Prediction of dosing behaviour based on powder properties using Lasso regression and Monte-Carlo techniques.

Research

Jean-Vincent Le Bé[1], Isabelle Castella[1], Vincent Girard[2], Marie Perrot[1] & Sophie Berçot[1]

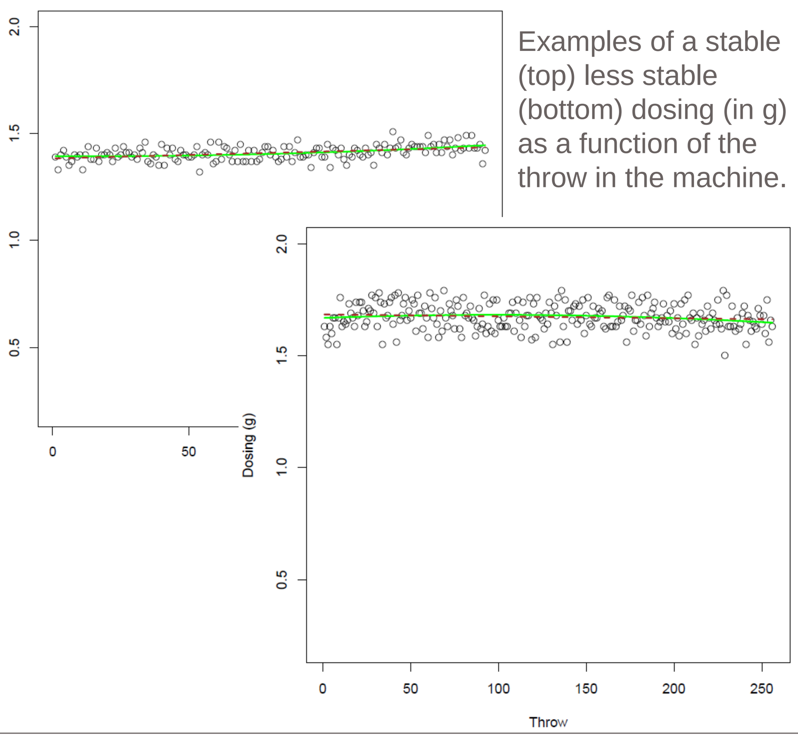
[1] Nestlé System Technology Centre, Orbe, Switzerland; [2] Nestlé Product Technology Centre Beverage, Orbe, Switzerland

Introduction

The dry powder dosing depends on a complex variety of parameters such as environment, physical and chemical characteristics.

The objective of this method is to find which parameters, called *predictors*, are the most influent and calculate the model predicting the dosing, called *variable,* based on the predictors.

A particularity of the datasets analyzed is that there was up to 31 predictors for only 19 products (or observations). A direct linear model could not be calculated. In addition the predictors measurement were given with an experimental uncertainty that needs to be considered.



Examples of a stable (top) less stable (bottom) dosing (in g) as a function of the throw in the machine.

4-steps method

Step 1: Simulate the predictors based on measurement and measurement uncertainty.

Step 2: Reduce the number of predictors to below the number of observations to be able to calculate an error and get the most occurring predictors with minimum error among all simulated datasets.

Step 3: Reduce further the number of predictors with expert based selection and resulting in «not too much» increased error.

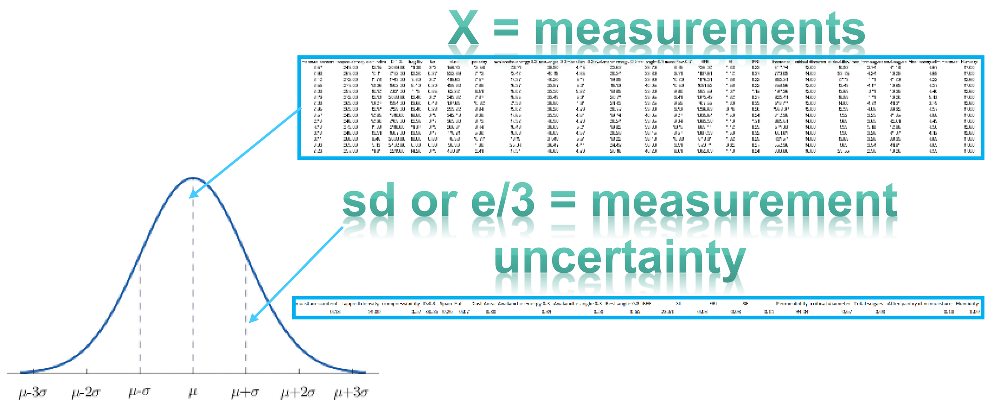
Step 4: Verify the model.

