International Journal of Reliability and Applications
Vol. 9, No. 1, pp. 53-70, 2008

Using R Software for Reliability Data Analysis

Leslie B. Shaffer
Forest Products Center
2506 Jacob Drive
University of Tennessee, Knoxville, TN 37996-4570 USA

Timothy M. Young*
Forest Products Center
2506 Jacob Drive
University of Tennessee, Knoxville, TN 37996-4570 USA

Frank M. Guess
Department of Statistics, Operations, and Management Science
335 Stokely Management Center
University of Tennessee, Knoxville, TN 37996-0532 USA

Halima Bensmail
Department of Statistics, Operations, and Management Science
334 Stokely Management Center
University of Tennessee, Knoxville, TN 37996-0532 USA

Ramón V. León
Department of Statistics, Operations, and Management Science
337 Stokely Management Center
University of Tennessee, Knoxville, TN 37996-0532 USA

Abstract. In this paper, we discuss the plethora of uses for the software package R, and focus specifically on its helpful applications in reliability data analyses. Examples are presented; including the R coding protocol, R code, and plots for various statistical as well as reliability analyses. We explore Kaplan-Meier estimates and maximum likelihood estimation for distributions including the Weibull. Finally, we discuss future applications of R, and usages of quantile regression in reliability.

Key Words : *R, R code, reliability, Kaplan-Meier, maximum likelihood estimation, Weibull, quantile regression.*

---

* Corresponding author.
E-mail address : tmyoung1@utk.edu

1. INTRODUCTION

The software package R provides an "Open Source" option for those interested in reliability and statistical analysis, plus R is freeware. 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 researchers and workers to create user-generated software content through either incremental individual effort, or team collaboration (http://en.wikipedia.org/wiki/Open_source). The software S, the precursor to R, was originally developed by John Chambers and others at Bell Laboratories (formerly AT&T, now Lucent Technology). R can be viewed as an alternative implementation of 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. Note R is capable of performing standard exploratory data analyses such as histograms, box plots and probability plots plus much 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, for example, 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 which is continually updated with the latest research and applied 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 reliability and statistical programming packages. The many advantages of R include its wide functionality. Also, R provides a tremendous value to the user when compared to the typically higher cost of software packages such as SAS, MATLAB, Statistica, S-PLUS, etc. On the other hand these other packages have strong advantages for those able to afford those costs for larger groups of people. Many industries and groups pressured with tight margins and cost issues, however, will find R useful with state of the art techniques. Some of these highly advanced techniques appear in R before other packages due to the research community that supports R.

There are many third-party, or independent, books written on R plus numerous useful websites such as one hosted by ORNL 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 other books on the market, plus 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 such courses offered through ASA visit http://www.amstat.org/education/index.cfm?fuseaction=learnstat.

The coding protocol for R is comparable to S-PLUS; however, R protocol may not be as 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, plus specific commands must be used for data retrieval.

Given that R is "Open Source" not all of its statistical analysis packages are automatically loaded with the original software download. When performing a particular statistical analysis using R, it is imperative that the user locates on-line the extra 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 1.

Table 1.1. General tutorial of installing R with code examples.

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

	A window will pop up and you write your function:
	
	function(){
	x=data[,1]
	y=data[,2]
	function (e.g., plot(x,y))
	}
	
	Note: always be sure at this step to use the original data name. In the example below the name of the data file in the R subfolder is 'data', not test.
	
	Example: data=read.table("test.txt")
	
	Then type:
	Function name ()
	in the R console to use the function you just created.
	
	Note: 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. Compare the Appendix.

In section 2, R output examples are presented and include Kaplain-Meier plots, probability plots of normal and Weibull data, histograms, descriptive statistics, and box plots. These tools can help in a variety of ways, such as exploratory data analyses or model validations. Next in section 3 (and the Appendix) we show how R can do maximum likelihood estimates for the Weibull and other distributions. Also, we expound more on Kaplan-Meier estimates in section 3. The helpful techniques in R for doing quantile regressions for reliability data are illustrated in section 4. Finally, in section 5 we provide concluding remarks on R and its many uses and future developments in reliability.

