

# A comparison of multiple linear regression and quantile regression for modeling the internal bond of medium density fiberboard

Timothy M. Young\*

Leslie B. Shaffer

Frank M. Guess\*

Halima Bensmail

Ramón V. León

---

## Abstract

Multiple linear regression (MLR) and quantile regression (QR) models were developed for the internal bond (IB) of medium density fiberboard (MDF). The data set that aligned the IB of MDF came from 184 independent variables that corresponded to online sensors. MLR models were developed for MDF product types that were distinguished by thickness in inches, i.e., 0.750-inch, 0.6875-inch, 0.625-inch, and 0.500-inch. A best model criterion was used with all possible subsets. QR models were developed for each product type for the most common independent variable of the MLR models for comparison. The adjusted coefficient of determination ( $R_a^2$ ) of the MLR models ranged from 72 percent with 53 degrees of freedom to 81 percent with 42 degrees of freedom. The Root Mean Square Errors (RMSE) ranged from 6.05 pounds per square inch (psi) to 6.23 psi; the maximum Variance Inflation Factor (VIF) was 5.6, and all residual patterns were homogeneous. A common independent variable for the 0.750-inch and 0.625-inch MLR models was "Refiner Resin Scavenger %." QR models for 0.750 inch and 0.625 inch indicate similar slopes for the median and average, with different slopes at the fifth and 95th percentiles. "Face Humidity" was a common independent variable for the 0.6875-inch and 0.500-inch MLR models. QR models for 0.6875-inch and 0.500-inch indicate different slopes for the median and average, and instability in IB in the outer fifth and 95th percentiles. The use of QR models to investigate the percentiles of the IB of MDF suggests significant opportunities for manufacturers for continuous improvement and cost savings.

---

The wood composites industry is undergoing unprecedented change in the forms of corporate divestitures and consolidation, real increases in the cost of raw material and energy, and extraordinary international competition. The forest products industry is an important contributor to the U.S. economy. In 2002, this sector contributed more than \$240 billion to the economy and employed more than one million Americans in 22,231 primary wood products manufacturing facilities (U.S. Census Bureau 2004). Sustaining business competitiveness by reducing costs and maintaining product quality will be essential for this industry. One of the challenges facing this industry is to develop a more advanced knowledge of the complex nature of process variables and quantify the causality between process variables and final product quality characteristics in the percentiles of the distribution. Information contained in the percentiles is a key measure for quality and safety concerns. This paper provides quantile regression (QR) statistical methods that can improve business competitiveness in the wood composites industry (Young and Guess 1994, 2002).

Some work has been initiated in data mining and predictive modeling of final product quality characteristics of forest products (Young 1997, Bernardy and Scherff 1997, 1998; Greubel 1999; Eriklsson et al. 2000, Young and Guess 2002,

---

The authors are, respectively, Research Associate Professor, Forest Products Center; former Graduate Research Assistant, Forest Products Center and Dept. of Statistics, Operations, and Management Sci. (currently Statistician with Eastman Chemical Company); and Professor, Associate Professor, and Associate Professor, Dept. of Statistics, Operations, and Management Sci.; all at the Univ. of Tennessee, Knoxville, Tennessee (tmyoung1@utk.edu, fgues@utk.edu, bensmail@utk.edu, rleon@utk.edu). This research was partially supported by The Univ. of Tennessee Agri. Expt. Sta. McIntire Stennis E112215 (MS-75); USDA Special Wood Utilization Grants R112219-150 and R112219-184. Funding was also provided by Univ. of Tennessee, Dept. of Statistics, Operations, and Management Sci. This paper was received for publication in May 2007. Article No. 10352.

\*Forest Products Society Member.

©Forest Products Society 2008.

Forest Prod. J. 58(4):39-48.

Young and Huber 2004, Clapp et al. 2007). Much work has been published on simulating process variables and using theoretical models to predict final product quality characteristics (Barnes 2001, Humphrey and Thoemen 2000, Shupe et al. 2001, Wu and Piao 1999, Xu 2000, Zombori et al. 2001). The use of quantile regression to investigate the percentiles of product quality for wood composites has not been documented in the literature.

A data set from a large-capacity North American manufacturer of medium density fiberboard (MDF)<sup>1</sup> was obtained in 2002. The data set aligned process measurements from online sensors with the Internal Bond (IB) analyzed during periodic destructive testing. For example, online sensor measurements were available for measuring press temperature, press closing time, resin content, moisture, weight, etc. The goal of any wood products manufacturer is to efficiently produce a high quality end product. To this end, it is imperative that the manufacturer has an advanced knowledge of the process and causality.

A common goal for statistical research is to investigate and quantify causality between independent variables ( $X$ ) and response variables ( $Y$ ) with a high level of scientific inference. In quantifying causality, de Mast and Trip (2007) note the important distinction between exploratory and confirmatory data analysis, which they attribute to Tukey's (1977) work. As Tukey (1977) pointed out, confirmatory data analysis is concerned with testing a prespecified hypothesis. The purpose of exploratory data analysis is hypothesis generation (de Mast and Trip 2007). This research was undertaken in the spirit of exploratory data analysis and hypothesis generation. A significant problem for MDF manufacturers is the quantification of known and unknown sources of variation. The objective of this research was to explore causality of sources of MDF process variation beyond the mean of the distribution of internal bond. The paper directly compares the use of Multiple Linear Regression (MLR) and Quantile Regression (QR) for modeling the IB of MDF. MLR and QR were used on the same MDF data set to model process variables and the process variables level of influence on IB. MLR develops models based on the mean of the response variable (e.g., IB), while QR develops models for any percentile of the response variable. Modeling beyond the mean of IB may greatly improve a MDF manufacturers understanding of the process. An improved understanding of process variables and the process variables' level of influence on IB can help MDF manufacturers identify and quantify unknown sources of process variation. Identifying and quantifying process variation can facilitate continuous improvement and improve competitiveness (Deming 1986, 1993).

As Mosteller and Tukey (1977) note in their influential text, as recently cited by Koenker (2005): "... the regression curve

gives a grand summary for the averages of the distributions corresponding to the set of  $X$ s ... and so regression often gives a rather incomplete picture. Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributions."

## Methods

MLR was used to study the relationship between various independent variables and the mean or average of the distribution for a response variable with an important goal of making useful predictions of the response variable. There is a plethora of literature on regression analysis and MLR, and many tomes are available on the method (Box and Cox 1964, Draper and Smith 1981, Neter et al. 1996, and Myers 1990, Kutner et al. 2004, etc.). For situations where the data are drawn from reasonably homogeneous populations and the response ( $Y$ ) is normally distributed, traditional methods such as MLR can yield insightful analyses. The usefulness of MLR can breakdown quickly if these stringent assumptions are not met. MLR has three important assumptions: 1) linearity of the coefficients; 2) normal or Gaussian distribution for the response errors ( $\epsilon$ ); and 3) the errors  $\epsilon$  have a common distribution. In many industrial settings when modeling a quality characteristic such as IB, these assumptions may not be valid.

