

**Examining Regression Analysis Beyond the Mean of the  
Distribution using Quantile Regression**

---

**A Case Study of Modeling the Internal Bond  
of Medium Density Fiberboard  
using Multiple Linear Regression and Quantile Regression  
with an Example of Reliability Methods using R Software**

**Master's Thesis  
Presented for the  
Master of Science Degree  
The University of Tennessee, Knoxville**

**Leslie Brooke Shaffer  
August 2007**

## DEDICATION

This master's thesis is dedicated to everyone who has believed in me throughout my life – especially recently – and to those who have shown me unconditional and unfaltering love. My parents have been a constant source of love and support since the day I was born. There is no doubt that their influence has shaped my life tremendously. My friends – my very best friends – Jack and Ginger Holder are like my second family and are a source of constant support; in some cases both emotionally and physically. This is also dedicated to Rachel Jackson whose homemade curry has tremendous healing properties and to Becki Stratton who once rescued me from a rabid tick and can always make me laugh. Also, my college roommate, and lifelong friend, Emily Briley taught me to think outside the box and also frequently reminded me that she thinks I am a “freakin’ genius” (she lied, but it made me feel good). To Eric, who kissed me on New Year’s and made me remember what life is really all about, and to buddha (not the god) who reminded me that we all have the power to make our lives exactly what we want. And lastly, this thesis is dedicated to Casey, my ever-faithful companion and constant source of affection. I think I love a good walk or visit to the dog park almost as much as he does.

I must always thank God for providing the great comfort and peace we all need to succeed.

Without His presence in my life none of this would be possible.

## **PROFESSIONAL ACKNOWLEDGEMENTS**

This research was partially supported by The University of Tennessee Agricultural Experiment Station McIntire-Stennis E112215 (MS-75); USDA Special Wood Utilization Grants R112219-150 and R112219-184. Funding was also provided by University of Tennessee, Department of Statistics, Operations, and Management Science.

I would also like to thank the members of my committee who have helped to shape my graduate career, and without whom this master's thesis would not be possible: Timothy M. Young, Dr. Frank M. Guess, and Dr. Ramón V. León. Other professors who have had a profound influence on my education include Dr. William Seaver, Dr. Mary Leitnaker, Dr. Mary Sue Younger, Dr. Hamparsum Bozdogan and Dr. Timothy G. Rials.

Special thanks goes to Dr. Yang Wang (M.D. China) for helping with proofreading various papers as well as this document. Also, I would like to thank Diane Perhac for all of her helpful wisdom, and Amanda Silk for all of her professional and personal support throughout her employment at the University of Tennessee over the past year. Rebecca Walker and Jane Moser have also been instrumental in coordinating the administrative portions of my education.

## ABSTRACT

The thesis examines the causality of the central tendency of the Internal Bond (IB) of Medium Density Fiberboard (MDF) with predictor variables from the MDF manufacturing process. Multiple linear regression (MLR) models are developed using a best model criterion for all possible subsets of IB for four MDF thickness products reported in inches, e.g., 0.750", 0.625", 0.6875", and 0.500". Quantile Regression (QR) models of the median IB are also developed.

The adjusted coefficient of determination ( $R^2_a$ ) of the MLR models range from 72% with 53 degrees of freedom to 81% with 42 degrees of freedom, respectively. The Root Mean Square Errors (RMSE) range from 6.05 pounds per square inch (p.s.i.) to 6.23 p.s.i. A common independent variable for the 0.750" and 0.625" products is "Refiner Resin Scavenger %". QR models for 0.750" and 0.625" have similar slopes for the median and average but different slopes for the 5<sup>th</sup> and 95<sup>th</sup> percentiles. "Face Humidity" is a common predictor for the 0.6875" and 0.500" products. QR models for 0.6875" and 0.500" indicate different slopes for the median and average with different slopes for the outer 5<sup>th</sup> and 95<sup>th</sup> percentiles.

The MLR and QR validation models for the 0.750", 0.625" and 0.6875" product types have coefficients of determination for the validation data set ( $R^2_{validation}$ ) ranging from 40% to 60% and RMSEP ranging from 26.5 p.s.i. to 27.85 p.s.i.. The MLR validation model for the 0.500" product has a  $R^2_{validation}$  and RMSEP of 64% and 23.63 p.s.i. while the QR validation model has a  $R^2_{validation}$  and RMSEP of 66% and 19.18 p.s.i. The IB for 0.500" has departure from normality which is reflected in the results of the validation models. The thesis results provide further evidence that QR is a more defensible method for modeling

the central tendency of a response variable when the response variable departs from normality.

The use of QR provides MDF manufacturers with an opportunity to examine causality beyond the mean of the distribution. Examining the lower and upper percentiles of a distribution may provide significant insight for identifying process variables that influence IB failure or extreme IB strength.

**Keywords.** -- multiple linear regression, quantile regression, model building, best model criterion, medium density fiberboard, internal bond, independent variables.

# TABLE OF CONTENTS

	<u>Page</u>
<b>CHAPTERS</b>	
<b>1. Introduction.....</b>	<b>1</b>
<b>2. Literature Review.....</b>	<b>5</b>
2.1 MEDIUM DENSITY FIBERBOARD.....	5
2.2 MULTIPLE LINEAR REGRESSION.....	8
2.3 QUANTILE REGRESSION .....	11
<b>3. Modeling the Internal Bond of Medium Density Fiberboard     using Quantile Regression.....</b>	<b>17</b>
3.1 INTRODUCTION AND MOTIVATION.....	17
3.2 METHODS.....	19
Relational database.....	20
Classical linear regression.....	21
3.3 MODEL BUILDING AND BEST MODEL CRITERIA.....	22
Model building.....	22
Best model criteria.....	23
3.4 SAS CODE FOR MIXED STEPWISE REGRESSION.....	25
3.5 QUANTILE REGRESSION.....	25
3.6 RESULTS AND DISCUSSION.....	28
Product types 0.750” and 0.625”.....	30
Product types 0.6875” and 0.500”.....	37
3.7 CONCLUSIONS FOR CHAPTER 3.....	42
<b>4. Predictive Modeling using Quantile Regression .....</b>	<b>46</b>
4.1 COMPARING PREDICTIVE MODELING OF MULTIPLE LINEAR REGRESSION WITH QUANTILE REGRESSION MODELS FOR THE IB OF MEDIUM DENSITY FIBERBOARD.....	46
4.2 METHODS.....	47
4.3 RESULTS AND DISCUSSION.....	48
Product type 0.750”.....	48
Product type 0.625”.....	52
Product type 0.6875”.....	56
Product type 0.500”.....	59
4.4 CONCLUSIONS FOR CHAPTER 4.....	62

# TABLE OF CONTENTS

## CHAPTERS

<b>5. Using R software for Reliability Data Analysis</b> .....	66
5.1 INTRODUCTION AND MOTIVATION.....	66
5.2 EXPLORATORY DATA ANALYSIS FOR RELIABILITY.....	70
5.3 MAXIMUM LIKELIHOOD ESTIMATES FOR THE WEIBULL DISTRIBUTION AND OTHERS .....	74
Weibull distribution.....	74
Reliability/Survival function and the Kaplan-Meier estimator..	75
Maximum likelihood estimation (MLE).....	77
5.4 CONCLUSIONS FOR CHAPTER 5.....	79
<b>6. Summary and Concluding Remarks</b> .....	80
<b>BIBLIOGRAPHY</b> .....	84
<b>APPENDIX A. SAS Code for Mixed Stepwise Regression for All Possible Subsets</b> .....	89
<b>APPENDIX B. R Code for Multiple Quantile Regression</b> .....	91
<b>APPENDIX C. R Code for Weibull Distribution MLE Estimation</b> .....	92

# LIST OF TABLES

	<u>Page</u>
Table 1. MLR models for product types 0.750” and 0.625”.....	31
Table 2. MLR and QR models for product type 0.750”.....	35
Table 3. MLR and QR models for product type 0.625”.....	36
Table 4. MLR models for product types 0.6875” and 0.500”.....	38
Table 5. MLR and QR models for product type 0.6875”.....	41
Table 6. MLR and QR models for product type 0.500”.....	42
Table 7. MLR and QR models for product types 0.750”.....	49
Table 8. MLR and QR models for product types 0.625”.....	53
Table 9. MLR and QR models for product types 0.6875”.....	57
Table 10. MLR and QR models for product types 0.500”.....	60
Table 11. General tutorial of installing R with code examples.....	68
Table 12. Exploratory data analysis- basic statistics.....	73
Table 13. Exploratory data analysis- plots.....	73

# LIST OF FIGURES

	<u>Page</u>
Figure 1. Modular cabinet unit constructed of MDF with a veneer overlay.....	5
Figure 2. Quantile regression $\rho$ function.....	27
Figure 3. Adjusted $R^2$ for all possible subsets explored for 0.750”.....	28
Figure 4. Adjusted $R^2$ for all possible subsets explored for 0.625”.....	29
Figure 5. Adjusted $R^2$ for all possible subsets explored for 0.6875”.....	29
Figure 6. Adjusted $R^2$ for all possible subsets explored for 0.500”.....	30
Figure 7. Comparison of MLR fit (red dashed line) with median (blue line) and other percentile fits (from bottom to top: 5 <sup>th</sup> , 10 <sup>th</sup> , 25 <sup>th</sup> , 75 <sup>th</sup> , 90 <sup>th</sup> , and 95 <sup>th</sup> ) for 0.750” product type.....	33
Figure 8. Comparison of MLR fit (red dashed line) with median (blue line) and other percentile fits (from bottom to top: 5 <sup>th</sup> , 10 <sup>th</sup> , 25 <sup>th</sup> , 75 <sup>th</sup> , 90 <sup>th</sup> , and 95 <sup>th</sup> ) for 0.625” product type.....	34
Figure 9. Comparison of MLR fit (red dashed line) with median (blue line) and other percentile fits (from bottom to top: 5 <sup>th</sup> , 10 <sup>th</sup> , 25 <sup>th</sup> , 75 <sup>th</sup> , 90 <sup>th</sup> , and 95 <sup>th</sup> ) for 0.6875” product type.....	39
Figure 10. Comparison of MLR fit (red dashed line) with median (blue line) and other percentile fits (from bottom to top: 5 <sup>th</sup> , 10 <sup>th</sup> , 25 <sup>th</sup> , 75 <sup>th</sup> , 90 <sup>th</sup> , and 95 <sup>th</sup> ) for 0.500” product type.....	40
Figure 11. MLR validation of 0.750” actual and predicted IB.....	50
Figure 12. QR (median) validation of 0.750” actual and predicted IB.....	50
Figure 13. Histogram and quantile plot for 0.750” training data set.....	51
Figure 14. MLR validation of 0.625” actual and predicted IB.....	54
Figure 15. QR (median) validation of 0.625” actual and predicted IB.....	54
Figure 16. Histogram and quantile plot for 0.625” training data set.....	55
Figure 17. MLR validation of 0.6875” actual and predicted IB.....	58

# LIST OF FIGURES

	<u>Page</u>
Figure 18. QR (median) validation of 0.6875” actual and predicted IB.....	58
Figure 19. Histogram and quantile plot for 0.6875” training data set.....	59
Figure 20. MLR validation of 0.500” actual and predicted IB.....	61
Figure 21. QR (median) validation of 0.500” actual and predicted IB.....	61
Figure 22. Histogram and quantile plot for 0.500” training data set.....	62
Figure 23. Example of summary output from R of descriptive statistics. ....	70
Figure 24. Example of normal Q-Q plot of internal bond of MDF using R code.....	71
Figure 25. Example of histogram of internal bond of MDF using R code.....	71
Figure 26. Example of box plot of internal bond of MDF using R code.....	72
Figure 27. Example of Weibull Q-Q plot of internal bond of MDF using R code.....	72
Figure 28. Example of Kaplan-Meier Plot of internal bond of MDF using R code.....	73
Figure 29. Illustration of Weibull PDF with altering values of $\lambda$ and $\kappa$ .....	76
Figure 30. Example of Weibull MLE of internal bond of MDF with Q-Q plot using R code.....	78

# CHAPTER 1

## Introduction

Medium Density Fiberboard (MDF) is a non-structural engineered wood product that has gained recent popularity due to its many desirable characteristics. These characteristics include: 1) surface consistency; 2) uniform core density; and 3) lack of irregularities that natural grown wood cannot always offer. This highly demanded product can also be machined to produce many aesthetically pleasing varieties of cabinetry and other home furnishing at very reasonable prices. These products may then be covered with veneers or painted to add to their appeal. Clearly, the MDF market has made its mark in the United States, as the domestic production of MDF increased by 32.3% in 2004 (Howard 2006). Globally, China's MDF industry has rapidly expanded since 2001 with 492 MDF manufacturers and 609 production lines in 2005 (<http://www.asiawoodweb.com/news.asp>). Major recent capital expansions in MDF have made China the No.1 producer in the world, surpassing all of Europe ([http://www.nbmda.org/Member\\_Center/Export\\_Resources](http://www.nbmda.org/Member_Center/Export_Resources)).

To ensure consistent product quality from all manufacturers, MDF quality standards are determined by the Composite Panel Association (CPA). Guidelines for product characteristics such as Modulus of Rupture (MOR), Modulus of Elasticity (MOE), Screw-Holding, Thickness Swell, and Internal Bond (IB) are all measured and documented by manufacturers. In this thesis, we concentrate on the important characteristic of IB. IB is a measure of the tensile strength that is calculated by a pulling apart two inch by two inch MDF blocks using a destructive testing process. IB is the standard of quality for MDF manufacturers. Identifying the key independent variables that most significantly impact IB

strength is crucial in maintaining quality, production efficiency and lowering costs, all vital for sustaining competitiveness in the industry.

The methods and research of this thesis provide MDF manufacturers with important techniques for quantifying unknown sources of variation that will facilitate variation reduction, cost savings and continuous improvement. The theme of the thesis is consistent with general strategies outlined by many notable scholars (Box 1993, Deming 1986, Deming 1993, Feigenbaum 1991, Ishikawa 1976, Juran and Gryna 1951, Shewhart 1931, and Taguchi 1993).

In Chapter 2 of the thesis, a literature review is presented. The literature review has three sections. First, a brief history of MDF manufacture and its applications are presented. Second, a brief review of the popular data-mining tool, Multiple Linear Regression (MLR) with a discussion of its origins is presented. Tersely, we examine a relatively new data analysis technique known as Quantile Regression (QR). We hope this literature review will incrementally improve the knowledge of these subjects for a broad audience of readers.

In Chapter 3, the data set used in this study is discussed. The data set came from a large-capacity North American Medium Density Fiberboard (MDF) manufacturer. The data set aligns 184 on-line process readings with IB measurements obtained from periodic destructive testing creating the real-time relational database used in this thesis. As previously discussed, MDF manufacturers strive to increase efficiency and lower costs; to this end, it is imperative that the manufacturer has an advanced knowledge of the process and causality of IB variation. This chapter focuses on a comparison of MLR and QR for modeling the IB of MDF. Four MDF thickness product types are analyzed and reported in inches are 0.750", 0.625", 0.6875", and 0.500". One data subset is created for each product type using SAS

Business Intelligence and Analytics Software (Appendix A) and a best model criterion is used to create both MLR and QR models. While MLR develops models based on the mean of the IB response variable, QR models can be developed for any percentile of the response variable. The thesis develops QR models using R software for the median IB or 50<sup>th</sup> percentile. Modeling beyond the mean of the IB distribution may provide greater insight into the manufacturing process and help MDF manufacturers identify and quantify unknown sources of process variation.

Chapter 4 builds upon the research presented in Chapter 3 by comparing MRL and QR predictive models. Currently, the biggest challenge facing MDF manufacturers in North America is identifying, quantifying and controlling sources of variation within their processes. Given that hundreds of process variables may influence the IB of MDF, it is imperative for sustaining competitiveness that manufacturers understand the structure of causality and are able to model it appropriately in order to improve quality, increase process efficiency, lower defects, lower energy usage and lower raw material costs.

A traditional predictive modeling method is MLR. However, this method can be problematic when important assumptions are not met. These assumptions include: 1) linearity of the coefficients; 2) normal or Gaussian distribution for the response errors ( $\varepsilon$ ); and 3) the errors  $\varepsilon$  have a common distribution. In a MDF industrial manufacturing setting, these assumptions may not always be valid; therefore, a QR predictive modeling method may be a more appropriate option for modeling the IB of MDF.

In this thesis, all QR analyses and the reliability analyses presented in Chapter 5 are performed using the software package R (Appendices B and C). This package is an “Open Source” option for those interested in statistical analysis and is free. “Open Source” refers

to the package being made available to the general public with relaxed intellectual property restrictions, allowing the users to create user-generated software content through either incremental individual effort, or collaboration. The following website may be visited for more information on this matter: ([http://en.wikipedia.org/wiki/Open\\_source](http://en.wikipedia.org/wiki/Open_source)). Although much code is specific to the R package, several S-PLUS commands will run in R without modification.

In Chapter 5 the R software package is used to perform various reliability analyses ranging from descriptive statistics and graphics to survival analysis and Maximum Likelihood Estimation (MLE). Limited documentation for R software exists; however, various references are listed to assist those readers interested in learning more about this versatile software package.

The purpose of this thesis is education and exploration. It is imperative for manufacturers to utilize all available analytical tools to enable them to produce the highest quality products as efficiently as possible. Real prices of manufactured wood products like MDF are declining in spite of higher raw material and energy costs (Howard 2006). MDF manufacturers will be forced to lower production costs in order to remain profitable and stay in business. Adopting new low-cost software packages coupled with the most current analytical techniques may provide manufacturers with some additional tools for sustaining competitiveness in today's highly competitive global economy.

# CHAPTER 2

## Literature Review

### 2.1 MEDIUM DENSITY FIBERBOARD

Large-scale production of Medium Density Fiberboard (MDF) began in the 1980s. MDF is an engineered wood product formed by combining wax and resin with broken down wood fibers 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)). Recently, MDF has become one of the most popular composite wood materials given its excellent uniformity and versatility. MDF is an excellent base for veneers and laminates as well as non-structural constructions such as shelving, furniture, and decorative molding (**Figure 1**). As with solid wood, MDF can be nailed, glued, screwed, stapled, or attached with dowels (<http://www.wisegeek.com/what-is-mdf.htm>).

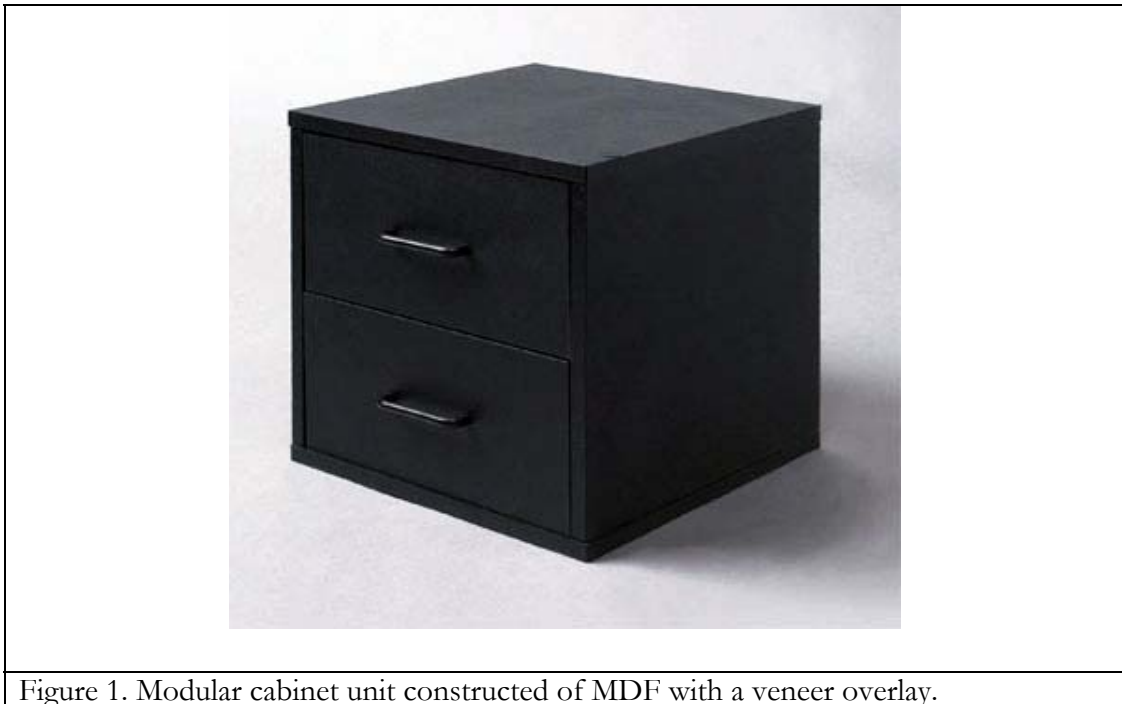


Figure 1. Modular cabinet unit constructed of MDF with a veneer overlay.

The use of MDF as a relatively low-cost non-structural building material has recently gained popularity as an alternative to more expensive solid wood building material. According to the “U.S. Annual Market Review and Prospectus 2002-2006” (Howard 2006), the domestic production of MDF increased by 32.3% in 2004, and is projected to continue this trend. MDF imports, consumption, and exports are also expected to increase in the following years. Since 1998, the real prices of manufactured wood products have declined and are expected to continue to do so (Howard 2006). This may force the manufacturers to lower their production costs in order to remain profitable in the competitive global economy.

As with any standardized building material, there are published industry product standards specifying the quality requirements of MDF. The quality of MDF is assessed based on several physical destructive test measurements. These include Modulus of Rupture (MOR), Modulus of Elasticity (MOE), Screw-Holding, Thickness Swell, and as analyzed in this thesis, Internal Bond or IB (Composite Panel Association 2006). Each destructive test measurement is highly important when assessing the quality of MDF. However, the challenge faced by many MDF manufacturers is to consistently produce high quality MDF using the aforementioned metrics as a measure of quality. The goal of this thesis is to identify and quantify causality between the IB of MDF and process variables that may be important during the manufacture of MDF. Process modeling, and detection of process differences, is vital the forest products manufacturing industry.

The use of statistical methods to examine sources of variation for the IB of MDF is not new. Steele (2006) discusses the use of Mean Residual Life (MRL) functions, and more specifically, unique “function domain sets” confidence intervals. This different breed of

confidence interval allows the practitioner to identify opportunities for quality improvement as well as make novel statements about the process. Steele's (2006) work was an extension of previous research utilized in a plethora of processes other than MDF. Steele (2006) insightfully discusses the use of the software package, MAPLE 10, and generously provides the code used for analysis.

Chen (2005) built upon the work of Edwards (2004) by exploring the use and effectiveness of estimating extremely small percentiles, or early failures, of strength measurements for MDF (i.e., IB). Chen (2005) observed that the distribution of strength failure data for IB does not follow a perfectly Gaussian distribution, and notes that forcing a Gaussian model on these data sets may lead to erroneous conclusions and profit loss. Chen (2005) proposes a forced censoring technique to closer fit the tails of strength distributions. The information obtained from these new fits may reduce the number of field failures, improve product safety, and even reduce the cost of destructive testing. More information on these reliability methods as applied to MDF can be found in the published work of Chen et al. (2006) and Guess et al. (2004).

