Quantile Regression for Capital Acquisition Cost Estimating

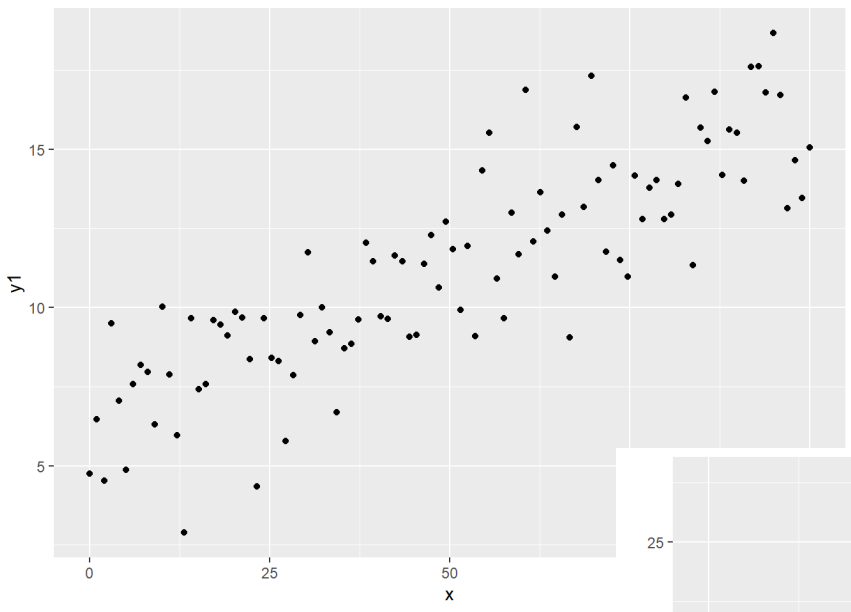
Zachary Matheson, Jeff Beck,
Charles Loelius, Greg Stamp

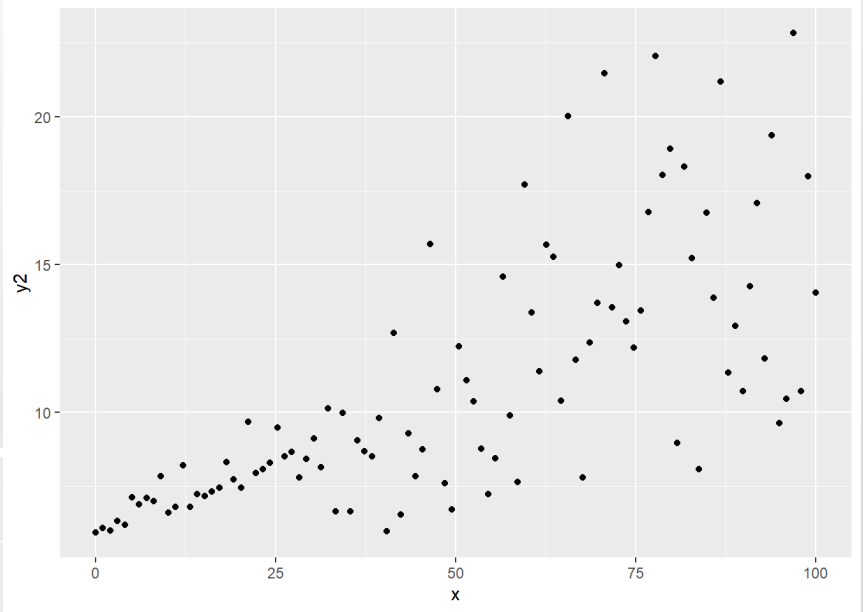
National Nuclear Security Administration

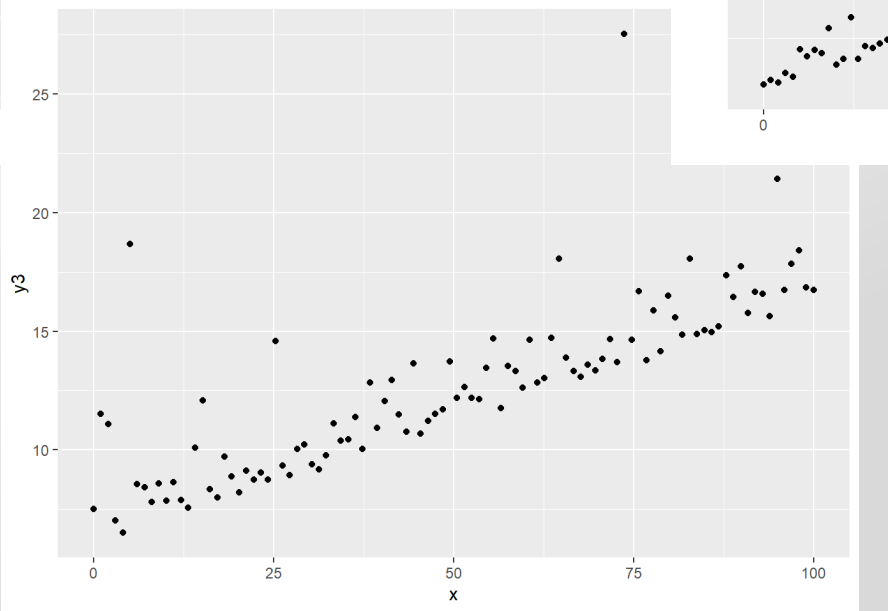
Why quantile regression?

- PA&E's customers typically ask for estimates at the 70% or 85% confidence level
- How to obtain an estimate at the 85th percentile?
  - OLS regression → prediction → prediction interval
    - Assumes normally-distributed error term
  - OLS regression → many predictions with inputs selected from a distribution → S-curve
    - Assumes some distribution for the inputs (often triangular or normal)
- Is there a distribution-agnostic (purely data-driven) way to generate prediction intervals/s-curves?

Why quantile regression?