QR is an approach that allows us to examine the behavior of the response variable ( $Y$ ) beyond its average of the Gaussian distribution, e.g., median (50th percentile), 10th percentile, 80th percentile, 90th percentile, etc. Examining the behavior of the regression curve for the response variable ( $Y$ ) for different quantiles with respect to the independent variables ( $X$ ) may result in very different conclusions relative to examining only the average of  $Y$ . In regard to the IB of MDF, examining the lower percentiles using QR may be more important for understanding IB failures (or very strong IBs) and be more beneficial for continuous improvement and cost savings.

## Relational database

An automated relational database was created by aligning real-time process sensor data with IB readings (Young and Guess 2002). The real-time process data were collected with Wonderware IndustrialSQL™ 8.0 ([www.wonderware.com](http://www.wonderware.com)). The readings were combined with IB by product type at the instant when a panel was extracted from the production line for testing. The process data were collected using a median value from the last 100 sensor values (e.g., for most of the 184 different sensor variables this represented a 2- to 3-minute time interval). The process data were collected and stored using IndustrialSQL. The lag times corresponding to the time required for the product to travel through the process from the point where a given parameter has an influence to the point where the panel was extracted for IB destructive testing were taken into account. A unique number (idnum) was generated when the panel was extracted from the process, and was later used to match process data with corresponding IB results.

When the IB results were matched with the process data, the combined data were recorded in two tables that appeared in a combined SQL database, i.e., a relational database of real-time sensor data and destructive test lab data. The real-time relational database was automatically updated as new lab samples were taken using Microsoft Transact SQL code with Microsoft SQL "Jobs" and "Stored Procedures."

<sup>1</sup> "Large-scale production of MDF began in the 1980s. MDF is an engineered wood product formed by breaking down softwood into wood fibers, often in a defibrator (i.e. "refiner"), combining it with wax and resin, and forming panels by applying high temperature and pressure ([http://en.wikipedia.org/wiki/Medium-density\\_fibreboard](http://en.wikipedia.org/wiki/Medium-density_fibreboard)). MDF has become one of the most popular composite materials in recent years. MDF is uniform, dense, smooth, and free of knots and grain patterns, and is an excellent substitute for solid wood in many applications. Its smooth surfaces also make MDF an excellent base for veneers and laminates. Builders use MDF in many capacities, such as in furniture, shelving, laminate flooring, decorative molding, and doors. MDF can be nailed, glued, screwed, stapled, or attached with dowels, making it a versatile product" ([www.wisegeek.com/what-is-mdf.htm](http://www.wisegeek.com/what-is-mdf.htm)).

The names used in this manuscript associated with the process variables for the online sensors were nondescriptive at the request of the manufacturer and given the terms of a legal confidentiality agreement. Definitions for the names of the process variables were not allowed under the terms of the legal confidentiality agreement.

### Classical linear regression

The classical first-order simple linear regression model has the form (Draper and Smith 1981),

$$Y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i, \quad [1]$$

where  $Y_i$  is the value of the response variable in the  $i$ th observation,

$\beta_0$  is the intercept parameter,

$\beta_1$  is a slope parameter,

$x_{1i}$  is the value of the independent variable in the  $i$ th observation,

$\varepsilon_i$  is a random error term of the  $i$ th observation with mean  $E(\varepsilon_i) = 0$  and variance  $\sigma^2\{\varepsilon_i\} = \sigma^2$ , with the error terms being independent and identically distributed,  $i = 1, \dots, n$ .

Most practitioners use multiple linear regression (MLR) first-order models of the form:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \dots + \beta_k x_{ki} + \varepsilon_i, \quad [2]$$

where  $Y_i$  is the value of the response variable in the  $i$ th observation,

$\beta_0$  is the intercept parameter,

$\beta_k$  is the slope parameter associated with the  $k$ th variable,

$X_{ki}$  is the  $k$ th independent variable associated with the  $i$ th observation,

$\varepsilon_i$  is a random error term with mean  $E(\varepsilon_i) = 0$

and variance  $\sigma^2\{\varepsilon_i\} = \sigma^2$ , with the error terms being independent and identically distributed,

$$i = 1, \dots, n.$$

The least squares method is a common method in simple regression and MLR and is used to find an affine function that best fits a given set of data.<sup>2</sup> Recall that a strength of the least squares method is that it minimizes the sum of the  $n$  squared errors (SSE) of the predicted values on the fitted line ( $\hat{y}_i$ ) and the observed value ( $y$ ).<sup>3</sup>

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad [3]$$

### Model building and best model criteria

*Model building.* — Model building using MLR is quite popular due to the refinement of user-friendly, inexpensive statistical software and real-time data warehousing. Many in

the forest products industry use MLR as a basic method for data mining. A popular model building method for MLR is “stepwise regression.” In this paper stepwise regression was used to develop first-order linear models of the IB for MDF.

Stepwise regression was introduced by Efron (1960). This method is an automated procedure used to select the most statistically significant variables from a large pool of explanatory variables. The method does not take into account industrial knowledge about the process, and therefore other variables of interest may be later added to the model if necessary. Three approaches can be used in stepwise regression: 1) backward elimination; 2) forward selection; and 3) mixed selection. The backward elimination method begins with the largest regression, using all variables, and subsequently reduces the number of variables in the equation until an acceptable model is developed (Draper and Smith 1981). The forward selection procedure attempts to achieve a similar conclusion working from the other direction, i.e., starting with one variable and inserting variables in turn until the regression is satisfactory (Draper and Smith 1981). The order of insertion is determined by using the partial correlation coefficient as a measure of the importance of variables not yet in the equation (Neter et al. 1996). The basic procedure is to select the most correlated independent variable ( $X$ ) with  $Y$  and find the first-order linear regression equation. This continues by finding the next most correlated independent variable ( $X$ ) with  $Y$ , and so forth. The overall regression is checked for significance; improvements in the  $R^2$  value and the partial F-values for all independent variables in the model were noted. The F-values are used for the F-test, which is used to calculate the one sided probability of the likelihood that two variances are different. The partial F-values are compared with an appropriate F percentage point, and the corresponding independent variables are retained or rejected from the model according to whether the test is significant or not significant. This continues until a suitable first-order linear regression equation is developed; see Kutner et al. (2004), Neter et al. (1996), and Myers (1990).

In stepwise regression it is important to note that the user specifies the probabilities ( $\alpha$ ) for an independent variable ( $X$ ) “to stay” and also the probabilities “to leave” the model. The mixed selection procedure is a combination of the aforementioned procedures. In this paper, the mixed stepwise regression procedure was used with a best model criteria.

*Best model criteria.* — There is much literature written on “Best Model Criteria” in model building using MLR. We use SAS® Business Intelligence and Analytics Software (www.sas.com). For our model we used the following seven criteria in selecting the best model of IB. The criteria include: 1) maximum adjusted  $R_a^2$ ; 2) minimum Akaike’s Information Criterion (AIC); 3) Variance Inflation Factor (VIF) < 10; 4) significance of  $p$ -value < 0.05 for selected independent variables; 5) residual pattern analysis; 6) absence of heteroscedasticity (i.e., equal variance of residuals); and 7) no bias in the residuals, i.e.,  $E(\varepsilon_i) = 0$ .