Edwards (2004) also applies reliability techniques to improve production quality and safety of MDF. Edwards (2004) is also concerned with the extremely small percentiles, or early failures, of MDF. Edwards (2004) discusses the applications of Akaike's Information Criteria or AIC (Akaike 1974) and Bozdogan's Information Complexity Criteria (ICOMP) (Bozdogan 1988) to the extremely small percentiles of MDF. Modeling these failures can be challenging given the small amounts of data in the tails of the MDF failure distributions. Given the small sample size Edwards (2004) discusses the use of bootstrap techniques to provide more accurate estimation of lower percentile strength data.

## 2.2 MULTIPLE LINEAR REGRESSION

Elementary statistics texts tell us that the method of least squares was first discovered about 1805 (Stigler 1986). 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 (1777-1855) 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). Stigler (1986) notes that Sir Francis Galton discovered regression about 1885 in studies of heredity. Any contemporary course in regression analysis today starts with the methods of least squares and its variations.

Multiple Linear Regression (MLR) is one of the most commonly used data mining techniques, and can provide insightful information in cases where the rigid assumptions associated with MLR are met. The assumptions include 1) linearity of the coefficients; 2) normal or Gaussian distribution for the response errors ( $\varepsilon$ ); and 3) the errors  $\varepsilon$  have a common distribution. MLR is a very versatile tool and can be applied to almost any process, system, or area of study. Much has been published regarding this subject, and the following text may be useful to the reader: Kutner et al. (2004), as well as Myers (1990), provide thorough accounts of MLR and will be indispensable for most readers.

A key step in developing an appropriate MLR model is selecting a method of model building and a set of best model criteria. As used in this thesis, stepwise regression is commonly used for model building. Introduced by Efroymson (1960), stepwise regression was intended to be an automated procedure that selects the most statistically significant variables from a finite pool of independent variables. There are three separate stepwise

regression procedures, including 1) forward selection; 2) backward selection; and 3) mixed selection. Mixed selection is the most statistically defensible type of stepwise regression, and is a mixture of the forward and backward procedures. For more information on this procedure see Kutner et al. (2004), Neter et al. (1996), and Draper and Smith (1981).

A set of best model criteria are commonly used in conjunction with stepwise regression in order to select the optimal model. Due to the nature of MDF manufacturing, some specific concerns must be addressed. As cited by Young and Guess (2002), and Young and Huber (2004), multicollinearity and heteroscedasticity can be significant problems when modeling the IB of MDF using industrial data. Young and Guess (2002) used the following best model criteria: 1) maximum Adjusted  $R^2$ ; 2) parameters ( $p$ )  $\approx$  Mallow's  $C_p$  (Mallow 1973); 3) minimum Akaike's Information Criterion (AIC), Akaike (1974); 4) Variance Inflation Factor (VIF)  $< 10$ ; 5) significance of independent variables  $p$ -value  $< 0.10$ ; 6) absence of heteroscedasticity in residuals,  $E(\epsilon_i) = 0$ .

For this thesis, we focus on the aforementioned criteria, but due to a lack of data records for each product type we do not use Mallow's  $C_p$  (Mallow 1973). We also use a  $p$ -value  $< 0.05$  for significance among the independent variables. The adjusted  $R^2$  statistic,  $R_a^2$ , is a better measure of fit for MLR models built with the potential to contain significantly more independent variables than data records. As additional independent variables are added to a regression model,  $R^2$  will always increase regardless of the fit. The  $R_a^2$  statistic only increases if the residual sum of squares decreases (Draper and Smith 1981). The  $R_a^2$  statistic minimizes the risk of, and penalizes for, using too many independent variables. AIC measures the complexity of the model and guards against model bias. VIFs

are reported to protect against multicollinearity, and redundancy in the model. Models with  $VIF < 10$  can be said to be relatively free of these effects (Kutner et al. 2004).

As noted by Kutner et al. (2004), model validation is the final step in the regression modeling-building process. Kutner et al. (2004) point to three main methods associated with model validation, as follows: 1) collection of new data to validate the current model and its predictability; 2) comparison of current results with other theoretical values, empirical and simulation results; and 3) use of a cross-validation sample to validate and assess the predictive power of the current model.

For this thesis, we use the cross-validation approach to assess the validity and predictability of the regression models constructed, i.e., we remove the most current twenty records from the model-building process, and then use the constructed model to estimate their computed values. A general rule of thumb in regression model building is to use 80 percent of the data set for the development of the training model and the remaining 20 percent for validation of the model (Kutner et al. 2004). Validation records can be selected at random from the entire data set or in the case of data that are a time series the validation set can be the most current 20 percent (Kutner et al. 2004). Adequate regression models are expected to yield estimates reasonably close to the actual data values.

A plethora of statistics are available to aid in assessing the predictive power of regression models. A popular statistic for assessing this predictability is the Root Mean Squared Error of the Prediction (RMSEP) statistic (André et al. 2006). This statistic is computed by calculating the square root of the Sum Squared Errors (SSE) for the withheld records divided by the corresponding degrees of freedom. Lower RMSEP values indicate better model predictability. Another common model validation statistic is the classic

coefficient of determination, or  $R^2$ , statistic. This value is also computed for the withheld sample, and provides some insight into the predictability of the model. By definition, higher  $R^2$  values are preferred, i.e., the  $R^2$  statistic indicates the amount of variation explained by the regression model.

## 2.3 QUANTILE REGRESSION

Response data in the tails, or outer quantiles, of a distribution may behave differently than data in the inner quantiles of the distribution in response to the predictor variables. Traditionally, MLR is used to study causality between independent variables and the central tendency of a response variable as measured by the mean or average, with an important goal of making useful predictions of the response variable. However, several stringent aforementioned assumptions must be met in order for a MLR model to perform well. In contrast to MLR, Quantile Regression (QR) does not impose any strict parametric assumptions (Koenker 2005).

QR seeks to estimate conditional quantile functions, i.e., the varying values of covariates are estimated based on the quantile's asymmetrically weighted absolute residuals of the median rather than the mean of the distribution (Buhai 2004). Quantile Regression (QR) is an approach that allows us to examine the behavior of the target variable (Y) beyond its average of the Gaussian distribution, e.g., median (50th percentile), 10<sup>th</sup> percentile, 80<sup>th</sup> percentile, 95<sup>th</sup> percentile, etc. Examining these quantiles may provide greater insight into the process being studied, and allows the manufacturer to make more informed production decisions. Given the nature of the median statistic, this results in a more accurate and robust representation of the relationship between the covariates and their response variable. Buhai (2004) eloquently states, "Instead of assuming that covariates shift only the location or the

scale of the conditional distribution, Quantile Regression looks at the potential effects on the shape of the distribution as well.” The effect of the shape of the distribution on modeling the response variable IB of MDF using QR is discussed in Chapter 4 of this thesis.

Examining the relationship between key quality characteristics and the independent variables associated with processes is imperative in the wood products industry. This is especially true in MDF manufacture as they have a vested interest in understanding the lower or higher percentiles of the distribution of the key quality metric IB strength.

Quantile Regression (QR) was introduced by Koenker and Bassett (1978) and is intended to offer a comprehensive strategy for completing the regression picture (Koenker 2005). As Mosteller and Tukey (1977) note in their influential text, as cited by Koenker (2005): “...the regression curve gives a grand summary for the averages of the distributions corresponding to the set of Xs...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.”

The tome by Koenker (2005) should prove to be fairly comprehensive for most readers. This book outlines the fundamental theory of QR, and also provides some code to be used with Koenker’s package in the R software package entitled “quantreg”. For more information on using this insightful package, as well as the R software package, refer to the following website: <http://www.econ.uiuc.edu/%7Eroger/research/rq/rq.html>. Some other thoughtful chapters of interest in Koenker (2005) include the following titles: Inference for Quantile Regression, Asymptotic Theory of Quantile Regression, Computational Aspects of Quantile Regression, and Quantile Regression in R: A Vignette.

As with MLR, QR has many applications, and was originally developed for economic use, as the first QR publication was in *Econometrica* (Koenker and Bassett 1978). Koenker and Bassett insightfully envisioned a more robust regression approach capable of modeling conditional quantile functions beyond the classic MLR least squares approach to model building. Koenker and Bassett (1978) note, “estimators are suggested, which have comparable efficiency to least squares for Gaussian linear models while substantially outperforming the least-squares estimator over a wide class of non-Gaussian error distributions”.

Gorr and Hsu (1985) began applying these new techniques to the Management Science field of study and introduce an adaptive filtering procedure for exploring regression quantiles. These models are used as part of their Quantile Estimation Procedure (QEP) and are utilized to signal preventative actions and therefore avoid undesirable system states (Gorr and Hsu 1985).

Young and Easterling (1994) investigate QR as applied to reliability data analysis. Typically, in reliability applications, the practitioner is most interested in the outer quantiles of distributions being studied, i.e., products that have an extremely short or long lifespan. Young and Easterling (1994) use QR techniques to explore the outer quantiles of sensitivity test distributions. Various sample sizes are examined, as well as quantiles, and the effect of assuming different specified models is noted. Young and Easterling (1994) find that QR provides better models for their data when quantiles are estimated as a function of specific model parameters as opposed to tests developed in order to estimate a specific quantile.

Buchinsky (1998) provides a basic guide for empirical research, focusing on cross-section applications, while summarizing the most significant aspects of QR and filling in some noted literature gaps. Several alternative covariance matrix estimators are presented,

and Buchinsky (1998) eloquently discusses useful procedures for testing QR models for homoskedasticity and symmetry of the error distribution. A generous empirical example is presented using data obtained from a current population survey, and the paper concludes with a brief discussion on the application of censored QR models.

In 1999, Koenker and Machado introduce goodness-of-fit procedures for QR. These statistics are quite similar to the  $R^2$  statistic applied to classical regression techniques. Various inference processes designed to assess the adequacy of the regression model are presented (Koenker and Machado 1999). Koenker and Machado (1999) then illustrate their findings using empirical economic growth models, hypothetical examples, and conclude with Monte Carlo evidence.

The idea of computing regression quantiles with the use of a conditional quantile function is further articulated by Koenker and Hallock (2001). Koenker and Hallock (2001) discuss the undeniable link between quantile and the “operations of ordering and sorting the sample observations that are usually used to define them” (Koenker and Hallock 2001). The innate symmetry of the absolute value function ensures that there is the same number of observations both below and above the median (Koenker and Hallock 2001). Koenker and Hallock (2001) note there is high demand for more specialize QR models in the finance industry.

In their useful text, Fitzenberger et al. (2002) discussed the practical application of QR as compared to the methodology of least-squares regression. They note the important MLR assumption of constant error, and insightfully articulate a useful example pertaining to wage distributions, acknowledging the importance of proper distribution modeling.

Recently, censored regression models have received substantial attention in economic literature, both theoretical and applied (Honore et al. 2002). Honore et al. (2002) note that, to date, most estimation procedures for panel data models or cross-sectional models are constructed using fixed censoring. Honore et al. (2002) suggest a new procedure for adapting these fixed censoring models to perhaps more applicable random censoring models.

Some other interesting applications include those in ecological and environmental studies (Cade and Noon 2003). As noted by Cade and Noon (2003), it is extremely difficult to identify, document, and measure every ecological independent variable. As a result, using classical MLR methods and others, it is sometimes impossible to arrive at a statistically significant model. However, models built using only portions of the response variable distribution may be more useful (Cade and Noon 2003). Cade and Noon (2003) explore various ecological QR applications, and thoughtfully estimate prediction intervals.

Interestingly, Green and Kozek (2003) use an approximate QR method to model weather data. These models are approximate because they are formed by applying quantile functions onto parametric models (Green and Kozek 2003). Parametric weather distributions are modeled as they vary over time, and regression quantiles are then applied to the models (Green and Kozek 2003). Five-curve summaries are obtained for the probability distributions of the weather data and the results are quite interesting.

Buhai (2004) provides an introduction to QR, discussing basic models and interpretations as well as computational and theoretical aspects of the algorithm. By concentrating on only two applications of QR: 1) survival analysis; and 2) recursive structural equation models, Buhai (2004) is able to articulate a thorough summary of each.

Although many QR applications have been explored, and utilized in practice, the literature does not yet support QR as applied to MDF manufacture. Profit loss and inefficiency in the composite wood products industry generally is a result of product whose quality characteristic is substandard or unnecessarily over engineered. These classifications may correspond to the lower and upper quantiles of the quality characteristic distribution. Articulating a method to detect and model these extreme IB readings, if adopted, could result in an improved knowledge of wood composite strength and lead to cost savings and increased efficiency. The QR method as applied to the IB of MDF is explored in the next chapter.

# CHAPTER 3

## Modeling the Internal Bond of Medium Density Fiberboard using Quantile Regression

### 3.1 INTRODUCTION AND MOTIVATION

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 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, 1998, 1999, Greubel, 1999, Eriklsson et al. 2000, Young and Guess 2002, 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).

We are not aware of any published literature that uses quantile regression to investigate the percentiles of product quality for wood composites.

A data set from a large-capacity North American Medium Density Fiberboard (MDF)<sup>1</sup> manufacturer was obtained in 2002. The data set aligned process measurements from on-line sensors with the Internal Bond (IB) analyzed during periodic destructive testing. For example, on-line sensor measurements are 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.

This paper directly compares the use of Multiple Linear Regression (MLR) and Quantile Regression (QR) for modeling the IB of MDF. The purpose of the study is to use MLR and QR 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

---

<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” (<http://www.wisegeek.com/what-is-mdf.htm>).

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."

### 3.2 METHODS

Traditionally, one uses MLR to study the relationship between various independent variables and the average of the distribution for a response variable with an important goal of making useful predictions of the response variable. MLR has three important assumptions: 1) linearity of the coefficients; 2) normal or Gaussian distribution for the response errors ( $\varepsilon$ ); and 3) the errors  $\varepsilon$  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), 10<sup>th</sup> percentile, 80<sup>th</sup> percentile, 90<sup>th</sup> 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 is created by aligning real-time process sensor data with IB readings (Young and Guess 2002). The real-time process data are collected with Wonderware Industrial SQL 8.0 ([www.wonderware.com](http://www.wonderware.com)). The readings are combined with IB by product type at the instant when a panel is extracted from the production line for testing. The process data are collected using a median value from the last 100 sensor values (e.g., for most of the 184 different sensor variables this represents a two to three minute time interval). The process data are collected and stored using Industrial SQL. 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 is extracted for IB destructive testing are taken into account. A unique number (idnum) is generated when the panel is extracted from the process, and is later used to match process data with corresponding IB results.

When the IB results are matched with the process data, the combined data are recorded in two tables that appear 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 is automatically updated as new lab samples are taken using Microsoft Transact SQL code with Microsoft SQL “Jobs” and “Stored Procedures”.

The names used in this manuscript associated with the process variables for the on-line sensors are non-descriptive at the request of the manufacturer and given the terms of a legal confidentiality agreement. Definitions for the names of the process variables are not allowed under the terms of the legal confidentiality agreement.

## Classical linear regression

The first-order simple linear regression model is (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  $i^{\text{th}}$  observation,  
 $\beta_0$  is the intercept parameter,  
 $\beta_1$  is a slope parameter,  
 $x_{1i}$  is the value of the independent variable in the  $i^{\text{th}}$  observation,  
 $\varepsilon_i$  is a random error term of the  $i^{\text{th}}$  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  $i^{\text{th}}$  observation,  
 $\beta_0$  is the intercept parameter,  
 $\beta_k$  is the slope parameter associated with the  $k^{\text{th}}$  variable,  
 $x_{ki}$  is the  $k^{\text{th}}$  independent variable associated with the  $i^{\text{th}}$  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 a strength of the least

---

<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

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>

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

### 3.3 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 is used to develop first-order linear models of the IB for MDF.

Stepwise regression was introduced by Efraymson (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

---

a subset that is closed under linear combinations, e.g., linear regression equation of a linear subspace (<http://mathworld.wolfram.com/AffineFunction.html>. 2006).

| This footnote is also earlier.<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 (1777-1855) 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).

in the equation until a decision is reached on the equation to use (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 are noted. 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), 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 is used. We also use the “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](http://www.sas.com)) and seven criteria in selecting the best model of IB. The criteria include:

1) maximum Adjusted  $R^2_a$ ; 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.,  $E(\epsilon_i) = 0$ .

Adjusted  $R^2$ , or  $R^2_a$ , 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  $R^2_a$  will only increase if the residual sum of squares decreases.  $R^2_a$  minimizes the risk of “over-fitting” and penalizes for model saturation, i.e., the model is penalized if additional independent variables do not reduce the residual sum of squares. The formula for  $R^2_a$  is:

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

where,

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

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).

The diagnostic tool used to check the impact of multicollinearity in the MLR model is referred to as the VIF. 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).

### **3.4 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 is 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 are explored using the most recent data and then moving backward in time. Initial models are developed for the 50 most recent data records and additional models are developed for each additional record moving backward in time through the data. The aforementioned best model criteria are used in selecting the best model from the subsets provided by SAS. The SAS code for mixed stepwise regression exploring all possible subsets is given in Appendix A.

### **3.5 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 non-parametric median statistic

may offer additional insight in the analysis of a 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),

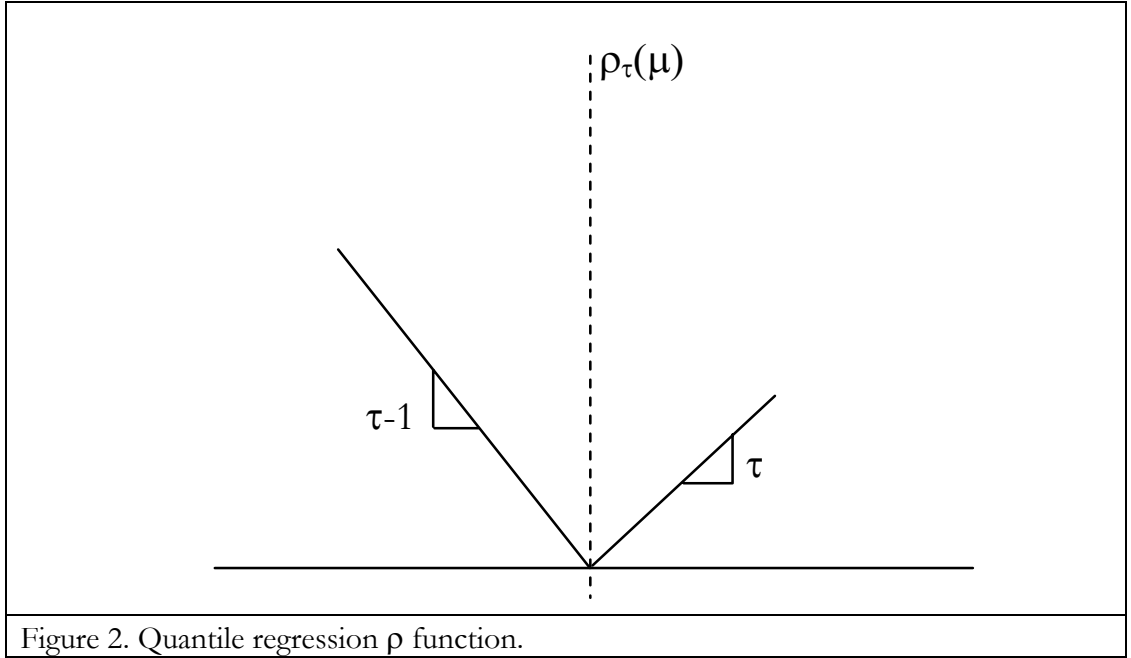
$$Q_{y_i}(\tau|x) = \beta_0 + \beta_i x_i + F_u^{-1}(\tau) \quad [8]$$

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 are 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 equation [9], is the tilted absolute value function appearing in **Figure 2** 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 equation [9] by the parametric function  $\xi(x_i, \beta)$  and set  $\tau$  to  $1/2$ .<sup>4</sup> To obtain estimates of the other conditional quantile functions, replace absolute values by  $\rho_\tau(\cdot)$ , e.g., equation [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. The R code for the QR models developed in this chapter is given in Appendix B.

---

<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).

### 3.6 RESULTS AND DISCUSSION

The Internal Bonds of four different product types of MDF are analyzed. Each product type represents a different board thickness in inches (i.e., 0.750", 0.625", 0.6875", 0.500"). All possible subset MLR models are explored for the four product types using  $R_a^2$  as a key indicator for determining the best subset model (Figures 3, 4, 5 and 6).

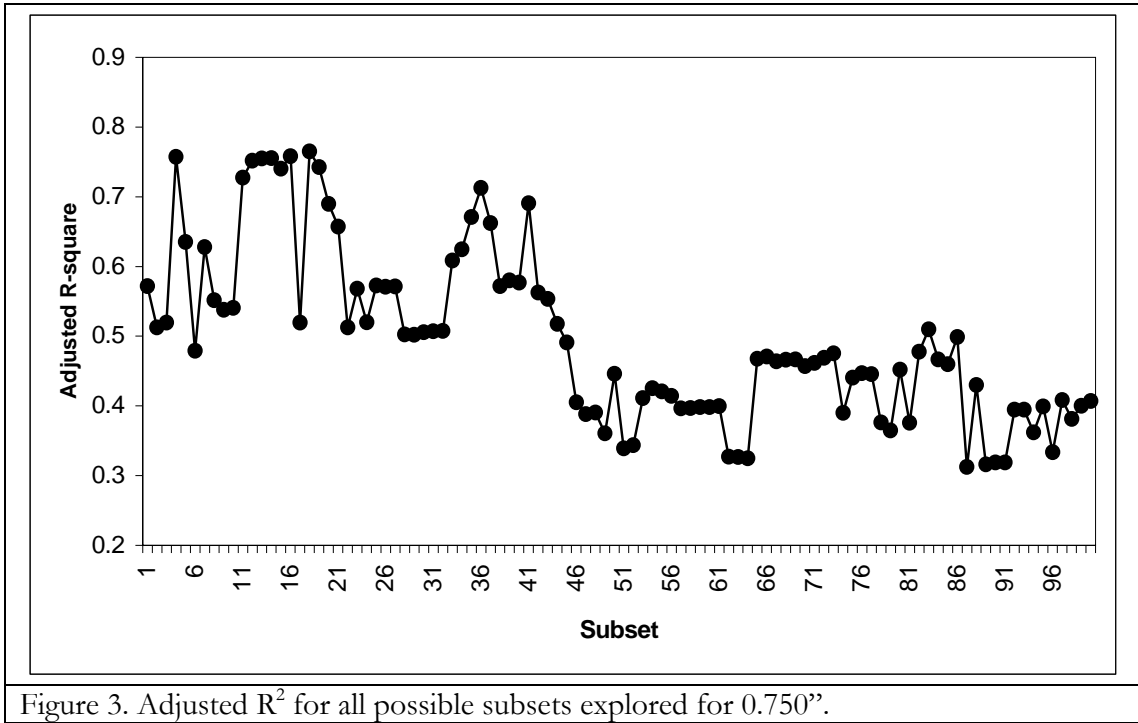


Figure 3. Adjusted  $R^2$  for all possible subsets explored for 0.750".

