9

Combining Model and Test Data for Optimal Determination of Percentiles and Allowables: CVaR Regression Approach, Part I

Stan Uryasev[1] and A. Alexandre Trindade[2]

[1] American Optimal Decisions, Inc.
and Department of Industrial and Systems Engineering
University of Florida
uryasev@ufl.edu
[2] Department of Statistics
University of Florida

Summary. We propose a coherent methodology for integrating various sources of variability on properties of materials in order to accurately predict percentiles of their failure load distribution. The approach involves the linear combination of factors that are associated with failure load, into a statistical factor model. This model directly estimates percentiles of the failure load distribution (rather than mean values as in ordinary least squares regression). A regression framework with CVaR deviation as the measure of optimality, is used in constructing the estimates. We consider estimates of confidence intervals for the estimates of percentiles, and adopt the most promising of these to compute A-Basis and B-Basis values. Numerical experiments with the available dataset show that the approach is quite robust, and can lead to a significant savings in number of actual testings. The approach pools together information from earlier experiments and model runs, with new experiments and model predictions, resulting in accurate inferences even in the presence of relatively small datasets.

Executive Summary

We propose a coherent methodology for integrating various sources of variability on properties of materials in order to accurately predict percentiles of their failure load distribution. The approach involves the linear combination of factors that are associated with failure load, into a statistical regression or *factor model*. The methodology can take as inputs various factors which can predict percentiles. For instance, for a given model, the difference between the 90th and 50th percentiles of the distribution generated by this model via Monte Carlo, can be viewed as a factor predicting the 90th percentile of the actual failure load distribution. Also, several model estimates, expert opinion estimates, and actual measurement estimates, can be pooled to estimate percentiles. However, in this particular study we were limited to the available failure dataset on one test type for a composite material, including the Model S predicted failure load limits, and actual experimental failure load data, for various stacking sequences. Specifically, and in the context of a supplied dataset we fit a model of the form,

$$Y = c_0 + c_1\mu + c_2\sigma + c_3 m + c_4 s + \varepsilon,$$

where Y is the failure load, $c_0, \ldots, c_4$ are unknown coefficients to be estimated from the data, and the pairs (μ, σ) are the midpoints and spreads of Model S limits (upper and lower bounds of failure loads) and (m, s) are estimates of the mean and standard deviation of the test data. ε is the residual error that accounts for the variability in failure loads that cannot be

explained through the factors. We develop the statistical tool of *CVaR regression*, a special case of which is *quantile regression*, first introduced by Koenker and Basset (1978). These allow us to estimate various percentile characteristics of the failure load distribution as a function of the factors. Recent related work on quantile regression provides methods for calculating confidence intervals for estimates of such percentiles. We adopt the most promising of these methods for the calculation of one-sided upper confidence intervals for the first and tenth percentiles. The lower limits of such intervals are by definition A-Basis and B-Basis values, respectively.

The so called *CVaR deviation measure* commonly used in financial risk management, is becoming increasingly popular in other areas such as supply chain and military risks. For a random value Y representing losses, Conditional Value-at-Risk (CVaR) measures the average of some fraction (typically 5%) of the highest losses. We show how CVaR emerges naturally in the objective function whose minimizer determines the quantile regression estimators. This means that it is equivalent to think about minimizing residual error tail risk when procuring percentile estimates. For A-Basis and B-basis calculation, it is more natural to think in terms of controlling the risk on the lower tail of the failure load distribution, and the CVaR regression objective function can be readily adapted for that purpose. As a consequence, we obtain a minimum that is more appropriate for assessing the goodness-of-fit of the data to the factor model.

The methodology is applied to the provided dataset in order to incorporate information from analytical model (Model S) predictions and actual open-coupon failure test data. For this dataset, the results, based on a combination of in-sample analyses, out-of-sample analyses, and Monte Carlo simulations, suggest the following.

(i) Model S information analyzed via CVaR regression, provides plausible percentile estimates, even in the absence of any experimental test information.
(ii) For each stacking sequence, the available dataset contained measurements only from one lab, which resulted in relatively small variability of failure loads among experiments. For such a dataset with (a small variability in actual measurements) the contribution from Model S predictions becomes insignificant in the presence of experimental test data.
(iii) Out-of-sample 10th percentile estimates based on Model S individually, and on Model S plus 5 test points, are close to their true values. B-basis values are also close to nominal values based on actual experiments.
(iv) For a dataset involving a larger variability in test data (generated by Monte Carlo simulations), results suggest that the Model S contribution is approximately equivalent to 4 test data points.
(v) For datasets with large variability in the test data, the number of actual test data points used can be significantly reduced (about 2 times) by using Model S predictions in out-of-sample calculations.

Advantages of the proposed methodology:

- Nonparametric; only mild distributional assumptions are made. Leads naturally to robust estimates.
- Direct estimation of percentiles as a function of the factors.
- Relatively simple approach; results are readily interpretable.
- It is possible to quantify the contribution to the response from different sources of information: physical models, experimental data, expert opinion, etc.
- Allows for pooling of data across different stacking sequences, resulting in only moderate sample size requirements for useful inferences.
- Allows for pooling of data across various experimental setups, resulting in the possibility to reduce the number complicated (expensive) experiments in lieu of simple (cheap) ones.
- Numerical efficiency: implemented via linear programming, leads to fast and stable numerical procedures.

Recommendations for implementation:

1. Certify a given analytical model (Model S etc.) for some area of application.
2. Build the factor model by splitting the available data into two portions: *training set* and *test set*. Use the first portion for model-building; the second for model cross-validation. Also, validate the factor model with Monte Carlo simulation of possible scenarios.
3. Use the resulting factor model combining modelling and experimental data for calculation of the 10th (1st) percentile and B-Basis (A-Basis) for the new variations of material (out-of-sample applications of the factor model).

9.1 Introduction

A basic problem in engineering design is how to manage uncertainty and risk in the properties of materials. Albeit costly, physical testing of the materials has traditionally been the primary method of quantifying the uncertainty. Recent advances in analytical physics-based engineering models and associated hardware/software technology, have made this an increasingly important contributor. From a monetary perspective, the lower cost of the modelling method makes it more desirable than the test method.

Rather than opt for one method over the other, a more effective risk management strategy might conceivably be the effective integration of the analytical models and the physical test program. The main objective of this study is to develop a sound methodology for such an integration. The resulting procedure will allow inferences in the form of point and interval estimates of percentiles to be made. Special cases include A-Basis and B-Basis values, defined in Appendix A.

Bayesian hierarchical modelling is one possibility when attempting to integrate heterogenous sources of data. Problems with this approach include the parametric assumptions about prior distributions that must be made, and the resulting large estimation errors for small data sets.

Financial engineering has recently made significant progress in working with percentile statistics and related optimization techniques. At the core of these developments is CVaR (Conditional Value-at-Risk), a new risk measure with appealing mathematical properties. The concepts and related methodology, are widely applicable to other engineering areas, and especially the military.

Drawing from our experience with these risk measures, we propose the use of factor models in combination with CVaR regression, as an alternative approach to integrate analytical model and physical test data. This involves treating a subset of the test data as the response variable, with sufficient statistics extracted from the remaining test data and model data as explanatory variables. CVaR regression, which includes the popular *quantile regression* of Koenker and Basset (1978) as a special case, is used to estimate the desired quantile of the response variable. In a case study, we demonstrate that CVaR regression is a viable methodology for combining analytical model and experimental test data.

9.2 Combining Model and Experimental Data: Factor Models and CVaR Regression

As outlined in the Introduction, we will use factor models to integrate model (or several models) predictions and experimental test data. In the conducted case study we integrated Model S predictions and experimental test data (henceforth *model* and *test* data) on the strength of composite materials.

Although beyond the scope of the present work, the approach is immediately generalizable to the integration of various other sources of heterogeneity, such as expert opinion.

Model setup

The idea is to use each of the test data values as the response in turn, with the remaining test data values as well as the model data, as explanatory variables or *factors*. Let Y_{ij} denote the jth test data point corresponding to the ith stacking sequence, $i = 1, \ldots, I$, and $j = 1, \ldots, n_i$.

We suppose that factors can be evaluated with the available model runs and experimental data, which can directly predict the percentiles of the failure distribution. For instance, we may have a structural reliability model which generates with a Monte Carlo simulation a histogram of failure load distribution for a specific stacking sequence and a specific experiment setup. Various statistical characteristics for this distribution can be considered as factors: mean, standard deviation, difference between 90th and 50th percentile, and others. Also a nonlinear transformation of data may be conducted (e.g. the logarithmic transformation for lognormally distributed data) to improve performance of the factor model.

Here, for demonstration purposes, we exemplify the approach with the factor model for the available dataset where for each stacking sequence only upper and lower limits of the failure distribution were generated by the Model S. Let the pair (X_{i1}, X_{i2}) be the Model S data for the ith stacking sequence. In this particular study, to restrict the number of factors, we condense the data from each source, test and model, into a pair of summary statistics: the mean and standard deviation. In statistical parlance, this pair is *sufficient* for the corresponding unknown parameters when sampling from a normal distribution (however, we *do not assume normality* of data; the approach is *distribution-free*). Other data reduction measures of location (e.g. median) and dispersion (e.g. lower semi-deviation) could also be used for datasets with asymmetrical distributions. Let (m_i, s_i) and (μ_i, σ_i) denote the sample mean and standard deviation for the test and model data, respectively, in stacking sequence i; that is

$$m_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}, \qquad s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i} (Y_{ij} - m_i)^2,$$

$$\mu_i = (X_{i2} + X_{i1})/2, \qquad \sigma_i = (X_{i2} - X_{i1})/2.$$

We now fit the regression or factor model:

$$Y_{ij} = c_0 + c_1\mu_i + c_2\sigma_i + c_3 m_i + c_4 s_i + \varepsilon_{ij}, \tag{9.1}$$

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 test data point in the ith stacking sequence, $i = 1, \ldots, I$, $j = 1, \ldots, n_i$. We develop a nonstandard *CVaR regression* technique, estimating percentiles of the response, rather than its mean value.

Factor model (9.1) essentially combines various estimates of strength in a linear way. In this framework, we use the regression model (see [12]) with the *Conditional Value-at-Risk (CVaR)* measure (see, [10],[11]). There are several variants of the methodology; CVaR can either be viewed as an optimality criterion, or as a constraint in the linear regression optimization problem. Here we consider a special case which leads to the same estimates as *quantile regression* due to Koenker and Bassett (1978), which directly estimates the quantiles or percentiles of the response variable Y, as a function of the generic factors μ, σ, m, and s.

The technical details behind the approach considered in the report are outlined in Appendices B and C. Appendix B on "Risk Measures" defines *deviation CVaR* measure of dispersion, which we use as the optimality criterion in this linear regression percentile estimation framework.