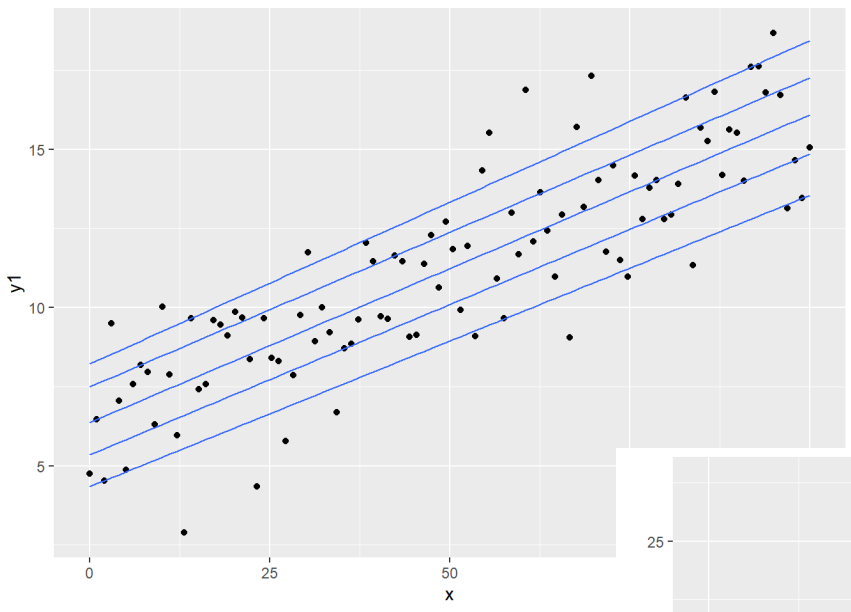


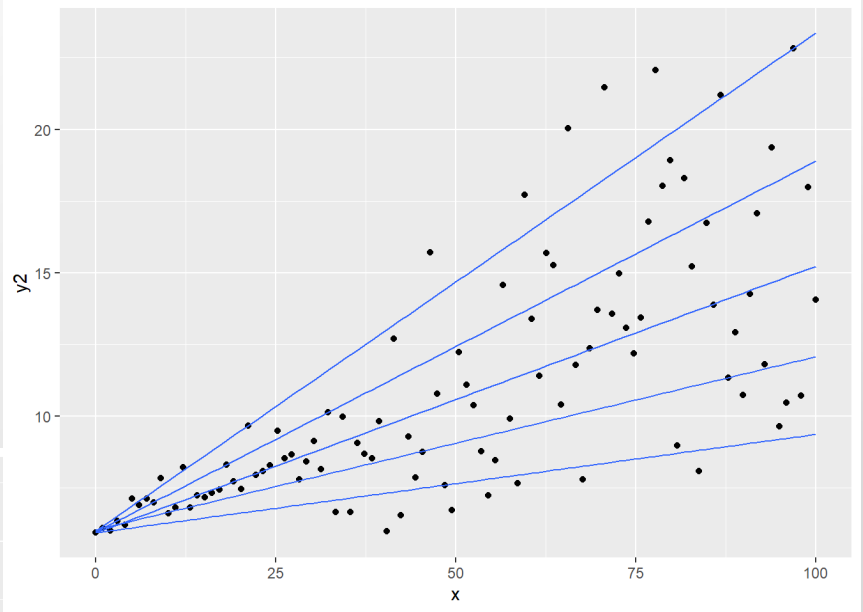


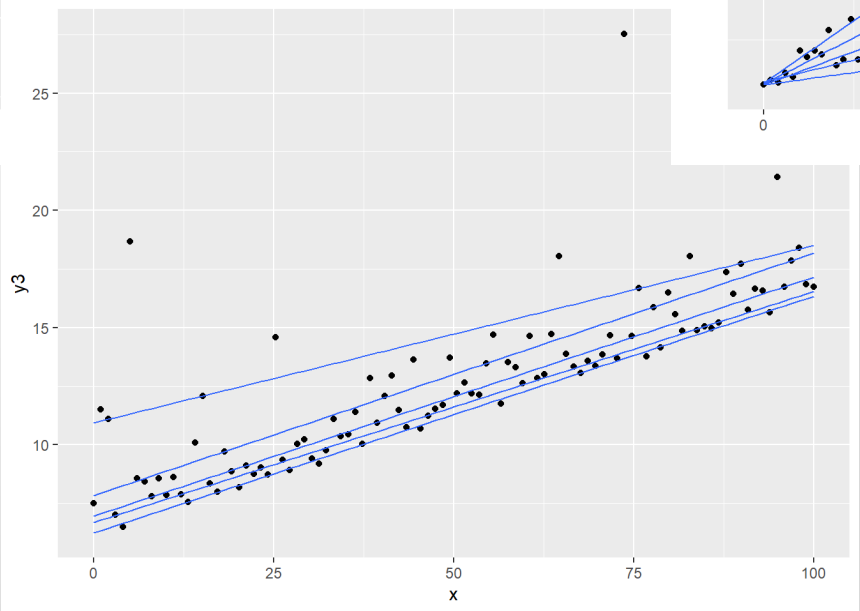


National Nuclear Security Administration

Why quantile regression?







National Nuclear Security Administration

Properties of quantile regression

Ordinary Least Squares	Linear Quantile Regression
$E(y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, i = 1, \ldots, n$	$Q_\tau(y_i) = \beta_0(\tau) + \beta_1(\tau) x_{i1} + \ldots + \beta_p(\tau) x_{ip}, i = 1, \ldots, n$
$MSE = \min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j(\tau)\right)^2$	$MAD = \min_{\beta_0(\tau),\ldots,\beta_p(\tau)} \sum_{i=1}^{n} \rho_\tau \left(y_i - \beta_0(\tau) - \sum_{j=1}^{p} x_{ij}\beta_j(\tau)\right)$
	where $\rho_\tau(r) = \tau \max(r, 0) + (1 - \tau) \max(-r, 0)$
Predicts conditional mean $E(Y\|X)$	Predicts conditional quantiles $Q_\tau(Y\|X)$
Applies when n is small	Needs sufficient data
Assumes normally-dist. errors	Distribution agnostic
Does not preserve $E(Y\|X)$ under transformation	Preserves $Q_\tau(Y\|X)$ under transformation
Sensitive to outliers	Robust to outliers
Computationally inexpensive	Computationally intensive

Properties of quantile regression

Ordinary Least Squares	Linear Quantile Regression
$E(y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, i = 1, \ldots, n$	$Q_\tau(y_i) = \beta_0(\tau) + \beta_1(\tau) x_{i1} + \ldots + \beta_p(\tau) x_{ip}, i = 1, \ldots, n$
$MSE = \min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j(\tau) \right)^2$	$MAD = \min_{\beta_0(\tau),\ldots,\beta_p(\tau)} \sum_{i=1}^{n} \rho_\tau \left(y_i - \beta_0(\tau) - \sum_{j=1}^{p} x_{ij}\beta_j(\tau) \right)$
	where $\rho_\tau(r) = \tau \max(r, 0) + (1 - \tau) \max(-r, 0)$
Predicts conditional mean $E(Y\|X)$	Predicts conditional quantiles $Q_\tau(Y\|X)$
Applies when n is small	Needs sufficient data
Assumes normally-dist. errors	Distribution agnostic
Does not preserve $E(Y\|X)$ under transformation	Preserves $Q_\tau(Y\|X)$ under transformation
Sensitive to outliers	Robust to outliers
Computationally inexpensive	Computationally intensive

Source: “Five things you should know about quantile regression” Rodiguez, R. and Yao, Y., SAS Institute Inc., https://support.sas.com/resources/papers/proceedings17/SAS0525-2017.pdf

Project objectives

- Assess the utility of quantile regression methods on MB-90 datasets to answer key questions:
  - Do we have enough data?
  - How to protect against influential points?
  - Are there pitfalls to using QR?

How much data is enough? Attempt #1

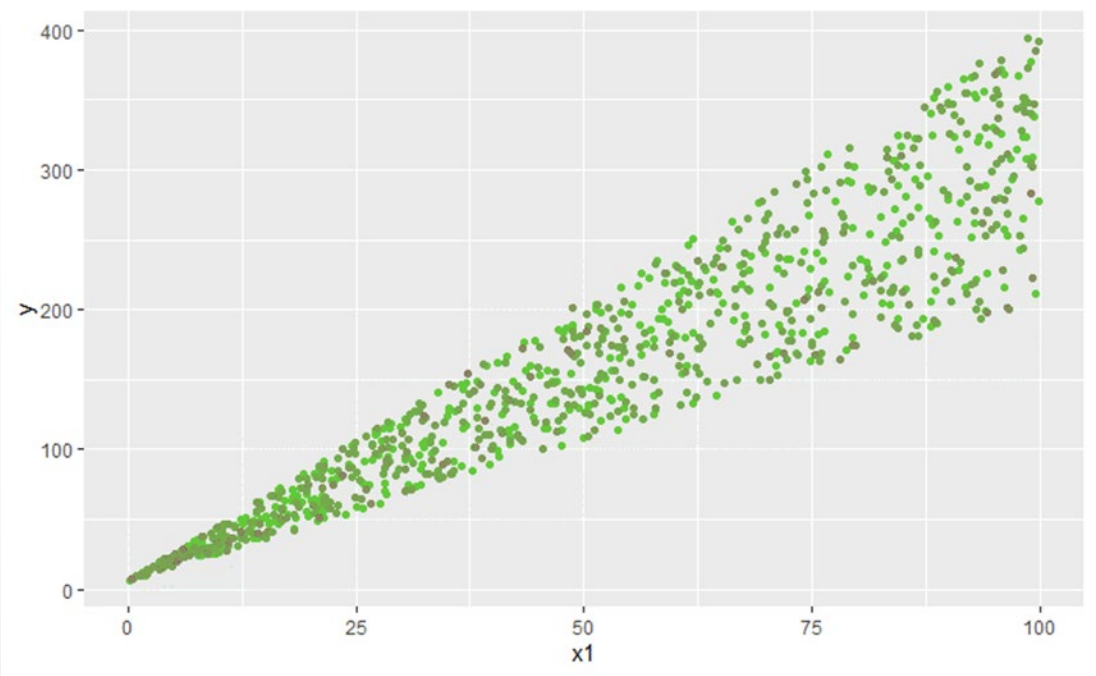
- To resolve the 60th and the 70th percentiles (a difference of 10%) you'd need a total of $^{100\%}/_{10\%} = 10$ datapoints **at minimum**
- Likewise, to differentiate between the 60th and 65th percentiles, you'd need $^{100\%}/_{5\%} = 20$ points **at minimum**
- Now we have a lower bound. Let's see if we can be a little more specific…

