Machine Learning to Inform Breast Cancer Post-Recovery Surveillance

Final Project Report
CS 229 Autumn 2017
Category: Life Sciences

Maxwell Allman (mallman)
Lin Fan (linfan)
Jamie Kang (kangjh)

1 Introduction

Women who recover from breast cancer are often given surveillance care for many years after their recovery, with the goal of early detection of relapses. Currently in the United States, there are no clear guidelines for physicians regarding the type or duration of this surveillance care, and patients with similar characteristics may receive different levels of surveillance. Furthermore, it is not known whether intense surveillance care is cost effective for reducing the medical costs (and patient suffering) in the event of a relapse. [4] [7] *The goal of our project is to use patient characteristics, and in particular the extent of surveillance, to predict relapse medical costs. Specifically, we wish to estimate the magnitudes of the predictive relationship (for instance, the sign and size of regression coefficients), from which cost-effectiveness can then be assessed.*

The input to our algorithms are patient medical characteristics during the treatment stage after initial breast cancer diagnosis such as age, cancer stage, surgeries, and biomarker testing results which are known to have strong clinical importance, as well as non-medical characteristics such as race, marital status and socioeconomic status quintiles. We also input patient medical characteristics post sucessful treatment and recovery from breast cancer, and in particular, the type and level of surveillance care received such as medical imaging and outpatient hospital visit cost per day. We then output predictions of cost of care in the event of a relapse to breast cancer. We explored four methods of prediction: (1) linear regression with forward step feature selection, (2) LASSO regression, (3) ridge regression, and (4) principle component analysis (PCA) and subsequent principle component regression (PCR) on the principle components deemed most significant. Our use of these more classical "parametric" methods is motivated by the desire of our project sponsors, Stanford Medicine oncologists, Dr. Douglas Blayney and Tina Seto, to have interpretable predictive models to inform policy making regarding recommended types and levels of surveillance care for patients who recover from breast cancer.

2 Related Work

Our search of the literature confirmed the opinion of Tina Seto, that this problem has not been well studied. In [7], the authors estimated ten-year survival and cost from breast cancer relapse, but these estimates were only conditional on the type of relapse, not on pre-relapse patient features and surveillance. In [4], the author indirectly assessed the value of surveillance by recording the proportion of a small cohort of patients who had an asymptomatic recurrence that was detected by regular surveillance. This study concluded that, "long term routine hospital follow up after treatment for breast cancer appears to be inefficient in detecting recurrence". We were unable to find work that directly predicts the cost of relapse from patient features and level of surveillance care, so our problem appears to be novel.

3 Dataset and Features

We obtained the data for this work from the Stanford Prevention Research Center's Oncoshare data resource through Tina Seto. This dataset includes medical information of over 11,000 breast cancer patients diagnosed at Stanford Hospital from 2000 to 2014, and over 6,000 patients who had an initial recovery from breast cancer, and subsequently were given varying levels of surveillance care. Among this set of patients, 362 patients relapsed. Because our goal is to predict the cost of care in the event of a relapse, we could only use the data from these 362 patients who relapsed. We randomly split this set of 362 examples via a 20/80% split into a test set of 72 (which we did not touch during model building) and a training/validation set of 290.

Initially, our dataset had over 50 features (both medical and non-medical) for each patient. After discussing with our oncologist data providers, we decided to use the following patient features. Medical patient features during the treatment phase include age, cancer stage, surgeries (lumpectomies and mastectomies), and biomarker testing results for estrogen receptor, progesterone receptor, and Her2 status, Charlson score (measure of overall patient health), and grade (measure of cancer severity in addition to stage). Non-medical features include race, marital status, and socioeconomic status quintiles. In the post recovery surveillance stage we used the two features: imaging and outpatient hopsital visit cost per day. So altogether we used 12 patient features in building predictive models of relapse cost of care.

Like most medical datasets, our dataset has many missing values. We used a common strategy to impute and account for missing values when appropriate. [1] If the missing value corresponded to a continuous feature, we imputed the missing value with the mean of the non-missing values for that feature. We included a dummy variable for each continuous feature to track whether or not each value is imputed. If the missing value corresponded to a categorical feature with multiple categories, an additional category was added to keep track of missing values. For the base set of 12 patient features, we thus ended up with 27 columns in our data matrix.

