Article from

Health Watch

April 2020

Quantile Regression—A New Actuarial Approach to Claims Estimation

By Jason Reed

Suppose, as a health insurance underwriter, you are confronted with the following two credible groups' experience for the past two years, with all numbers trended to a future rating period as shown in Table 1.

Table 1
Trended Two-year Claim Experience for Group Rating

	Year 1 Experience	Year 2 Experience	Average
Larry's Lumber	$370	$340	$355
Bob's Trucking	$360	$380	$370

How would you set expected claims for each group in that future rating period?

Traditional underwriting might indicate pricing at the group's average experience or a blend of that with a block-level per member per month (PMPM). But suppose you knew the entire distribution of claim costs for each group, which would vary based on the demographic and morbidity profile of each member in the group. Then, in addition to observing actual experience, you would know what percentile of the claims distribution that experience represented, as Table 2 illustrates.

Table 2
Percentile of Claims Distributions per Group

	Year 1 Experience	Percentile of Claims Distribution	Year 2 Experience	Percentile of Claims Distribution	Average
Larry's Lumber	$370	85th	$340	70th	$355
Bob's Trucking	$360	25th	$380	35th	$370

Based on this information, you might reason that Larry's Lumber would continue to regress to its mean with lower than average trend in Year 3, but that Bob's Trucking would likely experience higher than average trend in the future period, as claims regress upward toward the mean of that group's particular claim distribution.

