

CHRP Seminar

Combining Quantile Regression & Cubic Splines

Allan Garland, M.D., M.A.
Associate Professor of Medicine
Director, Medical Intensive Care Unit

3 Goals

1. To briefly review the techniques of quantile regression and restricted cubic splines
2. To demonstrate their combined value in HSR
3. To get feedback from the audience, regarding whether #2 is sufficiently interesting to warrant writing a "methods" paper on the topic

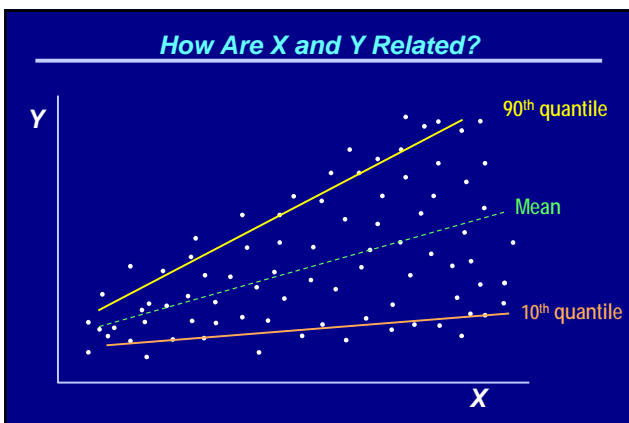
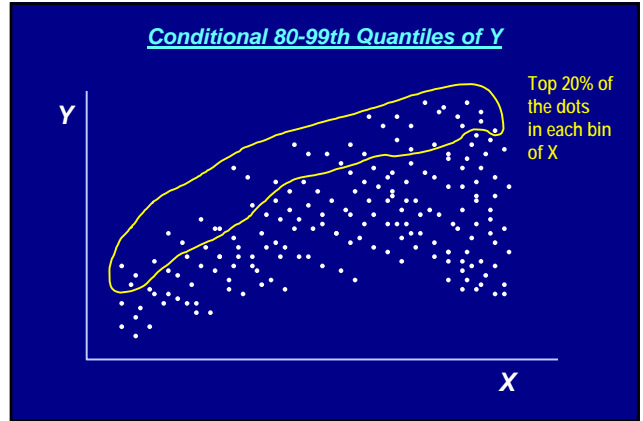
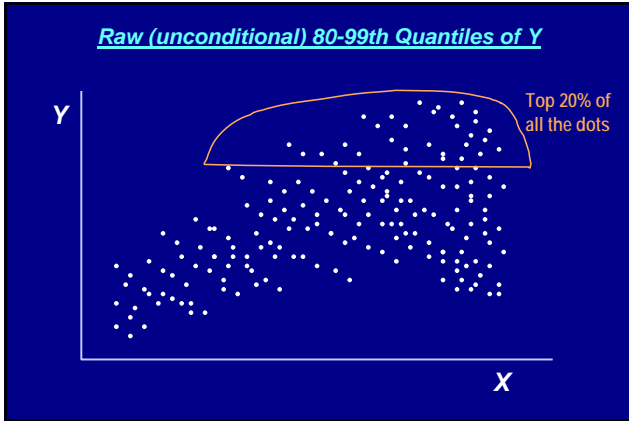
Introduction

- Much of HSR involves investigating relationships between metric parameters/variables (e.g. effect of age on QOL)
- Often done via linear regression
 - assesses how X relates to the *mean* of Y
- Linear regression is often performed without analysis of whether the relationship between X & Y is linear
- Biologically, there are no reasons to expect that:
 - Y relates linearly to X
 - the relationship of X to the mean of Y captures the full nature of the relationship between X and Y

Quantile Regression

(*Front Ecol Environ* 1(8):412-420, 2003)

- A numerical method that allows evaluation of the relationship of X across the range of Y
- It separately models the conditional quantiles of Y on X
 - not to be confused with the unconditional quantiles of Y



Quantile vs. OLS Regression

- QR makes no assumptions about about the shape or distribution of residuals
 - ∴ it is impervious to heteroscedasticity -- indeed such heterogeneous variance can be a result of having quantitatively different relationships between X & Y for different quantiles of Y
- Because it deals with quantiles, QR is much more robust to outliers
- Instead of minimizing $SSR = \sum_i \{Y_i - (\beta_0 + \beta_1 \cdot X_i)\}^2$ it minimizes a weighted version of $\sum_i \{Y_i - (\beta_0 + \beta_1 \cdot X_i)\}$ (weighted such that the total weight above the best-fit line is the same as that below it)

