Improved Tolerance Limits by Combining Analytical and Experimental Data: An Information Integration Methodology

A. Alexandre Trindade* Stan Uryasev†

February 16, 2005

Abstract

We propose a coherent methodology for integrating different sources of information on a response variable of interest, in order to accurately predict percentiles of its distribution. Under the assumption that one of the sources is more reliable than the other(s), the approach combines factors formed from the data into an additive linear regression model. Quantile regression, designed for quantifying the goodness of fit precisely at a desired quantile, is used as the optimality criterion in model-fitting. Asymptotic confidence interval construction methods for the percentiles are adopted to compute statistical tolerance limits for the response. The approach is demonstrated on a materials science case study that pools together information on failure load from physical tests and computer model predictions. A small simulation study assesses the precision of the inferences. The methodology gives plausible percentile estimates. Resulting tolerance limits are close to nominal coverage probability levels.

Keywords: Reliability, Quantile Regression, B basis, A basis

1 Introduction

The aim of this paper is to propose a method for integrating different sources of information on a response variable of interest, in order to make inferences on the percentiles of its distribution. The approach is general enough to be applicable in a variety of fields. To set the stage, consider the basic problem in engineering design of how to manage variability and risk in the properties of materials. Albeit costly, physical testing of the materials has traditionally been the primary method of quantifying uncertainty in vital characteristics like failure load. Recent advances in the sophistication of analytical physics-based engineering models and associated hardware/software technology, are steadily becoming increasingly

*To whom correspondence should be addressed at: Department of Statistics, University of Florida, P.O. Box 118545, Gainesville, FL 32611-8545, U.S.A. Phone: 352-392-1941 ext 216, Fax: 352-392-5175, E-mail: trindade@stat.ufl.edu

†Department of Industrial and Systems Engineering, University of Florida, P.O. Box 116595, Gainesville, FL 32611-6595, U.S.A., E-mail:uryasev@ufl.edu

important contributors in this endeavor. Although potentially less precise, the lower cost of the latter (the *indirect method*) makes it an attractive contender to the former (the *direct method*), since it does not involve physical manipulation of the material. A sound uncertainty management strategy in this context might arguably be the effective integration of the information obtained from each method. The aim of this paper is to propose a coherent methodology for such a purpose.

Fields as diverse as archaeology, astronomy, engineering, geophysics, medicine, military operations, and psychology, have developed and applied statistical methods to integrate heterogeneous sources of data (Draper et al., 1992). The concept is related to meta analysis, with an early effort by DuMouchel and Harris (1983) employing a Bayesian approach to combine the results from different studies. (See Hedges and Olkien, 1987, for a comprehensive account of meta analysis.) Craig et al. (2001), present a Bayesian forecasting approach that integrates data from computer models and expert opinions. Recently, a comprehensive methodology has been advanced by Reese et al. (2004), who show how information from computer models, expert opinions, and physical experiments, can be integrated via a multi-stage hierarchical Bayesian regression framework.

By their very nature, Bayesian approaches tend to suffer from prior sensitivity problems. Opting for a nonparametric frequentist stance, we propose a novel approach for the integration of heterogeneous sources of information, by combining data via a regression factor model. Our method is motivated by the need to analyze a specific setup where only test data and computer prediction data in the form of a range of values are available. In this context we seek to make inferences on specific extreme percentiles of the response. Selecting quantile regression (Koenker and Bassett, 1978) as our model-fitting criterion, provides us with an inferential mechanism that targets precisely the desired quantile. Associated confidence interval construction methods can then be used to calculate tolerance limits such as A basis and B basis values (see Definition 1) for the quantile(s) in question.

Although conceived in the context of a materials science problem, our approach has obvious potential applications in any field where percentile inference (extreme or otherwise) is the main goal. Nuclear Engineering safety assessment places a premium on quantifying the risk of extreme events such as reactor core meltdown. Environmental Engineering seeks to measure probabilities of high toxicity levels by combining data from various geobiotic sites. Civil Engineering, Coastal Engineering, the insurance industry, and several others, are all in the business of establishing safety margins to avert catastrophic events. It is often the case that such assessments involve the integration of information from more than one source.

This article is a refinement and condensation of the main findings in Trindade and Uryasev (2005), where specific connections with financial measures of risk are alluded to. The remainder of our paper is organized as follows. Section 2 introduces the regression modeling framework under which the data are to be combined. The model-fitting criterion, quantile regression, and associated inference techniques are covered in Section 3. The methodology is illustrated in Section 4 by applying it to a case study from the aerospace industry involving material failure loads. We conclude in Section 5 with a Monte Carlo

study on the precision of the inferences derived from the methodology.

2 The Modeling Framework

The setup for the problem we wish to address is as follows. Data are collected on a response variable of interest obtained through two mechanisms: *direct* observation and *indirect* observation. Assume that the **direct data** is more precise than the **indirect data**, that is to say, the direct measurements are believed to be closer to the truth. In the case study we will shortly consider, the response is material strength and the direct data is obtained by applying increasing loads to samples of such material until failure is observed. In addition, failure load predictions obtained via methods that do not involve physical testing of the material are also available, and comprise the indirect data. The latter can encompass a broad spectrum of methods, from analytical physics-based computer models and simulations, to subjective expert opinions. Primary questions of interest are: how to effectively integrate direct and indirect data in a coherent manner so as to use all available information, and how to quantify the contribution of the indirect data in the presence of the direct data. In the case study, the hope is that indirect model data which is cheaper to generate, can contribute to accurate prediction of failure loads, reducing the need for more expensive direct test data. The material may be available in different *formulations*, changes in its composition and/or structure deliberately introduced in the manufacturing process so as to affect the response.