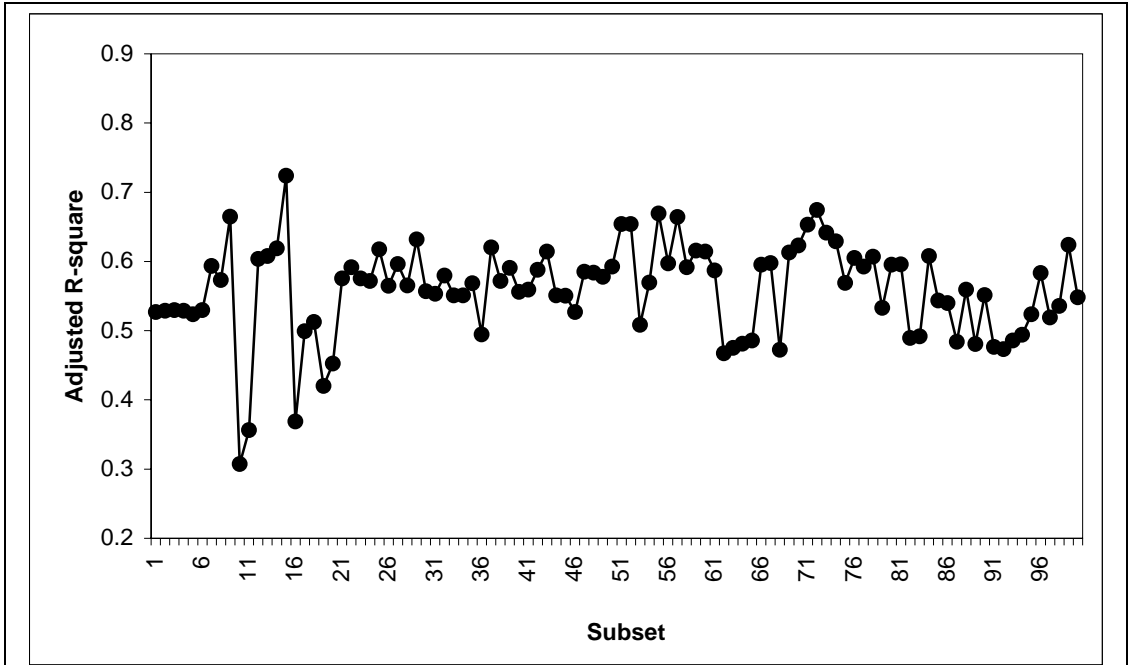


Figure 4. Adjusted  $R^2$  for all possible subsets explored for 0.625".

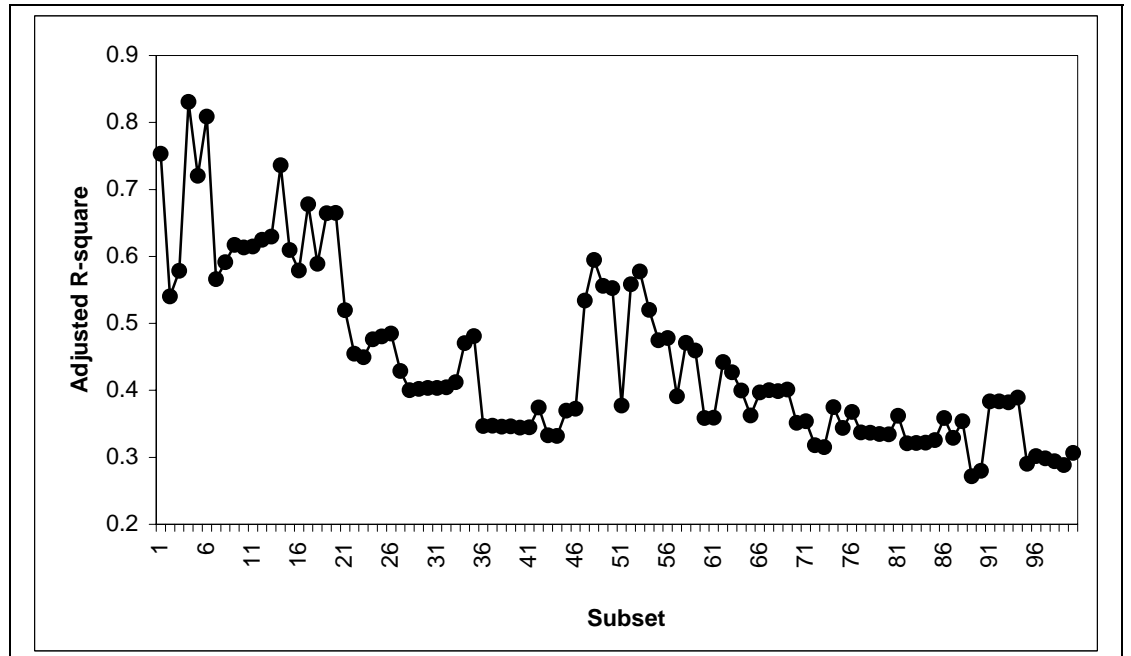


Figure 5. Adjusted  $R^2$  for all possible subsets explored for 0.6875".

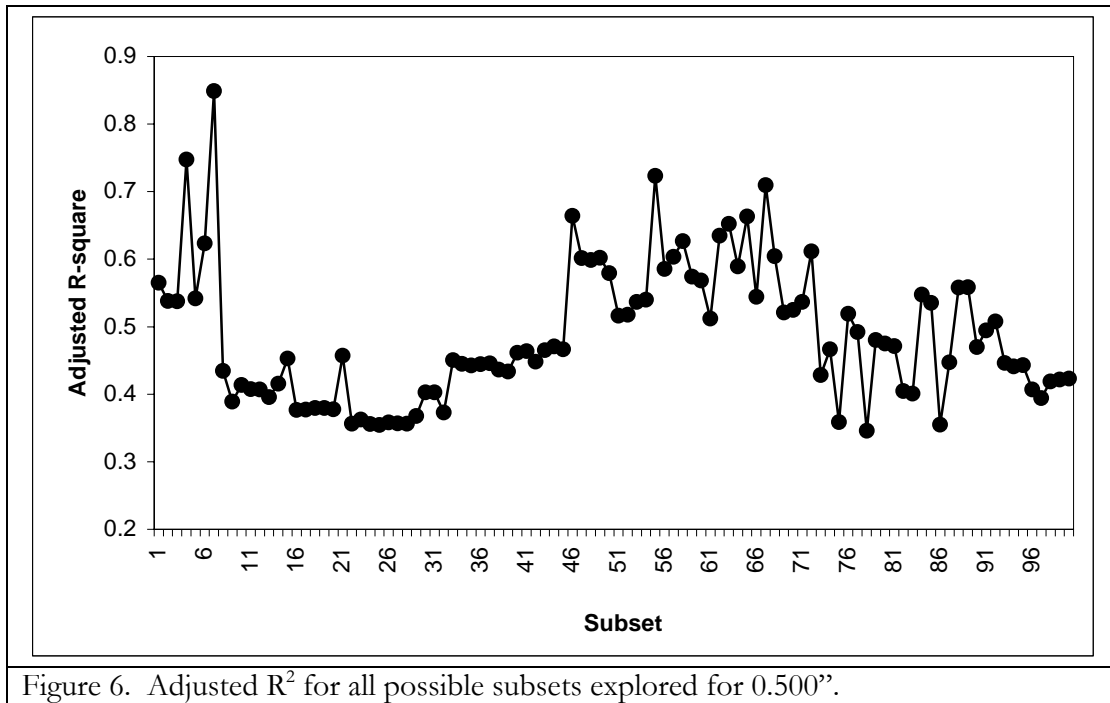


Figure 6. Adjusted  $R^2$  for all possible subsets explored for 0.500”.

The  $R^2_a$  for all possible subsets is 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  $R^2_a$  changes slowly as more records are 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 are obtained (i.e., using the best model criteria), commonalities in the independent variables are explored among the four product types.

**Product types 0.750” and 0.625”**

For the 0.750” product type a MLR model is developed with an  $R^2_a$  of 0.75, 50 degrees of freedom and 11 parameters. The RMSE of the model is 7.70 p.s.i. and the maximum VIF for any independent variable is 5.03. Residual patterns for the MLR model are homogeneous (Table 1).

Table 1. MLR models for product types 0.750” and 0.625”

	<b>0.750”</b>	<b>Scaled Estimate</b>	<b>p-value</b>	<b>0.625”</b>	<b>Scaled Estimate</b>	<b>p-value</b>
<b>P A R A M E T E R S</b>	Face MDF Temperature	-12.565	<.0001	Shavings Raw Weight	-15.872	<.0001
	Dryer S Fiber Moisture	-10.906	<.0001	<b>Refiner Resin Scavenger %</b>	8.396	0.0008
	<b>Refiner Resin Scavenger %</b>	-9.118	<.0001	Core Grinding Steam Flow	12.720	<.0001
	Core Dryer Outlet Temperature	18.498	0.0092	Core Resin to Wood %	22.473	<.0001
	Press Position Time	19.926	<.0001	Dryer Mass Flow	10.642	<.0001
	Dryer 1 Fan Current	23.662	<.0001	Resin Water Tank Temperature	-21.556	<.0001
	Dryer 2 Fan Current	-25.384	<.0001	Core Refiner Feeder Screw Speed	4.868	0.0110
	Refiner S Chip Level	10.666	<.0001	<b>Core Water to Wood</b>	-10.872	0.0077
	Refiner S Feeder Screw Speed	9.294	0.0017	Face Humidifier Temperature	13.583	<.0001
	<b>Core Water to Wood</b>	-21.043	<.0001	Relative Ambient Humidity	5.858	0.0100
	ESP Milliamps	-11.714	<.0001	Weight Actual	12.205	<.0001
<b>Important Regression Statistics</b>						
<b>R<sub>a</sub><sup>2,5</sup></b>	0.751646		<b>R<sub>a</sub><sup>2</sup></b>	0.723694		
<b>d.f.<sup>6</sup></b>	50		<b>d.f.</b>	53		
<b>P<sup>7</sup></b>	11		<b>P</b>	11		
<b>VIF<sub>max</sub><sup>8</sup></b>	5.0315819		<b>VIF<sub>max</sub></b>	5.603058		
<b>RMSE<sup>9</sup></b>	7.697272		<b>RMSE</b>	6.051464		
<b>Residual Pattern</b>	Homogeneous		<b>Residual Pattern</b>	Homogeneous		

<sup>5</sup> Adjusted coefficient of determination.

<sup>6</sup> Degrees of freedom.

<sup>7</sup> Number of explanatory variables.

<sup>8</sup> Maximum variance inflation factor.

<sup>9</sup> Root mean square error.

For the 0.625” product type a MLR model is developed with an  $R_a^2$  of 0.72, 53 degrees of freedom and 11 parameters. The RMSE of the model is 6.05 p.s.i. and the maximum VIF for any independent variable is 5.60. Residual patterns for the MLR model are homogeneous (**Table 1**).

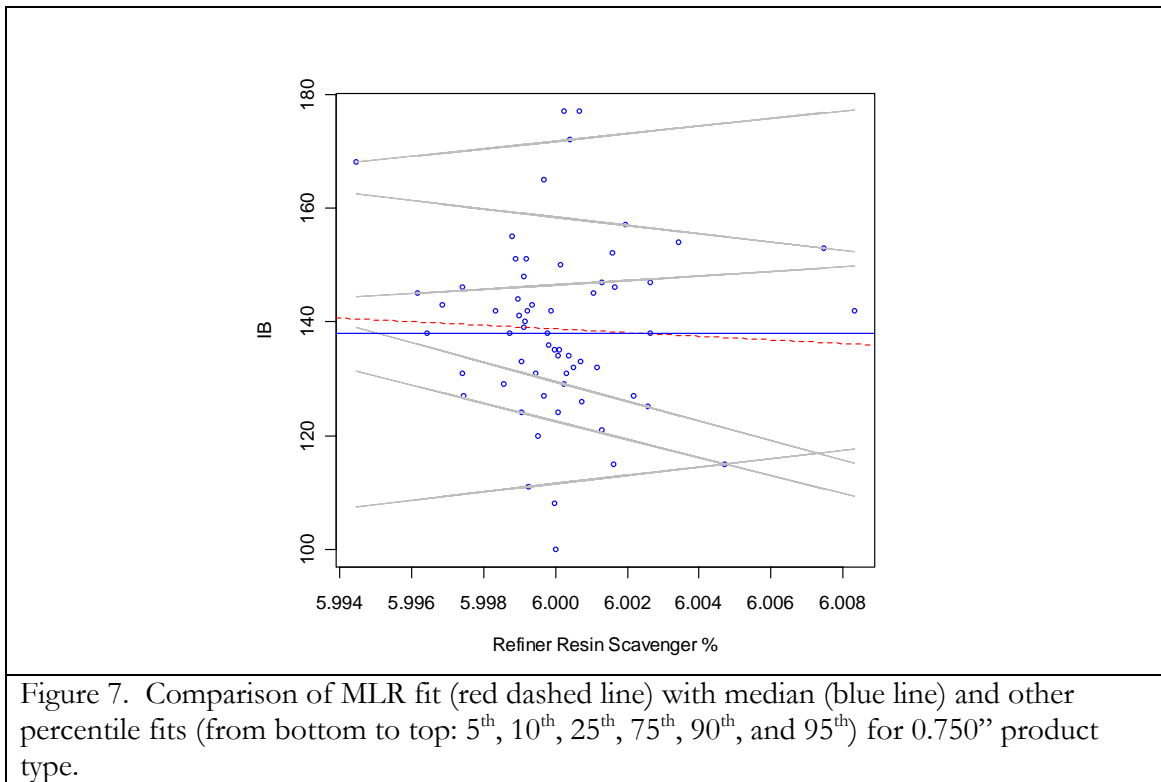
Common independent variables for the 0.750” and 0.625” MLR models are bolded in Table 1. “Refiner Resin Scavenger %” and “Core Water to Wood” were common for both 0.750” and 0.625” product types. It is surprising to see the scaled estimates for “Refiner Resin Scavenger %” differ in sign for each product type.<sup>10</sup> The “Refiner Resin Scavenger %” has a negative scaled estimate of approximately -9.12 p.s.i. on IB for 0.750” while the “Refiner Resin Scavenger %” has a positive scaled estimate of approximately 8.40 p.s.i. on IB for 0.625”. This may indicate that “Refiner Resin Scavenger %” is an important 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 is -21.04 p.s.i. for 0.750” and -10.87 p.s.i. for 0.625”. 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, i.e., the 0.750” product requires more wood to be refined because it is thicker than 0.650”; however the 0.750” IB is much more sensitive to changes in “Core Water to Wood”.

---

<sup>10</sup> Scaled estimate is a helpful statistic in MLR models in that 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.

To examine the influence of “Refiner Resin Scavenger %” beyond the mean effect on IB, QR is explored for this common parameter for both 0.750” and 0.625”.<sup>11</sup> We find that the influence of “Refiner Resin Scavenger %” on the lower percentiles of IB is quite different than the mid-range and higher percentiles (**Figures 7 and 8**).



<sup>11</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 only look at the single parameter case.

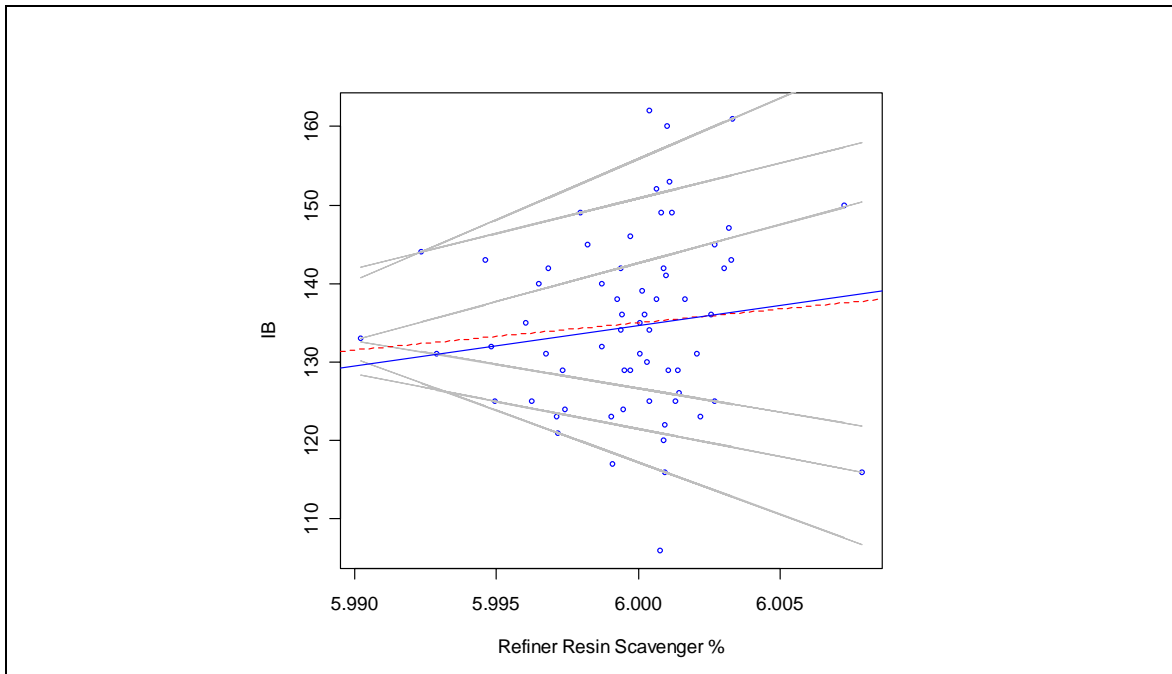


Figure 8. Comparison of MLR fit (red dashed line) with median (blue line) and other percentile fits (from bottom to top: 5<sup>th</sup>, 10<sup>th</sup>, 25<sup>th</sup>, 75<sup>th</sup>, 90<sup>th</sup>, and 95<sup>th</sup>) for 0.625” product type.

The red dashed line represents the MLR fit, the solid deep blue line represents the median fit, and the gray lines correspond to the 5<sup>th</sup>, 10<sup>th</sup>, 25<sup>th</sup>, 75<sup>th</sup>, 90<sup>th</sup>, and 95<sup>th</sup> percentiles, respectively. For the 0.750” product type (**Figure 7**), the slopes of the percentiles are quite different depending on percentile. The median and average have similar slopes. The 5<sup>th</sup> percentile (possible IB failures) and 95<sup>th</sup> 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 are the other operational settings and factors occurring during these events?

For the 0.625” product type (**Figure 8**), the slopes of the percentiles are 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 50<sup>th</sup>

percentile (median) the effect of “Refiner Resin Scavenger %” has a stronger positive influence on IB the higher the percentile. For percentiles below the 50<sup>th</sup> 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 are influencing IB in concert with “Refiner Resin Scavenger %” or that the quality of the “Refiner Resin Scavenger %” itself is 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 is used for illustration purposes, the quantile regression algorithm in R can also be applied to multiple independent variable models. Further analysis is conducted to examine the differences between the MLR and QR median fits for all of the MLR independent variables. For the 0.750” 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 are 39.84%, 34.37%, and 28.2%, respectively.

Table 2. MLR and QR models for product type 0.750”

0.750” Variables	Coefficients			
	MLR Average	QR Median	QR 10 <sup>th</sup> percentile	QR 90 <sup>th</sup> 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
<b>Refiner Resin Scavenger %</b>	-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
<b>Core Water to Wood</b>	-835.05	-651.32	-825.34	-957.74
ESP Milliamps	-0.16	-0.17	-0.15	-0.19

For the 0.625” product type (**Table 3**), the largest discrepancies between coefficients occur in “Shavings Raw Weight”, “Relative Ambient Humidity” and “Weight Actual”. The percent discrepancies are 42.96%, 16.16%, and 12.78%, respectively. These discrepancies reflect significant differences between modeling the mean and the median (50<sup>th</sup> 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 productions 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 10<sup>th</sup> and 90<sup>th</sup> 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.

Table 3. MLR and QR models for product type 0.625”

Variables	Coefficients			
	MLR Average	QR Median	QR 10 <sup>th</sup> percentile	QR 90 <sup>th</sup> percentile
Intercept	-1029.56	-2063.15	1896.63	-1745.51
Shavings Raw Weight	-1.55	-1.09	-1.38	-1.64
<b>Refiner Resin Scavenger %</b>	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
<b>Core Water to Wood</b>	-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

### **Product types 0.6875” and 0.500”**

For 0.6875” a MLR model is developed with an  $R_a^2$  of 0.81, 42 degrees of freedom and 13 parameters. The RMSE of the model is 6.23 p.s.i. and the maximum VIF for any independent variable is 4.54. Residual patterns for the MLR model are homogeneous (**Table 4**).

For 0.500” a MLR model is developed with an  $R_a^2$  of 0.75, 43 degrees of freedom and 10 parameters. The RMSE of the model is 6.57 p.s.i. and the maximum VIF for any independent variable is 5.55. Residual patterns for the MLR model are homogeneous (**Table 4**). “Face Humidity” is the common independent variable for both 0.6875” and 0.500” MLR models (**Table 4**).

