# Assignment 7 – PCR

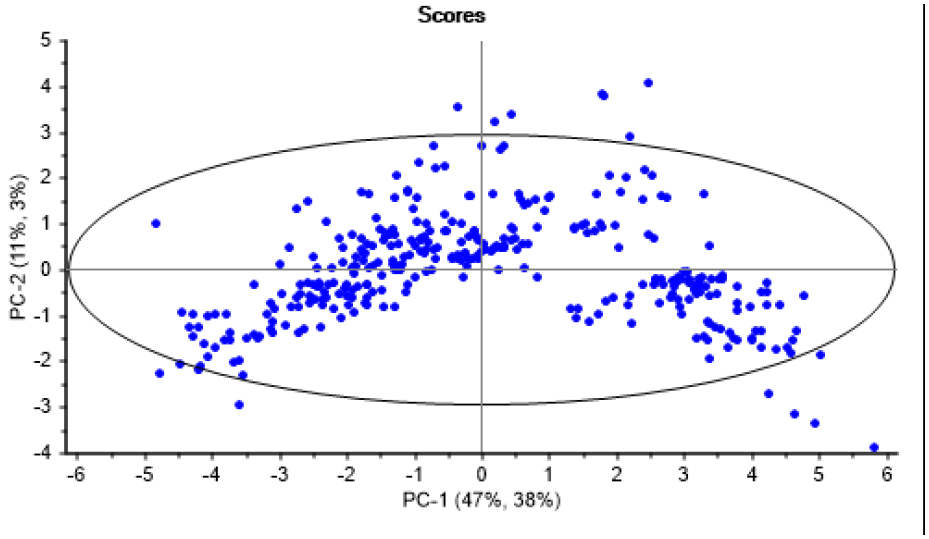
Principal components regression (PCR) is a regression technique based on principal component analysis (PCA). The basic idea is to calculate the principal components and then use some of these components as predictors in a linear regression model fitted using the typical least squares procedure. In some cases a small number of principal components are enough to explain the vast majority of the variability in the data. Say you have a dataset of 50 variables that you would like to predict a single variable. By using PCR you might find that 4 or 5 PCs are enough to explain 90% of the variance of your data.

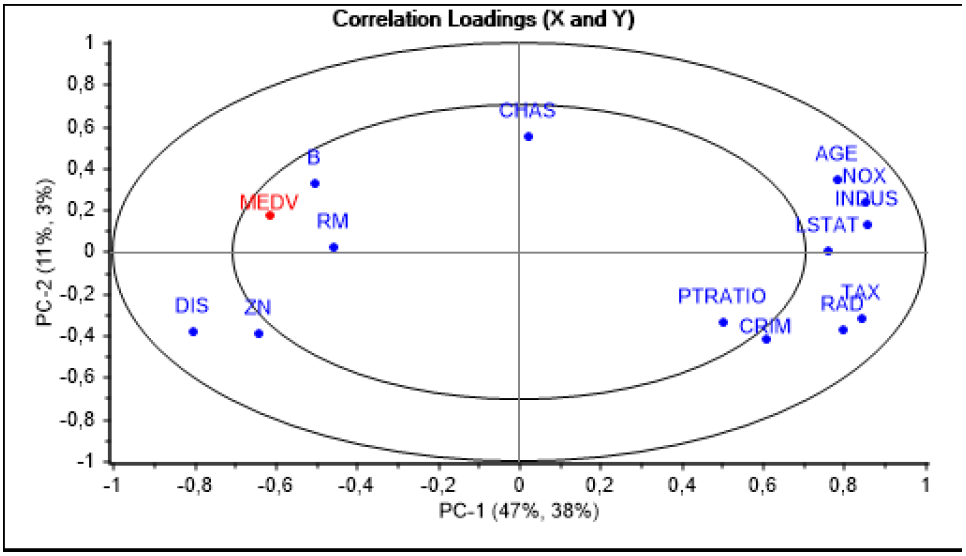
A core assumption of PCR is that the directions in which the predictors show the most variation are the exact directions associated with the response variable. On one hand, this assumption is not guaranteed to hold 100% of the times, however, even though the assumption is not completely true it can be a good approximation and yield interesting results.

**Assignment:**

The Boston Housing data set was analyzed by Harrison and Rubinfeld (1978) who wanted to find out whether ``clean air'' had an influence on house prices. The objective with the analysis: From the available data, can this hypothesis be confirmed?

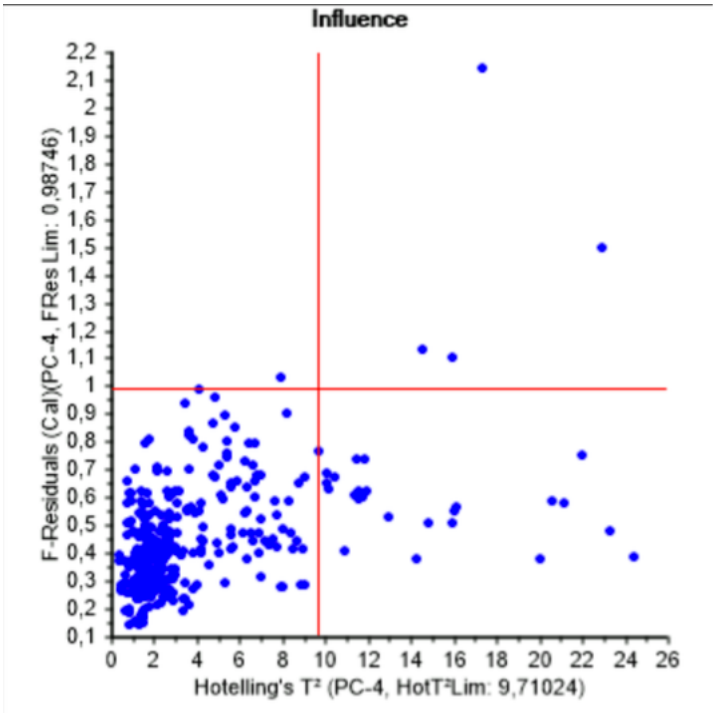
Dividing the system into test and training sets as described by the assignment, these are the results:



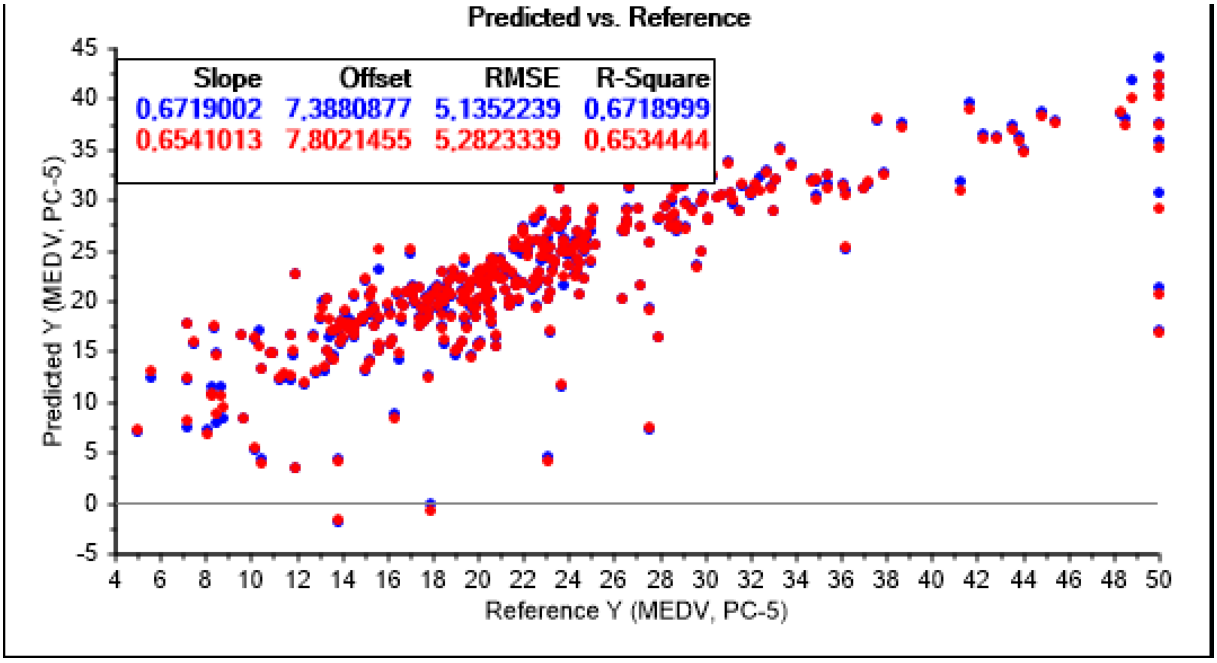



**Scores:** TAX, PTRATIO, RAD and CRIM seem to be important variables for the group in the lower right corner based on the loadings plot. Also the ellipse shows the outliers. A 95% confidence interval is shown, i.e. 5% of the measurements will lie outside.