To introduce deviation CVaR, let the residual value ε be a continuous random variable with expectation $\mathsf{E}\varepsilon$ (the general case, including discrete distributions, is discussed in Appendix B). For a fixed probability level $0 \le \tau \le 1$, the *Value-at-Risk (VaR)* is defined to be the τth quantile of ε. For a given level τ, $\mathrm{VaR}_\tau(\varepsilon)$ is a lower bound on the largest $(1 - \tau)$ fraction of

values of ε. First, let us consider the upper tail of the distribution, i.e., $\tau > 0.5$. For continuously distributed ε, the deviation CVaR denoted by $\mathrm{CVaR}_{\tau}^{\Delta}(\varepsilon)$, is a conditional average of values $\varepsilon - \mathsf{E}\varepsilon$ under condition that ε is in the τ tail of the distribution, i.e.

$$\mathrm{CVaR}_{\tau}^{\Delta}(\varepsilon) \equiv \mathsf{E}[\,\varepsilon - \mathsf{E}\varepsilon \mid \varepsilon \geq \mathrm{VaR}_{\tau}(\varepsilon)]\ , \qquad \tau > 0.5.$$

Deviation CVaR at level τ provides an assessment of the $(1-\tau)$ fraction of the largest differences of failure loads and their mean values. It is a dispersion measure satisfying certain desirable axioms, see [11]. From the definition, deviation CVaR is the average distance between the mean value and the values in the τ tail of the distribution, and as such, is a one-sided measure of the width of such distribution. Compared, to [11] we have changed the sign of VaR and CVaR to avoid negative values for the positive failure loads ε, similarly to [10]. Deviation CVaR is convenient in statistical evaluations of percentiles, since linear regression with deviation CVaR (CVaR regression) lends itself naturally to estimation of such percentiles. For the lower tail, deviation CVaR is defined as

$$\mathrm{CVaR}_{1-\tau}^{\Delta}(-\varepsilon) \equiv \mathsf{E}[\,\mathsf{E}\varepsilon - \varepsilon \mid \varepsilon \leq \mathrm{VaR}_{\tau}(\varepsilon)]\ , \qquad \tau \leq 0.5.$$

Here, deviation CVaR is the average distance between the mean value and the values in the lower τ tail of the distribution.

CVaR regression then entails using the following loss function or optimality criterion in linear regression:

$$\mathcal{P}_{\tau}^{2} = \begin{cases} \mathrm{CVaR}_{\tau}^{\Delta}(\varepsilon)\,, & \text{if } \tau \geq 0.5, \\ \mathrm{CVaR}_{1-\tau}^{\Delta}(-\varepsilon)\,, & \text{if } \tau < 0.5. \end{cases} \tag{9.2}$$

For discrete distributions, the definition of the loss function is somewhat more complicated, but the intent is the same. The loss is approximately (or exactly) equal to the conditional expected value of the tail of the distribution. In the discrete case on hand, the minimization of the optimality criterion can be effectively achieved through linear programming, a robust and highly efficient numerical optimization procedure.

Minimizing deviation measure $\mathcal{P}_{\tau}^{2}$ of regression residuals (see, also, equation (9.23) in Appendix C), leads to the following estimated equation for the τth quantile of the failure load ε, as a function of the generic factors $\{\mu, \sigma, m, s\}$:

$$\hat{Q}_{Y}(\tau) = \hat{c}_0(\tau) + \hat{c}_1(\tau)\mu + \hat{c}_2(\tau)\sigma + \hat{c}_3(\tau)m + \hat{c}_4(\tau)s. \tag{9.3}$$

However, it can be observed that $\mathcal{P}_{\tau}^{2}$ in (9.2) depends on parameters c_1, c_2, c_3, c_4, but not on c_0. Therefore, the minimization procedure determines only $\hat{c}_1(\tau), \ldots, \hat{c}_4(\tau)$. The coefficient $\hat{c}_0(\tau) = \mathrm{VaR}_{\tau}(\hat{\varepsilon}(\tau))$ is chosen so as to ensure $\hat{Q}_{Y}(\tau)$ is an unbiased quantile estimate, where $\hat{\varepsilon}(\tau)$ is the residual in an optimal point. This procedure leads to the same weights as the quantile regression, see Appendix C. All parameters, $c_0, ..., c_4$, can be found in one run of the linear programming problem [Uryasev and Rockafellar 2000, 2002].

One important advantage of quantile/CVaR regression is that only mild distributional assumptions ((i) and (ii) in Appendix C) are made, making the approach essentially nonparametric. In particular, different stacking sequences are allowed to have different standard deviations. Note that:

> The A-basis ($\tau = 0.01$) and B-basis ($\tau = 0.10$) estimates for factor vector $\mathbf{x}' = [1, \mu_i, \sigma_i, m_i, s_i]$, are given by equation (9.17) in Appendix C, with $\alpha = 0.05$.

Appendix C demonstrates the connection between Koenker and Bassett's quantile regression idea, and our *CVaR regression*. Essentially, the objective function used to estimate the regression model parameters in the former, is equivalent to a CVaR objective function in the latter. The equivalence is in the sense that one obtains exactly the same parameter estimates, but the objective functions evaluated at these estimates differ. Since the parameter estimates are the minimizers of these objective functions, the attained minimum values are natural candidates

for measures of goodness-of-fit. We argue that the measure of goodness-of-fit in CVaR regression (equation (9.2)), being normalized by sample size, is better suited for making comparisons across fits from models with different sample sizes and confidence levels. Additionally, and of relevance in A and B-bases where lower bounds on estimates are more important than upper bounds, it takes into account the lower tail behavior of the residual error distribution.

Problems and extensions

Although not done in this study, factor model (9.1) can be easily extended to incorporate other factors contributing to differences in subpopulations. In particular, material/batch indicator variables for open-hole tension and compression, could be included.

Another important issue we have chosen to circumvent, is the potential need to incorporate weighting of coefficients when levels of factors have unequal numbers of replicates. This is the subject of *weighted regression*, and techniques used there could also be applied here. Simple solutions might be to assign weights proportional to the number of replicates in that level; or inversely proportional to the level standard deviation. For demonstration purposes, we consider only the simplest case of equal replication within levels of each material/batch.

Finally, perhaps the most complicated extension concerns the area of optimal design of experiments. Specifically: optimal choice of level combinations of factors with sufficient data to calibrate the factor model; selection of samples from different materials/batches for testing; and allocation of treatments to these experimental units. Again we are not addressing this here, but design of experiment principles should receive careful attention in any future work that might be contemplated.

The suggested CVaR regression technique is based on minimization of the deviation CVaR measure of the regression error. Alternatively, we can consider techniques which estimate percentiles by linear regression with the CVaR constraints, see, [13]. In further studies we can compare the presented in this paper approach versus the approach in [13].

Section 9.4 describes in more detail the issues which are suggested to be addressed in future studies.

9.3 Case Study

In this section we apply the methodology (CVaR regression) outlined in the previous section to the supplied data set, henceforth referred to as the *original dataset*. The original dataset consists of 31 rows corresponding to stacking sequences, each with a lower and upper Model S prediction, but unequal numbers of test data points. A sample of two rows from the dataset is seen if Table 9.1. The original dataset contained one stacking sequence with the very large number of 294 test points. This was omitted from the in-sample analyses, and held in reserve for subsequent out-of-sample cross validation. The test and model failure loads (ksi) for the resulting *full dataset* are plotted in Figure 9.1 against stacking sequence number. There are 186 test points in this dataset.

From the full dataset, a subset consisting of the 15 stacking sequences that had at least 5 test data points was selected. (There are a total of 143 test points in this dataset.) From these, to avoid introduction of coefficients compensating unequal number of experiments for different stacking sequences, exactly 5 test points were (randomly) selected within each stacking sequence. (Issues relating to statistical analyses with unequal numbers of datapoints will be addressed in subsequent studies.) The resulting *reduced dataset*, containing $5 \times 16 = 80$ test points, is seen in Figure 9.2.

Table 9.1. Two rows of failure loads from the original dataset.

Row Number	Test Data Points	Model S Upper Limit	Model S Average	Model S Lower Limit
1	137.34 118.96 119.41 110.64 103.67	115.6		109
mean	118	mean	112.3	
2	48.30 48.36 49.54 49.88 46.02	56.64		51.68
mean	48.42	mean	54.16	

In-sample results

We perform two in-sample analyses with these data. The first will use only Model S model factors; the second will use Model S plus five test points in each stacking sequence.

Analysis 1. In the first analysis, we fit a factor model using only the Model S predictors as factors. That is, we fit the model,

$$Y_{ij} = c_0 + c_1\mu_i + c_2\sigma_i + \varepsilon_{ij},$$

to the full dataset, where $i = 1, \ldots, 31$, and the range of j varies for each i. Total sample size (number of test points) here is $n = 186$. The 10th and 90th percentile estimated surfaces are:

$$\hat{Q}_Y(0.10) = -2.427 + 0.974\mu - 1.008\sigma,$$

and

$$\hat{Q}_Y(0.90) = -6.870 + 1.138\mu + 0.789\sigma,$$

with deviation measures $\mathcal{P}^2_{0.10} = 8.906$ and $\mathcal{P}^2_{0.90} = 13.337$. The estimated percentiles are shown on Figure 9.3 along with the test data points, as a function of the (ordered) stacking sequence number. Overall, 10.8% (10.8%) of test points fell above (below) the estimated 90th (10th) percentiles in their respective stacking sequence. The fitted values are seen to provide a reasonably good in-sample fit, even in the absence of any test point information.

Analysis 2. In the second analysis, we fit the complete factor model using both Model S and test point data as factors. That is, we fit regression model (9.1) to the reduced dataset, where $i = 1, \ldots, 16$, and $j = 1, \ldots, 5$. The pairs of stacking sequences (2,23) and (3,24) were also merged. Total sample size here is $n = 5 \times 16 = 80$. The 10th and 90th percentile estimated surfaces are:

$$\hat{Q}_Y(0.10) = 0.430 - 0.015\mu - 0.075\sigma + 1.013m - 1.129s,$$

and

$$\hat{Q}_Y(0.90) = -0.167 + 0.058\mu - 0.080\sigma + 0.943m + 1.345s,$$

with deviation measures $\mathcal{P}^2_{0.10} = 4.901$ and $\mathcal{P}^2_{0.90} = 5.225$. A substantial drop in the respective deviation measures from Analysis 1 has occurred, indicative of an improvement in the fit. The estimated percentiles are shown on Figure 9.4 along with all 143 test data points, as a function of the (ordered) stacking sequence number. From (9.16), 95% confidence intervals for the model parameters of $Q_Y(0.10)$, are as follows:

$$c_0 : (-1.58, 2.44), \quad c_1 : (-0.10, 0.07), \quad c_2 : (-3.55, 0.20),$$

$$c_3 : \quad (0.93, 1.10), \quad c_4 : (-1.31, -0.95),$$

so that only c_3 and c_4 are significant. Overall, taking all 143 test points into consideration, 16.8% (15.4%) of test points fell above (below) the estimated 90th (10th) percentiles in their respective stacking sequence. The fitted values are seen to provide a reasonably good in-sample fit, even in the absence of any test point information.

