Assignment 7

Interpret the various plots (scores, loadings, influence, predicted vs. reference, regression coefficients, residuals). Decide on the optimal number of PCs to use for prediction.

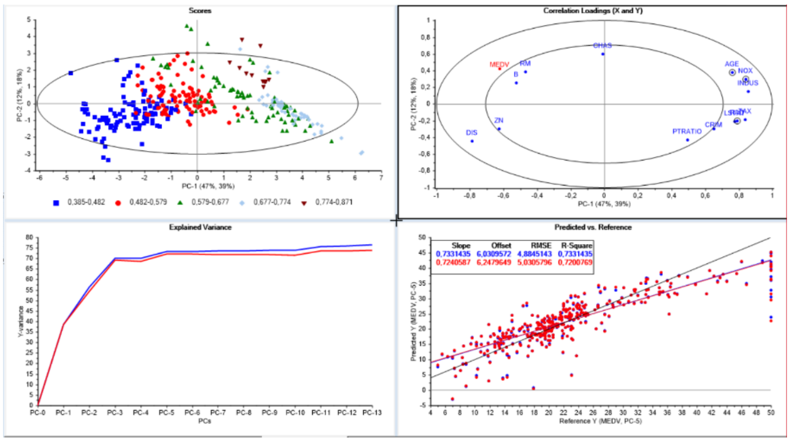
Scores: See that PC1 describes the change in NOX value

Loadings: Grouping for NOX, AGE, TAX, LSTAT,RAD,INDUS – Describe the same thing, have a negative correlation between these variables and the Y variable.

Pred vs. ref: High variation in the predicted values for the same ref y. Really bad at predicting values at the 50 value. Do not hit target line.

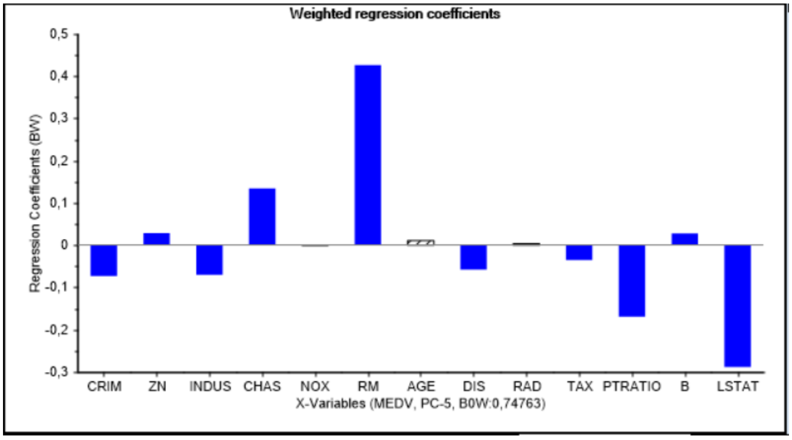
Reg. coeff: See that NOX has a small negative to the Y variable. RM has large pos.link, and LSTAT large neg.stat.

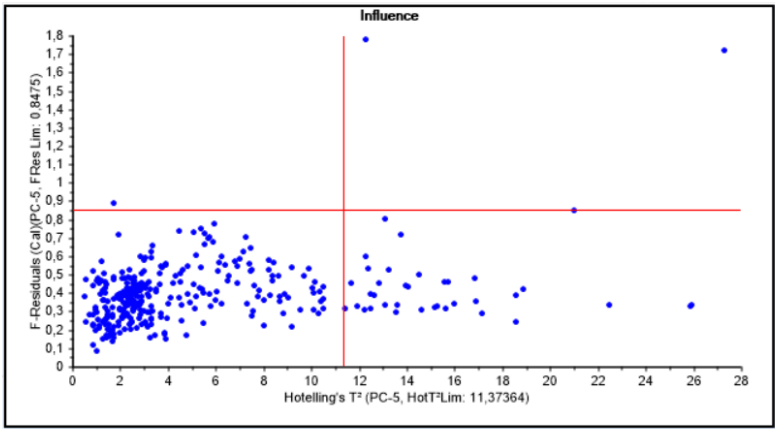
Residuals: See that we have a lot of samples outside the limits of that we estimated. The cause of this could be external factors that gives us samples deviate from the rest.



