# Kendall Brown r0773111 Chemometrics, Fall 2019 Assignment 3

# Task

To build several PLSR models for the purpose of implementing and comparing various variable selection algorithms. A genetic algorithm, interval PLS and recursive PLS method will be used for the comparison. The data used will be milk-protein reflectance spectra from 50 samples (32 allocated for calibration, 15 for validation, and 3 for prediction). The data shall undergo a first derivative pre-processing prior to model building and evaluation. Outlaying samples discovered through PCA will be excluded from the calibration data.

# PCA

Prior to undergoing a principle component decomposition, the spectra were mean centered. The first four principle components and diagnostic plots have been provided below. From the results it is apparent that sample 5 is an outlaying sample. For the purposes of model building sample 5 will be removed from the calibration data. Sample 10 is a borderline outlier and shall remain within the calibration data.

## Model

Principal Components Analysis Model

Developed 21-Nov-2019 00:22:26.223

Author: kebro@DESKTOP-B4UA9E4

X-block: 32 by 371 (kebro@DESKTOP-B4UA9E4@20191121T001319.40889196 m:20191121001319.677)

Included: [ 1-32 ] [ 1-371 ]

Preprocessing: Mean Center

Num. PCs: 4

Algorithm: SVD

Cross validation: venetian blinds w/ 10 splits and blind thickness = 1

RMSEC: 0.000522548

RMSECV: 0.00127891

## SSQ Table

Percent Variance Captured by PCA Model

Principal Eigenvalue % Variance % Variance

Component of Captured Captured

Number Cov(X) This PC Total

--------- ---------- ---------- ----------

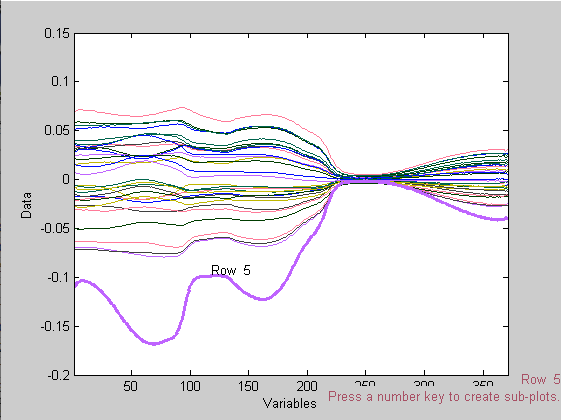
1 3.88e-01 98.74 98.74

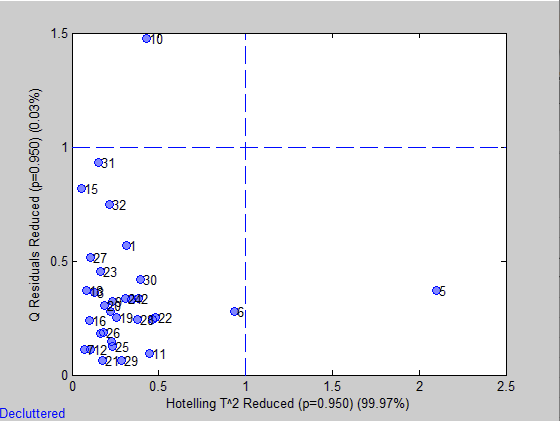
2 4.06e-03 1.03 99.78

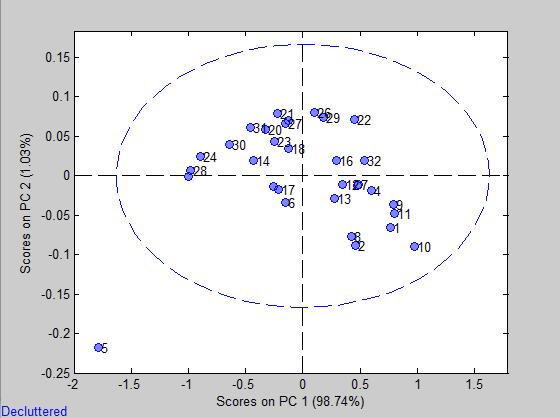
3 5.53e-04 0.14 99.92

4 2.18e-04 0.06 99.97

## Figures associated with the analysis:







# Baseline Model

Prior to variable selection, it is important to establish a baseline model for which to compare all models to. This model shall not undergo any variable selection apart from having sample five removed from the calibration data set. Model diagnostics can be seen below.

