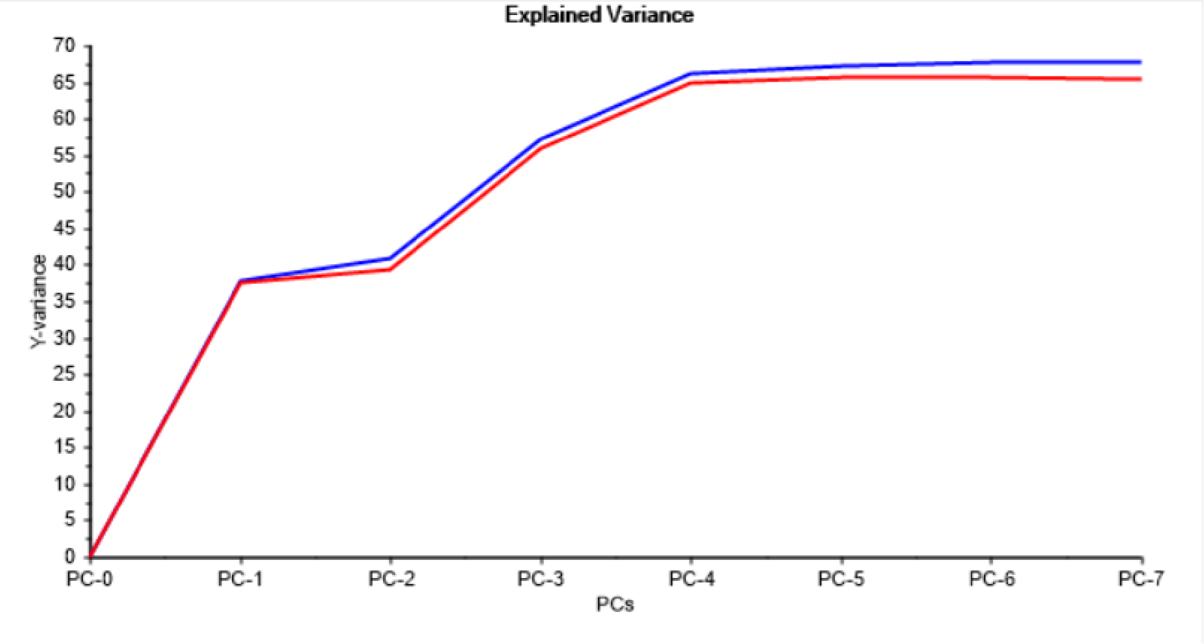
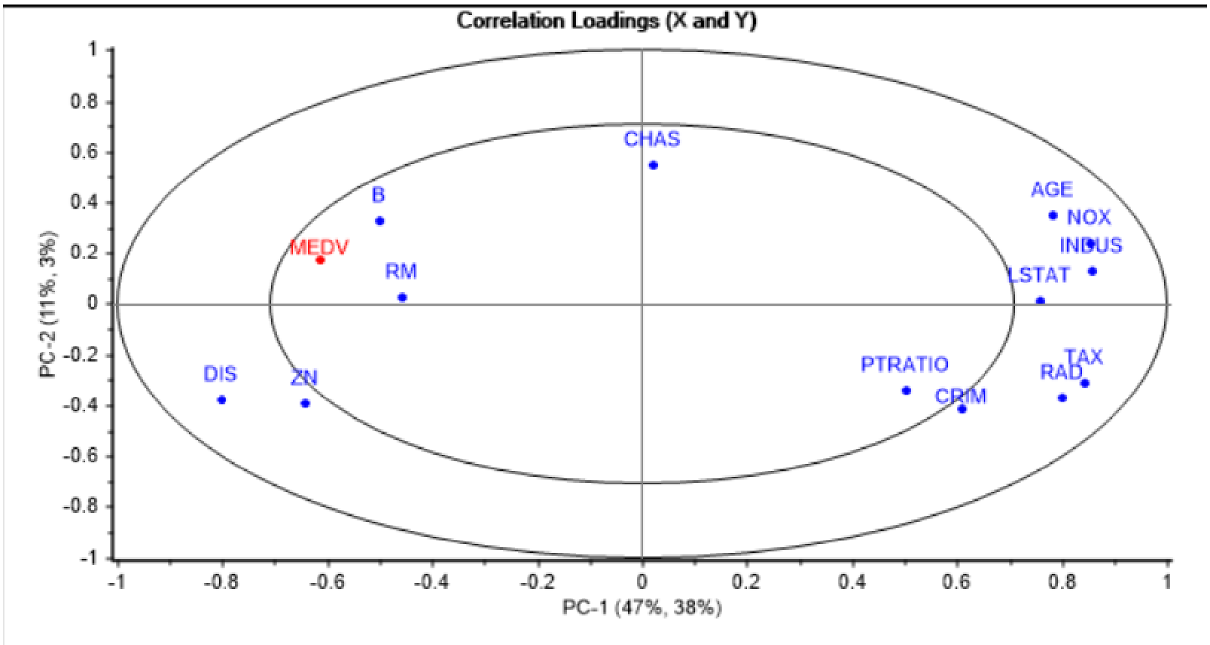
Assignment 7

PCR

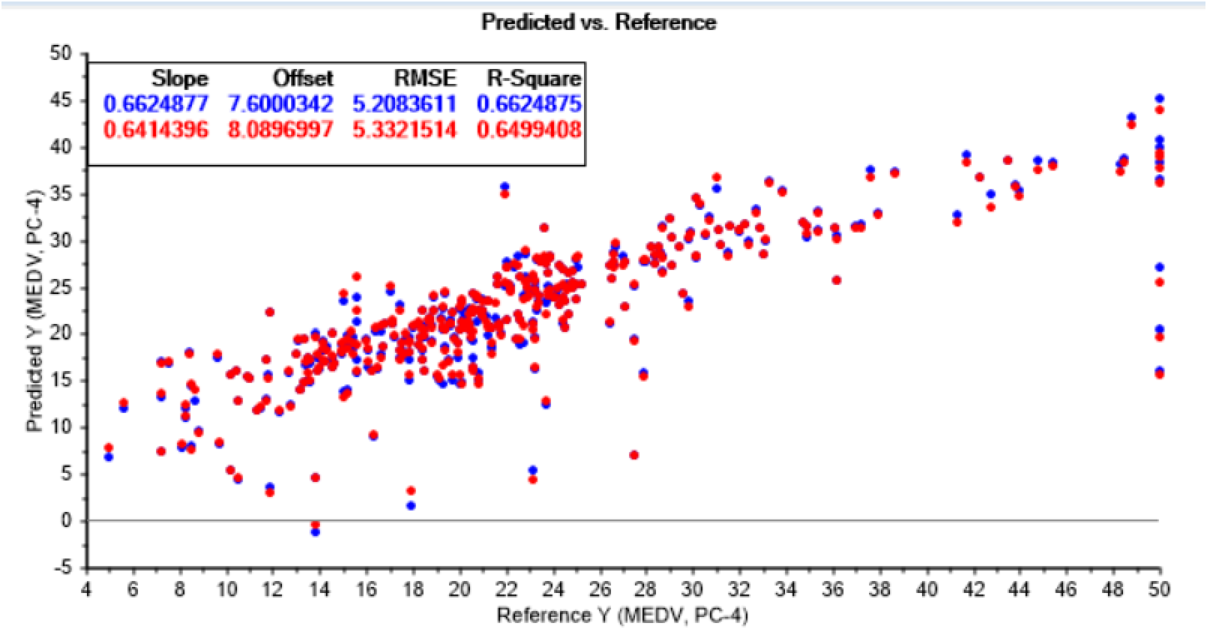
The model is calculated with scaling to unit variance.