Slope	Offset	RMSE	R-Square
0,7331435	6,0309572	4,8845143	0,7331435
0,7240587	6,2479649	5,0305796	0,7200769

Weighted regression coefficients



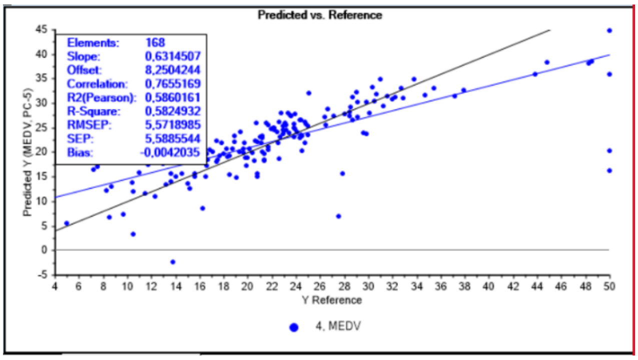


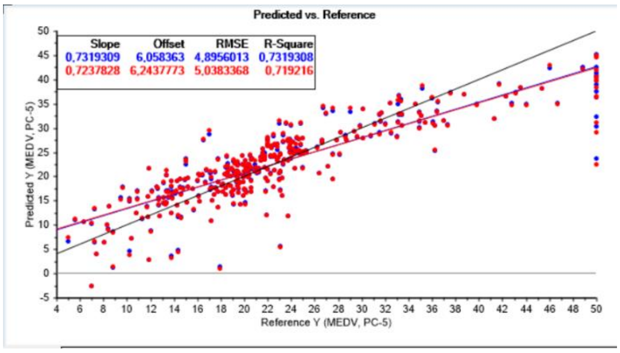
Recalculate without some of the not-so-important variables and compare the models. Mark variables and do Tasks – Recalculate - Without marked - Variables. Compare the results, can you improve the model?

Recalculate without NOX, AGE, RAD: See no difference in the model performance, except for a smoother curve for the explained variance. This is what we want, as we can simplify the model without losing any performance.

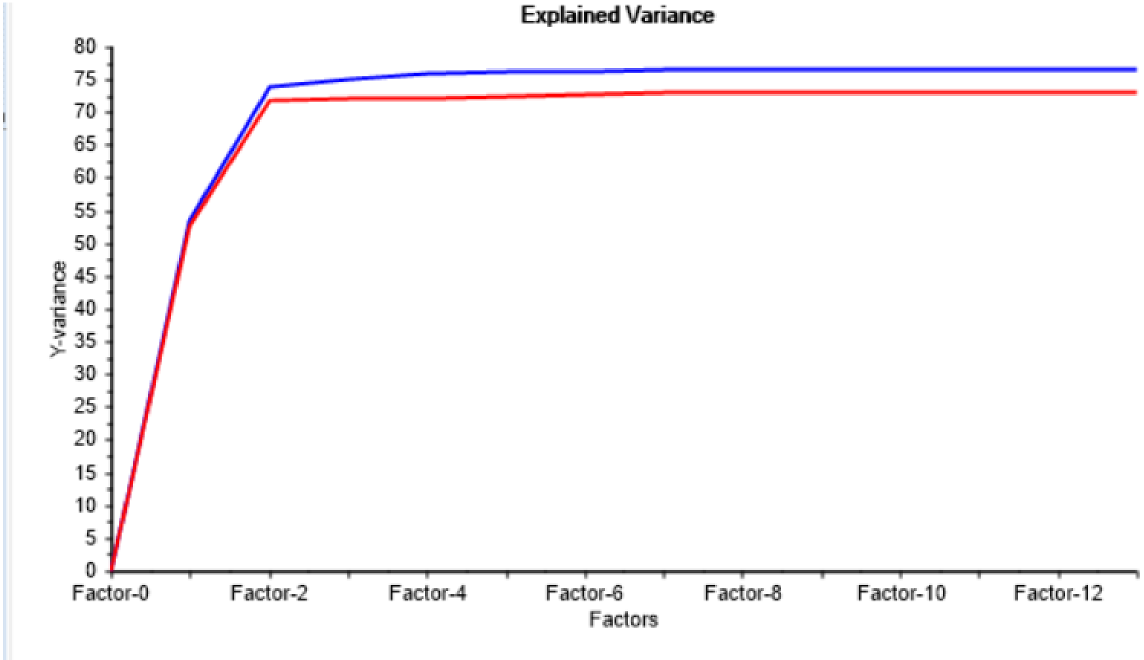
Now predict the test set: Tasks-Predict-Regression. Is the RMSEP similar to the model on the training set?