In order to capture non-linearities in the data, we included transformed features raised to the second power as well as first-order interaction terms. This resulted in a total of 624 columns in our data matrix, well above our number of training examples (290 in the training/validation set). Therefore, feature selection, dimensionality reduction and/or regularization are vitally important for model building. We did not pursue higher order transformations as the number of columns in our data matrix would quickly become unmanageable. As we know from function approximation theory (e.g. in the field of optimization), second-order transformations well approximate non-linearities in general, and continuing to higher order transformations result in diminishing benefits.

4 Methods

4.1 Rationale for Chosen Methods

As discussed in the Dataset and Features section, with polynomial mapping up to second order, we ended up with 624 columns in our data matrix, well above our number of training examples (290 in the training/validation set). Hence, feature selection, dimensionality reduction and/or regularization are a necessity for model identifiability and sensibility. As discussed in the introduction, our oncologist data providers desire an interpretable regression model, and hence, we decided to try the four methods: linear regression with forward step feature selection, LASSO regression, ridge regression and PCA/PCR, which allow us to build interpretable regression models with the necessary feature selection, dimensionality reduction and/or regularization properties.

4.2 Linear Regression

The first method we implemented to predict relapse cost was linear regression. In this model, for patient i the relapse cost is $y_i \in \mathbb{R}$, patient features $x_i \in \mathbb{R}^p$, and regression coefficients $\hat{\beta} = (\hat{\beta}_0, ..., \hat{\beta}_p)$, our prediction is

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}$$

where β is chosen so as to minimize the mean squared error (MSE), which for m patients is

$$J(\hat{\beta}) = \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2$$

Letting X be the design matrix such that x_{ij} is the jth feature of the ith patient, (where the first feature of every patient always has value 1), this model has the closed form solution

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

which we could efficiently compute for the size of X in this problem. Since the total number of features we used was large relative to the number of training examples we had, to avoid over fitting the data we used forward search and leave-one-out cross-validation (CV) error on the training set to select a subset of important features. Starting with a set containing only the first feature (the constant feature), forward search tests each of the remaining features, to find which one gives the smallest CV error when added to the set. The CV error is determined by computing how well the current set of features predicts the relapse cost of patient i, when the model is trained on every patient in the train set except patient i, for every choice of i. This forward step process will then give a sequence of increasing sets of features, and the one which gives the smallest CV error is chosen. Then, the model is trained on the full training set, using these selected features.

4.3 LASSO Regression

LASSO (least absolute shrinkage and selection operator) regression [3] is a technique for regularizing linear regression and simultaneously performing variable selection, i.e., setting coefficients corresponding to seemingly insignificant variables to zero. Consider the data $\{y_i, x_i\}_{i=1}^{m}$, where we wish to regress the $y_i \in \mathbb{R}$ on the features $x_i \in \mathbb{R}^p$. The LASSO optimization problem is as follows.

$$\hat{\beta}_{\text{lasso}} = \underset{\beta \in \mathbb{R}^{p+1}}{\text{argmin}} \left[\sum_{i=1}^{m} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + 2\lambda \sum_{j=1}^{p} |\beta_j| \right] \tag{1}$$

The first term corresponds to an ordinary least squares objective function, but the second term introduces an ℓ^1 penalty scaled by $\lambda > 0$ on all of the components of coefficient vector $\beta \in \mathbb{R}^{p+1}$ except for the first component which is the bias term β_0. The penalty parameter λ may be determined by CV, and once set, the objective is to find the β vector minimizing the two-part convex objective function. Typically, one efficiently computes an entire LASSO path for a sequence of λ values using the method of cyclical coordinate descent with warm starts. The method performs regularized regression as it penalizes large coefficient values which acts to prevent overfitting. The variable selection property is due to the geometry of the ℓ^1 norm such that what would otherwise be small coefficient values (ideally, insignificant ones) are set exactly to zero.