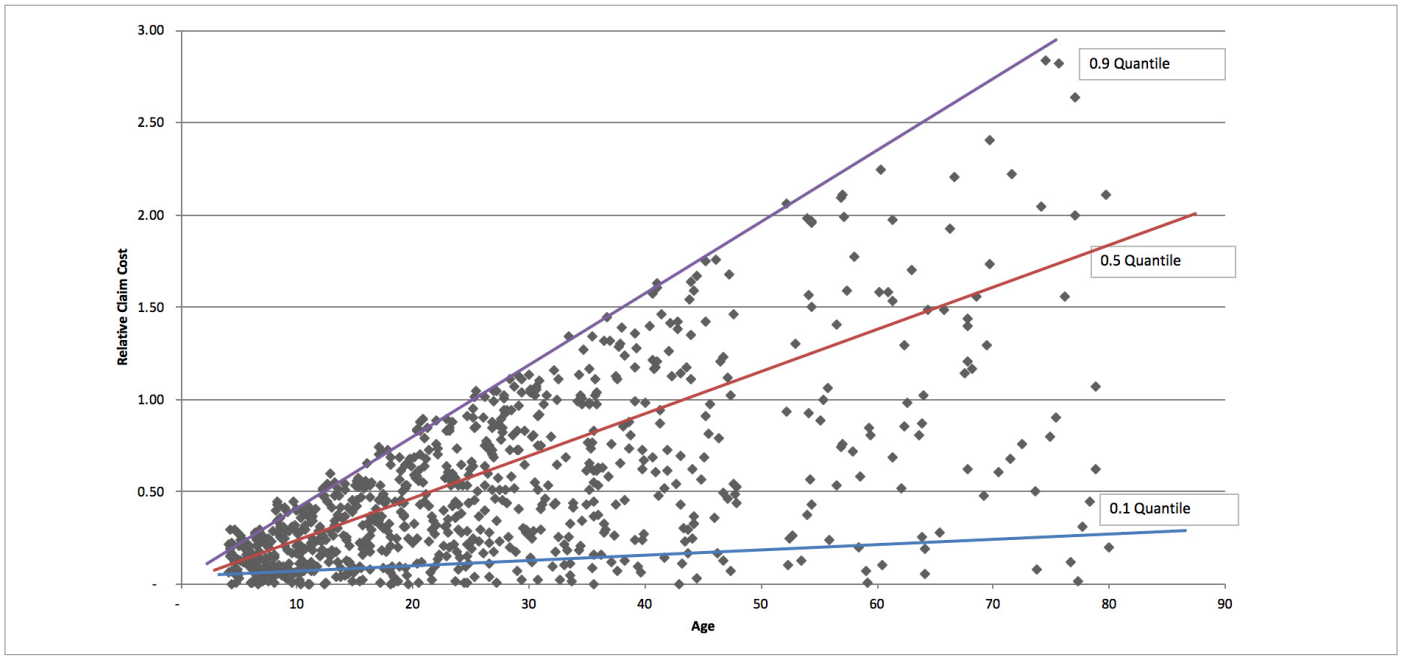
EVOLUTION FROM POINT ESTIMATES TO CONFIDENCE INTERVALS

Much more powerful inferences can be made from historical data when probabilities can be assigned to the range of possible future claims costs. Quantile regression, a statistical technique that involves estimating the percentiles of claims based on demographics and a group's unique morbidity profile, enables estimation with this level of precision and insight.

Like a traditional risk score, quantile regression estimates a linear function of a member's age, gender and morbidity profile. But expanding on the information in a risk score (which estimates only the mean given these predictors), it develops a distinct function at each percentile, so differing impacts of demographics and medical conditions at different parts of the distribution can be observed. For example, females may cost three times as much as males at the 50th percentile of the distribution (due to better compliance with preventive care and maternity costs) but may not have a significant difference in cost at the 98th percentile (since care for many catastrophic diseases may not vary significantly by gender).

Quantile regression adds a new dimension to the underwriting process. Traditionally one would compare normalized experience across similar groups within a block. Quantile regression takes a different perspective, allowing each group's experience to be evaluated relative to the entire distribution of possible claims outcomes *for that group*. So if a group had experience that was at the 20th percentile of expected claims given its morbidity profile, we might anticipate a regression to the mean and, therefore, higher trend than expected in a future period. This allows for

Figure 1
10th, 50th and 90th Quantile Regressions for the Age Predictor



more accurate claims projections and the observation of different patterns of volatility across groups.

As a grossly simplified example, you can see from the following quantile regression equations how the impacts of age, gender and morbidity could change at different percentiles of the claims distribution:

$$y_{25th} = 0.6 * age + 1.2 * gender + 0.4 * morbidity$$
$$y_{50th} = 1.1 * age + 3.2 * gender + 4.8 * morbidity$$
$$y_{98th} = 0.3 * age + 0.5 * gender + 6.7 * morbidity$$

Note how the magnitude of coefficients changes at different percentiles, which would not be possible with traditional risk scores that just estimate the expected value.

$$E(y) = 0.9 * age + 3 * gender + 5.2 * morbidity$$

Quantile regression has historically been disregarded due to computational intractability, but with the development of more efficient algorithms and increased computing power, it can now be used on large blocks of business. The simplified graph in Figure 1 shows a quantile regression on just the predictor of age.

In this one-dimensional case, quantile regression quantifies how the slopes (impact of a change in one year of age) of the best-fit line can change as we move from the 10th to the 50th and 90th percentiles, showing the differing impact of the predictor. Figure 1 suggests that the impact of increasing age is less for healthier members than for the sickest, and the slopes of the lines quantify that difference.

There are often good reasons for using traditional least squares regression and estimating only the expected value. When statistical assumptions are satisfied, the least squares estimates obtained are optimal, and it is solved with a computationally easy matrix equation. Indeed, assuming constant variance, the least squares estimates yield the "best linear unbiased estimate" of the expected value of claims, meaning that among all unbiased linear models that estimate the mean, the least squares estimate has minimum variance.

Unfortunately, the assumptions that are needed for those optimal properties to hold are strict: errors must be uncorrelated and have constant variance at all levels of the predictor. *The constant variance assumption is not satisfied by health claims*—variance at high levels of claim costs is higher than at lower levels, as Figure 1 illustrates for the age predictor.

ADVANTAGES OF QUANTILE REGRESSION

Quantile regression enables a deeper understanding relative to traditional claims estimation techniques.

1. **Quantile regression allows estimation of the whole distribution at different levels of covariates, so you can build confidence intervals.** Variability of claims at different levels of the predictors (for example, males aged 30 with hypertension versus females aged 50 with knee pain) may differ substantially. Quantile regression allows these estimates to differ by estimating different percentiles for each value of the predictors. Correspondingly, confidence intervals with different midpoints and widths can be produced that reflect those differing levels of precision.