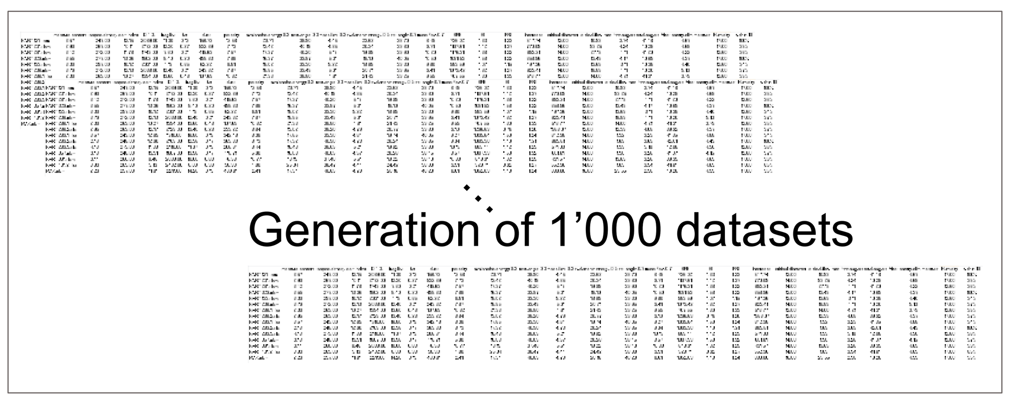
Step 1: Generation of simulated datasets

The measured value *X* of the predictors has a measurement variation *sd.*
It can be based either on the measurement error *e* : *sd = e/3* or on a standard deviation from repeated measurements of the same sample.
With **S samples** (or observations) and **P predictors** the simulated predictors values are: $\tilde{X}_{ij} = X_{ij} + \varepsilon_{ij}$ where $i = 1, \dots, P\ ; j = 1, \dots, S\ ;\ \varepsilon_{ij} \sim \mathcal{N}(0, sd_i)$

and sd_i is the *sd* of i^th^ predictor.

N = 1'000 simulated datasets were generated in the current study.