We consider only the simplest case when direct data and only one type of indirect data are available. The basic idea underlying the approach is to use a regression framework, with each of the direct data values as the response in turn; the remainder and all the indirect data serving as explanatory variables (covariates). Let Y_{ij} denote the response obtained from the jth direct data value corresponding to the ith formulation, $i = 1, \ldots, I$, and $j = 1, \ldots, N_i$. We condense the data from each source, direct and indirect, into a pair of summary statistics such as the mean and standard deviation. Other data reduction measures of location (e.g. median) and dispersion (e.g. lower semi-deviation) could also be used for more skewed data. Let $(m_{i(j)}, s_{i(j)})$ denote respectively the sample mean and standard deviation for the direct data in formulation i, obtained from $Y_{i,1}, \ldots, Y_{i,j-1}, Y_{i,j+1}, \ldots, Y_{i,N_i}$. Letting (μ_i, σ_i) denote the indirect data mean and standard deviation in formulation i, we fit the regression model:

$$Y_{ij} = c_0 + c_1\mu_i + c_2\sigma_i + c_3 m_{i(j)} + c_4 s_{i(j)} + \varepsilon_{ij}, \tag{1}$$

where $c_0, \ldots, c_4$ are unknown regression coefficients to be estimated from the data, and ε_{ij} is the residual error corresponding to the jth direct data value in the ith formulation. Evidently, if only the indirect data mean, μ_i is available, then the $c_2\sigma_i$ term will be absent. This basic setup can easily be extended to accommodate data from several (K say) indirect data sources $(\mu_i^{(k)}, \sigma_i^{(k)})$, $k = 1, \ldots, K$, which can occur if competing approaches are being considered, e.g. two numerical models plus expert opinions ($K = 3$). Extensions to additional sources of variability are likewise immediate. Of course, the basic linear model

can also be made more flexible by the inclusion of nonlinear terms and interactions. Specific details and an example on how to construct the resulting regression design matrix and response vector are given in the Appendix.

The motivation behind model (1) is two-fold and can be heuristically argued as follows. Firstly, it provides a way to combine the direct and indirect information; leaving Y_{ij} out of the $(m_{i(j)}, s_{i(j)})$ calculation in a sense justifies its use as the (obvious) response at that level of i and j. Secondly, the approach "pools" together information from all sources when making inferences on a particular source, in much the same way as the *Stein effect* improves estimation by using information from all coordinates when estimating each coordinate (Berger, 1982).

In reliability applications where failure load is the response, accurate inference on low quantiles of its distribution are often desired. To ensure ample safety margins, engineering specifications usually call for estimates of the 1st and 10th quantiles, along with associated 95% lower confidence bounds. The resulting *tolerance limits* are repectively known as *A basis* and *B basis*. Precise definitions of these terms are as follows.

Definition 1 (A basis and B basis) *Let $\{X_1, \ldots, X_n\}$ be a random sample of observations from random variable X with continuous distribution function $F(x) \equiv P(X \leq x)$. For probability level $0 \leq \tau \leq 1$, let $\xi_\tau \equiv \xi_\tau(X_1, \ldots, X_n)$ be a statistic (a function of the random sample) satisfying,*

$$\begin{aligned} 0.95 &= P\left[1 - F(\xi_\tau) > 1 - \tau\right] && (2) \\ &= P\left[F(\xi_\tau) < \tau\right] = P\left[\xi_\tau < F^{-1}(\tau)\right]. && (3) \end{aligned}$$

Representation (3) means that (ξ_τ, ∞) is a one-sided confidence interval for $F^{-1}(\tau)$, so that ξ_τ is a lower 95% confidence *bound for the τth quantile of X. Equivalently, representation (2) means that ξ_τ is a lower 95%* tolerance *bound for the $(1-\tau)$th quantile of X. An* A basis *value is defined to be $\xi_{0.01}$, and a* B basis *value, $\xi_{0.10}$. Thus in repeated sampling, the calculated A basis (B basis) value would fall below the true 1st (10th) percentile, 95% of the time.*

With inference on such extreme quantiles of the response in mind, minimization of residual error criteria such as squared or absolute deviations with their focus on measures of central tendency, may be inappropriate. Quantile regression, to be discussed next, provides a promising alternative.

3 Model Fitting: Quantile Regression

Consider observations

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \tag{4}$$

for $i = 1, \ldots, n$, from the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon, \tag{5}$$

where the errors $\{\varepsilon_i\}$ are independent and identically distributed (i.i.d.) with cdf F_ε and pdf f_ε. Since $P(Y \leq y|\mathbf{x}) = F_\varepsilon(y - \mathbf{x}'\boldsymbol{\beta})$, it follows immediately that the *conditional quantile function* of Y given $\mathbf{x} = [1, x_1, \ldots, x_p]'$, $Q_Y(\tau|\mathbf{x})$, and the quantile function of ε, are related according to,

$$Q_Y(\tau|\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta} + F_\varepsilon^{-1}(\tau) \equiv \mathbf{x}'\boldsymbol{\beta}(\tau), \tag{6}$$

where $\beta_0(\tau) = \beta_0 + F_\varepsilon^{-1}(\tau)$, and $\beta_j(\tau) = \beta_j$, for $j = 1, \ldots, p$. Viewed as a function of the design point $\mathbf{x}$, $Q_Y(\tau|\mathbf{x})$ describes how the τth quantile surface of Y varies as a function of the factors $X_1, \ldots, X_p$.