## Model

Linear regression model using

Partial Least Squares calculated with the SIMPLS algorithm

Developed 21-Nov-2019 20:31:12.134

Author: kebro@DESKTOP-B4UA9E4

X-block: 31 by 371 (kebro@DESKTOP-B4UA9E4@20191121T194801.52689196 m:20191121201923.375)

Included: [ 1-4 6-32 ] [ 1-371 ]

Preprocessing: 1st Derivative (order: 2, window: 15 pt, tails: polyinterp)

Y-block: 31 by 1 (kebro@DESKTOP-B4UA9E4@20191121T194813.57563904 m:20191121201923.388)

Included: [ 1-4 6-32 ] [ 1 ]

Preprocessing: Autoscale

Num. LVs: 10

Cross validation: random samples w/ 10 splits and 5 iterations

RMSEC: 0.0952841

RMSECV: 0.174859

Bias: 0.000374122

CV Bias: -0.00675456

R^2 Cal: 0.971876

R^2 CV: 0.911783

## SSQ Table

Percent Variance Captured by Regression Model

-----X-Block----- -----Y-Block-----

Comp This Total This Total

---- ------- ------- ------- -------

1 97.35 97.35 1.22 1.22

2 2.60 99.95 34.89 36.11

3 0.02 99.97 13.59 49.70

4 0.01 99.98 13.81 63.50

5 0.01 99.99 5.69 69.19

6 0.00 99.99 8.97 78.17

7 0.01 100.00 3.24 81.41

8 0.00 100.00 11.35 92.76

9 0.00 100.00 3.50 96.26

10 0.00 100.00 0.92 97.19

## Prediction

Linear regression model using

Partial Least Squares calculated with the SIMPLS algorithm

Developed 21-Nov-2019 20:31:12.383

Author: kebro@DESKTOP-B4UA9E4

X-block: 15 by 371 (kebro@DESKTOP-B4UA9E4@20191121T202433.76071190 m:20191121202433.762)

Included: [ 1-15 ] [ 1-371 ]

Preprocessing: 1st Derivative (order: 2, window: 15 pt, tails: polyinterp)

Y-block: 15 by 1 (kebro@DESKTOP-B4UA9E4@20191121T202439.62379766 m:20191121202439.625)

Included: [ 1-15 ] [ 1 ]