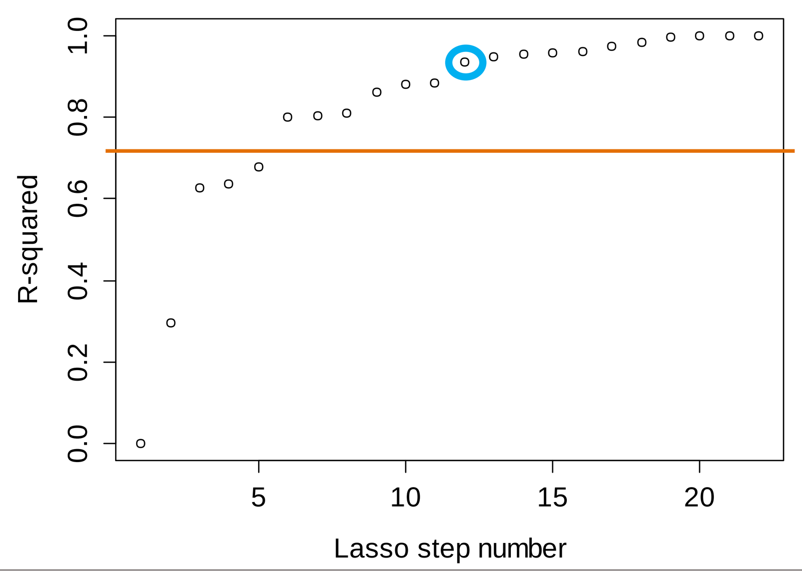
Step 2: Lasso Regressions

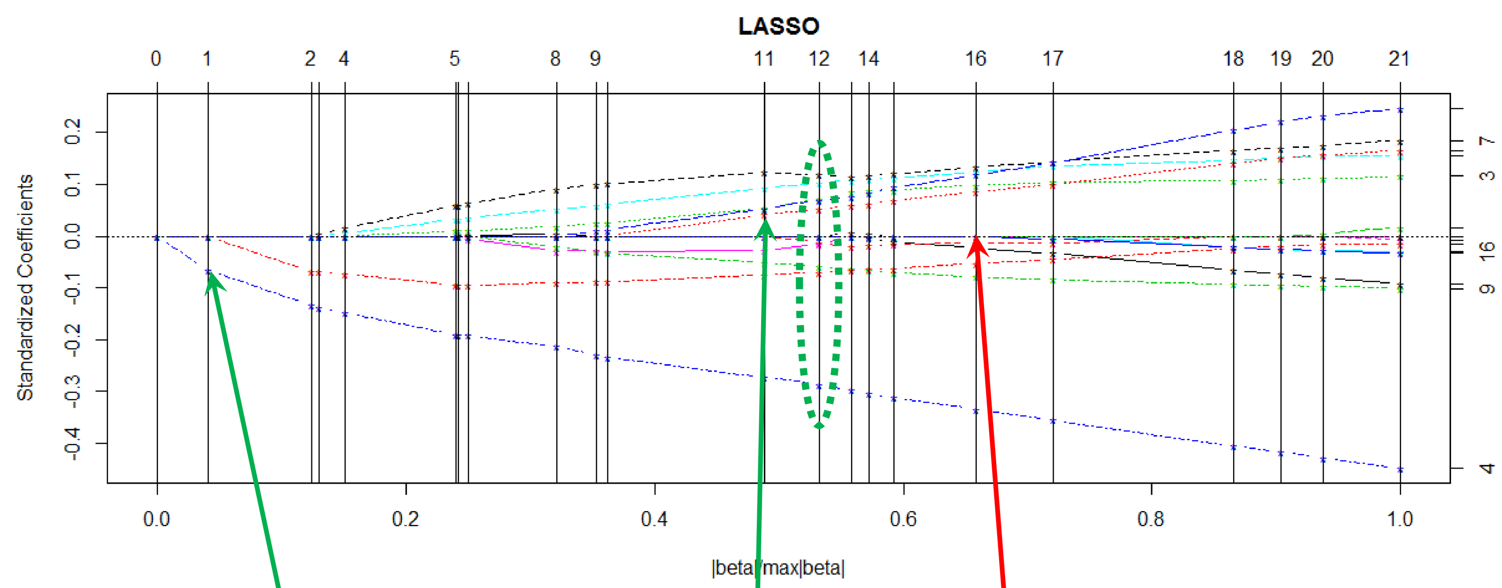
The identification of the relevant predictors using Lasso Regressions is done in two steps.

a) Reduction of the number of predictors to have less predictors than observations

If P > S, the LARS algorithm cannot calculate a Cp to select the best model.
The evolution of R^2 as a function of Lasso steps is then used, and the cutoff in the Lasso sequence is done when the R^2 reaches a plateau, on the graph below.
To prevent a too strict selection only $R^2 > 0.75 \cdot range(R^2)$ are considered, on the graph below.