Let $\mathbf{y}' = [y_1, \ldots, y_n]$ denote the observed vector of responses, so that $\mathbf{y} = \mathrm{X}\boldsymbol{\beta} + \varepsilon$. For any real number z, define $(z)^+ = \max\{0, z\}$ and $(z)^- = \min\{0, z\}$. In analogy with the fact that any number $b(\tau)$ that satisfies,

$$b(\tau) = \arg\min_{b \in \mathbf{R}} \frac{1}{n} \sum_{i=1}^{n} \left[\tau(y_i - b)^+ - (1-\tau)(y_i - b)^-\right],$$

is an estimator of the τth quantile of Y based on the sample $y_1, \ldots, y_n$, Koenker and Bassett (1978) proposed that a natural estimator of $\boldsymbol{\beta}(\tau)$ can be obtained by minimizing the criterion,

$$\mathcal{V}(\mathbf{b}) = \frac{1}{n} \sum_{i=1}^{n} \left[\tau(y_i - \mathbf{x}_i'\mathbf{b})^+ - (1-\tau)(y_i - \mathbf{x}_i'\mathbf{b})^-\right], \tag{7}$$

over all $\mathbf{b} \in \mathbf{R}^{p+1}$. The resulting estimators, $\hat{\boldsymbol{\beta}}(\tau)$, called *regression quantiles* by Koenker and Bassett (1978), are therefore given by,

$$\hat{\boldsymbol{\beta}}(\tau) = \arg\min_{\mathbf{b} \in \mathbf{R}^{p+1}} \mathcal{V}(\mathbf{b}).$$

The methodology that results in the regression quantiles is called **Quantile Regression**. Using it to fit model (1), leads to the following estimated equation for the τth quantile of the failure load, as a function of the generic design vector, $\mathbf{x}' = \{1, \mu, \sigma, m, s\}$:

$$\hat{Q}_Y(\tau) = \hat{c}_0(\tau) + \hat{c}_1(\tau)\mu + \hat{c}_2(\tau)\sigma + \hat{c}_3(\tau)m + \hat{c}_4(\tau)s = \mathbf{x}'\hat{\mathbf{c}}(\tau). \tag{8}$$

A basis and B basis values are obtained as a 95% lower confidence bound for the respective quantile, $\tau = 0.01$ or $\tau = 0.10$. Several methods are available for the construction of such a confidence bound; we present three of the most promising.

(i) *Studentization.* This is due to Koenker (1994), and is based on the asymptotic distribution of the regression quantiles derived by Koenker and Bassett (1978) under mild regularity conditions. With $\Omega = \mathrm{X}'\mathrm{X}$ and $z_{.05}$ denoting the 5th percentile from a standard normal distribution, the studentization method results in the A basis ($\tau = 0.01$) or B basis ($\tau = 0.10$) value

$$\mathbf{x}'\hat{\mathbf{c}}(\tau) + z_{.05}\hat{s}_n(\tau)\sqrt{\tau(1-\tau)\mathbf{x}'\Omega^{-1}\mathbf{x}}, \tag{9}$$

where $\hat{s}_n(\tau)$ is an estimate of the sparsity function $s(\tau) = 1/f_\varepsilon(F_\varepsilon^{-1}(\tau))$, obtained by using for example the Hall-Sheather bandwidth.

(ii) *Direct.* A nonparametric approach making direct use of the empirical quantile function, proposed by Zhou and Portnoy (1996). This results in the A basis ($\tau = 0.01$) or B basis ($\tau = 0.10$) value

$$\mathbf{x}'\hat{c}\left(\tau + z_{.05}\sqrt{\tau(1-\tau)\mathbf{x}'\Omega^{-1}\mathbf{x}}\right). \tag{10}$$

(iii) *Resampling.* Various bootstrap methods have also been suggested by Koenker (1994), and Zhou and Portnoy (1996). Generally speaking, the bootstrap B basis (A basis) value would be the 5th empirical quantile of the sampling distribution of the estimator of the 10th (1st) quantile.

Approaches (i) and (ii) were originally presented in the context of homoscedastic errors, that is, all ε_i having the same variance. By introducing weights $w_1, \ldots, w_n$ in the criterion function (7), which becomes

$$\mathcal{V}(\mathbf{b}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{w_i}\left[\tau(y_i - \mathbf{x}_i'\mathbf{b})^+ - (1-\tau)(y_i - \mathbf{x}_i'\mathbf{b})^-\right], \tag{11}$$

Zhou and Portnoy (1998) generalize (i) and (ii) to accommodate heteroscedasticity (different error variances). The only adjustment needed in (9) and (10), is to redefine $\Omega = \mathrm{X}'\mathrm{Diag}(w_1, \ldots, w_n)\mathrm{X}$ with the weights vector $\mathbf{w} = [w_1, \ldots, w_n]'$ estimated either via least absolute deviations or least squares. The latter permits a closed form representation for the estimator as

$$\hat{\mathbf{w}} = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'|\hat{\boldsymbol{\xi}}|,$$

where $\hat{\boldsymbol{\xi}}$ is the vector of quantile regression residuals resulting from a median quantile regression fit ($\tau = 0.5$).

Finally, and in analogy with R^2 of ordinary regression, Koenker and Machado (1999) point out that a measure of goodness-of-fit relative to the intercept-only model can also be defined in the framework of quantile regression. That is, if $\tilde{\beta}_0(\tau)$ denotes the estimated regression quantile in the intercept-only model, and $\hat{\boldsymbol{\beta}}(\tau)$ the estimated regression quantiles in the full model (8), we define

$$R^1(\tau) \equiv 1 - \mathcal{V}\left(\hat{\boldsymbol{\beta}}(\tau)\right)/\mathcal{V}\left(\tilde{\beta}_0(\tau)\right). \tag{12}$$

The $R^1(\tau)$ criterion thus assess the appropriateness of the fitted conditional quantile function of Y at τ, on a scale of 0 to 1. Values close to 0 (1) suggest a poor (good) fit.