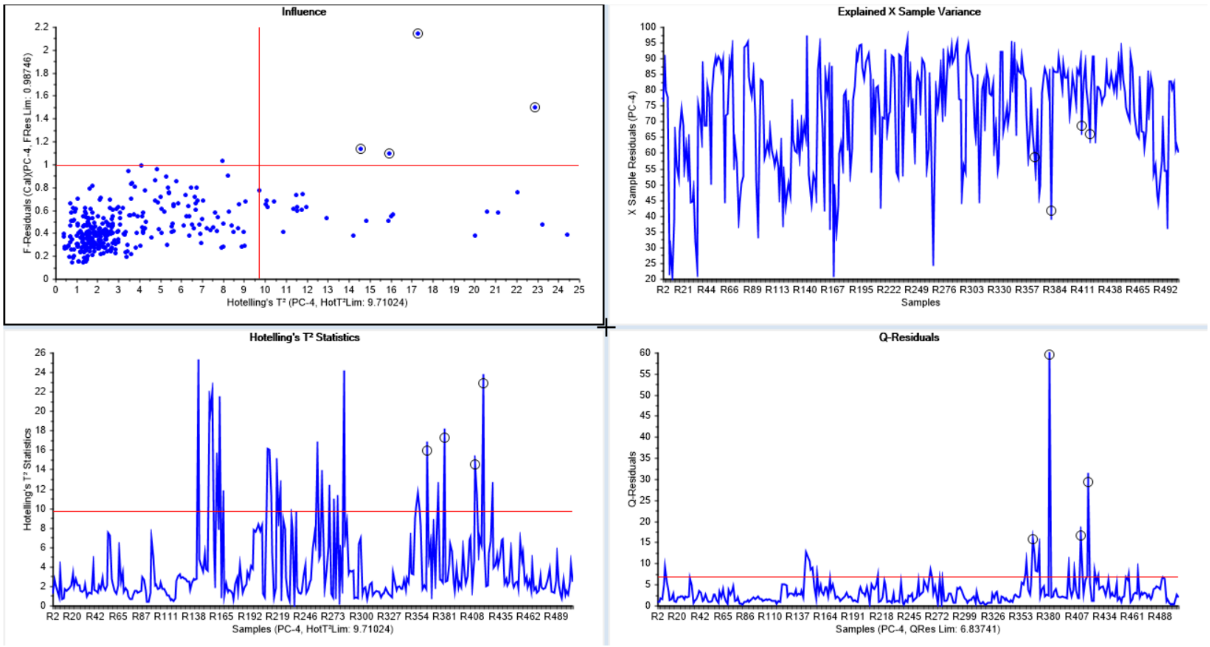
The explained variance plot stops improving after 4 PCs, and this can be a cut-off point. We also see very little improvement from PC-1 to PC-2 meaning that PC-2 might not include much information.



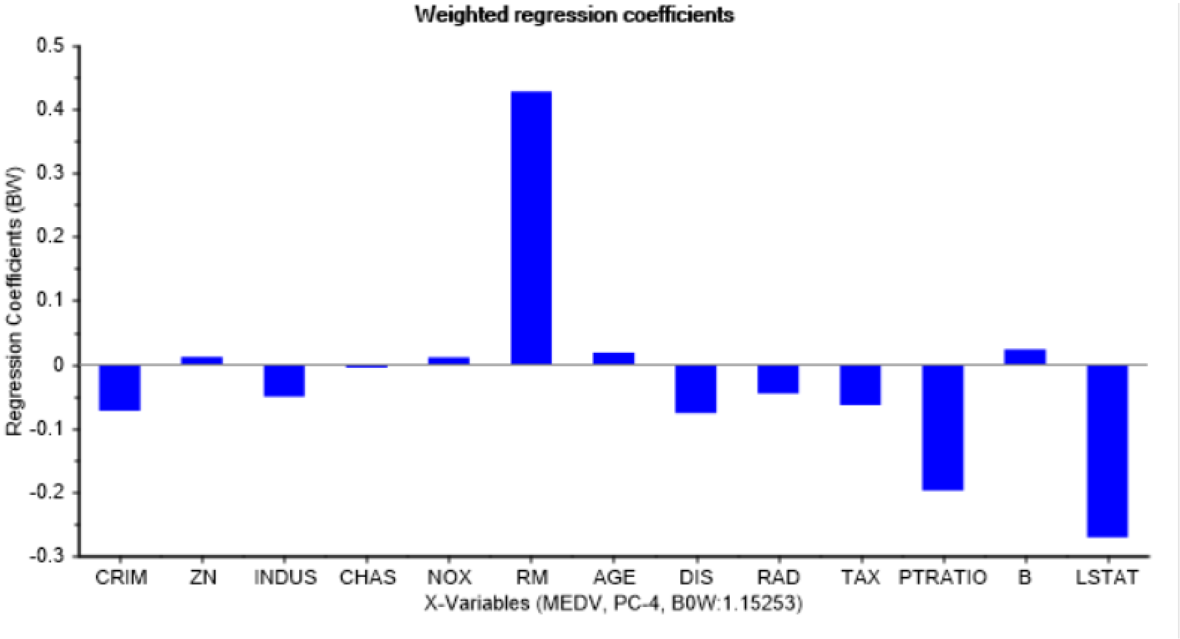
The correlation loadings shows that MEDV is inversely correlated to many variables, and correlated to DIS and ZN. From the description of the variables inversely correlated it can make sense that these are inversely correlated to MEDV. E.g. more old houses (AGE) will reduce the value of the homes.