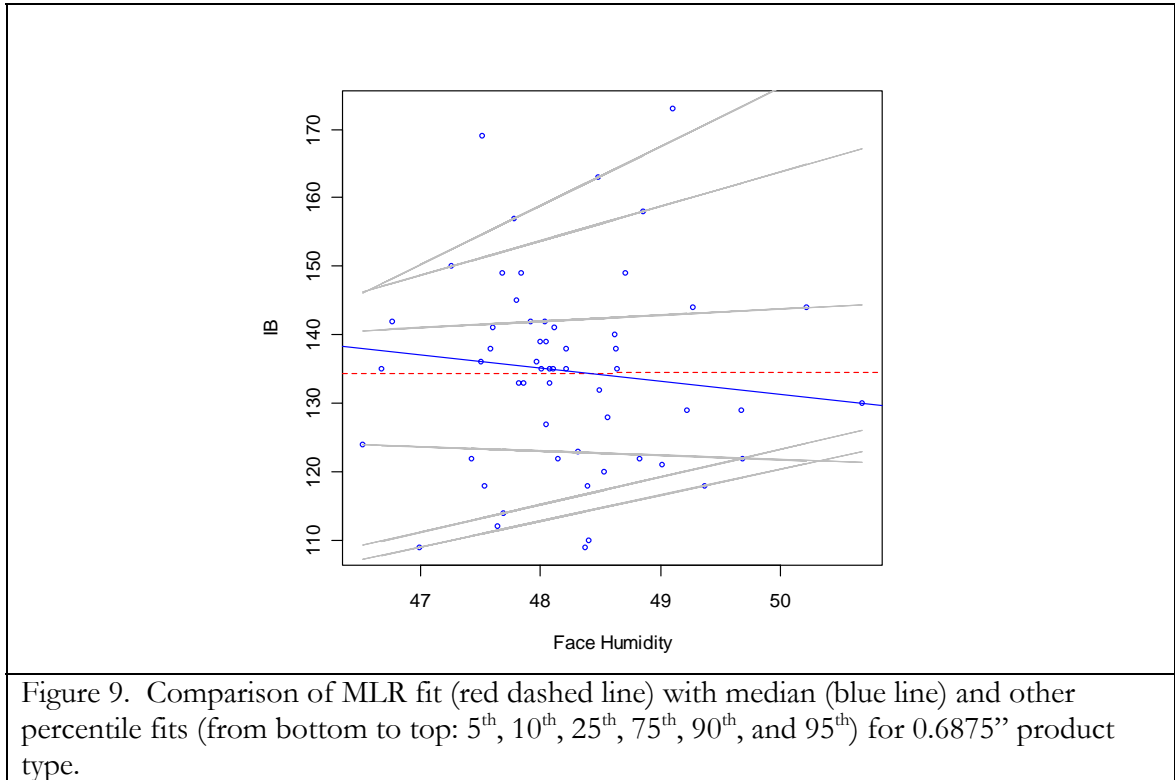
It is surprising to see the scaled estimates for “Face Humidity” differed in sign for each product type. The “Face Humidity” has a negative scaled estimate of -10.02 p.s.i. on IB for 0.6875” while the “Face Humidity” has a positive scaled estimate of 4.81 p.s.i. on IB for 0.500”. This may signify that “Face Humidity” is an important source of variability acting on IB that the manufacturer needs to investigate, i.e., it has a negative effect on IB for 0.6875” which requires more process control and can positively effect 0.500” with increases.

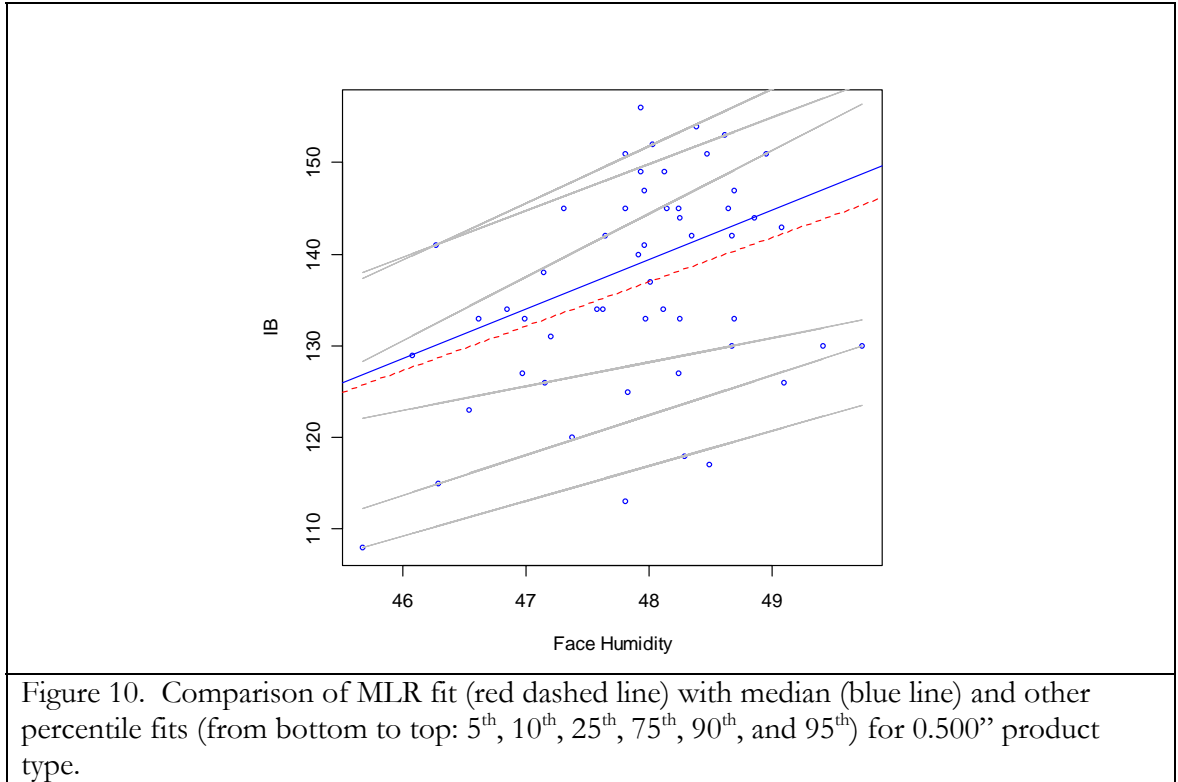
To examine the influence of “Face Humidity” beyond the mean effect on IB, QR is explored for this common parameter for both 0.6875” and 0.500”. The average and median fits for 0.6875” for “Face Humidity” have different slopes, which may indicate lack of normality in the response variable IB. We found that influence of “Face Humidity” on the outer 5<sup>th</sup> and 95<sup>th</sup> percentiles of IB is quite different than the inner percentiles (**Figure 9**).

Table 4. MLR models for product types 0.6875” and 0.500”

	<b>0.6875”</b>	<b>Scaled Estimate</b>	<b>p-value</b>	<b>0.500”</b>	<b>Scaled Estimate</b>	<b>p-value</b>
<b>P A R A M E T E R S</b>	Face Scavenger Resin %	25.479	<.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	<.0001	Weight Target	7.938	0.0194
	Mat Shave off Level	-15.408	<.0001	Core Blow Line Pressure	19.091	<.0001
	Refiner S Chip Level	14.655	<.0001	Face Digester Pressure	-9.494	0.0004
	Refiner S Grinding Steam Flow	21.066	<.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	<.0001
	Dryer ESP Outlet Temperature	-13.138	<.0001	<b>Face Humidity</b>	4.811	0.0460
	<b>Face Humidity</b>	-10.016	0.0031			
	Press Open Time	5.471	0.0056			
Face Humidifier Temperature	19.560	<.0001				
<b>Important Regression Statistics</b>						
<b>R<sup>2</sup><sub>a</sub></b>	0.808614			<b>R<sup>2</sup><sub>a</sub></b>	0.747666	
<b>d.f.</b>	42			<b>d.f.</b>	43	
<b>P</b>	13			<b>P</b>	10	
<b>VIF<sub>max</sub></b>	4.5371586			<b>VIF<sub>max</sub></b>	5.5493187	
<b>RMSE</b>	6.233895			<b>RMSE</b>	6.573086	
<b>Residual Pattern</b>	Homogeneous			<b>Residual Pattern</b>	Homogeneous	

For the 0.500” product type (**Figure 10**), the slopes of the IB percentiles are 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” 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” product type. It may also indicate that the product is 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 is used for illustration purposes, the quantile regression algorithm in R can also be applied to multiple independent variable models. Further analysis is conducted to examine the differences between the MLR and the QR median, 10<sup>th</sup> and 90<sup>th</sup> percentile fits. For the 0.6875” 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 are 66.53%, 23.16%, and 16.14%, respectively.

Table 5. MLR and QR models for product type 0.6875”

Variables	Coefficients			
	MLR Average	QR Median	QR 10 <sup>th</sup> percentile	QR 90 <sup>th</sup> percentile
Intercept	-1231.75	-1556.92	-692.31	-1693.05
<b>Face Scavenger Resin %</b>	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
<b>Face Humidity</b>	-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

For the 0.500” 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 are 36.58%, 22.39%, and 15.60%, respectively. A comparison of the 10<sup>th</sup> and 90<sup>th</sup> percentiles (**Tables 5 and 6**) of the coefficients gives a good method for 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.

Table 6. MLR and QR models for product type 0.500”

0.500” Variables	Coefficients			
	MLR Average	QR Median	QR 10 <sup>th</sup> percentile	QR 90 <sup>th</sup> 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
<b>Face Humidity</b>	2.37	1.74	0.31	4.58

A 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., 1<sup>st</sup>, 5<sup>th</sup>, 99<sup>th</sup>, etc.) with respect to similar variables. A more detailed examination of each quantile may provide useful insight for root-cause analysis of sources of variation.

### 3.7 CONCLUSIONS FOR CHAPTER 3

The wood composites industry is undergoing unprecedented change in the forms of corporate divestitures and consolidation, real increases in the costs of raw material and energy, and extraordinary international competition. The forest products industry is important to the U.S. economy. The challenge for this industry for maintaining business competitiveness is to develop a more advanced knowledge of causality between the complex nature of process variables and final product quality characteristics. It may be very important to examine this causality in the percentiles of final product quality characteristics. This chapter provides Quantile Regression (QR) statistical methods that can improve business competitiveness in the wood composites industry.

Multiple Linear Regression models (MLR) and QR models are developed for the Internal Bond (IB) of Medium Density Fiberboard (MDF). The models are developed from a manufacturing data set for a North American MDF producer. The data set aligned the IB of MDF with 184 different independent variables that are on-line sensors located throughout the process, i.e., from refining to final pressing. MLR models are developed for MDF product types that are distinguished by thickness in inches, i.e., 0.750", 0.6875", 0.625" and 0.500". A best model criterion is used with all possible subsets. QR models are developed for each product type for the most common independent variable of the MLR models.

Common independent variables for the 0.750" and 0.625" MLR models are "Refiner Resin Scavenger %" and "Core Water to Wood". The scaled estimates for "Refiner Resin Scavenger %" differed in sign for each product type. The "Refiner Resin Scavenger %" has a negative scaled estimate of -9.12 p.s.i. on IB for 0.750" while the "Refiner Resin Scavenger %" has a positive scaled estimate of 8.40 p.s.i. on IB for 0.625". This may indicate some volatility in IB for these product types for this common independent variable. We found that the influence of "Refiner Resin Scavenger %" on the lower percentiles of IB is quite different than the mid-range and higher percentiles. For the 0.750" product type, the median and average models fit have similar slopes. The 5<sup>th</sup> percentile (possible IB failures) and 95<sup>th</sup> percentile (extreme IB strength) behave quite differently from the inner percentiles. For the 0.625" product type the slopes of the percentiles are 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 50<sup>th</sup> percentile (median) the effect of "Refiner Resin Scavenger %" has a stronger positive influence on IB the higher the

percentile. For percentiles below the 50<sup>th</sup> percentile (median) the effect of “Refiner Resin Scavenger %” has a stronger negative influence on IB the higher the percentile. The QR analyses suggest that opportunities exist for additional root cause investigation of the sources of IB variability from “Refiner Resin Scavenger %”.

For the MLR and QR models that included all significant variables, it appears that for the “0.750” product type more investigation is needed to determine the true effects of “Dryer 1 Fan Current”, “Dryer 2 Fan Current” and “Core Water To Wood”. These variables have discrepancies of 39.84%, 34.47%, and 28.2%, respectively. For the “0.625” product type more investigation is needed to determine the true effects of “Shavings Raw Weight”, “Relative Ambient Humidity” and “Weight Actual”. These variables have discrepancies of 42.96%, 16.16%, and 12.78%, respectively. These discrepancies reflect the opportunity for key parameters to be incorrectly modeled, possibly resulting in inefficiency and a higher overall cost for the producer. Further analysis could be conducted to examine each quantile with respect to these same key variables, and would perhaps provide further insight into the process of interest.

“Face Humidity” is common for both 0.6875” and 0.500” product types. The scaled estimates for “Face Humidity” differ in sign for each product type. The “Face Humidity” has a negative scaled estimate of -10.02 p.s.i. on IB for 0.6875” while the “Face Humidity” has a positive scaled estimate of 4.81 p.s.i. on IB for 0.500”. This may also indicate that “Face Humidity” is an important source of variability between the two product types that the manufacturer needs to investigate. The average and median fits for 0.6875” 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 5<sup>th</sup> and 95<sup>th</sup>

percentiles of IB is quite different than the inner percentiles. For the 0.500” product type, the slopes of the IB percentiles are very similar for all of the percentiles for “Face Humidity”. The median and average have different scales, which may imply non-normality in the response variable IB. The QR analysis for 0.500” 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” product type. It may also indicate that the product is easier to make between production runs as “Face Humidity” changes.

When QR models are compared with the significant variables of the MLR models for the “0.6875” product type there is a significant discrepancy in the influence of IB by “Face Humidifier Temperature”, “Core Scavenger Resin Flow” and “Dryer Mass Flow”. The discrepancies in the coefficients for these three process variables are as large as 66.53%, 23.16%, and 16.14%, respectively. This discrepancy between the mean and median influence on IB also exists for the 0.500” product type for the variables “Face Humidity”, “Mat Shave Off Target” and “Refiner S Steam Flow”. These variables have discrepancies in the coefficients of 36.58%, 22.39%, and 15.60%, respectively. These discrepancies further highlight the risk associated with making decisions on the mean of the distribution.

The aforementioned quantile regression methods used in conjunction with classical multiple linear regression analysis can improve forest products manufacturers’ knowledge of process variation. An improved knowledge of process variation can lead to variation reduction and costs savings, both vital for long-term sustained business competitiveness of this important industry.

# CHAPTER 4

## Predictive Modeling using Quantile Regression

### 4.1 COMPARING PREDICTIVE MODELING OF MULTIPLE LINEAR REGRESSION WITH QUANTILE REGRESSION MODELS FOR THE IB OF MEDIUM DENSITY FIBERBOARD

The biggest challenge facing North American Medium Density Fiberboard (MDF) manufacturers is identifying, quantifying and controlling sources of variation within their processes. There are hundreds of process variables that may impact the final output of any industrial process. It is vital for competitiveness that MDF manufacturers understand the causality of which process variables significantly influence final product quality characteristics (e.g., IB). Quantifying causality and possibly predicting final product quality outcomes is vastly important to the MDF manufacturer for improving process efficiency, lowering defects, lowering energy and raw material costs, and sustaining business competitiveness

A traditional and popular method of predictive modeling is Multiple Linear Regression (MLR). Recall from Chapter 3 that MLR has three important assumptions: 1) linearity of the coefficients; 2) normal or Gaussian distribution for the response errors ( $\varepsilon$ ); and 3) the errors  $\varepsilon$  have a common distribution. In a MDF industrial setting, when modeling a quality characteristic such as the IB of MDF, these assumptions may not always be valid. The Quantile Regression (QR) method may be a more appropriate modeling method for the IB of MDF because it does not have the stringent assumption of normality in the response variable along with the other critical assumptions associated with MLR. QR also allows MDF manufacturers to examine causality beyond the mean of the distribution.

Examining causality of the 50<sup>th</sup> percentile or median of IB may be more realistic for MDF manufacturers, as well as improving the understanding of causality in the outer 5<sup>th</sup> (possible failing IB) and 95<sup>th</sup> percentiles (extreme IB strength).

This chapter examines predictive modeling of IB of MDF for four product types using QR for the 0.5 quantile (or median). MLR predictive models of the mean IB are compared with QR predictive models of the median IB. This chapter builds upon the research presented in Chapter 3.

## 4.2 METHODS

The IB of four different product types of MDF are analyzed using both MLR and QR predictive models. Each product type represents a different board thickness in inches (i.e., 0.750", 0.625", 0.6875" and 0.500"). As previously discussed in Chapter 3, we use SAS Business Intelligence and Analytics Software ([www.sas.com](http://www.sas.com)) and seven criteria in selecting the best model of IB. There are 56 records of IB in the training set for 0.750", 51 records of IB in the training set for 0.625", 73 records of IB in the training set for 0.6875", and 74 records of IB in the training set for 0.500". The most recent set of 20 continuous records are held from each product type to be used as a model validation sample before selecting the best model.

This method is referred to as cross-validation (Kutner et al. 2004). A validation sample is simply a sample that is withheld from the estimation of a regression model. The model developed is then used to predict the true values of the records withheld. Statistics such as  $R^2_{validation}$  (coefficient of determination for the validation sample) and Root Mean Square Error of the Predicted (RMSEP) are calculated for the validation data set to compare the performance of the training models. The formula for  $R^2_{validation}$  is equivalent to that

mentioned in equation [5], and is only calculated for the validation set of records. The RMSEP statistic is:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n}} \quad [11]$$

where,  $n$  is the number of observations,  $Y_i$  is the  $i$ -th actual value, and  $\hat{Y}_i$  is the  $i$ -th predicted value.

### 4.3 RESULTS AND DISCUSSION

#### Product type 0.750<sup>12</sup>

For the 0.750<sup>12</sup> product type a MLR training model is developed with a  $R^2$  of 0.89, 44 degrees of freedom and 11 parameters. The RMSE of the model is 5.95 p.s.i. and the maximum VIF for any independent variable is 2.78. Residual patterns for the MLR training model are homogeneous (**Table 7**). A QR (median) training model is developed with a  $R_M^2$  of 0.86<sup>12</sup>, 44 degrees of freedom and 11 parameters. The RMSE of the QR (median) training model is 5.59 p.s.i. and residual patterns for the QR (median) training model are homogeneous (**Table 7**). The  $R_{validation}^2$  and RMSEP for the MLR validation model are 0.40 and 26.53 p.s.i., respectively (**Figure 11**). The  $R_{validation}^2$  and RMSEP for the QR (median) validation model are 0.40 and 26.54 p.s.i., respectively (**Figure 12**). These statistics are very similar given the normality of IB in the training data set (**Figure 13**). The p-value for the Shapiro-Wilks test for normality of the training data IB is 0.31, i.e., cannot reject the null hypothesis that IB is Gaussian.

---

<sup>12</sup> The  $R_M^2$  statistic for the QR regression model is calculated using the coefficient of determination formula or  $R^2$  and replacing the mean with the median statistic.

Table 7. MLR and QR models for product types 0.750”

<b>Training</b>					
	<b>MLR</b>	<b>Estimate</b>	<b>p-value</b>	<b>QR</b>	<b>Estimate</b>
<b>P A R A M E T E R S</b>	Face Fiber Temperature	-0.276534	<.0001	Face Fiber Temperature	-0.24617
	Dryer ESP Outlet Temperature	-0.325384	0.0005	Dryer ESP Outlet Temperature	-0.29706
	Core Humidifier Temperature	-0.853387	0.0009	Core Humidifier Temperature	-0.97901
	FaceFiber Mat Moisture	7.9621925	0.0010	FaceFiber Mat Moisture	8.13247
	Press Postion Time	3.9645979	<.0001	Press Postion Time	3.96844
	Press Temperature	0.4717754	0.0117	Press Temperature	0.40643
	Weight Target	187.35323	0.0010	Weight Target	219.80313
	Core Blow Line Pressure	-1.87737	<.0001	Core Blow Line Pressure	-1.63350
	Dryer S Outlet Temperature	-0.847441	0.0156	Dryer S Outlet Temperature	-0.65215
	Refiner S Steam Flow	-0.00345	0.0001	Refiner S Steam Flow	-0.00277
Core Refiner Valve Position	-0.493938	0.0008	Core Refiner Valve Position	-0.59418	
<b>Important Regression Statistics</b>					
<b>R<sup>2</sup></b>	0.890014		<b>R<sub>M</sub><sup>2</sup></b>	0.864207	
<b>R<sup>2</sup><sub>a</sub></b>	0.862517		<b>R<sup>2</sup><sub>a</sub></b>	0.830258	
<b>d.f.</b>	44		<b>d.f.</b>	44	
<b>P</b>	11		<b>P</b>	11	
<b>VIF<sub>max</sub></b>	2.7811752		<b>VIF<sub>max</sub></b>	N/A	
<b>RMSE</b>	5.946879		<b>RMSE</b>	5.587963	
<b>Residual Pattern</b>	Homogeneous		<b>Residual Pattern</b>	Homogeneous	
<b>Validation</b>					
<b>R<sup>2</sup></b>	0.401682		<b>R<sup>2</sup></b>	0.403263	
<b>RMSEP</b>	26.53421		<b>RMSEP</b>	26.53579	

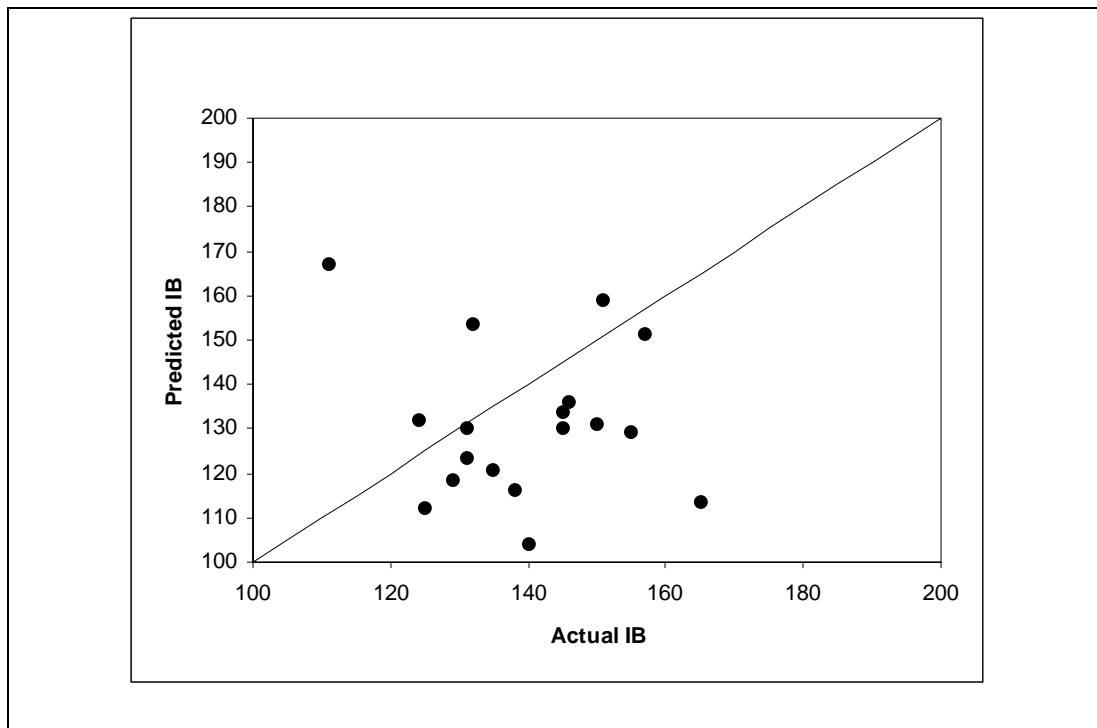


Figure 11. MLR validation of 0.750” actual and predicted IB.