4.4 Ridge Regression

Ridge regression [3] is similar to LASSO regression in that it is also a regularization technique. The objective function is similar to that of LASSO:

$$\hat{\beta}_{\text{ridge}} = \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \left[\sum_{i=1}^{m} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right],$$

except an ℓ^2 penalty is imposed on the non-bias components of the coefficient vector β. This convex optimization problem has a closed-form solution:

$$\hat{\beta}_{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T Y,$$

where X, Y and I are the matrix of features, vector of response variables and the identity matrix, respectively. From this closed-form solution, we see that ridge regression is simply "bumping" up the diagonal elements of what would otherwise be the matrix $X^T X$ in ordinary least-squares regression (i.e., linear regression) with the penalty parameter $\lambda > 0$. This has the effect of resolving rank-deficiency of $X^T X$ in cases where there are more columns of X than there are rows, i.e., more coefficients to fit than training examples. Like in the case of LASSO regression, here selection of the λ parameter is also typically done by CV. Typically, ridge regression offers better predictive power than LASSO regression, but does not have the variable selection property.

4.5 Principle Component Regression

PCA selects orthogonal dimensions that maximize the variance within data by iteratively solving the following optimization problem:

$$\underset{U_i}{\operatorname{argmax}} (Z^T U_i)^2$$

$$\text{s.t. } U_i U_j = 0, \quad U_i U_i = I.$$

where Z represents the original features and U directions of newly constructed PCs. We used MATLAB function called *pca* to obtain PCs [6]. Using these PCs, PCR fits a linear regression models as shown in [5]. This is an extremely convenient way to resolve any rank deficiency issue present in the data, which is quite common in empirical datasets like our Oncoshare data. Similarly, PCA also helps when there is any multicollinearity since the algorithm reorganizes the data using orthogonal dimensions. Most importantly, it is also an efficient dimensionality reduction method. In our case, we had a total of 624 features initially. Although forward search method within linear regression can also reduce the number of features, PCA is another practical method, given that we can also control for R^2 (coefficient of determination).

5 Results and Discussion

5.1 Metrics

Our primary metrics for evaluating the following four methods are the square root of MSE (RMSE) on the training set (290 examples) and on the hold-out test set (72 examples). Note that RMSE has the unit of dollars of relapse cost of care and is a more meaningful characterization of prediction error than MSE. We are also interested in the specific features selected by these methods since our primary goal is to evaluate cost-effectiveness of pre-relapse surveillance on post-relapse cost of care through building a predictive model.

5.2 Linear Regression

For the regression, we used the basic 12 patient features, as well as the monomial terms of order 2 (we created a new feature $x_i x_j$ for every feature i and j). This resulted in a design matrix with 625 columns in total. The result of the forward search, with 5-fold CV error as the termination criterion, was a set of 33 features. The resulting model had an RMSE on the training set of 2.93×10^4, and 4.52×10^4 on the test set. This clearly indicates the model is grossly overfitting, and that regularization is needed. In fact, using only the average relapse cost of the patients in the training set as the prediction, gave an RMSE on the test set of only 4.29×10^4, so the results of this model are unlikely to be meaningful.

Because of the poor performance of this method, we tried this same method on just the 12 basic patient features, without the polynomial feature map. The resulting prediction had a training set RMSE of 3.36×10^4, and a test set RMSE of 4.44×10^4. Again, the error on the test set was higher than the error from predicting just the average relapse cost of care, so the results of this model are unlikely to be meaningful.

5.3 LASSO Regression

We first performed LASSO regression with the full set of 624 columns of the section of the data matrix corresponding to the training/validation set, varying the weight of the ℓ^1 penalty λ and performing 5-fold CV with randomized partitioning of the data to calculate an average MSE at each value of λ. The goal is to find the value of λ with minimum average MSE, and using this optimal value of λ, compute the coefficient estimates as in (1). Standardization of the features, which is crucial for regularized methods, was performed after data splitting in the 5-fold CV to avoid "peeking" into the validation set. We used the MATLAB implementation of *glmnet* [2] to calculate the average 5-fold CV error for a range of λ values, and results are shown in Figure 1. To obtain the lowest MSE, LASSO regression ends up setting all coefficients to zero, except for the constant bias term. This was alarming for us, and perhaps indicative that we had started out with too many features.

