Paper 8240-2016

Real-Time Predictive Modeling of Key Quality Characteristics Using Regularized Regression: SAS® Procedures GLMSELECT and LASSO

Jon M. Lindenauer, Weyerhaeuser Company

ABSTRACT

This paper describes the merging of designed experiments and regularized regression to find the significant factors to use for a real-time predictive model. The real-time predictive model is used in a Weyerhaeuser modified fiber mill to estimate the value of several key quality characteristics. The modified fiber product is accepted or rejected based on the model prediction or the predicted values deviation from lab tests. To develop the model, a designed experiment was needed at the mill. The experiment was planned by a team of engineers, managers, operators, lab techs, and a statistician. The data analysis used the actual values of the process variables manipulated in the designed experiment. There were instances of the set point not being achieved or maintained for the duration of the run. The lab tests of the key quality characteristics were used as the responses. The experiment was designed with JMP® 64-bit Edition 11.2.0 and estimated the two-way interactions of the process variables. It was thought that one of the process variables was curvilinear and this effect was also made estimable. The LASSO method, as implemented in the SAS® GLMSELECT procedure, was used to analyze the data after cleaning and validating the results. Cross validation was used as the variable selection operator. The resulting prediction model passed all the predetermined success criteria. The prediction model is currently being used real time in the mill for product quality acceptance.

INTRODUCTION

The Weyerhaeuser Cellulose Fibers business runs two modified fiber production facilities. One in Columbus, MS and one in Gdansk, Poland. This paper focuses on the development of the quality predictive model for the Poland Modified Fiber mill (PMF). A similar predictive model is also used in Columbus. The modified fiber product is accepted or rejected based on the model prediction, the predicted value's difference from lab tests or the lab test itself. A bale is produced every few minutes and only lab tested once every 3 hours. If the predicted value, the difference between the predicted and lab value, or the lab test, is beyond the specification limits the bale is quarantined.

The details of the modified fiber production process, and the prediction model, are proprietary but every effort will be made to try and explain them. Figure 1 is a simplified schematic of the modified fiber production process flow. The modified fiber process involves impregnating a pulp sheet with a chemical and disintegrating the pulp sheet into fibers. The fiber is then dried, cured, cooled and baled. The process is continuous. The tests for product quality can take up to an hour to complete in the lab. Our customer proposed that we try to develop a predictive model of the lab quality tests. The model will assure that product made between lab tests has acceptable quality.

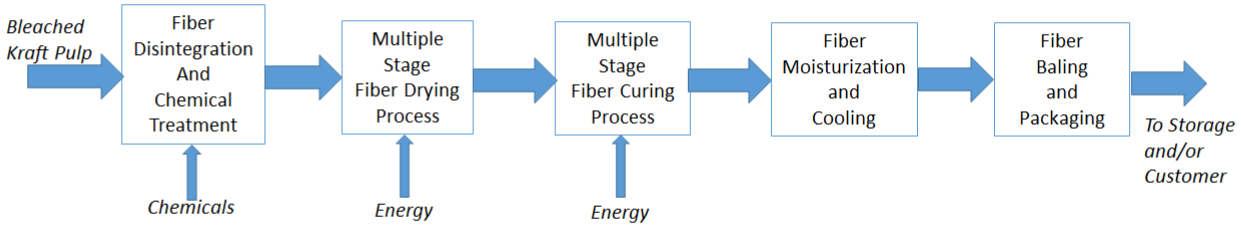

Bleached Kraft Pulp
Fiber Disintegration And Chemical Treatment
Chemicals
Multiple Stage Fiber Drying Process
Energy
Multiple Stage Fiber Curing Process
Energy
Fiber Moisturization and Cooling
Fiber Baling and Packaging
To Storage and/or Customer

Figure 1. Simplified schematic of the modified fiber production process.

Each of these process steps has multiple set points for their stage. One overarching set point, as in any continuous process, is the production rate. Since there are multiple stages with multiple set points, every effort must be made to reduce the number of variables in the final designed experiment for the prediction model. The focus of the team designing the experiment was to use its expertise to reduce the size of the experiment to something reasonable for the mill. This means getting the maximum amount of information in a minimum number of runs to achieve the least downtime and off quality production.

DESIGNING THE EXPERIMENT

The designed experiment is called a “trial” at the mill. A team made up of process engineers, a production supervisor, operators, a lab manager and a statistician worked on the trial plan and preparation. The limitation for the designed experiment was that it needed to be completed within two 12 hour shifts and could only be replicated once. The mill can do no more than 12 runs per shift. The trial could be run as a replicated completely randomized design within 24 hours.