From this plot we see that the RMSE for training and valiation sets are very similar, indicating that we have a good model using 4 PCs.

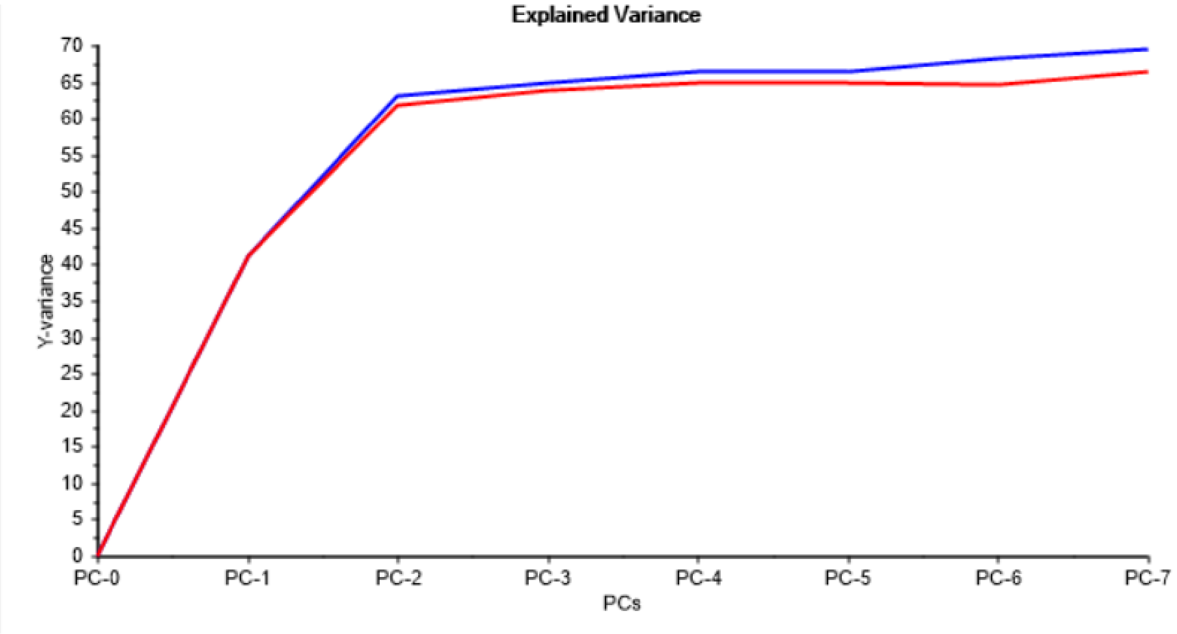


From the influence plot we see not too many outliers in the quadrant with high T2 and high F-res, i.e. the marked variables. But in T2 plot there are more above the critical limit, and the same is seen in the Q-res plot. These might be outliers.

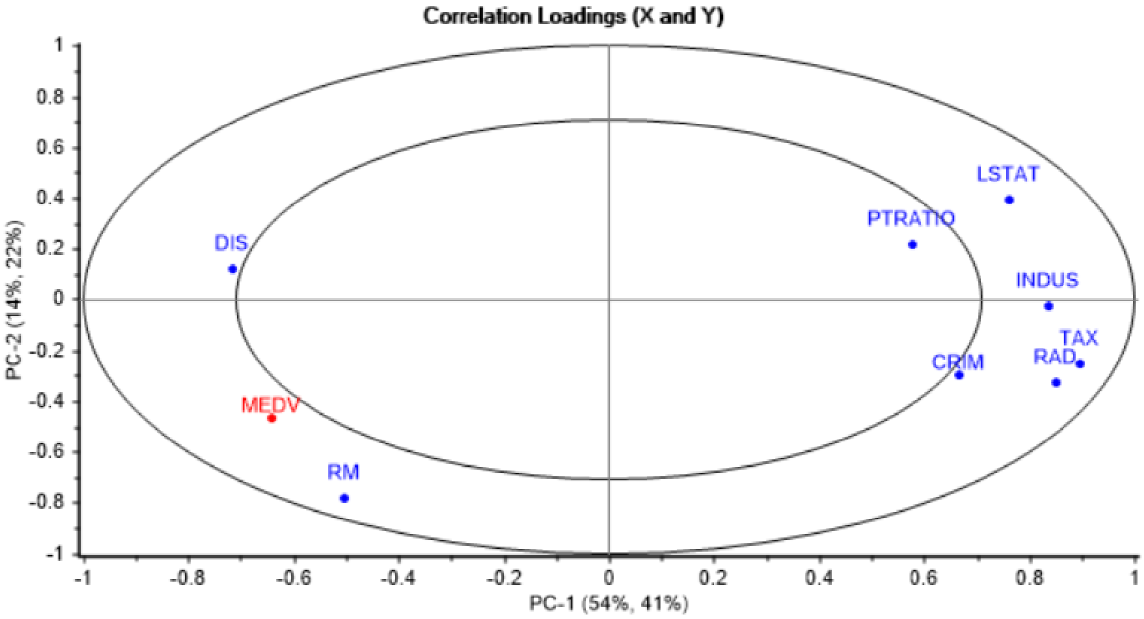


From the regression coefficients we see that it's primarily RM, PTRATIO and LSTAT that impact the MEDV response variable. We also see that ZN, CHAS, NOX, AGE and B has very little impact on MEDV. We can recalculate without these.