Adjusted  $R^2$ , or  $R_a^2$ , is a better measure for building models with the potential of a large number of independent variables than the Coefficient of Determination ( $R^2$ ).  $R^2$  will always increase as an additional independent variable is added to the model, where  $R_a^2$  will only increase if the residual sum of squares decreases.  $R_a^2$  minimizes the risk of “over-fitting” and penalizes for model saturation, i.e., the model is penalized

<sup>2</sup> An affine (from the Latin, affinis, “connected with”) subspace of a vector space (sometimes called a linear manifold) is a coset of a linear subspace. A linear subspace of a vector space is a subset that is closed under linear combinations, e.g., linear regression equation of a linear subspace (<http://mathworld.wolfram.com/AffineFunction.html>. 2006).

<sup>3</sup> There has been a dispute about who first discovered the method of least squares. It appears that it was discovered independently by Carl Friedrich Gauss (1777 to 1855) and Adrien Marie Legendre (1752-1833), that Gauss started using it before 1803 (he claimed in about 1795, but there is no corroboration of this earlier date), and that the first account was published by Legendre in 1805, see Draper and Smith (1981).

if additional independent variables do not reduce the residual sum of squares. The formula for  $R_a^2$  is:

$$R_a^2 = 1 - (1 - R^2) \left( \frac{n-1}{n-p-1} \right), \quad 0 \leq R_a^2 \leq 1 \quad [4]$$

where, 
$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{SSE}{SSTO}, \quad 0 \leq R^2 \leq 1 \quad [5]$$

The important AIC statistic is calculated as follows:

$$AIC = n \ln \left( \frac{SSE}{n} \right) + 2p, \quad [6]$$

where  $n$  is the number of observations, and  $p$  is the number of independent variables.

The goal is to balance model accuracy and complexity. This is achieved by finding the minimum value of AIC (Akaike 1974).

VIF is a diagnostic tool used to check the impact of multicollinearity in the MLR model. The VIF is calculated for each independent variable and is computed as follows:

$$(VIF_k) = (1 - R_k^2)^{-1} \quad [7]$$

where  $R_k^2$  is the coefficient of multiple determination for  $X_k$  when regressed on the remaining  $p - 2$  predictors in the model. High levels of multicollinearity ( $VIF > 10$ ) can falsely inflate the least squares estimates; therefore, lower VIF values are desired (Kutner et al. 2004).

Residual pattern analysis is key criteria for assessing model quality. Residuals should not have any pattern when plotted against  $Y$ , any  $X$ , or as a function of time. Heteroscedasticity is unequal variance of residuals and indicates that the model is ill-founded. Residual plots should not have any slope or bias with a  $E(\varepsilon_i) = 0$ .

*SAS code for mixed stepwise regression.* — When modeling manufacturing processes, it is important to consider the most recent data first, i.e., this data will be most informative for continuous improvement (Deming 1986, 1993). SAS code was used to develop the mixed stepwise regression MLR models for the four product types using the previously described Best Model Criteria. MLR models for all possible subsets were explored using the most recent data and then moving backward in time. Initial models were developed for the 50 most recent data records and additional models were developed for each additional record moving backward in time through the data. The aforementioned best model criteria were used in selecting the best model from the subsets provided by SAS.

### Quantile regression

QR is intended to offer a comprehensive strategy for completing the regression picture (Koenker 2005). It is different from the MLR approach in that it takes into account the differences in behavior a characteristic may have at different levels of the response variable by weighting the central tendency measure. Also, this method uses the median as the measure of central tendency rather than the mean. The nonparametric median statistic may offer additional insight in the analysis of

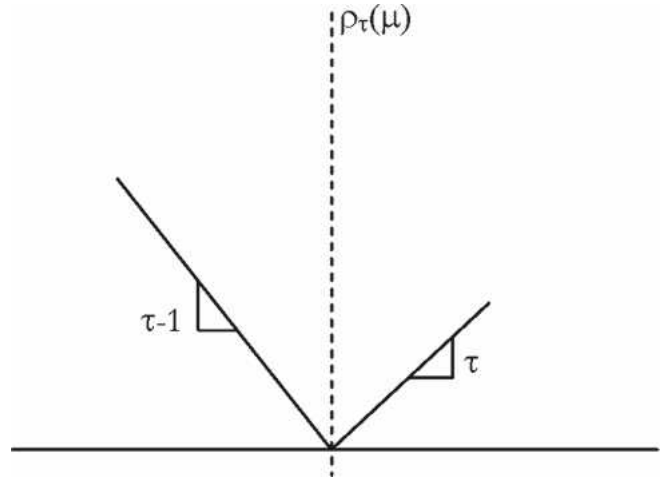


Figure 1. — Quantile regression  $\rho$  function.

data, especially when compared to the parametric mean or average statistic.

The QR model does not require the product characteristics or the response variable (IB in this study) to be normally distributed and does not have the other rigid assumptions associated with MLR. The first-order QR model has the form (Koenker 2005),

$$Q_{y_i}(\tau|x) = \beta_0 + \beta_i x_i + F_u^{-1}(\tau) \quad [8]$$

where,  $Q_{y_i}$  is the conditional value of the response variable given  $\tau$  in the  $i^{\text{th}}$  trial,  $\beta_0$  is the intercept,  $\beta_i$  is a parameter,  $\tau$  denotes the quantile (e.g.,  $\tau = 0.5$  for the median),  $x_i$  is the value of the independent variable in the  $i^{\text{th}}$  trial,  $F_u$  is the common distribution function (e.g., normal, Weibull, lognormal, other, etc.) of the error given  $\tau$ ,  $E(F_u^{-1}(\tau)) = 0$ , for  $i = 1, \dots, n$ , e.g.,  $F^{-1}(0.5)$  is the median or the 0.5 quantile.

Just as we can define the sample mean as the solution to the problem of minimizing a sum of squared residuals, we can define the median as the solution to the problem of minimizing a sum of absolute residuals (Koenker and Hallock 2001). The symmetry of the piecewise linear absolute value function implies that the minimization of the sum of absolute residuals must equate the number of positive and negative residuals, thus assuring that there were the same number of observations above and below the median (Koenker and Hallock 2001). Minimizing a sum of asymmetrically weighted absolute residuals yields the quantiles (Koenker and Hallock 2001). Solving

$$\min \sum_{i=1}^n \rho_{\tau}(y_i - \xi), \quad [9]$$

where the function  $\rho_{\tau}(\cdot)$ , e.g., in Eq. [9], is the tilted absolute value function appearing in Figure 1 that yields the  $\tau^{\text{th}}$  sample quantile as its solution (Koenker and Hallock 2001). To obtain an estimate of the conditional median function in quantile regression, we simply replace the scalar  $\xi$  in Eq. [9] by the parametric function  $\xi(x_i, \beta)$  and set  $\tau$  to  $1/2$ .<sup>4</sup> To obtain estimates

<sup>4</sup> Variants of this idea were proposed in the mid-eighteenth century by Boscovich and subsequently investigated by Laplace and Edgeworth (Koenker and Hallock 2001).

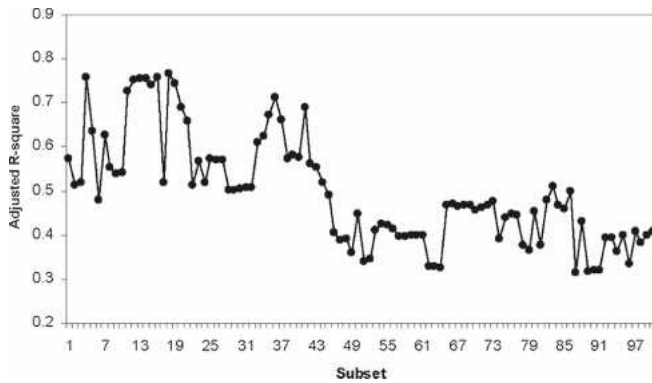


Figure 2. — Adjusted  $R^2$  for all possible subsets explored for 0.750-inch.