Preprocessing: Autoscale

Num. LVs: 10

Cross validation: random samples w/ 10 splits and 5 iterations

RMSEC: 0.0952841

RMSECV: 0.174859

RMSEP: 0.588501

Bias: 0.000374122

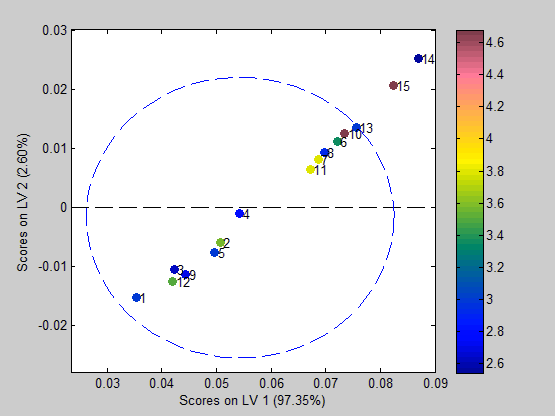
CV Bias: -0.00675456

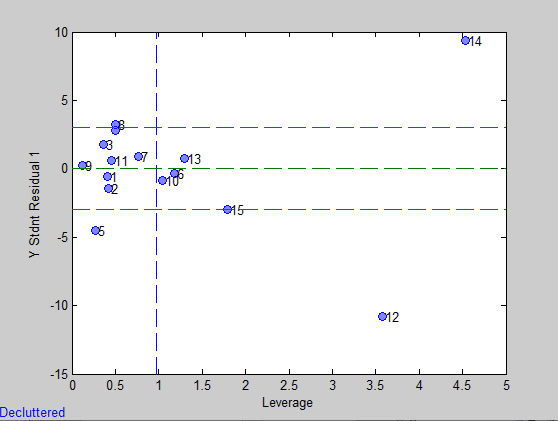
Pred Bias:-0.0172932

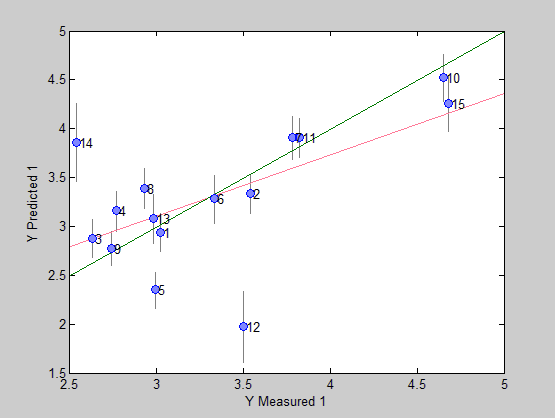
R^2 Cal: 0.971876

R^2 CV: 0.911783

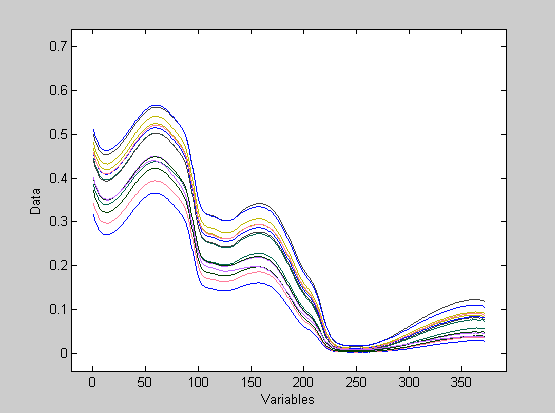
R^2 Pred: 0.366026





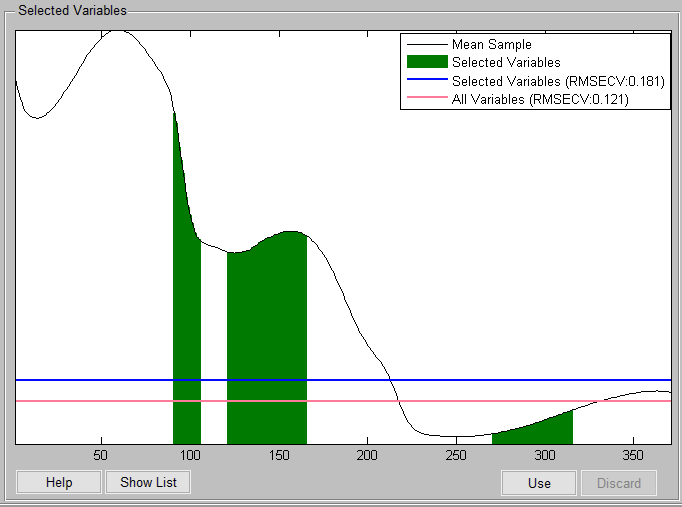


In a ten latent-variable model with no variable selection the resulting cross validation metrics are strong with an R^2 of .912 and a bias of -.007. However, the model is poorly generalizing to new data as the validation set appears to have been influenced quite heavily by outliers (samples 12 and 14). Since the effects of variable selection remain unknown currently and, unlike calibration sample 5, graphically there does not appear to be any odd behavior of the spectra. For these reasons they shall remain within the validation set until after variable selection occurs. Plot of validation spectra shown below



# Genetic Algorithm

With a variable block width of 15, a genetic algorithm is to be used to trim the spectra of unwanted measurements. As seen below the highlighted variables are the surviving traits resulting from the algorithm. Full diagnostics seen below.