Remark. Trindade and Uryasev (2005) demonstrate that whereas the fitted quantile regression criterion $\mathcal{V}(\hat{\boldsymbol{\beta}}(\tau))$ is not easily interpretable, the modified criterion $\mathcal{V}^{\mathrm{C}}(\hat{\boldsymbol{\beta}}(\tau))$ defined as,

$$\mathcal{V}^{\mathrm{C}}(\hat{\boldsymbol{\beta}}(\tau)) = \frac{\max(\tau, 1-\tau)}{\tau(1-\tau)}\mathcal{V}(\hat{\boldsymbol{\beta}}(\tau)),$$

can be visualized as the distance between the mean and the tail mean beyond the smaller of the τth or $(1-\tau)$th quantile, for the fitted residual error distribution,

$$\hat{\varepsilon} = Y - \hat{\beta}_0(\tau) - \hat{\beta}_1(\tau)X_1 - \ldots - \hat{\beta}_p(\tau)X_p.$$

That is,

$$\mathcal{V}^{\mathrm{C}}(\hat{\boldsymbol{\beta}}(\tau)) = \begin{cases} \mathrm{CVaR}_\tau\left(\hat{\varepsilon} - \mathbf{E}\hat{\varepsilon}\right), & \text{if } \tau \geq 0.5, \\ \mathrm{CVaR}_{1-\tau}\left(\mathbf{E}\hat{\varepsilon} - \hat{\varepsilon}\right), & \text{if } \tau < 0.5, \end{cases}$$

where for random variable Z, $\mathrm{CVaR}_\tau(Z) = \mathbf{E}\left[Z \mid Z \geq F_Z^{-1}(\tau)\right]$.

4 Integrating Data on Composite Material Strength: A Case Study from the Aerospace Industry

In this section we apply the data integration and quantile regression methodology to a dataset supplied by The Boeing Company involving the strength of a composite material available in several different formulations. The information available within each formulation consisted of: (i) an upper and lower failure load prediction limit for each formulation stemming from an analytical model (model data); and (ii) actual observed failure loads obtained through physical testing of formulation samples (test data). The test and model data comprise the direct and indirect data respectively, and we refer to it as such in the context of this case study. The dataset considered consisted of 18 formulations with exactly 5 test points per formulation.

Figure 1 shows the case study data augmented with 10th and 90th percentile lines, estimated via weighted quantile regression (least squares weights). If U_i and L_i denote respectively the upper and lower model failure load prediction limits in the ith formulation, the model mean and standard deviation is computed as $\mu_i = (U_i + L_i)/2$ and $\sigma_i = (U_i - L_i)/2$. As is evident, the method gives plausible percentile estimates. Information is pooled from all formulations when making inferences on a particular formulation.

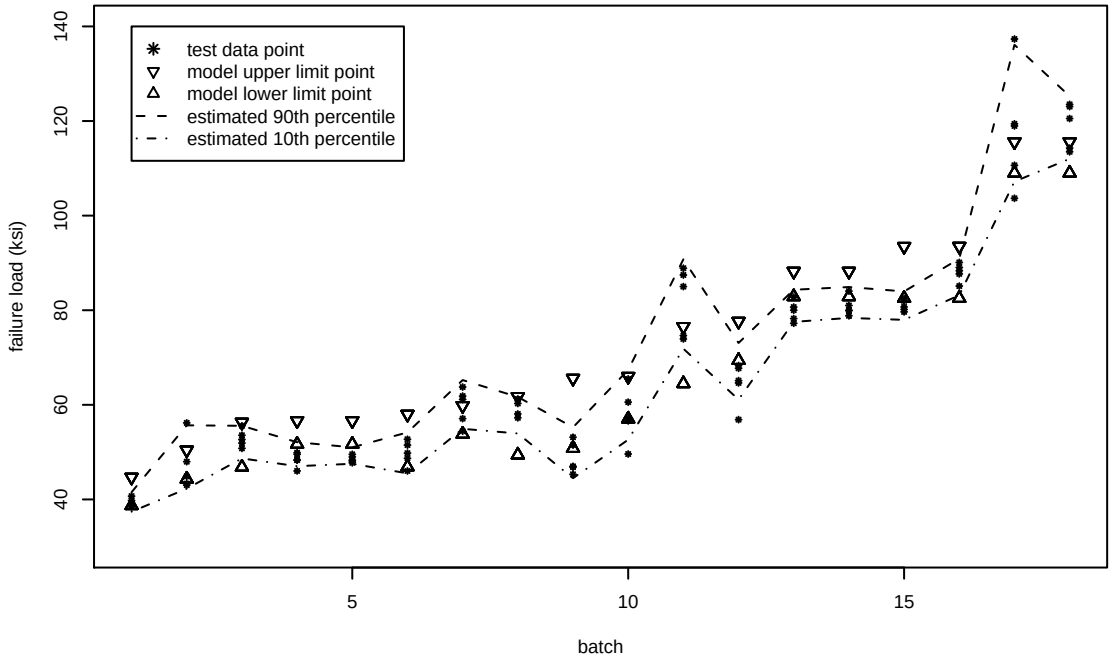
In direct analogy with ordinary regression, and to quantify the contribution from a subset of covariates given a subset already in the model, we can compute the *partial* $R^1(\tau)$. Simplify the notation by suppressing dependence on the fitted parameters $\hat{\boldsymbol{c}}(\tau)$, and let the subscripts m, t, and mt denote respectively results of weighted quantile regression runs using as covariates model only data, test only data, and both model and test data. For example, $\mathcal{V}^{\mathrm{C}}_{mt}(\tau)$ denotes the weighted quantile regression goodness-of-fit criterion attained under model (1), while $\mathcal{V}^{\mathrm{C}}_{t}(\tau)$ denotes the same attained under the subset model

$$Y_{ij} = c_0 + c_3 m_{i(j)} + c_4 s_{i(j)} + \varepsilon_{ij}.$$

The partial $R^1(\tau)$ of the model data given the test data can then be defined as

$$R^1_{m|t}(\tau) \equiv 1 - \mathcal{V}^{\mathrm{C}}_{mt}(\tau)/\mathcal{V}^{\mathrm{C}}_{t}(\tau) = \frac{R^1_{mt}(\tau) - R^1_t(\tau)}{1 - R^1_t(\tau)}.$$