Thus, we then performed LASSO regression with the 27 columns of the data matrix corresponding to the base set of 12 features, i.e., no transformations or interaction terms were included. Running the same procedure to find the value of λ corresponding to the minimum average 5-fold CV MSE, we obtained a value of $\lambda = 2690$ (see Figure 2). With this optimal penalty, *LASSO found imaging cost per day, ER biomarker and Her2 biomarker to be the only features predictive of relapse cost of care. However, the part of the imaging cost per day feature that turned out to be significant according to LASSO was the dummy variable indicating when data values are missing. The coefficients corresponding to the actual values of imaging and outpatient hospital visit costs per day were not deemed significant by LASSO.*

The RMSEs of this method on the training set and test set are reported in Table 1 below. We see that the test set RMSE of the LASSO model built starting with the full set of transformed features is the lowest at 4.29×10^4, but recall that this LASSO model ends up simply predicting using the average relapse cost of care. We also see that starting with the 12 base features, LASSO builds a model that is essentially as predictive as when starting with the full set of transformed features. There is indeed some overfitting due to the test errors being larger than the training errors. Our overall conclusion is that more data is needed, perhaps data of higher quality, to build a more meaningful predictive model.

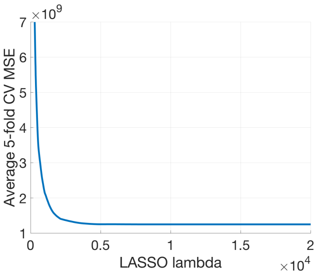

Figure 1: LASSO: full set of 624 columns

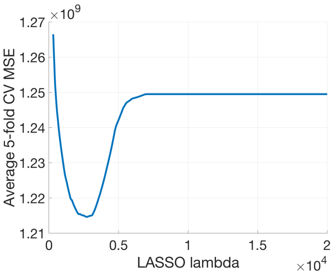

Figure 2: LASSO: base set of 27 columns

5.4 Ridge Regression

We performed the exact same analysis as with LASSO regression, except here we implemented ridge regression ourselves. We show plots of average 5-fold CV MSE with the ridge penalty λ varying in Figures 3 and 4. The values of λ are again chosen to minimize the average 5-fold CV MSE. The errors of this method on the training set and test set are reported in Table 1 below. The ridge and LASSO results are very similar and indicate that additional features besides a constant bias term do not help prediction, i.e., predicting using the average of the relapse cost of care results in the lowest RMSE.

5.5 Principle Component Regression

We also modeled our data using PCR. As mentioned earlier, PCR fits a linear regression model using principal components (PCs) of the inputs as new features. We first had to select the number of PCs we would regress on. This was done by comparing the R^2 values from PCR models using different number of PCs (1 to 624 PCs). R^2 is a conventional measure used as a model selection criterion as it captures the portion of variability in the prediction that the PCs can explain.

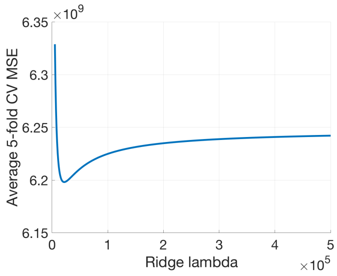

Figure 3: Ridge: full set of 624 columns

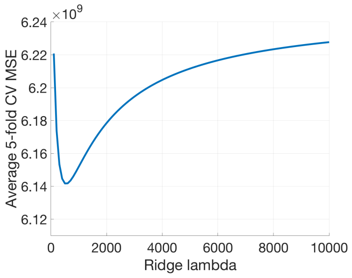

Figure 4: Ridge: base set of 27 columns

As in Figure 5, we noticed that the R^2 can be increased to values over 80%, but this would definitely overfit the model. Hence, we choose R^2 around 50%, which is achieved with 141 PCs, to mitigate the overfitting issue. Again, we fit the PCR model using the training set, which substantially lowered the training error to 2.49×10^4 for a model including second-order polynomial terms. However, the PCR model suffered from high test error of 5.20×10^4, implying that further reducing the feature dimension could not fully resolve the overfitting issue.

