Multivariat øving 7

Siawash Naqibi

April 4, 2021

Make PCR model and interpret various plots

The scores plot shows us the distribution of the data on the principal component frames. PC1 accounts for 33% of the total variation of the data and i guess 37% of the total variation of the target. PC2 accounts for 18% of the data and 28% of the target and so on.

The correlation loading plot shows helps to visualize the stricture of the data. The correlation loading also shows how much of the variance for each variable is represented by a certain principal component. We can observe that 64% of the total variation of the variable PTRATIO is captured by pc2. We can also observe that 56% of the total variation of LSTAT is captured by pc1. The correlation loading plot also shows how variables that are close to each other are positively correlated. Based on this we can see that the variables NOX, INDUS, and AGE say more or less the same thing as they are very closed to each other.

The influence plot shows the F-residuals vs. Hotelling's T2 statistics. They represent two different kinds of outliers. The residual statistics describe the sample distance to the model, whereas the hotellings statistics describe how well the sample is described by the model. Samples that are lying in the upper region are poorly estimated by the model. In our case we do have 2-3 samples that are poorly estimated and maybe should be taken out from the training set. However, the sample with the highest residual value (R103) has also quite low leverage on the model and will therefore not affect the model that much. The sample with the highest leverage on the model (R156) should be taken out because it has a high influence on the model.

The predicted vs. reference plot shows the predicted y value with the measured y value. It is a good way to check the quality of the regression model. If the model gives a good fit, then the plot will shows points close to a straight line through the origin and with slope equal to 1. In our model for up to 5 principal components, we have an offset value of 4.7, slope of 0.81 and RMSE of 3.7. Although this is good, maybe certain values should be removed. These are the values R165 R156 and R157.

The regression coefficients summarizes the relationship between all predictors and the response. For PCR the regression coefficients can be computed for any number of principal components. For for example 3 principal components, the regression coefficients will summarize the relationship between the predictors and the response as three principal components models the data. The weighted regression coefficients informs about the importance of the X variables. For pc1, the highest positive variable was DIS with a value of 0.057 with the highest negative value was NOX at -0.087.

Based on the observed plots, especially the scores plot we can see that to capture all the variation in the target value and to capture as much variation the data as possible we will need 5 principal components. This is also backed up the the preicted vs. reference plot.

Recalculate without some of the variables

Took out the variables RAD, NOX, INDUS, and AGE. Because NOX, INDUS, and AGE variables had a strong positive correlation and could there be taken out. Furthermore, RAD and CHAS had a strong correlation as well. If we look at the predicted vs. reference plot we can get an indication on how well the new plot went. The slope has now changed to 0.82 and the RMSE value has changed to 3.56. So the whole model improved by reducing the number of variables. The model can still be improved because there are certain outliers in the model that should be taken out. I frankly don't know how to take out certain values and recalculate the model.

Predict the test set

The prediction on our first model with 5 principal components over the test set yielded a slope of 0.6 and an RMSE of 8.15 so it got much worse.
The prediction on our second model with 5 principal components over the test set yielded a slope of 0.73 and an RMSE value of 10.46 so the prediction over our secodn model yeilded a much better slope but also a worse RMSE value.

Do the same with PLS regression

Comparing the PCR with the PLS we can observe clear differences. In PCR with 5 PCs we had a slope of 0.81 and RMSE of 3.7. IN PLS with 2 PCs we had a slope of 0.82 and RMSE of 3.54. The reason for this is because whereas in PCR we needed 5 PCs to explain 81% of the variance of Y, in PLS however with can explain 82% of the variance of Y with only 2 PCs. Our observation is in line with the theory behind PLS because PLS tries to find the principal components such that the we capture as much of the variance of Y as possible.

Are the regression coefficients the same?

The regression coefficients are not the similar. For PCR the most important variable for pc1 is NOX with a value of -0.087, whereas for PLS the most important variables for pc1 is RM with a value of 0.269. Both PCR and PLS however put CHAS and RAD as one of their least important coefficients. The strctural differences between the different methodologies exist because of what they are trying to maximize. PCR tries to maximize the variation of the data X whereas PLS maximizes the variation of the target Y.