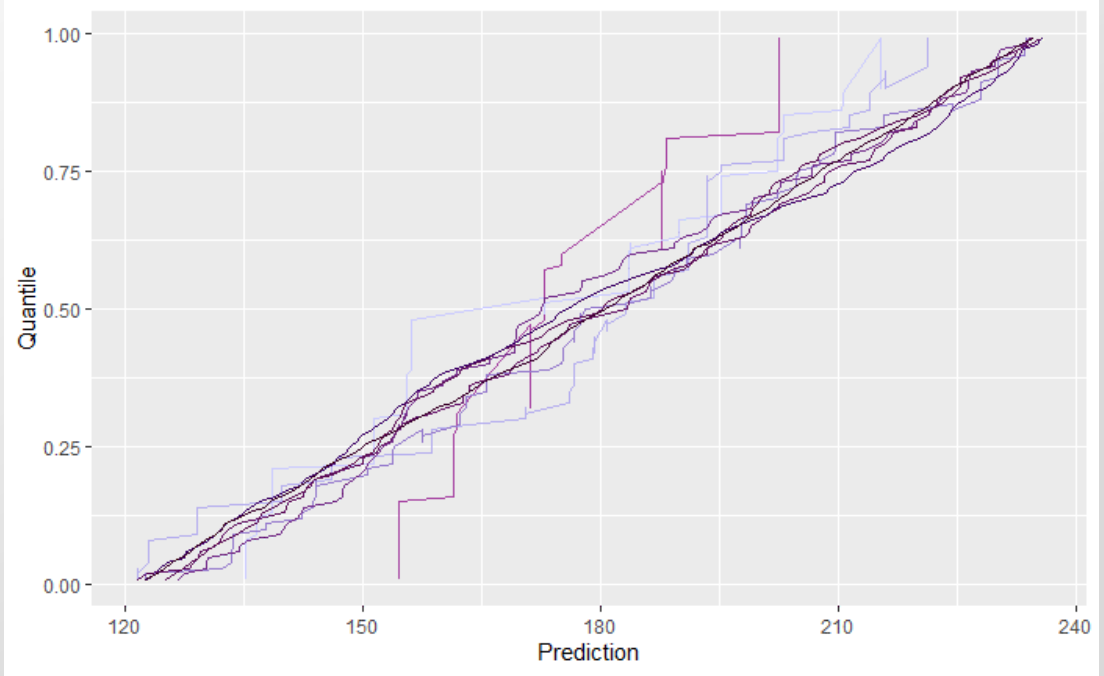
How much data is enough? Attempt #2

- Measured convergence by comparing theoretical vs. empirical CDFs as a function of the number of points used in the regression
  - Uniform, Triangular, Normal, and Log-normal Distributions
- Represented visually on the following slides