Figure 1: The Case Study Data With Estimated 10th and 90th Percentiles. Each formulation has an upper and lower failure load prediction limit stemming from an analytical model (model points), and actual observed failure loads obtained through physical testing of formulation samples (test points). The percentile lines are fitted via weighted quantile regression, using both model and test data as covariates.



$\mathcal{V}_t^{\mathrm{C}}(\tau)$ and $R_t^1(\tau)$ are replaced by $\mathcal{V}_m^{\mathrm{C}}(\tau)$ and $R_m^1(\tau)$ in the definition of $R_{t|m}^1(\tau)$. Figure 2 shows a schematic of all the partial $R^1(.10)$ values, as a function of the terms present in the quantile regression model.

The diagram in Figure 2 suggests that the model data contributes little (1.8%) in the presence of the test data, whereas the test data contributes 19.1% in the presence of the model data. Table 1 shows the change in the fitted coefficients (and standard errors) as we add more covariate terms into the 10th quantile regression model. The signs of the coefficients remain consistent throughout; positive (negative) for the mean (standard deviation) terms, agreeing with the intuitively obvious fact that quantile estimates should increase with increasing mean failure loads, and decrease with increasingly variable failure loads. Only the mean terms in the subset models (model only and test only) are significant; sample sizes are generally too small to accommodate a more complex regression model.

Figure 2: Partial $R^1(0.10)$ Values for the Case Study Data. Values are computed from weighted quantile regression fits using test only (t), model only (m), and both test and model (mt) data as covariates.

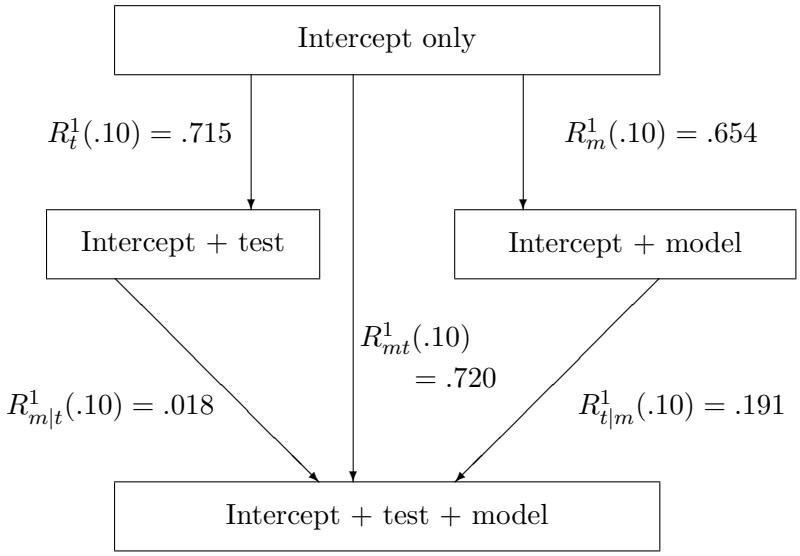


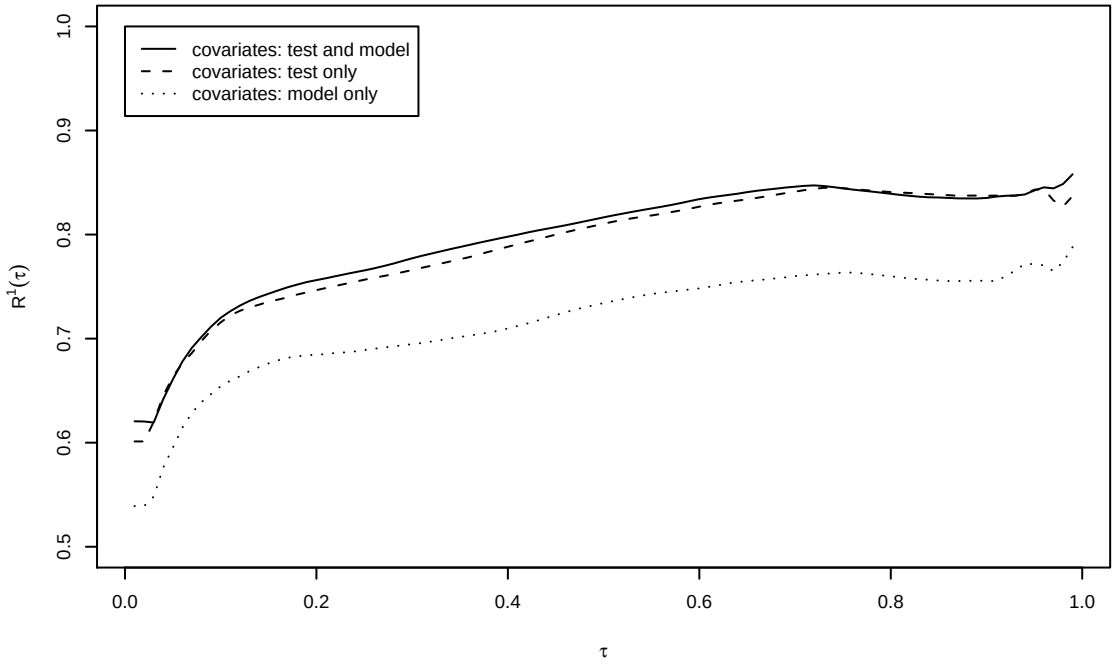
Table 1: Fitted coefficient values under 10th quantile regression modeling with different sets of covariates. The first set includes only model mean (c_1) and standard deviation (c_2) data; the second only test mean (c_3) and standard deviation (c_4) data; the third both model and test data. Standard errors appear in parentheses.