There was more than 20 process settings that could be manipulated during production. Reducing the number of variables for the mill trial was important. Process experience and previous model development knowledge was key to the trial variable reduction. The number of potential experimental factors was reduced to 8 by applying this knowledge.

A sequence of small pre-trial experiments were run to assess some of the process stages and settings effect on the product quality. The pre-trial runs were also used to establish the minimum and maximum levels for the set points of the process variables to be included in the trial. The settings are chosen so we are sure that the quality measures will vary when the mill operating conditions change. Some off quality product should be expected as we explore the limits of the process.

The team was able to select 4 process variables from the remaining 8 that were key levers for product quality. The designed experiment focused on the 4 selected process variables. It is believed that one of the process variables may be curvilinear over its proposed settings. The team also wanted to test if some of the process variables interacted with each other. Given the experiments’ constraints, JMP® was used to design the experiment. There was also a covariate that was monitored for inclusion in the initial model.

Three center points were added per replication to increase the precision in the estimate of the experimental error (Myers 2009). Even though the experiment is being replicated once, we have found that adding center points is necessary because of the variable/noisy mill environment. Another reason for replicating and using center points is the quality testing is destructive, therefore the measurement system variability cannot be separated from the product and process variability.

The D-optimal design was used as the optimality criterion. D-optimality minimizes the variance of the parameter estimates. This is important when trying to discover which factors are significant because the standard errors of the estimates are minimized (Wu 2009). There is more functionality in the JMP “Custom Design” (Bailey 2010), which is omitted for brevity.

One of the trial’s constraints listed above was 12 runs per shift. The design above specified 11 runs. The design above was re-randomized and repeated for the replication runs.

The experiment and sampling are performed as follows: The set points for the 4 factors are dialed in by the operator. The process is stabilized for 30 minutes. The modified fiber samples are taken every 5 to 10 minutes after the 30 minute process stabilization period. There are 3 samples taken per treatment combination. Each of the samples is tested for several key quality characteristics. One of the quality tests is time sensitive, so this test is done first. The experiment’s run and test process repeats itself for 24 consecutive hours. Since the process parameter set points can be reset for each treatment combination and the production run for 30 minutes to stabilize for sampling, the experiment’s design structure is completely randomized (Milliken 2009).

TEAMWORK

The key to success for this project is teamwork. The most important outcome of our trial is that it is run safely. The quality manager, lab manager, operations manager, process engineers, IT analysts, human resources manager, operators, technicians, our customer and a statistician all worked together to make a written, detailed plan.

The lab results and the actual process parameter values reside in an online system. The ease of data access and archiving in a historian enhance the data validation and cleaning process. In our case, two online systems were used: One to collect the quality test data and another to archive the process values being monitored or set in the DCS (distributed control system). Each data set is downloaded separately and then merged. The merging is done with SAS code, which is too large to put in the paper. One final note is that even with all of the automation in data collection, there were errors. It is important to clean and validate data even if it comes from a computer.

STATISTICAL ANALYSIS

The previous predictive models for the quality measures have been linear. Those experiments have been designed to include estimable interactions terms, but not quadratic ones. The interaction effects have not been significant predictors for the previous models. The team thought there might be interactions due to the construction of an additional drying stage. The team also thought that Factor X1 may have a curvilinear effect. The trial minimum and maximum interval for Factor X1 was 2.5 times wider than previous experimental designs. The full model was,

$$\begin{aligned} \textit{Quality Measure} = {} & \beta_0 + \beta_1 * X1 + \beta_2 * X2 + \beta_3 * X3 + \beta_4 * X4 + \beta_5 * X5 \\ & + \beta_6 * X2 * X3 + \beta_7 * X3 * X4 + \beta_8 * X1 * X1 \end{aligned}$$

The process set points are frequently not exactly the same as the actual process values during production. There may be instances of the set point not being achieved or maintained for the duration of the run. We have found that for model building, it is better to use the actual recorded values of process variables in the historian, not the recorded set points.

As mentioned above, there were 22 runs made for the trial. Each treatment combination had 3 samples taken and tested for a total of 66 observations. All 66 test results were used in the data analysis. It can be argued that the 3 samples within a treatment combination are sub-samples and should be averaged. However, the variability in the testing and the process, as well as our need to model that variability for prediction, led us to the decision to use all the data. The production of one bale usually takes less than 5 minutes, so we are sampling at the experimental unit that the customer would see at their plant.