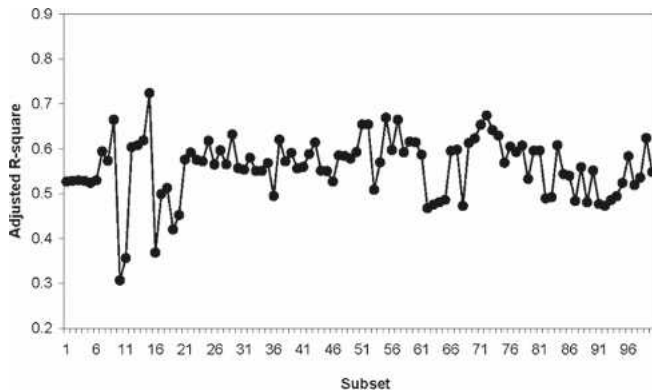


Figure 3. — Adjusted  $R^2$  for all possible subsets explored for 0.625-inch.

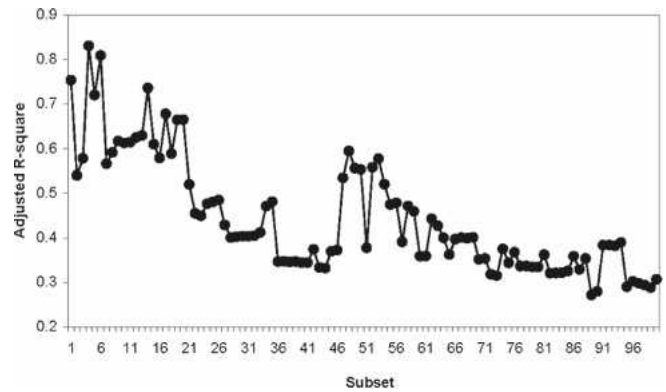


Figure 4. — Adjusted  $R^2$  for all possible subsets explored for 0.6875-inch.

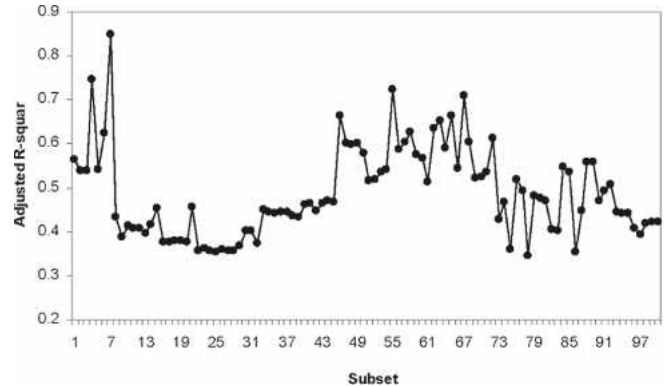


Figure 5. — Adjusted  $R^2$  for all possible subsets explored for 0.500-inch.

of the other conditional quantile functions, replace absolute values by  $\rho_\tau(\cdot)$ , e.g., Eq. [9], and solve

$$\hat{\beta}(\tau) = \min \sum_{i=1}^n \rho_\tau(y_i - \xi(x_i, \beta)). \quad [10]$$

For any quantile  $\tau \in (0,1)$ . The quantity  $\hat{\beta}(\tau)$  is called the  $\tau^{\text{th}}$  regression quantile.

### Results and discussion

The internal bonds of four different product types of MDF were analyzed. Each product type represents a different board thickness in inches (i.e., 0.750-inch, 0.625-inch, 0.6875-inch, and 0.500-inch). All possible subset MLR models were explored for the four product types using  $R_a^2$  as a key indicator for determining the best subset model (Figures 2, 3, 4, and 5). The  $R_a^2$  for all possible subsets was an indicator of a MDF manufacturer's stability in reproducing product quality from one production run to the next, i.e., product types where the  $R_a^2$  changes slowly as more records were added moving back in time may indicate less volatility in IB between production runs, and also that changes in processes occur less frequently between production runs of the product type. Once acceptable MLR models were obtained (i.e., using the best model criteria), commonalities in the independent variables were explored among the four product types.

#### Product types 0.750-inch and 0.625-inch

For the 0.750-inch product type a MLR model was developed with an  $R_a^2$  of 75 percent, 50 degrees of freedom and 11

parameters. The RMSE of the model was 7.69 psi and the maximum VIF for any independent variable was 5.03. Residual patterns for the MLR model were homogeneous (Table 1). Recall the RMSE estimates the SD of the residual error, which is the square root of the MSE which is the SSE divided by the degrees of freedom.

For the 0.625-inch product type a MLR model was developed with an  $R_a^2$  of 72 percent, 53 degrees of freedom and 11 parameters. The RMSE of the model was 6.05 psi and the maximum VIF for any independent variable was 5.60. Residual patterns for the MLR model were homogeneous (Table 1).

Common independent variables for the 0.750-inch and 0.625-inch MLR models were highlighted as bold in Table 1. "Refiner Resin Scavenger %" and "Core Water to Wood" were common for both 0.750-inch and 0.625-inch product types. It was surprising to see the scaled estimates for "Refiner Resin Scavenger %" differ in sign for each product type.<sup>5</sup> The "Refiner Resin Scavenger %" has a negative scaled estimate of approximately -9.12 psi on IB for 0.750-inch while the "Refiner Resin Scavenger %" has a positive scaled estimate of approximately 8.40 psi on IB for 0.625-inch. This may indicate that "Refiner Resin Scavenger %" was an important

<sup>5</sup> Scaled estimate is a helpful statistic in MLR models in that it illustrates the relative influence of independent variables on the response variable. The scaled estimate is the influence that an independent variable has on the response variable when the independent variable moves one-half its range used in the model.

Table 1. — MLR models for product types 0.750-inch and 0.625-inch.

0.750-inch			0.625-inch		
Parameters	Scaled estimate	p-value	Parameters	Scaled estimate	p-value
Face MDF temperature	-12.565	<0.0001	Shavings raw weight	-15.872	<0.0001
Dryer S fiber moisture	-10.906	<0.0001	Refiner resin scavenger %	8.396	0.0008
Refiner resin scavenger %	-9.118	<0.0001	Core grinding steam flow	12.720	<0.0001
Core dryer outlet temperature	18.498	0.0092	Core resin to wood %	22.473	<0.0001
Press position time	19.926	<0.0001	Dryer mass flow	10.642	<0.0001
Dryer 1 fan current	23.662	<0.0001	Resin water tank temperature	-21.556	<0.0001
Dryer 2 fan current	-25.384	<0.0001	Core refiner feeder screw speed	4.868	0.0110
Refiner S chip level	10.666	<0.0001	Core water to wood	-10.872	0.0077
Refiner S feeder screw speed	9.294	0.0017	Face humidifier temperature	13.583	<0.0001
Core water to wood	-21.043	<0.0001	Relative ambient humidity	5.858	0.0100
ESP milliamps	-11.714	<0.0001	Weight actual	12.205	<0.0001

Important regression statistics			
$R_a^{23}$	0.751646	$R_a^2$	0.723694
d.f. <sup>4</sup>	50	d.f.	53
P <sup>5</sup>	11	P	11
VIF <sub>max</sub> <sup>6</sup>	5.0315819	VIF <sub>max</sub>	5.603058
RMSE <sup>7</sup>	7.697272	RMSE	6.051464
Residual pattern	Homogeneous	Residual pattern	Homogeneous

<sup>3</sup>Adjusted coefficient of determination.

