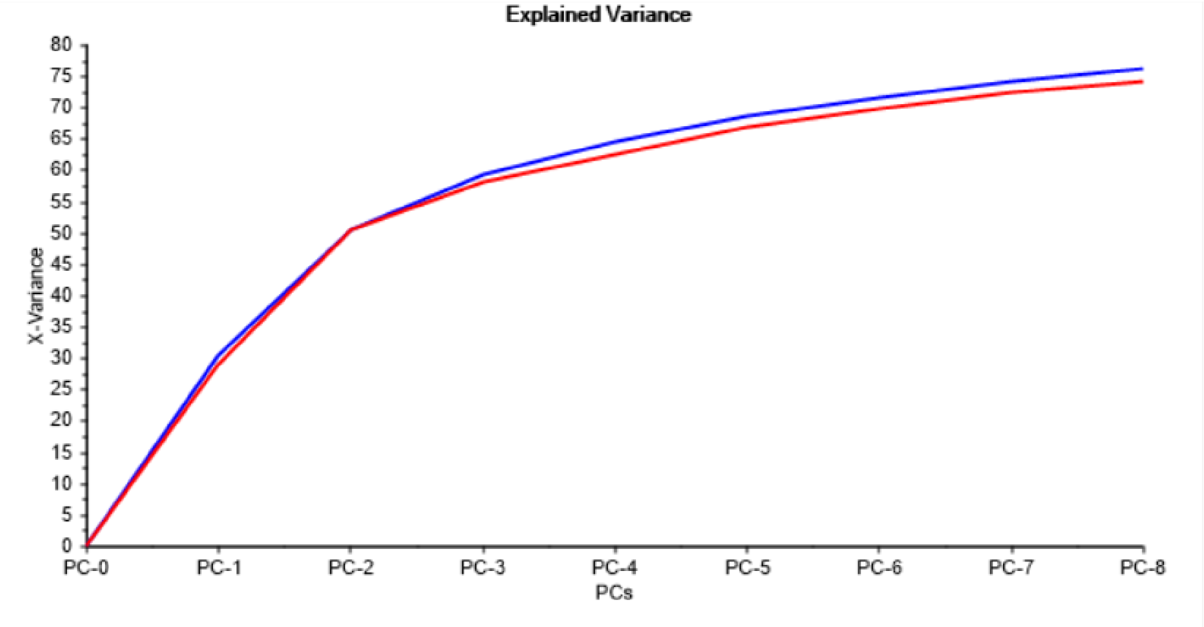
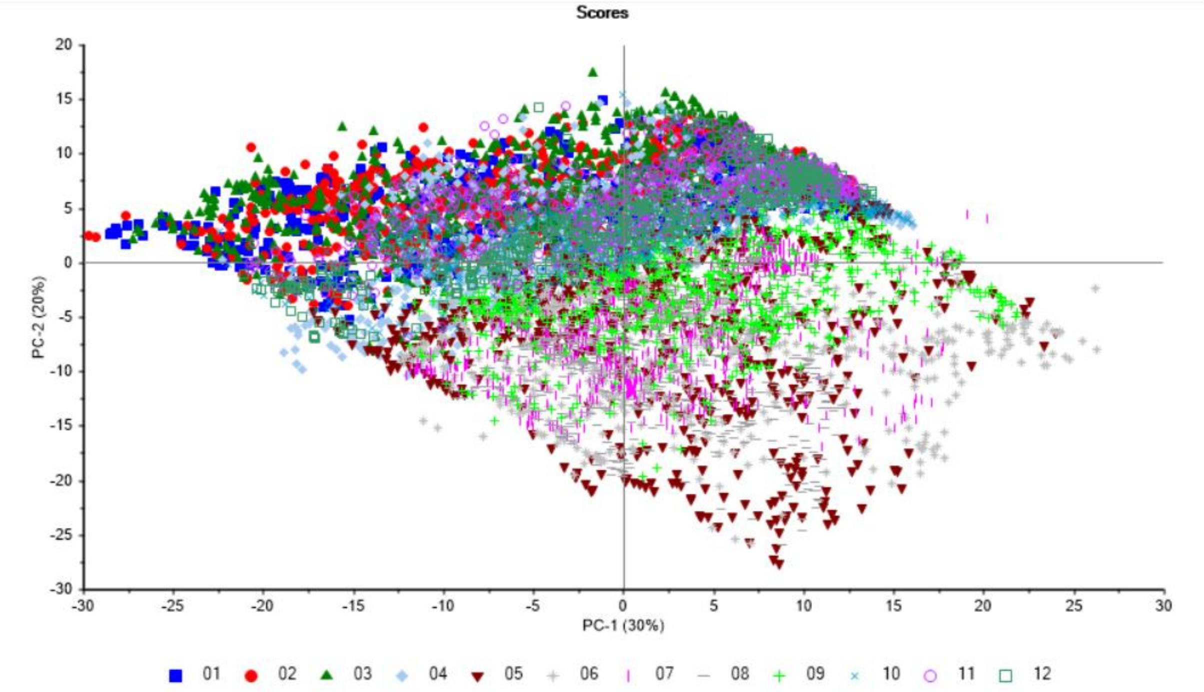
Assignment 8

PCA

Scaled to unit variance. Systematic CV across months. NIPALS algorithm.