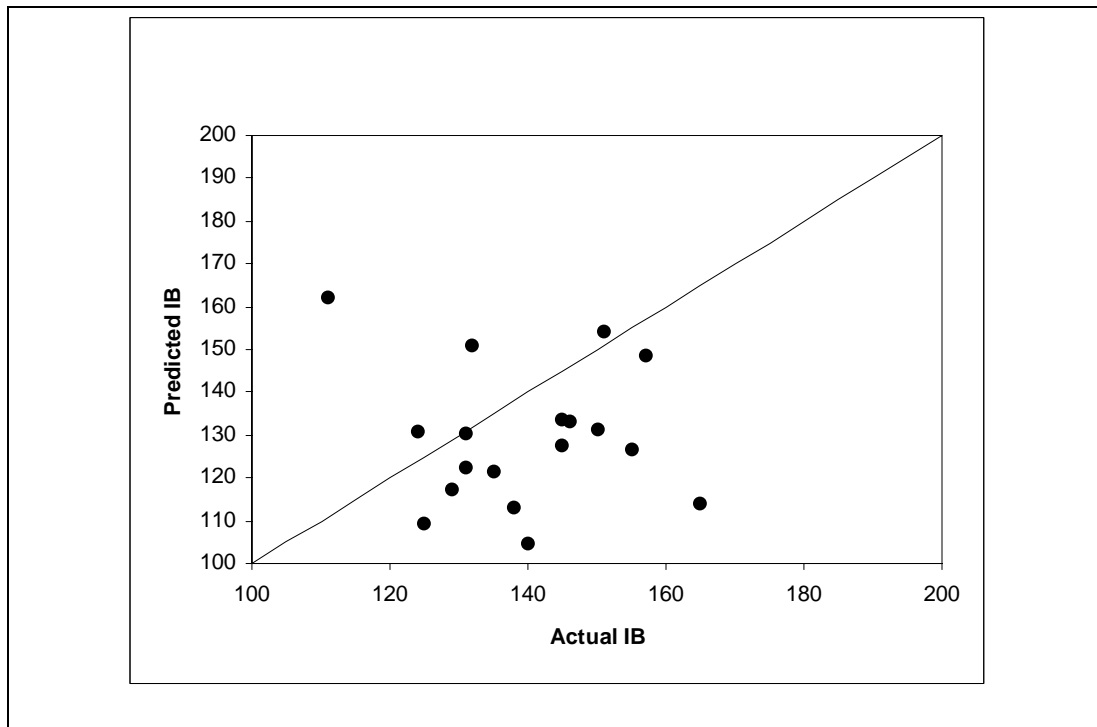


Figure 12. QR (median) validation of 0.750” actual and predicted IB.

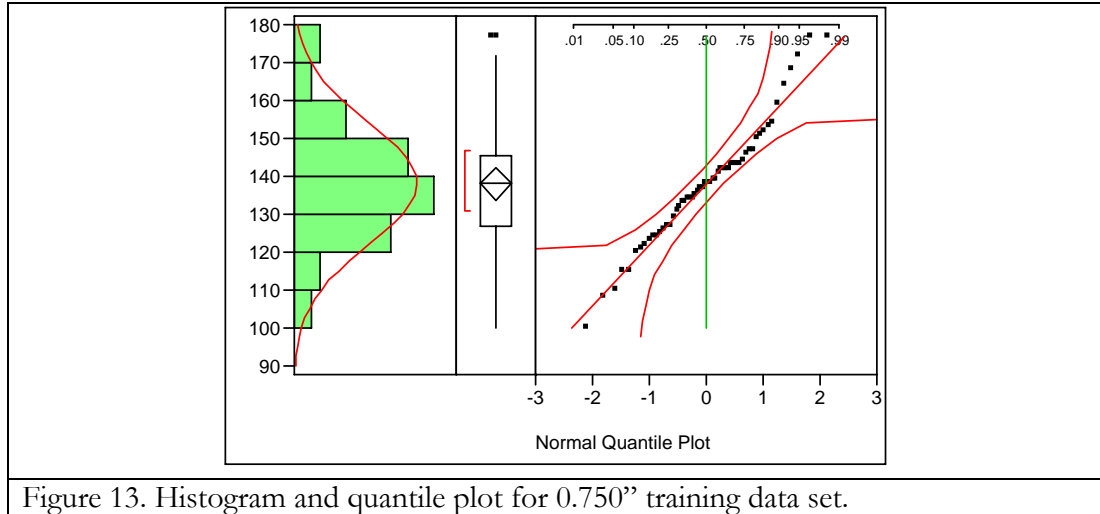


Figure 13. Histogram and quantile plot for 0.750” training data set.

As previously mentioned in Chapter 3, the MLR model is built by minimizing the sums of squares about the mean of the distribution. The QR model is built by minimizing the sums of absolute residuals about the median. Recall, when a data set is normally distributed, the mean and median are equivalent. Therefore, we would expect to see very similar regression models for the MLR and QR (median) fits for the 0.750” data set. The models are built using 11 parameters with only 44 degrees of freedom. Typically, one would like to see six to ten times as many data records as independent variables (parameters) (Kutner et al. 2004). However, in many industrial settings more parameters must be used in order to obtain a model with an acceptable  $R^2$  value. One risk associated with using too many parameters is known as “over-fitting”. This can result in data dependent models that may not predict well. This may explain why the prediction models for the 0.750” product type performed poorly and further investigation is warranted. One must also consider the process variation that may be present that is not measurable with current sensing technology, e.g., refiner plate wear, resin formation on fiber, etc.

### Product type 0.625”

For the 0.625” product type a MLR training model is developed with a  $R^2$  of 0.79, 62 degrees of freedom and 10 parameters. The RMSE of the model is 6.78 p.s.i. and the maximum VIF for any independent variable is 7.39. Residual patterns for the MLR training model are homogeneous (**Table 8**). A QR (median) training model is developed with a  $R_M^2$  of 0.78, 62 degrees of freedom and 10 parameters. The RMSE of the QR (median) training model is 6.57 p.s.i. and residual patterns for the QR (median) training model are homogeneous (**Table 8**). The  $R_{validation}^2$  and RMSEP for the MLR validation model are 0.60 and 26.92 p.s.i., respectively (**Figure 14**). The  $R_{validation}^2$  and RMSEP for the QR (median) validation model are 0.58 and 36.86 p.s.i., respectively (**Figure 15**). It is not surprising that the regression analysis descriptive statistics are similar given the normality of IB in the training data set (**Figure 16**). The p-value for the Shapiro-Wilks test for normality of the training data IB is 0.9738, i.e., cannot reject the null hypothesis that IB is Gaussian.

Table 8. MLR and QR models for product types 0.625”

<b>Training</b>					
	<b>MLR</b>	<b>Estimate</b>	<b>p-value</b>	<b>QR</b>	<b>Estimate</b>
<b>P A R A M E T E R S</b>	Face Resin to Wood Actual	15.400902	<.0001	Face Resin to Wood Actual	19.36181
	Main Motor Power	-0.036854	<.0001	Main Motor Power	-0.03641
	Former Thayer Weight	-0.629768	0.0414	Former Thayer Weight	-0.030011
	Press Steam Pressure	-1.311257	0.0001	Press Steam Pressure	-1.90398
	Weight Target	-284.6235	0.0007	Weight Target	-235.97566
	Resin Water Tank Temperature	-1.34079	<.0001	Resin Water Tank Temperature	-1.33890
	Swing Grinding Steam Flow	0.0101131	0.0003	Swing Grinding Steam Flow	0.01187
	Swing Main Motor Power	-0.016858	0.0007	Swing Main Motor Power	-0.01723
	Face Humidifier Temperature	0.8846549	<.0001	Face Humidifier Temperature	0.94011
	Weight Actual	218.06084	<.0001	Weight Actual	221.69205
<b>Important Regression Statistics</b>					
<b>R<sup>2</sup></b>	0.78516		<b>R<sub>M</sub><sup>2</sup></b>	0.780713	
<b>R<sup>2</sup><sub>a</sub></b>	0.750509		<b>R<sup>2</sup><sub>a</sub></b>	0.745344	
<b>d.f.</b>	62		<b>d.f.</b>	62	
<b>P</b>	10		<b>P</b>	10	
<b>VIF<sub>max</sub></b>	7.3850638		<b>VIF<sub>max</sub></b>	N/A	
<b>RMSE</b>	6.77959		<b>RMSE</b>	6.571053	
<b>Residual Pattern</b>	Homogeneous		<b>Residual Pattern</b>	Homogeneous	
<b>Validation</b>					
<b>R<sup>2</sup></b>	0.604067		<b>R<sup>2</sup></b>	0.583302	
<b>RMSEP</b>	26.92316		<b>RMSEP</b>	36.86166	

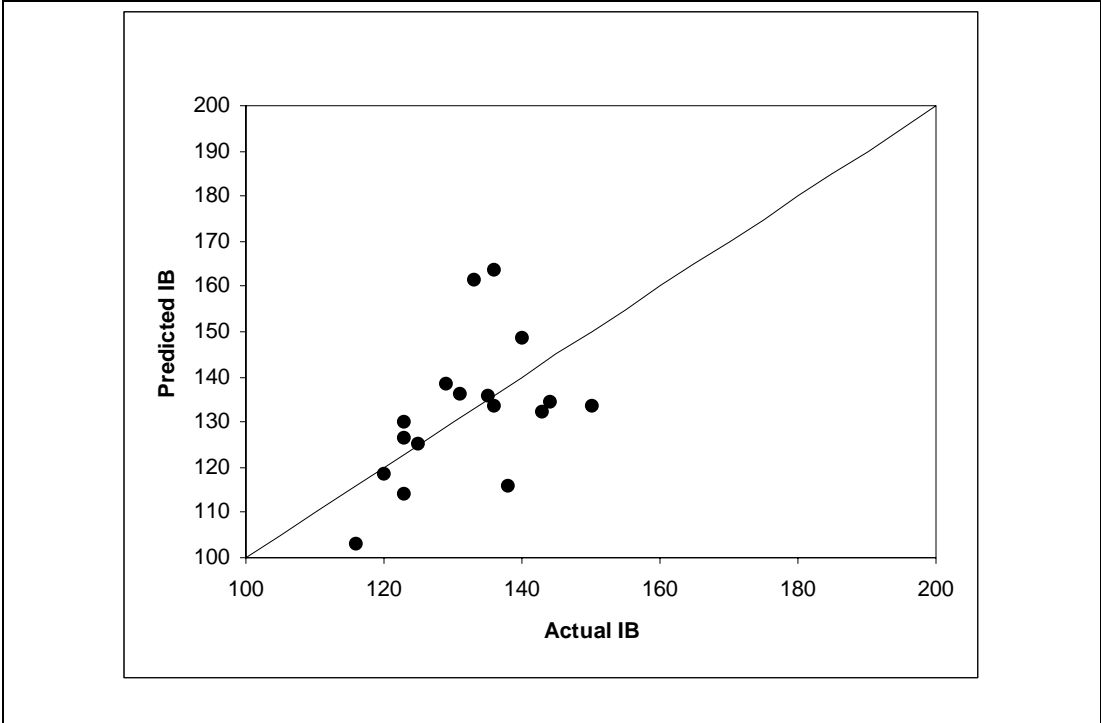


Figure 14. MLR validation of 0.625” actual and predicted IB.

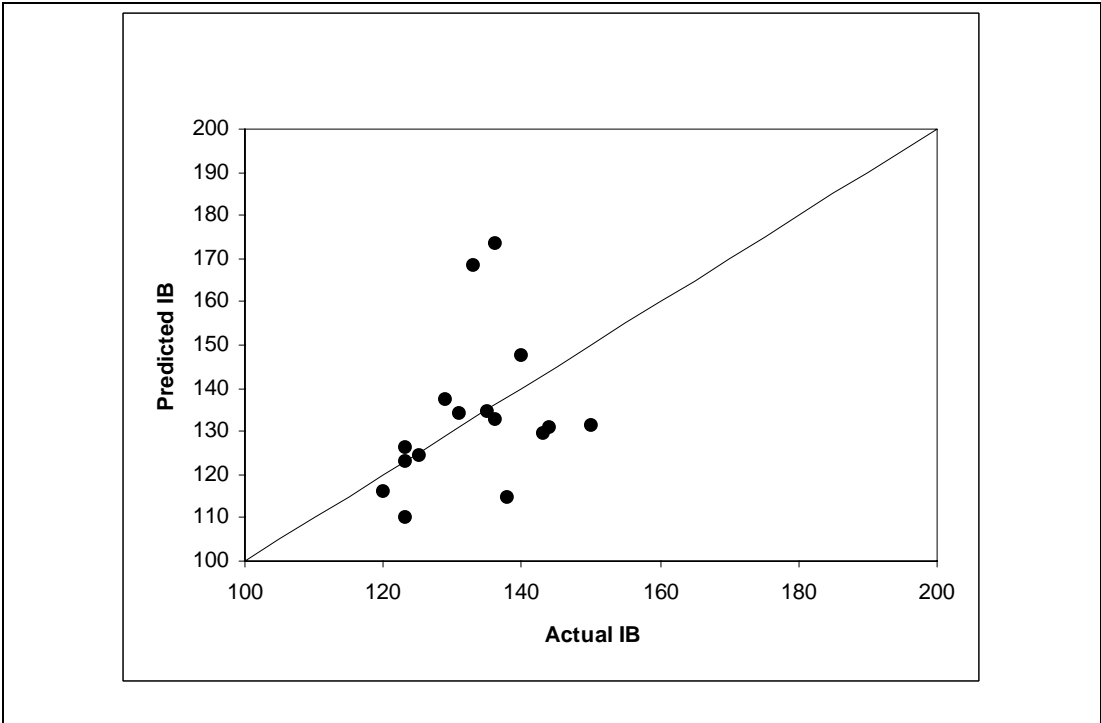
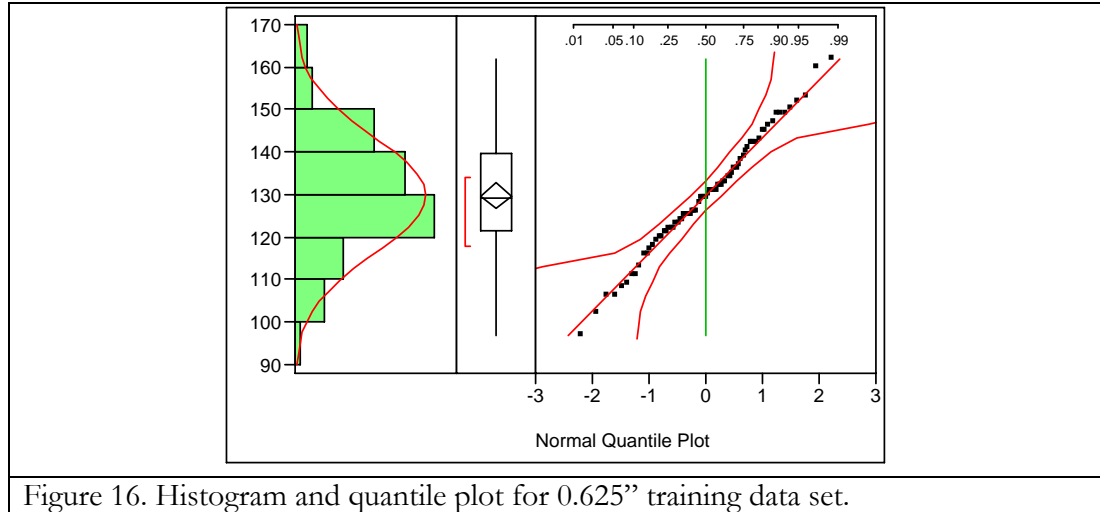


Figure 15. QR (median) validation of 0.625” actual and predicted IB.



Given the Gaussian characteristics of IB for the 0.625" MDF product, improvements in modeling using QR (median) is not possible for this data set. The MLR validation is slightly better ( $R^2_{\text{validation}} = 0.60$ ) than the QR (median) validation ( $R^2_{\text{validation}} = 0.58$ ). This is also reflected in the RMSEP statistic with MLR of 26.92 p.s.i. and QR (median) of 36.86 p.s.i. It is important for the practitioner to thoroughly understand the process being studied so the correct methods can be utilized when analyzing data. Given the normality of IB for 0.625" MDF, the mean may be a more efficient estimator of the central tendency of the distribution and MLR may be more appropriate for modeling this central tendency.

The training models for the 0.625" product are built using 73 records, 10 parameters and 62 degrees of freedom which is closer to the rule of thumb recommended by Kutner et al. (2004), i.e., six to ten times data records as many independent variables. However, as both Young and Guess (2002), and Young and Huber (2004) note, when working with real

world data from wood composite manufacturing environments it may not always be plausible to obtain a high  $R^2$  with few parameters.

### **Product type 0.6875”**

A MLR model for the 0.6875” product type is developed with a  $R^2$  of 0.64, 44 degrees of freedom and 6 parameters. The RMSE of the MLR model is 9.82 p.s.i. and the maximum VIF for any independent variable is 2.04. Residual patterns for the MLR training model are homogeneous (**Table 9**). A QR (median) training model is developed with a  $R_M^2$  of 0.62, 44 degrees of freedom and 6 parameters. The RMSE of the QR (median) training model is 9.52 p.s.i. and residual patterns for the QR training model are homogeneous (**Table 9**). The  $R_{validation}^2$  and RMSEP for the MLR validation model are 0.57 and 27.85 p.s.i., respectively (**Figure 17**). The  $R_{validation}^2$  and RMSEP for the QR (median) validation model are 0.55 and 25.90 p.s.i., respectively (**Figure 18**). The regression statistics are quite similar given the normality of IB in the training data set (**Figure 19**). The p-value for the Shapiro-Wilks test for normality of the training data IB is 0.8240.

There was strong evidence of model bias for the 0.6875” product upon examination of the validation plots, i.e., predictions of IB from both the MLR and QR (median) models over-estimate actual IB. It is not known given the original data records what causes the model bias for 0.6875” but it reflects that other variables not recorded in the data set are acting on the variability of IB. As previously noted, this may be refiner plate wear, fiber quality change, resin quality change, etc. Real-time sensing technology does not exist for process variables such as refiner plate wear, fiber quality change, and resin quality change.

Table 9. MLR and QR models for product types 0.6875”

<b>Training</b>					
	<b>MLR</b>	<b>Estimate</b>	<b>p-value</b>	<b>QR</b>	<b>Estimate</b>
<b>P A R A M E T E R S</b>	Refiner S Valve Position	-1.143532	0.0025	Refiner S Valve Position	-1.08402
	Core Fiber Wet Weight	-0.002447	<.0001	Core Fiber Wet Weight	-0.00191
	Core Humidifier Temperature	1.1782842	0.0346	Core Humidifier Temperature	1.65096
	Face Fiber Mat Moisture	16.350038	0.0005	Face Fiber Mat Moisture	12.65856
	Face Plug Feeder Screw Speed	-1.220197	0.0012	Face Plug Feeder Screw Speed	-1.37557
	E Emissions	1.2351385	0.0006	E Emissions	0.98901
<b>Important Regression Statistics</b>					
<b>R<sup>2</sup></b>	0.64452		<b>R<sup>2</sup></b>	0.616275	
<b>R<sup>2</sup><sub>a</sub></b>	0.596046		<b>R<sup>2</sup><sub>a</sub></b>	0.563949	
<b>d.f.</b>	44		<b>d.f.</b>	44	
<b>P</b>	6		<b>P</b>	6	
<b>VIF<sub>max</sub></b>	2.0449477		<b>VIF<sub>max</sub></b>	N/A	
<b>RMSE</b>	9.823333		<b>RMSE</b>	9.518499	
<b>Residual Pattern</b>	Homogeneous		<b>Residual Pattern</b>	Homogeneous	
<b>Validation</b>					
<b>R<sup>2</sup></b>	0.568123		<b>R<sup>2</sup></b>	0.54688	
<b>RMSEP</b>	27.84805		<b>RMSEP</b>	25.89948	

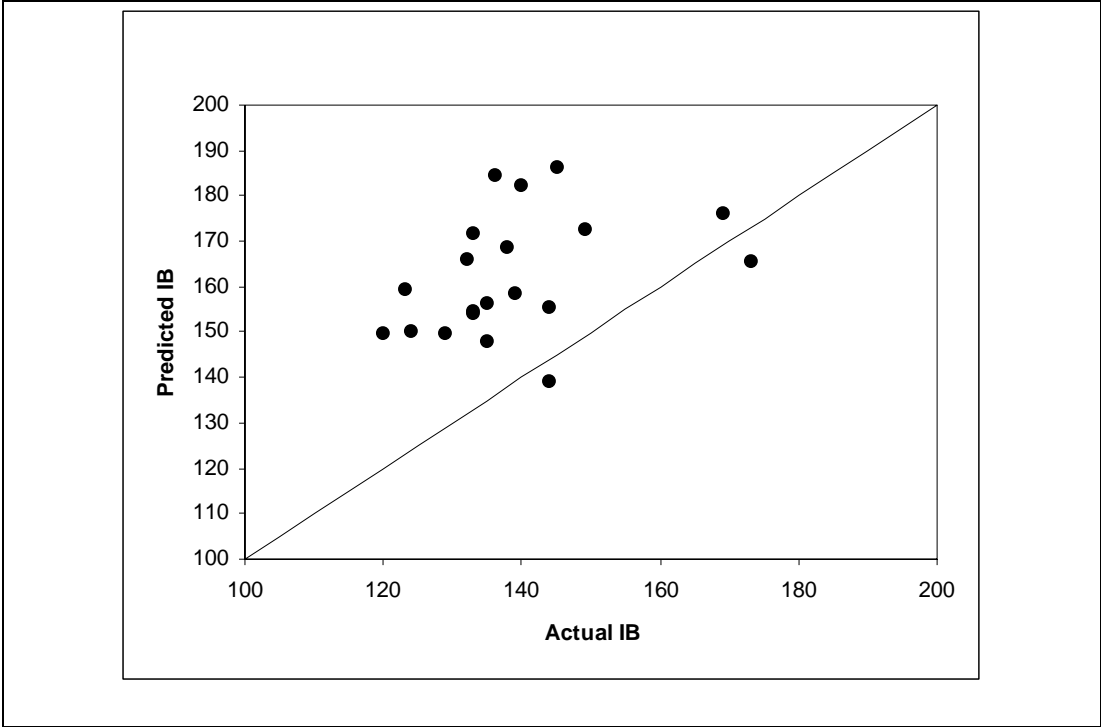


Figure 17. MLR validation of 0.6875” actual and predicted IB.

