3. Methods

In this section, we present the methods and procedure to build predive models and select the best model that predicts the target variable, balance. We fit and tune five different regression models namely OLS, Lasso, Ridge, PLSR and PCR.

3.1 Data preparation

We first consider to split the original dataset into train, validation and test. However, since modeling libraries internally support cross validation by splitting train into train and validation, we decide to split the original dataset into train and test data. It is important to hold out test data and fit models on train data because the held-out test data will later be used to measure the effectiveness of final models. If it is included in a training phase, the models overfit the data and bring about very optimistic MSE but do not generalize to future observations.

3.2 Evaluation

Since the goal of this analysis is to figure out which model works the best, we need to have an evaluation criterion to compare the effectiveness of each model. Thus, we choose a common evaluation criterion for regressiom model, Mean Square Error. To effectively calculate MSE for each model, we calculate MSE for each fold ten times using 10-fold cross-vaition and average them to obtain the averaged MSE for tuning parameters for each model. However, this process is automatically supported by libraries. Finally, we fit each model using test data and compare MSE from test data.

3.3 Model description and hypter-parameter tuning

While training each model, we need to find optimal paramters for each model. In order to effectively select hyper-parameters, we use 10-fold cross-validation. For lasso and ridge, lambda is the tuning parameter. It determines how much we will penalize models for high weights on predictors. If lambda is high, it penalizes models more and ends up generating sparse models. This kind of models is called shrinkage method because they shrinken weights or even remove predictors by penalizing models. These models are especially good options when there are many predictors. By penalziing or removing unnecessary predictors, they provide more interpretable results. Thus, they are often utilized in genomic and pharametical analysis. For PCR and PLSR, the number of principal components is the tuning parameter. They both internally use the precipal component analysis to obtain pricipal components, which are linear combination of original predictors in dataset. Based on spectral theorem, the eigenvector corresponding to highest eigenvalue is the direction that explains the most about the variability in data, which is the first principal component. We need to find the optimal number of subsets of PCs that summarize the entire data set without harming accuracy. Again, we use 10-kold cross validation to obtain the optimal number of principal components.

3.4 Model comparison

As mentioned on 3.2 evaluation, we use MSE to compare models and select the best model. Although MSE does not provide an absolute means of model accuracy, it provides a relative measure to compare models. Thus, we finalize our model with the lowest MSE.

In summary, the following is the the procedure for each model.

1. Split the data into train and test, 80% and 20% repectively.
2. Train a model using train data with 10-fold cross-valdation.
3. Pick the optimal hyper-parameters.
4. Predict balance using the model with the optimal parameters.

5. Calculate Mean Square Error.
6. Record both Mean Square Error and coefficients.