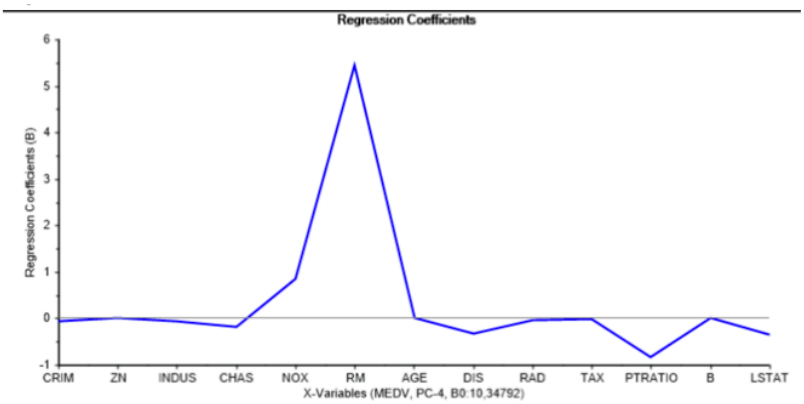
**Loadings:** Several of the variables are strongly correlated as we can see in the outer right part. Also DIS and ZN are negatively correlated, CHAS seems to be uncorrelated while the rest seems to be correlated to some of the other variables.




**Influence:** Most of the samples are inside both the Hotelling´s T^2 criteria and the F-residuals criteria.

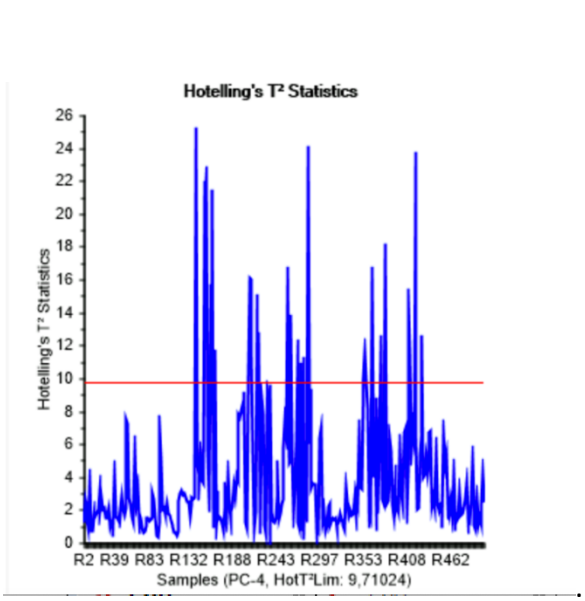


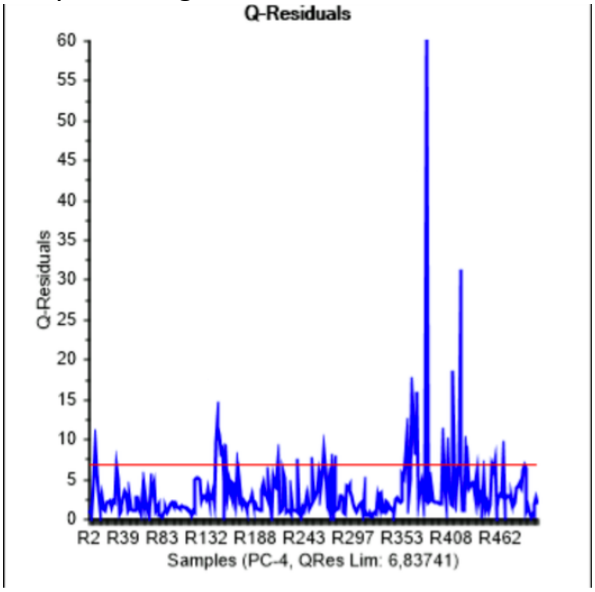

**Predicted vs. reference:** The deviation from the ideal model is shown in this plot. We want the R-square value to be as close to one as possible, as it is the explained variance. The plot also shows to what extent high values are underpredicting and low values are overpredicting.




| Beta Coeffi... | | CRIM | ZN | INDUS | CHAS | NOX | RM | AGE | DIS | RAD | TAX | PTRATIO | B | LSTAT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| MEDV | 1 | -0,0726 | 0,0046 | -0,0666 | -0,1884 | 0,8487 | 5,4515 | 0,0057 | -0,3223 | -0,0466 | -0,0034 | -0,8315 | 0,0023 | -0,3459 |

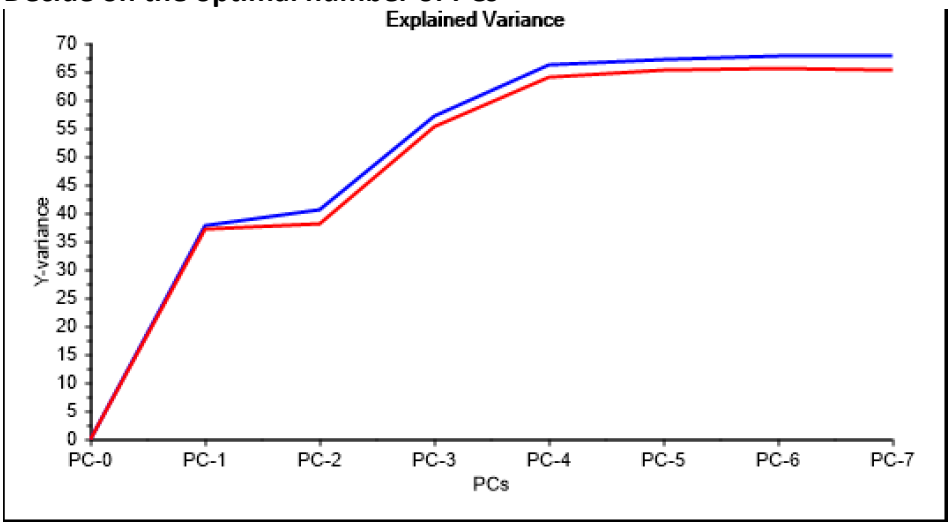
**Regression coefficients:** RM contributes the most, and next is NOX. The other values either have close to zero, they do not contribute, or they have negative contribution



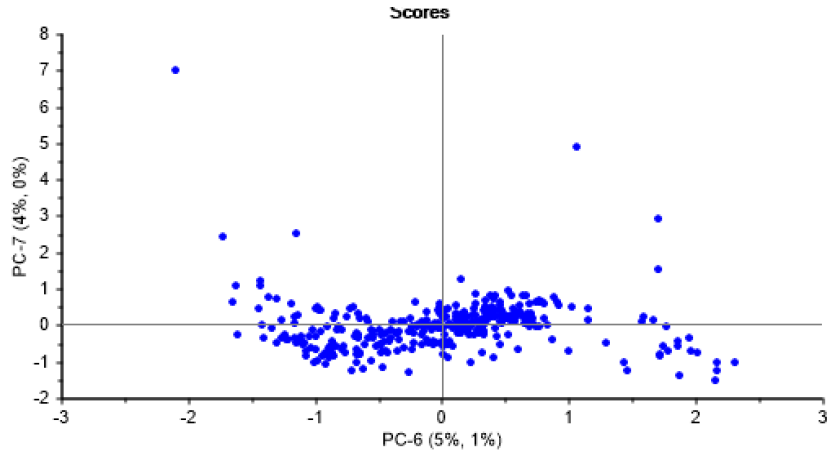




**Residuals:** The second plot shows that the error in the predicted values varies. We should investigate if there are some outliers or any other reason for the peaks over the critical limit.

