Exercise 2

Advanced Methods for Regression and Classification

November 10, 2022

Use again the `College` data (preferably transformed) as for the last exercise, and split it into training and test data. For the following tasks, always consider the RMSE as evaluation measure.

(1) Take the full model from the last exercise. Use the function `cvFit()` from the package `cvTools`. A look at the help file shows how you can use the function for model evaluation based on cross-validation (CV). Set the parameters `cost=rmspe` for RMSE calculations, and perform 5-fold CV with 100 replications. A plot of the result object shows the distribution of the resulting error measures. What do you conclude?

(2) *Best subset regression*:

(a) Use *best subset regression* which is implemented in the `library(leaps)` as the function `regsubsets()`, see help and course notes. To find the models, examine the best 3 models of each size, for a maximum model size of 10 regressors.

(b) Plot the results. Which model seems to be the best?

(c) Save the resulting `summary` as another object. Display the structure `str()` of this object and plot the size of models against the BIC values. Which is the best model? Apply `lm()` on the final best model and interpret the results of `summary()`. Compare the result using `cvFit()` with that obtained from (1).

(3) *Principal component regression (PCR)*:

(a) Apply PCR, which is implemented in the `library(pls)` as the function `pcr()`, see help. Perform cross-validation using 10 segments (see help of `pcr()`) and scale the data (`scale=TRUE`).

(b) Plot the obtained prediction errors from cross-validation, see lecture notes. How many components seem to be optimal? What is the resulting RMSE?

(c) Use the function `predplot()` to plot the measured y values against the cross-validated y values considering the optimal model.

(d) Do the same as in 3.(c) for the test data, and compute the RMSE.

(4) *Partial least squares regression (PLS)*:

(a) Apply PLS, implemented in the `library(pls)` as function `plsr()`, see help. Apply the function similarly as in 3.(a).

(b)-(d) Perform the same tasks as in 3.(b)-(d) for the PLS model. Compare the outcomes with the PCR model outcomes.

(5) *PCR "by hand"*:

Perform principal component analysis (PCA) on the scaled x-variables (of the training data), using the function `princomp()`. Use the PCA scores, available as list element `$scores` from the PCA result object, and take the same number of those components as in 3.(b) as input variables for regression on our response. Use the model to compute the RMSE for the test data. You should obtain the same result as in 3.(d).

Hint: The test data are obtained according to the formula $\mathbf{Z} = \mathbf{X}\mathbf{V}$, see course notes, where $\mathbf{V}$ is the list element `$loadings` from the PCA result object, and $\mathbf{X}$ needs to be centered and scaled appropriately.