Final Points About QR

- Mathematically there is higher power of finding a statistically significant effect of X on Y for an extreme quantile than a more central one
 - the greater differences from zero for the parameter estimate usually present at extremes offsets the greater sampling variation associated with more extreme quantiles
 - i.e. you can miss an effect if you look at the mean or median when in fact there is only a relationship between X and Y for more extreme values of Y
- If X & Y have the same functional relationship over the range of Y, then QR has no advantage over mean (OLS) regression

Dealing With Nonlinear Relationships

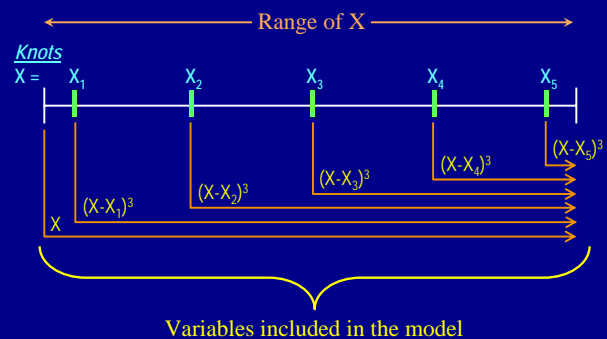
- Sometimes data transformation of X &/or Y will linearize their relationship -- but that relationship is often unknown
- Including polynomial terms (X^2 , X^3 , etc) in the model can help, but will not fit some functional forms well (e.g. logarithms), and high order polynomials behave poorly over large intervals
- Cubic splines is a technique of fitting 3rd order polynomials piecewise over the range of X
 - ensure that the 1st and 2nd derivatives are matched at the knots (joining points)
 - can smoothly fit a virtually arbitrary range of functional forms
 - the piecewise nature obviates the poor behavior of polynomials over long distances

Restricted Cubic Splines

(Harrell. Regression Modeling Strategies 2001, Springer)

- Divide the range of X into intervals with k knots
- For each knot, include a new variable that is a cubic polynomial in X that applies to all values of X above $X_{knot\#j}$
- Fit this group of variables to Y, constrained so that the function is linear in X beyond the 1st and last knots (such *restricted* cubic splines behave better in the tails; have k-2 new variables for k knots)
 - k is more important to the fit than is the location of knots
 - k=5 usu. gives good fit -- use fewer knots for small data sets
 - location based on percentiles of the range of X

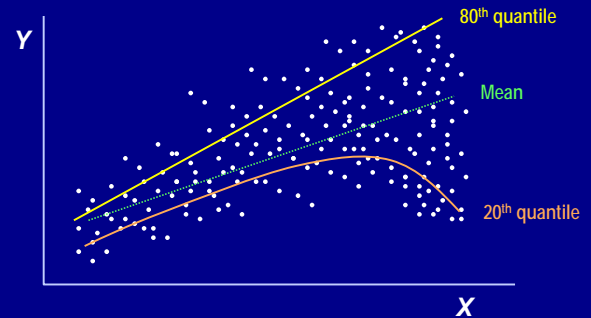
Restricted Cubic Splines - Schematic



Using Cubic Splines in Regression

- Create the (k-2) restricted cubic spline variables (S_1, \dots, S_{k-2})
(\exists functions in SAS and Stata)
- Include X itself, and its cubic spline terms in the model
 - $Y = \beta_0 + \beta_1 X + \beta_{S1} S_1 + \dots + \beta_{Sk} S_{k-2} + \text{all the other vars}$
- Assess whether X is a relevant predictor for Y by a chunk test (df=k-1) of: $\beta_1 = \beta_{S1} = \dots = \beta_{Sk-2} = 0$
- If X is a relevant predictor, assess whether X is linearly related to Y by a chunk test (df=k-2) of: $\beta_{S1} = \dots = \beta_{Sk-2} = 0$