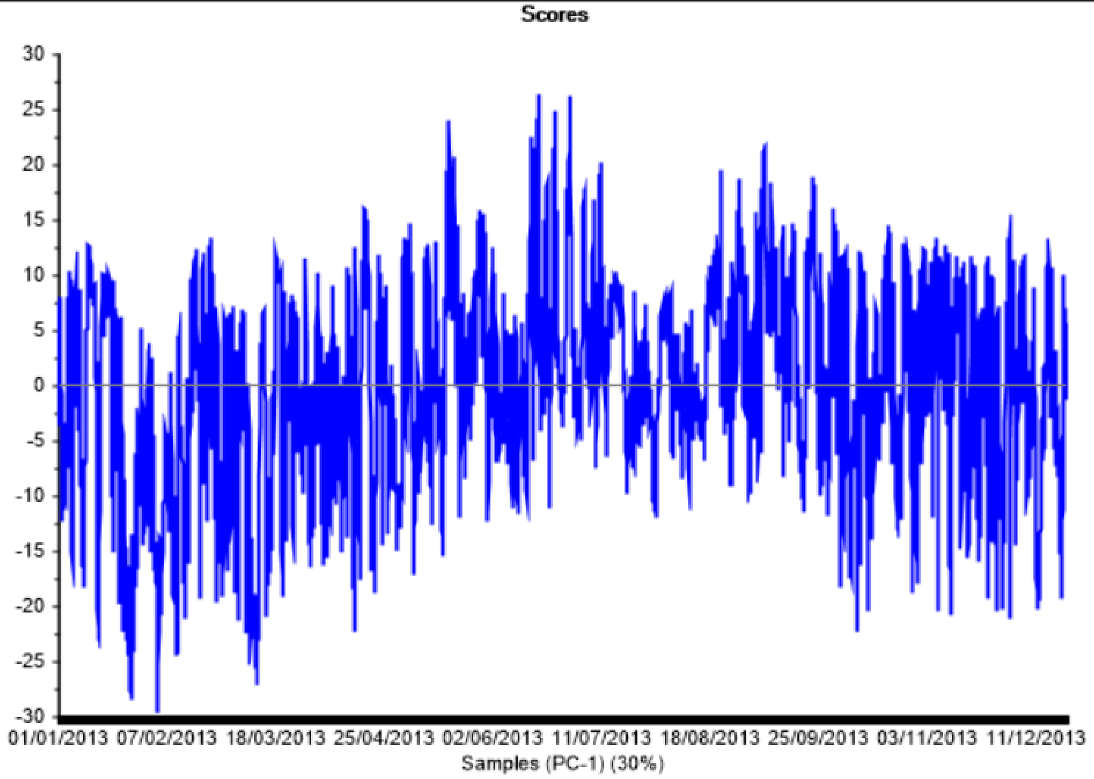


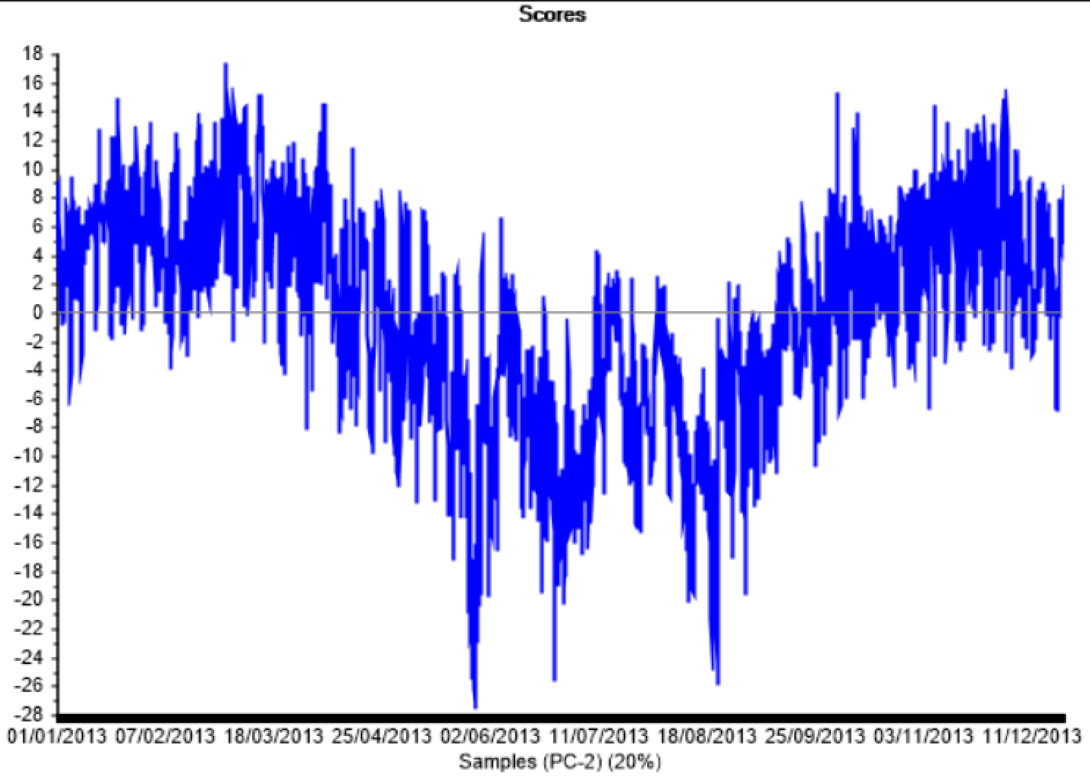
We see the validation follows the training good when cross validating across months which can indicate that there is little variation from month to month (i.e. making a model on one month will predict the next month well).