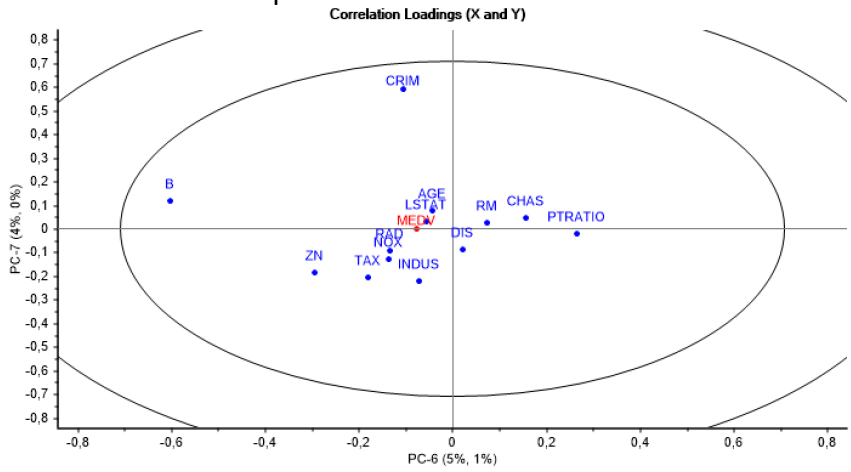
**Decide on the optimal number of PCs**



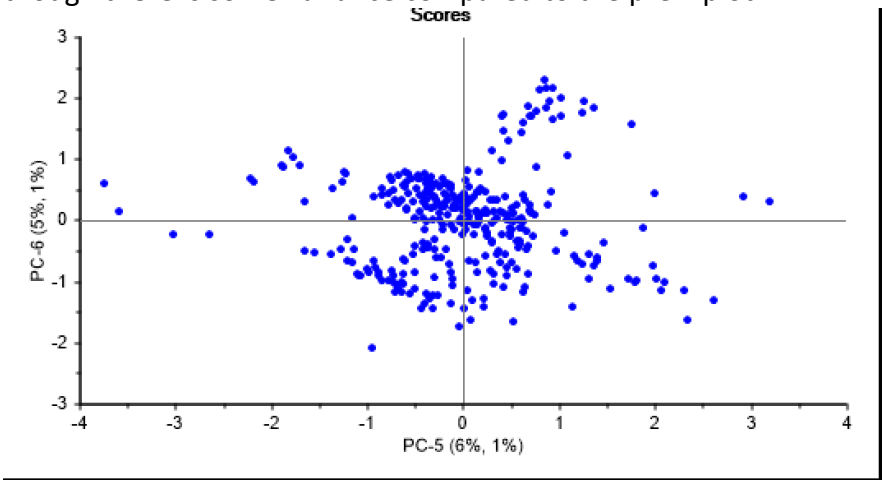

Based on the explained variance the optimal number of PCs seems to be 4. From PC-4 the explained variance does not get much higher, but stays at approximate 70%. Looking at the scores and loadings for the different PCs the scores plot show a low score and few outliers for PC-7:

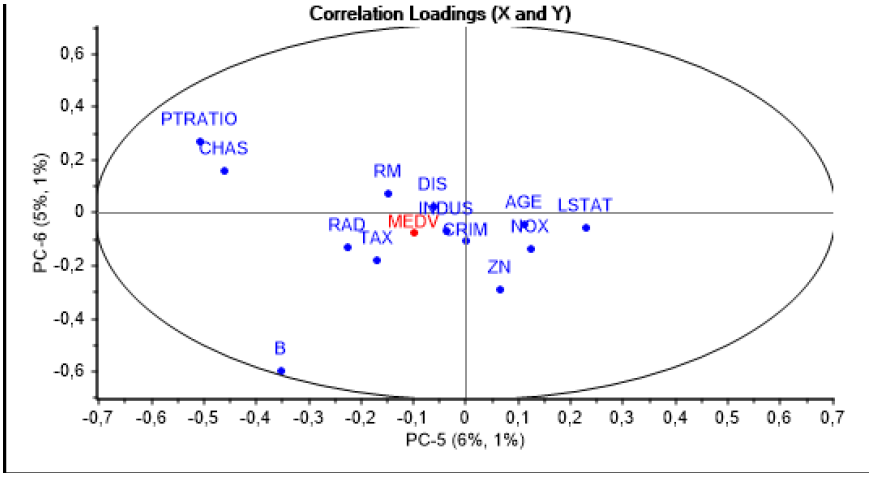



The loadings for PC-6 and PC-7 show that most of the loadings lie close to zero with a few closer to the outer ellipse.



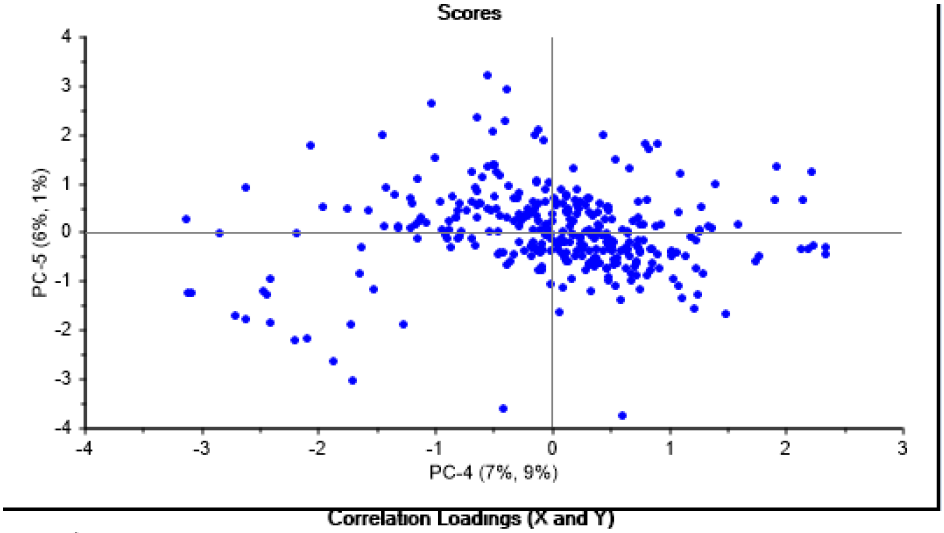

The scores and loadings for PC-5 and PC-6 are similar to the ones from PC-6 and PC-7, though there is some variance compared to the prev. plot.

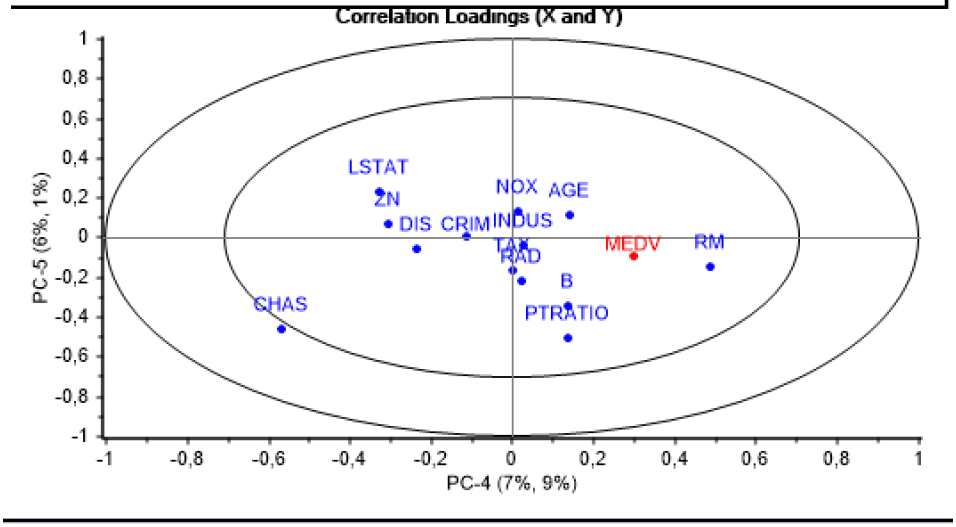





The plot is a bit zoomed in, therefore the outer ellipse does not appear. No points outside the inner ellipse.