How much data is enough? Attempt #2

Uniformly distributed error: $y = mx + b + \epsilon_{unif} \cdot x$





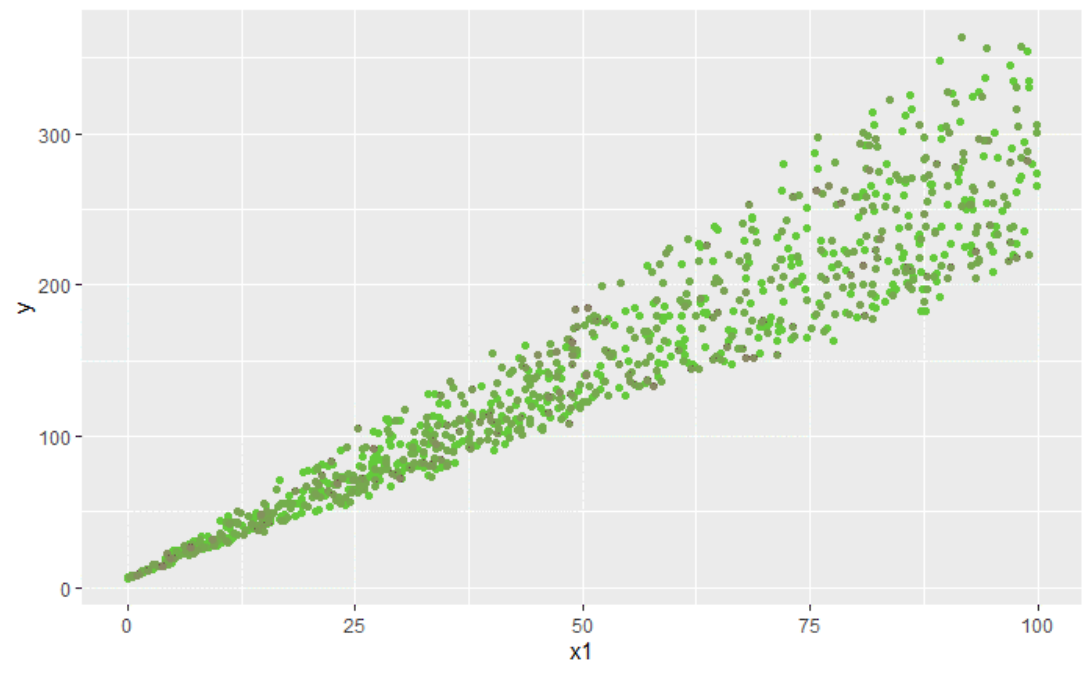
Few data — Many data

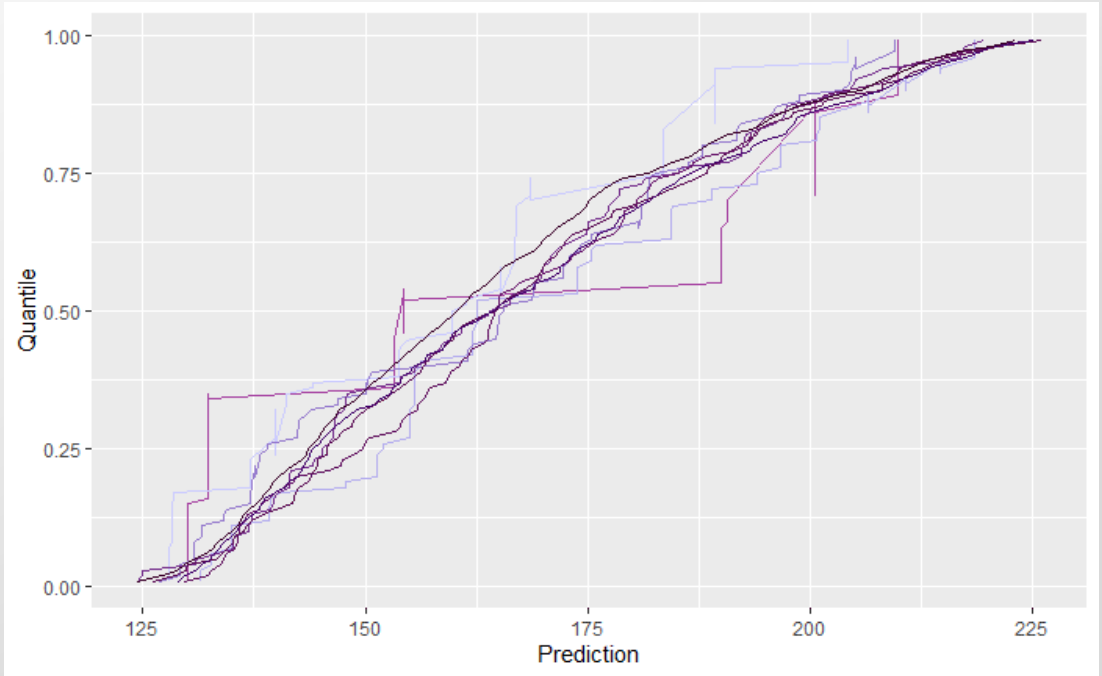
Few data — Many data

National Nuclear Security Administration

How much data is enough? Attempt #2

Triangular distributed error: $y = mx + b + \epsilon_{tri} \cdot x$





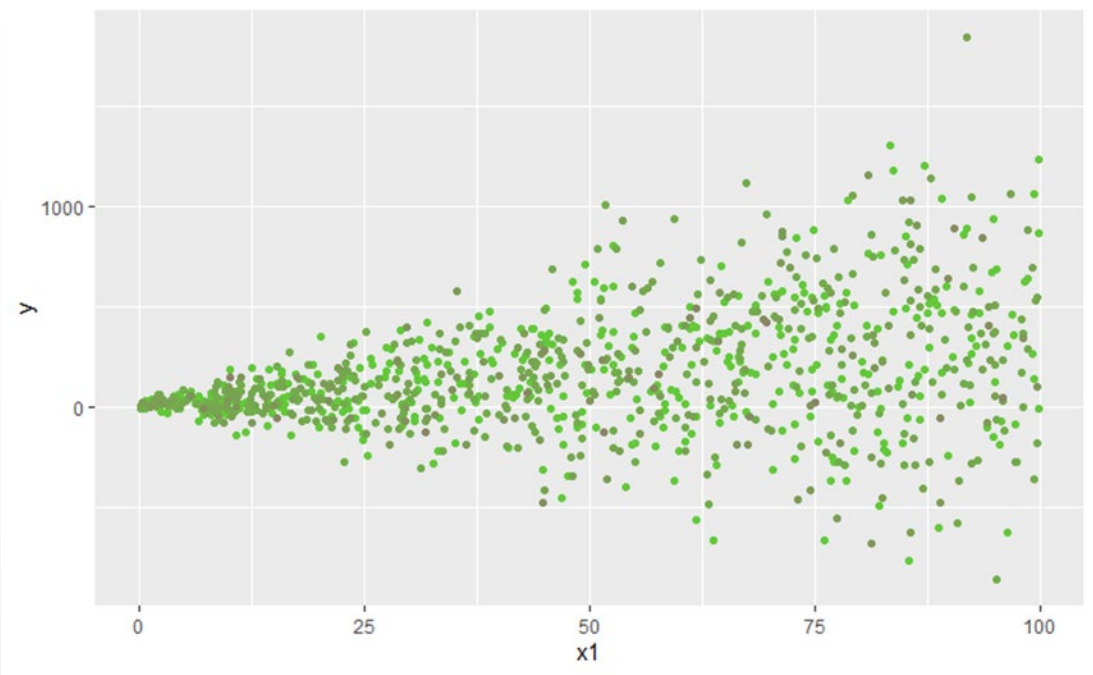
Few data — Many data

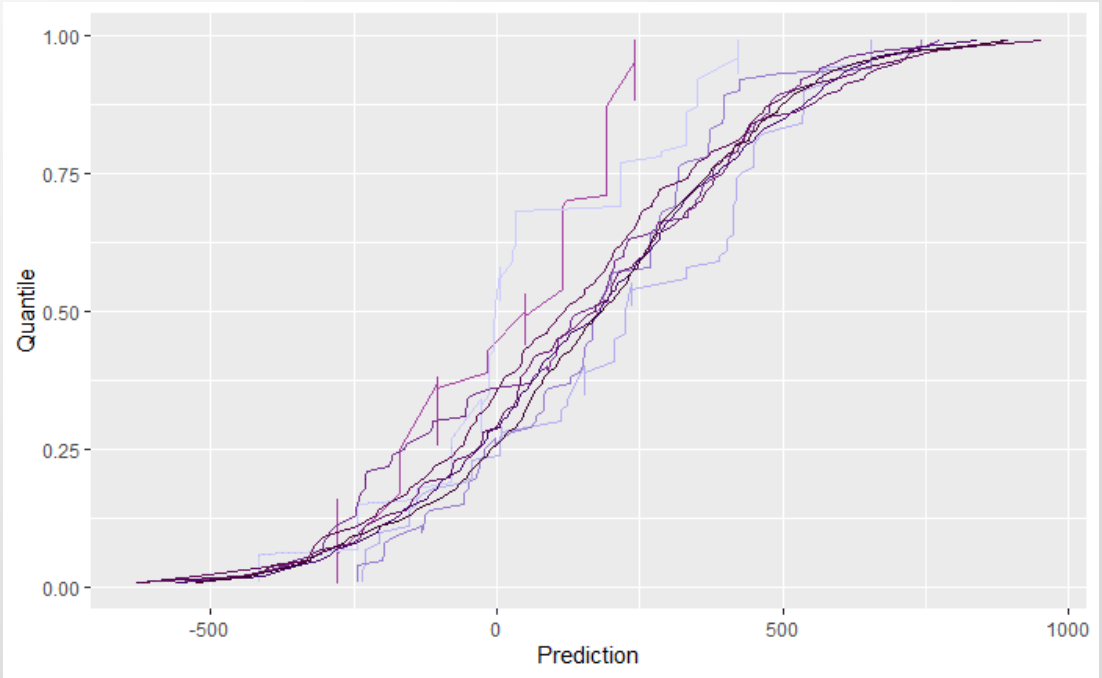
Few data — Many data

National Nuclear Security Administration