Standardizing the predictor variables is widely recommended in the literature (Draper 1998). The factors used as the predictors were all on a very similar scale. The process data was standardized for the statistical analysis.

LASSO MOTIVATION

The existing predictive model at the mill had **R^2** of the predicted versus observed data in the 0.45 range when the models were developed. There is more to the "success criteria" (this is discussed later) than **R^2**, but the idea of a better predictive model was desired. It was thought that using a machine learning algorithm like LASSO (Least Absolute Shrinkage and Selection Operator) would help.

The use of a modified LARS (least angle regression) algorithm to successfully analyze design of experiments data with a large number of candidate variables ($p > n$), for example a Plackett-Burman design, has been discussed in the literature (Yuan 2007). The use of a cross validation score as a selection criteria was also discussed. The algorithm that the PROC GLMSELECT uses for the LASSO is based on a variant of the LARS algorithm (Cohen 2006).

The LASSO is a shrinkage or penalty method arising for ordinary least squares (a type regularized regression). These methods effectively perform variable selection by reducing non-significant coefficients to zero. Since the LASSO is a machine learning technique, it greatly mitigates variable selection bias (Cohen 2006). The LASSO can yield a reduction in variance with only small increases in bias and this will generate more accurate predictions. LASSO performs well when there are a small number of predictors with large coefficients and the remaining predictors have very small coefficients (James 2013).

The LASSO coefficients minimizes the residual sum of squares (RSS) subject to a penalty. This can be written as a constraint problem. More formally, if you have a response ***y*** and a matrix of variables ***X*** that have been standardized, then for a tuning parameter ***s***, the LASSO coefficients $\boldsymbol{\beta_j}$ are the solution to

$$\min_{\beta} \|y - X\beta\|^2 \text{ subject to } \sum |\beta_j| < s$$

If the tuning parameter ***s*** is large enough, the least squares solution results. If ***s*** is small enough, then some regression coefficients will be zero and LASSO acts as a variable selection operator. The LASSO procedure increase ***s*** in small steps and calculates a sequence regression coefficients for the non-zero parameters at each step. The sequence of models created by LASSO can then be selected by a specified criteria (James 2013; Cohen 2006).

Predictive model building is about choosing variables that simultaneously achieve low variance and low bias. This is called the bias-variance trade-off. The mean square error (MSE) has two variance components:

$$MSE = (bias)^2 + \sigma^2$$

The bias = 0 for a “training” dataset, but not for the “validation” or “test” dataset. Large bias can occur for a model not being the true model. We want to choose a model to minimize MSE on the validation or test dataset.

If you don’t have a large dataset that can be split into training, validation and test data, then *k*-fold cross validation, using 5 or 10 folds, is recommended. 5-folds or 10-folds yield test error rates with low bias and low variance (James 2013). *K*-fold cross validation is available as an option for the LASSO.

To obtain coefficients based on regularized regression to calculate predicted residual sum of squares (PRESS) in GLMSELECT, select the CHOOSE=CVEX option for “external cross validation”. The *k*-fold cross validation method works by splitting the data into *k* equal datasets. The LASSO uses *k* – 1 datasets to fit a model. The CVEX PRESS statistic is calculating on the one remaining “validation” dataset. This process is repeated *k* times for a given tuning parameter ***s***. As LASSO iterates through the set of tuning parameters ***s***, the CVEX PRESS is stored to make *k* cross validated sets of CVEX PRESS values creating *k* solution paths. A complex algorithm is used merge the *k* model fits and compute CVEX PRESS scores for the set of tuning parameters. The model with the smallest CVEX PRESS score is selected (SAS®/STAT 13.2 User’s Guide).

More detail on LARS (Efron, 2004) and LASSO (Tibshirani 1996) is available at http://statweb.stanford.edu/~tibs/lasso.html . *K*-fold cross validation is detailed in James, 2013. The Cohen SUGI 31 paper has an excellent description of LASSO and GLMSELECT.

DATA ANALYSIS

The final dataset yielded one surprise. Variable *X1* and *X5* were highly correlated (***r*** = 0.62). Multicollinearity can cause problems for the LARS algorithm and most other regression methods (Yuan 2007). Recall that the variable *X5* was a covariate whose values were recorded as the trial was run. *X5* was removed from the analysis.

