in the AJCC 6th stage variable. We also excluded the grade variable, as it captured the same clinical information as the differentiate variable. Due to the skewness in the distribution of the tumor size, we applied a square root transformation to that variable (**Supplemental Figure 1**). We also added a `main_stage` variable, which groups all the stages in `6th_stage` so that “IIIA” and “IIIC” are under factor level “III” for example, and added a `white` variable that tells us if the patient is “White” or not, essentially grouping “Black” and “Other” together. Both variables were created in case the reduction of factor levels makes the fit better.

Model construction We decided to use logistic regression model to estimate the risk of patient death within the followup window. Formally, we assumed that an for an individual, with probability p to die after receiving a breast cancer diagnosis, the log-odds of p was linear, i.e.

$$\text{logit}(p_i) = \mathbf{X}\beta + \epsilon_i$$

Where \mathbf{X} is the $n \times p$ design matrix, and β is a vector in \mathbb{R}^p .

In addition to the covariates we included the interaction between `estrogen_status` and `progesterone_status` given that we found in our background research that having both “positive” increase the chances of breast cancer.

Model selection We considered all We used a criterion-based method, utilizing Akaike Information Criterion (AIC) to assess the performance of our models.

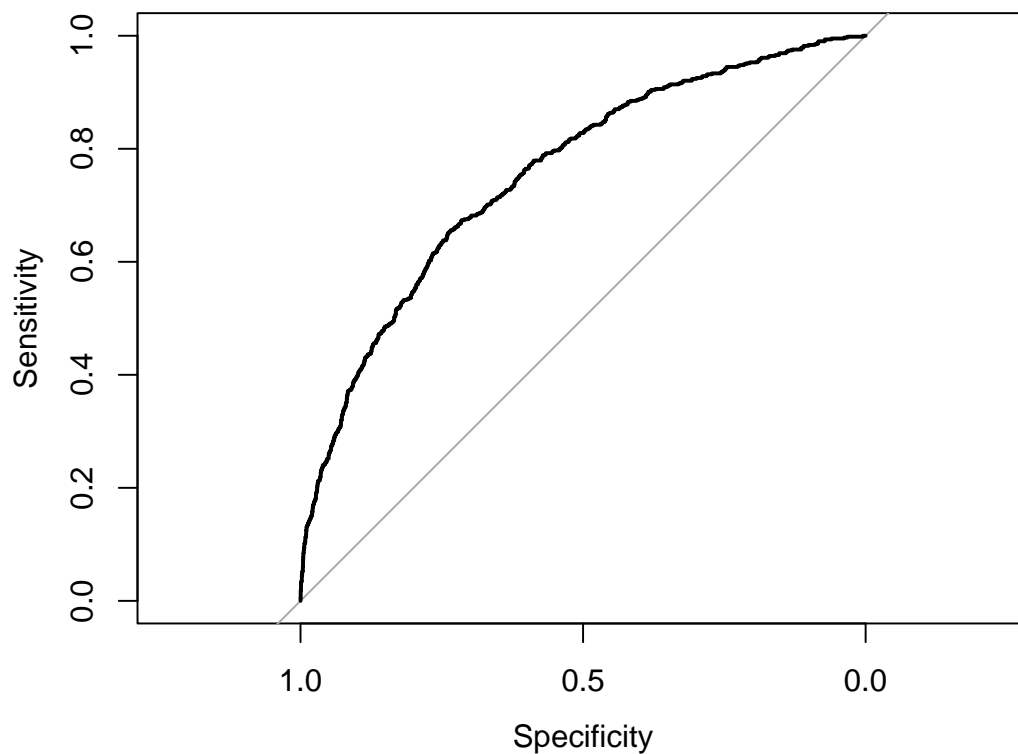
Model validation We performed 10 cross-validation to assess the performance of our model. Each observation in a dataset will be in 1 of 10 folds such that it gets used as training data 9 times, and as test data once. The predictions are then saved such that each row has an out-of-sample prediction that can be compared to the real value.