The scores and loadings are more spread when looking at PC-4 and PC-5, though the variance is still low.

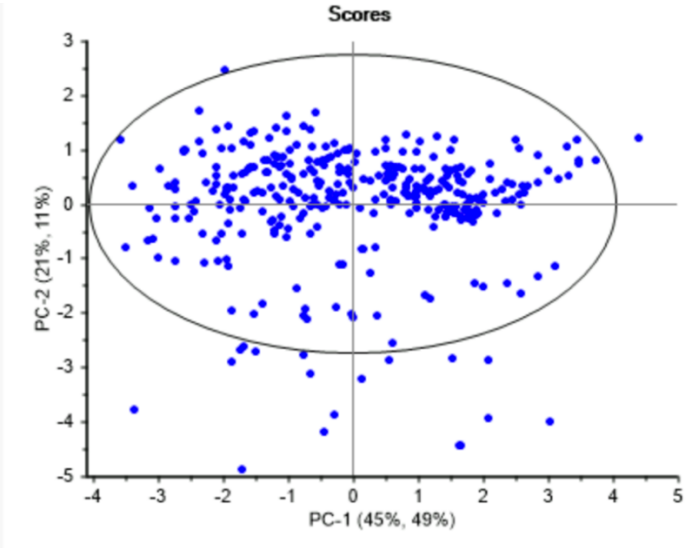


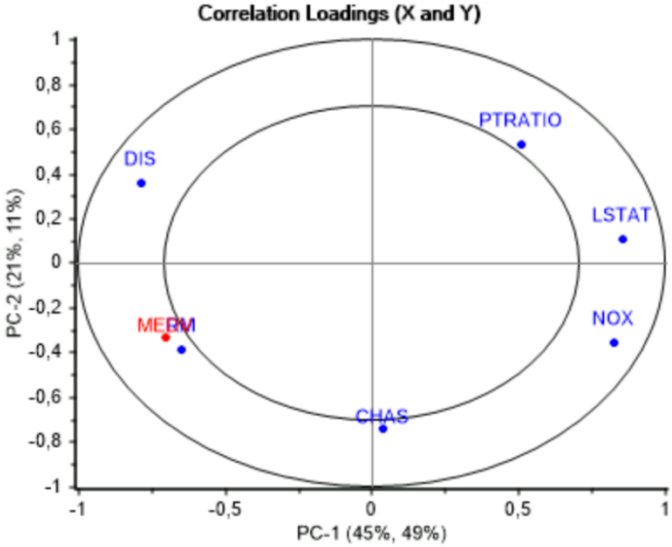

Looking at the explained variance and also the scores and loadings above I decide on 4 PCs as the best choice.

**New model – removed variables**

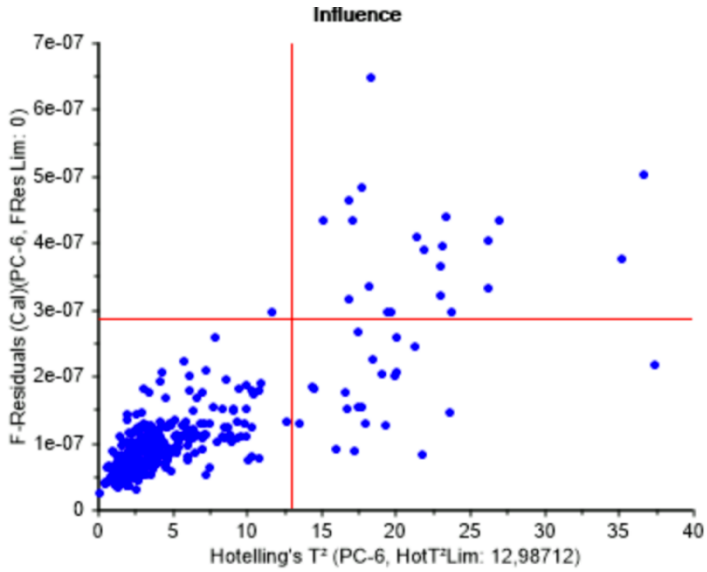
Now the variables AGE, TAX, B, CRIM, ZN, RAD and INDIUS are removed. The new model will be compared to the one presented above.



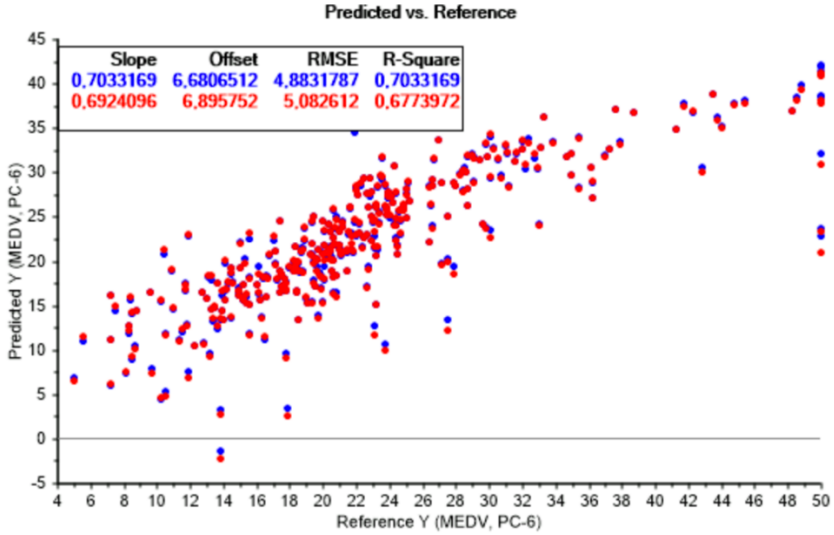

Looking at the scores plot for PC-1 and PC-2 we can see that more of the variance is being covered.



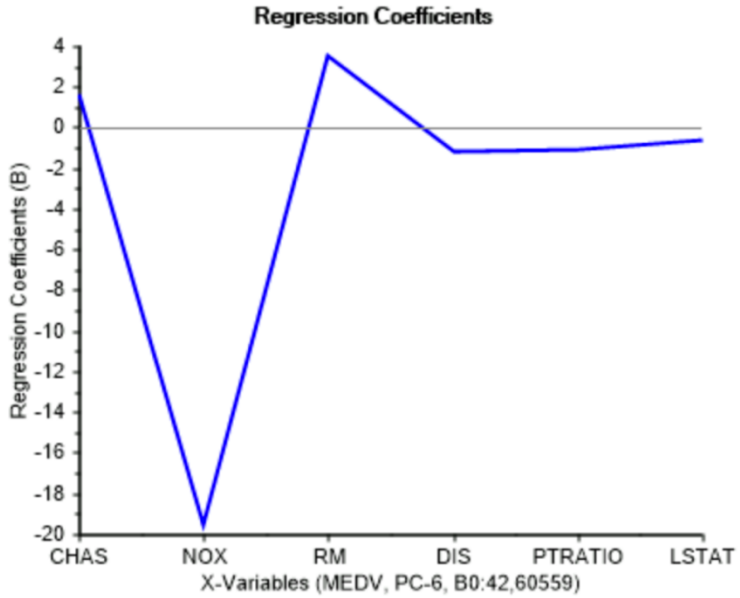

Comparing to the previous plot all variables are now less correlated. The variables are in the outer ellipse, meaning that they contain more information.