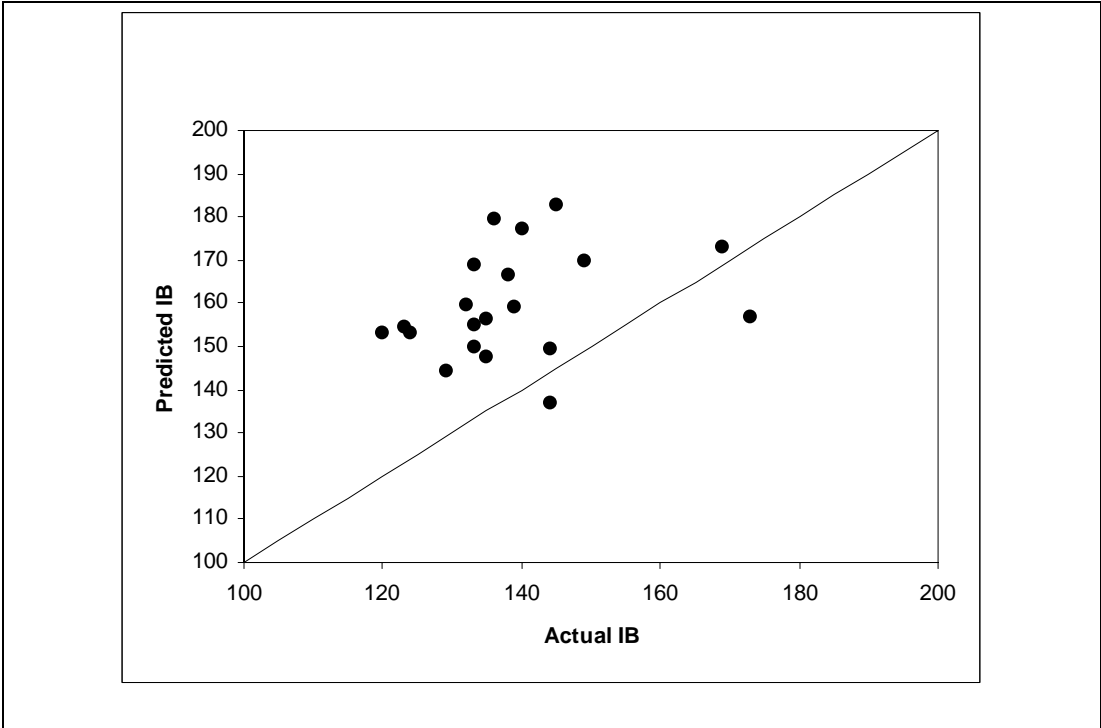
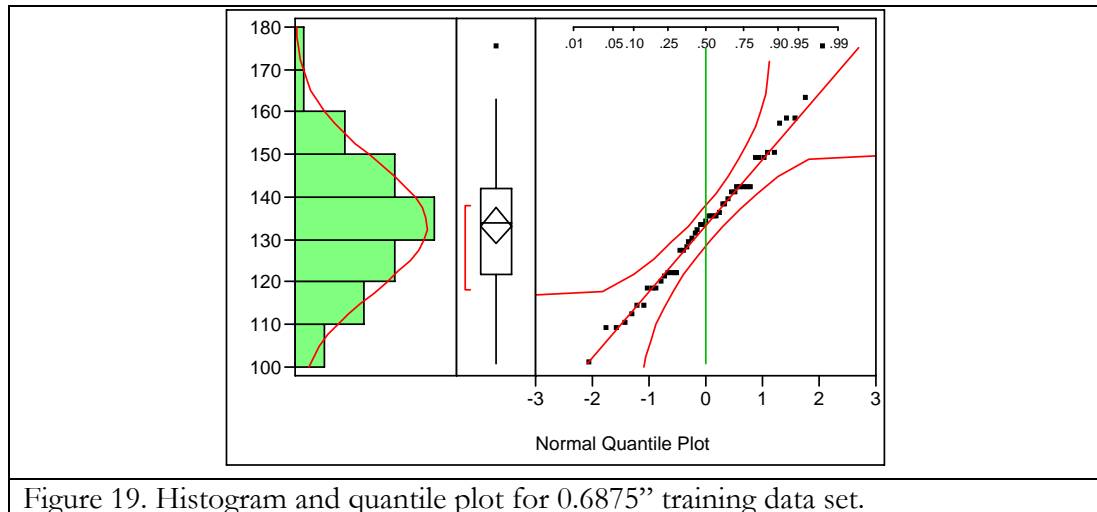


Figure 18. QR (median) validation of 0.6875” actual and predicted IB.



### Product type 0.500''

For the 0.500'' product type a MLR training model is developed with a  $R^2$  of 0.69, 63 degrees of freedom and 10 parameters. The RMSE of the model is 9.97 p.s.i. and the maximum VIF for any independent variable is 5.47. Residual patterns for the MLR training model are homogeneous (**Table 10**). A QR (median) training model is developed with a  $R_M^2$  of 0.67, 63 degrees of freedom and 10 parameters. The RMSE of the model is 6.64 p.s.i. and residual patterns for the QR (median) training model are homogeneous (**Table 10**). The  $R_{validation}^2$  and RMSEP for the MLR validation model are 0.64 and 23.63 p.s.i., respectively (**Figure 20**). The  $R_{validation}^2$  and RMSEP for the QR (median) validation model are 0.66 and 19.18 p.s.i., respectively (**Figure 21**). The p-value for the Shapiro-Wilks test for normality of the training data IB is 0.3837 (**Figure 22**). It is interesting to note that the IB for the 0.500'' product is the least Gaussian when compared to the other three product types.

Table 10. MLR and QR models for product types 0.500”

<b>Training</b>					
	<b>MLR</b>	<b>Estimate</b>	<b>p-value</b>	<b>QR</b>	<b>Estimate</b>
<b>P A R A M E T E R S</b>	Face Metering Bin Speed	-2.116445	<.0001	Face Metering Bin Speed	-1.69530
	Main Motor Power	0.044035	<.0001	Main Motor Power	0.03352
	Former Thayer Weight	-5.109695	0.0005	Former Thayer Weight	-4.56964
	Press Overall Time	1.2204688	<.0001	Press Overall Time	1.06912
	Press Temperature	0.9781079	<.0001	Press Temperature	1.08140
	Core Resin Pressure	-0.157047	<.0001	Core Resin Pressure	-0.13741
	Core Blow Valve Position	-0.290756	0.0265	Core Blow Valve Position	-0.21119
	Core Fiber Moisture	2.5739548	0.0003	Core Fiber Moisture	3.18516
	Core Refiner Feeder Screw Speed	-0.213207	0.0059	Core Refiner Feeder Screw Speed	-0.22404
Relative Humidity	3.5289234	<.0001	Relative Humidity	3.71690	
<b>Important Regression Statistics</b>					
<b>R<sup>2</sup></b>	0.688184		<b>R<sub>M</sub><sup>2</sup></b>	0.672784	
<b>R<sup>2</sup><sub>a</sub></b>	0.63869		<b>R<sup>2</sup><sub>a</sub></b>	0.620845	
<b>d.f.</b>	63		<b>d.f.</b>	63	
<b>P</b>	10		<b>P</b>	10	
<b>VIF<sub>max</sub></b>	5.469399		<b>VIF<sub>max</sub></b>	N/A	
<b>RMSE</b>	6.966834		<b>RMSE</b>	6.642721	
<b>Residual Pattern</b>	Homogeneous		<b>Residual Pattern</b>	Homogeneous	
<b>Validation</b>					
<b>R<sup>2</sup></b>	0.644817		<b>R<sup>2</sup></b>	0.660509	
<b>RMSEP</b>	23.63211		<b>RMSEP</b>	19.17969	

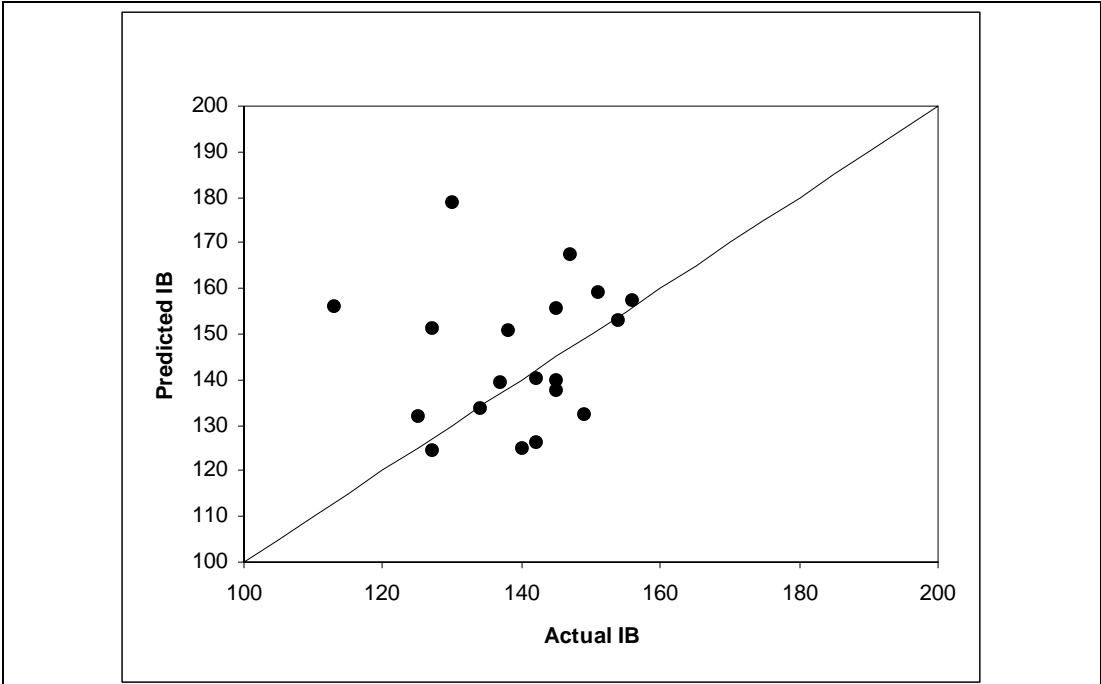


Figure 20. MLR validation of 0.500” actual and predicted IB.

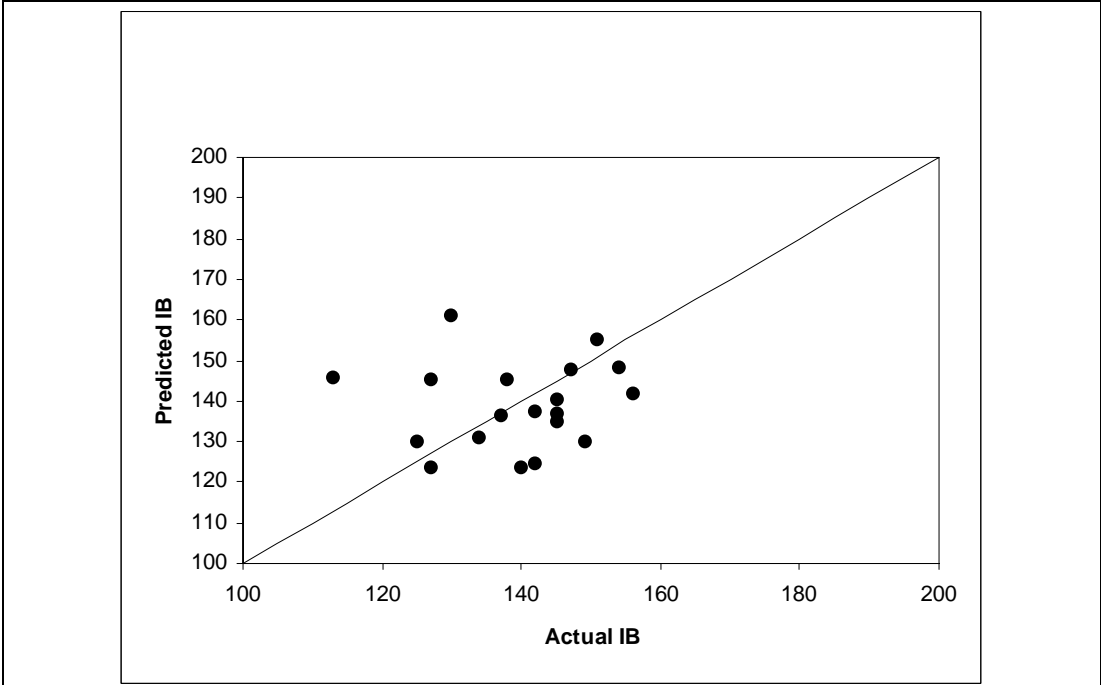


Figure 21. QR (median) validation of 0.500” actual and predicted IB.

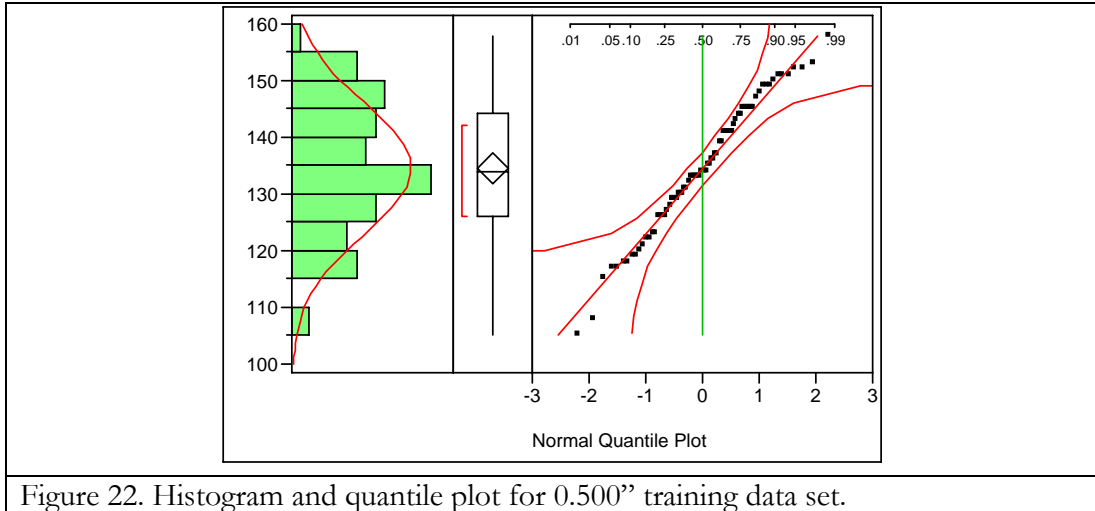


Figure 22. Histogram and quantile plot for 0.500” training data set.

The QR (median) validation model is slightly better than the MLR validation model as reflected by the discrepancies in the both the  $R^2_{validation}$  and RMSEP statistics (**Table 10**). Even though normality of IB for the 0.500” product cannot be rejected, the quantile plot of IB suggests that the product departs from normality in the upper and lower quantiles (**Figure 22**). In the case of 0.500” the results of the study indicate that QR (median) models may be better than MLR models when examining the central tendency of IB.

#### 4.4 CONCLUSIONS FOR CHAPTER 4

Chapter 4 compared MLR and QR (median) predictive models for the IB of MDF for the 0.750”, 0.625”, 0.6875” and 0.500” product types. The motivation for the chapter was driven by discussions with practitioners in the MDF industry.<sup>13</sup> Practitioners in the industry have a strong interest in real-time predictive modeling of the physical strength properties of MDF, e.g., IB. If feasible, real-time predictive modeling would improve the practitioners’ decision-making between destructive tests, which may be as long as two or three hours of production. In a modern large-capacity MDF plant two or three hours of

<sup>13</sup> Dougal Gillis, Technical Director of Langboard MDF, LLC, Willacoochee, Georgia. Ron Matthews, Technical Director of Langboard OSB, LLC, Quitman, Georgia.

production may represent hundreds of thousands of lineal feet of MDF product. Many practitioners have a working knowledge of MLR but are not familiar with the assumptions and limitations of MLR. This chapter highlights the limitations of MLR when modeling the central tendency of a response when the response departs from normality. QR models of the median may be better and more helpful for the practitioner.

For the 0.750” product type a MLR training model is developed with a  $R^2$  of 0.89, 44 degrees of freedom, 56 records and 11 parameters. The RMSE of the model is 5.95 p.s.i. and the maximum VIF for any independent variable is 2.78. Residual patterns for the MLR training model are homogeneous. A QR training model is developed with a  $R_M^2$  of 0.86, 44 degrees of freedom, 56 records and 11 parameters. The RMSE of the model is 5.59 and residual patterns for the QR training model are homogeneous. The  $R_{validation}^2$  and RMSEP for the MLR validation model are 0.40 and 26.53 p.s.i., respectively. The  $R_{validation}^2$  and RMSEP for the QR validation model are 0.40 and 26.54 p.s.i., respectively. These descriptive statistics are very similar given the normality of IB in the training data set.

For the 0.625” product type a MLR training model is developed with a  $R^2$  of 0.79, 62 degrees of freedom, 73 records and 10 parameters. The RMSE of the model is 6.78 p.s.i. and the maximum VIF for any independent variable is 7.39. Residual patterns for the MLR training model are homogeneous. A QR training model is developed with a  $R_M^2$  of 0.78, 62 degrees of freedom, 73 records and 10 parameters. The RMSE of the model is 6.57 p.s.i. and residual patterns for the QR training model are homogeneous. The  $R_{validation}^2$  and RMSEP for the MLR validation model are 0.60 and 26.92 p.s.i. The  $R_{validation}^2$  and RMSEP for the QR validation model are 0.58 and 36.86 p.s.i. We would expect these statistics to be very similar

given the normality of IB in the training data set. In this case, the MLR validation model is superior to the QR validation model of the median, with the largest discrepancy being in the RMSEP statistic.

For the 0.6875” product type a MLR training model is developed with a  $R^2$  of 0.64, 44 degrees of freedom, 51 records and 6 parameters. The RMSE of the model is 9.82 p.s.i. and the maximum VIF for any independent variable is 2.04. Residual patterns for the MLR training model are homogeneous. A QR training model is developed with a  $R_M^2$  of 0.62, 44 degrees of freedom, 51 records and 6 parameters. The RMSE of the model is 9.52 p.s.i. and residual patterns for the QR training model are homogeneous. The  $R_{validation}^2$  and RMSEP for the MLR validation model are 0.57 and 27.85 p.s.i., respectively. The  $R_{validation}^2$  and RMSEP for the QR validation model are 0.55 and 25.90 p.s.i., respectively. Again, we would expect these statistics to be very similar given the normality of IB in the training data set. For 0.6875” the MLR validation model is slightly superior to the QR validation model with the largest discrepancy being in the  $R^2$  statistic.

For the 0.500” product type a MLR training model is developed with a  $R^2$  of 0.69, 63 degrees of freedom, 74 records and 10 parameters. The RMSE of the model is 9.97 p.s.i. and the maximum VIF for any independent variable is 5.47. Residual patterns for the MLR training model are homogeneous. A QR training model is developed with a  $R_M^2$  of 0.67, 63 degrees of freedom, 74 records and 10 parameters. The RMSE of the model is 6.64 p.s.i. and residual patterns for the QR training model are homogeneous. The  $R_{validation}^2$  and RMSEP for the MLR validation model are 0.64 and 23.63 p.s.i., respectively. The  $R_{validation}^2$  and RMSEP for the QR validation model are 0.66 and 19.18 p.s.i., respectively.

For the 0.500” product the QR validation model is slightly superior to the MLR validation model, with discrepancies in the both the  $R^2$  and RMSEP statistics. This may be the result of departures in normality in the quantiles of IB for 0.500”.

As noted earlier in the chapter an important criterion for predictive MLR models is to have six to ten times as many data records as independent variables (parameters). This criterion was met when MLR models for the 0.750”, 0.625”, 0.6875” and 0.500” product types were built using 56, 73, 51 and 74 data records, respectively. The challenge for most industrial practitioners is to not “overfit” MLR models that result in weak validation performance. This chapter highlights the capabilities of QR (median) models as an alternative to MLR models of the mean central tendency when the response variables departs from normality. Future research work may explore examining other quantiles of the IB of MDF using QR. Understanding causality of process variables that influence IB in the outer percentiles (e.g., 5<sup>th</sup> percentile representing possible IB failures or 95<sup>th</sup> percentiles representing extreme IB strength) may be very important for the practitioner.

# CHAPTER 5

## Using R Software for Reliability Data Analysis

### 5.1 INTRODUCTION AND MOTIVATION

The software package R provides an “Open Source” option for those interested in statistical analysis and R is also free. The term “Open Source” is commonly applied to the source code of software that is made available to the general public with either relaxed or non-existent intellectual property restrictions. This allows users to create user-generated software content through either incremental individual effort, or collaboration ([http://en.wikipedia.org/wiki/Open\\_source](http://en.wikipedia.org/wiki/Open_source)). The package was originally developed by John Chambers at Bell Laboratories (formerly AT&T, now Lucent Technology) and can be viewed as an alternative implementation to the software package S-PLUS, <http://www.insightful.com/>. While much code is specific to the R package, there are also many S-PLUS commands that will run in R without being modified. R is capable of performing the standard exploratory data analyses such as histograms, box plots and probability plots as well as more complex analyses such as those involved in the study of reliability and quantile regression. Currently, R is being used at many prestigious universities including the University of California at Los Angeles (UCLA), <http://www.jstatsoft.org/v13/i07/v13i07.pdf> and is also being implemented at the Oak Ridge National Lab (ORNL) in Oak Ridge, TN (<http://www.ornl.gov/>). While there are numerous excellent statistical software packages on the market today, R provides a very economical option while continually updating with the latest tools. For more information about downloading R, visit <http://www.r-project.org/>.

Given that R software is “Open Source”, there is modest formal documentation for the package. This lack of documentation may create a steeper learning curve for the novice to moderate-level software programmer/user than other statistical programming packages. The advantages of R include its functionality at no direct cost. R provides a tremendous value to the user when compared to the sometimes higher cost of software packages such as SAS, MATLAB, Statistica, S-PLUS, etc.

There are several third-party, or independent, books written on R as well as many useful websites such as one hosted by the ORNL that can be found at <http://www.csm.ornl.gov/esh/aoed/>. Some excellent choices for an introduction to the R package include: *A Handbook of Statistical Analysis Using R* (Everitt and Hothorn 2006), and *Introduction to Statistics through Resampling Methods and R/S-plus* (Good 2005). There are also other books on the market, as well as online training courses that are devoted solely to the instruction of the R software package.

The American Statistical Association (ASA) often provides information about these online courses on the ASA website. For more information on the courses offered through ASA visit <http://www.amstat.org/education/index.cfm?fuseaction=learnstat>.

The coding protocol for R has been compared to S-PLUS, however R protocol may not be intuitive for the novice or moderate-level programmer. Data files for use by R must be stored in the specific subfolders for access to the data, and specific commands must be used for data retrieval.

Given that R is “Open Source” not all statistical analysis packages are automatically loaded with the original software download. When performing statistical analysis using R, it is imperative that the user locates on-line the specific statistical package of interest. This on-

line location information is found in the R documentation, <http://cran.r-project.org/src/contrib/PACKAGES.html>.

The “survival” package is used in this paper for the reliability analyses. Examples of downloading the “survival” package, importing data, loading data, “Create” function and “Write” function are presented in **Table 11**.

Table 11. General tutorial of installing R with code examples

<b>Step</b>	<b>Protocol</b>
<b>1: Install the appropriate package</b>	<pre>install.packages("survival")</pre> <p>You will be prompted to select a “Cran Mirror”. Choose a location close to your geographic location to ensure faster download speed.</p>
<b>2: Load the appropriate package</b>	<p>Click the “Packages” tab on the R console, then select “Load Packages” and “Survival”</p> <p><b>Note:</b> R log may instruct you to load an additional package to use the one you have originally requested. Load that package in the same way you attempted to load “survival”.</p>
<b>3: Load data</b>	<pre>data title=read.table("file name.txt")</pre> <p><b>Note:</b> No zeros or null fields are allowed in predictor variables. Also, the file must be in the R directory on your computer:</p> <p><b>Example:</b> C:/program files/R/rw2011</p>
<b>4: Create Function</b>	<p>Create function:</p> <pre>function name=function() {}</pre>

Table 11. Continued.