The LASSO method, with *k*-fold cross validation is implemented in the SAS GLMSELECT procedure. The following code was used to perform that LASSO with 5-fold (default) cross validation for the model specified earlier in the paper:

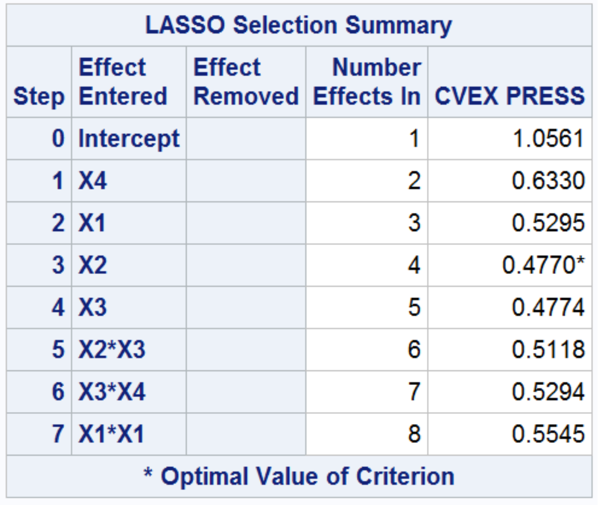
```
proc glmselect data=CenteredData plots=all seed=193794473 ;
  model QM = X1 X2 X3 X4 X2*X3 X3*X4 X1*X1
           /selection=lasso(stop=none choose=cvex) ;
          output out=diag pred resid;
run;
```

The **selection=lasso(…)** option implements the LASSO method with options. The option **stop=none** allows the procedure to complete full model (all variables specified). This allows for the visualization of the entire solution path. The option **choose=cvex** specifies the use of PRESS with *k*-fold external cross validation (CVEX PRESS) method for LASSO. This performs the requested regularized regression and chooses the model that minimizes CVEX PRESS. External cross validation uses the coefficients based on regularized least squares regression to compute the PRESS statistic. The CVEX PRESS is computed at each step of the model selection and the step that has the smallest CVEX PRESS is selected.

Note that the **choose=cv** option for cross validation does not use coefficients based on regularized regression results (SAS®/STAT 13.2 User's Guide). The **output** statement saves the dataset and adds the predicted and residual values. Specifying a **seed=** allows the same stream of random numbers to be used every time you run the procedure.

Selected output for the GLMSELECT procedure are displayed and described below. Output 1 shows the results of the LASSO variable selection process that minimizes CVEX PRESS using 5-fold cross validation. The procedure selected the model below (no interaction or quadratic terms):

$$QualityMeasure = \beta_0 + \beta_1 * X1 + \beta_2 * X2 + \beta_4 * X4$$



LASSO Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	CVEX PRESS
0	Intercept		1	1.0561
1	X4		2	0.6330
2	X1		3	0.5295
3	X2		4	0.4770*
4	X3		5	0.4774
5	X2*X3		6	0.5118
6	X3*X4		7	0.5294
7	X1*X1		8	0.5545

* Optimal Value of Criterion

Output 1. Output results of the LASSO variable selection process.

Notice that there is very little difference in CVEX PRESS between adding *X2* (PRESS = 0.4770) and adding *X3* (PRESS = 0.4774) to the model. This represents the "art" of model building. Do you go with the more parsimonious model or do you add *X3* to the model? A simple model is usually better, but in the past X3 has been included in the prediction model. Fortunately, GLMSELECT gives more output to help make this decision. Figure 2. shows the progression of the standardized coefficients and the CVEX PRESS statistic after each iteration of variable inclusion in the model for QM. Clearly variables *X1*, *X2* and *X4* belong in the model as PRESS is reduced and they all have non-zero coefficients. *X3* has a non-zero coefficient, but PRESS increases slightly at its inclusion in the model.

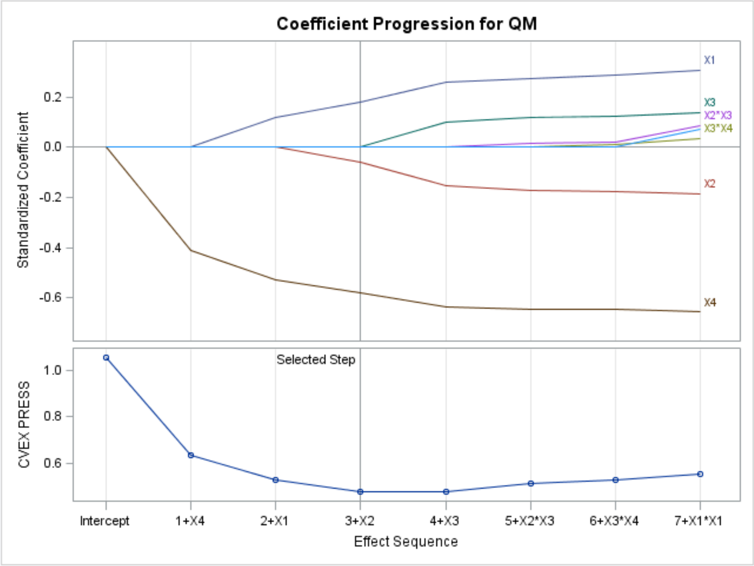


Figure 2. Graph of the output results for variable inclusion in the model.