<sup>4</sup>Degrees of freedom.

<sup>5</sup>Number of explanatory variables.

<sup>6</sup>Maximum variance inflation factor.

<sup>7</sup>Root mean square error.

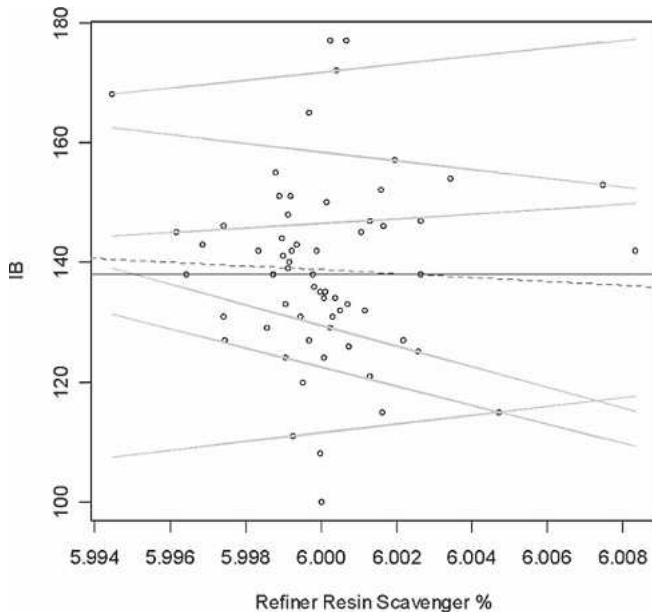


Figure 6. — Comparison of MLR fit (dashed line) with median (dark line) and other percentile fits (from bottom to top: 5th, 10th, 25th, 75th, 90th, and 95th) for 0.750-inch product type.

source of variability between the two product types that the manufacturer needs to further investigate.

“Core Water to Wood” has a large scaled estimate for both product types and has a negative influence on IB. The influence of “Core Water to Wood” as measured by the scaled estimate was -21.04 psi for 0.750-inch and -10.87 psi for 0.625-inch. This may reflect a difference in scale for this process variable as related to the refining process for different

product types that have varying throughput levels at the refiner.

To examine the influence of “Refiner Resin Scavenger %” beyond the mean effect on IB, QR was explored for this common parameter for both 0.750-inch and 0.625-inch.<sup>6</sup> We find that the influence of “Refiner Resin Scavenger %” on the lower percentiles of IB was quite different than the midrange and higher percentiles (Figs. 6 and 7). The dashed line represents the MLR fit, the solid dark line represents the median fit, and the gray lines correspond to the 5th, 10th, 25th, 75th, 90th, and 95th percentiles, respectively. For the 0.750-inch product type (Fig. 6), the slopes of the percentiles were quite different depending on percentile. The median and average have similar slopes. The 5th percentile (possible IB failures) and 95th percentile (extreme IB strength) behave quite differently than the inner percentiles. This may be helpful to a MDF producer in analyzing occurrences of IB failures, i.e., Why does IB decrease at a faster rate for the lower percentiles? What were the other operational settings

and factors occurring during these events?

For the 0.625-inch product type (Fig. 7), the slopes of the percentiles were extremely different depending on percentile and on scale of the level of “Refiner Resin Scavenger %.” The median and average have similar slopes. However, for percentiles above the 50th percentile (median) the effect of “Refiner Resin Scavenger %” has a stronger positive influence on IB the higher the percentile. For percentiles below the 50th percentile (median) the effect of “Refiner Resin Scavenger %” has a stronger negative influence on IB the higher the percentile. This may indicate that other factors were influencing IB in concert with “Refiner Resin Scavenger %” or that the quality of the “Refiner Resin Scavenger %” itself was varying. The QR analysis for the common parameter “Refiner Resin Scavenger %” indicates opportunities for additional root cause investigation by the manufacturer in sources of variability in “Refiner Resin Scavenger %” that influence IB.

Although only one independent variable was used for illustration purposes, the quantile regression algorithm in R can also be applied to multiple independent variable models. Further analysis was conducted to examine the differences between the MLR and QR median fits. For the 0.750-inch product type (Table 2), the largest discrepancies between coefficients occur in “Dryer 1 Fan Current,” “Dryer 2 Fan Current,” and “Core Water to Wood.” The percent differences were 39.84 percent, 34.37 percent, and 28.2 percent, respectively.

<sup>6</sup> It is important to note that multiple parameter models can be built using quantile regression, but for the purposes of illustration we chose to look only at the single parameter case.

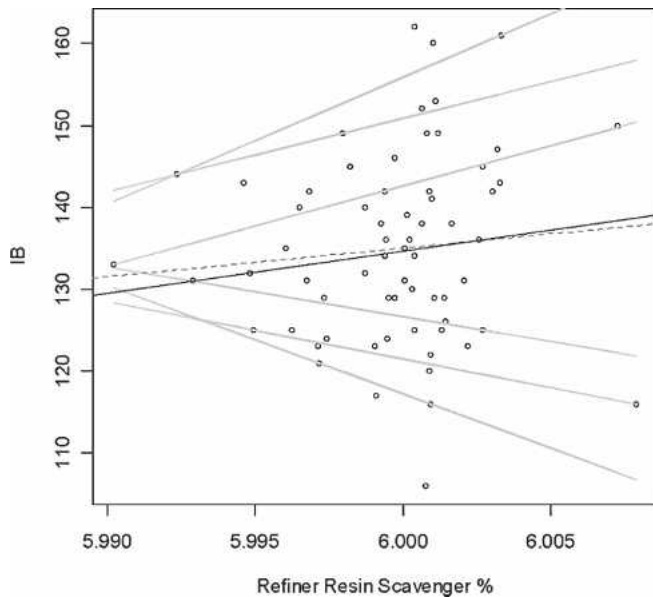


Figure 7. — Comparison of MLR fit (dashed line) with median (dark line) and other percentile fits (from bottom to top: 5th, 10th, 25th, 75th, 90th, and 95th) for 0.625-inch product type.

Table 2. — MLR and QR models for product type 0.750-inch.