## Model

Linear regression model using

Partial Least Squares calculated with the SIMPLS algorithm

Developed 21-Nov-2019 21:23:19.065

Author: kebro@DESKTOP-B4UA9E4

X-block: 31 by 105 (kebro@DESKTOP-B4UA9E4@20191121T194801.52689196 m:20191121212308.885)

Included: [ 1-4 6-32 ] [ 91-105 121-165 271-315 ]

Preprocessing: 1st Derivative (order: 2, window: 15 pt, tails: polyinterp)

Y-block: 31 by 1 (kebro@DESKTOP-B4UA9E4@20191121T194813.57563904 m:20191121205557.852)

Included: [ 1-4 6-32 ] [ 1 ]

Preprocessing: Autoscale

Num. LVs: 10

Cross validation: random samples w/ 10 splits and 5 iterations

RMSEC: 0.077101

RMSECV: 0.14784

Bias: -0.00032791

CV Bias: 0.000480295

R^2 Cal: 0.981586

R^2 CV: 0.938098

## SSQ Table

Percent Variance Captured by Regression Model

-----X-Block----- -----Y-Block-----

Comp This Total This Total

---- ------- ------- ------- -------

1 99.54 99.54 0.30 0.30

2 0.45 99.99 44.49 44.79

3 0.01 99.99 13.69 58.48

4 0.00 100.00 1.97 60.45

5 0.00 100.00 27.83 88.28

6 0.00 100.00 5.15 93.43

7 0.00 100.00 0.30 93.73

8 0.00 100.00 3.24 96.97

9 0.00 100.00 0.80 97.77

10 0.00 100.00 0.39 98.16

## Prediction

Linear regression model using

Partial Least Squares calculated with the SIMPLS algorithm

Developed 21-Nov-2019 21:23:22.842

Author: kebro@DESKTOP-B4UA9E4

X-block: 15 by 105 (kebro@DESKTOP-B4UA9E4@20191121T202433.76071190 m:20191121202433.762)