How much data is enough? Attempt #2

Normally distributed error: $y = mx + b + \epsilon_{norm} \cdot x$





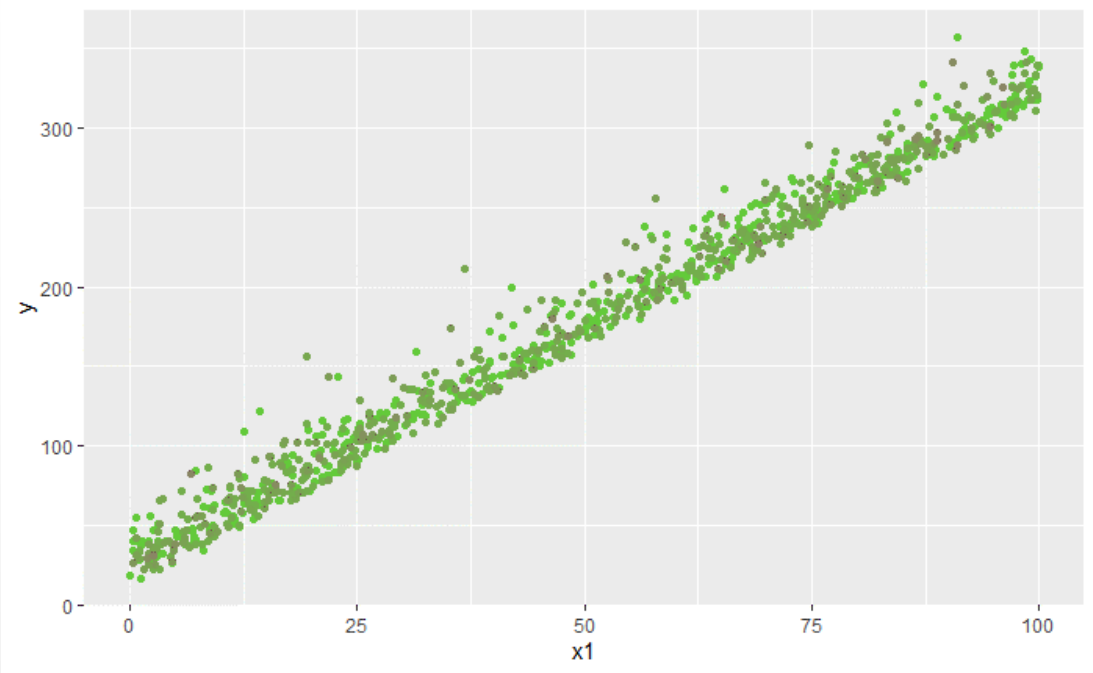
Few data — Many data

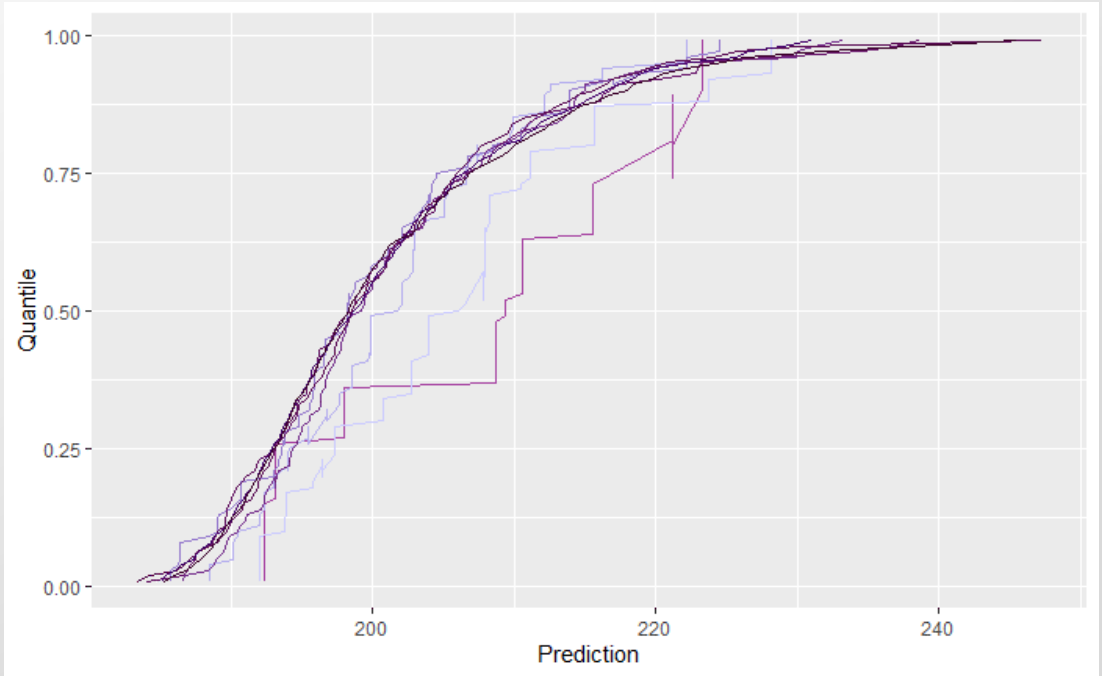
Few data — Many data

National Nuclear Security Administration

How much data is enough? Attempt #2

Log-normally distributed error: $y = mx + b + \epsilon_{lnorm}$





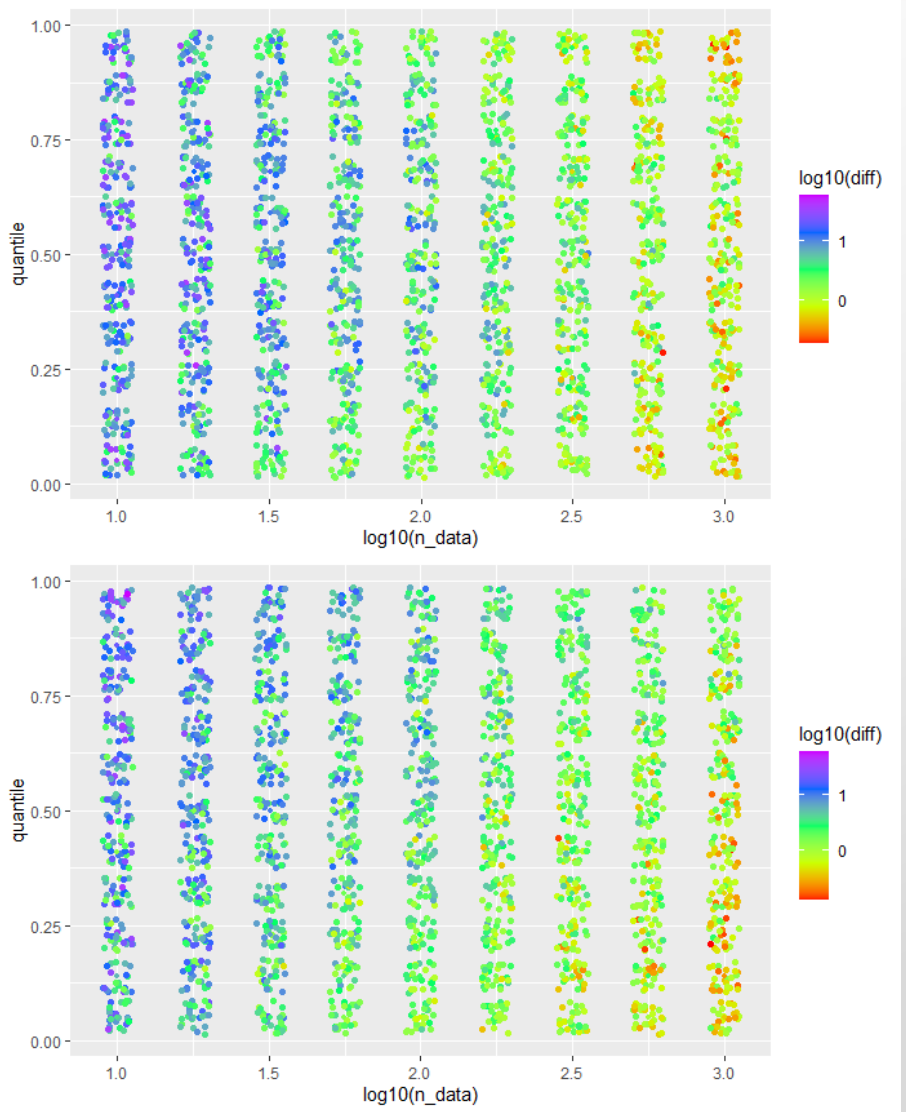
Few data — Many data

Few data — Many data

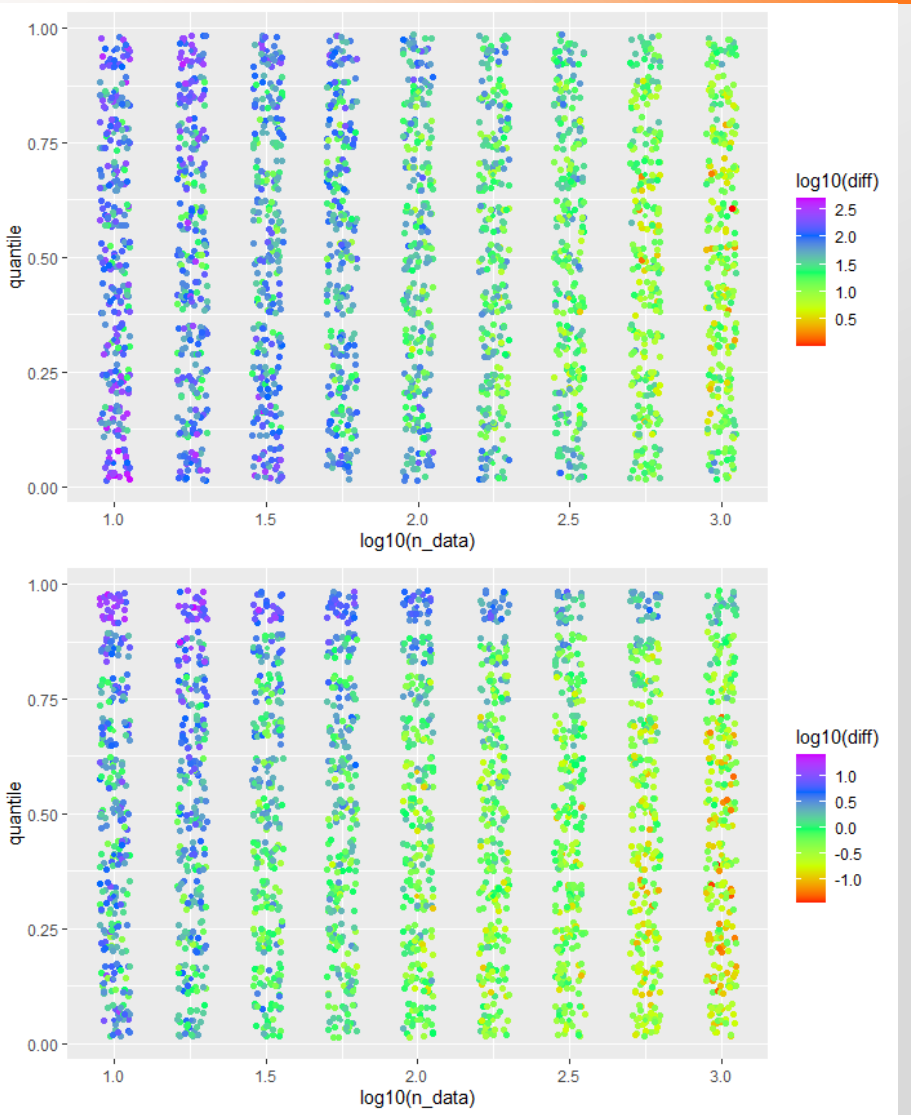
National Nuclear Security Administration