Using Quantile Regression with Splines

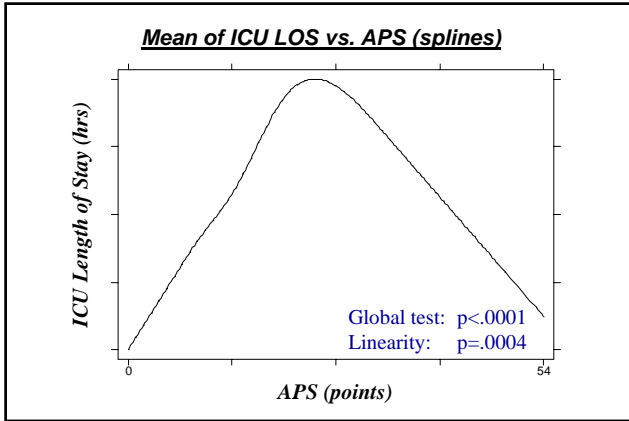


Example 1: MICU Data

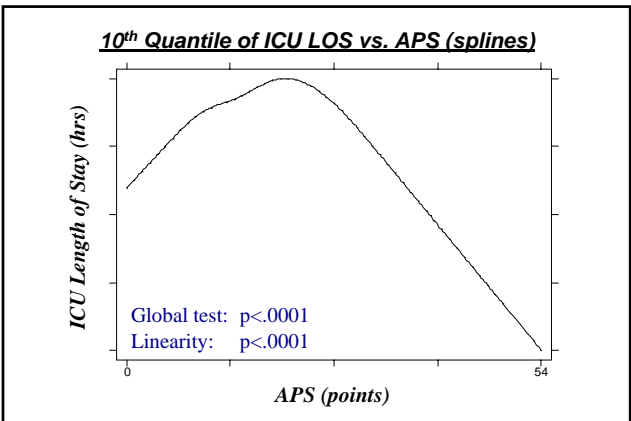
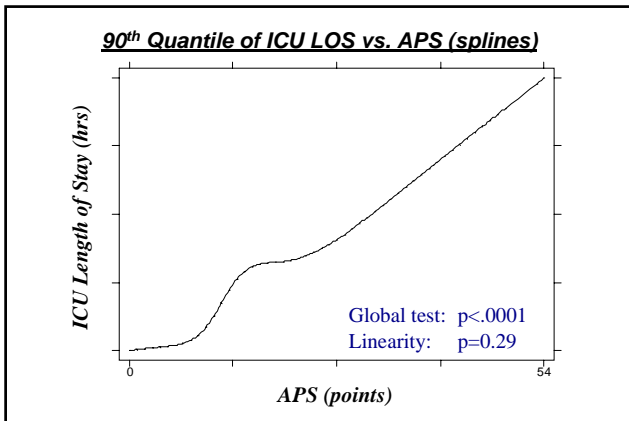
- Goal: Identify relationship of APACHE II acute physiology score (APS) to ICU LOS (hrs) using QR with cubic splines
- Covariates:
 - age, gender, race (minority vs. not), insurance (private, none, medicare, medicaid, medicare + medicaid), source of ICU admission (ED, floor, other ICU, outside hospital, others), \pm ICU readmission, acute admission diagnostic group (resp, CV, GI, neuro, misc. medical, surgical/trauma), \pm intubated, # chronic comorbid conditions (Elixhauser)
- Patients: N=2398 admitted 2004-2006 to MICU at Metro
ICU LOS: Mean=82 \pm 99 hrs Median=49 hrs (IQR=29-88)
APS: Mean=15.7 \pm 8.6 Median=14 (IQR=9-20)

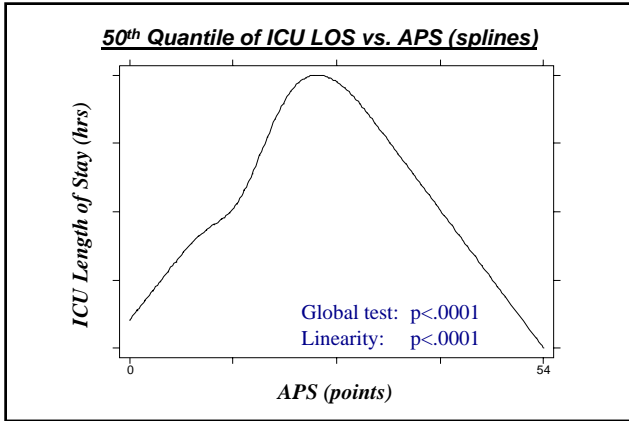
Result: OLS (Mean) Model

- Using APS linearly (robust SE)
 - $R^2=0.18$, $p<.0001$
 - significant predictors: ICU source, diagnostic category, comorbidity, intubated, **APS**
 - APS: $\beta=1.15$, $p=.001$
(69 mins longer ICU stay for each rise of 1 in the APS)
- Entering APS into the model as cubic splines
 - 5 knot cubic spline of APS, robust SE
 - $R^2=0.19$, model $p<.0001$
 - Linearity test: $p=0.0004$ (i.e. the relationship between mean length-of-stay and severity of illness is nonlinear)