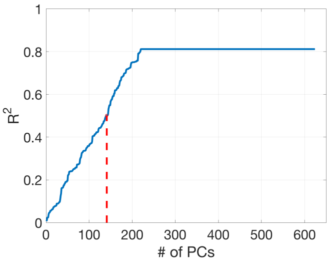

Figure 5: Coefficient of determination R^2 of PCR

5.6 Comparison of Prediction Errors of Different Methods

Table 1 below compares the errors of the four different models. We see that LASSO and ridge regression have the best test set predictive ability using both the complete set of transformed features and the minimal set of 12 base features. Linear regression and PCA/PCR seem to overfit more than LASSO and ridge regression due to the larger discrepancies between training and test errors. However, the performance on the test set of every method is upper bounded by the performance of prediction using the average relapse cost of care in the training/validation set (4.29×10^4 RMSE), which means that our models do not have much predictive ability.

Table 1: Comparison of RMSE (units in dollars) for four methods on training and test sets.

	Up to 2nd-Order Transformations		12 Base Features	
	Training Error	Test Error	Training Error	Test Error
Linear Regression	2.93e4	4.52e4	3.36e4	4.44e4
LASSO	3.52e4	4.29e4	3.43e4	4.30e4
Ridge	3.35e4	4.30e4	3.74e4	4.34e4
PCA/PCR	2.49e4	5.20e4	3.16e4	4.58e4

6 Conclusion and Future Work

Because each of the four methods we implemented performed more poorly on the test set than just using the average relapse cost of care in the training/validation set as the prediction, our models do not appear to be meaningful. This may be due to the large amount of noise that is inherent in relapse costs, and we do not have enough data to find a signal among the noise. In addition to the difficulty posed by the small number of patients in our data set, missing values were rampant in the patient features. With a larger and perhaps more complete data set, our methods and code could be used to possibly find a more predictive model. As follow-up to this work, we might also investigate other methods of data imputation as well as build models using other patient characteristics and/or surveillance procedures. Moreover, many patient characteristics are categorical with a natural ordering to the categories. Using more sophisticated methods that take into account the ordinal information might result in better predictive power.

Acknowledgments

We thank Dr. Douglas Blayney and Tina Seto of Stanford Medicine for providing the data for us and for several helpful discussions during the course of this project. We would also like to thank Professor Dan Boneh and Professor Andrew Ng as well as of course the dedicated TAs for making CS 229 an inspiring course.

References

[1] Cismondi, F., et al. *Missing data in medical databases: Impute, delete or classify?*, Artificial Intelligence in Medicine 58.1 (2013): 63-72.

[2] Friedman, J., Hastie, T., Tibshirani, R., *Regularization Paths for Generalized Linear Models via Coordinate Descent*, Journal of Statistical Software, 33.1 (2010): 1-22.

[3] Hastie, T., Tibshirani, R., Friedman, J., *The Elements of Statistical Learning*, Second Edition, Springer (2009).

[4] Hiramanek, N. *Breast cancer recurrence: follow up after treatment for primary breast cancer*, Postgraduate Medical Journal 80.941 (2004): 172-176.

[5] Jolliffe, I.T. *A note on the use of principal components in regression*, Applied Statistics (1982): 300-303.

[6] MathWorks, Statistics and Machine Learning Toolbox: User's Guide (R2017a). Retrieved December 1, 2017 from https://www.mathworks.com/help/stats/pca.html (2017).

[7] Stokes, M.E., et al. *Ten-Year Survival and Cost Following Breast Cancer Recurrence: Estimates from SEER-Medicare Data*, Value in Health 11.2 (2008): 213-220.

Contributions

- Data exploration and troubleshooting: Maxwell Allman, Lin Fan, Jamie Kang
- Data cleaning and processing: Maxwell Allman, Lin Fan
- Linear regression: Maxwell Allman, Jamie Kang
- LASSO and ridge regression: Lin Fan
- PCA/PCR: Jamie Kang