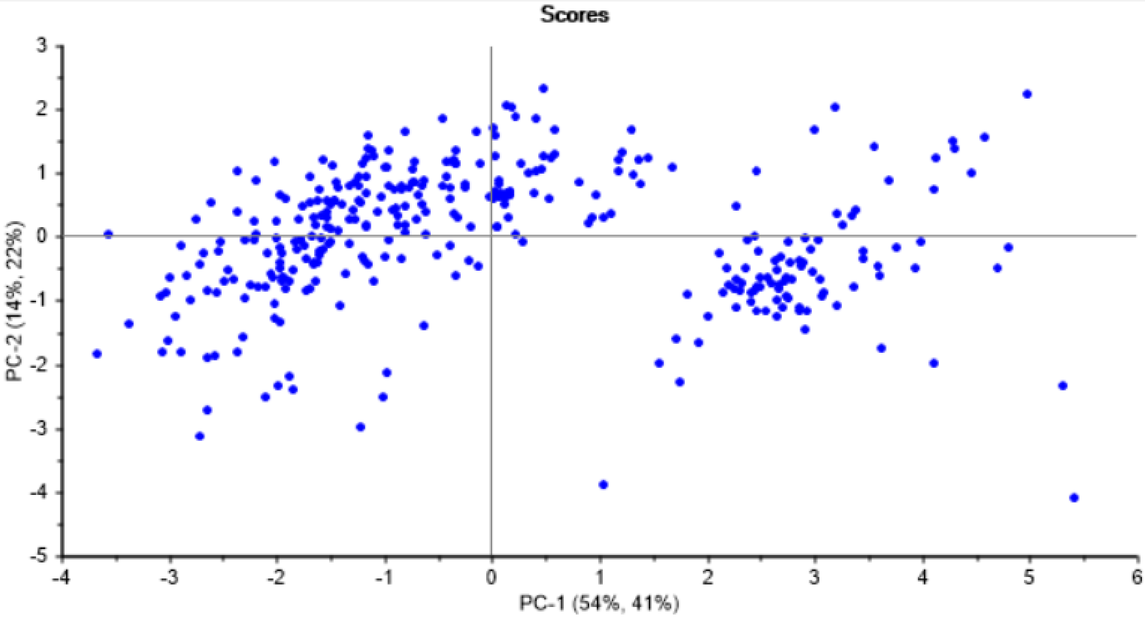
Without marked



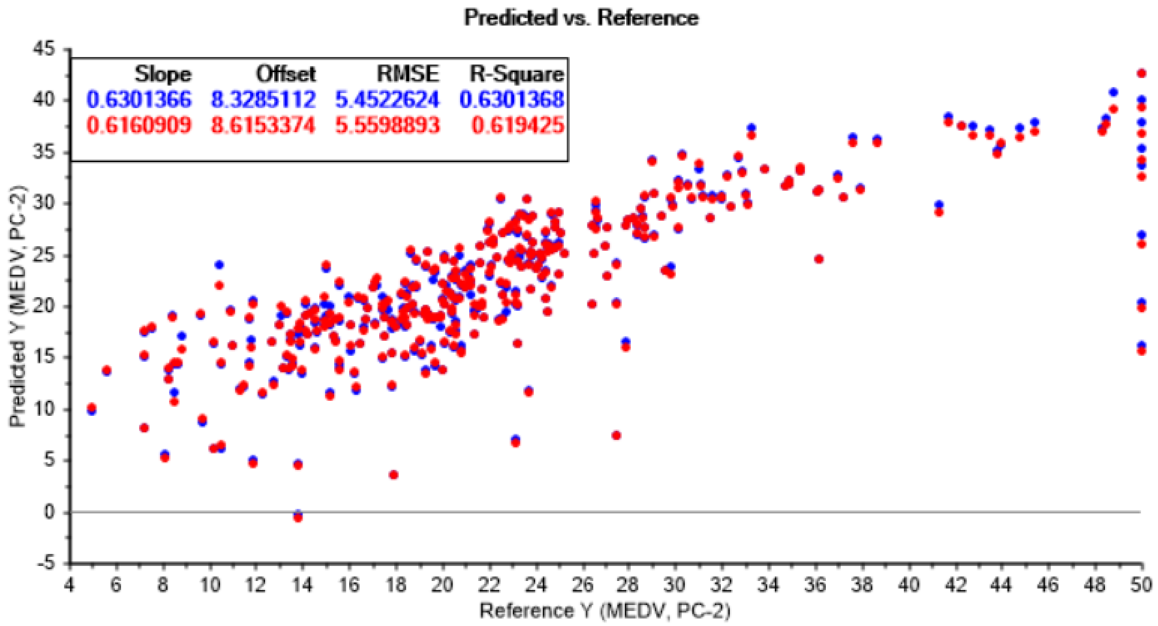
Now we see that without the marked we can maybe get by just using two PCs. The later PCs has marginal increase in variance.



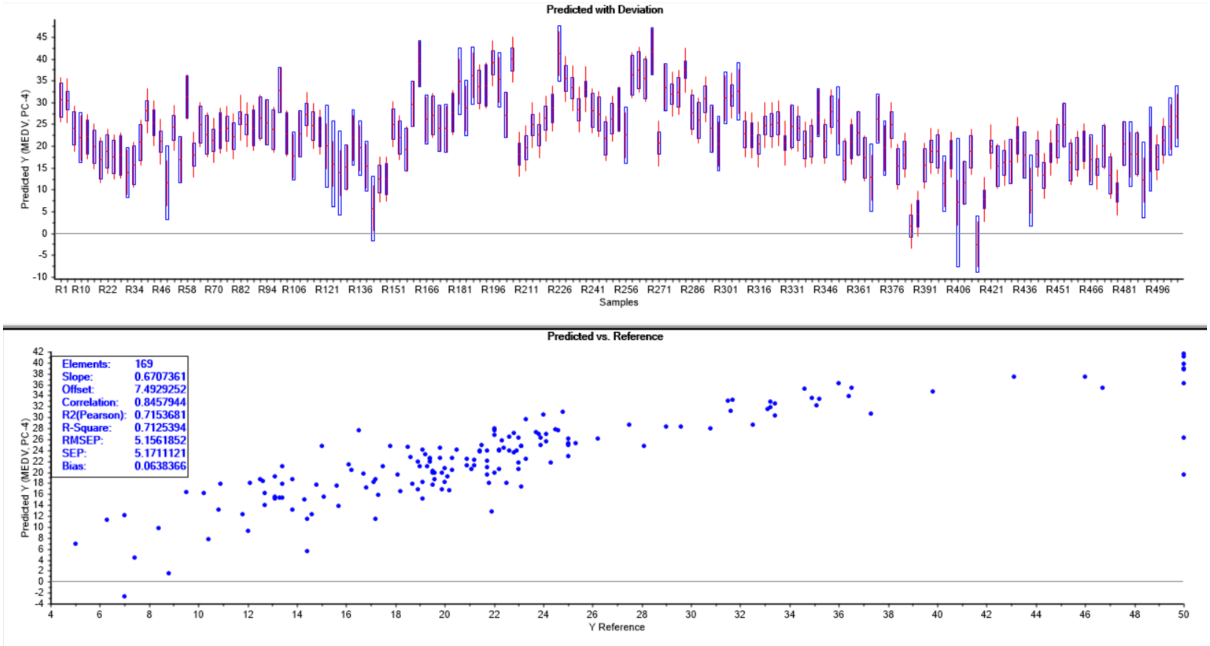
We still see the same relationships between variables in the correlation loadings as we saw before.