<b>Step</b>	<b>Protocol</b>
<p>5: <b>Write Function</b></p>	<p>Write function:  function name=edit(function name)</p> <p>A window will pop up and you write your function:</p> <pre>function() { x=test[,1]</pre> <p><b>Note:</b> always be sure at this step to use the original data name.</p> <p><b>Example:</b> data=read.table("test.txt")</p> <p><b>Note:</b> In the example above the name of the data file in the R subfolder is 'data', not test.</p> <pre>y=test[,2]</pre> <pre>function (the function you choose) }</pre> <p>Then type:  Function name ()  in the R console to use the function you just created.</p> <p><b>Note:</b> Any of the functions listed later in the paper can be used in this manner or typed directed into the program. However, it is much easier to manipulate the functions when they are stored in this form.</p>

## 5.2 EXPLORATORY DATA ANALYSIS FOR RELIABILITY

R software utilizes basic functions to allow for easy computation of descriptive statistics such as mean, median, minimum, maximum, quantiles and variance (**Figure 23**).

```
>summary(y)
  V1
Min.   : 97.0
1st Qu.:127.0
Median:137.0
Mean   :137.3
3rd Qu.:147.0
Max.   :185.0

>quantile (y$V1)
 0%   25%   50%   75%  100%
 97   127   137   147   185

>var (y$V1)
[1] 195.7928
```

Figure 23. Example of summary output from R of descriptive statistics.

R has excellent plotting functionality for exploratory statistical and reliability analyses such as: normal probability plots, histograms, box plots, Weibull plots and Kaplan-Meier estimators (**Figures 24, 25, 26, 27 and 28**). The R code and commands for exploratory statistical analysis are quite intuitive (**Tables 12 and 13**).

The reliability tools discussed in the previous examples can be downloaded using the R “survival” package. More information on the “survival” package in R can be found in Chapter 9 of *A Handbook of Statistical Analysis Using R* (Everitt and Hothorn 2006) or on: <http://stat.ethz.ch/R-manual/R-patched/library/survival/html/survfit.html>. The R package is capable of analyzing many types of reliability data including censored and uncensored observations. The package can also perform hazard and survival analyses.

For illustrating R functionality in this paper we will use a data set that contains the tensile strength known as Internal Bond (IB) for Medium Density Fiberboard (MDF).

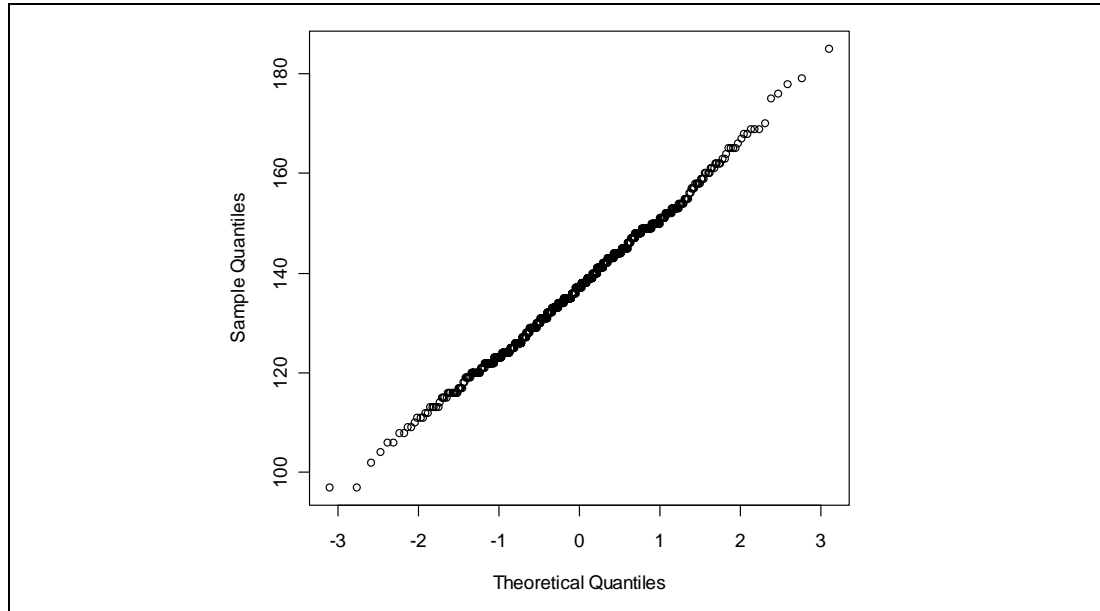


Figure 24. Example of normal Q-Q plot of internal bond of MDF using R code.

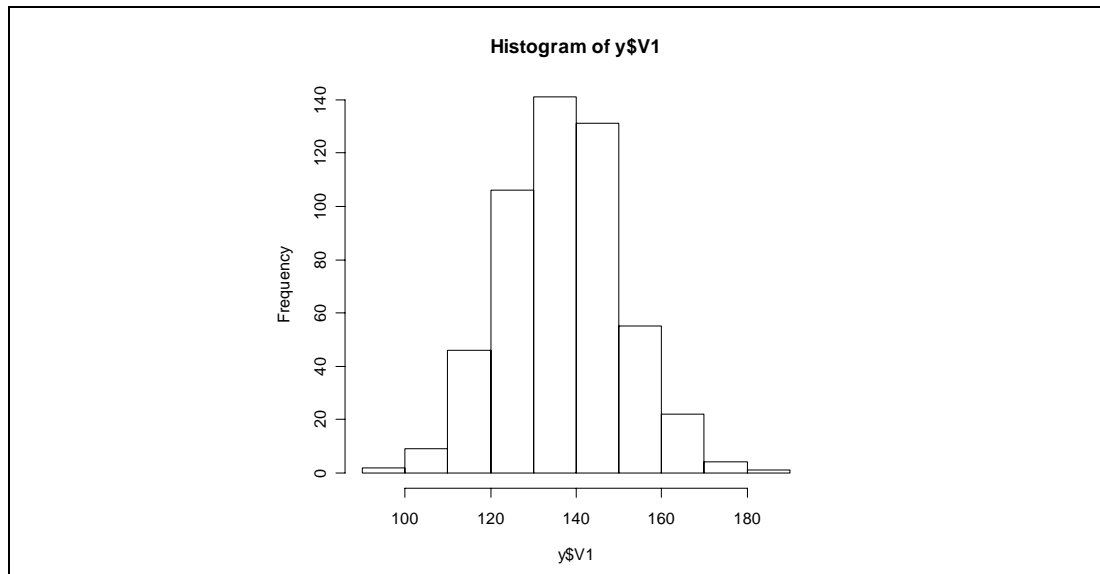


Figure 25. Example of histogram of internal bond of MDF using R code.

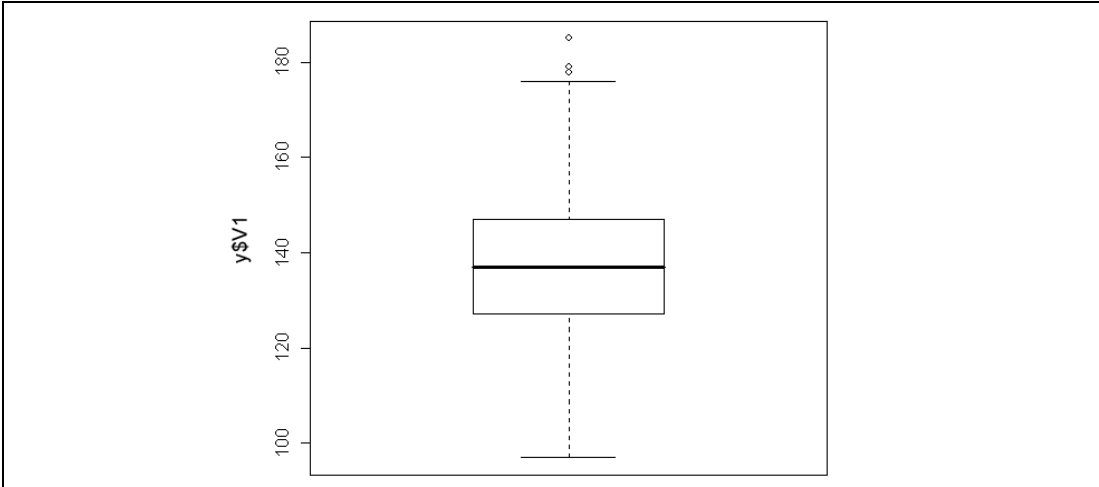


Figure 26. Example of box plot of internal bond of MDF using R code.

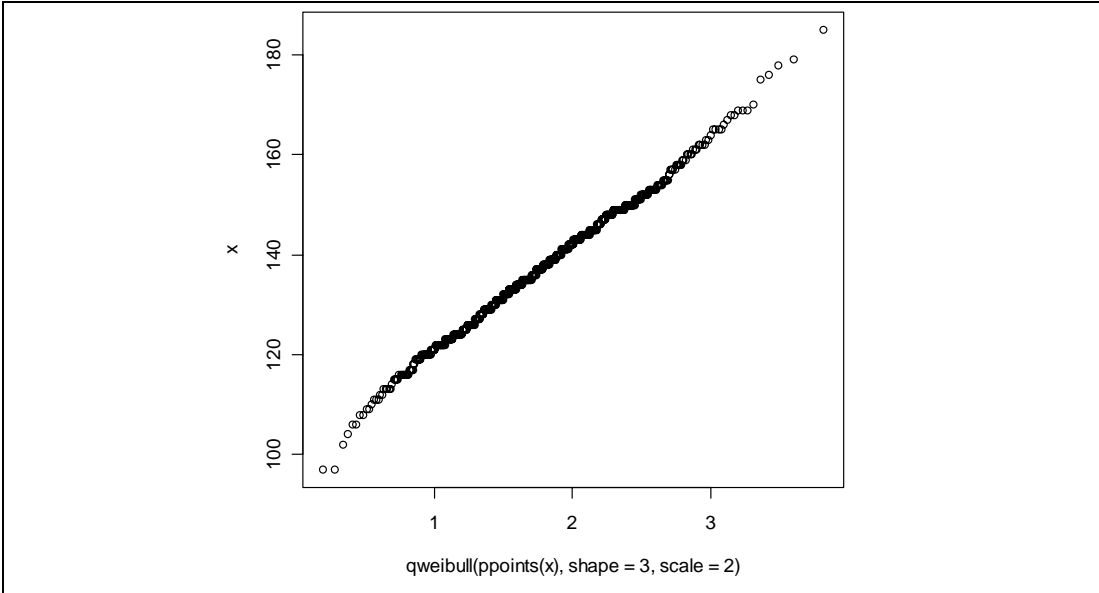


Figure 27. Example of Weibull Q-Q plot of internal bond of MDF using R code.

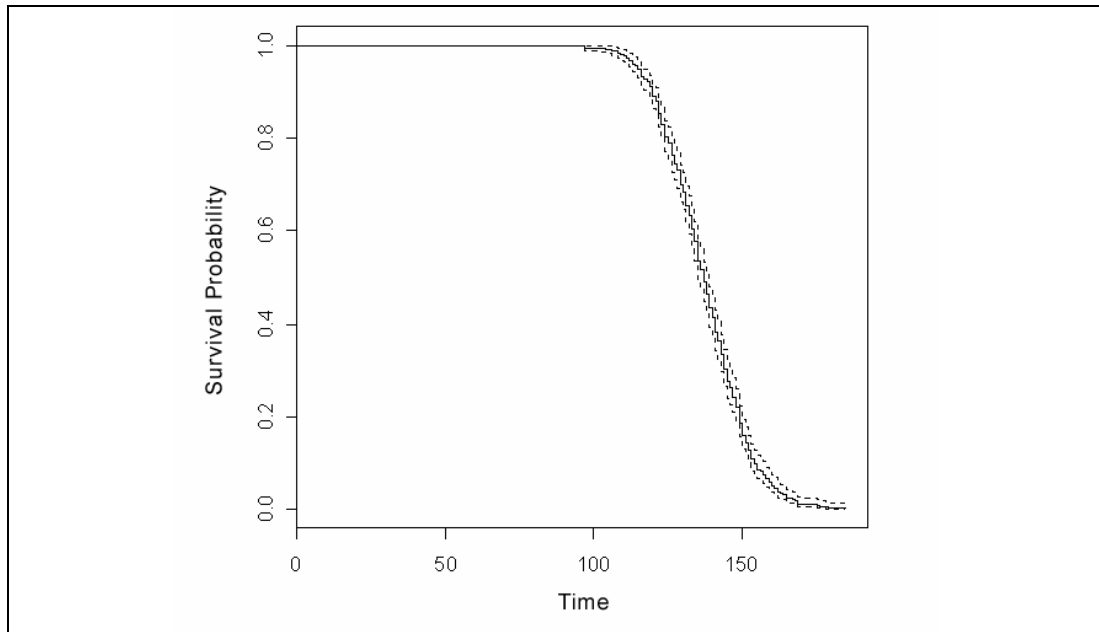


Figure 28. Example of Kaplan-Meier plot of internal bond of MDF using R code.

Table 12. Exploratory data analysis- basic statistics

Statistic	R Command
Mean & Median	<code>summary(DataName)</code>
Quantile	<code>quantile(DataName\$ColumnName)</code>
Variance	<code>var(DataName)</code>

Table 13. Exploratory data analysis- plots

Plot	R Command
Normal Probability Plot	<code>qqnorm(DataName\$ColumnName)</code>
Histogram	<code>hist(DataName\$ColumnName)</code>
Box plot	<code>boxplot(DataName\$ColumnName)</code>
Weibull Probability Plot	<code>x=sort(y\$V1)</code> <code>pp=ppoints(x)</code> <code>qqplot( qweibull(ppoints(x),</code> <code>shape=numeric, scale=numeric), x )</code>
Kaplan Meier Plot (uncensored data)	<code>fit &lt;- survfit(Surv(time))</code> <code>plot(fit)</code>
Kaplan Meier Plot (censored data)	<code>fit&lt;-survfit(Surv(time, status)~x,</code> <code>data=DataName)</code> <code>plot(fit)</code>

MDF is an engineered wood product formed by breaking down softwood into wood fibers, often in a defibrator, combining it with wax and resin, and forming panels by applying high temperature and pressure. MDF is a wood composite sheathing material similar in uniformity to plywood, but MDF is made up of separated fibers, not wood veneers and therefore doesn't have the structural strength properties of plywood. MDF is used for interior non-structural applications such as furniture, cabinets, non-structural doors, etc. MDF is denser than a complimentary interior, non-structural wood composite known as particleboard ([http://en.wikipedia.org/wiki/Medium-density\\_fibreboard](http://en.wikipedia.org/wiki/Medium-density_fibreboard)).

IB is a destructive tensile strength metric of product quality used by MDF producers reported in pounds per square inch (p.s.i.) or kilograms per cubic meter ( $\text{kg}/\text{m}^3$ ). Testing of the MDF product does not require any censoring.

### **5.3 MAXIMUM LIKELIHOOD ESTIMATES FOR THE WEIBULL DISTRIBUTION AND OTHERS**

#### **Weibull distribution**

The Weibull Distribution is often used in the analysis of lifetime, or reliability, data because of its ability to mimic the behavior of other distributions such as the normal or exponential simply by altering the parameters (Weibull 1939, 1951, 1961). The Weibull distribution is the most frequently used model for time (or pressure) to failure, perhaps followed by the lognormal distribution. The Weibull cumulative distribution function (cdf) giving the probability that a unit will fail by time  $t$  (or at pressure  $p$ ) is:

$$F(t) = 1 - \exp[-(t/\lambda)^k]. \quad [12]$$

The probability density function (pdf) of the Weibull is:

$$f(t) = (\kappa / \lambda)(t / \lambda)^{\kappa-1} e^{-(t/\lambda)^\kappa} . \quad [13]$$

The parameter  $\lambda$  is the distribution scale parameter and approximately equals the sixty-third percentile of the distribution. The parameter  $\kappa$  is the shape parameter (**Figure 29**).

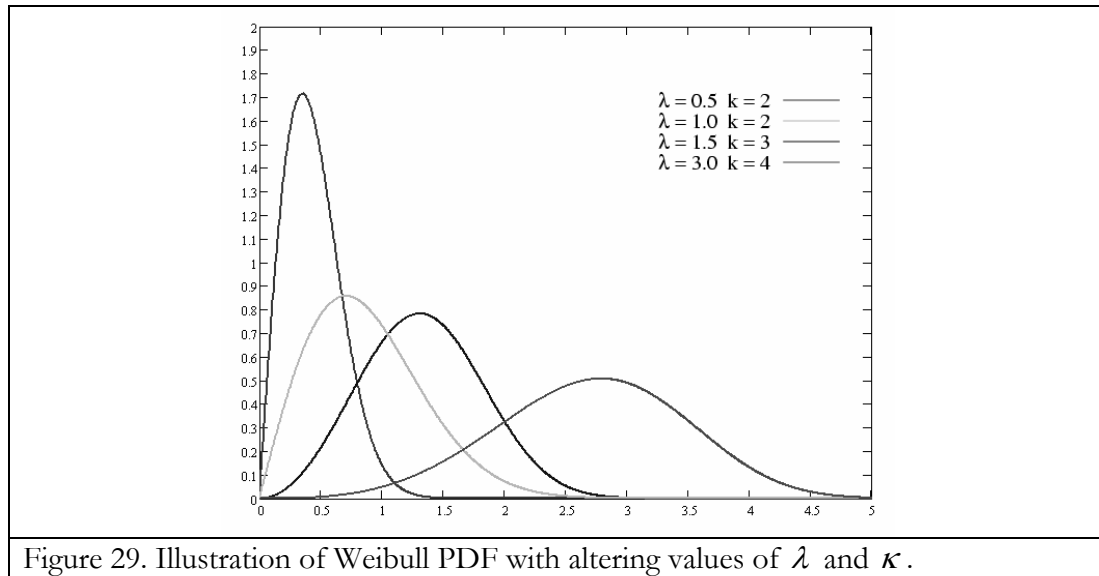
The hazard or instantaneous failure rate function for various Weibull distributions is:

$$h(t) = f(t)/[1 - F(t)] = (\kappa/\lambda)(t/\lambda)^{\kappa-1} . \quad [14]$$

The instantaneous failure rate is a measure of proneness to failure as a function of age (or pressure). The website, [www.weibull.com](http://www.weibull.com), is a very helpful resource for learning more about this distribution and its application to various reliability problems.

### **Reliability/Survival function and the Kaplan-Meier estimator**

The reliability/survival function captures the probability that the system will survive beyond a specified time (or pressure) to failure. Kaplan-Meier plots are one of the most popular survival plots. The Kaplan-Meier estimator (origin of Product Limit Estimator) estimates the [survival function](#) from life-time (or pressure to failure) data (Kaplan and Meier 1958). A plot of the Kaplan-Meier estimate of the survival function is the percent survival (Y) and life (or pressure) of the product at failure (t). The function is typically a declining function, i.e., as the products age or as pressure increases, the chance of survival declines.



For large enough samples it approaches the true survival function for that population. An important advantage of the Kaplan-Meier curve is that the method can take into account censored data, i.e., project failures before the final outcome is observed (Kaplan and Meier 1958).

Guess et al. (2003) published the first known work of applying reliability methods (e.g., Kaplan-Meier estimator) to the IB of MDF. Guess et al. (2003) discovered unusual crossings of the Kaplan-Meier estimators for similar products of MDF. These crossings represented differences in product quality that were not anticipated by the manufacturer. Guess et al. (2004) used forced censoring reliability methods to estimate bootstrap confidence intervals for the IB for MDF under different probability model assumptions. Bootstrap confidence intervals varied greatly depending on the model assumption, which is an important consideration for the manufacturers of MDF. Guess et al. (2004) also discovered that the lower percentiles of the IB for MDF fit different probability models when compared to the model fits for entire distribution. Guess et al. (2005) used reliability

methods and the mean residual life function for the IB for MDF to discover an unusual “J-shaped” mean residual life function that identified the inertia strength of MDF.

Chen et al. (2006) built upon the work by Guess et al. (2004) and investigated the lower percentiles of the IB for MDF. Chen et al. (2006) discovered that the best fit for the lowest one percentile of IB was the Weibull model and estimated 95% bootstrap confidence bounds for this lower one percentile of 91.8 p.s.i. and 97.4 p.s.i., respectively. Guess et al. (2006) further developed empirical mean residual life functions to discover crossing points as a method for establishing potential data driven specification limits (see Young and Guess, 1994 and Deming’s 1986, 1993 comments on specification limits).

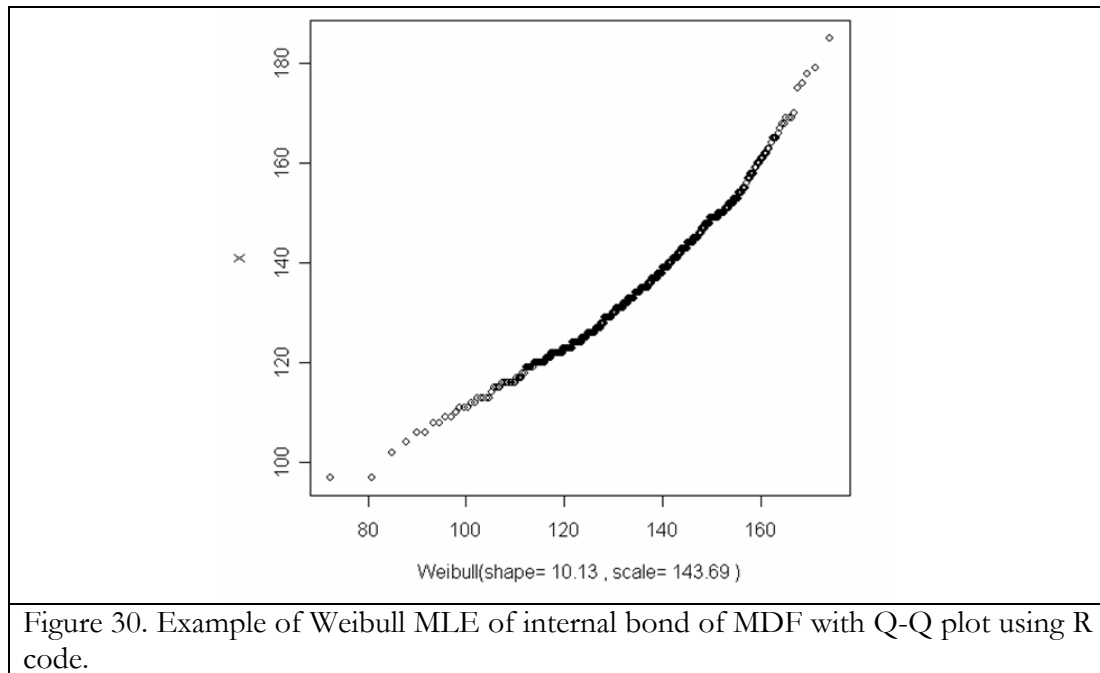
Wang et al. (2006) applied the Kaplan-Meier estimator to oriented strand board (OSB) destructive test data. Wang et al. (2006) found that 50% of the Parallel Elasticity Index (EI) of OSB can survive 57,856 pounds per inch (p.s.i.) and only 5% of the Parallel EI of OSB can survive at 65,435 p.s.i. Five percent of the IB for OSB failed before 33 p.s.i and 95% of OSB failed before a pressure of 68 p.s.i. The Kaplan-Meier estimator indicated that pressure to failure for the IB of OSB decreases at increasing rates between 35 p.s.i and 65 p.s.i.

