A Probabilistic Model for the Haberman Cancer Dataset

Introduction

The Haberman Dataset (Haberman, 1976) contains patient cases from a study conducted in the late-1950s and 1960s at the University of Chicago Billings Hospital, observing the long-term survival of patients who had undergone surgery for breast cancer. The data comprises details on 306 patients between the ages of 30 and 83, whose surgeries were performed between 1958 and 1969 (inclusive). For each patient, cancer spread at the time of the surgery and whether the patient survived for 5 years or longer post-surgery are also reported. Though the analysis presented here is certainly not applicable for patients who underwent surgeries for cancers other than breast cancer in the 1950s and 1960s, nor for patients who are undergoing surgery for breast cancer today, understanding the dataset and the influence of patient age and cancer spread on survival rates is an important contribution to the medical field.

Data & Methods

In this work, we aim to develop a probabilistic model which predicts patient survival probability (for longer than 5 years after surgery), based on age, year in which the surgery was performed, and cancer spread – measured by the number of detected positive axillary nodes. Based on preliminary analysis of the data, shown in Figure 1, age and cancer spread are the two most important factors impacting survival probability for breast cancer patients in this sample.

We find that patients who live longer than 5 years after the surgery have a mean age of 52 ± 11 years, while those who died within 5 years of the surgery tend to be older, with a mean age of 53 ± 10 years at the time of the surgery. While the difference in age between the two groups is not statistically significant, applying a two-sample Kolmogorov-Smirnov (KS) test, we find that the distributions of the number of positive axillary nodes detected for each subset are not drawn from the same parent distribution (i.e., the two distributions are statistically different, $p < 0.001$). Patients who died within 5 years have a higher mean number of positive axillary nodes detected as compared to those patients who survived for longer than 5 years after their surgeries – 7.5 nodes versus 2.8 nodes, respectively. Further, the means and distributions of the year of the surgery for both groups are identical, so year in which patients underwent surgery is likely not a significant factor in determining survival probability.

Based on this cursory overview of the dataset, it is likely that patients who are younger with fewer positive axillary nodes at the time of surgery have a better chance of being part of the long-term survival group. However, the goal of this work is to provide a method for predicting the probability of a patient's survival. We use a logistic regression model, trained on the entire dataset, to predict the probability that a patient of a given age, with a certain number of positive axillary nodes, undergoing surgery in a given year will be a part of the long-term survival group. Details of the development and evaluation of this model can be found in the code repository.[1]

Age	Year	$N_{\text{pos. nodes}}$	P_{survival}	Status
66	58	0	81.01%	[1]
52	69	3	71.22%	[2]
36	60	1	86.10%	[1]
38	60	0	89.83%	[1]
53	59	3	68.34%	[2]

Table 1: Results of model test for five patients. Column 1, 2, and 3 are the age, year of surgery, and number of positive axillary nodes detected, respectively, available in the dataset. Column 4 gives the predicted probability of that patient being part of the long-term (5 years or longer) survival group, and Column 5 indicates that patient's actual survival status, where [1] indicates that the patient was part of the subset of patients that survived 5 years or longer and [2] indicates that the patient was part of the group that died within 5 years.

Results

To test the predictive model, we randomly select five patients and apply the model to predict the probability of long-term survival, and compare to the known the outcome (whether they survived for 5 years or longer after the surgery, or not). The results of this test are summarized in Table 1. In cases where we know that the patient selected did live for 5 years or longer after the surgery (Status [1]), the model predicts a probability of survival $\gtrsim 80\%$, where the predicted probability of survival for patients who did not live for 5 years after the surgery (Status [2]) was $\lesssim 70\%$.

To fully understand the probability of survival for the entire parameter space, including for "new" patients who may not be included in the dataset, we apply the model to a grid of patients with ages from 30 to 83 years old (intervals of 1 year), with 0 to 52 positive axillary nodes (intervals of 1 node), and surgeries performed between the years of 1958 and 1969 (inclusive, intervals of 1 year). Note, the parameter space is defined by the limits of each parameter within the dataset; we do not attempt to extrapolate our model to e.g., patients younger or older than the patients in the sample or to patients undergoing surgeries today (> 50 years after these data were collected), as the model may not accurately predict long-term survival probability beyond the parameters space in which