Results

Model construction and selection We used a logistic model coupled with criterion-based stepwise regression to determine which variables were useful in predicting the risk of death in breast cancer patients. The variables that were identified as important were age, race, marital status, AJCC 6th stage, differentiate, estrogen status, progesterone status, tumor size, and regional node positive proportion. The variables that were not identified as important by the model were whether the tumor was Stage A and the interaction between estrogen and progesterone status. For a list of model coefficients see **Supplemental Table 1**.

Diagnostics ROC, Brier score

ROC for full model on in-sample data:



Area under the curve: 0.7542

Table 1. This table shows the confusion matrix made from our 10 fold cross-validation, such that the predicted values are out-of-sample predictors. We can see that the model is very good at correctly predicting that somebody survived, and rarely predicts the individual died when it didn't. However, the model makes

a lot of errors when trying to predict if the individual survived: more than 90% of the total error happens when the model predicts the individual died, and it actually survived. This reflects the fact that the model prefers to predict 0 as a result of the `status` variable being 0 inflated.

	Predicted Died	Predicted Survived
Actually Died	3354	54
Actually Survived	532	84

«««< HEAD Separation plots =====

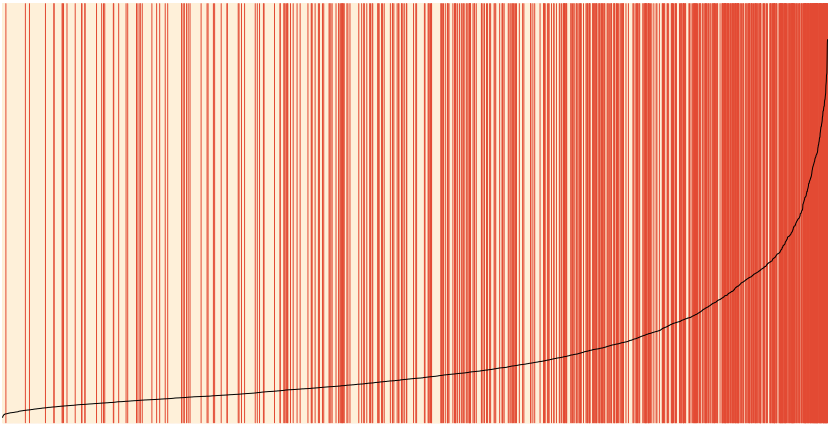


Figure X. Separation plot of model. Values are stripes, arranged in increasing predicted probability of death. The stripes are colored yellow if the patient survived, and red if they died. The black line indicates the predicted probability of death.

This separation plot shows that the model is able to reasonably effectively distinguish the patients that survived from those that died, though the low value of the predicted probability across the sample shows that it gives low probabilities of dying to all subjects. »»»> 19eea427594f950975b935ec060bbaf09fdf5638

Model performance by race Make a table with ROC AUC and Brier scores by race. Discuss

race	roc_auc	brier_score
White	0.760	0.109
Black	0.704	0.171
Other	0.650	0.088

Discussion

We constructed a model to predict the survival of breast cancer patients, and optimized it using a criterion based-selection approach.

Our model was effective at classifying patients who survived, but performed less well at classifying patients who died when using a $\hat{p} = 0.5$ as the cutoff. The separation plot we constructed suggests that a lower cutoff, such as 0.35 may be more appropriate, and further tuning may improve model performance.

Author Contributions

References

Software The aforementioned analyses were carried out using R 4.3.1 and RStudio Version 2023.06.2+561.