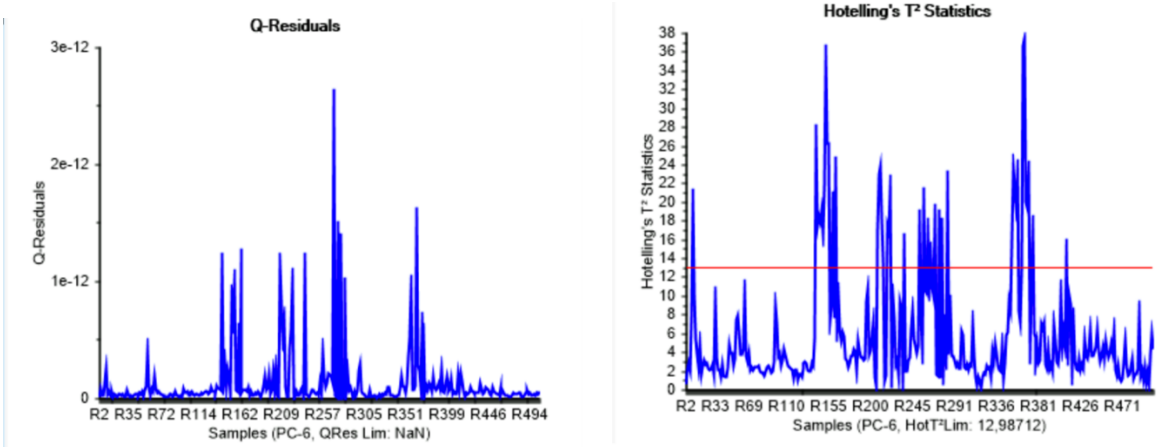

Compared to the previous model there are more variables crossing the critical lines, though most of them are inside both the critical lines



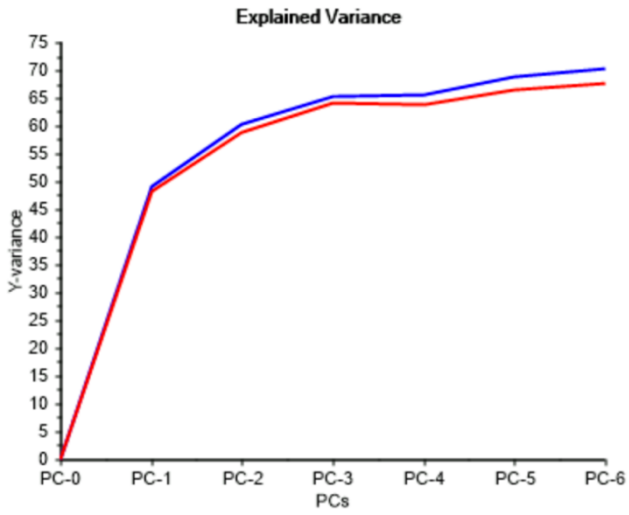

The R-square value is higher, and RMSE is lower, which is what we want.




RM still has the highest contribution in positive direction, but it is a bit lower in the new model. NOX has now a negative contribution, which is interesting as it was the value contributing the second most in a positive direction for the previous model.



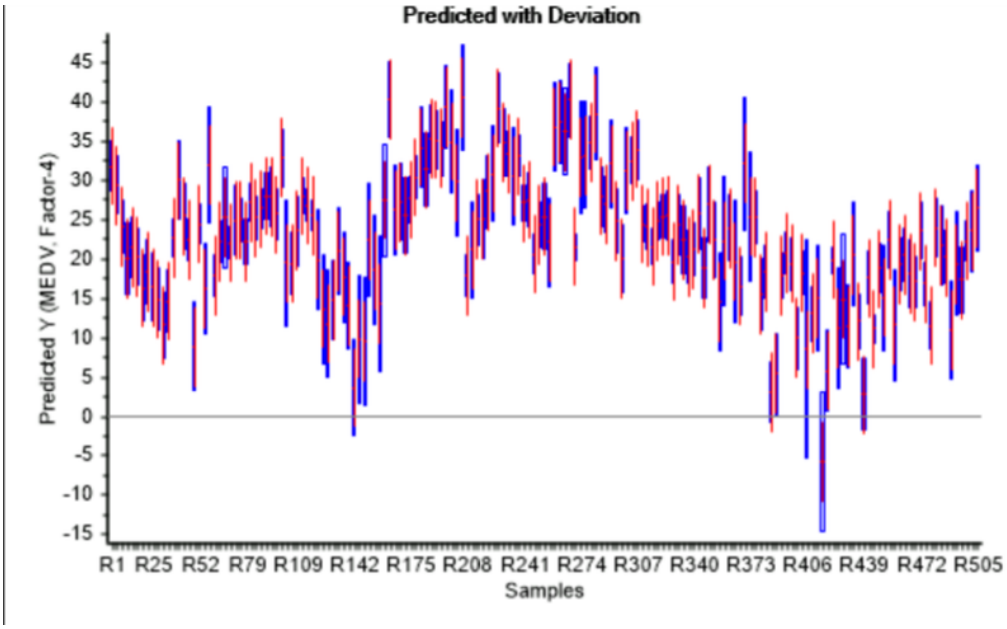

Compared to the previous model the Q-residuals are now small and under the critical limit. The Hotelling´s T^2 has increased, but it still has the same shape.



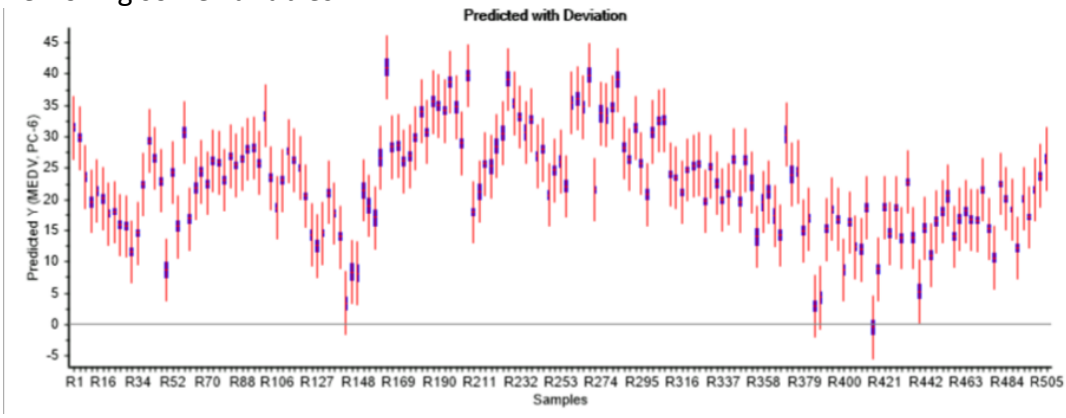
Looking at the explained variance is similar to the one in the previous model, where 4 PCs = approx. 70%. One difference is that the first PCs have a higher explained variance, e.g. PC-2 has a new explained variance at about 58% while in the old plot the explained variance for PC-2 was ca. 36%.