Covariates	Estimates of quantile regression coefficients				
used	$c_0(0.1)$	$c_1(0.1)$	$c_2(0.1)$	$c_3(0.1)$	$c_4(0.1)$
model only	-1.59	0.96	-1.02		
	(10.94)	(0.16)	(2.36)		
test only	1.72			0.95	-1.33
	(7.88)			(0.15)	(1.70)
model & test	0.24	0.20	-0.54	0.79	-0.58
	(14.73)	(0.92)	(2.32)	(0.93)	(2.69)

A broader perspective of the extent of the goodness-of-fit attained for these data is presented in Figure 3, which shows the value of $R^1(\tau)$ plotted as a function of τ. These values are obtained as in Figure 2, for all values of $0 < \tau < 1$. The different line types indicate whether only test, only model, or both test and model, data are used as covariates. The small difference in fits obtained by using test only vs. model and test as covariates, with model only fits lagging about 10% below these, is consistent with our earlier statements.

Slightly better fits for all are attained in the right tails ($\tau > 0.5$).

Figure 3: The entire quantile regression process for the case study data. The goodness-of-fit criterion $R^1(\tau)$ from weighted quantile regression fits, plotted as a function τ. The line types indicate whether only test, only model, or both test and model, data are used as covariates.



5 Assessing the Precision of the Inferences on Simulated Data

As a check on the precision of the results, the performance of the methodology was assessed in a controlled environment consisting of failure data simulated from two-parameter Weibull distributions. The choice of shape (α) and scale (β) parameters, constituting a formulation, is made randomly and uniformly over the rectangle, $10 < \alpha < 80$ and $40 < \beta < 120$. (This region encompasses plausible values of the Weibull parameters for the case study data, determined by a maximum likelihood fit to the test data within each formulation.) A random draw of fixed sample size $N_i \equiv n$ is then made in each of I formulations, for all combinations of $n = 5, 10, 20, 30,$ and $I = 5, 10, 20, 50$. The n points constitute the test

data used to form $(m_{i(j)}, s_{i(j)})$. The true mean and standard deviation of the formulation Weibull distribution form the model data,

$$\mu_i = \beta_i \Gamma(1 + 1/\alpha_i), \quad \text{and} \quad \sigma_i = \beta_i \sqrt{\Gamma(1 + 2/\alpha_i) - \Gamma(1 + 1/\alpha_i)^2}.$$

Regression model (1) is then fitted via weighted quantile regression, and the B basis calculated using the studentization approach. This is replicated 1,000 times for each (n, I) combination. The percentage of time that the calculated B basis fell below the true 10th percentile within each formulation is recorded. Averaging these percentages across all formulations gives the B basis percent coverage probabilities presented on Table 2.

Table 2: Percent coverages for B basis values computed via studentized weighted quantile regression. The data within a formulation are simulated from a Weibull distribution, with shape (α) and scale (β) parameters selected randomly and uniformly over the rectangle, $10 < \alpha < 80$, $40 < \beta < 120$. Coverages are based on 1,000 replications. Nominal level is 95%.

Number of	Number of test points per formulation			
formulations	5	10	20	30
5	32.6	39.4	39.9	41.5
10	72.2	78.1	82.8	85.8
20	89.9	92.0	94.6	95.4
50	95.4	96.6	97.6	98.0

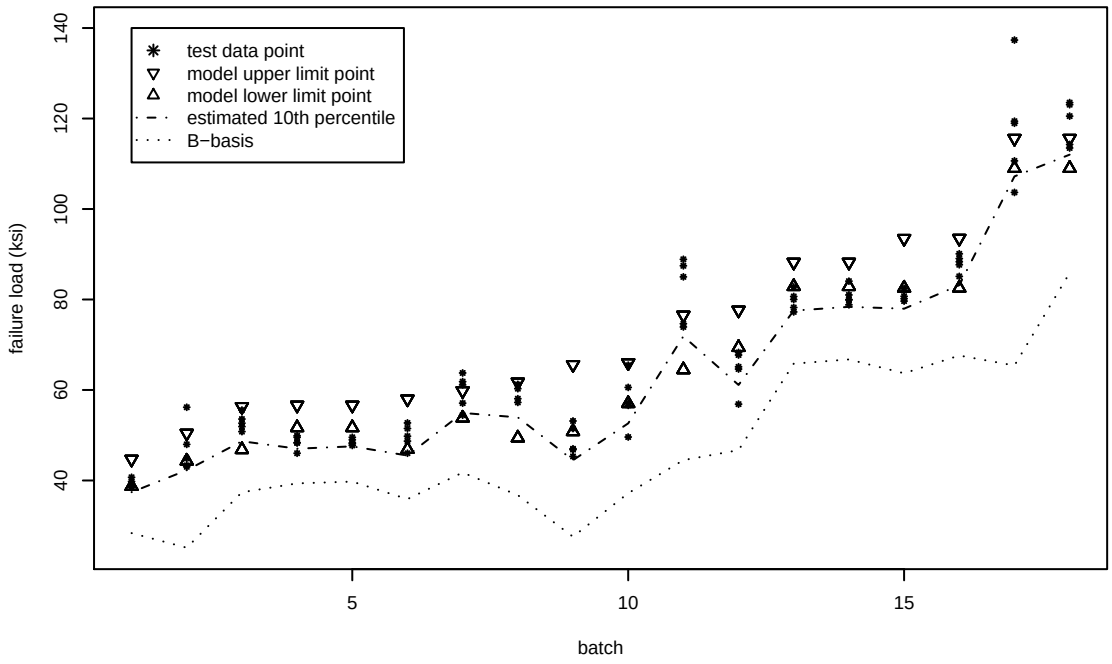
For small formulation and test point numbers, coverages are well below the nominal 95% level. The nominal level is attained at about 20 formulation, each with 20 test points. The direct method was found to be unsuitable for extreme quantiles like the 10th; the value in parenthese of equation (10) was often negative. Because of the reasonable studentization coverages obtained in the range of sample sizes present in the study, the computationally intensive resampling methods were not implemented.