0.750-inch Variables	Coefficients			
	MLR average	QR median	QR 10th percentile	QR 90th percentile
Intercept	40264.84	34655.89	40679.39	44452.38
Face mdf temperature	-0.27	-0.27	-0.33	-0.09
Dryer S fiber moisture	-5.10	-4.87	-5.76	-2.33
Refiner resin scavenger percent	-1314.71	-1535.49	-1488.00	-1373.25
Core dryer outlet temperature	1.91	1.63	1.97	1.46
Press position time	1.96	2.21	2.02	1.93
Dryer 1 fan current	75.06	53.67	78.09	95.56
Dryer 2 fan current	-65.80	-48.93	-67.03	-77.23
Refiner S chip level	4.00	3.16	3.37	6.14
Refiner S feeder screw speed	0.31	0.32	0.31	0.33
Core water to wood	-835.05	-651.32	-825.34	-957.74
ESP milliamps	-0.16	-0.17	-0.15	-0.19

For the 0.625-inch product type (Table 3), the largest discrepancies between coefficients occur in “Shavings Raw Weight,” “Relative Ambient Humidity,” and “Weight Actual.” The percent discrepancies were 42.96 percent, 16.16 percent, and 12.78 percent, respectively. These discrepancies reflect significant differences between modeling the mean and the median (50th percentile) of IB. These differences may illustrate the risk of incorrect decision-making about process variables that influence the mean of IB when the distribution is not Gaussian. Incorrect decisions lead to higher operating targets, unexpected IB failures and ultimately higher overall production costs. Further analysis could be conducted for other IB quantiles that may be invaluable to the producer for understanding low or failing IBs. A comparison of the 10th and 90th percentiles of the coefficients (Tables 2 and 3) may also give a good method for the practitioner on the relative

comparisons of the influence of a process variable on IB. The discrepancies in coefficients highlight the importance of examining the percentiles of a distribution.

### Product types 0.6875-inch and 0.500-inch

For 0.6875-inch a MLR model was developed with an  $R_a^2$  of 81 percent, 42 degrees of freedom and 13 parameters. The RMSE of the model was 6.23 psi, and the maximum VIF for any independent variable was 4.54. Residual patterns for the MLR model were homogeneous (Table 4).

For 0.500-inch a MLR model was developed with an  $R_a^2$  of 75 percent, 43 degrees of freedom, and 10 parameters. The RMSE of the model was 6.57 psi, and the maximum VIF for any independent variable was 5.55. Residual patterns for the MLR model were homogeneous (Table 4).

“Face Humidity” was the common independent variable for both 0.6875-inch and 0.500-inch MLR models (Table 4). It was surprising to see the scaled estimates for “Face Humidity” differed in sign for each product type. The “Face Humidity” has a negative scaled estimate of -10.02 psi on IB for 0.6875-inch while the “Face Humidity” has a positive scaled estimate of 4.81 psi on IB for 0.500-inch. This may signify that “Face Humidity” was an important source of variability acting on IB that the manufacturer needs to investigate.

To examine the influence of “Face Humidity” beyond the mean effect on IB, QR was explored for this common parameter for both 0.6875-inch and 0.500-inch. The average and median fits for 0.6875-inch for “Face Humidity” have different slopes, which may indicate lack of normality in the response variable IB. We found that influence of “Face Humidity” on the outer 5th and 95th percentiles of IB was quite different than the inner percentiles (Fig. 8). This may indicate more volatility in IB for the MDF producer for this product type in the presence of changes in “Face Humidity.” For the 0.500-inch product type (Fig. 9), the slopes of the IB percentiles were very similar for all of the percentiles for “Face Humidity.” The median and average have different scales, which may also indicate non-normality in the response variable IB. The QR analysis for 0.500-inch may indicate that this product type has less volatility in IB in the presence of changes in “Face Humidity” when compared to the 0.6875-inch product type. It may also indicate that the product was easier to make between production runs in the presence of changes in “Face Humidity.” The QR models for “Face Humidity” may reveal an opportunity for further root cause analysis by the manufacturer.

Although only one independent variable was used for illustration purposes, the quantile regression algorithm in R can also be applied to multiple independent variable models. Further analysis was conducted to examine the differences between the MLR and the QR median, 10th and 90th percentile fits. For the 0.6875-inch product type (Table 5), the largest discrepancies between the coefficients of median and average fits occur in “Face Humidifier Temperature,” “Core Scavenger Resin Flow,” and “Dryer Mass Flow.” The percent discrepancies were 66.53 percent, 23.16 percent, and 16.14 percent, respectively. For the 0.500-inch product type (Table 6), the largest discrepancies between the coefficients of median and average fits occur in “Face Humidity,” “Mat Shave Off Target,” and “Refiner S Steam Flow.” The percent discrepancies were 36.58 percent, 22.39 percent, and 15.60 percent, respectively. A comparison of the 10th and 90th percentiles (Tables 5 and 6) of the coefficients gives a good method for

Table 3. — MLR and QR models for product type 0.625-inch.

Variables	Coefficients			
	MLR Average	QR Median	QR 10th percentile	QR 90th percentile
Intercept	-1029.56	-2063.15	1896.63	-1745.51
Shavings raw weight	-1.55	-1.09	-1.38	-1.64
Refiner resin scavenger %	949.74	1084.13	588.90	889.98
Core grinding steam flow	0.34	0.37	0.38	0.28
Core resin to wood %	12.03	10.73	13.25	10.17
Dryer mass flow	0.68	0.76	0.65	0.67
Resin water tank temperature	-1.69	-1.81	-1.49	-2.01
Core refiner screw speed	0.14	0.14	0.26	0.04
Core water to wood	-133.48	-127.01	-150.40	-105.60
Face humidifier temperature	1.22	1.31	0.97	1.80
Relative ambient humidity	1.22	1.42	0.56	2.39
Weight actual	157.15	139.34	130.54	130.00

Table 4. — MLR models for product types 0.6875-inch and 0.500-inch.

0.6875-inch			0.500-inch		
Parameters	Scaled estimate	p-value	Parameters	Scaled estimate	p-value
Face scavenger resin %	25.479	<0.0001	Core total weight	-5.191	0.0112
Dryer mass flow	-8.192	0.0005	Mat shave off target	6.823	0.0020
Core humidifier temperature	-10.683	0.0037	Press preposition time	10.060	0.0015
Face fiber mat moisture	26.949	<0.0001	Weight target	7.938	0.0194
Mat shave off level	-15.408	<0.0001	Core blow line pressure	19.091	<0.0001
Refiner S chip level	14.655	<0.0001	Face digester pressure	-9.494	0.0004
Refiner S grinding steam flow	21.066	<0.0001	Core resin pressure	-11.273	0.0013
Refiner S screw speed	-5.873	0.0030	Refiner s steam flow	-7.452	0.0078
Core scavenger resin flow	-6.914	0.0225	Core refiner screw speed	-21.777	<0.0001
Dryer ESP outlet temperature	-13.138	<0.0001	Face humidity	4.811	0.0460
Face humidity	-10.016	0.0031			
Press open time	5.471	0.0056			
Face humidifier temperature	19.560	<0.0001			
Important Regression Statistics					
$R_a^2$	0.808614		$R_a^2$	0.747666	
d.f.	42		d.f.	43	
P	13		P	10	
VIF <sub>max</sub>	4.5371586		VIF <sub>max</sub>	5.5493187	
RMSE	6.233895		RMSE	6.573086	
Residual pattern	Homogeneous		Residual pattern	Homogeneous	

relative comparisons of the influence of a process variable on IB. The discrepancies in coefficients highlight the importance of examining the percentiles of a distribution.