It seems like 05 May is more spread out in both PC-1 and PC-2 than others. It also seems like summer months are located down to the right while winter months are located in the center and up to the left.

Looking at the other combinations of PCs did not reveal any other structure.



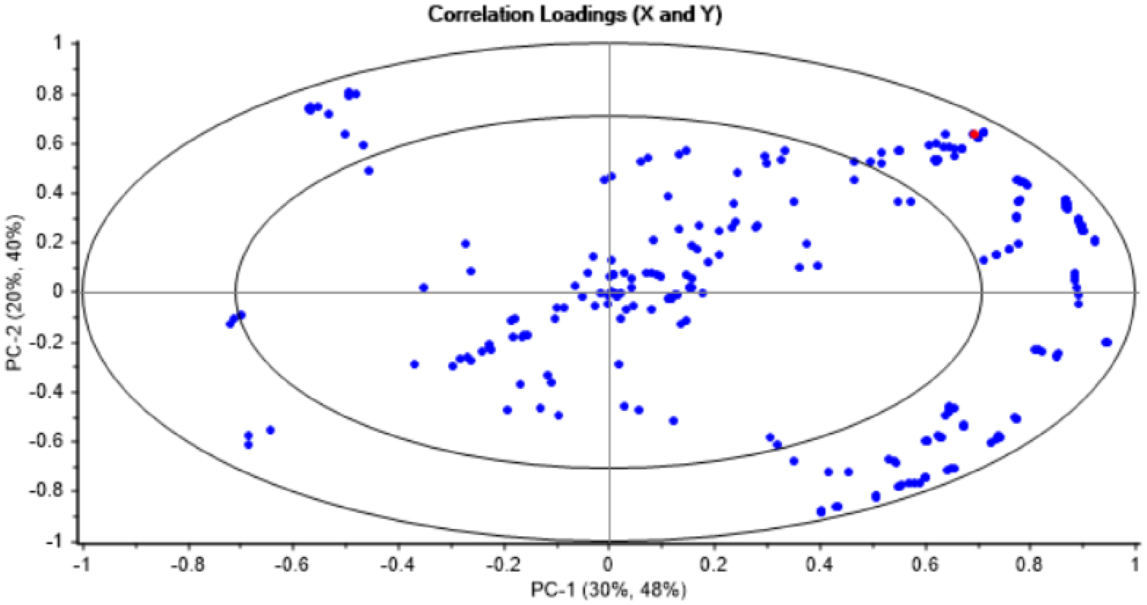


Looking at the scores for PC-1 and PC-2 as line plots, there can be some structure in PC-1 with the summer months a little bit higher, but there isn't much structure except some oscillations between days. In PC-2 we see that the summer months are smaller than the winter months, which are in line with what we saw in the scatter plot.

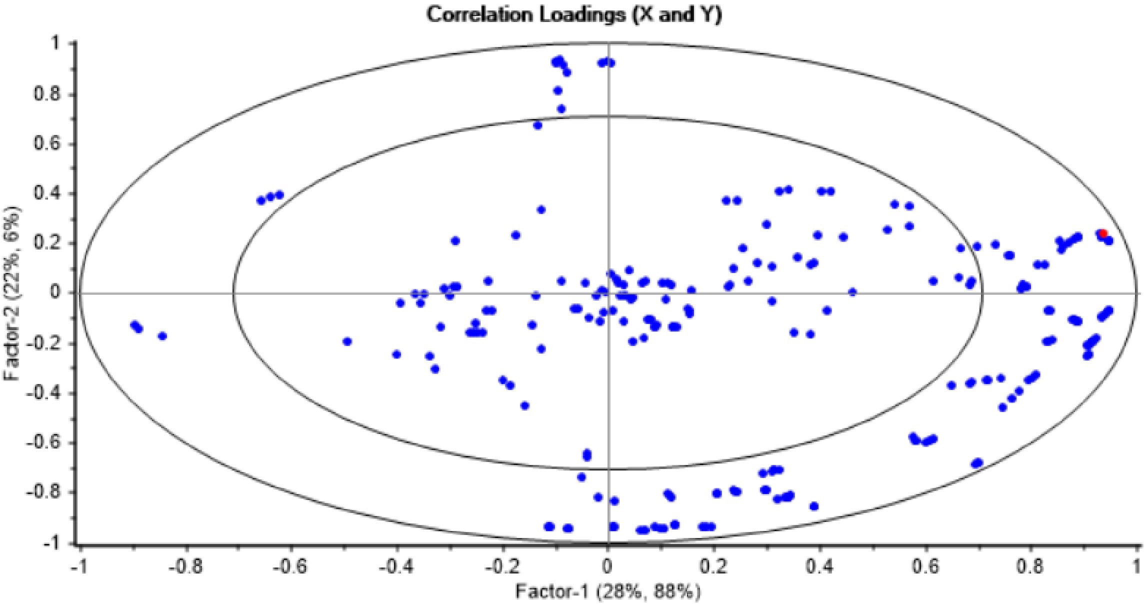
PCR and PLSR

Again, scaled to unit variance, CV'd across months, and used NIPALS for PCR and Kernel PLS for PLSR.