There's now a more clear grouping in the score plot than before. Matching it with the loadings there seem to be one group with high on DIS and RM, and another group with high on the other variables.

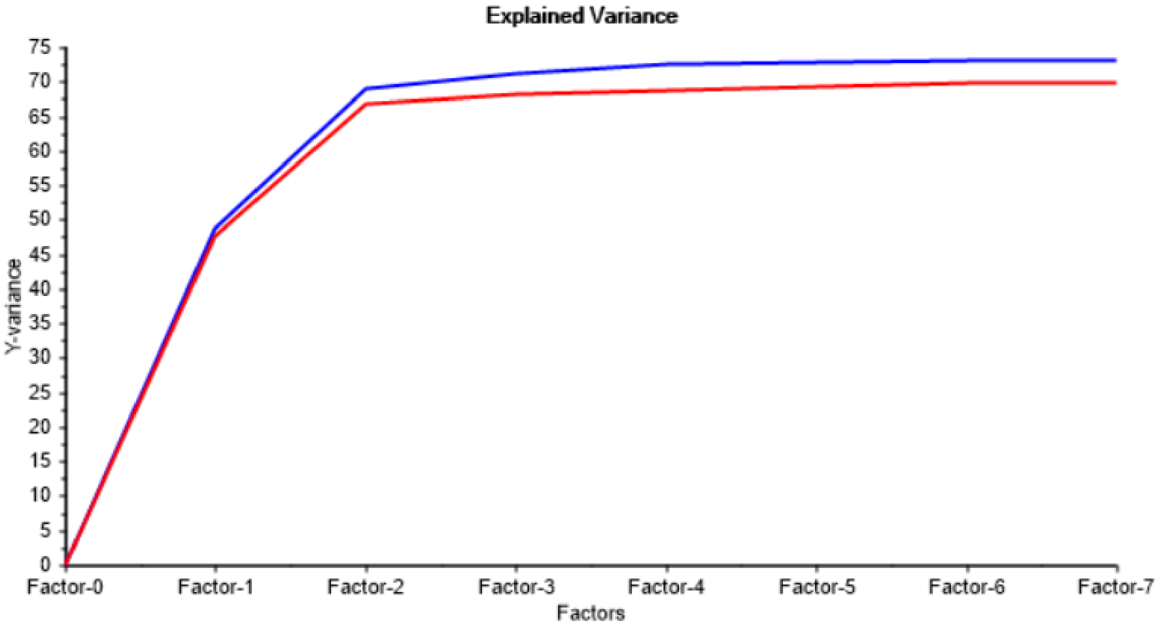


We also have good RMS using two PCs, and this is comparable to earlier.

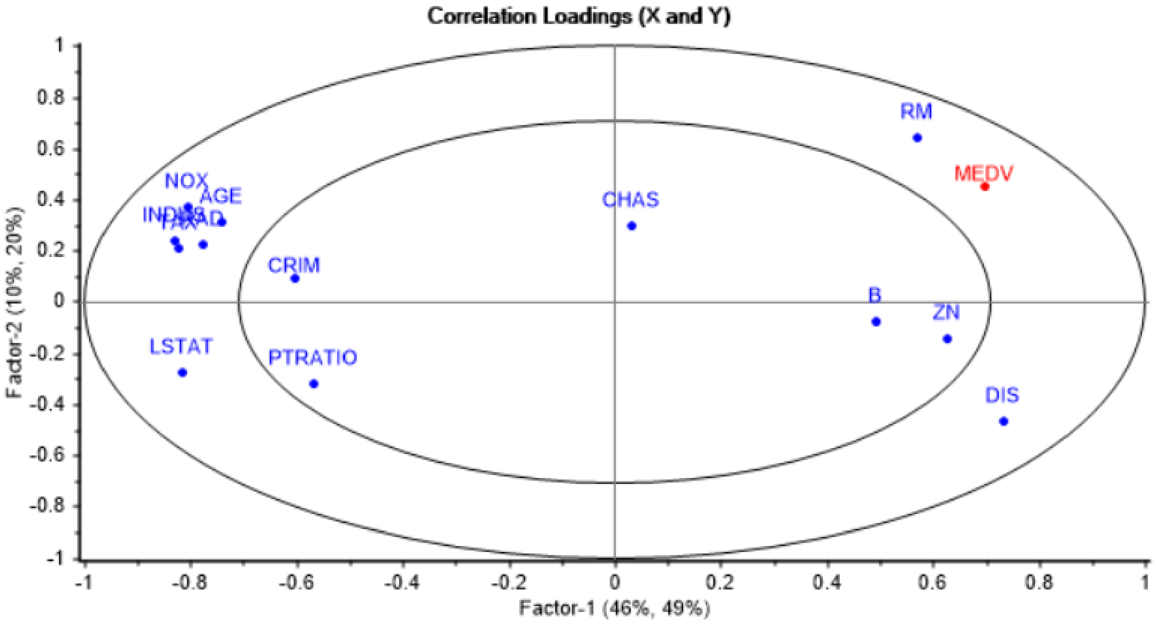


The test set seem to perform well on the model using 4 PCs. There are some samples that are more wrongly predicted, but overall it is within reasonable limits. The RMSE is also close to what we saw earlier so it seems like we have a good model.

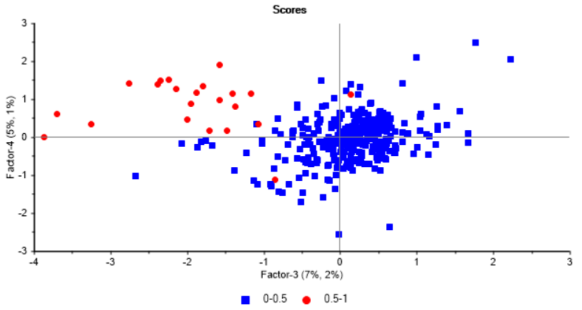
PLS

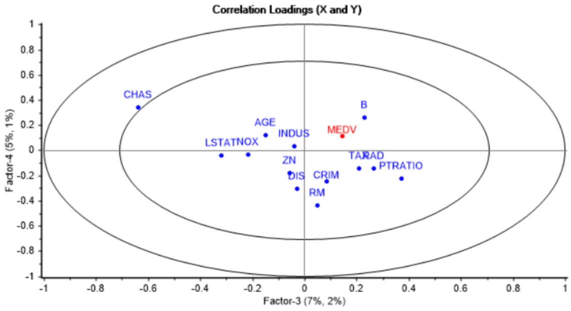


The explained variance plot indicates that most of the variance in the data is captured by the two first factors, but also that there are some information in factor 3 and 4 as there is a very slight increase there.