[1] `https://github.com/hmlewis-astro/haberman_ml`

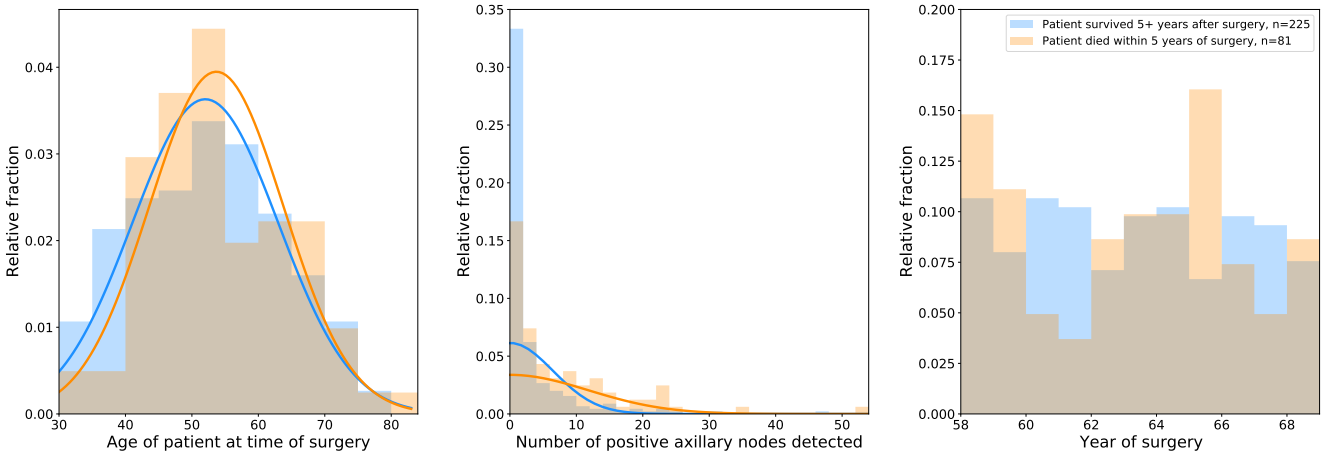

Figure 1: Left: distribution of the ages of patients at the time of the breast cancer surgery. Center: distribution of the number of positive axillary nodes detected prior to the breast cancer surgery. Right: distribution of the year in which the surgery was performed. The distributions shown in orange represent those patients who died within 5 years of surgery, those shown in blue represent those patients who survived 5 years or longer after the breast cancer surgery.

it was trained. The resulting probabilities are shown in Figure 2.

From this analysis, we find that age and number of positive axillary nodes are the most important factors in determining survival probability. Patients over the age of $\sim$45 with more than $\sim$10 positive nodes have a $\lesssim$50% probability of long-term survival, whereas patients who are younger or have fewer positive nodes tend to have a $>$60% change of long-term survival after surgery. Year of surgery also has some impact on survival probability – though not as significant an impact as age or number of nodes – with surgeries performed later (i.e., in 1968-69) having a slightly higher probability of success (for a patients of the same age and same number of positive nodes) than surgeries performed earlier (i.e., in 1958-59).

Conclusion

While these data are not current and the sample size is relatively small, such that we cannot expect the model to extrapolate to patients undergoing surgeries today, a similar model could be trained on a larger, more recent dataset to provide additional information to people considering undergoing surgery for breast cancer currently. Understanding the probability of long-term survival following a major surgery based on age and cancer spread is certainly an important consideration for breast cancer patients in weighing treatment (or non-treatment) options.

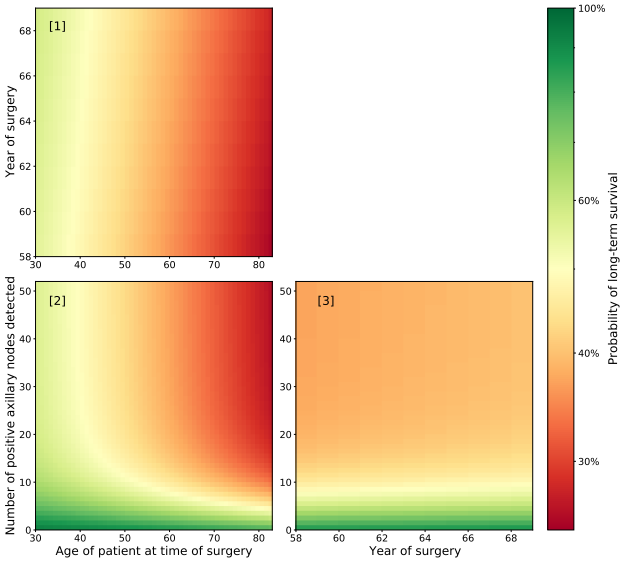

Figure 2: Probabilistic model applied to the full parameter space in age, year of surgery, and number of positive axillary nodes. Each plot shows two parameters (e.g., year vs. age) and is colored as a function of the median probability over the third (e.g., nodes). For example, plot [1] shows the year of the surgery versus the age of the patient, and is colored by the median probability of long-term survival for patients with 0 to 52 positive axillary nodes detected, where redder colors indicate a lower median survival probability and greener colors indicate a higher median probability.

References

S. J. Haberman. Generalized residuals for log-linear models. *Proceedings of the 9th International Biometrics Conference, Boston*, pages 104–122, 1976.