Convergence rate depends on distribution

Uniform



Normal



Triangular

Lognormal

National Nuclear Security Administration

How much data is enough? Attempt #2

- Used Cramer-Von Mises & Anderson-Darling goodness-of-fit tests to compare theoretical and empirical s-curves at 10 randomly-selected points on the s-curve
- Declared convergence by when the goodness of fit test was unable to distinguish between the theoretical and empirical CDF
- Performed 25 realizations per distribution
- In all cases, convergence occurred around 30 datapoints (~90% confidence) or 100 datapoints (~95% confidence)

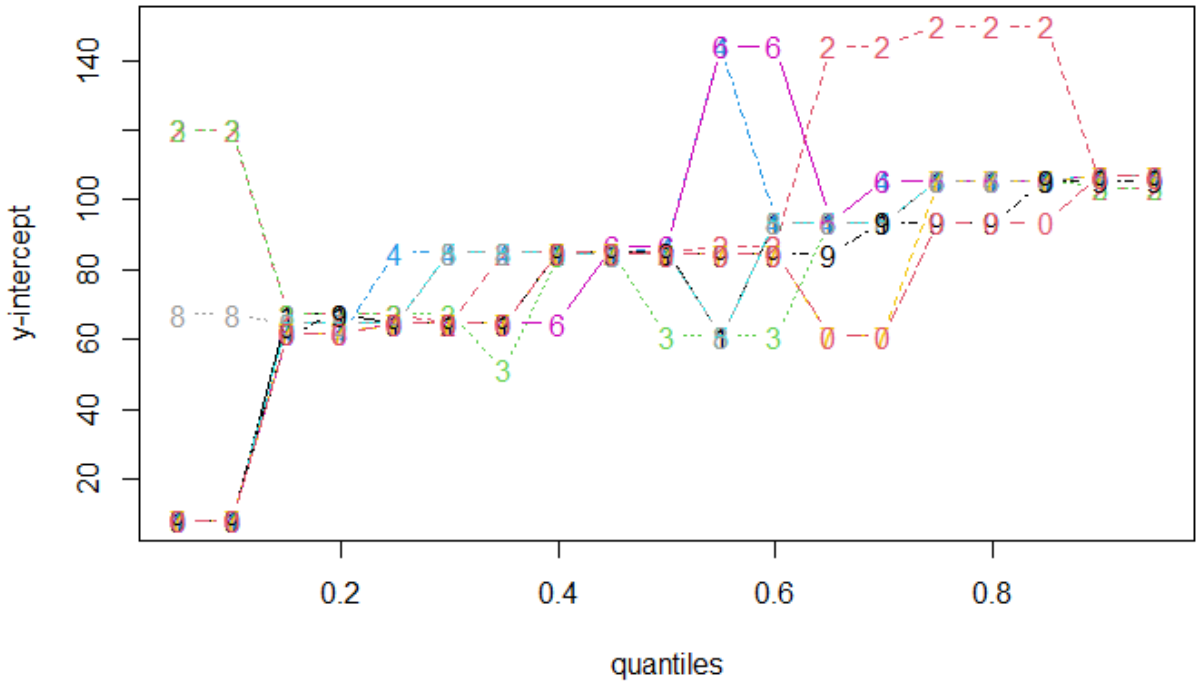
How much data is enough? Interlude

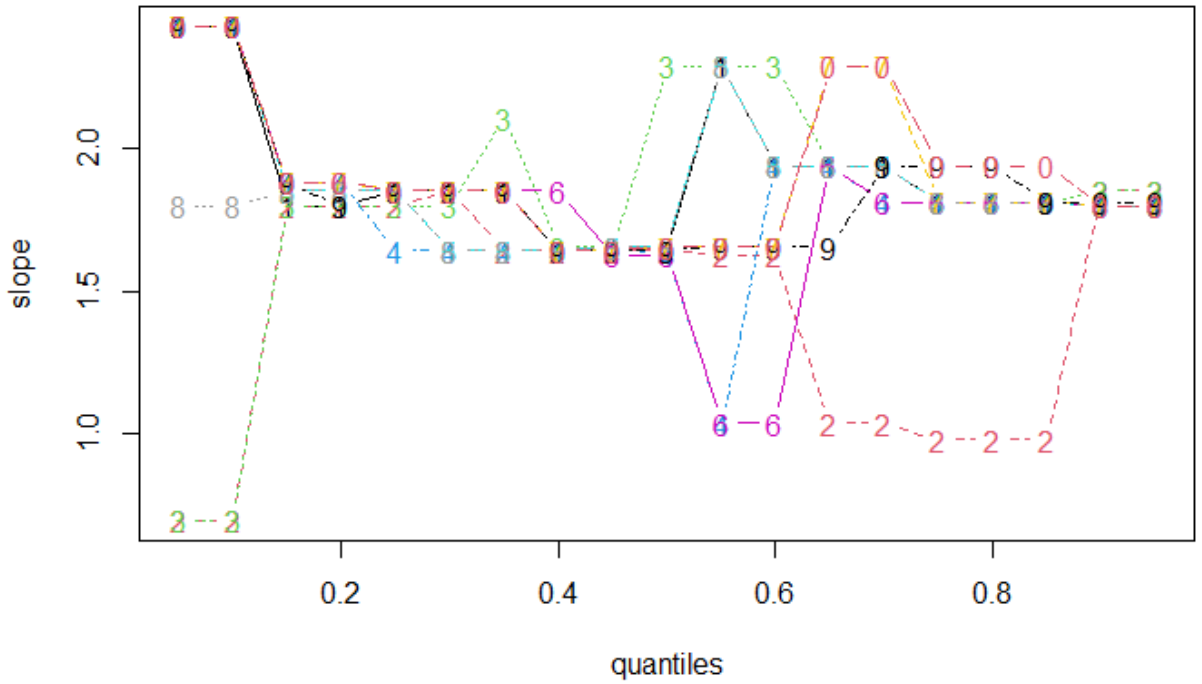
- Better, but still not a perfect solution
  - Analysis was limited to known, well-behaved distributions
  - Statistical test don't guarantee convergence, but rather "fail to reject the hypothesis that the two distributions are equivalent"
- Is there a way to test our own data for "sufficiency", without knowing the underlying distribution?