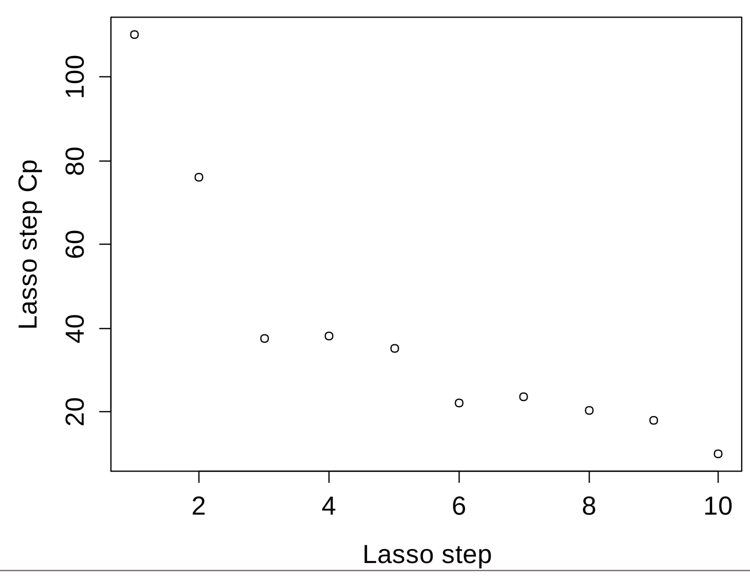
Finally, only the active predictors at this step are kept, on the "Lasso" graph.






```
Sequence of LASSO moves:
     X4 X14 X7 X5 X3 X1 X6 X9 X10 X1 X8 X2 X13 X1 X6 X13 X16 X11 X2 X15 X12
Var   4  14  7  5  3  1  6  9  10 -1  8  2  13  1 -6 -13  16  11 -2  15  12
Step  1   2  3  4  5  6  7  8   9 10 11 12  13 14 15  16  17  18 19  20  21
```

- Each LARS step either adds or suppresses a predictor in the model
- Lowest Cp indicates the best fit. The active variables at this step are recorded.

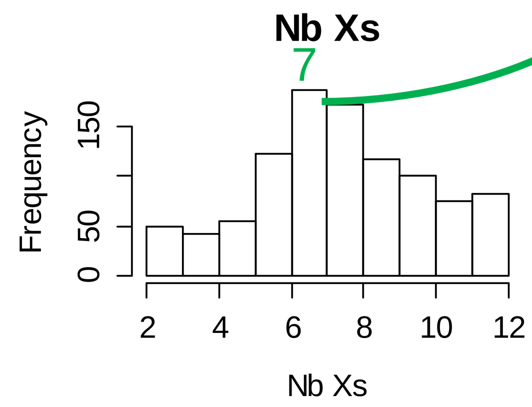


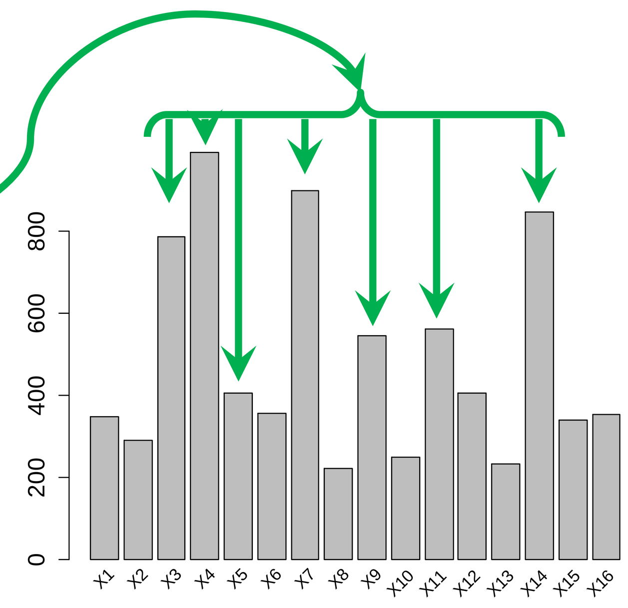
b) Get Models with minimum Cp

When a simulated dataset has P < S (either after step "a" or from start), the step with lowest Cp is identified and its active variables are recorded.

The number of active variables for the lowest Cp for each simulated dataset are compared and the most frequent number of predictors (Nb Xs) at lowest Cp step is selected as the real number of predictors needed.

The most frequent predictors select are selected up to most frequent number of predictors is reached. In the example shown, X3, X4, X5, X7, X9, X11, X14 are selected.