Appendices

a. tumor size transformation

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

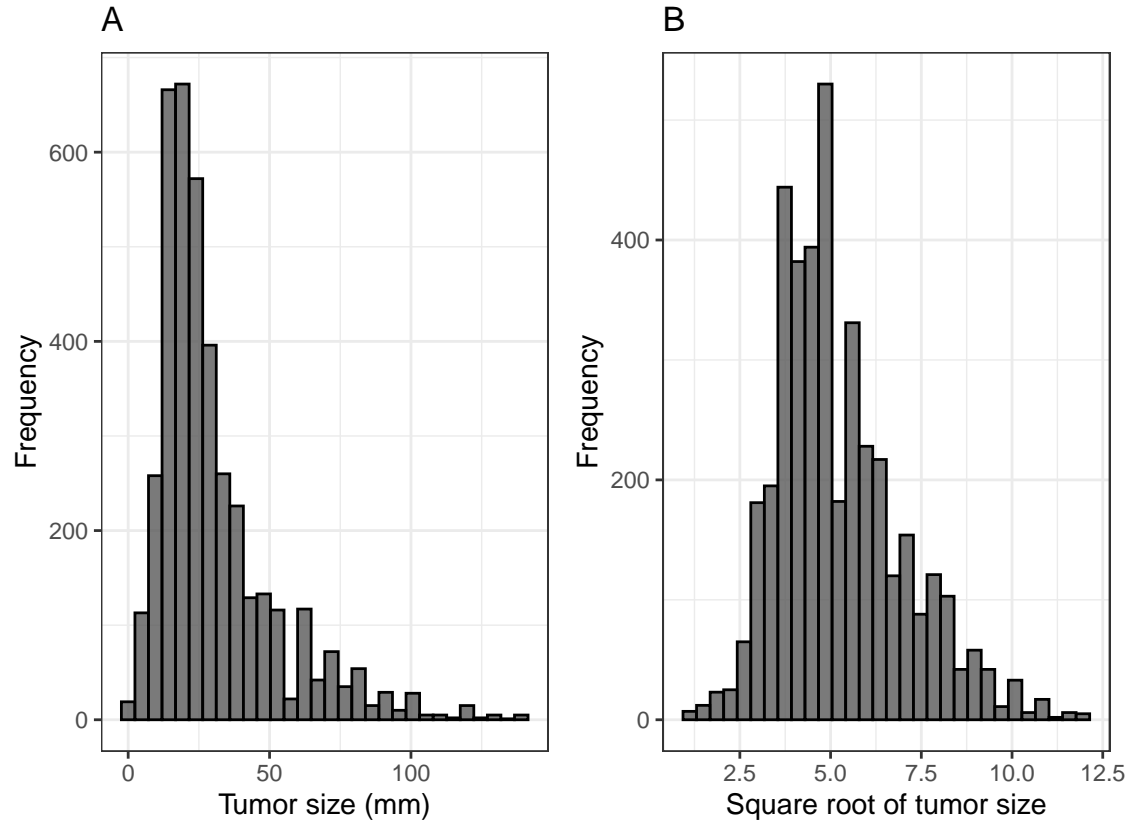


Figure S1. Transformation of tumor size variable. (A) Before transformation. (B) After square root transformation.

b. model coefficients

Table S1. Model Coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	-4.197	0.369	-11.382	0.000
age	0.023	0.006	4.180	0.000

term	estimate	std.error	statistic	p.value
raceBlack	0.506	0.162	3.129	0.002
raceOther	-0.431	0.202	-2.132	0.033
marital_statusDivorced	0.222	0.141	1.575	0.115
marital_statusSingle	0.153	0.134	1.146	0.252
marital_statusWidowed	0.225	0.192	1.171	0.242
marital_statusSeparated	0.847	0.366	2.316	0.021
x6th_stageIIIA	0.563	0.163	3.460	0.001
x6th_stageIIIC	1.064	0.193	5.505	0.000
x6th_stageIIB	0.413	0.154	2.676	0.007
x6th_stageIIIB	1.141	0.322	3.546	0.000
differentiateModerately differentiated	-0.389	0.104	-3.726	0.000
differentiateWell differentiated	-0.919	0.192	-4.776	0.000
differentiateUndifferentiated	0.961	0.529	1.816	0.069
estrogen_statusNegative	0.738	0.177	4.169	0.000
progesterone_statusNegative	0.571	0.127	4.485	0.000
root_tumor_size	0.048	0.031	1.523	0.128
regional_prop	1.237	0.185	6.686	0.000

b. ROC by race +/- sep plots?