Summary Remarks

- Model S information provides plausible percentile estimates, even in the absence of any experimental test information.
- In the considered dataset measurements only from one lab for each stacking sequence are available. This undoubtedly translates into smaller variability, both within and among stacking sequences, resulting in perhaps over-optimistic fits.
- Small variability in the experimental data may lead to incorrect estimation of the predictive capabilities of the model. The contribution from Model S information becomes insignificant in the presence of a small amount of experimental test information.
- We think that data from several labs needs to be incorporated into the analyses for a more realistic assessment of the predictive capabilities of each source of information.

Out-of-sample results

The out-of-sample predictive capabilities of the models fitted in Analyses 1 and 2 are now assessed on the omitted stacking sequence with 294 test points. The Model S lower and upper failure load bounds for this stacking sequence are 37.86 and 39.75 respectively, which from Figure 9.2 are seen to be at the lower extremum of all Model S bounds. Also, the actual 10th percentile of all 294 test points is 35.84, which is on the boundary of range of percentile estimates for other stacking sequences.

Analysis 3. Using only the Model S values, the fitted model from Analysis 1 yields an estimate of 34.04 for the 10th percentile of failure loads. The calculated value of the B-Basis is 31.75. As it is supposed to be, the estimated B-Basis value, 31.75, is lower than the actual experimental 10th percentile estimate which equals to 35.84. A histogram of the test failure loads for this out-of-sample stacking sequence is seen in Figure 9.5.

Analysis 4. Using Model S and the five randomly chosen test values 38.12, 37.214, 37.637, 37.707, and 35.63, the fitted model from Analysis 2 yields the estimate of 36.36 for the 10th percentile of failure loads. The calculated value of the B-Basis is 35.21. The estimated B-Basis value, 35.21, is lower than the actual experimental 10th percentile, 35.84.

A histogram of the test failure loads for this out-of-sample stacking sequence is seen in Figure 9.6.

Aiming at a more precise quantification, the procedure of randomly selecting 5 test points and fitting the model of Analysis 2, was repeated 5,000 times. In 78% of cases, the actual 10th percentile estimate of failure loads, 35.84, is above the B-Basis value. This is somewhat lower than the nominal value of 95%, but it must be remembered that this is an out-of-sample inference.

Monte Carlo results

For our final round of analyses, the Model S predictions for the 31 stacking sequences with at least 3 test points were selected from the full dataset. To simulate a more heterogenous scenario and incorporate between lab variability, 3 sets of 3 points (9 altogether) were then

generated by Monte Carlo for each stacking sequence. This was done so as to ensure a greater between lab than within one lab variability, each set of 3 points viewed as originating from a different lab. The resulting *artificial dataset*, containing $9 \times 31 = 279$ test points, is seen in Figure 9.7.

Analysis 5. In the first scenario, and in analogy with Analysis 1, we considered estimation of the 10th and 90th percentiles based on Model S only information. The 10th percentile estimated surface is:

$$\hat{Q}_Y(0.10) = 4.377 + 0.826\mu - 1.186\sigma,$$

with deviation measure $\mathcal{P}^2_{0.10} = 15.077$. The results are presented in Figure 9.8. 95% confidence intervals for the model parameters of $Q_Y(0.10)$, are as follows:

$$c_0 : (-0.96, 9.72), \quad c_1 : (0.76, 0.89), \quad c_2 : (-1.88, -0.49),$$

so that only c_1 and c_2 are significant.

Analysis 6. We now add test point information. We start by selecting 3 test points within each stacking sequence to form m and s; using the remaining 6 points as responses. We do this for each of the $\binom{9}{3} = 84$ possible ways to select 3 points from 9, thus creating a dataset of size $n = 6 \times 84 \times 31 = 15,624$. The 10th percentile estimated surface for this Model S plus 3 points dataset is:

$$\hat{Q}_Y(0.10) = 2.487 + 0.671\mu - 0.717\sigma + 0.202m - 0.377s,$$

with deviation measure $\mathcal{P}^2_{0.10} = 14.715$. 95% confidence intervals for the model parameters are as follows:

$$c_0 : (1.87, 3.11), \quad c_1 : (0.64, 0.70), \quad c_2 : (-0.80, -0.64),$$

$$c_3 : (0.18, 0.23), \quad c_4 : (-0.41, -0.34),$$

Although all coefficients are significant, the addition of 3 test points per stacking sequence does not appreciably change the model based on Model S only information. $\mathcal{P}^2_{0.10}$ has decreased by 2.4% (from 15.077 to 14.715) and the coefficients from the Model S contribution outweigh those from the test point contribution. This demonstrates that the Model S for this particular dataset has a high predictive power and adding several actual measurements do not bring a lot of new information.

Analysis 7. In an attempt to quantify the Model S contribution in terms of test data points, we consider a factor model with only test point information. The calculation of m and s within each stacking sequence is based on selecting 7 test points; the remaining 2 points used as responses. Since there are $\binom{9}{7} = 36$ possible ways to select 7 points from 9, the resulting dataset has size $n = 2 \times 36 \times 31 = 2,232$. The 10th percentile estimated surface for this 7 points dataset is:

$$\hat{Q}_Y(0.10) = 0.725 + 0.933m - 0.967s,$$

with deviation measure $\mathcal{P}^2_{0.10} = 14.484$.

The deviation measure for the 7 test points dataset, 14.484, is approximately the same as the deviation measure for the Model S plus 3 test points dataset, 14.715. Thus the *Model S contribution can be roughly equated to 4 test data points*.

Analysis 8. Lastly, we mimic Analysis 2 by incorporating Model S and all 9 test points into the construction of the factors. m and s are now calculated from all 9 test points within each stacking sequence; the same 9 points used as responses. The resulting dataset has size $n = 9 \times 31 = 279$. The 10th percentile estimated surface for this Model S plus 9 test points dataset is:

$$\hat{Q}_Y(0.10) = 0.000 + 0.041\mu - 0.140\sigma + 0.965m - 1.214s,$$

with deviation measure $\mathcal{P}^2_{0.10} = 11.697$. The results are presented in Figure 9.9. 95% confidence intervals for the model parameters are as follows:

$$c_0 : (-2.63, 2.63), \quad c_1 : (-0.13, 0.22), \quad c_2 : (-0.48, 0.20),$$

$$c_3 : (0.79, 1.14), \quad c_4 : (-1.46, -0.97),$$

so that the Model S contribution is not significant.

9.4 Summary and Recommendations

We have proposed a risk management strategy for integrating various sources of information into one coherent factor model. The sound methodology of CVaR Regression was developed to enable direct estimation of percentile characteristics of the response, as a function of explanatory variables. This was applied to the provided dataset in order to incorporate information from analytical model (Model S) predictions and physical test data. The results, based on a combination of in-sample analyses, out-of-sample analyses, and Monte Carlo simulations, suggest the following.

(i) Model S information provides plausible percentile estimates, even in the absence of any experimental test information.
(ii) In the available dataset, one lab measurement, is available for each stacking sequence, which results in a low variability of actual measurements. This condition makes the contribution from Model S insignificant in the presence of even small amounts of experimental data.
(iii) For one out-of-sample stacking sequence, 10th percentile estimates based on Model S as well as on Model S plus 5 test points are close to their true values. However, coverage probabilities for B-Basis estimates deviated about 20% from their nominal 95% levels.
(iv) With Monte Carlo simulations variability in measurements between different labs was introduced. For the simulated dataset, in the presence of Model S plus experimental test information, Model S contribution is approximately equivalent to 4 test data points.
(v) The number of test data points needed can be significantly reduced by using Model S predictions; both in-sample and out-of-sample. Experiments showed that Model S provides acceptable estimates of percentiles and B-Basis even without any actual measurements.

Advantages of the proposed methodology

We list some of the benefits of the proposed factor model with CVaR regression approach.

- Nonparametric; only mild distributional assumptions are made. Leads naturally to robust estimates.
- Leads to direct estimation of percentiles as a function of the factors.
- Relatively simple approach; results are readily interpretable.
- It is possible to quantify the contribution to the response from different sources information: physical models, experimental data, expert opinion, etc.
- Allows for pooling of data across different stacking sequences, resulting in only moderate sample size requirements for useful inferences.
- Allows for pooling of data across various experimental setups, resulting in the possibility to reduce the number of sophisticated (expensive) experiments in lieu of simple (cheap) ones.
- Numerical efficiency: implemented via linear programming, leads to fast and stable numerical procedures.

Recommendations for implementation and future research

We conclude with some guidelines for applying the proposed methodology to new situations. This report briefly discussed the factor model approach and suggested implementation steps. However, for practical applications, each step should be elaborated in detail, tested and documented in user-friendly format. Firstly, we outline the basic steps in applying the methodology. Secondly, we discuss several issues which need to be addressed in further studies.

Basic Steps of the Factor Analysis Approach

1. Certify a given analytical model (Model S etc.) for some area of application.
2. Build the factor model by splitting the available data into in-sample and out-of-sample portions. Use the former for model-building (the *training set*); the latter for model cross-validation (the *test set*).
3. Validate the choice of selected model on the out-of-sample portion, and by using Monte Carlo simulation of possible scenarios.
4. Use the resulting factor model for calculation of the 1st (10th) percentile and A-Basis (B-Basis).

Issues to be Addressed in Further Studies

1. The model certification process should be investigated in detail. Certification issues have been discussed in the framework of [1]. However, much more research in this area is needed.
2. Model S predictive capabilities have been studied in this report. Other structural reliability models can be investigated and compared with Model S.
3. Guidelines for the planning of the experiments should be developed. How much data is needed for designing training and test sets? How to allocate resources between activities on conducting experiments and model developments? How to take into account existing data which may provide unbalanced information on different variations of material?
4. *Optimality considerations.* We have suggested factor models for estimating percentiles of failure load distribution and CVaR regression methodology for finding the optimal weighting coefficients. Calculations for the considered example show that optimal coefficients are quite sensitive to variability in the actual measurements. We think that weighting coefficients for various factors may not need to be optimal for achieving acceptable precision of predictions. It may be reasonable to identify a set of factors and fix some coefficients for these factors for certain areas of application. These factors and *fixed coefficients* can be tested with various datasets and with extensive Monte Carlo calculations.
5. Procedures for calculating A-basis and B-basis in combination with factor models should be studied. In this report, percentile confidence intervals were calculated under a mild parametric assumption on the kernel bandwidth. Other promising approaches such as bootstrapping (which is nonparametric), also need to be evaluated.
6. Issues relating to robustness of the approach and sensitivity to various factors, e.g. amount of data, variability of data inputs, impact of outliers, etc., should be investigated.
7. The current case study estimated 10th and 90th percentiles of the failure load distribution. The amount of data in the supplied dataset is not sufficient for estimating the 1st and 99th percentiles in the framework of the CVaR regression methodology. Parametric approaches that are more suited for high percentile estimation, such as *extreme value theory*, are worthy of further investigation.
8. Performance of the suggested CVaR regression approach can be compared with other approaches, including the Bayes approach, by testing with several datasets and Monte Carlo simulations.