Figure 4 shows the resulting B basis line for the case study data, computed via this studentized approach at each of the fitted 10th percentile lines. The large variability present in the estimates is reflected in the distance between the two lines. This is not however inconsistent with the small amount of data available, especially test points within formulations. One would naturally expect better results in larger studies.

6 Summary

We have shown how predictions on material strength from computer or analytical models can be coherently integrated with physical test data, in order to improve the accuracy of estimates for percentiles of the failure load distribution. The method is general enough to be applied to any situation where one seeks to combine quantitive data on a particular

Figure 4: B basis Values for Case Study Data. The estimated 10th percentile and B basis lines are based on quantile regression fits, with both test and model data as covariates.



response variable from a more trustworthy source, with data obtained from one or more less reliable sources. The information integration is accomplished in the farmework of a regression model, the parameters of which can be estimated via weighted quantile regression, a criterion specifically designed to achieve a good fit at a chosen percentile. In the case of Weibull-distributed failure loads, tolerance limits calculated via a one-sided asymptotic confidence bound were found to achieve good coverage probabilities at moderate sample sizes. An argument analogous to the partial R^2 of ordinary regression indicates that the model information by itself explains about 65% of the "variability" seen in percentiles of the failure load distribution. The incorporation of the test data in the form of summary statistics serving as covariates, accounts for approximately 20% of the remaining variability.

Acknowledgments

This effort was jointly accomplished by the Boeing led Accelerated Insertion of Materials - Composites (AIM-C) team and the United States Government under the guidance

of NAVAIR. This work was funded by DARPA/DSO and administered by NAVAIR under Technology Investment Agreement N00421-01-3-0098. The program acknowledges the guidance and support of Dr. Leo Christodoulou of DARPA/DSO for this effort. The technical monitor for the program is Dr. Ray Meilunas of NAVAIR. In addition, the authors acknowledge collaboration with Dr. Raj Rajagopal of The Boeing Company. We thank Saveliy Uryasev at American Optimal Decisions, Inc., for programming assistance.

A Details on Formation of the Response Vector and Design Matrix from Available Test and Model Data

We provide specifics on the formation of response vector $\mathbf{y}$ and design matrix X of Section 2, used for quantile regression fitting in Section 3. Let Y_{ij} be jth response for the ith formulation, obtained from direct data, $i = 1, \ldots, I$, $j = 1, \ldots, N_i$. The total number of direct data values is $N = \sum_{i=1}^{I} N_i$. To build the regression data, the direct data in each formulation are partitioned into two disjoint sets; values in one set are used as the response while those in the other are used to form the covariates. This is repeated for all remaining possible ways to partition the direct values so as to result in different response and covariate sets. The details are as follows:

1. Repeat the following loop for each formulation i, where $i = 1, \ldots, I$:

   (a) Form all $p_i = \binom{N_i}{r}$ possible partitions of the N_i direct values into two disjoint sets consisting of r and $n_i = N_i - r$ values, respectively, where $0 \le r \le N_i$, is constant for all i. The r values will be used to form the covariates (the *covariate set*), while the n_i values will be used as responses (the *response set*).

   (b) Let $Y^{(q)}_{i(1)}, \ldots, Y^{(q)}_{i(n_i)}$ denote the direct values in the qth response set, $q = 1, \ldots, p_i$. Form $\mathbf{y}_i$, the vector of length $n_i p_i$ consisting of the responses ordered as follows,

   $$\mathbf{y}_i = [Y^{(1)}_{i(1)}, \ldots, Y^{(1)}_{i(n_i)}, Y^{(2)}_{i(1)}, \ldots, Y^{(2)}_{i(n_i)}, \ldots, Y^{(p_i)}_{i(1)}, \ldots, Y^{(p_i)}_{i(n_i)}]'.$$

   (c) Let $Y^{(q)}_{i(n_i+1)}, \ldots, Y^{(q)}_{i(N_i)}$ denote the direct data values in the qth covariate set, $q = 1, \ldots, p_i$. Form $m_i^{(q)}$ and $s_i^{(q)}$, the sample mean and standard deviation of the direct data values in the qth covariate set:

   $$m_i^{(q)} = \frac{1}{r} \sum_{j=n_i+1}^{N_i} Y^{(q)}_{i(j)}, \qquad s_i^{(q)} = \sqrt{\frac{1}{r-1} \sum_{j=n_i+1}^{N_i} \left(Y^{(q)}_{i(j)} - m_i^{(q)} \right)^2}.$$

   If $r = 0$, the direct data is not used as covariates (only indirect data). If $r = 1$, then $s_i^{(q)} = 0$. If $r = N_i$, we use all the direct data as both response and covariates.