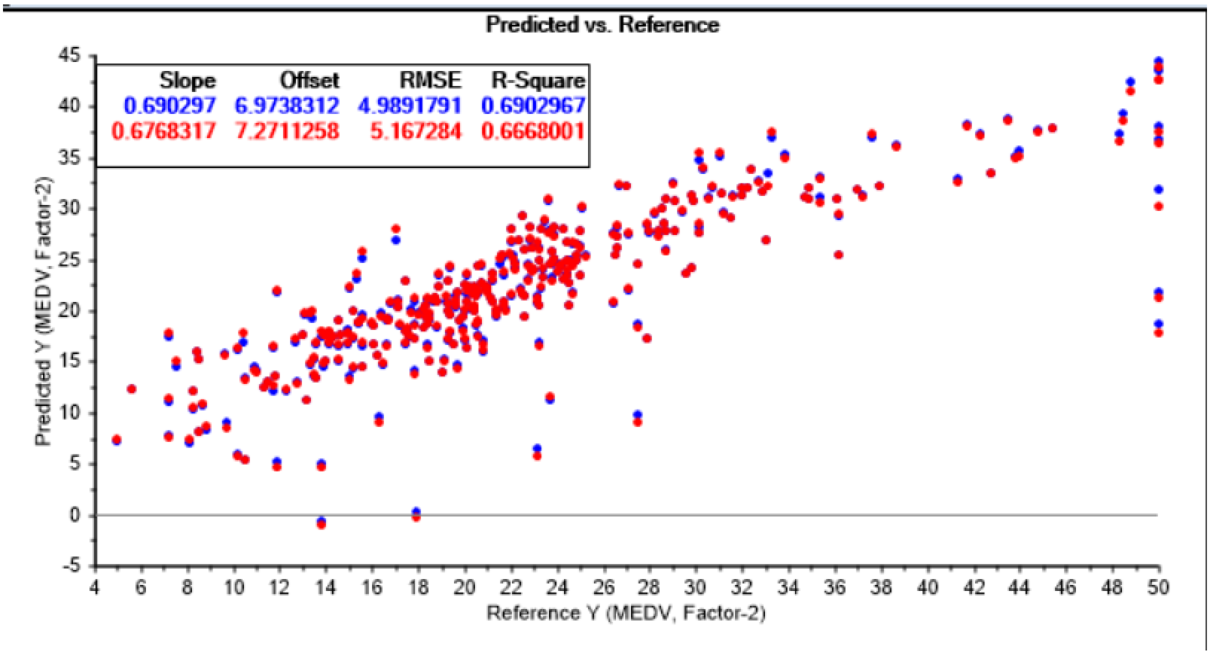
We also see very similar correlation loadings as for PCR, but with the first factor/PC reversed. However, the relationships between the variables are the same, and the interpretation is also the same.



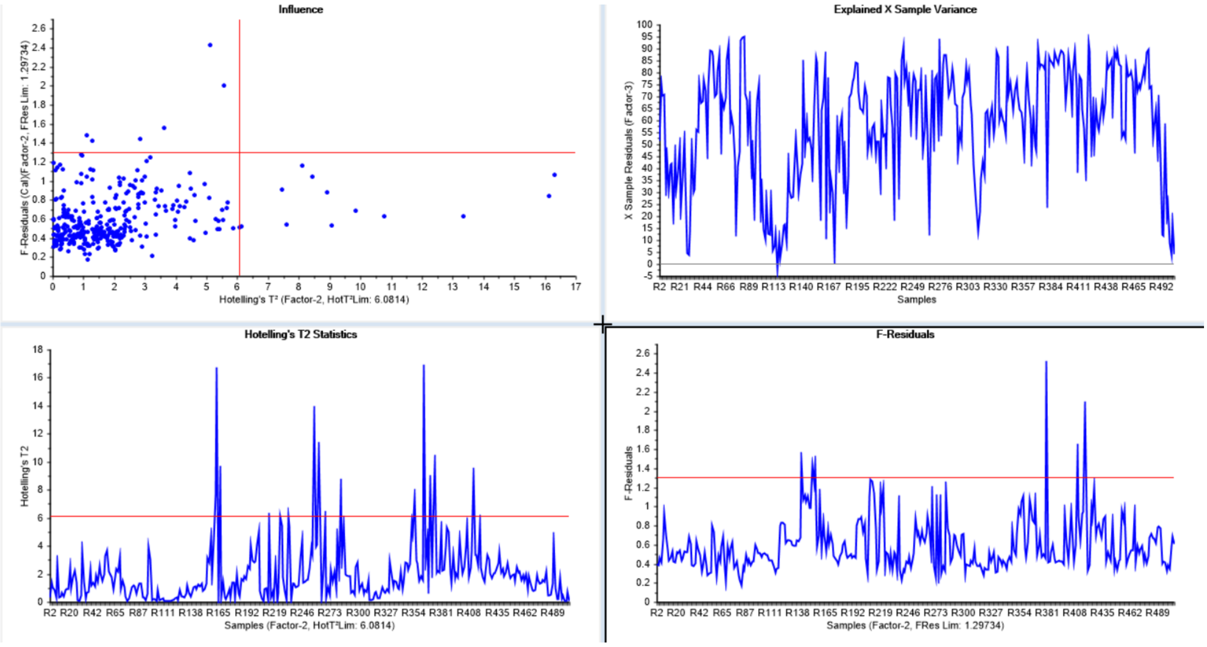


From correlation loadings for factor 3 and 4 it can seem that CHAS is slightly relevant. In the scores there can seem to be some clustering/grouping, but I wouldn’t say it is very pronounced.