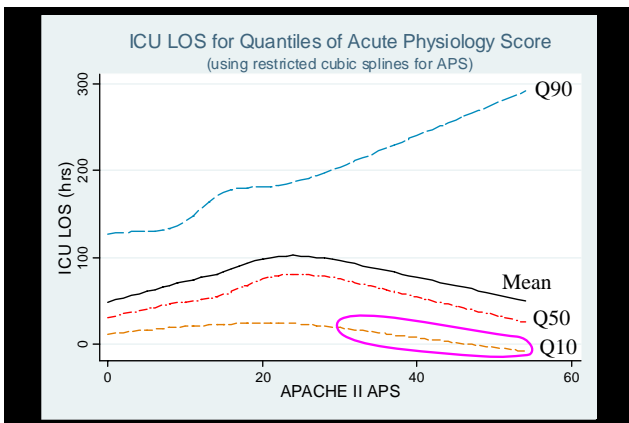
Now:
Quantile Regression
Using
Cubic Splines for APS
(600 reps)





Results: Comparing QR Models

- Comparing the influence of the 2 continuous (splined) variables between the 10th and 90th quantiles
 - use `-iqreg-` in Stata to generate the differences in all the β 's between the 2 regressions \Rightarrow
 - do a Wald test that all the difference terms for the spline variables of APS are equal to zero (i.e. $\Delta\beta_1 = \Delta\beta_{S1} = \dots = \Delta\beta_{Sk-2} = 0$)
- Global test of whether the influence of APS on ICU LOS is different between 10th vs. 90th quantiles: **Yes** ($p = .0005$)



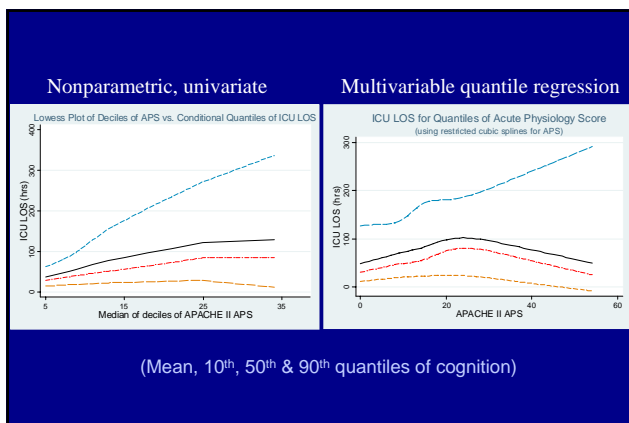
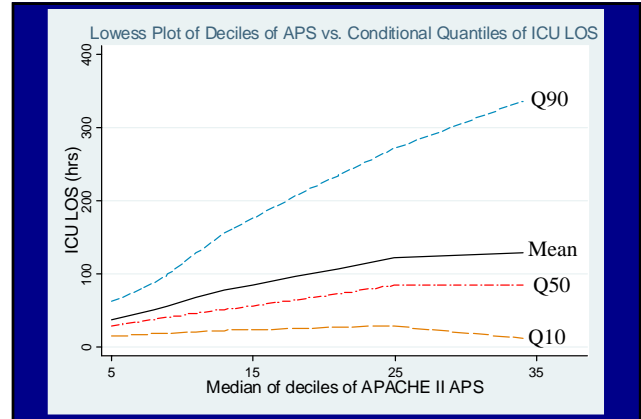
Use this Info to Understand What's Going On

- Note what happens among those in the lower quartiles of LOS --- i.e. LOS first rises with increasing severity of illness, then it peaks and declines with further increase in severity of illness.
- Is it because these people come in and die quickly?
- YES, ICU mortality rate of the 34 patients with ICULOS<30 hrs and APS>30 is 94.1%, vs. 8.5% for all patients

Is There an Easy Way to See if We Need QR?