Included: [ 1-15 ] [ 91-105 121-165 271-315 ]

Preprocessing: 1st Derivative (order: 2, window: 15 pt, tails: polyinterp)

Y-block: 15 by 1 (kebro@DESKTOP-B4UA9E4@20191121T202439.62379766 m:20191121202439.625)

Included: [ 1-15 ] [ 1 ]

Preprocessing: Autoscale

Num. LVs: 10

Cross validation: random samples w/ 10 splits and 5 iterations

RMSEC: 0.077101

RMSECV: 0.14784

RMSEP: 0.243474

Bias: -0.00032791

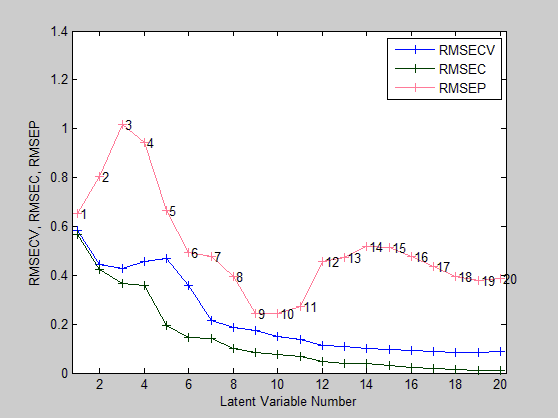
CV Bias: 0.000480295

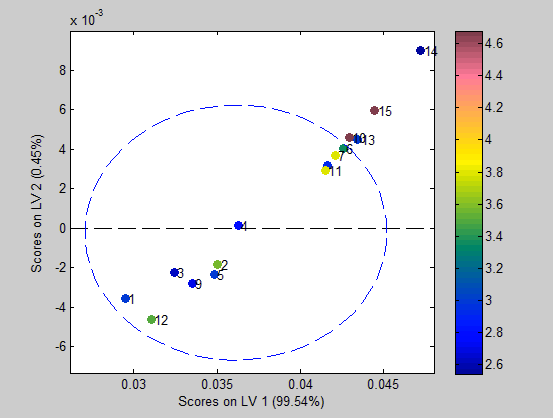
Pred Bias:-0.0789127

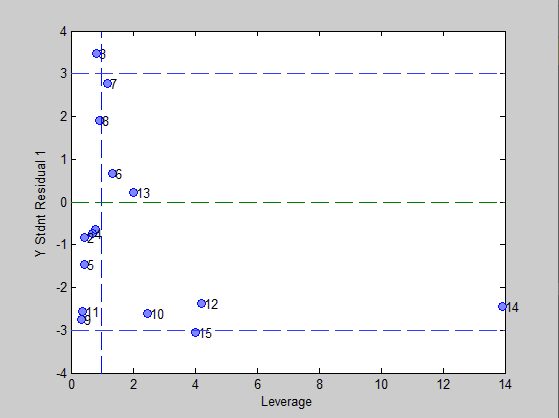
R^2 Cal: 0.981586

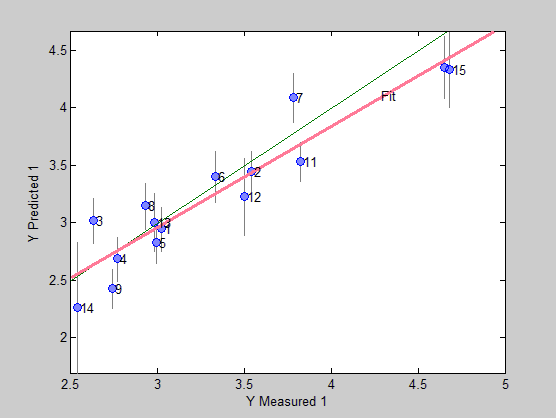
R^2 CV: 0.938098

R^2 Pred: 0.874708









After variable selection, the power of the model has seen rather significant improvements. Cross-validation R^2 and bias both see improvements with R^2 increasing to .938 and bias reducing in magnitude to .0005. Prediction results also see an improvement with its respective R^2 increasing to .874 and bias magnitude unfortunately increasing to .079, with observations 3 and 7 appearing to be the true outliers of the validation data. Removing samples 3 and 7 from the validation table results in the following diagnostics where the prediction results see drastic improvement over its unfiltered iteration.