A significant finding of the research was the identification of similar regressors across all product types for both the MLR and QR models of the median. Results suggest that the regressors “Core Scavenger Resin %,” “Face Humidity,” and “Core Water to Wood” require further root cause investigation. The variable “Core Scavenger Resin %” was related to the application rate of scavenger resin and its influence on IB which was not contrary to the literature (Maloney 1977 and Suchsland and Woodson 1986). However, “Face Humidity” was related to the amount of humidity in the face fiber layer of MDF during mat formation, and “Core Water to Wood” was an indicator of the amount of H<sub>2</sub>O added during the

defibrillation process of wood to fiber. The significant influence of both of these process variables in the MLR and QR models has not been previously documented in the literature. This finding further strengthens the justification for conducting more exploratory research when investigating sources of variation in the MDF manufacturing process.

Further hypothesis-based research using a designed experiment would provide additional quantification of the causality of these sources of variation relative to IB. Even though industrial experimentation is difficult given the many possible nuisance sources of variation and potential measurement error, the inductive research described in this paper provides insight for hypothesis generation, i.e., hypothesis tests and deductive scientific reasoning may provide more detailed scientific inference of causality.

The exploratory analysis of this research further highlighted that a narrow-view or focus only on the mean of the distribution may lead to incorrect conclusions, operational inefficiency and ultimately higher cost of manufactured product. Further analysis could also be conducted to examine each quantile (e.g., 1st, 5th, 50th, 99th, etc.) with respect to similar variables, e.g., “Core Scavenger Resin %,” “Face Humidity,” and “Core Water to Wood.” A more detailed examination of each quantile using a designed experiment may also provide additional validation of important sources of variation for the manufacturer.

## Conclusions

Multiple Linear Regression models (MLR) and Quantile Regression (QR) models were developed and compared for the Internal Bond (IB) of Medium Density Fiberboard (MDF). The models were developed from a manufacturing data set for a North American MDF producer. Models were developed for MDF product types that were distinguished by thickness in inches, i.e., 0.750 inch, 0.6875 inch, 0.625 inch, and 0.500 inch. MLR models have stringent assumptions and investigate causality of the mean of the response. QR models do not have the stringent assumptions and limitations of MLR and allow for the investigation of causality beyond the mean for any percentile of the response (e.g., 1st, 25th, 50th, 80th, 99th, etc.)

Common independent variables or regressors for the 0.750-inch and 0.625-inch MLR models were “Refiner Resin Scavenger %” and “Core Water to Wood.” The scaled estimates for “Refiner Resin Scavenger %” differed in sign for

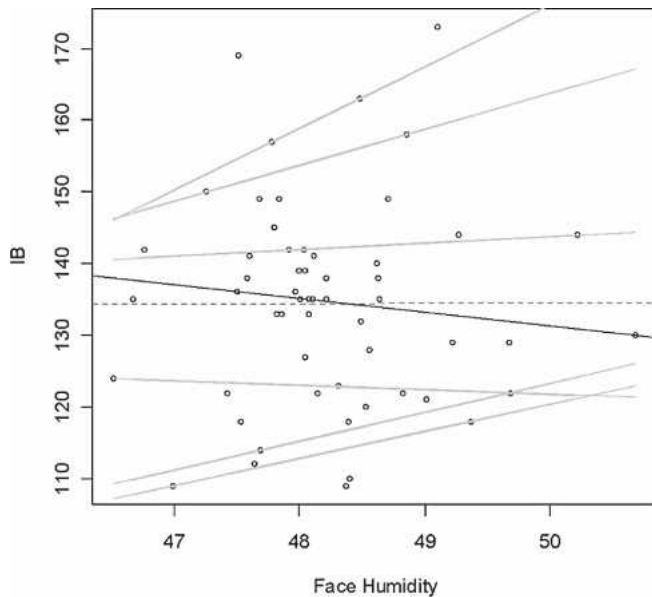


Figure 8. — Comparison of MLR fit (dashed line) with median (dark line) and other percentile fits (from bottom to top: 5th, 10th, 25th, 75th, 90th, and 95th) for 0.6875-inch product type.

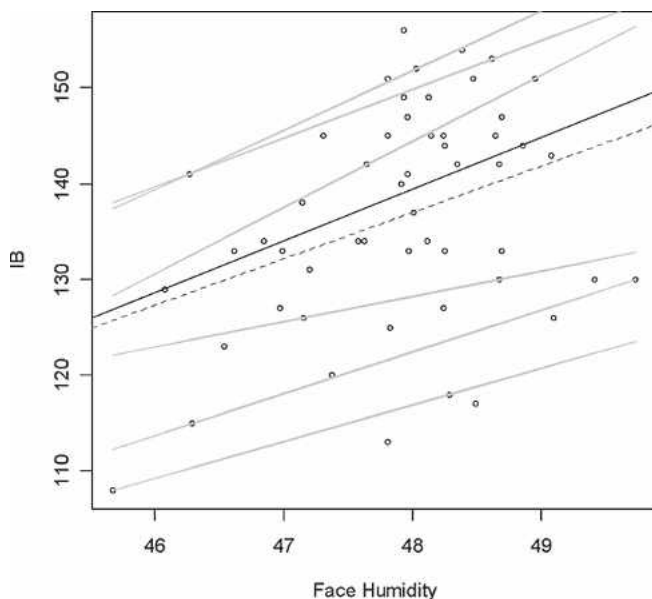


Figure 9. — Comparison of MLR fit (dashed line) with median (dark line) and other percentile fits (from bottom to top: 5th, 10th, 25th, 75th, 90th, and 95th) for 0.500-inch product type.

each product type. The influence of “Refiner Resin Scavenger %” on the lower percentiles of IB was quite different than the midrange and higher percentiles. The research identified opportunities for additional investigation using designed experimentation on the “Refiner Resin Scavenger %” source of variation.

“Face Humidity” was common for both 0.6875-inch and 0.500-inch product types. The scaled estimates for “Face Humidity” differ in sign for each product type and the average and median fits for 0.6875-inch for “Face Humidity” have different slopes, suggesting a lack of normality in the response variable IB. The QR analysis for 0.500-inch may indicate that this product type has less volatility in IB in the

Table 5. — MLR and QR models for product type 0.6875-inch.

0.6875-inch Variables	Coefficients			
	MLR average	QR median	QR 10th percentile	QR 90th percentile
Intercept	-1231.75	-1556.92	-692.31	-1693.05
Face scavenger resin %	280.75	314.06	227.50	345.93
Dryer mass flow	-0.61	-0.53	-0.69	-0.85
Core humidifier temperature	-1.54	-1.57	-1.86	-2.16
Face fiber mat moisture	24.50	22.78	27.19	17.18
Mat shave off level	-16.18	-15.63	-17.63	-16.17
Refiner S chip level	1.91	1.84	1.68	2.24
Refiner S grinding steam flow	0.04	0.04	0.04	0.03
Refiner S screw speed	-0.18	-0.19	-0.21	-0.09
Core scavenger resin flow	-3.76	-3.05	-5.72	-0.29
Dryer ESP outlet temperature	-0.68	-0.63	-0.74	-0.73
Face humidity	-4.81	-4.84	-6.10	-2.38
Press open time	0.34	0.30	0.45	0.26
Face humidifier temperature	2.38	3.96	2.93	4.29

Table 6. — MLR and QR models for product type 0.500-inch.