Additional model fit criteria graphs are also available in the output. Figure 3 displays the fit criteria AIC (Akaike Information Criterion), AICC (Corrected Akaike Information Criterion), SBC (Schwarz Bayesian Criterion), Adj R-Sq (Adjusted R-Square) and the CVEX PRESS (Cohen 2006).

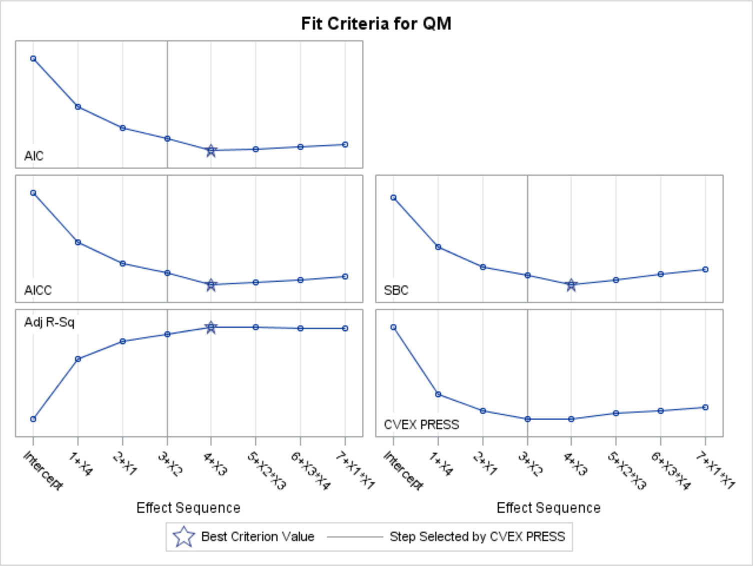


Figure 3. Graph of output showing most commonly used fit criteria for model building.

We want to minimize AIC, AICC, SBC and CVEX PRESS, and maximize Adj R-Sq (Cohen 2006). Similar to the previous plot, the fit criteria are plotted after the inclusion of each variable. The optimal fit criteria for a chosen model is highlighted by a star. Notice that all of the fit criteria, except CVEX PRESS, choose a model with *X3* included.

Figure 4 shows the residual versus predicted quality measure and residual versus run order plots. There are no apparent trends or patterns in these model diagnostics plots. The run order does show lower residual values at the end of the trial, but the team felt that it was due to the treatment combinations that were run at that point of the experiment.

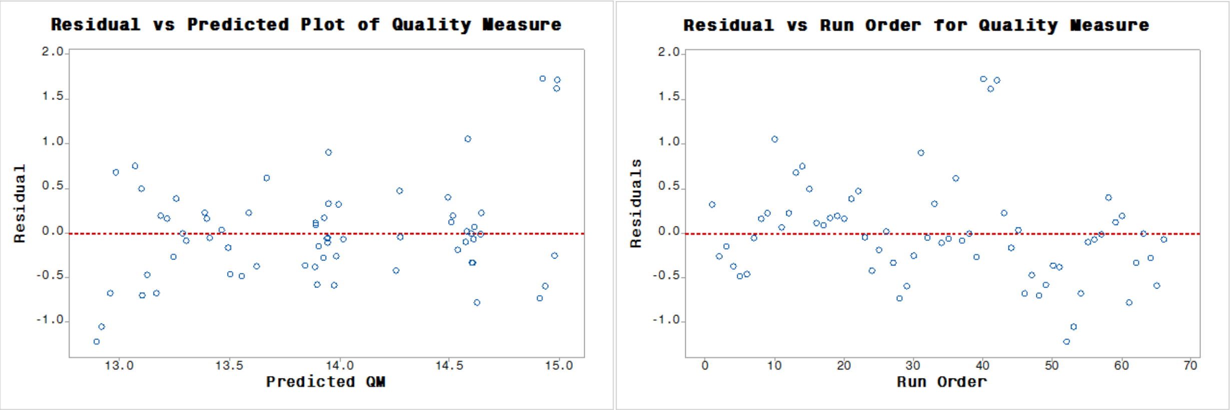


Figure 4 Plots of the residual versus predicted quality measure and residual versus run order.