## Prediction

Linear regression model using

Partial Least Squares calculated with the SIMPLS algorithm

Developed 21-Nov-2019 21:56:43.379

Author: kebro@DESKTOP-B4UA9E4

X-block: 13 by 105 (kebro@DESKTOP-B4UA9E4@20191121T202433.76071190 m:20191121215518.312)

Included: [ 1-2 4-6 8-15 ] [ 91-105 121-165 271-315 ]

Preprocessing: 1st Derivative (order: 2, window: 15 pt, tails: polyinterp)

Y-block: 13 by 1 (kebro@DESKTOP-B4UA9E4@20191121T202439.62379766 m:20191121215518.293)

Included: [ 1-2 4-6 8-15 ] [ 1 ]

Preprocessing: Autoscale

Num. LVs: 11

Cross validation: random samples w/ 10 splits and 5 iterations

RMSEC: 0.0663708

RMSECV: 0.136722

RMSEP: 0.255688

Bias: 1.51846e-05

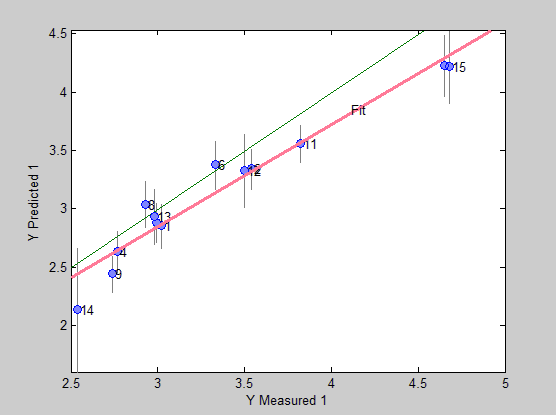
CV Bias: -0.00282611

Pred Bias:-0.192791

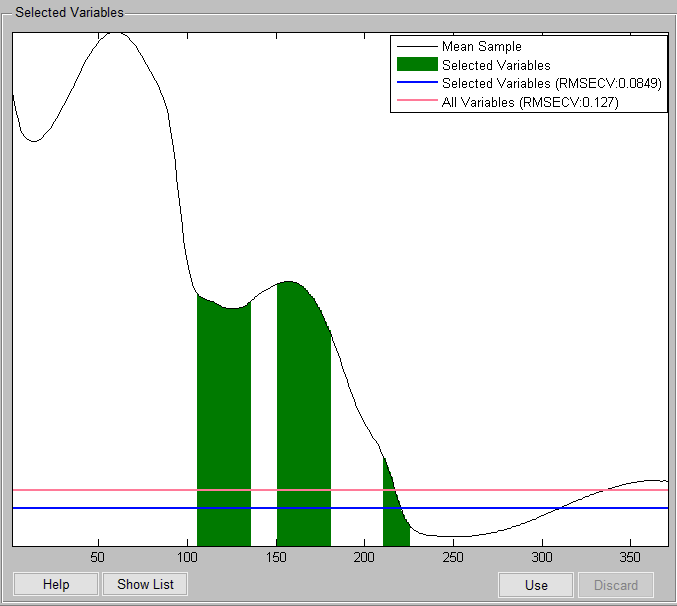
R^2 Cal: 0.986354

R^2 CV: 0.94823

R^2 Pred: 0.939526



# Interval PLS

A forward selection method using bands of width 15 an iPLS variable selection method is to be employed. Shown below, the highlighted range of values are to be used as the training set. As seen, much the data from the tails of the spectra been trimmed away in entirety with only the central third maintaining mostly intact.  


## SSQ Table

Percent Variance Captured by Regression Model

-----X-Block----- -----Y-Block-----

Comp This Total This Total

---- ------- ------- ------- -------

1 99.10 99.10 1.59 1.59

2 0.86 99.96 38.06 39.65

3 0.02 99.98 17.42 57.07

4 0.01 100.00 7.08 64.15

5 0.00 100.00 19.89 84.04

6 0.00 100.00 8.89 92.93

7 0.00 100.00 3.53 96.46

8 0.00 100.00 1.63 98.09

9 0.00 100.00 0.60 98.69

10 0.00 100.00 0.22 98.90

## Prediction

Linear regression model using

Partial Least Squares calculated with the SIMPLS algorithm

Developed 21-Nov-2019 23:00:11.700

Author: kebro@DESKTOP-B4UA9E4

X-block: 15 by 75 (kebro@DESKTOP-B4UA9E4@20191121T202433.76071190 m:20191121230007.896)