**Now predicting the test set and looking at the RMSEP compared to the training model.**

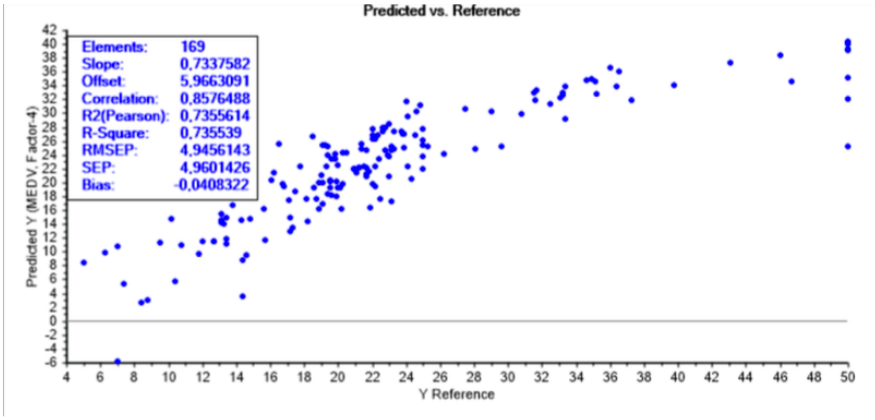
With all variables:



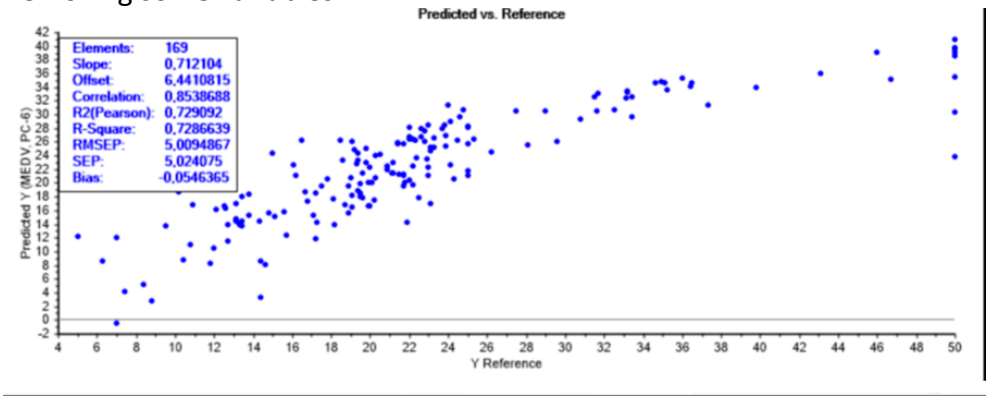

Removing some variables:



All variables:

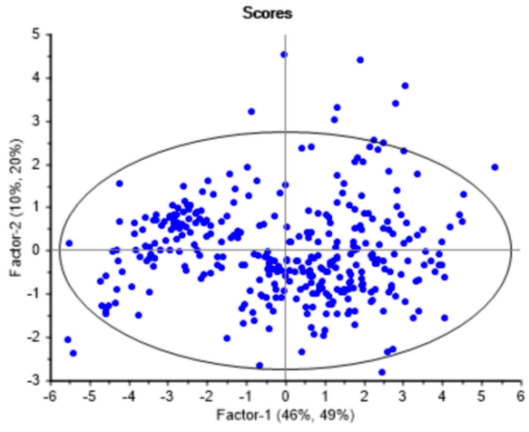


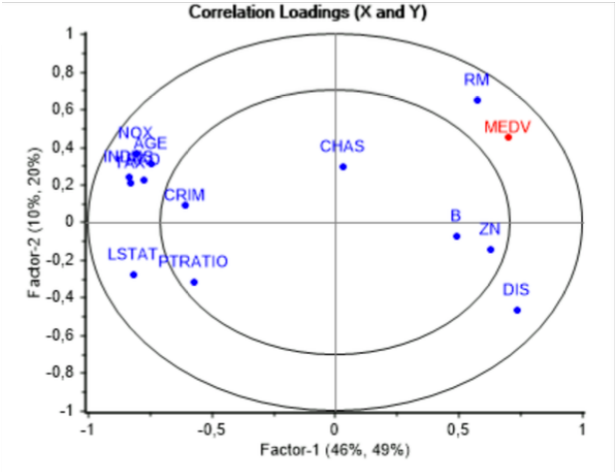

Removing some variables:



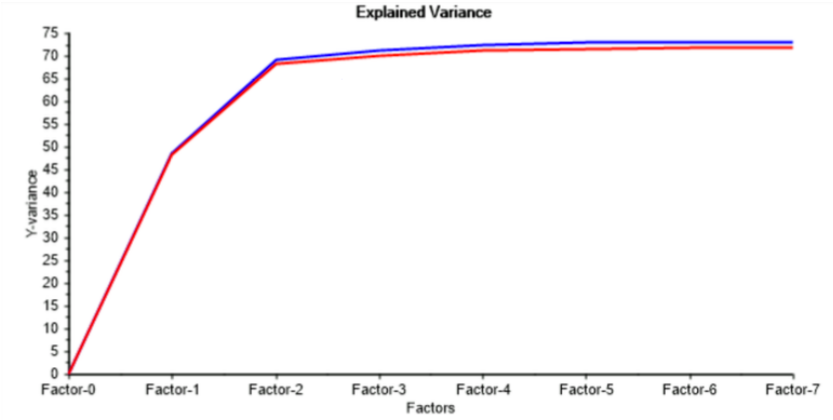

Looking at the model containing all variables the RMSEP has a value of 4.94, compared to the model where some of the variables are excluded, where the RMSEP has a value of 5.0. The first model is thus, based on the RMSEP the best, as it has the lowest value. On the other hand the model performs better on the test set where some of the variables are excluded.

**Doing the same for PLS regression:**

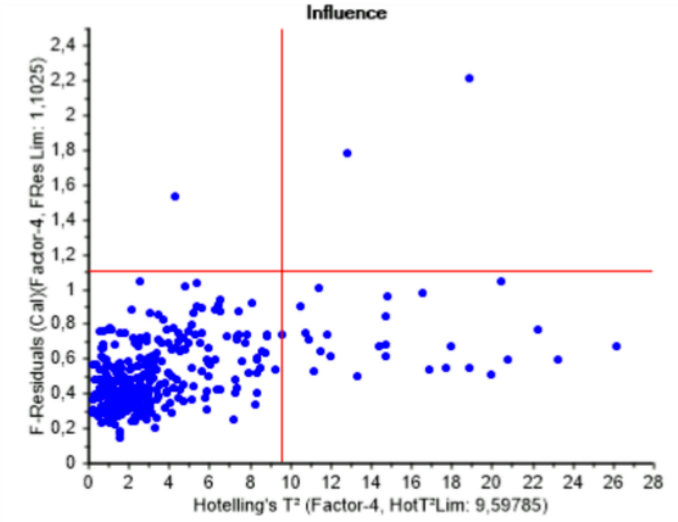


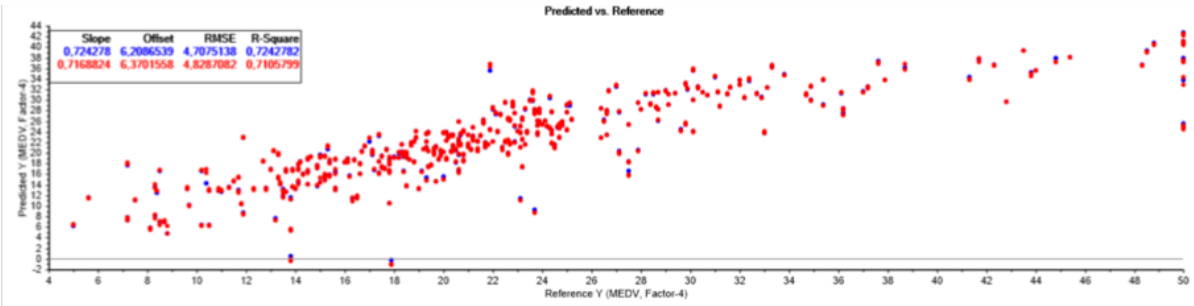
Some of the variables are inside the inner ellipse. Also a lot of the variables are strongly correlated to each other.



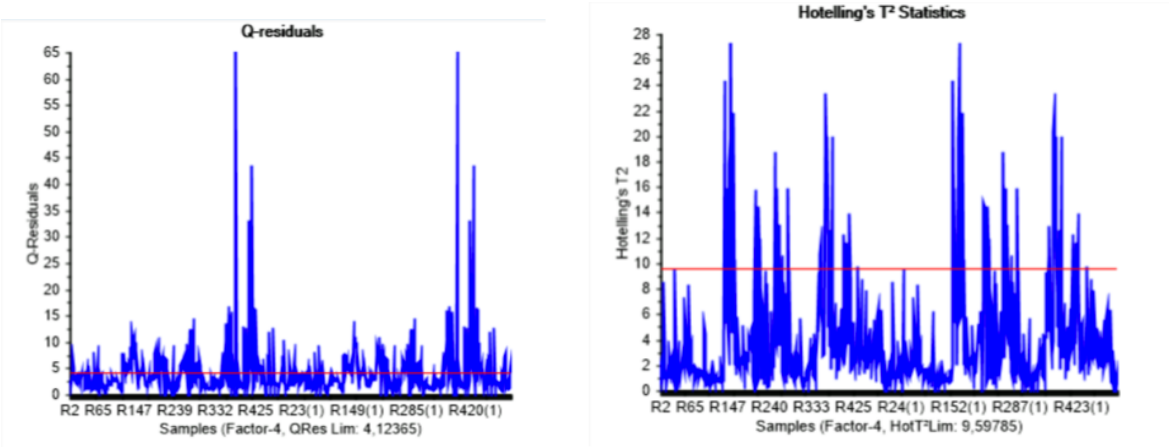

With 2 PCs the model is now able to cover 70% of the explained variance.