Other tabular output from GLMSELECT (not displayed) includes the ANOVA table, a summary of the fit statistics for the selected model and the parameter estimates of the regularized regression of that model. The parameter estimates are what we use for the predictive model.

One of the reasons we chose to analyze the trial design of experiment results with LASSO and *k*-fold cross validation was to improve **R²** for the observed versus predicted values. Recall that in the past that **R²** was in the 0.45 range. Figure 5 displays the selected model observed versus predicted values. The plot shows that **R**^{2} = 0.67, an improvement. While we can't say that the LASSO was responsible for the good fit, it may have been helpful.

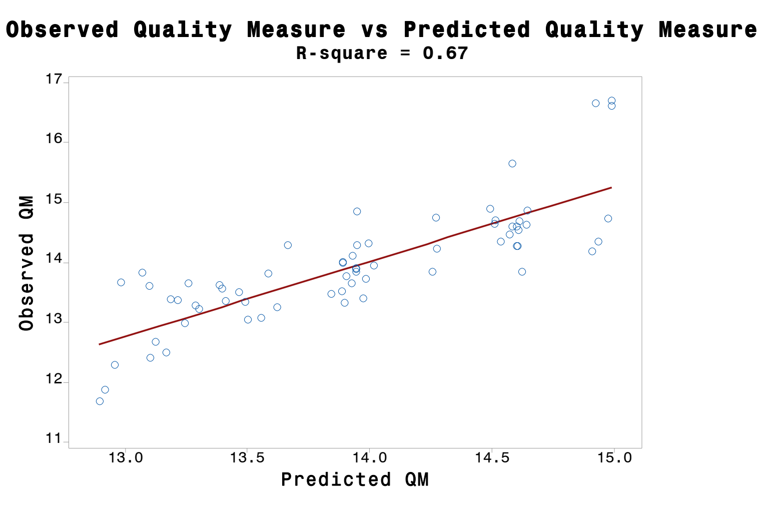


Figure 5. Plot of the observed quality measure versus predicted, with R^2 = 0.67.

Model Fitting Conclusion

The final model fit criteria conundrum: to include or not include *X3*? The process experts on the team reviewed the selected model to confirm variables in the model described known process cause and effects. It was decided we should wait for more actual run-time production data from normal mill operations to become available. The same process variables and product quality measures used for model development were collected and used as model validation data. We prefer a hundred or more independent lab quality measures, with time matching process data, for validation.

Models with *X3* excluded and included were used to predict nearly 200 results of the quality measure. The model without *X3* had an actual vs predicted **R**^{2} = 0.22. The model with *X3* had an actual vs predicted **R**^{2} = 0.20. It was decided to use the simpler model (chosen by LASSO) as the predictive model for the key quality characteristic. The selected predictive model will be validated using a set of "success criteria" that were agreed upon with our customer. If the parsimonious model did not pass the success criteria, we would then use the model that included *X3* and test if it passed the success criteria.

SUCCESS CRITERIA

There were four success criteria spelled out for the quality measure's predictive model to be validated for use in the mill by our customer. We will only discuss two of those criteria, as they are important for real-time process control. The model is first evaluated "off-line" with real-time data. The quality and statistics team members validate the predictive model before going "live". The following examples use 100 simulated data points.

Success Criteria #1 is that the predictive model is accurate: lab and model data are within range allowed by the specifications. The specifications are from past calculations based on mill process capability. There are both lower and upper specifications. Figure 6 is a time series plot of predicted data and lab data, predicted and lab averages as the centerline, and dashed lines for the lower and upper specifications. If the selected model has no points outside the specifications, or there are points outside the specification but there is an assignable cause, then Success Criteria #1 is passed. The LASSO chosen model passed this criteria for the customer validation.