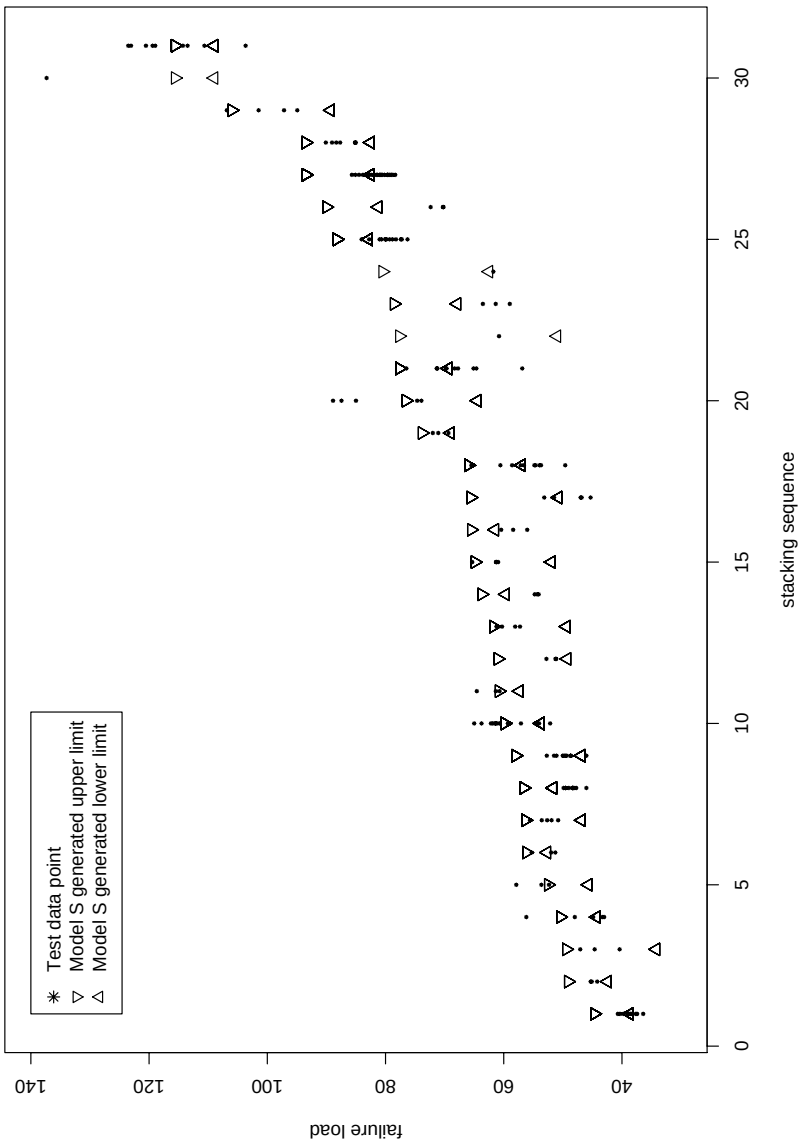

Fig. 9.1. Model S upper and lower failure load bounds for the full dataset.

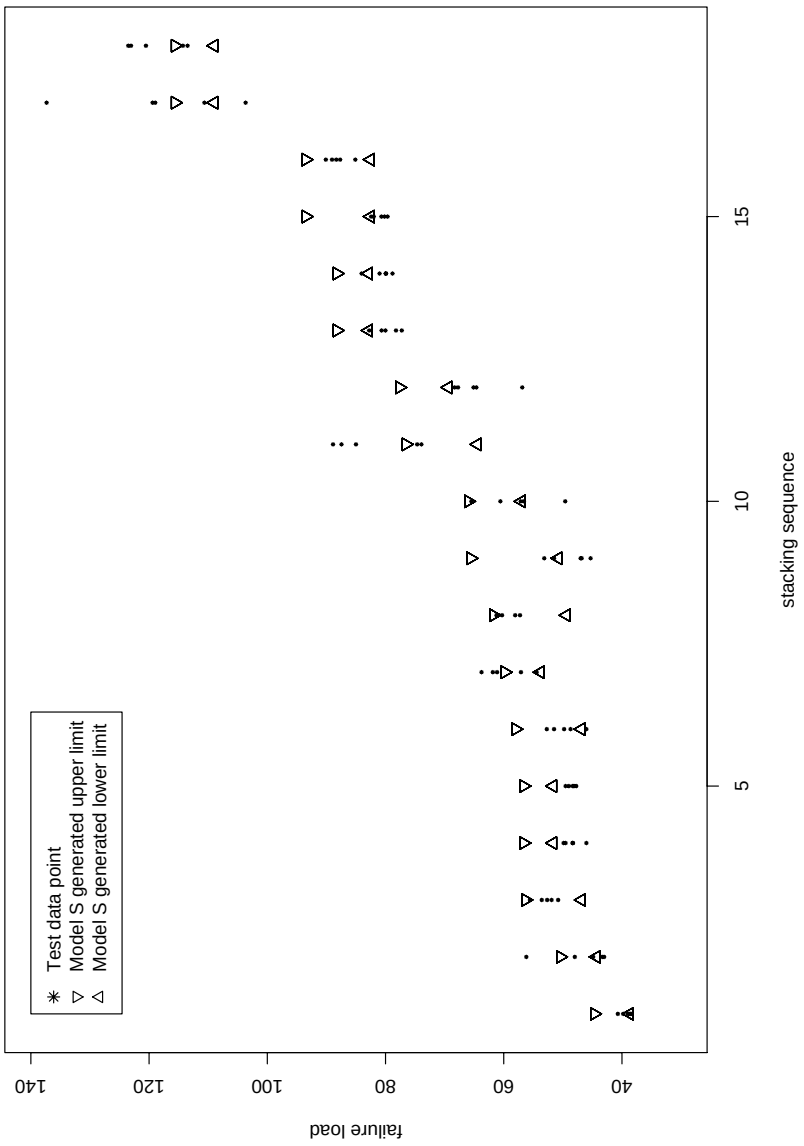

Fig. 9.2. Model S upper and lower failure load bounds for the reduced dataset.

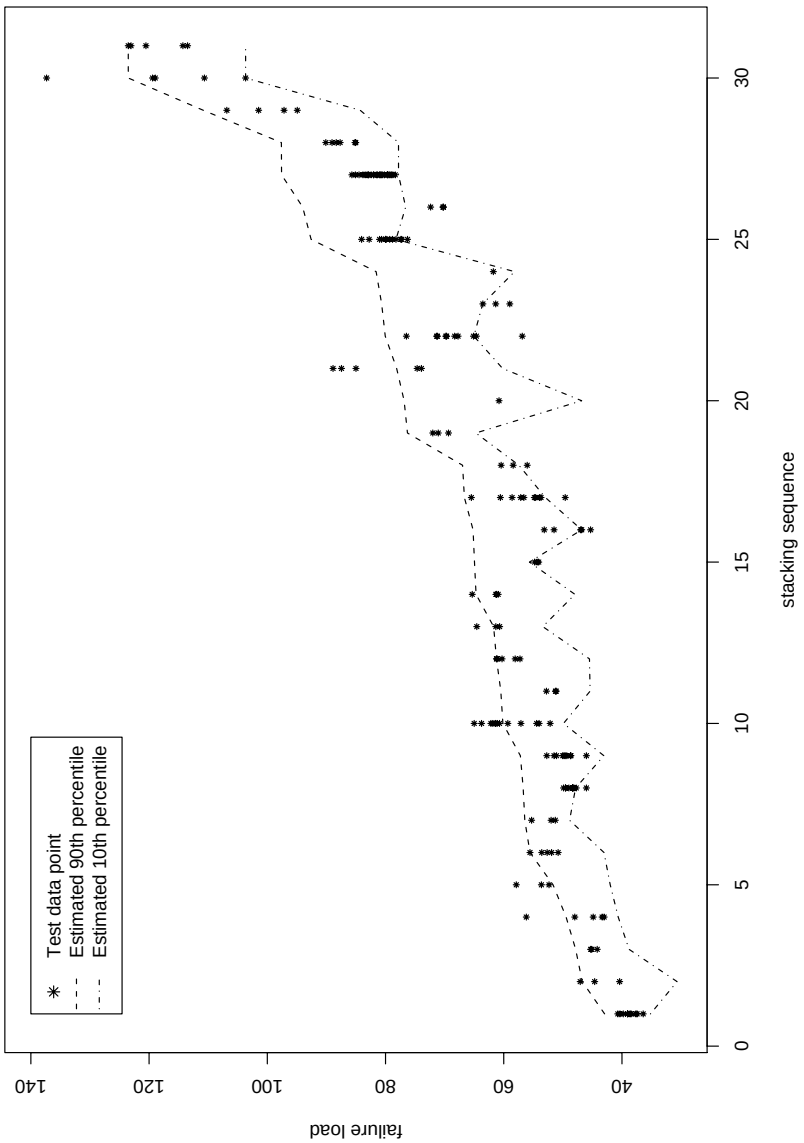

Fig. 9.3. Estimated 10th and 90th percentiles for the full dataset based on Model S only information.