Table 3
Claims Distribution Estimates Using Quantile Regression

	Percentile of Claims Distribution—Year 1				
	10th	25th	50th	75th	90th
Larry's Lumber	$337	$346	$352	$360	$375
Bob's Trucking	$345	$360	$366	$371	$377

Returning to our example with Bob's trucking, suppose our quantile regression estimates produced the distributions shown in Table 3.

For Larry's Lumber, a 50 percent confidence interval is ($346, $360) PMPM with a width of $14 PMPM. If we increase our confidence to 80 percent, the interval would be ($337, $375) with a width of $38 PMPM.

Conversely, the variability of claims in Bob's Trucking at 50 percent confidence is ($360, $371) with a width of $11 PMPM, and the 80 percent confidence interval is ($345, $377) with a width of $32. We observe that Bob's risk is lower, even though the median cost is higher.

2. **Quantile regression gives a distinct set of parameter estimates at different percentiles, so you know where they are significant and when they differ.** Using quantile regression estimates, the impact of covariates at different percentiles can be contrasted. Women may cost 3 times as much as men at the 50th percentile of claims but only 1.1 times as much at the 98th percentile. Using ordinary least squares, which has a fixed gender coefficient, would not allow the reflection of this variability.

Another application could be provider performance evaluation. For example, for each cardiologist you could determine the 10th, 20th, … 90th percentiles of cost corresponding to the panel of members attributed to that doctor, then measure the actual costs incurred by that physician against that distribution.

3. **Quantile regression enables us to observe at what parts of the distribution predictors are significant.** Using the wrong level of coefficients, or pricing with coefficients that are not really significant, can lead to mispricing across a block and anti-selection. It is critical to understand what part of the distribution is being priced to.

There are also disadvantages to the quantile regression approach. For estimating the mean of a group, traditional risk scores will be more accurate when the least squares assumptions are satisfied. Also, it is difficult to calculate these estimators correctly: a substantial amount of data is necessary to estimate many more parameters, and it must be verified that the algorithm found the "right" answer. For example, naive solutions in statistical packages sometimes lead to negative contribution to risk from some conditions. Finally, using quantile regression introduces ambiguity—it can be difficult to know whether an observed outlier at, say, the 90th percentile of the claims distribution implies that experience will regress to the mean or that a true shift in the distribution has happened.

Risk scores are a canonical example where quantile regression can be used to provide refinements to actuarial projections. Traditionally, actuarial risk scores compute the expected average contribution to future costs from age, gender and a set of conditions. So, for example, if a 55-year-old male member has stomach cancer, his risk score is a sum of demographics and conditions, such as 0.85 + 4.2 = 5.15. If instead we generated risk scores using quantile regression, we might see expected future costs as in Table 4, greatly enhancing our understanding of claim variability for this condition.

Table 4
Risk Scores Generated by Quantile Regression

Percentile	Demographic	Conditions	Total
25th	0.70	3.60	4.30
50th	0.85	4.75	5.60
75th	1.07	7.34	8.41

We could then compare the expected distribution of costs of groups of members with the given condition to the actual treatment costs by a panel of physicians to understand resource use deviating significantly from the average. Note that in practice an interaction term may be appropriate between age/gender and the condition at some percentiles of the distribution.

APPLICATIONS IN UNDERWRITING

Going back to our underwriting story, when broadly applied across a book of business, quantile regression can facilitate discount guidance to underwriters as well as improve operational efficiencies by identifying which groups are most likely to benefit from an underwriter's comprehensive review and which groups are straightforward and do not merit any additional attention. However, unlike traditional underwriting approaches that often focus on high dollar claims (which in many cases have resolved themselves), the focus is on groups where there are significant deviations, up or down, from expected.

The increase in computing power and development of efficient algorithms has opened up new frontiers for estimating complex probability distributions, allowing for more comprehensive and creative actuarial analysis that allows us to understand the true drivers of risk. Linear regression and its cousins allow for efficient estimation of the mean under some circumstances, but the flexibility, accuracy and empirical value of knowing the entire distribution at each level of a set of regression predictors argues strongly for a new statistical approach to actuarial claims estimation. ■

Jason Reed, FSA, MAAA, is a senior director, payer consulting, at Optum. He can be reached at *r.jason.reed@optum.com.*