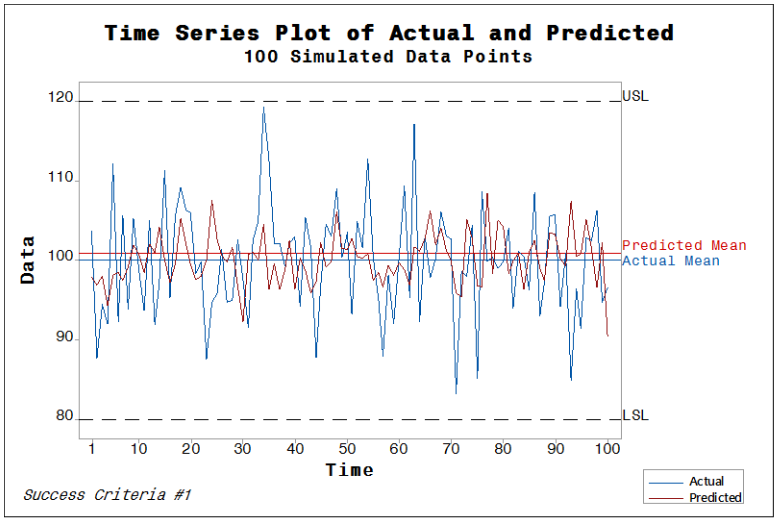


Figure 6. A time series plot of predicted data and lab data (and means) with historical specifications.

Success Criteria #2 is that the difference between predictions and the lab data is less than the historical calculation of the lower and upper specification. Figure 7 is a time series plot of (predicted – actual) data differences, the mean difference as the centerline, and lines for the lower and upper specification limits. If the selected model has no points outside the specifications, or there are points outside the specification but there is an assignable cause, then Success Criteria #2 is passed. The LASSO chosen model passed this criteria for the customer validation.

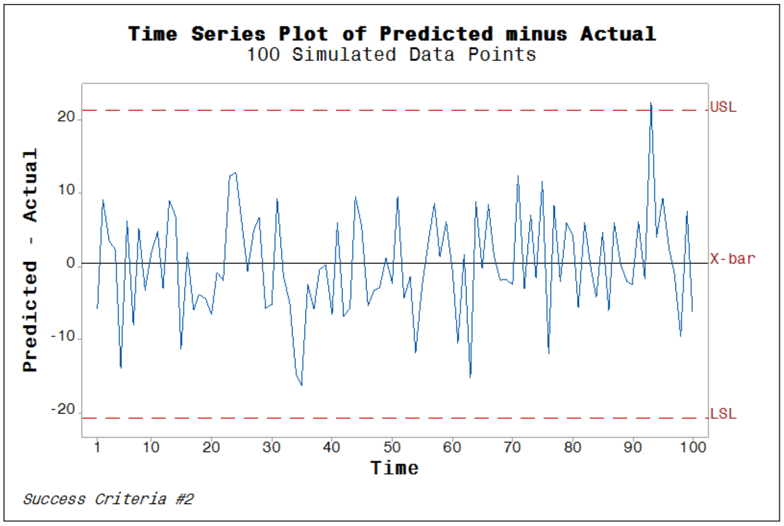


Figure 7. A time series plot of (predicted – actual) differences with mean difference (X-bar) and lower and upper specifications drawn.

The model chosen by LASSO also passed the other two customer specified success criteria.

PREDICTIVE MODEL USE

The LASSO selected predictive model passed all the customer success criteria. The quality, engineering and IT team members coded the model into the DCS computer. The predicted and actual product quality data is available online in real-time for the operators to use as process control tools. The operators have displays of the predicted values and actual lab values and the (predicted – actual) difference plots. The operators also use trend charts displayed to monitor the process data. The mill uses Parcview® as the display software of the real-time data. Figure 8 displays a trend chart of the real-time lab (blue) and predicted values (green). The predictive model average from the designed experiment is the centerline. The upper and lower limits are calculated based on historical data that preceded the current model.