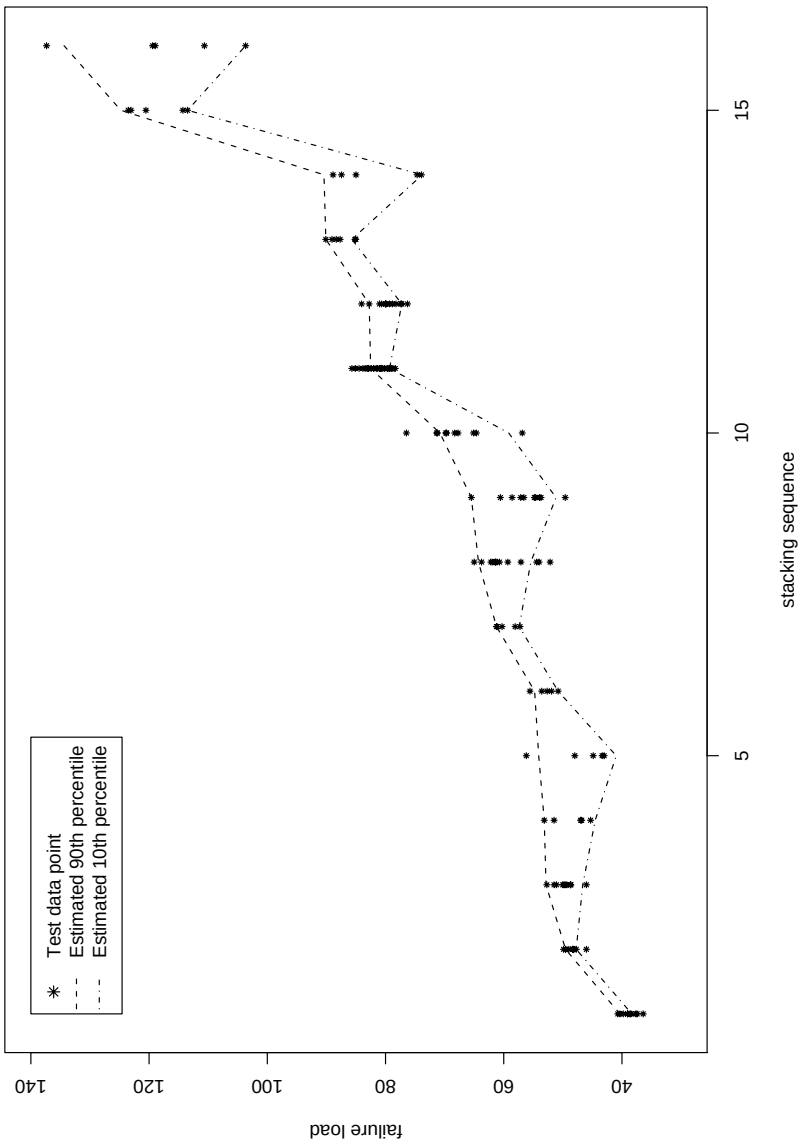

Fig. 9.4. Estimated 10th and 90th percentiles for the reduced dataset based on Model S plus 5 randomly selected test points.

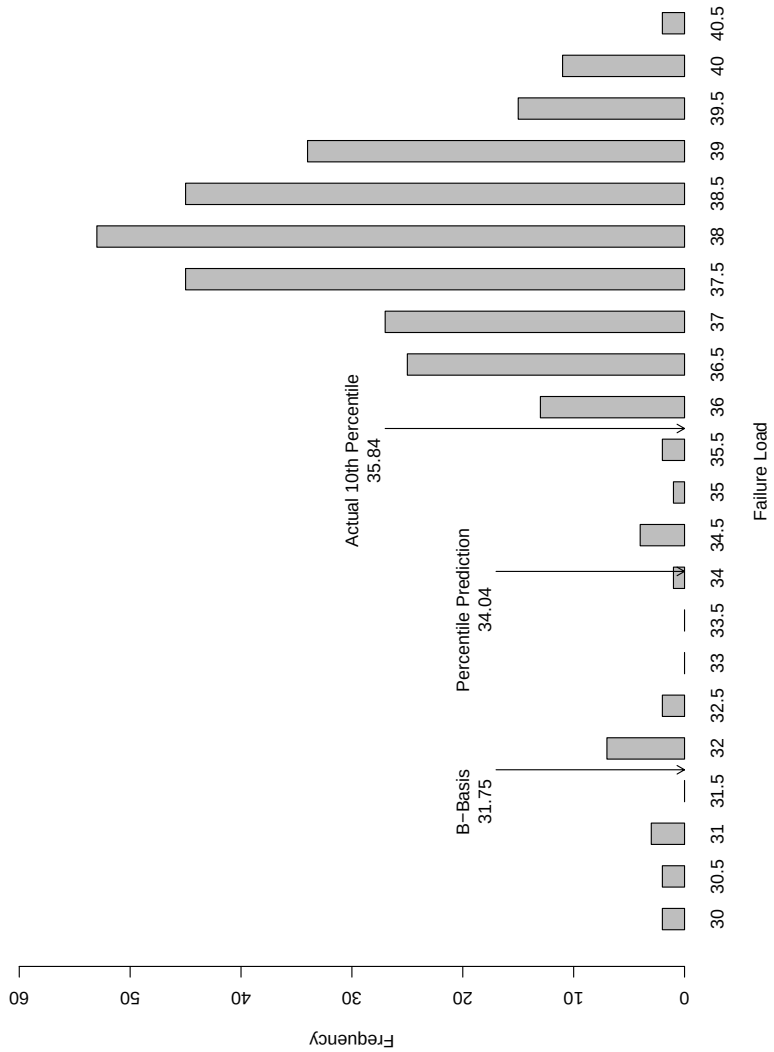

Fig. 9.5. Histogram of the 294 test failure loads for the out-of-sample stacking sequence. The actual and estimated 10th percentiles are shown, along with the B-Basis value. These are based on the fitted model of Analysis 1.

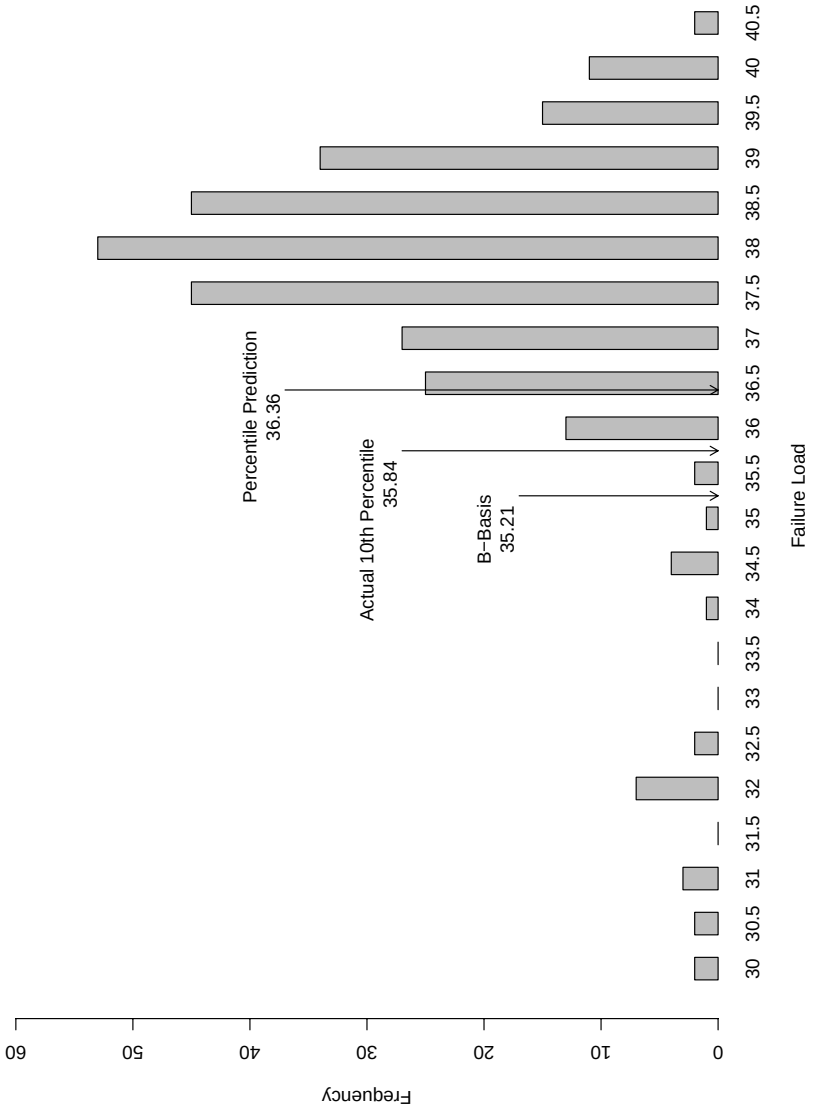

Fig. 9.6. Histogram of the 294 test failure loads for the out-of-sample stacking sequence. The actual and estimated 10th percentiles are shown, along with the B-Basis value. These are based on the fitted model of Analysis 2, with test points 38.12, 37.214, 37.637, 37.707, 35.63.

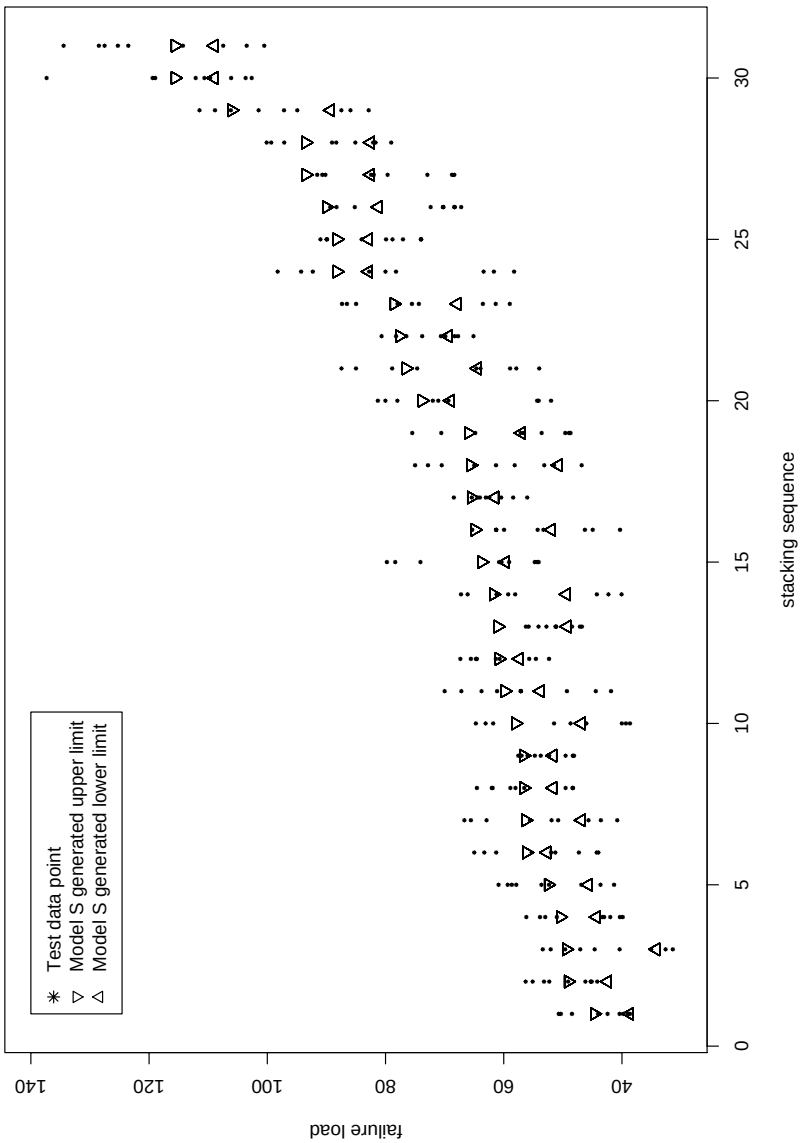

Fig. 9.7. Model S upper and lower failure load bounds for the artificial dataset, along with all 279 test data points.