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 1).

```
>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 2.1. Example of summary output from R of descriptive statistics.

R has excellent plotting functionality for exploratory reliability and statistical analyses such as: normal probability plots, also, called normal Quantile-Quantile (Q-Q) plots, Weibull Q-Q plots, histograms, box plots and Kaplan-Meier estimators (Figures 2, 3, 4, 5 and 6). The R code and commands for exploratory statistical analysis are quite intuitive (Tables 2 and 3).

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).

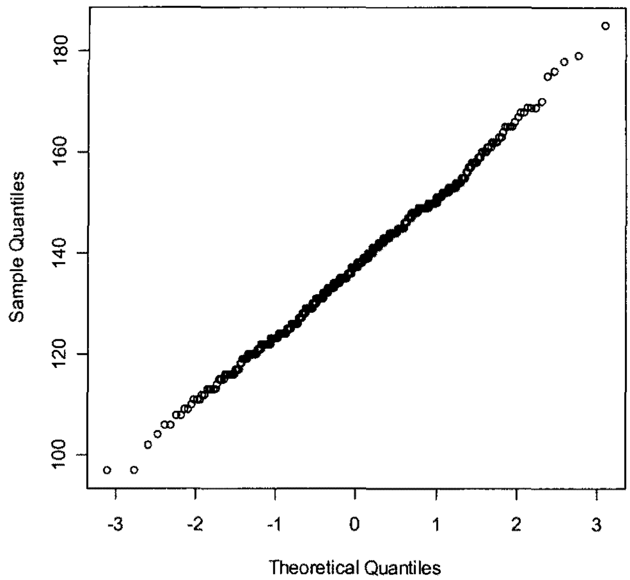

Figure 2.2. Example of normal Quantile-Quantile (Q-Q) plot of internal bond of MDF using R code.

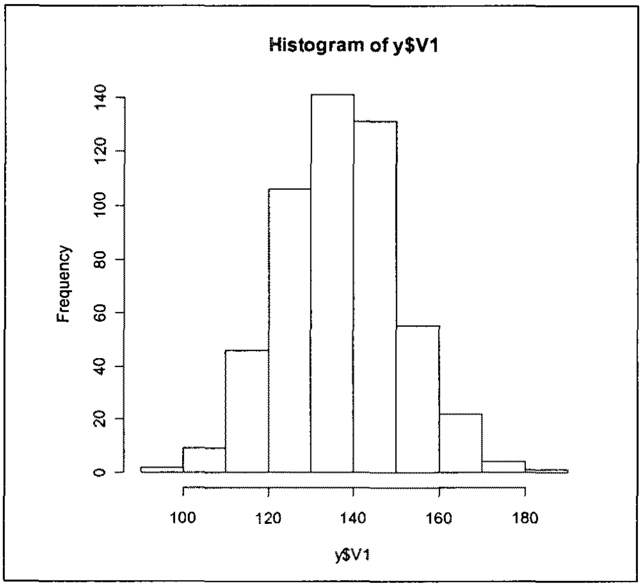

Figure 2.3. Example of histogram of internal bond of MDF using R code.

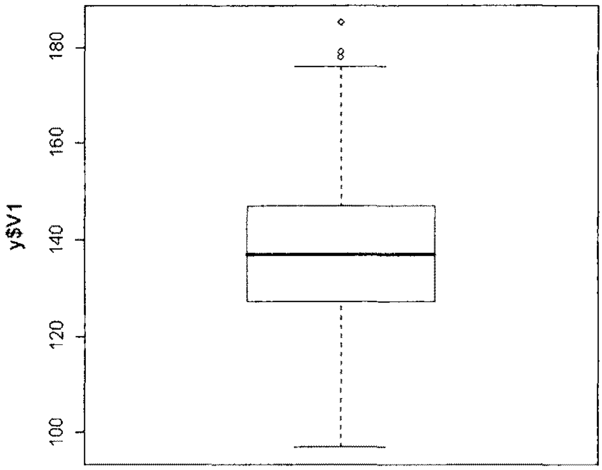

Figure 2. 4. Example of box plot of internal bond of MDF using R code.

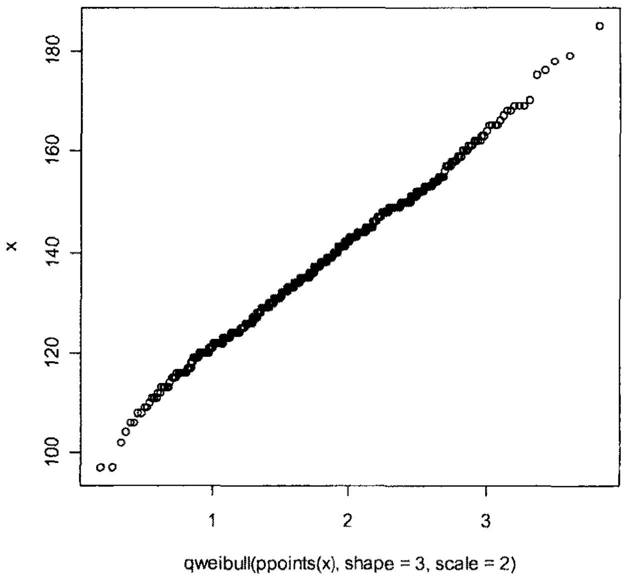

Figure 2.5. Example of Weibull Q-Q plot of internal bond of MDF using R code.

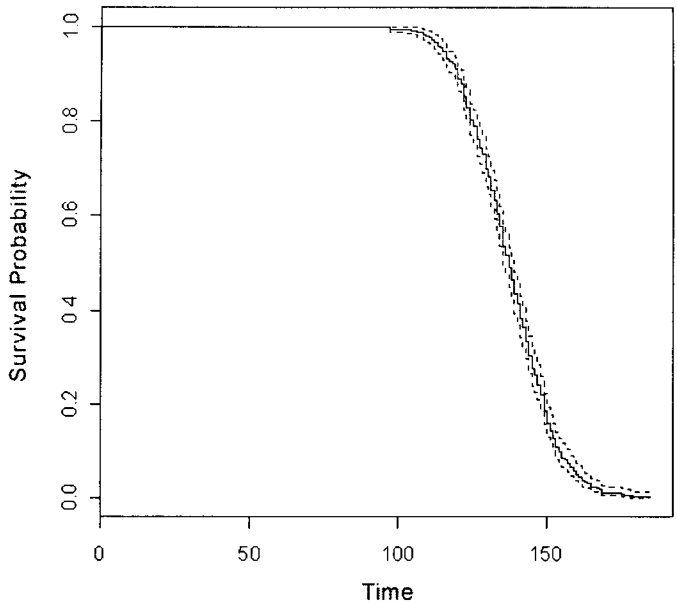

Figure 2.6. Example of Kaplan-Meier plot of internal bond of MDF using R code.

Table 2.1. Exploratory data analysis- basic statistics.

Statistic	R Command
Mean & Median	summary(DataName)
Quantile	quantile(DataName$ColumnName)
Variance	var(DataName)

Table 2.2. Exploratory data analysis- plots.

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

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 does not 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).

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 (kg/m^3). Testing of the MDF product in industry does not typically have any censoring, since testing is till failure in p.s.i. For more on MDF issues and setting up an automated data base see Young and Guess (2002).

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 model a wide range of naturally occurring reliability data from lifetime to strength data (Weibull 1939, 1951, 1961). The Weibull distribution is the most frequently used model for time (or strength via pressure) to failure,. The Weibull cumulative distribution function (cdf) giving the probability that a unit will fail by time t (or at pressure p) is:

$$F(t) = 1 - \exp[-(t/\lambda)]^{\kappa}. \tag{3.1}$$