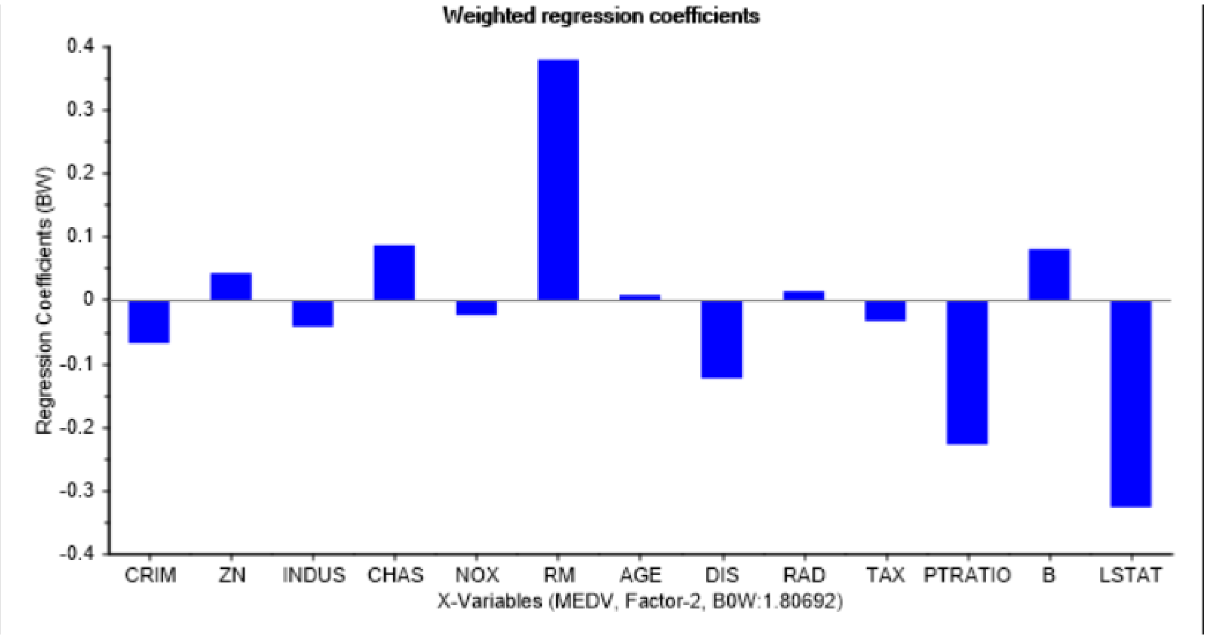
So, the optimal number of PCs are likely 2 or 3 in this case.



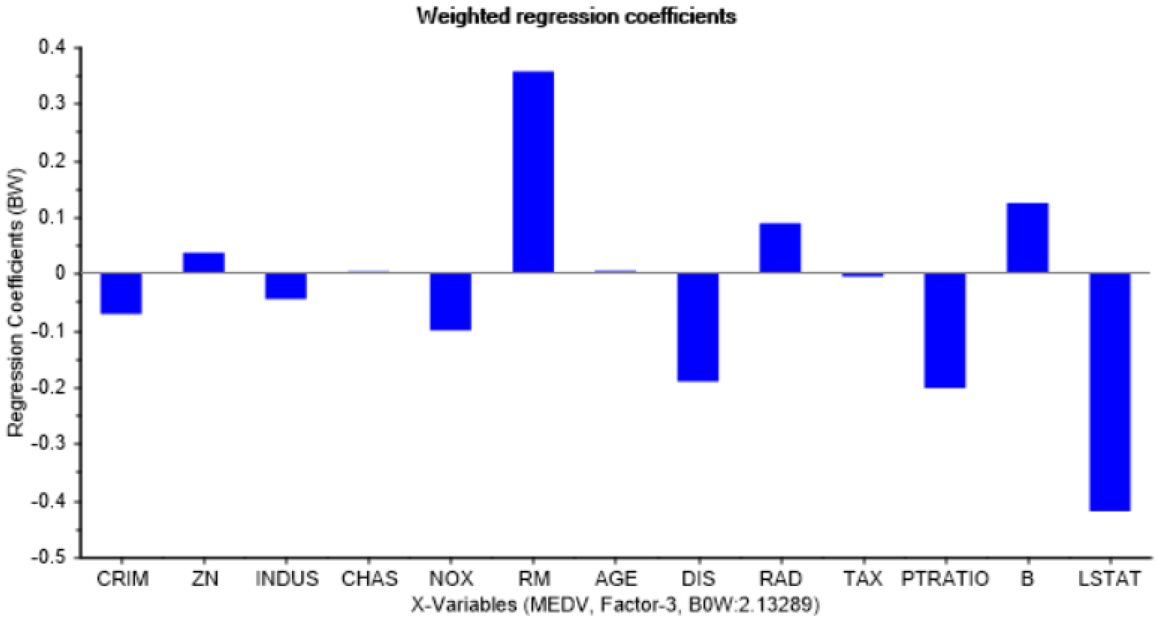
The RMSE is close between calibration and validation, and is similar to what we got for PCR, but a little smaller.



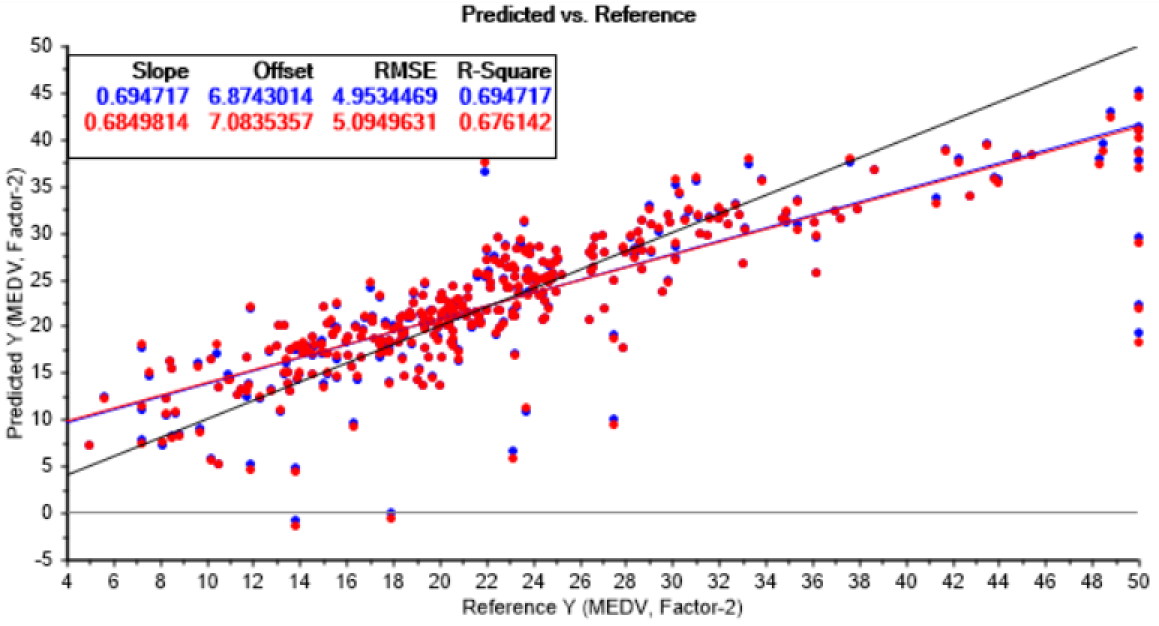
There does not seem to be any outliers with large T2 and F-res, but compared to PCR there are some more samples with larger F-residual.