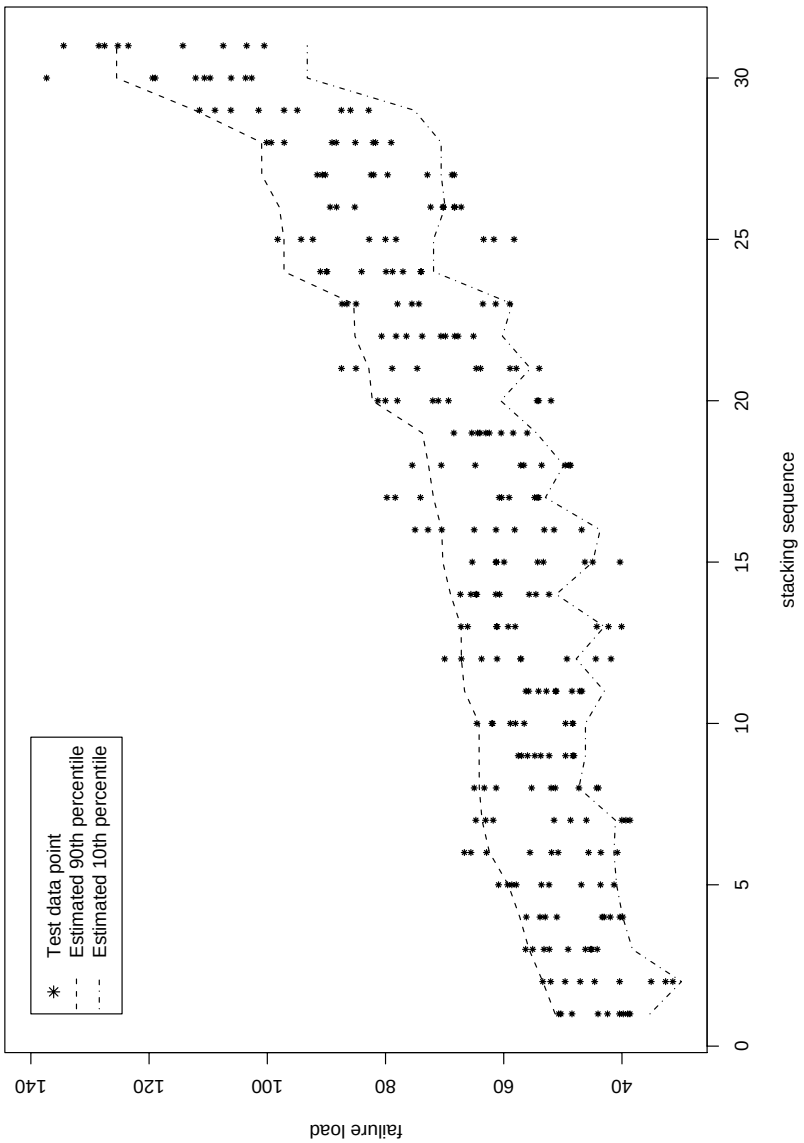

Fig. 9.8. Estimated 10th and 90th percentiles based on Model S only information for the artificial dataset.

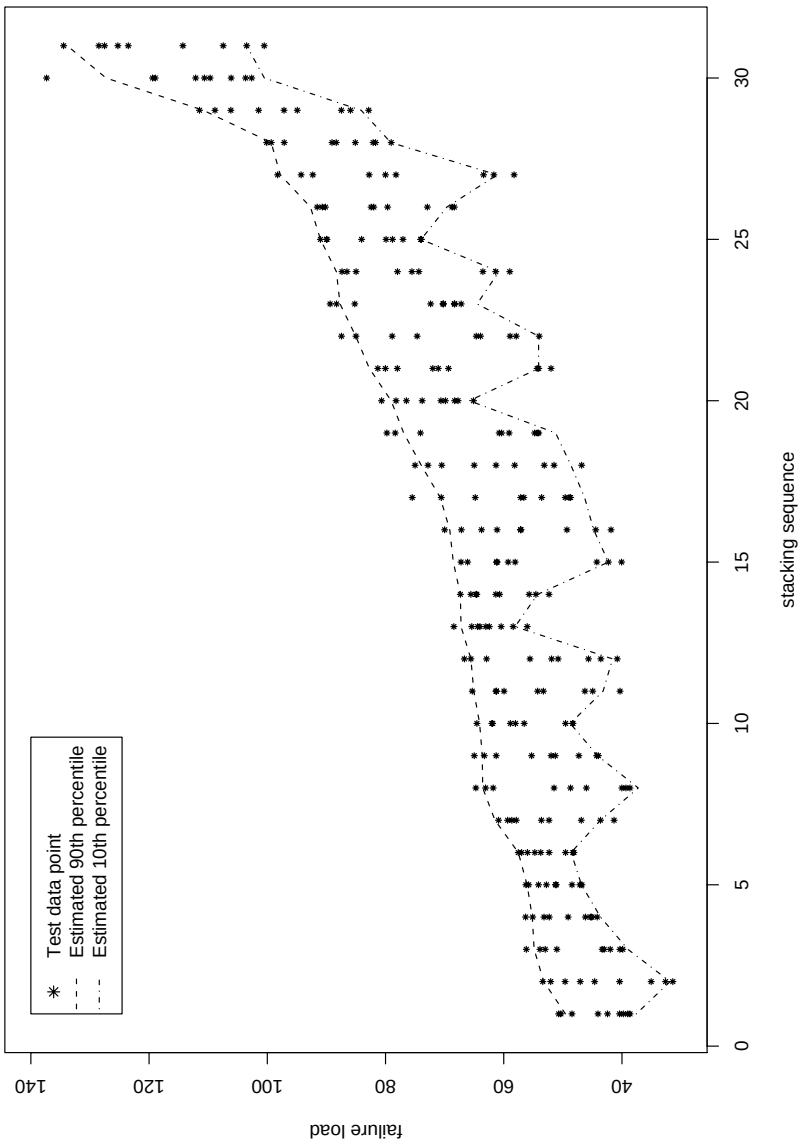

Fig. 9.9. Estimated 10th and 90th percentiles based on Model S plus 9 test points for the artificial dataset.

Acknowledgements

This effort was jointly accomplished by the Boeing led 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 would like to acknowledge the guidance and support of Dr. Leo Christodoulou of DARPA/DSO for this effort.

A Definition of A-basis and B-basis values

Let $\{X_1, \ldots, X_n\}$ be a random sample of observations from random variable X with 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,

$$\begin{aligned} 0.95 &= P\left[1 - F(\xi_\tau) > 1 - \tau\right] && (9.4)\\ &= P\left[F(\xi_\tau) < \tau\right] && (9.5)\\ &= P\left[\xi_\tau < F^{-1}(\tau)\right]. \end{aligned}$$

Representation (9.5) means that (ξ_τ, ∞) is 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 (9.4) 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, $\xi_{0.10}$.

B Risk Measures

Financial *risk measures* provide a useful class of objective functions with a wide spectrum of applications. To introduce these, let Y be a continuous random variable with $\mathsf{E}Y < \infty$, and distribution function $F(y)$. For a fixed probability level $0 \leq \tau \leq 1$, the *Value-at-Risk (VaR)* is defined to be the τth quantile of Y,

$$\text{VaR}_\tau(Y) \equiv F^{-1}(\tau).$$

With Y measuring profit/loss, the conceptual simplicity of VaR has resulted in its widespread incorporation into standard banking regulation, where it is imperative to assess credit risk.

However, note that for a given level τ, $\text{VaR}_\tau(Y)$ is merely a lower bound on the worst $(1-\tau)$ fraction of losses, and thus fails to capture any information about the magnitude of losses beyond that level. This and other undesirable mathematical properties inherent in VaR, are documented by Artzner *et al.* (1997, 1999), who introduce an axiomatic method for the proper construction of risk measures. With $(\nu)^+ = \max\{0, \nu\}$ for any real ν, they propose an alternative measure of risk, *Conditional Value-at-Risk (CVaR)*, defined as the solution of the minimization problem

$$\text{CVaR}_\tau(Y) \equiv \min_{a \in \mathscr{R}} \left\{ a + \frac{1}{1-\tau}\mathsf{E}(Y-a)^+ \right\}. \tag{9.6}$$

Rockafellar and Uryasev (2000) shows that the minimizer is actually $\text{VaR}_\tau(Y)$. Also, Rockafellar and Uryasev (2002) show that for continuous Y,

$$\text{CVaR}_\tau(Y) = \mathsf{E}\left[Y \mid Y \geq \text{VaR}_\tau(Y)\right],$$

i.e., CVaR equals conditional expectation of the τ tail of the distribution. For general distributions, including discrete distributions, CVaR approximately (or exactly) equals the conditional expectation of the τ tail of the distribution. In this sense, CVaR at level τ provides

a more realistic assessment of the worst $(1-\tau)$ fraction of losses, by evaluating their expected value. As it has been proved by Pflug (2000), CVaR is a *coherent* measure of risk in the sense of Artzner *et al.* (1997, 1999), and satisfies several desirable properties: translation invariance, sub-additivity, positive homogeneity, and monotonicity with respect to stochastic dominance. Figure B illustrates the relationship between VaR and CVaR. Note that one always has $\text{VaR}_\tau(Y) \leq \text{CVaR}_\tau(Y)$.

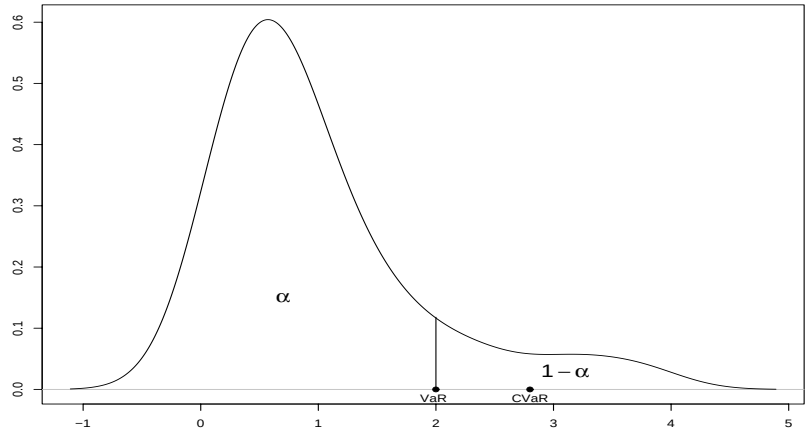

Fig. 9.10. Example of the relationship between VaR and CVaR at quantile α for a continuous distribution.

For discrete Y with probability mass points at $y_1, \ldots, y_n$, and corresponding masses $p_1, \ldots, p_n$, (9.6) becomes,

$$\text{CVaR}_\tau(Y) = \min_{a \in \mathscr{R}} \left\{ a + \frac{1}{1-\tau} \sum_{i=1}^{n} (y_i - a)^+ p_i \right\}. \tag{9.7}$$

In transitioning to an empirical setting when a sample of observations $\{y_1, \ldots, y_n\}$ from Y is available, VaR and CVaR can be estimated via method of moments:

$$\widehat{\text{VaR}}_\tau(Y) = \arg\min_{a \in \mathscr{R}} \left\{ a + \frac{1}{n(1-\tau)} \sum_{i=1}^{n} (y_i - a)^+ \right\} \tag{9.8}$$

$$\widehat{\text{CVaR}}_\tau(Y) = \min_{a \in \mathscr{R}} \left\{ a + \frac{1}{n(1-\tau)} \sum_{i=1}^{n} (y_i - a)^+ \right\}. \tag{9.9}$$

This amounts to assigning the uniform mass $p_i = 1/n$ to each y_i in (9.7).