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

$$f(t) = (\kappa/\lambda)(t/\lambda)^{\kappa-1} e^{-(t/\lambda)^{\kappa}}. \tag{3.2}$$

The parameter λ is the scale parameter and approximately equals the sixty-third lower percentile of the distribution. The parameter κ is the shape parameter (Figure 7).

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

$$h(t) = f(t)/[1 - F(t)] = (\kappa/\lambda)(t/\lambda)^{\kappa-1}. \tag{3.3}$$

The instantaneous failure rate is a measure of proneness to failure in a short interval of time (or pressure) near the current age (or pressure). The website, www.weibull.com, is a very helpful resource for learning more about this distribution and its application to various reliability problems.

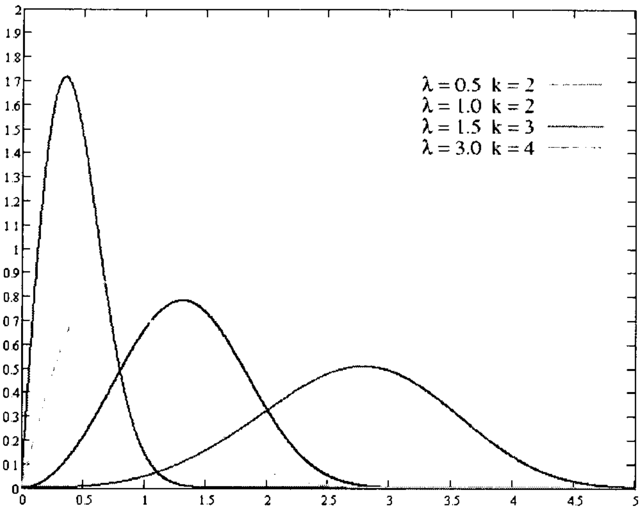

Figure 3.1. Illustration of Weibull PDF with altering values of λ and κ.

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 (originally called the 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 sample percent survival life (or pressure) of the product at failure time t over all t's. The function is a declining function, i.e., as the products age or as pressure increases, the chance of survival declines.

For large enough samples it will approach 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, (Kaplan and Meier 1958).

Guess et al. (2003) published work on 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, which proved quite helpful. Later Guess et al. (2004) used forced censoring reliability methods to better estimate bootstrap confidence intervals for the IB for MDF under different probability model assumptions.

Bootstrap confidence intervals can sometimes vary 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 than the probability models fit for the 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 (MRL) function that identified the inertia strength of MDF. 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).These empirical MRL's can be done in R, Matlab, etc.

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 and other lower tail 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.

Wang et al. (2006) applied the Kaplan-Meier estimator to oriented strand board (OSB) destructive test data and 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. Compare, also, the recent insightful work of Perhac et al. (2007) on wood-plastic composites.