Most of the samples are inside both critical lines, but we can see that there are some crossing the Hotelling´s T^2 critical line.

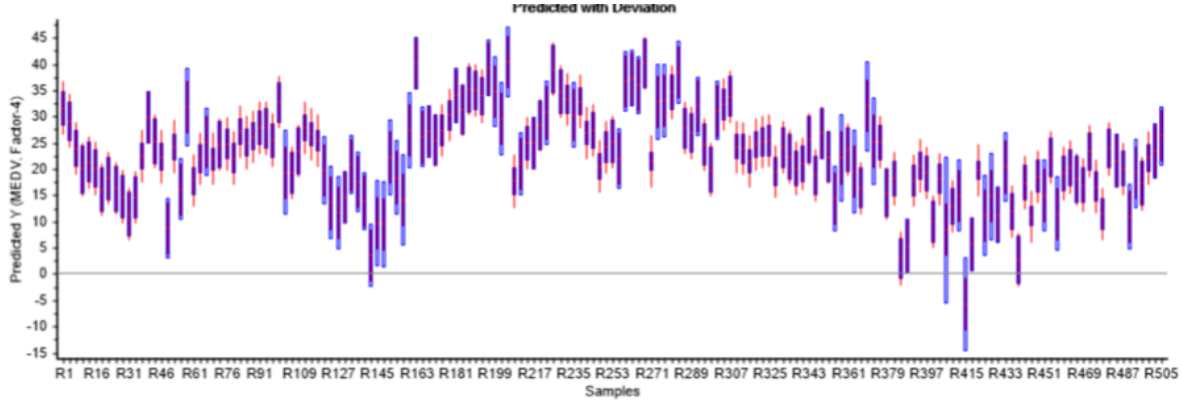


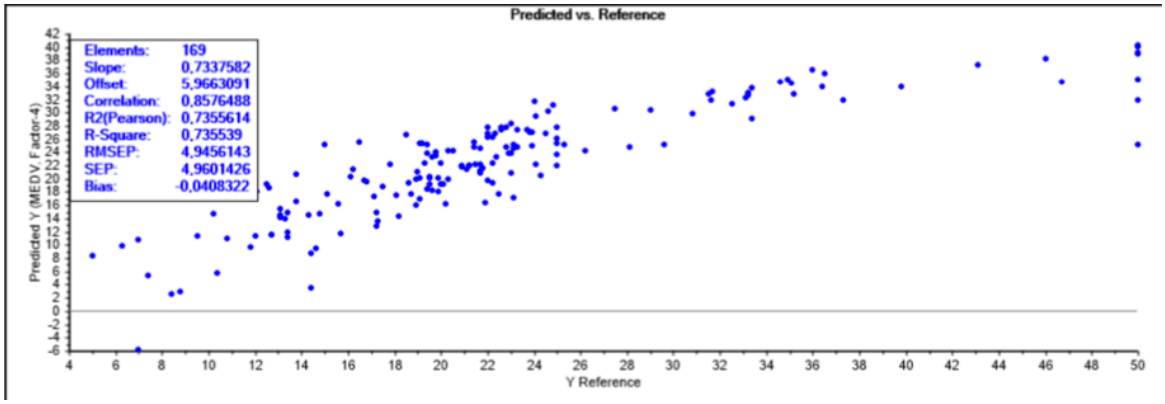

The predicted vs. reference plot shows that both the RMSE and R-square perform better than when using PCR.



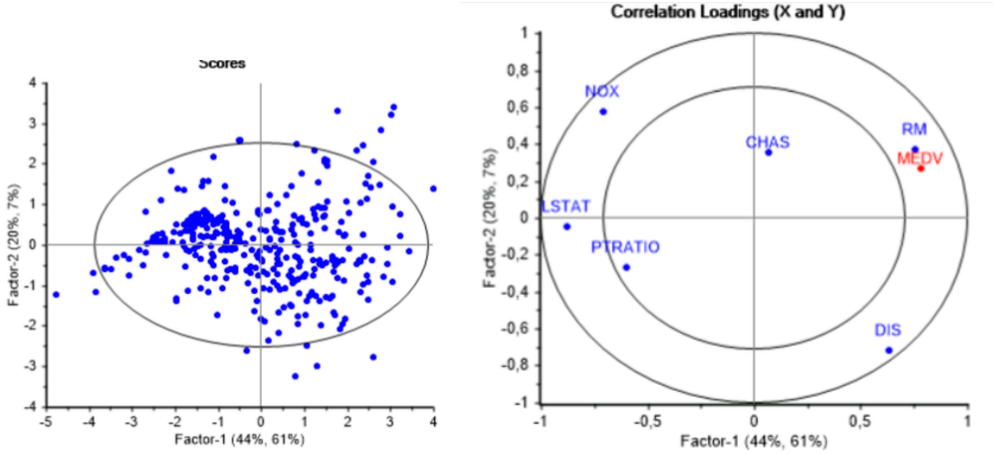

Looking at the Q-residuals we can see that there are several samples lying outside the critical limit.

Prediction:

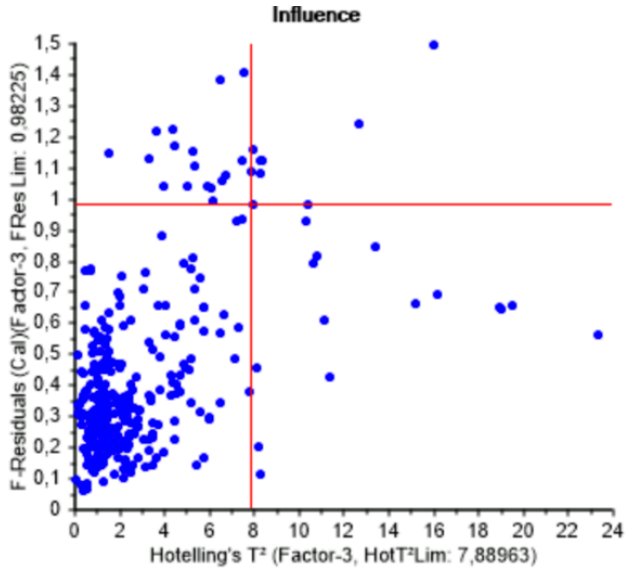


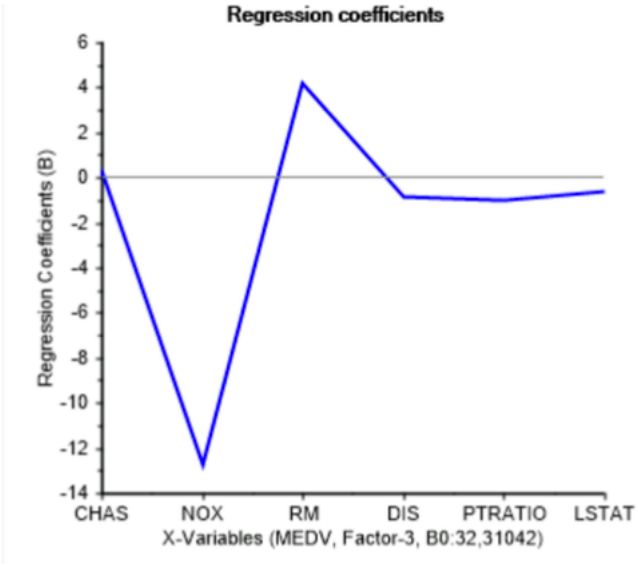
**Recalculating without the same variables as for PCR:**