Included: [ 1-15 ] [ 106-135 151-180 211-225 ]

Preprocessing: 1st Derivative (order: 2, window: 15 pt, tails: polyinterp)

Y-block: 15 by 1 (kebro@DESKTOP-B4UA9E4@20191121T202439.62379766 m:20191121230007.917)

Included: [ 1-15 ] [ 1 ]

Preprocessing: Autoscale

Num. LVs: 10

Cross validation: random samples w/ 10 splits and 5 iterations

RMSEC: 0.059474

RMSECV: 0.103678

RMSEP: 0.361018

Bias: 5.74995e-06

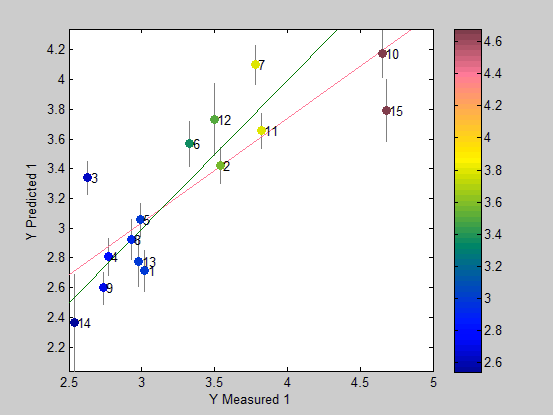
CV Bias: -0.00750002

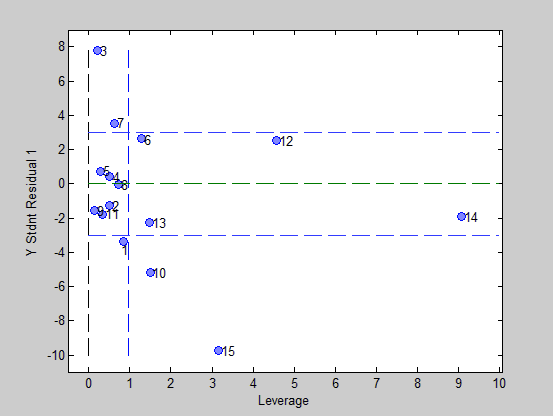
Pred Bias:-0.0567255

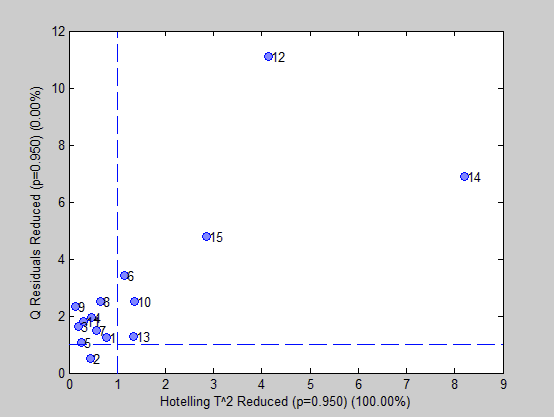
R^2 Cal: 0.989043

R^2 CV: 0.968546

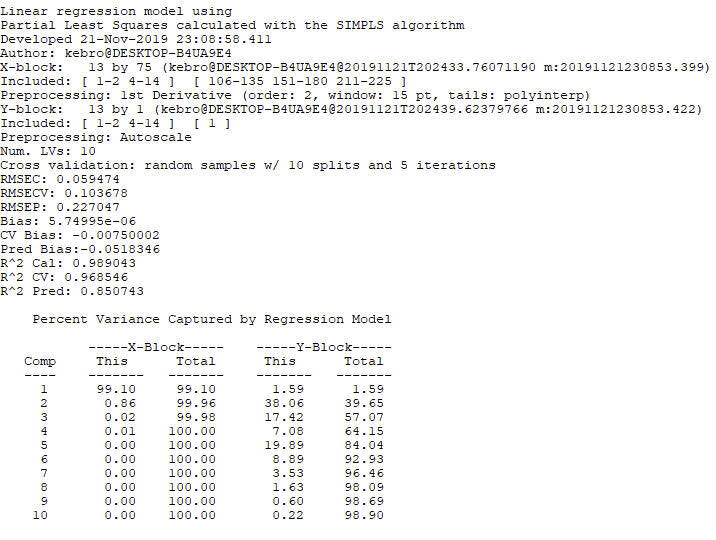
R^2 Pred: 0.699747

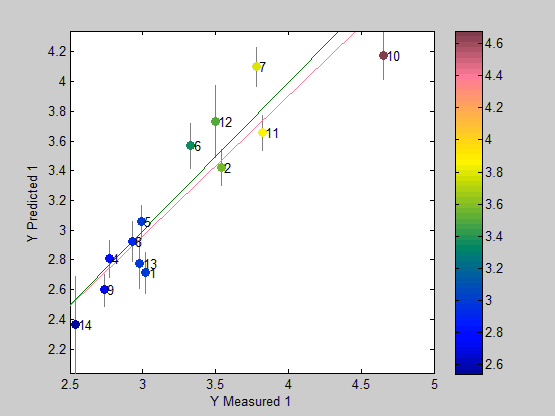






Much like with the genetic algorithm cross-validation metrics are strong with an R^2 of .969 and a bias magnitude of .008. Prediction results leave something to be desired with an R^2 <.7. From the diagnostic plots samples 3 and 15 are behaving poorly. Removing them from the validation set results in the following table which shows that the model does generalize quite well, however outliers are still rather common.





# Recursive PLS

A backward selection method generating upwards of 20 latent variables is to be employed. Shown below, the highlighted range of values are to be used as the training set. A recursive method has trimmed much of the spectra away. The resulting spectra appears to be somewhat uniformly distributed across its original domain.

## SSQ Table

Percent Variance Captured by Regression Model

-----X-Block----- -----Y-Block-----

Comp This Total This Total

---- ------- ------- ------- -------

1 95.35 95.35 1.94 1.94

2 4.55 99.90 33.73 35.68

3 0.04 99.94 11.99 47.66

4 0.03 99.97 10.97 58.63

5 0.01 99.98 9.30 67.93

6 0.01 99.98 12.71 80.64

7 0.01 100.00 5.03 85.66

8 0.00 100.00 11.92 97.58

9 0.00 100.00 0.36 97.94

10 0.00 100.00 0.96 98.90

## Prediction

Linear regression model using

Partial Least Squares calculated with the SIMPLS algorithm

Developed 21-Nov-2019 22:27:06.780

Author: kebro@DESKTOP-B4UA9E4

X-block: 15 by 56 (kebro@DESKTOP-B4UA9E4@20191121T221520.72508174 m:20191121222700.522)

Included: [ 1-15 ] [ 2 5 15 30 33 37-39 57-58 71-72 83-86 105-110 120-122 124-125 127 135-136 147-148 183-184 188-189 210-211 217 227 253 255 263 291-292 309 320-322 350 357-358 362 365 369 371 ]

Preprocessing: 1st Derivative (order: 2, window: 15 pt, tails: polyinterp)

Y-block: 15 by 1 (kebro@DESKTOP-B4UA9E4@20191121T221526.62183381 m:20191121222700.544)

Included: [ 1-15 ] [ 1 ]