Idea: Use the type of simple, graphical test favored by Hosmer & Lemeshow for examining linearity in logistic regression

- Divide the X-axis into deciles \Rightarrow
- For each decile of X, calculate the 90th percentile of Y \Rightarrow
- Collect the 90th percentile of Y for each of the 10 bins of X \Rightarrow
- Plot the median of each bin of X vs. the 90th conditional percentile of Y for that bin of X \Rightarrow
- Smooth the curve with Lowess \Rightarrow
- Do the same thing for any quantile of Y desired, or the mean

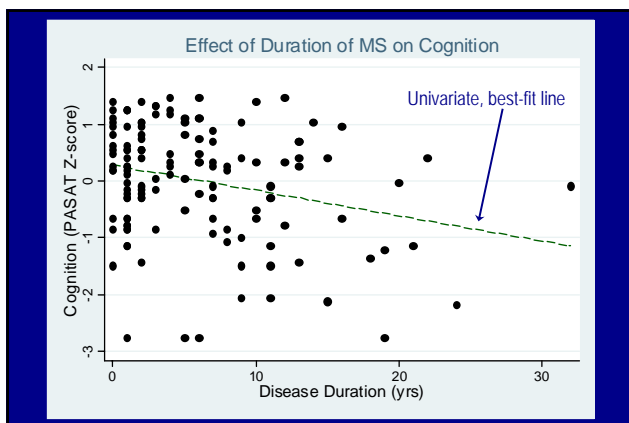
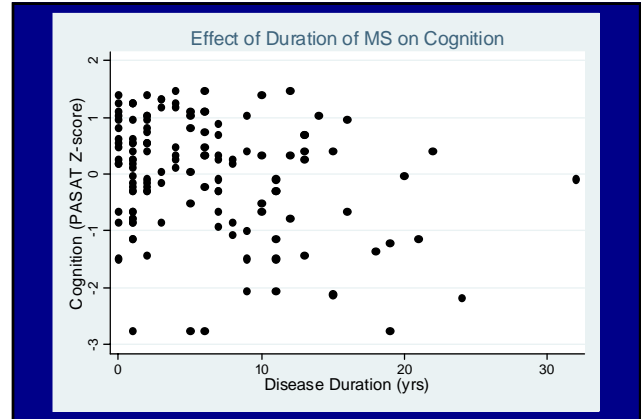


Interpretation of This Example

- The relationship between acute severity of illness and ICU length-of-stay is complex and nonlinear
- Furthermore, there is not a single such relationship
 - qualitative nature of the relationship varies across the range of Y
 - the mean or median do not tell the story accurately
- Correctly identifying these complex relationships can lead to added insights about patient subgroups
- A simple, univariate, nonparametric graph gives a good sense of the differential quantile effects

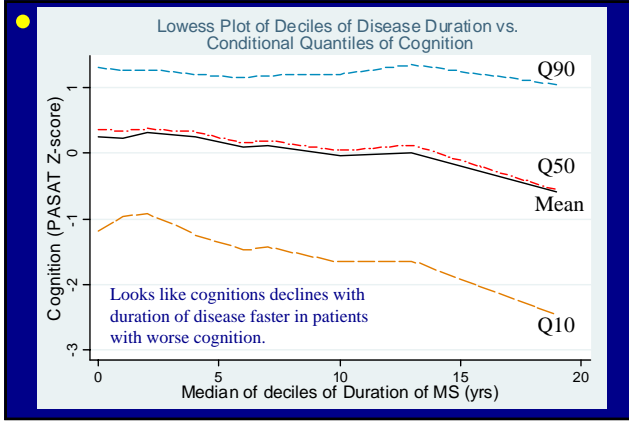
Example 2: Multiple Sclerosis Data

- Goal: Identify relationship between the duration of MS and cognitive function, using QR with cubic splines
- Cognitive function = Z-score for the Paced Auditory Serial Addition Test (lower [more negative] is worse)
- Covariates:
 - age, gender, race (minority vs. not), years of education, \pm married, \pm has children, \pm employed
 - \pm depressed (Beck Depression Inventory ≥ 13 vs. < 13)
 - type of disease course (2° progressive vs. relapsing-remitting)
- Patients: X-sectional study of 123 CCF outpatients studied 2001-2003 --- Mean duration=5.9 \pm 6.0 yrs Median=4 yrs (IQR = 1-9)