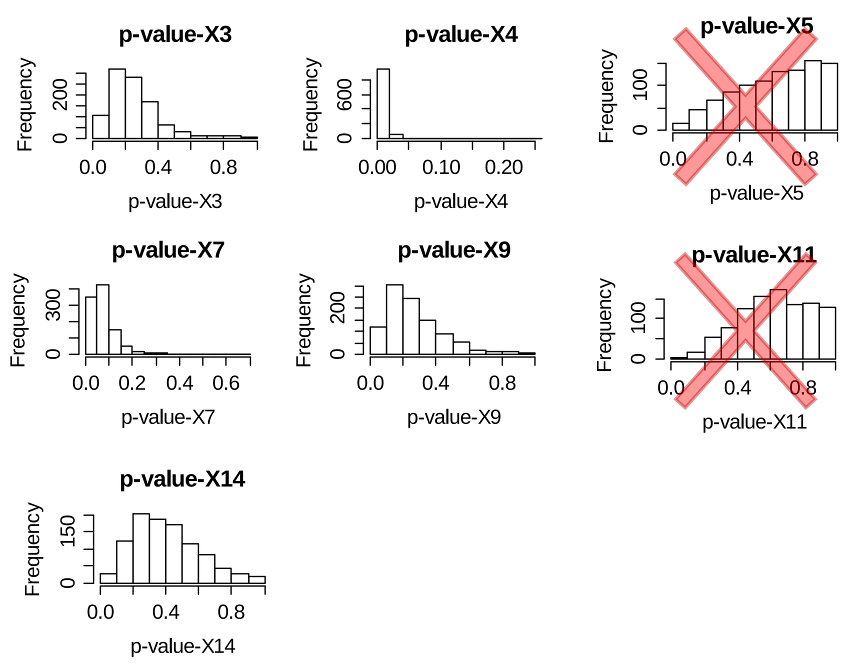


Step 3: Multi-linear Regression and Final Model

Selected Predictors (Xs) by the Lasso step:
X3, X4, X5, X7, X9, X11, X14

A linear model is run on each simulated dataset (1'000 models) and the p-values for each predictor is displayed.

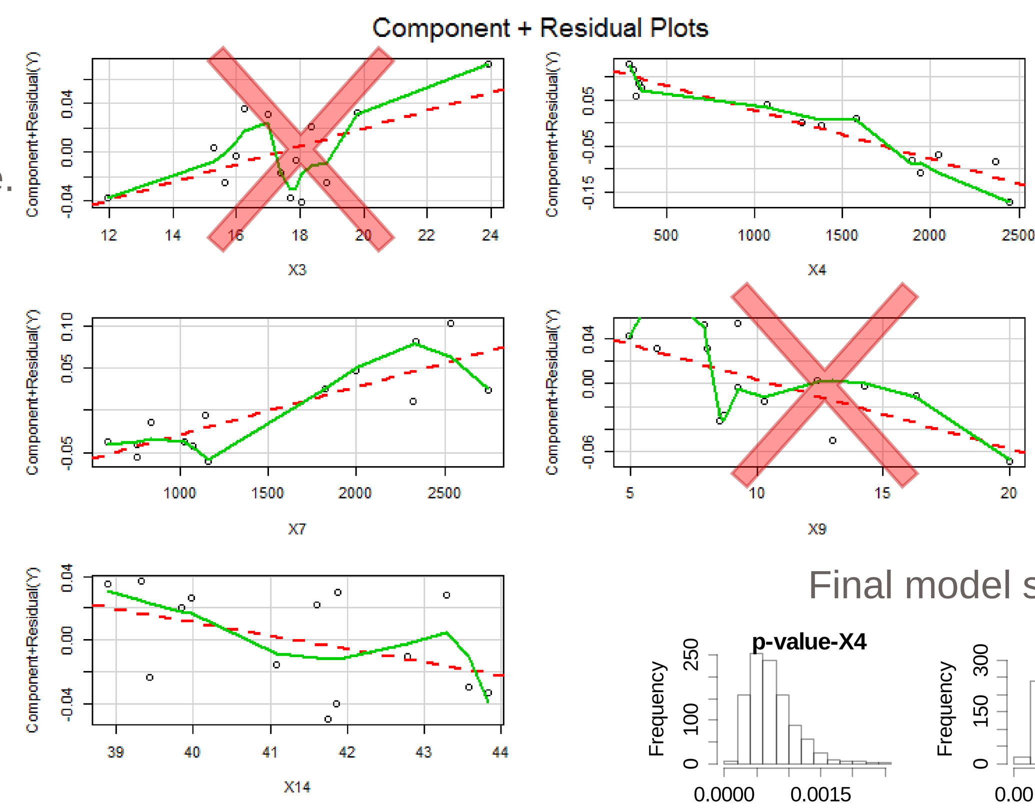
The predictors with p-values too evenly distributed or towards 1 are discarded.