How much data is enough? Attempt #3

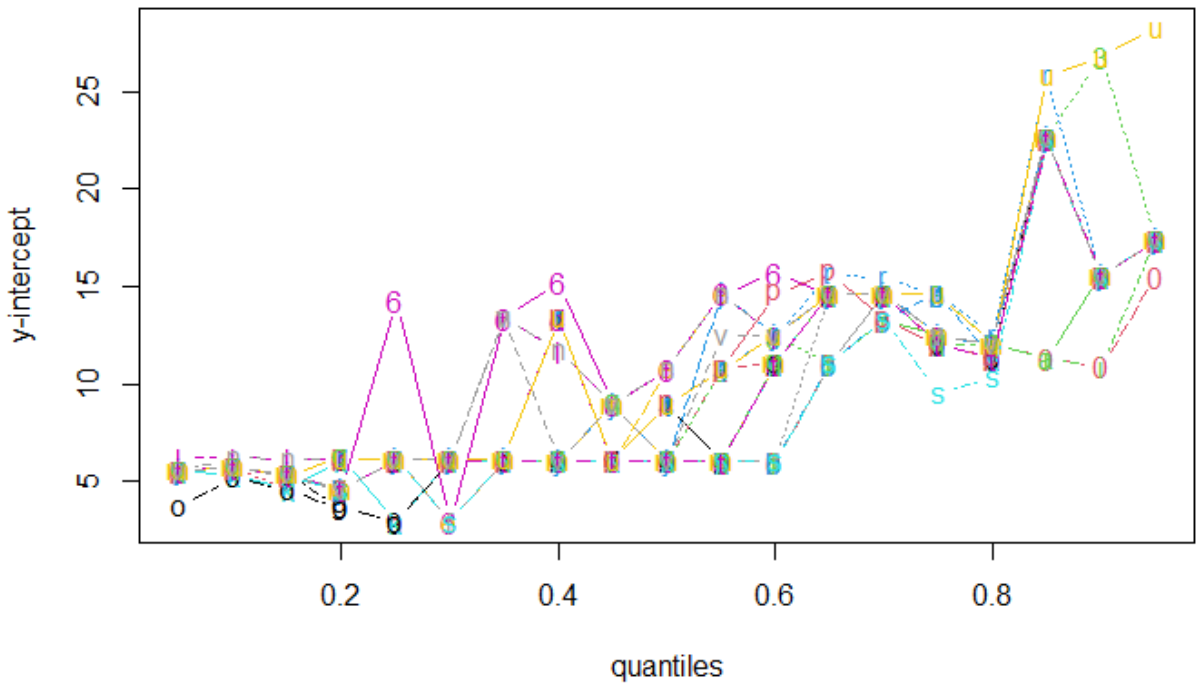
- Use a regression technique called jackknife, in which you remove one point at a time from the data set and perform the regression
- Jackknifing helps us to identify influential points that strongly affect the regression