As already introduced in Section 9.2, for general random variable Y the quantity

$$\text{CVaR}_\tau^{\Delta}(Y) \equiv \text{CVaR}_\tau(Y - \mathsf{E}Y), \tag{9.10}$$

is a *deviation measure*, as defined by Rockafellar, Uryasev and Zabarankin (2002). This class of functionals was introduced to generalize the properties of symmetric deviation measures such as standard deviation, thus allowing for the definition of more asymmetric ones. In words, (9.10) equals approximately (or exactly) to the expectation of the right tail of the centered

random variable $Y - \mathsf{E}Y$, and as such can be viewed as a measure of dispersion of its right tail.

When $\tau < 0.5$ (A and B-basis), it is more meaningful to think about minimizing left-tail CVaR, rather than right-tail. We therefore propose the measure:

$$\mathcal{P}_\tau^2 = \begin{cases} \mathrm{CVaR}_\tau^{\Delta}(Y), & \text{if } \tau \geq 0.5, \\ \mathrm{CVaR}_{1-\tau}^{\Delta}(-Y), & \text{if } \tau < 0.5, \end{cases} \tag{9.11}$$

as the optimality criterion or measure of goodness-of-fit in quantile regression estimation. This gives rise to what we are calling *CVaR regression*. Since Pflug (2000) shows that for any random variable Z,

$$\mathrm{CVaR}_\tau(Z - \mathsf{E}Z) = \frac{\tau}{1-\tau}\mathrm{CVaR}_{1-\tau}(\mathsf{E}Z - Z),$$

we can rewrite (9.11) as,

$$\mathcal{P}_\tau^2 = \begin{cases} \mathrm{CVaR}_\tau^{\Delta}(Y), & \text{if } \tau \geq 0.5, \\ \frac{1-\tau}{\tau}\mathrm{CVaR}_\tau^{\Delta}(Y), & \text{if } \tau < 0.5. \end{cases}$$

In this way, the emphasis is on adequately accounting for under-estimates of quantiles, the major source of concern in this study.

In case of discrete distributions, minimization of CVaR with loss functions which linearly depend upon control parameters can be reduced to linear programming, see, Rockafellar and Uryasev, 2000, 2002.

In the next section, we will demonstrate that there is a connection between CVaR regression as we have defined it, and quantile regression as defined by Koenker and Bassett (1978). This will lead naturally to the use of CVaR regression as a generalization of quantile regression.

C Quantile Regression and CVaR Regression

Consider the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon = \boldsymbol{X}'\beta + \varepsilon, \tag{9.12}$$

where $\beta = [\beta_0, \ldots, \beta_p]'$, and $\boldsymbol{X} = [1, X_1, \ldots, X_p]'$. The factors $\{X_1, \ldots, X_p\}$ are viewed as independent random variables, and independent from ε which has cdf F and pdf f. A random sample of observations from this model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i = \boldsymbol{X}_i'\beta + \varepsilon, \quad i = 1, \ldots, n,$$

with the $\{\varepsilon_i\}$ i.i.d., can be concisely written as,

$$\mathbf{Y} = \mathrm{X}\beta + \varepsilon, \tag{9.13}$$

where X denotes the n by $p+1$ design matrix whose ith row is $\boldsymbol{X}_i'$. When transitioning to the empirical setting, we will use lower case, $\{y_i\}$ and $\{x_{ij}\}$, to denote realized values of the corresponding random variables.

Now, note that for random variable Z and constant c, the quantile functions of $Z + c$ and Z are related by $F_{Z+c}^{-1}(\tau) = F_Z^{-1}(\tau) + c$. Since

$$P(Y \leq y | \boldsymbol{X} = \mathbf{x}) = F\left(y - \mathbf{x}'\beta\right),$$

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,

$$Q_Y(\tau|\mathbf{x}) = \mathbf{x}'\beta + F^{-1}(\tau) \equiv \mathbf{x}'\beta(\tau), \tag{9.14}$$

where $\beta_0(\tau) = \beta_0 + F^{-1}(\tau)$, 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.

Quantile Regression

If $\mathbf{y}' = [y_1, \ldots, y_n]$ is the observed vector of responses, Koenker and Bassett (1978) proposed the following estimator of $\beta(\tau)$:

$$\begin{aligned}\hat{\boldsymbol{\beta}}(\tau) &= \arg\min_{\mathbf{b}\in\mathscr{R}^{p+1}} \sum_{i=1}^{n} \left\{\tau(y_i - \mathbf{x}_i'\mathbf{b})^+ - (1-\tau)(y_i - \mathbf{x}_i'\mathbf{b})^-\right\} \qquad (9.15)\\ &\equiv \arg\min_{\mathbf{b}\in\mathscr{R}^{p+1}} \mathcal{P}_\tau^0(\mathbf{b}),\end{aligned}$$

where $(\nu)^+ = \max\{0, \nu\}$ and $(\nu)^- = \min\{0, \nu\}$. This is a natural generalization to linear regression, of the consistent estimator of the τth quantile

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

in the special case of (9.12) when $\beta_j = 0$, $j = 1, \ldots, p$ (the *location model*). For any design point $\mathbf{x}$, the empirical conditional quantile function is then defined by Koenker and Bassett (1982) as

$$\hat{Q}_Y(\tau|\mathbf{x}) = \mathbf{x}'\hat{\boldsymbol{\beta}}(\tau).$$

In case (9.15) does not have a unique solution, the minimum value of $\hat{Q}_Y(\tau|\mathbf{x})$ over all such $\hat{\boldsymbol{\beta}}(\tau)$ is taken.

Koenker and Bassett (1978) also show that under the mild design conditions,

(i) $\{\varepsilon_i\}$ i.i.d. with continuous cdf F, and continuous and positive density, f, at $F^{-1}(\tau)$,
(ii) $\lim_{n\to\infty} n^{-1}\mathrm{X}'\mathrm{X} \equiv D$, is a positive definite matrix,

with $\Omega \equiv \tau(1-\tau)/f^2\left(F^{-1}(\tau)\right) D^{-1}$, we have the asymptotic result:

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}(\tau) - \beta(\tau)\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \Omega\right).$$

Defining $\Omega_n \equiv n^{-1}\Omega$, this means $\hat{\boldsymbol{\beta}}(\tau)$ and $\mathbf{x}'\hat{\boldsymbol{\beta}}(\tau)$ are asymptotically normal

$$\hat{\boldsymbol{\beta}}(\tau) \sim \mathrm{AN}(\beta(\tau), \Omega_n), \quad \text{and} \quad \mathbf{x}'\hat{\boldsymbol{\beta}}(\tau) \sim \mathrm{AN}(\mathbf{x}'\beta(\tau), \mathbf{x}'\Omega_n\mathbf{x}).$$

The only obstacle to immediate construction of asymptotics-based confidence intervals for regression quantiles therefore, is the estimation of the *sparsity function*,

$$s(\tau) = \left[f\left(F^{-1}(\tau)\right)\right]^{-1}.$$

Koenker (1994), argues that a natural candidate is

$$\hat{s}_n(\tau) = \left[\hat{F}_n^{-1}(\tau + h_n) - \hat{F}_n^{-1}(\tau - h_n)\right]/(2h_n),$$

with $\hat{F}_n^{-1}(\cdot)$ the usual empirical quantile function based on the quantile regression residuals. The most difficult question of course is selection of the bandwidth h_n. The heuristic argument and simulations in Koenker (1994), indicate that the normal model version of the Hall and Sheather (1988) method is promising. For construction of $(1-\alpha)100\%$ confidence intervals, this method identifies

$$h_n = \left[\frac{3z_{1-\alpha/2}^2 e^{-z_\tau^2}}{4n\pi(2z_\tau^2 + 1)}\right]^{1/3},$$

as the optimal bandwidth. (For probability r, z_r denotes the rth quantile of a standard normal random variable.) With the above choices, Ω_n can be estimated by,

$$\hat{\Omega}_n = \tau(1-\tau)(\mathrm{X}'\mathrm{X})^{-1}\hat{s}_n(\tau)^2.$$

This leads immediately to the following $(1-\alpha)100\%$ two-sided confidence interval bounds for $\mathbf{x}'\beta(\tau)$:

$$\mathbf{x}'\hat{\boldsymbol{\beta}}(\tau) \pm z_{1-\alpha/2}\sqrt{\mathbf{x}'\hat{\Omega}_n\mathbf{x}}, \tag{9.16}$$

and the lower $(1-\alpha)100\%$ one-sided confidence interval bound for $\mathbf{x}'\beta(\tau)$:

$$\mathbf{x}'\hat{\boldsymbol{\beta}}(\tau) + z_{\alpha}\sqrt{\mathbf{x}'\hat{\Omega}_n\mathbf{x}}. \tag{9.17}$$

(Confidence bounds for the kth element of $\beta(\tau)$, are obtained by setting the kth element of $\mathbf{x}$ to 1, zeroes elsewhere.) Note that the A-basis and B-basis values for Y at $\mathbf{x}$ are given by the last formula with $\tau = 0.01$ and $\tau = 0.10$, respectively, when $\alpha = 0.05$.