Result: OLS (Mean) Model

- Using Disease Duration linearly (robust SE)
 - $R^2=0.17$, model $p=0.002$
 - significant predictors: disease course, **disease duration**
 - Disease duration: $\beta = -0.037$, $p=.015$
(0.37 SD lower cognitive function for each decade of disease)
- Mean model with cubic splines
 - OLS model (robust SE) using 4 knot cubic spline of Disease duration
 - $R^2=0.18$, model $p=0.005$
 - Linearity test: $p=0.56$ (i.e. no evidence the relationship is nonlinear)

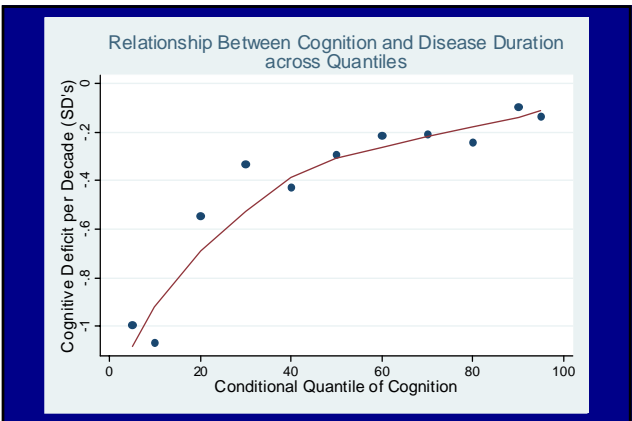
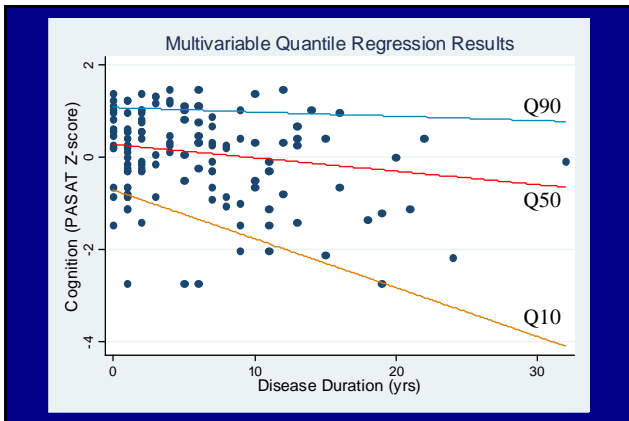


Quantile Regression Models Using Cubic Splines

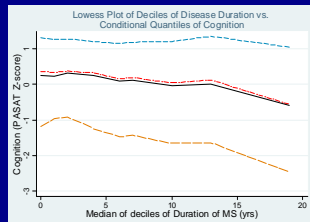
- Linearity checking: No QR model had significant nonlinearity \Rightarrow make QR models with linear term only (bootstrapped SE, 600 reps)

Quantile	Coefficient (per decade)	p-value
5	-1.00	.01
10	-1.07	.009
20	-0.544	.12
30	-0.331	.26
40	-0.425	.07
50	-0.290	.18
60	-0.214	.27
70	-0.207	.21
80	-0.242	.11
90	-0.094	.57
95	-0.133	.49
Mean	-0.370	.017

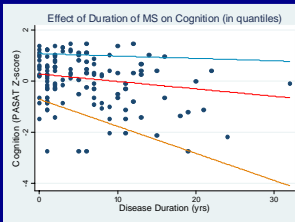
The coefficients indicate by how many SD's cognition is lower for each decade of disease.



Nonparametric, univariate



Multivariable quantile regression



(10th, 50th and 90th quantiles of cognition)

Interpretation of This Example

- In this small study, patients with longer duration of MS have lower cognition; this effect is linearly related to disease duration.
- The influence of disease duration on cognition was greater for those with worse cognition.
- Assessing the influence of disease duration on the mean patient substantially underestimates its influence on those with the worst cognition, and overestimates its influence on those with relatively intact cognition
- Again, the simple, univariate, nonparametric graph gives an good sense of these quantile effects.

Overall Conclusions

- Using cubic splines allows us to screen for nonlinear relationships
- Using quantile regression allows us to assess whether there is a uniform relationship between X and Y
- Use of the simple graphical technique appears to be a useful way to screen for the need to use QR

QUERY

Is this topic, with example, worthy of a "methods paper"?