### **Maximum likelihood estimation (MLE)**

Maximum likelihood estimation (MLE) is highly important in reliability analysis because it allows the practitioner to approximate the true parameters of the distribution and make inferences about the process, system, or component being studied. Statistical theory demonstrates that maximum likelihood estimators are both consistent and asymptotically efficient (Meeker and Escobar 1998). The R software packages allows user to easily calculate these estimates for both complete and censored data. For more information on censored

data analysis, visit: <http://www.csm.ornl.gov/esh/statoed/>. The forest products industry uses destructive testing, therefore we concentrate on complete data. However, with some manipulation of the R code, R can also calculate MLEs for censored data.

We use the Weibull distribution as an example given the findings of Chen et al. (2006) where the Weibull distribution was used to model IB pressure to failure for MDF. The MLE output for the IB data set was  $\kappa = 10.13$  (shape) and  $\lambda = 143.69$  (scale) and is included in **Figure 30**. The R code used for this analysis was presented in *Modern Applied Statistics with S* (Venables and Ripley 2002) and it can be found at [http://www.wessa.net/rwasp\\_fitdistrweibull.wasp?outtype=Browser%20Blue%20-%20Charts%20Whiten](http://www.wessa.net/rwasp_fitdistrweibull.wasp?outtype=Browser%20Blue%20-%20Charts%20Whiten) (Wessa 2006), also see Appendix C for the MLE R code.



## 5.4 CONCLUSIONS FOR CHAPTER 5

The R software package is a very powerful analytical tool and can be used for several different types of data analysis. R provides an “Open Source” option for those interested in statistical analysis while also being free. “Open source” describes the principles and methodologies to promote open access to the production and design process for various goods, products, resources and technical conclusions or advice. One of the most important facts is that R is user-generated and is created through collaboration. Therefore, R is constantly being updated with the most current functions and techniques.

The great advantage of the R software package is its ability to adapt to the ever-changing needs of the software user. Through collaboration of software programmers and insightful documentation, R is capable of meeting the needs and filling the niches of several separate software packages while remaining highly cost effective.

# CHAPTER 6

## Summary and Concluding Remarks

The purpose of this thesis is to explore just a few of the seemingly endless applications of Quantile Regression (QR), as well as uses of the easily accessible software package R. Often, practitioners or industries become comfortable with a particular set of statistical analyses using specific or company-directed software packages and are hesitant to investigate more advanced methods. The thesis seeks to illuminate some practical uses of new methods, which can be readily applied to many industrial processes. The methods and research of this thesis may provide MDF manufacturers with important techniques for quantifying unknown sources of variation in order to facilitate variation reduction, cost savings and continuous improvement.

Chapter 2 provides a concise account of the current literature pertaining to Medium Density Fiberboard (MDF), Multiple Linear Regression (MLR), and Quantile Regression (QR). Large-scale production of MDF began in the 1980s and has become one of the most highly-demanded composite wood materials. Given its excellent uniformity and versatility, MDF is an excellent base for veneers and laminates as well as non-structural constructions such as shelving, furniture and decorative molding. In 2004, the domestic production of MDF increased by 32.3% and is projected to continue this trend (Howard 2006). However, real prices of manufactured wood products are declining in an environment of higher energy and raw material costs (Howard 2006). This will pressure the competitive MDF manufacturer to focus on high quality, high production efficiency and lower costs of

manufacturing. The use of statistical techniques for continuous improvement is low risk and highly defensible.

One of the most popular and commonly used data mining techniques is Multiple Linear Regression (MLR). Much is published on MLR and its popularity may be due to it being a core course for undergraduates majoring in math, engineering or science. MLR can provide insightful information in cases when the rigid assumptions associated with it are met.

In chapter 3, we explore modeling the Internal Bond (IB) of MDF using Quantile Regression (QR) methods as compared to classical MLR models. MLR and QR models are developed for the IB of MDF. The data set used for the analysis aligns the IB of MDF with 184 different independent variables that are on-line sensors located throughout the process, i.e., from refining to final pressing. The MLR and QR models are developed using a best model criterion for all possible subsets of IB for four MDF thickness products reported in inches, e.g., 0.750", 0.625", 0.6875", and 0.500". The QR models are developed for the 50<sup>th</sup> percentile or median.

The adjusted coefficient of determination ( $R^2_a$ ) of the MLR models range from 72% with 53 degrees of freedom to 81% with 42 degrees of freedom, respectively. The Root Mean Square Errors (RMSE) range from 6.05 pounds per square inch (p.s.i.) to 6.23 p.s.i. A common independent variable for the 0.750" and 0.625" products is "Refiner Resin Scavenger %". QR models for 0.750" and 0.625" have similar slopes for the median and average but different slopes for the 5<sup>th</sup> and 95<sup>th</sup> percentiles. "Face Humidity" is a common predictor for the 0.6875" and 0.500" products. QR models for 0.6875" and 0.500" indicate

different slopes for the median and average with different slopes for the outer 5<sup>th</sup> and 95<sup>th</sup> percentiles.

Discrepancies between the coefficients derived from the MLR models and those derived from QR models of the median indicate QR may be more appropriate when the response variable departs from normality. These discrepancies signal a need for additional research of the sources of variation acting on the percentiles of IB. Improved knowledge of causality of the percentiles of IB may lead to variation reduction, costs savings and competitive advantage.

Chapter 4 examines the validity and predictability of QR (median) models as compared to MLR models for MDF. The MLR and QR validation models for the 0.750”, 0.625” and 0.6875” product types have  $R^2_{validation}$  ranging from approximately 40% to 60% and RMSEP ranging from 26.53 p.s.i. to 27.85 p.s.i.. The MLR validation model for the 0.500” product has a  $R^2_{validation}$  and RMSEP of 64% and 23.63 p.s.i. while the QR validation model has a  $R^2_{validation}$  and RMSEP of 66% and 19.18 p.s.i. The IB for 0.500” has the greatest departure from normality which is reflected in the results of the validation models. The results of this chapter provide further evidence that QR is a more defensible method for modeling the central tendency of a response variable when the response variable departs from normality.

In Chapter 5, reliability applications using the R software package are presented. This software package is an extremely powerful analytical tool and can be utilized for several different analyses. R is a user-generated, “Open Source”, option for those interested in statistical analysis while also being free. As R is created through collaboration, it is continuously being updated with the most current functions and techniques as they come

available. In this thesis, we utilize R for computing basic descriptive statistics, various data plots, and to calculate Maximum Likelihood Estimates (MLE) for the useful Weibull distribution. The great advantage of the R software package is its ability to adapt to the ever-changing needs of the software user.

In the rapidly changing and highly competitive global economy it is imperative that the wood composite industry utilize all analytical and statistical tools available in order to produce the highest quality products as efficiently as possible. This thesis highlights some of these important statistical tools. Improved product quality and more efficient use of valuable forest resources not only benefit the wood composites industry but also benefit the economy and society.

# BIBLIOGRAPHY

- Akaike, H. 1974. Factor analysis and AIC. *Psychometrika*. 52:317-332.
- André, N, T.M. Young, and T.G. Rials. 2006. On-line monitoring of the buffer capacity of particleboard furnish by near-infrared spectroscopy. *Applied Spectroscopy*. 60:1204-1209.
- Barnes, D. 2001. A model of the effect of strand length and strand thickness on the strength properties of oriented wood composites. *Forest Products Journal*. 51(9):36-46.
- Bernardy, G. and B. Scherff. 1998. Saving costs with process control, engineering and statistical process optimization. Proceedings 2<sup>nd</sup> European Panel Production Symposium (EPPS). Llandudno, Wales.
- Bernardy, G. and B. Scherff. 1999. Process modeling provides on-line quality control and process optimization in particle and fiberboard production. ATR Industrie-Elektronik GmbH&Co.KG – TextilstraBe. D-41751 Viersen Germany.
- Box, G. 1993. Quality improvement - the new industrial revolution. *International Statistical Review*. 61(1):3-19.
- Bozdogan, H. 1988. ICOMP: a new model selection criterion. *In: Bock, H.H., editor, Classification and Related Methods of Data Analysis*. Amsterdam, North-Holland. pp.599-609.
- Buchinsky, M. 1998. Recent advances in quantile regression models: a practical guideline for empirical research. *The Journal of Human Resources*. 33(1):88-126.
- Buhai, S. 2004. Quantile regressions: overview and selected applications. Unpublished manuscript. Tinbergen Institute and Erasmus University Rotterdam.
- Cade, B. and B. Noon. 2003. A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*. 1(8):412-420.
- Chen, W. 2005. A reliability case on estimating extremely small percentiles of strength data for continuous improvement of medium density fiberboard product quality. M.S. Thesis. University of Tennessee. Knoxville, TN.
- Chen, W., R.V. León, T.M. Young, and F.M. Guess, F.M. 2006. Applying a forced censoring technique with accelerated modeling for improving estimation of extremely small percentiles of strengths. *International Journal of Reliability and Application*. 7(1):27-39.

- 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 Products Journal*. *In Press*.
- Composite Panel Association. 2006. Second wave, the new generation of composite panel products. Composite Panel Association. Gaithersburg, MD.
- Deming, W.E. 1986. *Out of the Crisis*. Massachusetts Institute of Technology's Center for Advanced Engineering Design. Cambridge, MA.
- Deming, W.E. 1993. *The New Economics*. Massachusetts Institute of Technology's Center for Advanced Engineering Design. Cambridge, MA.
- Draper, N.R. and H. Smith. 1981. *Applied Regression Analysis*, 2<sup>nd</sup> Ed. John Wiley and Sons, Inc. New York, NY.
- Edwards, D.J. 2004. An applied statistical reliability analysis of the internal bond of medium density fiberboard. M.S. Thesis. University of Tennessee. Knoxville, TN.
- Efroymson, M.A. 1960. *Multiple Regression Analysis*. *In: Ralston, A. and Wilf, HS, editors, Mathematical Methods for Digital Computers*. John Wiley and Sons, Inc. New York, NY.
- Eriilsson, L., P. Hagberg, E. Johansson, S. Rannar, O. Whelehan, A. Astrom and T. Lindgren. 2000. Multivariate process monitoring of a newsprint mill. Application to modeling and predicting COD load resulting from de-inking of recycled paper. *Journal of Chemometrics*. 15:337-352.
- Everitt, B.S. and T. Hothorn. 2006. *A Handbook of Statistical Analysis using R*. Chapman and Hall. Boca Raton, FL.
- Feigenbaum, A.V. 1991. *Total Quality Control*. McGraw-Hill, Inc. New York, NY.
- Fitzenberger, B., R. Koenker, and J.A.F. Machado (editors). 2002. *Economic Applications of Quantile Regression*. Physica-Verlag Heidelberg. New York, NY.
- Good, P.I. 2005. *Introduction to Statistics through Resampling Methods and R/S-PLUS*. John Wiley and Sons, Inc. Hoboken, NJ.
- Gorr, W.L., and C. Hsu. 1985. An adaptive filtering procedure for estimating regression Quantiles. *Management Science*. 31(8):1019-1029.
- Green, H.M., and A.S. Kozek. 2003. Modeling weather data by approximate regression. *Quantiles*. *Anziam*. 44:C229-C248.
- Greubel, D. 1999. Practical experiences with a process simulation model in particleboard

- and MDF production. Proceedings 3<sup>rd</sup> European Wood-based Panel Symposium. Hanover, Germany.
- Guess, F.M., D.J. Edwards, T.M. Pickerell, and T.M. Young. 2003. Exploring graphically and statistically the reliability of medium density fiberboard. *International Journal of Reliability and Application*. 4(4):97-109.
- Guess, F.M., R.V. León, W. Chen, and T.M. Young, T.M. 2004. Forcing a closer fit in the lower tails of a distribution for better estimating extremely small percentiles of strength. *International Journal of Reliability and Application*. 6(4):79-85.
- Guess, F.M., X. Zhang, T.M. Young, and R.V. León. 2005. Using mean residual life functions for unique insights into strengths of materials data. *International Journal of Reliability and Application*. 6(4):79-85.
- Guess, F.M., J.C. Steele, T.M. Young, and R.V. León. 2006. Applying novel mean residual life confidence intervals. *International Journal of Reliability and Application*. 7(2):27-39.
- Honore, B, J. Powell, and S. Khan. 2002. Quantile regression under random censoring. *Journal of Econometrics*. 109(1):67-109.
- Howard, J. L. 2006. U.S. forest products annual market review and prospects, 2002–2006 Research Note FPL-RN-0302. Madison, WI: U.S. Department of Agriculture, Forest Service, Forest Products Laboratory. 9 pages.
- Humphrey, P.E. and H. Thoemen. 2000. The continuous pressing of wood-based composites: a simulation model, input data and typical results. In *Proceedings Pacific Rim Bio-Based Composites Conference* Australian National University Press. Canberra, Australia. pp. 303-311.
- Ishikawa, K. 1976. Guide to quality control. JUSE Press Ltd. Tokyo, Japan.
- Juran, J.M. and F.M. Gryna. 1951. *Juran's Quality Control Handbook*. McGraw-Hill Book Company. New York, NY.
- Kaplan, E.L. and P. Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*. 53:457-481.
- Koenker, R. 2005. *Quantile Regression*. Cambridge University Press. New York, NY.
- Koenker, R. and G. Bassett, Jr. 1978. Regression quantiles. *Econometrica*. 46(1):33-50.
- Koenker, R. and K.F. Hallock. 2001. Quantile regression. *Journal of Economic Perspective*. 15(4):143-156.

- Koenker, R. and J.A.F. Machado. 1999. Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*. 94(448):1296-1310.
- Kutner, M.H., C. J. Nachtsheim and J. Neter. 2004. *Applied Linear Regression Models*. 4<sup>th</sup> Ed. McGraw-Hill Irwin, Inc. Boston, MA.
- Mallow, C.L. 1973. Some comments on Cp. *Technometrics* 15:661-675
- Meeker, W. Q. and Escobar, L. A. 1998. *Statistical Methods for Reliability Data*. John Wiley and Sons. New York, NY.
- Mosteller, F. and J. Tukey. 1977. *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley. Reading, MA.
- Myers, R.H. 1990. *Classical and Modern Regression with Applications*. PWS-Kent Publishing Company. Boston, MA.
- Neter, J., M.H. Kutner, C.J. Nachtsheim and W. Wasserman. 1996. *Applied Linear Regression Models*. 3<sup>rd</sup> Ed. Irwin, Inc. Chicago, IL.
- Shewhart, W.A. 1931. *Economic Control of Quality of Manufactured Product*. D. Van Nostrand Company. New York, NY.
- Shupe, T.F., C.Y. Price and E.W. Price. 2001. Flake orientation effects on physical and mechanical properties of sweetgum flakeboard. *Forest Products Journal*. 51(9):38-43.
- Steele, J.C. 2006. "Function domain sets" confidence intervals for the mean residual life Functions with applications in production of medium density fiberboard. M.S. Thesis. University of Tennessee. Knoxville, TN.
- Stigler, S.M. 1986. *The History of Statistics: The Measurement of Uncertainty before 1900*. The Belknap Press of Harvard University Press: Cambridge.
- Taguchi, G. 1993. *Taguchi on Robust Technology Development*. The American Society of Mechanical Engineers. American Society of Mechanical Engineers (ASME) Press. New York, NY.
- U.S. Census Bureau. 2004. 2002 Economic census: Table 1. Advance summary statistics for the United States 2002 NAICS basis. Washington, D.C.  
<http://www.census.gov/econ/census02/advance/TABLE1.HTM>
- Venables, W.N. and Ripley B.D. 2002. *Modern Applied Statistics with S*. Springer. United States.

- Wang, Y., T.M. Young, F.M. Guess, and R.V. León. 2006. Exploring reliability of oriented strand board's tensile and stiffness strengths. *International Journal of Reliability and Application*. *In Print*.
- Weibull, W. 1939. A statistical theory of the strength of materials. *Ing. Vetenskaps Akad. Handl.* 151(1):1-45.
- Weibull, W. 1951. A statistical distribution function of wide applicability. *Journal of Applied Mechanics*. 18(1):293-297.
- Weibull, W. 1961. *Fatigue Testing and Analysis of Results*. Pergamon Press. London, UK.
- Wessa, P. 2006. Maximum-Likelihood Weibull Distribution Fitting (v1.0.0) in Free Statistics Software (v1.1.21). Office for Research Development and Education, URL: [http://www.wessa.net/rwasp\\_fitdistrweibull.wasp/](http://www.wessa.net/rwasp_fitdistrweibull.wasp/).
- Wu, Q. and C. Piao. 1999. Thickness swelling and its relationship to internal bond strength loss of commercial oriented strandboard. *Forest Products Journal*. 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 Products Journal*. 50(7/8):88-98.
- Young, T.M. 1997. Process improvement through “real-time” statistical process control in MDF manufacture. *Proc. Process and Business Technologies for the Forest Products Industry*. Forest Products Society Proceedings No. 7281. pp. 50-51.
- Young, L.J. and R.G. Easterling. 1994. Estimation of extreme quantiles based on sensitivity tests: a comparative study. *Technometrics*. 36(1):48-60.
- Young, T.M. and C.W. Huber. 2004. Predictive modeling of the physical properties of wood composites using genetic algorithms with considerations for distributed data fusion. *Proceedings of the 38th International Particleboard/Composite Materials Symposium*. Washington State Univ., Pullman, WA. pp. 145-153.
- Young, T.M. and F.M. Guess. 1994. Reliability processes and structures. *Microelectronics and Reliability*. 34:1107-1119.
- Young, T.M. and F.M. Guess. 2002. Developing and mining higher quality information in automated relational databases for forest product manufacture. *International Journal 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 Science*. 33(4):564-579.

# APPENDIX A

## SAS Code for Mixed Stepwise Regression for All Possible Subsets

\*First, we import our data into SAS using the “import” option under “file”. We name the data file “base”, and use the following command to load it into SAS memory;  
sasfile base load;

\*Here, we are extracting our data from the new SAS file “base”.

```
proc sql noprint;  
  select nobs into :nobs from sashelp.vtable  
  where libname eq 'WORK' and memname eq 'BASE';
```

\*The following code breaks the data “base” out into subsets, starting at 50 and adding one record at a time;

```
data subsets / view=subsets;  
  do samplesize = 50+subset;  
    do subset = 1 to 100;  
      start = 1;  
      end = 50+subset;  
      if end le &nobs then do obs = start to end;  
        set base point=obs;  
        output;  
      end;  
    end;  
  end;  
stop;
```

\*Below, stepwise regression is performed for each subset for the target variable, ib;

```
proc reg noprint outest=estimates rsquare adjrsq aic data=subsets;  
  by samplesize subset;  
  model ib=independent variables  
/selection=stepwise slentry=0.05 slstay=0.05;  
run;
```

\*These lines of code plot the corresponding adjusted r-sq by sample size plot;

```
proc gplot data=estimates;  
  symbol1 value=dot i=join color=black;  
  symbol2 value=dot i=join color=blue;  
  symbol3 value=dot i=join color=green;  
  legend1 label=(j=1 "Sample Size:");  
  plot _ADJRSQ_ *subset=samplesize / frame legend=legend1;  
run; quit; run;
```

\*Lastly, we filter the subsets that are listed in the output table by adding some qualifiers;  
**data** results; **set** estimates; **if** \_edf\_ ge **35**; **if** \_adjrsq\_ ge **0.50**; **if** \_p\_ le **20**; **if** \_aic\_ le **300**;  
**run**;

# APPENDIX B

## R Code for Multiple Quantile Regression

\*First, we create a function in R, then open it up to edit;

```
function() {
```

\*Here, we declare variable names (i.e., x is the first column of the data table “test”);

```
x=test[,1]
```

```
y=test[,2]
```

```
z=test[,3]
```

\*When plotting more than one variable, use the “+” sign;

```
plot(x+z,y)
```

```
points(x+z,y,cex=.5,col="blue")
```

\*We declare the percentiles of interest using the “taus” command.

```
taus <- c(.05,.1,.25,.75,.9,.95)
```

```
f <- coef(rq((y)~(x+z),tau=taus))
```

```
yy <- cbind(1,x+z)%*%f
```

```
for(i in 1:length(taus)) {
```

```
  lines(x+z,yy[,i],col = "gray")
```

```
}
```

\*Below, we add the multiple linear regression line and the quantile regression lines to the plot;

```
abline(lm(y~x+z),col="red",lty = 2)
```

```
abline(rq(y~x+z),col="blue")
```

\*Lastly, we ask for a summary of the quantile output;

```
summary(rq(y~x+z, ci=FALSE, tau=taus))
```

```
}
```

# APPENDIX C

## R Code for Weibull Distribution MLE Estimation

```
*First, we create a function in R, then open it up to edit;
function()
{

*Here, we declare variable name (i.e., x is the first column of the data table “test”);
data=read.table("ib.txt")
x=data[,1]

*We declare the Weibull parameters;
par1=1
par2=8

*The Weibull function is calculated and output is sorted;
PPCCWeibull <- function(shape, scale, x)
{
x <- sort(x)
pp <- ppoints(x)
cor(qweibull(pp, shape=shape, scale=scale), x)
}
par1 <- as.numeric(par1)
par2 <- as.numeric(par2)
if (par1 < 0.1) par1 <- 0.1
if (par1 > 50) par1 <- 50
if (par2 < 0.1) par2 <- 0.1
if (par2 > 50) par2 <- 50
par1h <- par1*10
par2h <- par2*10
sortx <- sort(x)
c <- array(NA,dim=c(par2h))
for (i in par1h:par2h)
{
c[i] <- cor(qweibull(ppoints(x), shape=i/10,scale=2),sortx)
}

*Plots the Q-Q plot;
plot((par1h:par2h)/10,c[par1h:par2h],xlab='shape',ylab='correlation',main='PPCC Plot -
Weibull')
dev.off()
f<-fitdistr(x, 'weibull')
f$estimate
f$sd
```

```
*Lastly, the following code labels the Q-Q plot;  
xlab <- paste('Weibull(shape=',round(f$estimate[[1]],2))  
xlab <- paste(xlab,', scale=')  
xlab <- paste(xlab,round(f$estimate[[2]],2))  
xlab <- paste(xlab,')')  
qqplot(qweibull(ppoints(x), shape=f$estimate[[1]], scale=f$estimate[[2]]), x, main='QQ plot  
(Weibull)', xlab=xlab )  
}
```