Connection with CVaR Regression

It can be shown that for any real numbers τ and ν,

$$\tau(\nu)^+ - (1-\tau)(\nu)^- = (1-\tau)\left[\frac{1}{1-\tau}(\nu)^+ - \nu\right].$$

If ν is viewed as a random variable, this equality will continue to hold when the expectation of both sides is taken, i.e.

$$\mathsf{E}\left[\tau(\nu)^+ - (1-\tau)(\nu)^-\right] = (1-\tau)\left[\frac{1}{1-\tau}\mathsf{E}(\nu)^+ - \mathsf{E}(\nu)\right].$$

Defining $\bar{\mathbf{b}} = [b_1, \ldots, b_p]'$ and $\bar{\mathbf{x}}_i = [x_{i1}, \ldots, x_{ip}]'$, the empirical version of the above result applied to (9.15) gives,

$$\begin{aligned}
\mathcal{P}^0_\tau(\mathbf{b}) &= n(1-\tau)\left[\frac{1}{n(1-\tau)}\sum_{i=1}^n (y_i - \mathbf{x}_i'\mathbf{b})^+ - \frac{1}{n}\sum_{i=1}^n (y_i - \mathbf{x}_i'\mathbf{b})\right] \\
&= n(1-\tau)\left[\frac{1}{n(1-\tau)}\sum_{i=1}^n (y_i - \bar{\mathbf{x}}_i'\bar{\mathbf{b}} - b_0)^+ - \frac{1}{n}\sum_{i=1}^n (y_i - \bar{\mathbf{x}}_i'\bar{\mathbf{b}} - b_0)\right] \\
&= n(1-\tau)\left[b_0 + \frac{1}{n(1-\tau)}\sum_{i=1}^n (y_i - \bar{\mathbf{x}}_i'\bar{\mathbf{b}} - b_0)^+ - \frac{1}{n}\sum_{i=1}^n (y_i - \bar{\mathbf{x}}_i'\bar{\mathbf{b}})\right].
\end{aligned}$$

Let us denote

$$\mathcal{P}^1_\tau(\bar{\mathbf{b}}) \equiv \left\{\widehat{\mathrm{CVaR}}_\tau\left(Y - \sum_{j=1}^p X_j b_j\right) - \hat{\mathsf{E}}\left(Y - \sum_{j=1}^p X_j b_j\right)\right\},$$

where $\hat{\mathsf{E}}(Y - \sum_{j=1}^p X_j b_j) = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{\mathbf{x}}_i'\bar{\mathbf{b}})$. Then,

$$\begin{aligned}
\mathcal{P}^0_\tau(\mathbf{b}) &= n(1-\tau)\left[b_0 + \frac{1}{n(1-\tau)}\sum_{i=1}^n (y_i - \bar{\mathbf{x}}_i'\bar{\mathbf{b}} - b_0)^+ - \hat{\mathsf{E}}\left(Y - \sum_{j=1}^p X_j b_j\right)\right] \\
&= n(1-\tau)\left[b_0 + \frac{1}{n(1-\tau)}\sum_{i=1}^n (y_i - \bar{\mathbf{x}}_i'\bar{\mathbf{b}} - b_0)^+ \right. \\
&\qquad \left. - \widehat{\mathrm{CVaR}}_\tau\left(Y - \sum_{j=1}^p X_j b_j\right) + \mathcal{P}^1_\tau(\bar{\mathbf{b}})\right].
\end{aligned}$$

Now, since

$$\min_{b_0 \in \mathscr{R}}\left\{b_0 + \frac{1}{n(1-\tau)}\sum_{i=1}^n (y_i - \bar{\mathbf{x}}_i'\bar{\mathbf{b}} - b_0)^+\right\} = \widehat{\mathrm{CVaR}}_\tau\left(Y - \sum_{j=1}^p X_j b_j\right),$$

with minimizer $b_0 = \widehat{\mathrm{VaR}}_\tau(Y - \sum_{j=1}^p X_j b_j)$, the minimizers of (9.15) are

$$\hat{\bar{\boldsymbol{\beta}}}(\tau) \;=\; \left[\hat{\beta}_1(\tau), \ldots, \hat{\beta}_p(\tau)\right]' = \arg\min_{\bar{\mathbf{b}} \in \mathscr{R}^p} \mathcal{P}_\tau^1(\bar{\mathbf{b}}), \tag{9.18}$$

$$\hat{\beta}_0(\tau) = \widehat{\mathrm{VaR}}_\tau \left(Y - \sum_{j=1}^p X_j \hat{\beta}_j(\tau)\right). \tag{9.19}$$

The attained minimum is

$$\mathcal{P}_\tau^0 \left(\hat{\boldsymbol{\beta}}(\tau)\right) = n(1-\tau)\mathcal{P}_\tau^1 \left(\hat{\bar{\boldsymbol{\beta}}}(\tau)\right), \tag{9.20}$$

which can be viewed as a measure of goodness-of-fit. However, we propose that a more natural measure is the normalized quantity

$$\mathcal{P}_\tau^1 \left(\hat{\bar{\boldsymbol{\beta}}}(\tau)\right) = \frac{1}{n(1-\tau)} \mathcal{P}_\tau^0 \left(\hat{\boldsymbol{\beta}}(\tau)\right). \tag{9.21}$$

If $\hat{\varepsilon}(\tau) \equiv Y - \hat{\beta}_0(\tau) - \sum_{j=1}^p X_j \hat{\beta}_j(\tau)$ denotes the fitted residual random variable, then by the translation invariance of CVaR,

$$\begin{aligned}
\mathcal{P}_\tau^1 \left(\hat{\bar{\boldsymbol{\beta}}}(\tau)\right) &= \widehat{\mathrm{CVaR}}_\tau \left(\hat{\beta}_0 + \hat{\varepsilon}(\tau)\right) - \hat{\mathsf{E}} \left(\hat{\beta}_0 + \hat{\varepsilon}(\tau)\right) \\
&= \widehat{\mathrm{CVaR}}_\tau \left(\hat{\beta}_0 + \hat{\varepsilon}(\tau)\right) - \hat{\beta}_0 - \hat{\mathsf{E}} \left(\hat{\varepsilon}(\tau)\right) \\
&= \widehat{\mathrm{CVaR}}_\tau \left(\hat{\varepsilon}(\tau) - \hat{\mathsf{E}}\hat{\varepsilon}(\tau)\right) \\
&= \widehat{\mathrm{CVaR}}_\tau^{\,\Delta} \left(\hat{\varepsilon}(\tau)\right).
\end{aligned} \tag{9.22}$$

In CVaR regression, as discussed in Appendix B, we use the measure of goodness-of-fit:

$$\begin{aligned}
\mathcal{P}_\tau^2 \left(\hat{\bar{\boldsymbol{\beta}}}(\tau)\right) &= \begin{cases} \widehat{\mathrm{CVaR}}_\tau^{\,\Delta} \left(\hat{\varepsilon}(\tau)\right), & \text{if } \tau \geq 0.5, \\ \widehat{\mathrm{CVaR}}_{1-\tau}^{\,\Delta} \left(-\hat{\varepsilon}(\tau)\right), & \text{if } \tau < 0.5. \end{cases} \\
&= \begin{cases} \widehat{\mathrm{CVaR}}_\tau^{\,\Delta} \left(\hat{\varepsilon}(\tau)\right), & \text{if } \tau \geq 0.5, \\ \frac{1-\tau}{\tau} \widehat{\mathrm{CVaR}}_\tau^{\,\Delta} \left(\hat{\varepsilon}(\tau)\right), & \text{if } \tau < 0.5. \end{cases}
\end{aligned} \tag{9.23}$$

In this manner, the emphasis is on adequately accounting for under-estimates of $Q_Y(\tau|\mathbf{x})$, the major source of concern. The minimizer, $\hat{\bar{\boldsymbol{\beta}}}(\tau)$, continues to be the same, but the new minimum (measure of goodness-of-fit) is

$$\mathcal{P}_\tau^2 \left(\hat{\bar{\boldsymbol{\beta}}}(\tau)\right) = \begin{cases} \frac{1}{n(1-\tau)} \mathcal{P}_\tau^0(\hat{\boldsymbol{\beta}}(\tau)), & \text{if } \tau \geq 0.5, \\ \frac{1}{n\tau} \mathcal{P}_\tau^0(\hat{\boldsymbol{\beta}}(\tau)), & \text{if } \tau < 0.5. \end{cases} \tag{9.24}$$

Remark 1. CVaR regression in its most general form doesn't determine an estimator for $\beta_0(\tau)$. The value in (9.19) was chosen so as to agree with the estimate from quantile regression. This is an unbiased estimate as we will now show. From the linear regression model formulation, we know

$$Y = \sum_{j=1}^p X_j \beta_j + \beta_0 + \varepsilon,$$

which by (9.14) can be written as,

$$Y = \sum_{j=1}^p X_j \beta_j(\tau) + \beta_0(\tau) - F_\varepsilon^{-1}(\tau) + \varepsilon.$$

This means that

$$Y - \sum_{j=1}^{p} X_j \beta_j(\tau) = \beta_0(\tau) - F_{\varepsilon}^{-1}(\tau) + \varepsilon \equiv \bar{\varepsilon}(\tau).$$

Applying the same argument used to obtain (9.14) with $Z = \varepsilon$ and $c = \beta_0(\tau) - F_{\varepsilon}^{-1}(\tau)$, gives

$$F_{\bar{\varepsilon}(\tau)}^{-1}(\tau) = \beta_0(\tau) - F_{\varepsilon}^{-1}(\tau) + F_{\varepsilon}^{-1}(\tau) = \beta_0(\tau).$$

Thus any unbiased estimator for $F_{\bar{\varepsilon}(\tau)}^{-1}(\cdot)$ will result in an unbiased $\hat{\beta}_0(\tau)$.

References

[1] "Accelerated Insertion of Materials - Composites (AIM-C): Methodology", Boeing Phantom Works Report Number 2004P0020, V 1.2.0, 12 May 2004.

[2] Artzner P., Delbaen, F., Eber, J., and Heath, D. (1997), "Thinking Coherently", *Risk*, 10, 68-71.

[3] Artzner P., Delbaen, F., Eber, J., and Heath, D. (1999), "Coherent measures of risk", *Mathematical Finance*, 9, 203-228.

[4] Hall, P. and Sheather, S. (1988), "On the distribution of a studentized quantile", *J. Royal Statist. Soc. B*, 50, 381-391.

[5] Koenker, R. and Bassett, G. (1978), "Regression quantiles", *Econometrica*, 46, 33-50.

[6] Koenker, R., and Bassett, G. (1982), "An empirical quantile function for linear models with iid errors", *Journal of the American Statistical Association*, 77, 407-415.

[7] Koenker, R. (1994), "Confidence intervals for quantile regression", *Proceedings of the 5th Prague Symposium on Asymptotic Statistics*, P. Mandl and M. Huskova (eds), Heidelberg: Physica-Verlag.

[8] Pflug, G. (2000), "Some remarks on the Value-at-Risk and the Conditional Value-at-Risk", in *Probabilistic Constrained Optimization: Methodology and Applications*, S. Uryasev (ed), Kluwer Academic Publishers.

[9] Rockafellar, R., and Uryasev, S., (2000), "Optimization of conditional value-at-risk", *Journal of Risk*, 2, 21-41.

[10] Rockafellar R.T. and S. Uryasev. (2002), "Conditional Value-at-Risk for General Loss Distributions", *Journal of Banking and Finance*, 26-7, 1443-1471.

[11] Rockafellar, R.T., Uryasev S., and Zabarankin, M. (2002), "Deviation Measures in Risk Analysis and Optimization.", *Research Report 2002-7*, ISE Dept., University of Florida.

[12] Rockafellar, R.T., Uryasev S., and Zabarankin, M. (2002), "Deviation Measures in Generalized Linear Regression", *Research Report 2002-9*, ISE Dept., University of Florida.

[13] Trindade, A.A., Uryasev S., and Zrazhevsky, G. (2003), "Controlling Risk via Asymmetric Rezidual Error Tail Constraints with an Application to Financial Returns ", *Research Report 2003*, ISE Dept., University of Florida, to appear.