Maximum likelihood estimation (MLE)

Maximum likelihood estimation (MLE) is highly important in reliability analysis because it allows practitioners 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 allows users to easily calculate these estimates for both complete and censored data. For more information on certain censored data analysis, visit: http://www.csm.ornl.gov/esh/statoed/. The forest products industry uses destructive testing to complete failure; therefore, we concentrate in this paper on complete data. However, with some manipulation of the R code, R can also readily calculate MLE's for censored data.

We use the Weibull distribution as an example to model IB pressure to failure for MDF given how often it is used in practice, plus the Weibull plot seems to support it as a potential model. The MLE output for the IB data set was $\kappa = 10.13$ (shape) and $\lambda = 143.69$ (scale) and is included in Figure 8. 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 (Wessa 2006). Also, see our Appendix for the R code to do Q-Q Weibull plots and MLE's.

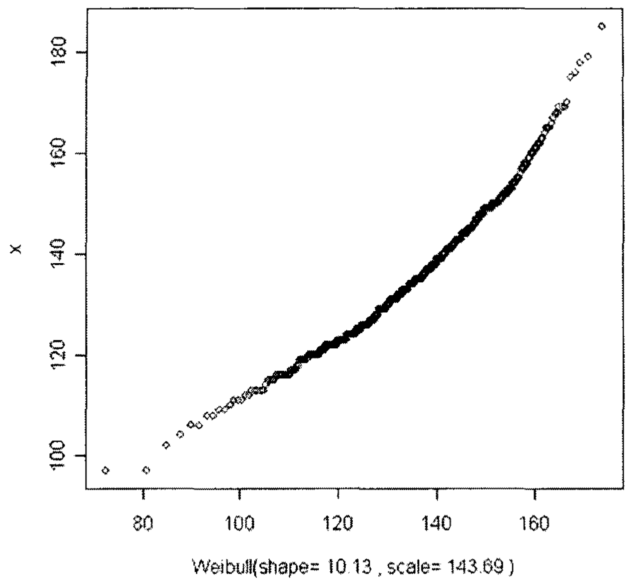

Figure 3.2. Example of Weibull MLE of internal bond of MDF with Q-Q plot using R code.

4. QUANTILE REGRESSION USING R

Koenker and Bassett (1978) introduced quantile regression (QR), and it has recently gained popularity through the need for more specialized regression techniques and extensive publications. QR is intended to offer a comprehensive strategy for completing the regression picture (Koenker 2005). It is different from the multiple linear regression (MLR) approach in that it takes into account the differences in behavior a characteristic may have at different levels of the target variable. Also, this method uses the median as the more robust measure of central tendency rather than the mean. The nonparametric median statistic may offer additional insights in the analysis of a data set, especially when compared to the mean.