PCR:



PLSR:



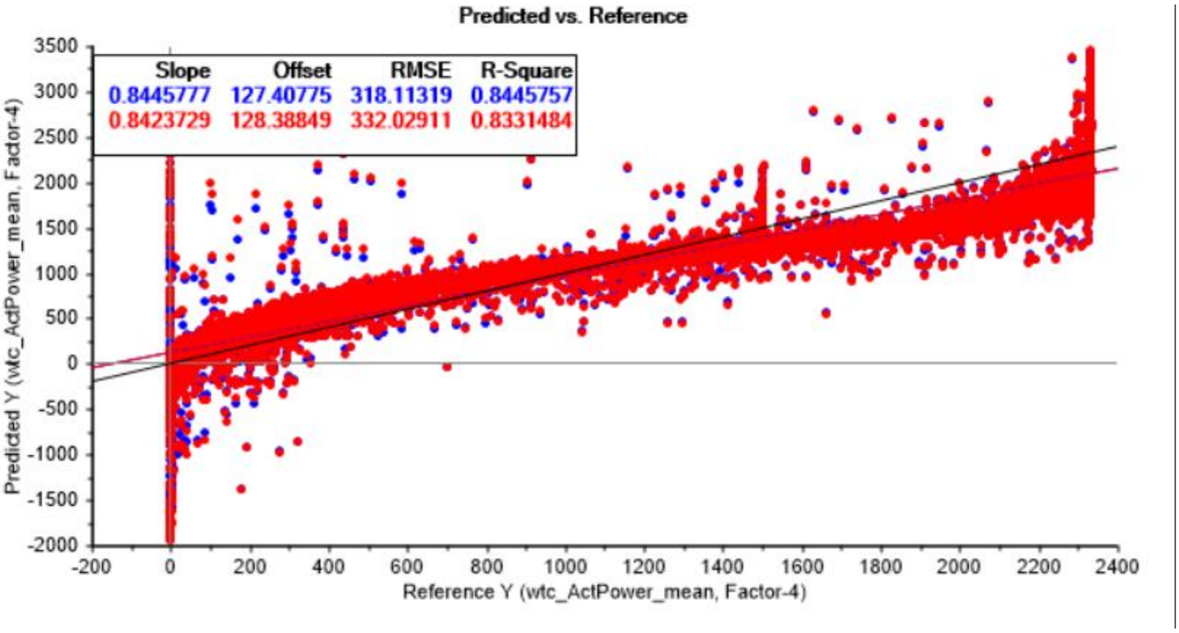
The two correlation loadings look similar, but the PLSR one appear slightly rotated from the PCR one. This can be because PLS includes correlation to the Y data. It is also striking that factor 1 explains more than PC-1 (88% vs 48%). In fact, PLSR model explains more of the variance with fewer factors than PCR. Also, it seems like since PLSR explains more of the variance with factor 1, the correlation loadings are more "compressed" along the factor 2 axis.

Different subsets

Turbine

PCR: close RMSEs, 427 on validation
PLSR: larger difference in RMSEs, but smaller on validation with 332

In both cases, the plot for predicted vs reference looks more like a cubic than linear:



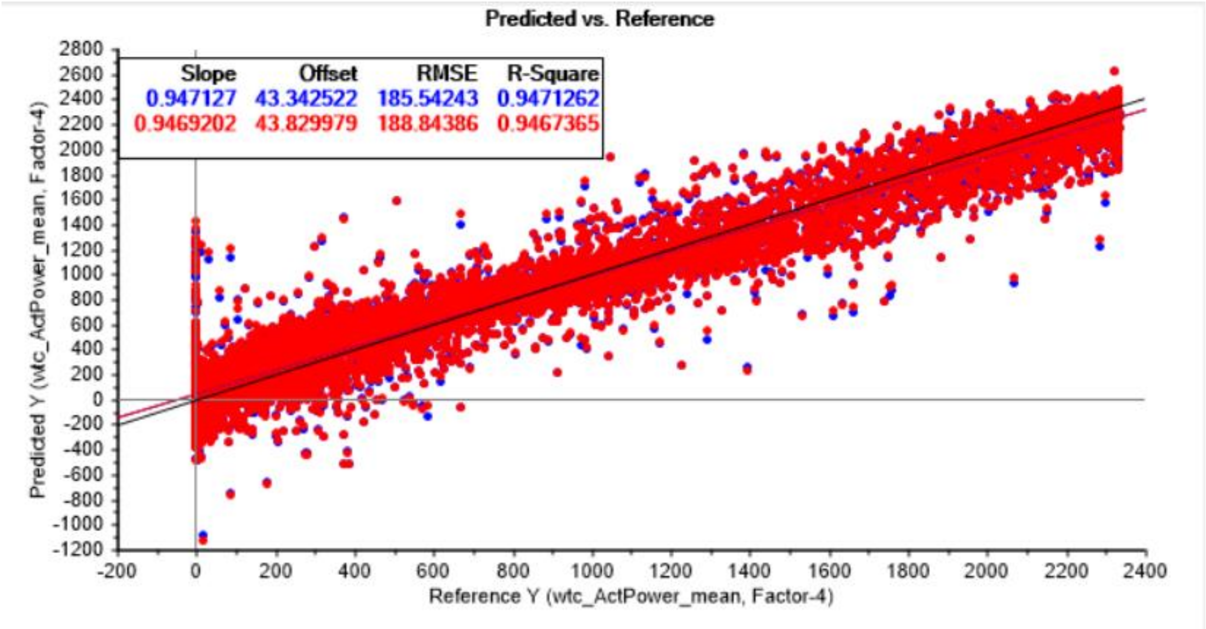
Slope	Offset	RMSE	R-Square
0.8445777	127.40775	318.11319	0.8445757
0.8423729	128.38849	332.02911	0.8331484