Preprocessing: Autoscale

Num. LVs: 10

Cross validation: random samples w/ 10 splits and 5 iterations

RMSEC: 0.0597233

RMSECV: 0.120943

RMSEP: 0.397349

Bias: -5.27657e-06

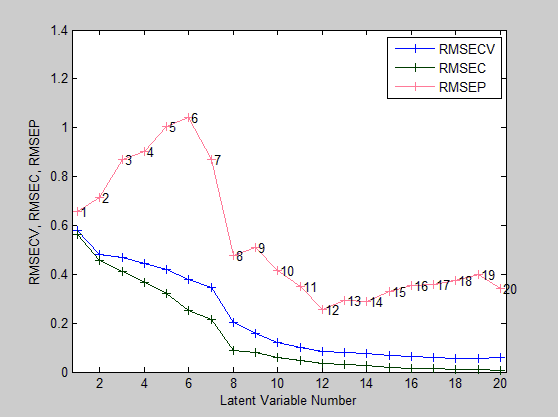
CV Bias: -0.00559091

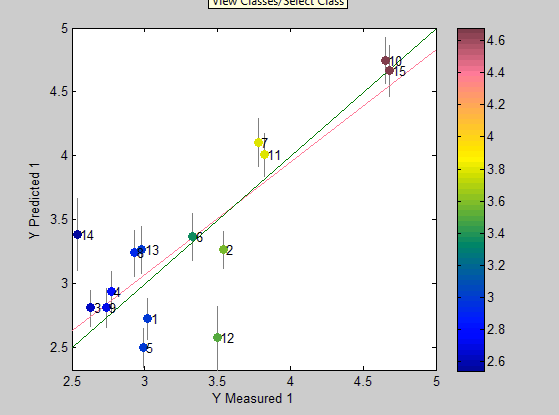
Pred Bias:0.0360293

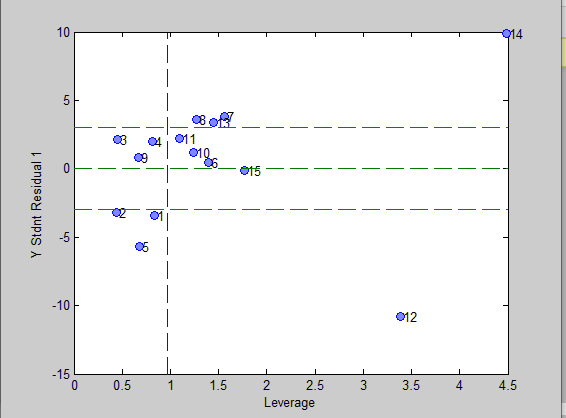
R^2 Cal: 0.988951

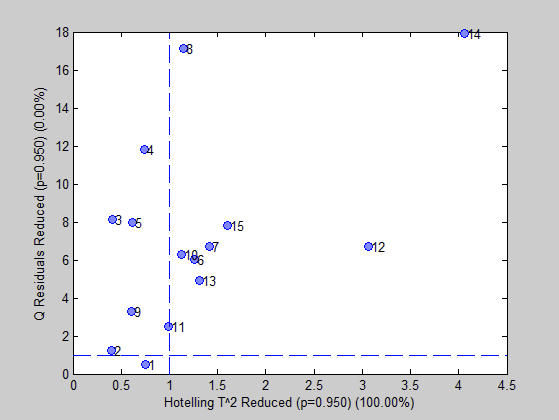
R^2 CV: 0.958454

R^2 Pred: 0.685439









The results are quite strong for the cross-validation measurements. Strong CV R^2 and bias magnitude measurements of .96 and .006 show the model is behaving well during cross validation. Unfortunately, the prediction results are quite poor and from the diagnostic plots above it is quite clear that observations 12 and 14 are outliers. Removing them from validation set results much stronger measurements and can be seen in the following table and plot.

## Prediction

Linear regression model using

Partial Least Squares calculated with the SIMPLS algorithm

Developed 21-Nov-2019 22:46:10.515

Author: kebro@DESKTOP-B4UA9E4

X-block: 13 by 56 (kebro@DESKTOP-B4UA9E4@20191121T221520.72508174 m:20191121224600.060)

Included: [ 1-11 13 15 ] [ 2 5 15 30 33 37-39 57-58 71-72 83-86 105-110 120-122 124-125 127 135-136 147-148 183-184 188-189 210-211 217 227 253 255 263 291-292 309 320-322 350 357-358 362 365 369 371 ]

Preprocessing: 1st Derivative (order: 2, window: 15 pt, tails: polyinterp)

Y-block: 13 by 1 (kebro@DESKTOP-B4UA9E4@20191121T221526.62183381 m:20191121224600.081)

Included: [ 1-11 13 15 ] [ 1 ]

Preprocessing: Autoscale

Num. LVs: 10

Cross validation: random samples w/ 10 splits and 5 iterations

RMSEC: 0.0597233

RMSECV: 0.120943

RMSEP: 0.248248

Bias: -5.27657e-06

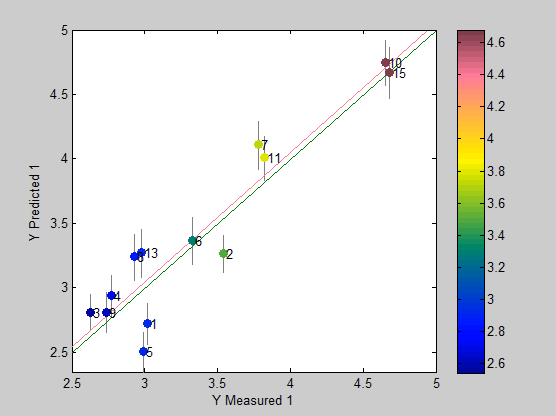
CV Bias: -0.00559091

Pred Bias:0.0472805

R^2 Cal: 0.988951

R^2 CV: 0.958454

R^2 Pred: 0.881563



# Conclusions

A genetic algorithm appears to have resulted in the most powerful model. Variable selection methods appear to provide a way towards acquiring more powerful models. Unfortunately, the validation set appears to be influenced rather severely by outliers, accounting for their presence results in models performing about as well as they do under cross validation circumstances. Applying the 3 samples which were allocated for raw prediction into the model generated by the genetic algorithm gives the following results.

Sample Y Predicted 1

1 4.748267934084

2 2.968819327946

3 3.896504298617