0.500-inch Variables	Coefficients			
	MLR average	QR median	QR 10th percentile	QR 90th percentile
Intercept	-225.75	-173.05	-305.28	-86.06
Core total weight	-0.07	-0.08	-0.03	-0.09
Mat shave off target	9.52	7.78	9.45	9.95
Press preposition time	0.93	0.90	1.36	0.60
Weight target	158.76	153.68	219.71	56.18
Core blow line pressure	1.65	1.71	1.69	0.82
Face digester pressure	-2.06	-2.21	-2.18	-1.72
Core resin pressure	-0.12	-0.13	-0.13	-0.07
Refiner S steam flow	-0.01	-0.01	-0.01	-0.002
Core refiner screw speed	-0.55	-0.55	-0.41	-0.36
Face humidity	2.37	1.74	0.31	4.58

presence of changes in “Face Humidity” when compared to the 0.6875-inch product type. This may suggest the product 0.500-inch was easier to manufacture across production runs in the presence of variability in “Face Humidity” The discrepancies in regressors and the magnitude of scaled estimates further highlighted limitations of investigating only the mean of the distribution.

The research questions generated from the study relate to identifying significant and common sources of variation in MDF manufacture beyond the mean of the distribution of the response variable IB. Important research questions relate to the relative influence of “Core Scavenger Resin %,” “Face Humidity,” and “Core Water to Wood” in the manufacture of the IB of MDF. Additional designed experimentation in an industrial setting may lead to additional scientific inference and quantification of causality for these process variables.

Quantile regression methods can improve forest products manufacturers’ knowledge of process variation. Improved

knowledge of process variation facilitates variation reduction and costs savings, both vital for the long-term sustained business competitiveness of this important forest products industry.

### Literature cited

- Akaike, H. 1974. Factor analysis and AIC. *Psychometrika* 52:317-332.
- Barnes, D. 2001. A model of the effect of strand length and strand thickness on the strength properties of oriented wood composites. *Forest Prod. J.* 51(9):36-46.
- Bernardy, G. and B. Scherff. 1998. Saving costs with process control, engineering and statistical process optimization. *In: Proc. 2nd European Panel Prod. Symp. (EPPS)*. Llandudno, Wales.
- \_\_\_\_\_ and \_\_\_\_\_. 1999. Process modeling provides on-line quality control and process optimization in particle and fiberboard production. ATR Industrie-Elektronik GmbH and Co., Viersen, Germany.
- Clapp, N.E., Jr., T.M. Young, and F.M. Guess. 2007. Predictive modeling the internal bond of medium density fiberboard using principal component analysis. *Forest Prod. J.* (In print).
- de Mast, J. and A. Trip. 2007. Exploratory data analysis in quality-improvement projects. *J. Qual. Tech.* 39(4):301-311.
- Deming, W.E. 1986. *Out of the Crisis*. Massachusetts Inst. of Tech.'s Center for Advanced Engineering Design, Cambridge, Massachusetts.
- \_\_\_\_\_. 1993. *The New Economics*. Massachusetts Inst. of Tech.'s Center for Advanced Engineering Design, Cambridge, Massachusetts.
- Draper, N.R. and H. Smith. 1981. *Applied Regression Analysis*, 2nd Ed. John Wiley and Sons, Inc. New York.
- Efroymson, M.A. 1960. Multiple Regression Analysis. *In: Mathematical Methods for Digital Computers*. Ralston, A., and H.S. Wilf, Eds. John Wiley and Sons, Inc. New York.
- Erilsson, L., P. Hagberg, E. Johansson, S. Rannar, O. Whelehan, A. Astrom, and T. Lindgren. 2001. Multivariate process monitoring of a newsprint mill. Application to modeling and predicting COD load resulting from de-inking of recycled paper. *J. of Chemometrics* 15(4): 337-352.
- Greubel, D. 1999. Practical experiences with a process simulation model in particleboard and MDF production. *In: Proc. 2nd European Wood-Based Panel Symp.*, Hanover, Germany.
- Humphrey, P.E. and H. Thomen. 2000. The continuous pressing of wood-based composites: A simulation model, input data and typical results. *In: Proc. Pacific Rim Bio-Based Composites Conf.*, Australian National Univ. Press., Canberra, Australia. pp. 303-311.
- Koenker, R. 2005. *Quantile Regression*. Cambridge Univ. Press, New York.
- \_\_\_\_\_ and K.F. Hallock. 2001. Quantile regression. *J. Econ. Perspect.* 15(4):143-156.
- Kutner, M.H., C.J. Nachtsheim, and J. Neter. 2004. *Applied Linear Regression Models*, 4th Ed. McGraw-Hill Irwin, Inc. Boston.
- Maloney, T.M. 1977. *Modern particleboard and dry-process fiberboard manufacturing*. Miller Freeman, Inc. San Francisco.
- Mosteller, F. and J. Tukey. 1977. *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley, Reading, Massachusetts.
- Myers, R.H. 1990. *Classical and Modern Regression with Applications*. PWS-Kent Publishing Company. Boston.
- Neter, J., M.H. Kutner, C.J. Nachtsheim, and W. Wasserman. 1996. *Applied Linear Regression Models*, 3rd Ed. Irwin, Inc. Chicago.
- Shupe, T.F., C.Y. Price, and E.W. Price. 2001. Flake orientation effects on physical and mechanical properties of sweetgum flakeboard. *Forest Prod. J.* 51(9):38-43.
- Suchsland, O. and G.E. Woodson. 1986. *Fiberboard manufacturing practices in the United States*. Agri. Handbook No. 640. USDA Forest Serv., Government Printing Office, Washington, D.C.
- Tukey, J.W. 1977. *Exploratory Data Analysis*. Addison-Wesley, Reading, Massachusetts.
- U.S. Census Bureau. 2004. 2002 Economic Census. Table 1, Advance summary statistics for the United States 2002 NAICS basis. Washington, D.C. [www.census.gov/econ/census02/advance/TABLE!HTM](http://www.census.gov/econ/census02/advance/TABLE!HTM)
- Wu, Q. and C. Piao. 1999. Thickness swelling and its relationship to internal bond strength loss of commercial oriented strandboard. *Forest Prod. J.* 49(7/8):50-55.
- Xu, W. 2000. Influence of percent alignment and shelling ratio on linear expansion of oriented strandboard: A model investigation. *Forest Prod. J.* 50(7/8):88-98.
- Young, T.M. 1997. Process improvement through "real-time" statistical process control in MDF manufacture. *In: Proc. Process and Business Technologies for the Forest Products Industry*. Forest Prod. Soc. Proc. No. 7281. pp. 50-51.
- \_\_\_\_\_ and C.W. Huber. 2004. Predictive modeling of the physical properties of wood composites using genetic algorithms with considerations for distributed data fusion. *In: Proc. of the 38th Inter. Particleboard/Composite Materials Symp.* Washington State Univ., Pullman, Washington. pp. 145-153.
- \_\_\_\_\_ and F.M. Guess. 1994. Reliability processes and structures. *Microelectron. Reliab.* 34:1107-1119.
- \_\_\_\_\_ and \_\_\_\_\_. 2002. Developing and mining higher quality information in automated relational databases for forest product manufacture. *Inter. J. of Reliability and Application* 3(4):155-164.
- Zombori, B.G., F.A. Kamke, and L.T. Watson. 2001. Simulation of the mat formation process. *Wood and Fiber Sci.* 33(4):564-579.