04-analysis.Rmd

Analysis

Once my partner and I felt like we had a solid understanding of these regression methods, we felt like we were ready to run these regressions on our dataset. One of us was going to work on Partial Least Squares Regression, Ordinary Least Squares Regression, and Ridge Regression. Meanwhile, the other would work on Lasso Regression and Principal Component Regression.

Despite the differences between each regression, the formatting and workflow for each regression script remained quite similar. We would begin by first running a specified regression on the training set. We would use built-in cross-validation functionality in R's regression functions to find the optimal regression parameter. Afterwards, we would take our model and our optimal parameter and use them to predict values from our test dataset. We would then calculate the model's mean squared error.

Amidst all of this, we would also save certain images, computations, and R objects that we thought might be informative. For instance, for each regression, we saved a plot, which showed how prediction accuracy varies with a regression's parameter. We also saved each regression's estimated beta coefficients, which came from running a regression and its optimal parameter on the whole dataset.

In order to run all these regressions, my partner and I heavily relied on two key packages in R: **glmnet** and **pls**. The **glmnet** package contained functions that helped with our shrinkage regressions (Lasso and Ridge) while the **pls** package contained functions that helped with our dimension reduction regressions (PLSR and PCR).