Further selection is done with experts in the field considered to keep only variables making sense.

Component + Residual plots are used to verify that there is no tendency (such as quadratic or exponential) and to discard predictors with a too small amplitude on the CR-plot.

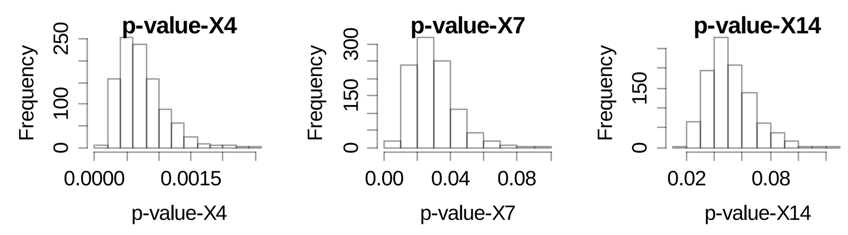
The resulting model is not having the absolute lowest Cp but the Adjusted-R^2 reduction must be «reasonable».



Final model

	CoeffsMeas
(Intercept)	1.549247e+00
X4	-8.761529e-05
X7	4.914006e-05
X14	-1.516741e-02
R-squared	8.001594e-01
Adj R-squared	7.402073e-01
p-value-X4	6.191382e-04
p-value-X7	2.527405e-02
p-value-X14	4.834882e-02

Final model simulated p-values

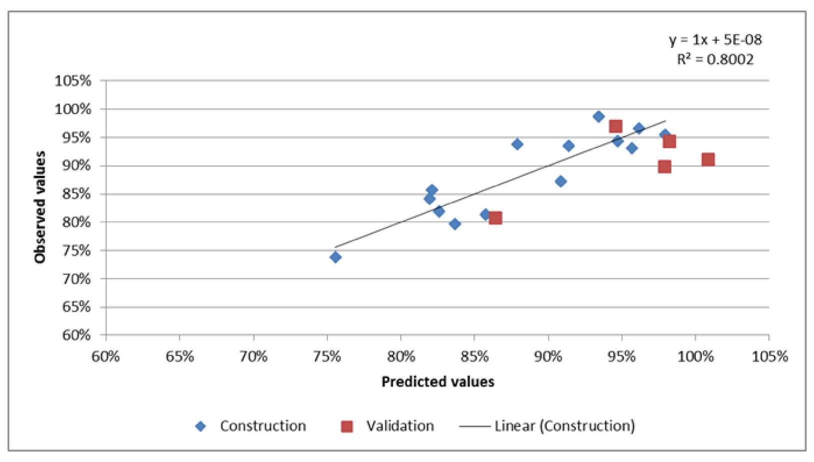


Step 4: Model validation

14 products were used to construct the model

5 products were measured later for validation

If the validation points are within two standard deviations of repeated measurements of Y, the model is considered as valid.



Conclusion

The method presented was developed by iteration during the project. It considers «Error in variables» problem with Monte-Carlo simulation. The original computation time was 8' per variable and case, and after R-code optimization, the final computation time was decreased to 1'25''. It would be now recommended to do 10'000 simulations at once instead of 3x1'000 which was done in the project. In addition a convergence indicator could be developed.

Further improvements could also consider the variation in output (Y) for the simulated datasets.

In addition, other methods could be used for comparison such as Bootstrap, coefficient over predictor uncertainty ratio or clustering.