Results

By: Lily Thomson and Alysha Akhtar

Using the METABRIC dataset, which contains clinical attributes and gene Z-scores of patients with breast cancer, we wanted to create a model that could predict a patient's overall survival in months. Our prediction question was: Can we predict a breast cancer patient's overall survival time in months based on clinical attributes, gene Z-scores, and receptor statuses using partial least squares regression (PLSR)? This could be used by healthcare professionals to identify critical predictors or estimate survival times to ultimately guide treatment decisions. We decided to use PLSR to analyze the data since we want to model a response variable (overall survival in months) given a large number of predictor variables that are highly correlated. The predictor variables are a mix of 498 numerical features like age_at_diagnosis, tumor_size, mutation count, gene expression data, and dummy variables like er_status, pr_status, and her2_status. Since the receptor statuses are categorical variables, they were converted into binary columns for our analysis. The data was split into an 80/20 training/testing set.

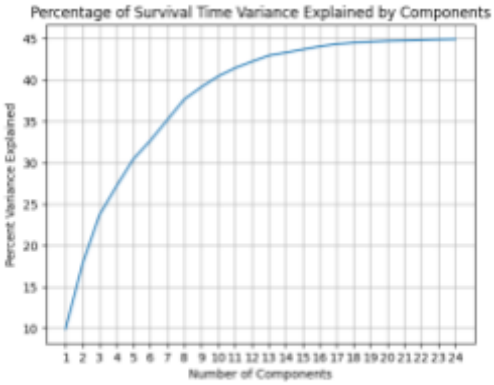

Figure 1. Percentage of response variance explained by the number of components in the PLSR

We chose to use 13 components for our PLSR model. The PLSR model constructs these new predictor variables, or components, as linear combinations of the original predictor variables (1). Cross validation was used to evaluate the percentage of variance in survival time that was explained when a given number of components was used. As shown by Figure 1, as more components were added, the percentage of variance explained increases. We selected 13 components since this is where the graph started leveling off more steadily with 42.9% variance explained, and adding more components would not significantly improve the model's ability to explain the target variable. PLSR was conducted with 13 components using the training data to construct a model.

Index	VIP Score
age_at_diagnosis	2.54055
tumor_size	2.310454
pdpk1	1.753507
sf3b1	1.645189
hsd3b7	1.555044
acvr1b	1.532486
ep300	1.532028
zfp36l1	1.529072
gsk3b	1.52183

Figure 2. Highest VIP scores of features of PLSR model.

We then wanted to determine the Variable Importance in Projection (VIP) scores of the model. These scores represent the importance of each feature or predictor variable in the components created by the model. Now only using the 13 components and the 211 features calculated to have a VIP score above 1, the PLSR model was retrained. Figure 2 shows features with the highest VIP scores, with age_of_diagnosis seen to be the most important feature in the components and the pdpk1 gene seen to be the most important gene feature.

Using the retrained model, the predicted versus observed survival time in months was plotted on the test data as shown in Figure 3. The model is very inaccurate in predicting the test data. The R^2 is only 0.01, so the model only explains 1% of the variance in the target variable while the rest of the variance is unexplained by the model. As shown by the graph, there is no clear relationship between observed survival time and survival time as predicted by the model. The predicted versus observed survival time was then plotted on the training dataset as shown in Figure 4. The model performs better on the data that it is trained on as indicated by the R^2 of 0.32 which is greater. However, the performance is still not great and predictive power is limited.

Figure 3. Scatter plot of predicted vs observed survival time on testing dataset. R^2: 0.0104. MSE: 0.971.

Figure 4. Scatter plot of predicted vs observed survival time on training dataset. R^2: 0.325. MSE: 0.678.

The model does not do well on predicting survival time on unseen data. This is likely due to the complexity of breast cancer and the factors that contribute to it. The relationship between features and survival time may be too complicated for the PLSR model to capture well. Since the model performed poorly on training and testing data, this indicates underfitting which means the model is too simplistic to capture the underlying patterns in the response data. Thus, using the predictor variables chosen, we were not effectively able to predict a breast cancer patient's overall survival time based on data for clinical attributes, gene expression, and receptor statuses.

References

1. "Partial Least Squares Regression and Principal Components Regression." *MathWorks*, www.mathworks.com/help/stats/partial-least-squares-regression-and-principal-components-regression.html. Accessed 9 Dec. 2024.