Overestimates lower values, underestimates higher values. Also notice that there are very many samples at exactly 0 reference.

Temp

PCR: again close in RMSE, 228 for validation
PLSR: smaller difference than PCR and earlier, 188 for validation

In both cases, the predicted vs reference plot looks much more linear

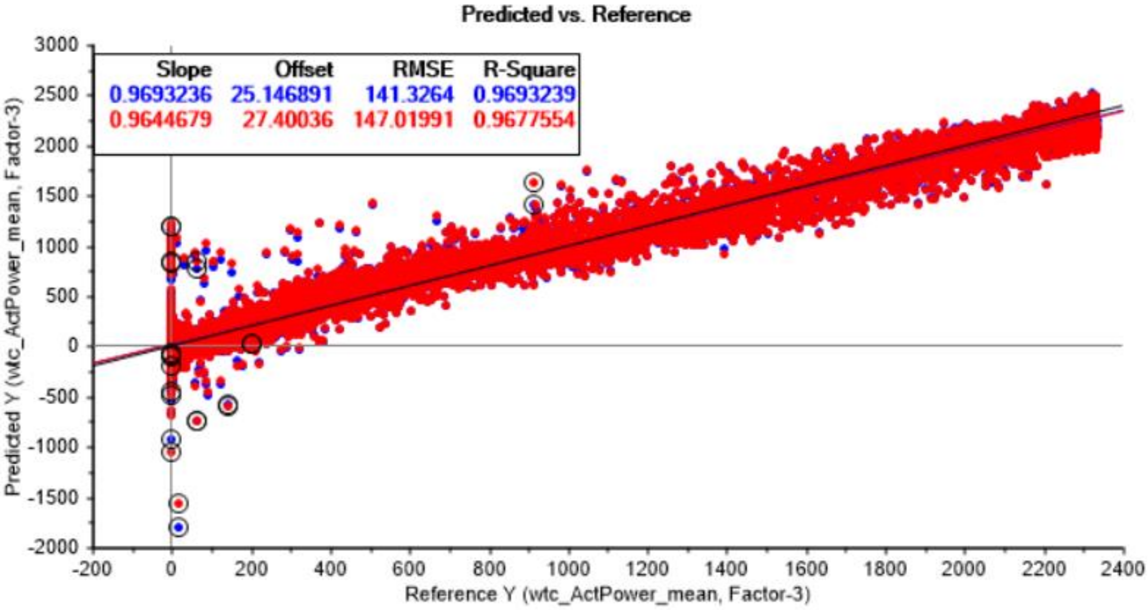


And it seems to more correctly predict both large and small values.

AllX

PCR: RMSEs close with validation at 208
PLSR: a bit larger delta between RMSEs, but close enough, and 147 for validation

In both cases the predicted vs reference plot is quite linear, but slightly more so for PSLSR.