RMSEP: No, the RMSEP increased from 5 on the model on the training set, to 5,57 on the test set. This coud be because we have already used the training set to find the optimal nr of PCs, and therefore cant also use this set get the lowest error for the test set.



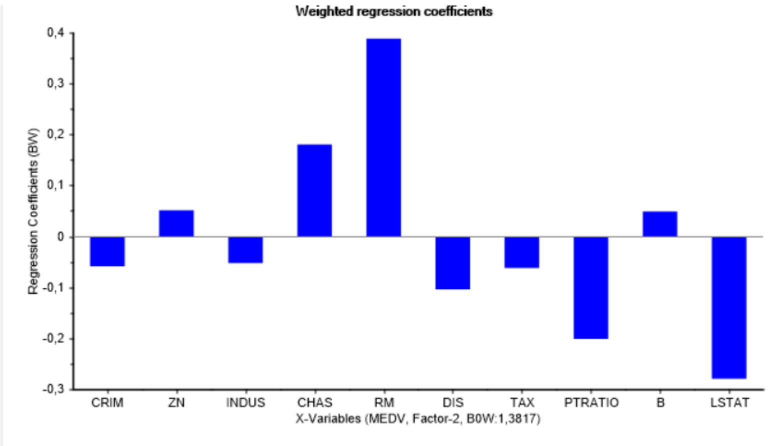


Now do the same with PLS regression. Compare the number of PCs needed for modelling Y. Why is there a difference? Are the final regression coefficients similar?

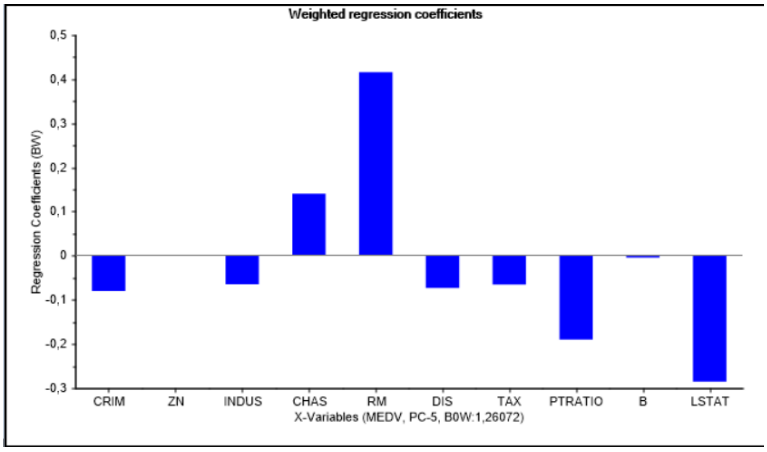


See that the nr of PCs needed are a lot fewer than for PCR, we choose that the optimal nr of PCs is 2. PLS has fewer factors than PCR, as PLS only picks out the needed variables to maximize the correlation between X and Y. While PCR contain factors that doesn't necessarily help correlate X and Y.

Reg. Coeff PLS

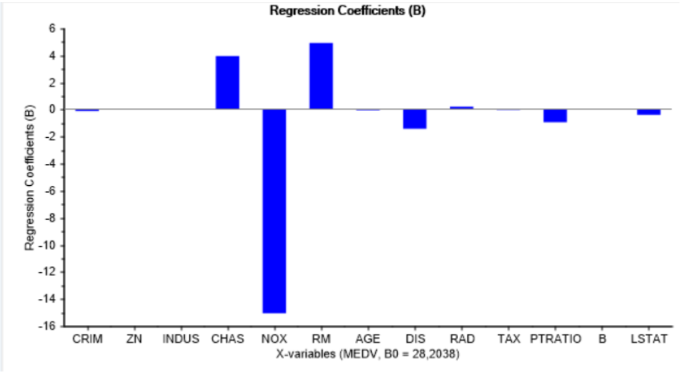


Reg. Coeff PCR



See that the regression coefficients are similar for PCR and PLS, this means that even tho we use different methods for find the correlation between X and Y, we end up finding the same model.

Compare the regression coefficients from MLR with PCR and PLSR. Why are they different?



See that there is a large difference in reg.coeff for PCR/PLS and MLR. This is because we earlier saw that there is a large correlation between several x-variables (NOX, AGE, TAX, LSTAT,RAD,INDUS), meaning we have collinearity among x-variables, which for MLR makes the regression coefficients unreliable.