   (d) Form $\mu_i^{(k)}$ and $\sigma_i^{(k)}$, the sample mean and standard deviation of the indirect data from the kth source in formulation i, for all $k = 1, \ldots, K$.

(e) Form the design matrix, X_i, for formulation i. This consists of the direct and indirect data summary statistics arranged in a matrix of $n_i p_i$ rows and $2K+3$ columns (the first column consists of the number 1). The rows correspond to those in $\mathbf{y}_i$. If $\# = \lceil v/n_i \rceil$ denotes the largest integer greater than or equal to v/n_i, the vth row of X_i is just

$$\left[1, \mu_i^{(1)}, \sigma_i^{(1)}, \mu_i^{(2)}, \sigma_i^{(2)}, \ldots, \mu_i^{(K)}, \sigma_i^{(K)}, m_i^{(\#)}, s_i^{(\#)}\right].$$

2. Form $\mathbf{y}$, the regression response vector of length $n = \sum_{i=1}^{I} n_i p_i$ comprised of the vertical concatenation of the formulation response vectors, $\mathbf{y}' = [\mathbf{y}_1', \ldots, \mathbf{y}_I']$.

3. Form the regression design matrix, X, with n rows and $l+1 = 2K+3$ columns, comprised of the vertical concatenation of the formulation design matrices X_i, $\mathrm{X}' = [\mathrm{X}_1', \ldots, \mathrm{X}_I']$.

Note that throughout the paper we used $r = 4$. An example follows.

Example. Suppose formulation i has $\{10.1, 12.3, 14.5, 16.7\}$ as direct data, and $\{12.4, 16.8\}$ as indirect data. We will illustrate the calculations in 1(a). Here $K = 1$, $N_i = 4$, and suppose we choose $r = 2$. Then $p_i = \binom{4}{2} = 6$, and X_i will be of dimension (12×5). We obtain,

$$\mu_i^{(1)} = (12.4 + 16.8)/2 = 14.6, \qquad \text{and} \qquad \sigma_i^{(1)} = 16.8 - 12.4 = 4.4.$$

For the first covariate set, $\{14.5, 16.7\}$, we obtain

$$m_i^{(1)} = (14.5 + 16.7)/2 = 15.6, \qquad \text{and} \qquad s_i^{(1)} = 16.7 - 14.5 = 2.2.$$

Computing $m_i^{(2)}, \ldots, m_i^{(6)}$ and the corresponding $s_i^{(2)}, \ldots, s_i^{(6)}$ similarly, gives eventually,

$$\mathbf{y}_i' = [10.1, 12.3, 10.1, 14.5, 10.1, 16.7, 12.3, 14.5, 12.3, 16.7, 14.5, 16.7].$$

The first 3 columns of X_i consist of the numbers 1, 14.6, and 4.4, respectively, while the 4th and 5th columns are

$$[15.6, 15.6, 14.5, 14.5, 13.4, 13.4, 13.4, 13.4, 12.3, 12.3, 11.2, 11.2]$$

and

$$[2.2, 2.2, 4.4, 4.4, 2.2, 2.2, 6.5, 6.5, 4.4, 4.4, 2.2, 2.2].$$

References

[1] Berger, J. (1982). Bayesian robustness and the Stein effect. *Journal of the American Statistical Association* **77**, 358-368.

[2] Craig, P., Goldstein, M., Rougier, J. and Seheult, A. (2001). Bayesian Forecasting for Complex Systems Using Computer Simulators. *Journal of the American Statistical Association* **96**, 717-729.

[3] Draper, D., Gaver, D., Goel, P., Greenhouse, J., Hedges, L., Morris, C., Tucker, J. and Waternaux, C. (1992). Selected Statistical Methodology for Combining Information (CI). *Combining Information: Statistical Issues and Opportunities for Research*, eds. D. Cochran and J. Farrally. National Academy Press, Washington DC.

[4] DuMouchel, W. and Harris, J. (1983). Bayes Methods for Combining the Results of Cancer Studies in Humans and Other Species (with discussion). *Journal of the American Statistical Association* **78**, 313-315.

[5] Hedges, L. and Olkien, I. (1987). *Statistical Methods for Meta Analysis*. Wiley, New York.

[6] Koenker, R. (1994). Confidence Intervals for Quantile Regression. *Proceedings of the 5th Prague Symposium on Asymptotic Statistics*, eds. P. Mandl and M. Huskova. Physica-Verlag, Heidelberg.

[7] Koenker, R. and Bassett, G. (1978). Regression Quantiles. *Econometrica* **46**, 33-50.

[8] Koenker, R. and Machado, J. (1999). Goodness of Fit and Related Inference Processes for Quantile Regression. *Journal of the American Statistical Association* **94**, 1296-1310.

[9] Reese, C., Wilson, A., Hamada, M., Martz, H. and Ryan, K. (2004). Integrated Analysis of Computer and Physical Experiments. *Technometrics* **46**, 153-164.

[10] Trindade, A. and Uryasev, S. (2005). Combining Model and Test Data for Optimal Determination of Percentiles and Allowables: CVaR Regression Approach, (a two part report). *Robust Optimization-Directed Design*, eds. A.J. Kurdila, P.M. Pardalos and M. Zabarankin. Kluwer Academic Publishers.

[11] Zhou, K. and Portnoy, S. (1996). Direct use of regression quantiles to construct confidence sets in linear models. *The Annals of Statistics* **24**, 287-306.

[12] Zhou, K. and Portnoy, S. (1998). Statistical inference on heteroscedastic models based on regression quantiles. *Journal of Nonparametric Statistics* **9**, 239-260.