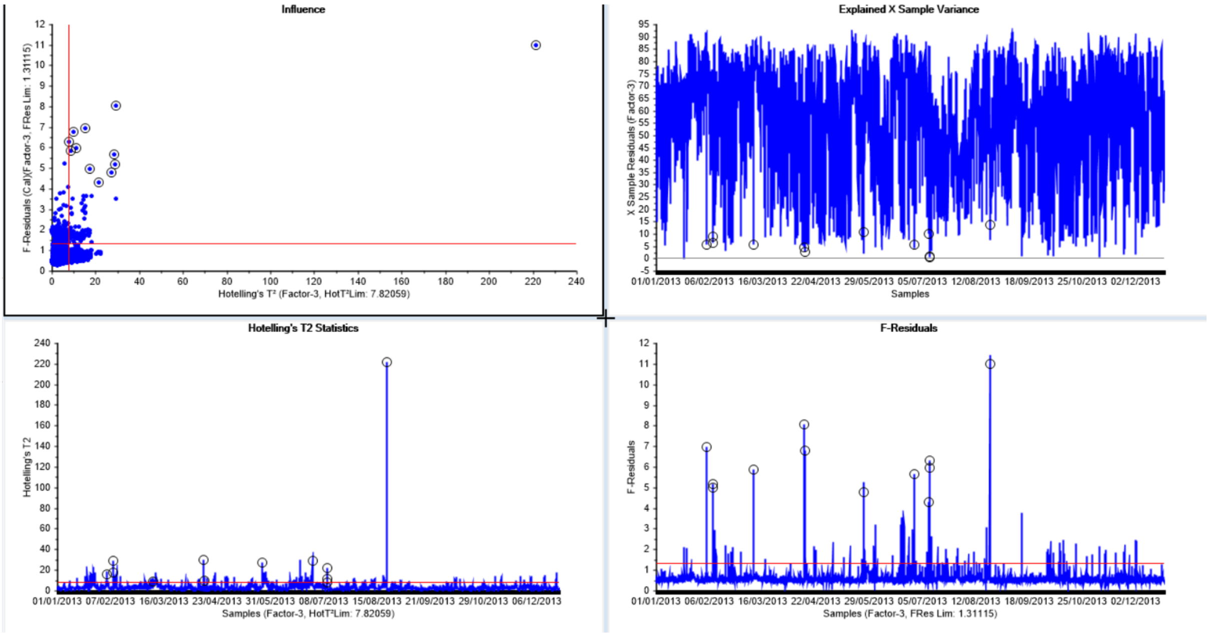
We also see some outliers at 0 reference, but not as many as in for Turbine.

Another interesting observation is that PCR has little explained variance in the first PC, and much more in the second, while PLSR has a lot of explained variance in the first factor.

Based on inspection of the explained variance and loadings plot, it seems that 3 factors for PSLR model, for AllX, Temp and Turbine, are the optimal number of factors. Likewise for PCR models, 4 is seems to be the optimal number of PCs.

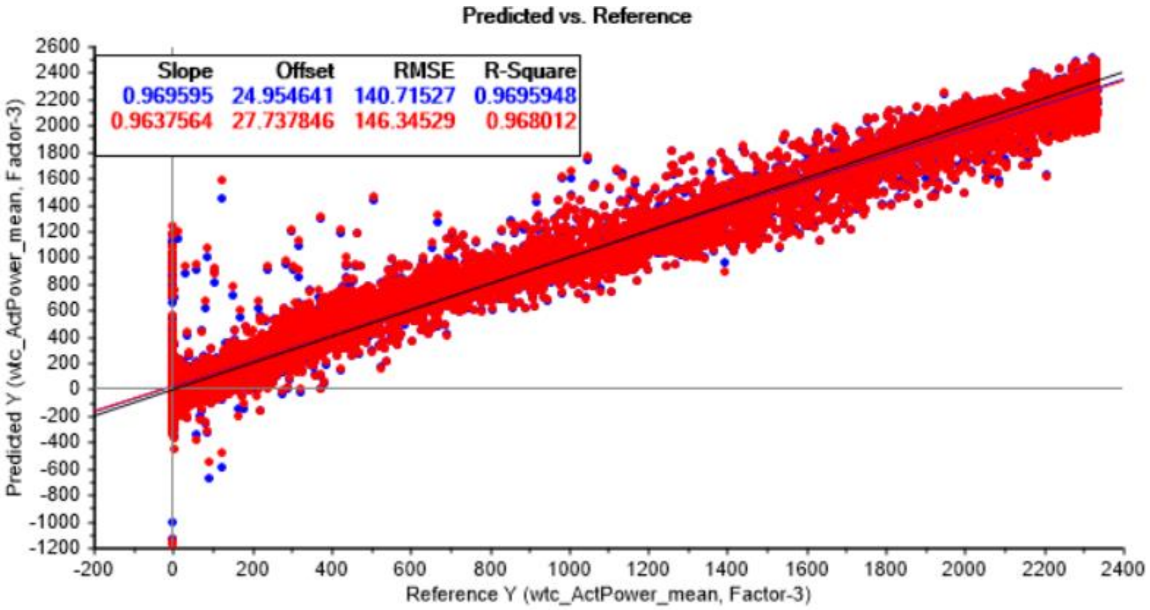
Outliers

Doing this for PLSR on AllX.