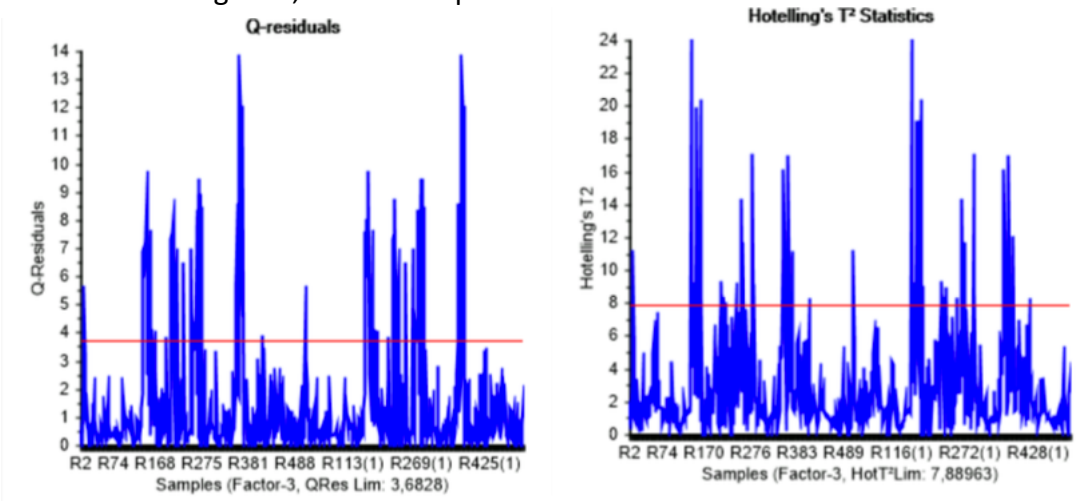

The number of variables inside the inner ellipse have been reduced, showing that removing some of the variables leads to less correlation between the remaining variables.



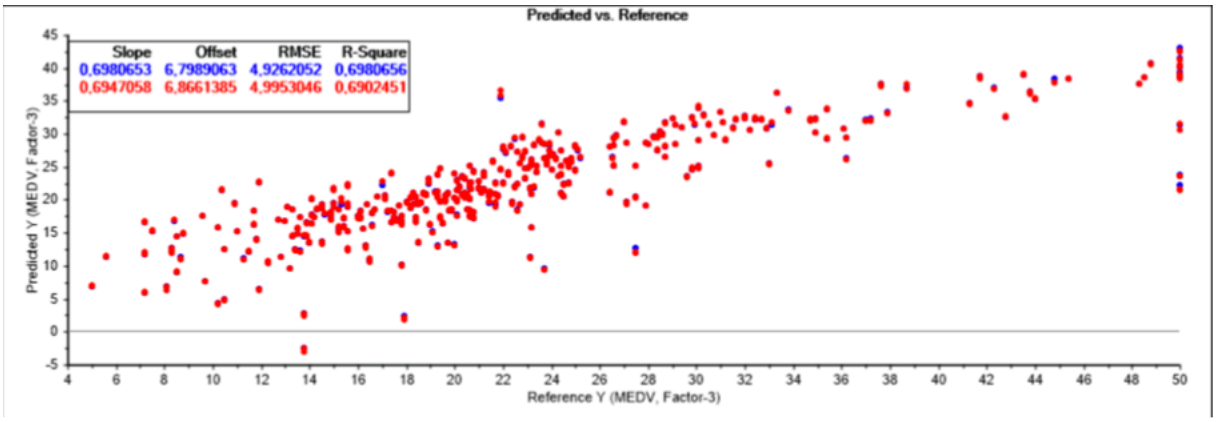

Compared to the model with all the variables, the new one has less samples crossing the Hotelling´s T^2 critical line. There are a few more crossing the F-residuals critical line, but the majority are still inside both lines.




NOX is not as negative, but the shape is the same.

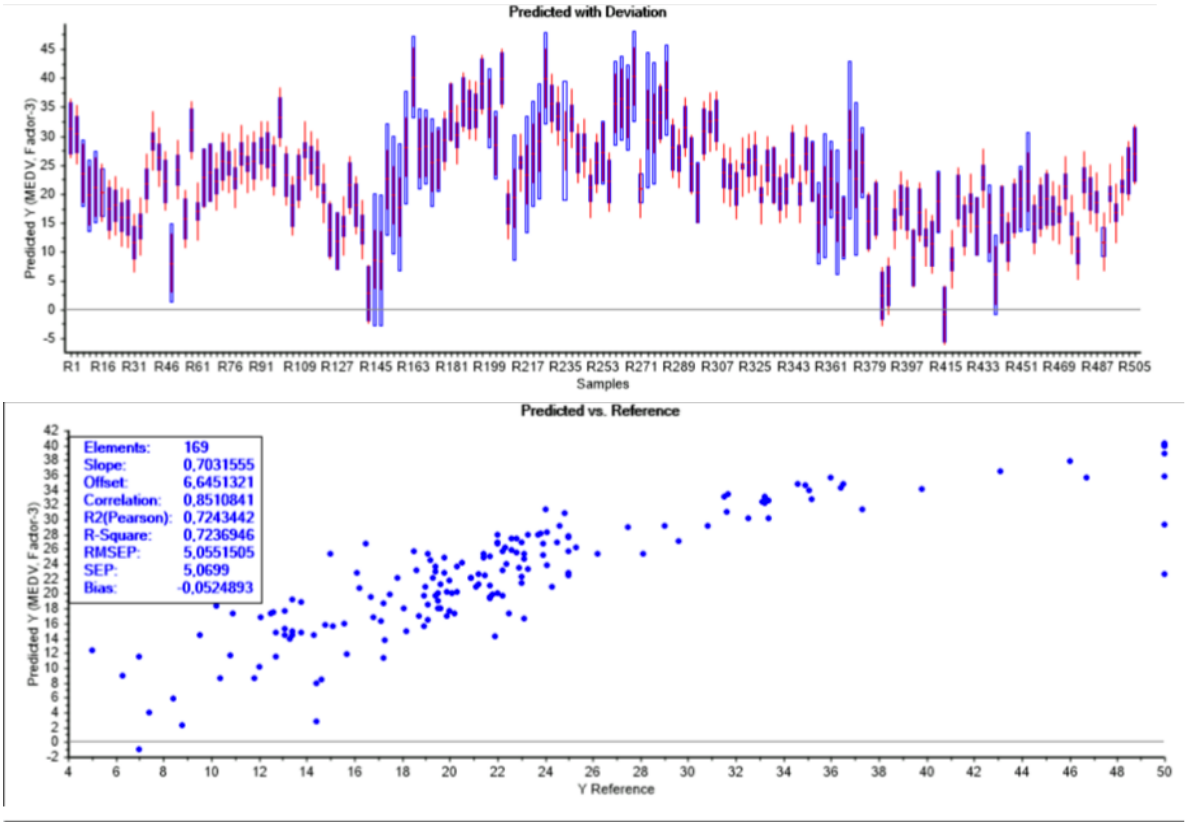



The number of tops outside the critical limit has increased in the Q-residuals.




Comparing to the model with all variables we can see that RMSE has increased and the R-value has decreased. This is not desired, and shows that the model might not have gotten better from removing variables.

Prediction:




As mentioned the number of PCs in the PCR model should be 4, and 2 in the PLS model. The difference is due to that the scores and loadings will be used to model y in the next step, but this is unknown to the PCA algorithm. Since the main systematic variance in X is not relevant to model y, PCR needs more PCs, while the loading vector in the PLS is calculated as the individual correlation between X and y. PLS will rotate the scores towards modeling the variance in bot X and y.

Looking at the regression coefficients they are not that different, though MLR is more different than the coefficients for PLS and PCR.