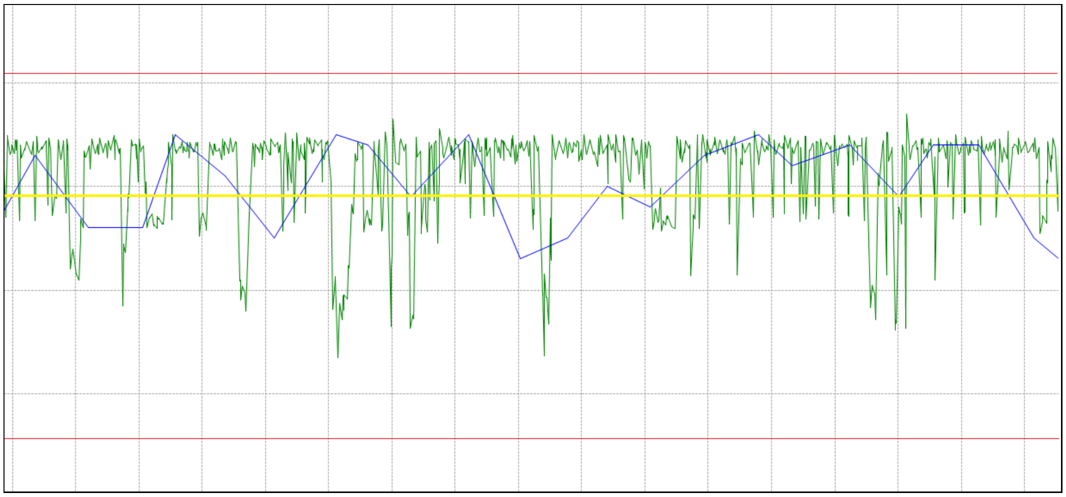


Figure 8. Trend chart of the real-time lab (blue) and predicted (green) values with centerline and historical limits for the predicted values.

UPDATING THE PREDICTIVE MODEL

The final topic of the predictive model use is to know when the model needs updating. The best way to understand if the model needs updating is to monitor the residuals.

Residual = actual lab value – predicted value

To do so, use a control chart on the residuals or a time series plot of the actual lab values and the predicted values to look for long term trends. The most common update is to change the intercept constant. There may be drift of the measurement device over time or a re-calibration.

If the model is impacted there are two ways to update it:

1. Use existing process and lab data.
2. Run a new designed experiment.

CONCLUSION

The LASSO using *k*-fold cross validation as implemented in the GLMSELECT procedure was successful in quickly choosing models for the mill's key quality measures. All LASSO selected models passed the customer prescribed "success criteria". Previous predictive model building exercises took significantly longer to complete. There are currently three predictive models developed with the LASSO that are being used successfully in real-time at the mill. The ability to use these predictive models to approve material release reduces the testing expenses by more than 66%. It also allows for immediate material release for delivery, which reduces inventory costs.

What worked to make the predictive model development a success?

- Have the human resources available to do the model development.
  - Engineers, operators, testers, statistician, IT, etc.
- Make sure what you are trying to predict is both measurable and controllable.
- Have online sensors and a data historian.
- Have a very detailed trial plan.
- Be willing to work with the customer as a teammate.
- Be flexible.
- Use existing data if you have it.

Remember that George Box said, "All models are wrong, but some are useful." The predictive models are not supposed to predict individual measurements. Instead, they are designed to predict the average of future measurements. The goal is for the predicted future average to be equal to the specified target (average) to ensure quality product.

REFERENCES

Bailey, M. 2010. “*Custom Designs for Experiments Course Notes”*. Cary, NC: SAS® Institute.

Cohen, R. A. 2006. “Introducing the GLMSELECT PROCEDURE for Model Selection”. SAS Institute Inc. 2006. Proceedings of the Thirty-first Annual SAS® Users Group International Conference. Cary, NC: SAS Institute Inc.

Draper, N. R. and Smith, H. 1998. *Applied Regression Analysis*. New York, NY: Wiley.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. 2004. “Least Angle Regression”. The Annals of Statistics 2004, Vol. 32, No. 2.

James, G., Witten, D., Hastie, T. and Tibshirani, R. 2013. *An Introduction to Statistical Learning: with Applications in R*. New York, NY: Springer.

Milliken, G. A. and Johnson, D. E. 2009. *Analysis of Messy Data: Volume 1, Designed Experiments*, 2nd ed. Boca Raton, FL: Chapman and Hall.

Myers, R. H., Montgomery, D. C. and Anderson-Cook, C. M. 2009. *Response Surface Methodology: Process and Product Optimization using Designed Experiments*, 3rd ed. Hoboken, NJ: Wiley.

Tibshirani, R. 1996. “Regression Shrinkage and Selection via the Lasso”. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 58, Issue 1.

Wu C.F.J. and Hamada M.S. 2009. *Experiments: Planning, Analysis and Optimization*, 2nd ed. Hoboken, NJ: Wiley.

Yuan, M., Joseph, V. R. and Lin, Y. 2007. “An Efficient Variable Selection Approach for Analyzing Designed Experiments”. Technometrics, November 2007, Vol. 49, No. 4.

ACKNOWLEDGMENTS

Jack Giovanini, PhD, for help with editing and content.
Kent Robarge, MS ChemE, for help with editing and provided the process schematic.
Dick Kleinknecht, PhD, for motivating the use of LASSO to analyze the designed experiment.
Chuck Miller, PhD, for help with editing.

RECOMMENDED READING

- *SAS®/STAT 13.2 User’s Guide*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jon Lindenauer
Weyerhaeuser Company
253-924-6672
jon.lindenauer@weyerhaeuser.com or jmlstat@comcast.net