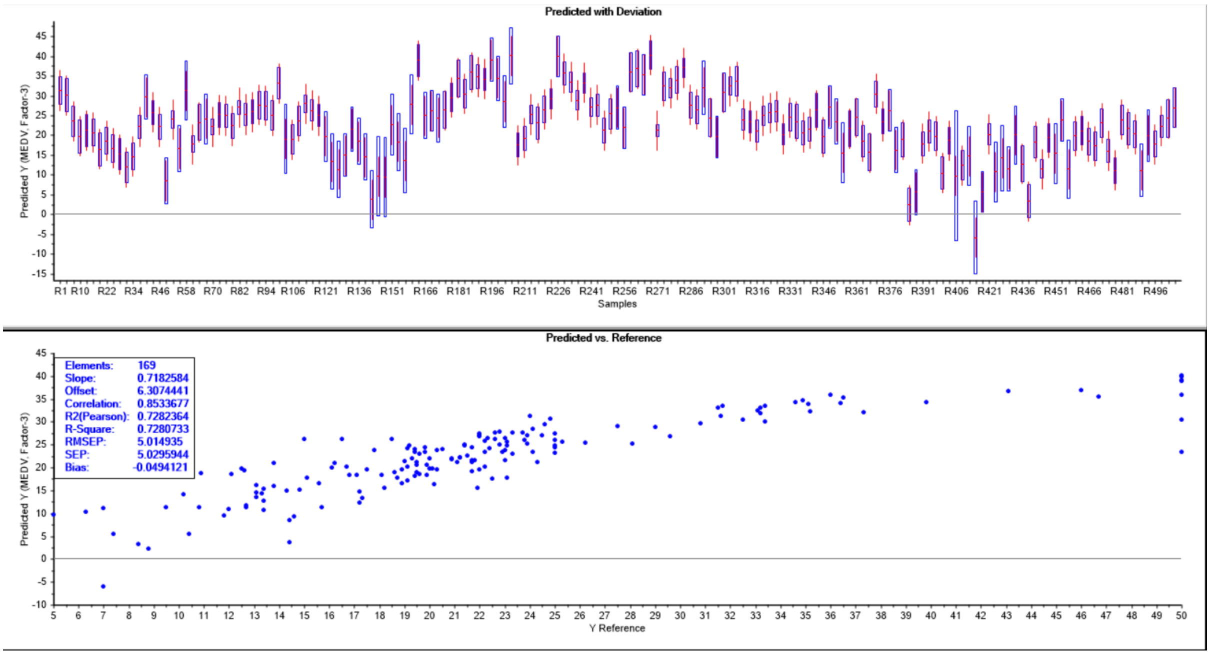
The regression coefficients using two factors show much of the same story as for PCR, but gives a little more weight to some variables compared to PCR (e.g. CHAS is noticeably larger).



An interesting note is that including factor 3 makes CHAS coefficient more or less zero, while it puts more weight to RAD. We can now try to remove CHAS, AGE and TAX and recalculate.



We see the RMSE is slightly improved using two factors. But we still see that the regression line is not too close to the optimal, so it will overestimate low values and underestimate high values.

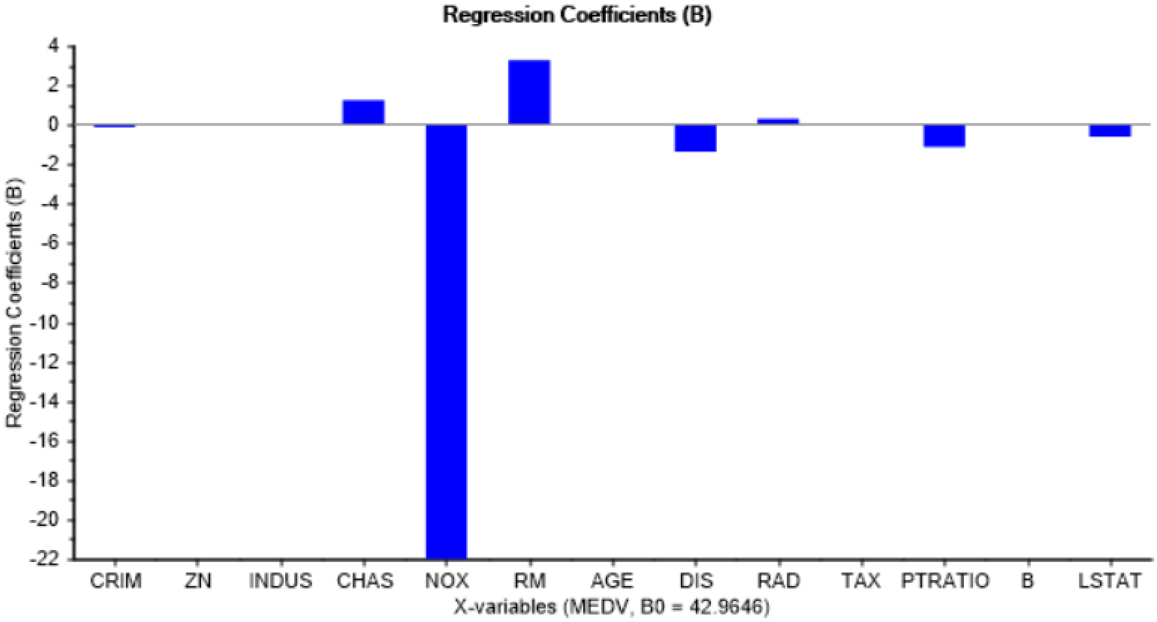


```
Elements:        169
Slope:           0.7182584
Offset:          6.3074441
Correlation:     0.8533677
R2(Pearson):     0.7282364
R-Square:        0.7280733
RMSEP:           5.014935
SEP:             5.0295944
Bias:           -0.0494121
```

Regressing the test set using the PLSR model with some variables removed gave almost identical results to PCR in this case.

Comparing coefficients: MLR, PCR, PLSR

The coefficients for MLR looks like this



As we see, these are *very* different from what the PCR and PLSR models gave us. MLR assumes orthogonal variables which is not the case here, and PCR and PLSR fixes this by exploiting the latent structure in the data to create orthogonal variables.