The QR model does not require the product characteristics or the target variable (IB in this study) to be normally distributed, plus it does not have the other rigid assumptions associated with MLR. The "rq" function, for performing QR, can easily be downloaded and utilized in the R software package. Compare the helpful paper by Young et al. (2007).For more information, visit: http://www.econ.uiuc.edu/~roger/research/rq/vig.pdf. After downloading this package, one may perform the quantile regression analysis (Table 4, Figure 9).

Table 4.1. Quantile Regression Commands.

Function	*R command*
Quantreg	*Rq*
Plot	R command
QR and MLR fit	plot(x,y) points(x,y,cex=.5,col="blue") taus <- c(.05,.1,.25,.75,.9,.95) f <- coef(rq((y)~(x),tau=taus)) yy <- cbind(1,x)%*%f for(i in 1:length(taus)) { lines(x,yy[,i],col = "gray")} abline(lm(y~x),col="red",lty = 2) abline(rq(y~x),col="blue")
Statistic	R command
QR Summary	Summary(rq(y~x, ci=FALSE, tau=taus))

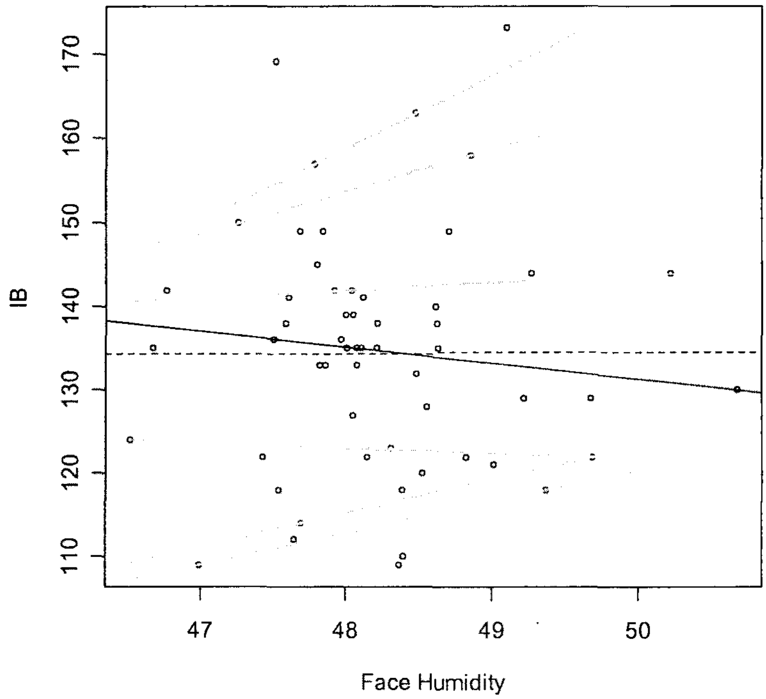

Figure 4.1. Example of a comparison of MLR fit (dotted line) with median (solid black line) and other percentile fits (gray lines) for MDF.

The aforementioned quantile regression methods used in conjunction with classical multiple linear regression analysis can improve a 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 any industry. For more information on QR, and on using the "rq" package in R, refer to *Quantile Regression* (Koenker 2005) or any of the R publications previously mentioned. Compare, also, Young et al (2007) for connections of QR with the forest products industries.

5. CONCLUSIONS

The R software package is a very powerful analytical tool and can be used for many 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 by a strong community of workers and researchers around the world.

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 freeware, R is capable of meeting the needs and filling the niches of several separate software packages while remaining highly cost effective. We, also, view other packages, such as Matlab, SAS, SPSS, Minitab, and S-PLUS as still having helpful roles.

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.

REFERENCES

Chen, W., León, R.V., Young, T.M. and 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.

Deming. W.E. (1986). *Out of the Crisis.* Massachusetts Institute of Technology's Center for Advanced Engineering Study, Cambridge, MA.

Deming, W.E. (1993). *The New Economics*. Massachusetts Institute of Technology's Center for Advanced Engineering Study, Cambridge, MA.