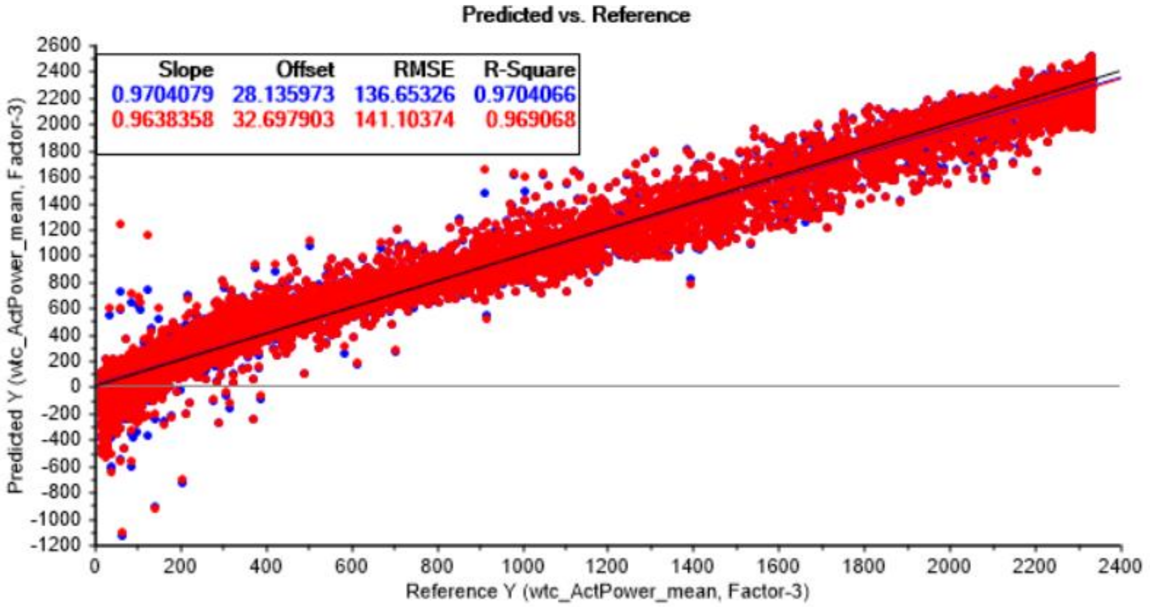


There are some clear outliers like the one point lying all the way to the right in the influence plot. There are also some extra points in the F-residuals plot that could be outliers. These outliers are also seen in the predicted vs reference plot on the previous page that many of them are at 0 reference.

Removing the selected outliers above gave little improvement in RMSE and R-square.



Now trying to remove all samples that are below 0.1 of actual power (Y) (this was done manually by zooming in on the line plot as I don't know a more efficient way to do it). These values can be thought of being so low that they are basically zero, and are connected to turbines that are offline and therefore shouldn't be part of the model.

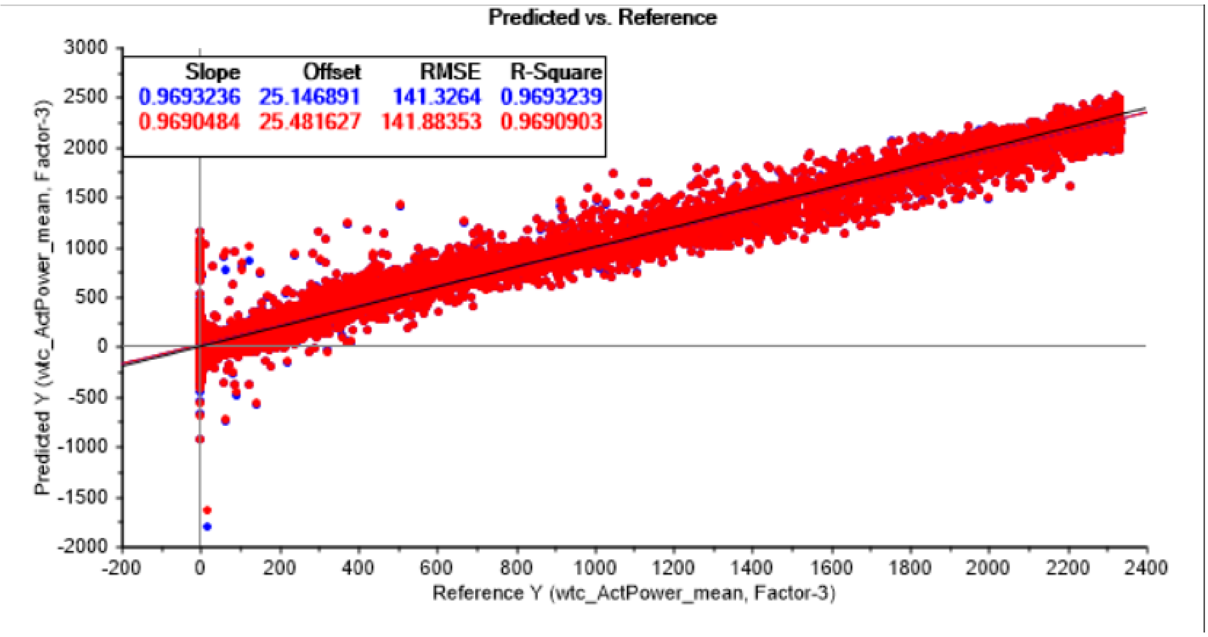


The RMSE is slightly improved to 141, but this is not a very significant change, and perhaps this indicates that the model indeed was good even though there were so many values close to zero.

Different validation schemes

Up until now we've always validated across months. Now lets see how PLS on AllX does with different CV schemes.

