Pre-Analysis Plan for DS 3001 Final Project

By: Lily Thomson and Alysha Akhtar

Our dataset is from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database, a Canada-UK project looking at a cohort of 1905 women for targeted sequencing data. The dataset contains clinical attributes of individuals with breast cancer and gene expression data. An observation in our study is a given patient with breast cancer in our data set. Attributes include clinical variables like age at diagnosis, cellularity, chemotherapy, and ER, PR, and HER2 status, z-scores for gene expression, and survival time in months.

We will do supervised learning to make predictions or decisions based on labeled training data. In this case, we know the response variable we want is survival in months, and we have a multitude of known input data points. We hope to train a model that can depict the relationship between these clinical and gene attributes and survival in months. Since we have data on patients' clinical attributes as well as their survival in months, we already understand the relationship and want to create a model to represent this known relationship and predict outcomes in the future. Thus, in this case, we are doing a regression analysis because we have a numerical output which is survival in months. We want our model to be able to use the input data about a patient in order to predict their survival in months.

We plan to use partial least squares regression (PLSR) for our analysis. Our predictor variable (X) will be gene expression data and clinical attributes like age at diagnosis, treatment method, and hormone receptor (ER, PR, HER2) characteristics. Our response variable (Y) will be survival in months. Since gene expression data is already z-scored, other forms of data cleaning can be conducted if needed like logarithmic conversion and handling of outliers. PLSR can be conducted using the corresponding algorithm in Python using many components and all gene data. The percentage of variance explained for Y versus the number of components can be plotted. Some threshold can be established for the amount of variance we want to be explained, and the number of components needed to explain this degree of variance can be determined. Based upon this value, we will repeat PLSR using this determined number of components. The weights obtained from the PLSR model can be normalized to ensure magnitudes do not influence VIP scores disproportionately. Then, VIP scores can be calculated to rank genes based on their importance, and genes with a VIP score greater than 1 will be selected for use in the model due to greater importance in predicting survival time. PLSR will be conducted once again, but this time using the determined number of components and the genes with high VIP scores. Then, PLSR results can be used to fit a model representing predicted survival time based on gene expression data. The conclusions made from the modeling will be visualized on multiple plots with the individual genes and their expression that were found to be most predictive of survival against the survival in months of the patients.

To know if the model developed using PLSR is accurate, the observed survival time can be plotted against the predicted survival time, and a line of best fit can be created to visualize the model's performance. The R^2 value will be computed to quantify how well the model fits the

data and determine how effectively gene expression and clinical variables predict survival time. Leave-one-out validation can also be performed by removing one patient at a time from data sets, predicting survival time, and then recomputing R^2 in order to determine if any patients are disproportionately affecting the output of the model. These R^2 will be visualized in a matrix of different leave-on-out validation simulations.

Potential issues are that our input data does not only contain one type of data, like just gene expression. Instead, there are a variety of other variables that will be included in addition to gene expression like age at diagnosis, treatment options, and receptor status. Thus, it is unclear whether our PLSR procedure will be able to handle these varied types of variables when predicting survival time. If there are issues with data compatibility in the input data, we can try to normalize the clinical attributes differently, or handle feature selection differently for clinical attributes and gene expression before creating the model. Worst case, we can choose to just use gene expression data to create our model. Another issue is the high dimensionality of gene expression data resulting in a model with a poor R^2. To remedy this issue, we can perform further cleaning or grouping of gene data through an algorithm like principal component analysis (PCA) to reduce the dimensionality of the data. Given that the gene expression data are numerical variables that are highly correlated, using a PCA might be necessary for more accurate outcomes. If our approach fails, we will learn that the provided input or data set may just not provide any meaningful predictive value. We will also learn more about the importance of feature selection to produce meaningful results with true predictive power through exploring new methods to choose what variables will be important for our model.

In order to prepare the data specifically for our analysis, we want to ensure that the data is normalized and scaled correctly. The gene expression data is already Z-scored, so we want to ensure that the clinical variables are standardized. For example, we want to make sure that categorical variables are converted into binary format when appropriate. We will apply one-hot encoding for categorical clinical attributes like ER, PR, and HER2 status (hormone receptor status).