Everitt, B.S. and Hothorn, T. (2006). *A Handbook of Statistical Analysis Using R*. Chapman and Hall, Boca Raton, FL.

Good, P.I. (2005). *Introduction to Statistics through Resampling Methods and R/S- PLUS*. John Wiley and Sons, Inc., Hoboken, NJ.

Guess, F.M., Edwards, D.J., Pickrell, T.M. and Young, T.M. (2003). Exploring graphically and statistically the reliability of medium density fiberboard. *International Journal of Reliability and Application*, **4**(4), 97-109.

Guess, F.M., León, R.V., Chen, W. and 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., Zhang, X., Young, T.M. and León, RV. (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., Steele, J. C., Young, T. M. and León, R. V. (2006). Applying novel mean residual life confidence intervals. *International Journal of Reliability and Application*, **7**(2), 177-186.

Kaplan, E.L. and Meier, P. (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 Jr., Bassett, G. (1978). Regression Quantiles. *Econometrica*, **46**(1), 33-50.

Meeker, W. Q. and Escobar, L. A. (1998). *Statistical Methods for Reliability Data*. John Wiley and Sons, New York, NY.

Perhac, D.G., Young, T.M., Guess, F.M., and León, R.L. (2007). Exploring reliability of wood-plastic composites: stiffness and flexural strengths. Submitted manuscript.

Venables, W.N. and Ripley, B.D. (2002). *Modern Applied Statistics with S*. Springer, New York, NY.

Wang, Y., Young, T.M., Guess, F.M. and León, R.V. (2006). Exploring reliability of oriented strand board's tensile and stiffness strengths. Submitted. *International Journal of Reliability and Application*.

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/.

Young, T.M. and Guess, F.M. (1994). Reliability processes and corporate structures. *Microelectronics and Reliability*, **34**(6), 1107-1119.

Young, T.M. and Guess, F.M. (2002). Mining information in automated relational databases for improving reliability in forest products manufacturing. *International Journal of Reliability and Application,* **3**(4), 155-164.

Young, T.M., Shaffer, L.B., Guess, F.M., Bensmail, H., and León, R.V. (2007). A comparison of multiple linear regression and quantile regression for modeling the internal bond of medium density fiberboard. Submitted manuscript.

APPENDIX

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='(
plot (Weibull)', xlab=xlab )
}
```