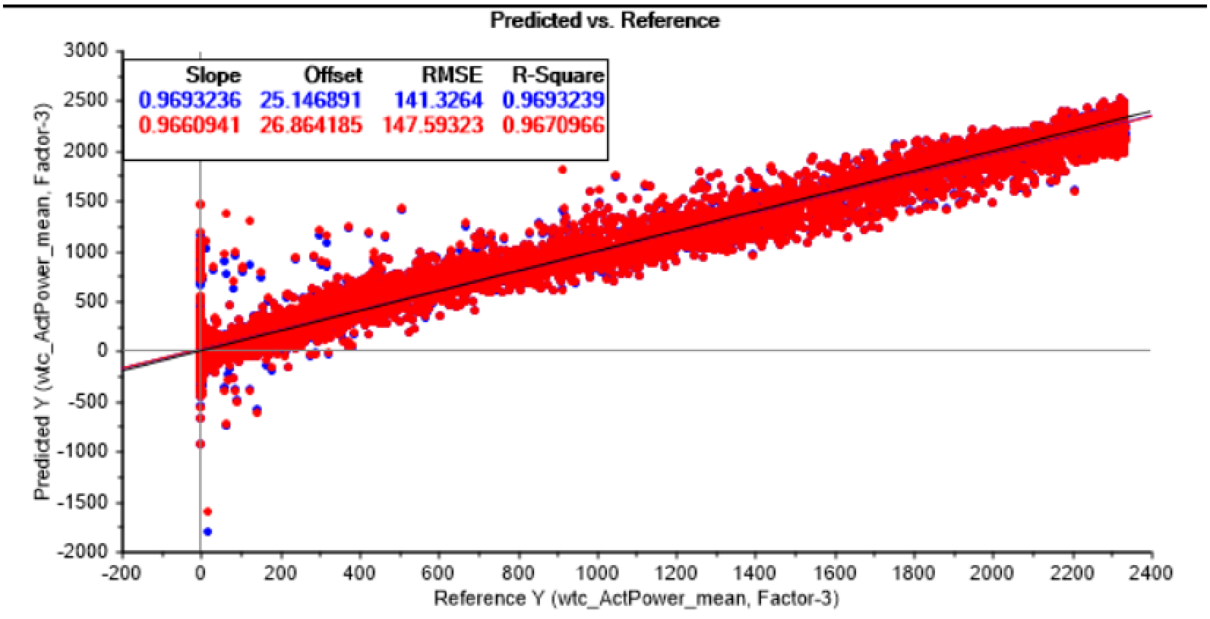
20-fold Random CV



Slope	Offset	RMSE	R-Square
0.9693236	25.146891	141.3264	0.9693239
0.9690484	25.481627	141.88353	0.9690903

RMSE for calibration and validation is more or less identical, and quite good R-square. In fact, these numbers are very similar to when making a model without zero power samples.

20 segments systematic 112233 CV

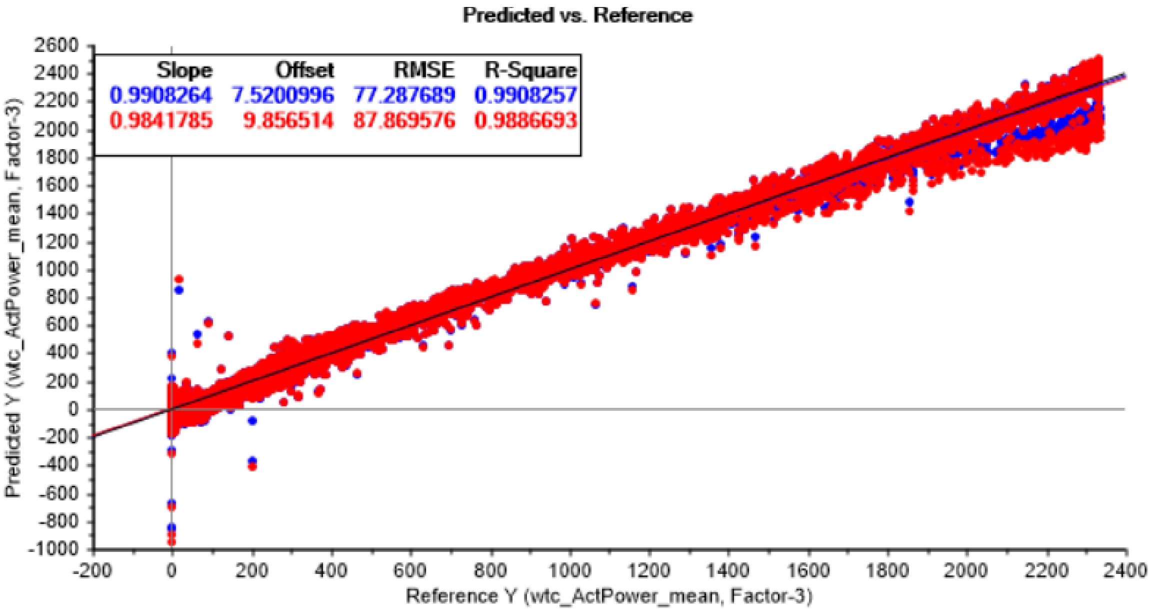


Slope	Offset	RMSE	R-Square
0.9693236	25.146891	141.3264	0.9693239
0.9660941	26.864185	147.59323	0.9670966

Slightly worse than above. But still quite good compared to some of the PCR models we saw earlier.

Without some variables

Using only the Grid column set gives this



Which is the best RMSE so far. This set only has 42 variables and shows that we only need some of the variables and that there is a lot of redundancy in the data.

To have about the same RMSE as for AllX we can use Temp which had RMSE of 188, but this also has 144 variables.