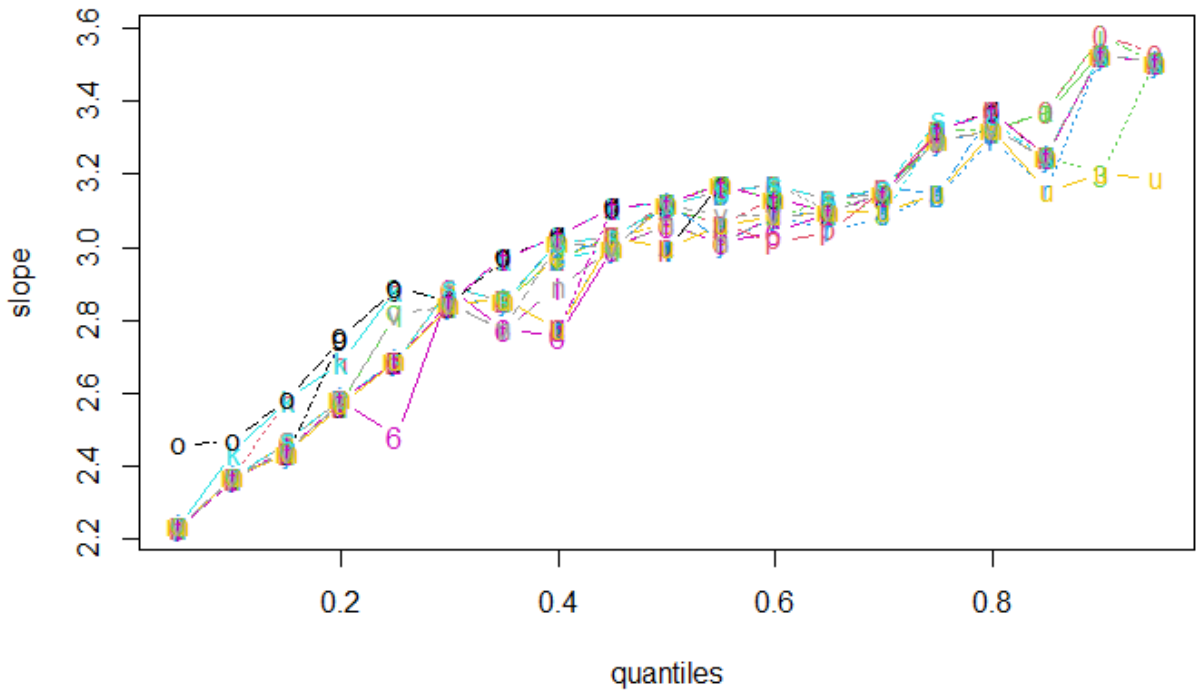
N=10, uniformly-distributed



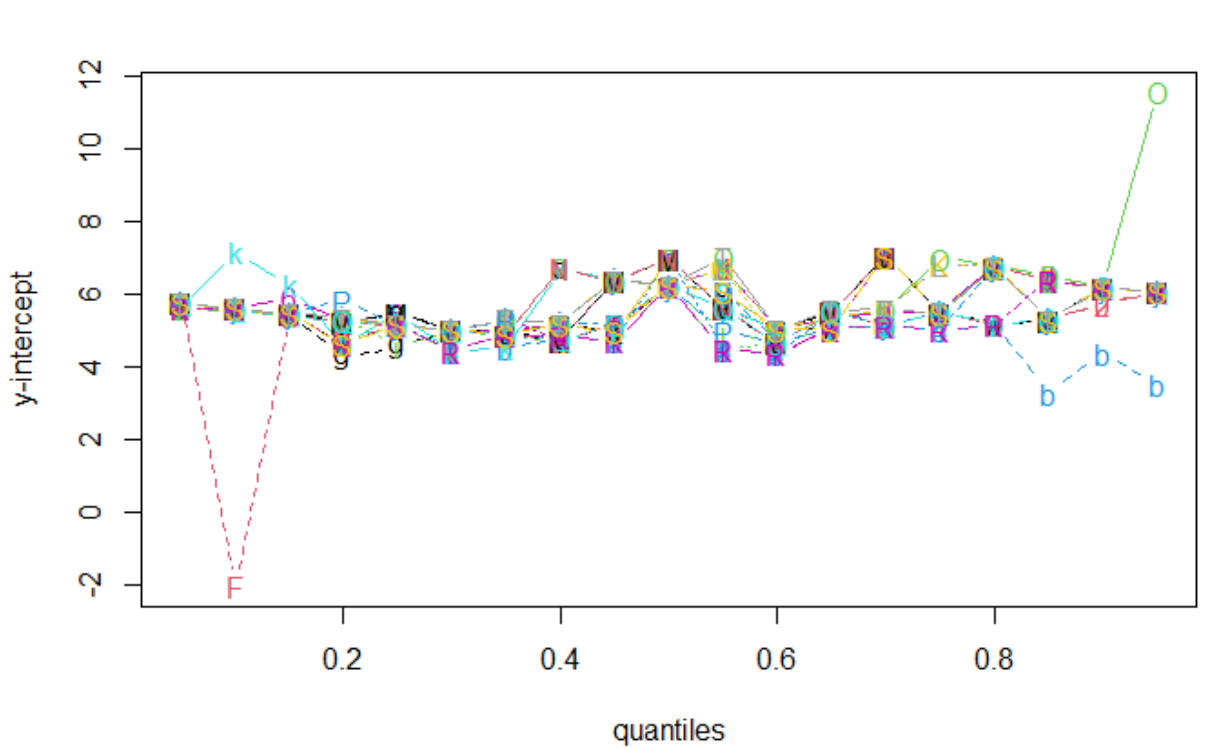


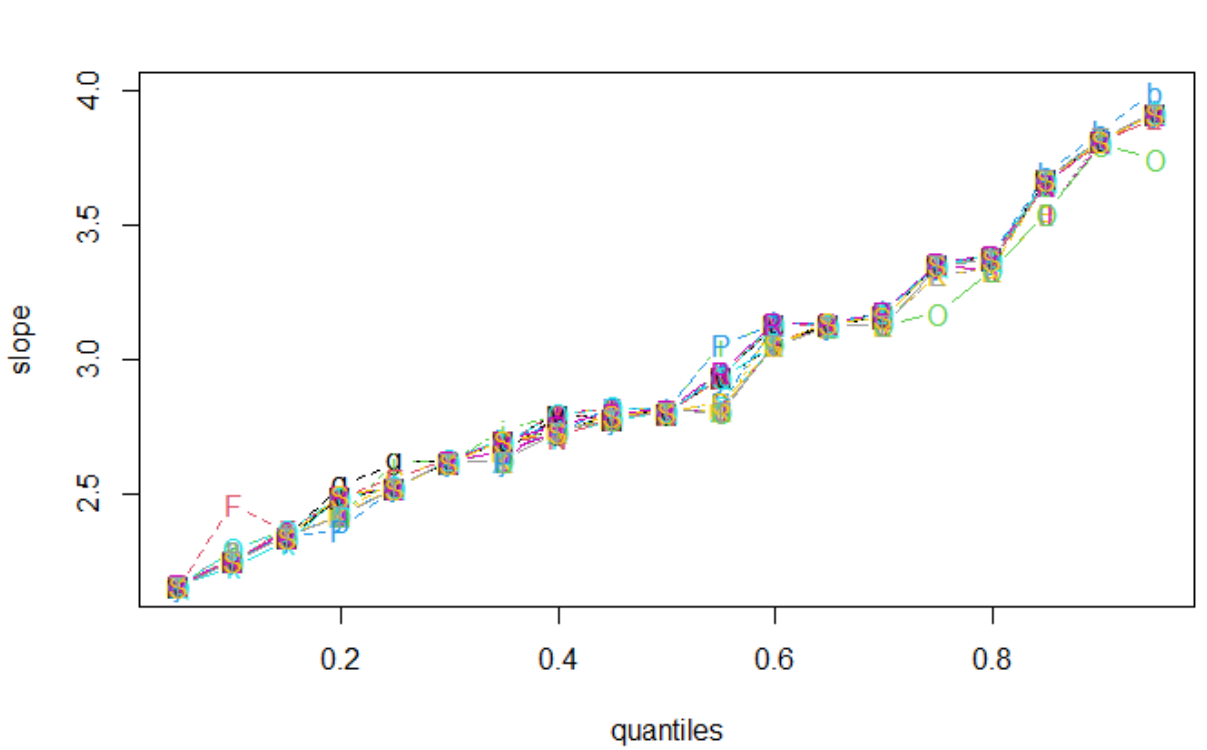
N=32, uniformly-distributed



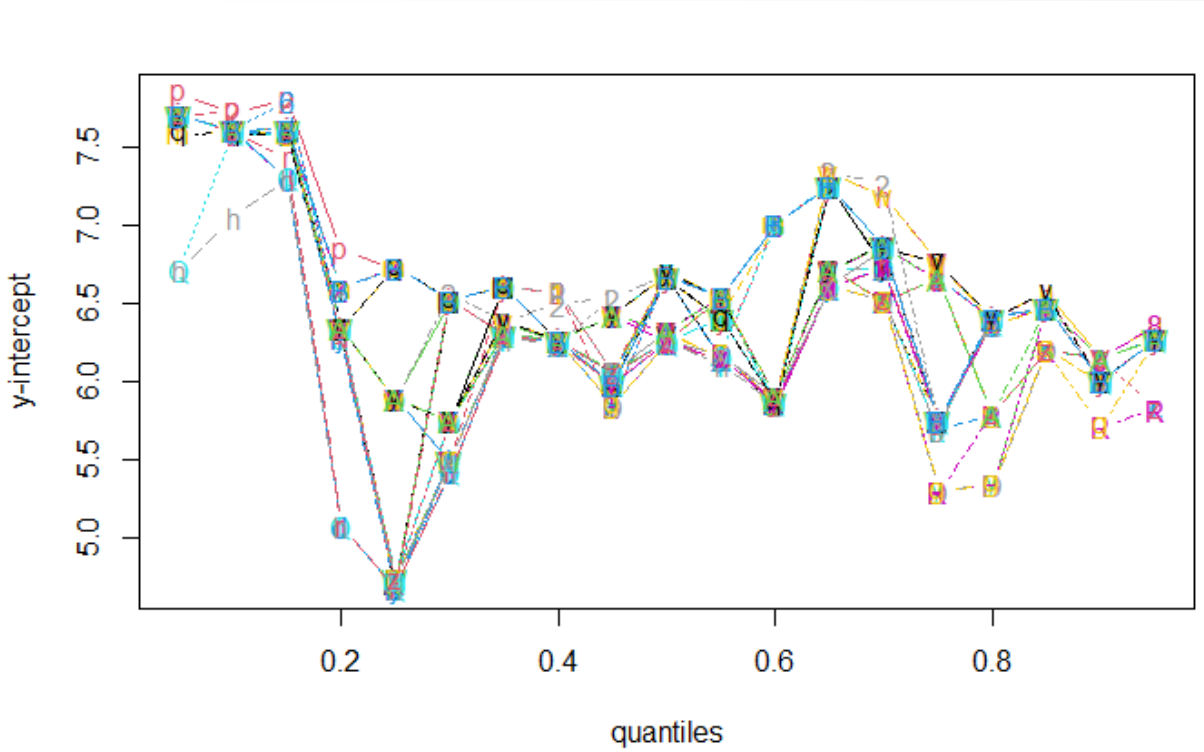


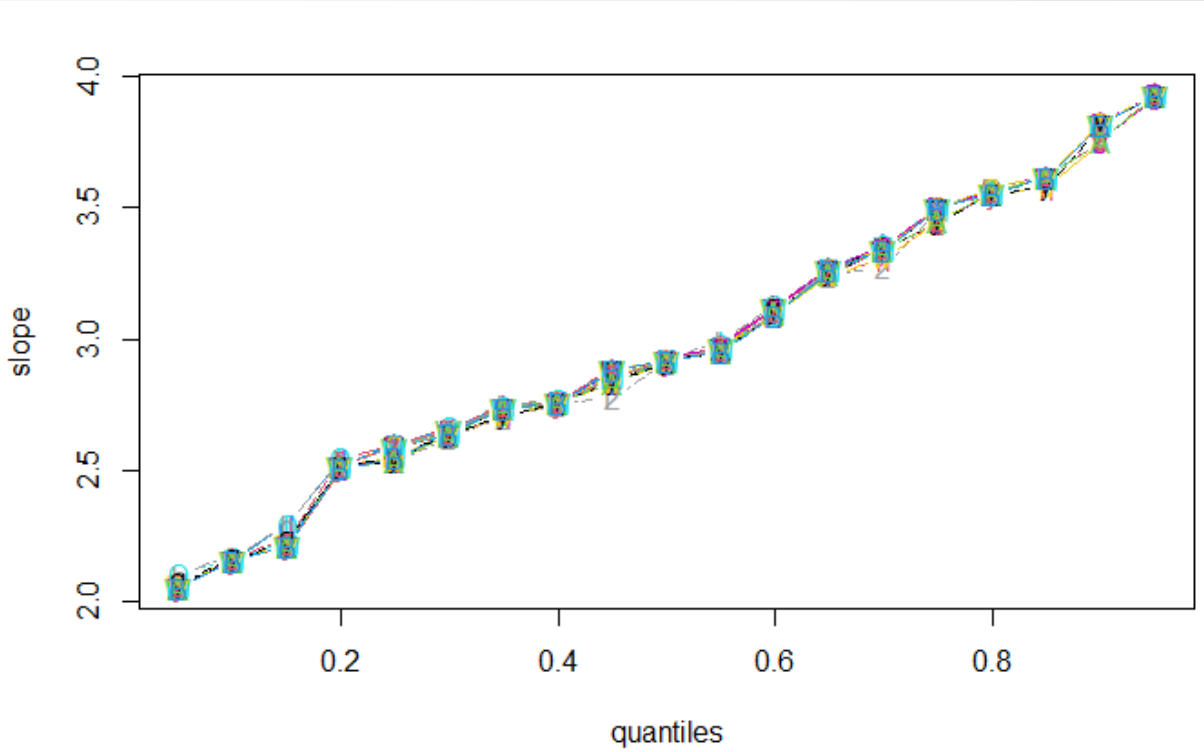
N=56, uniformly-distributed



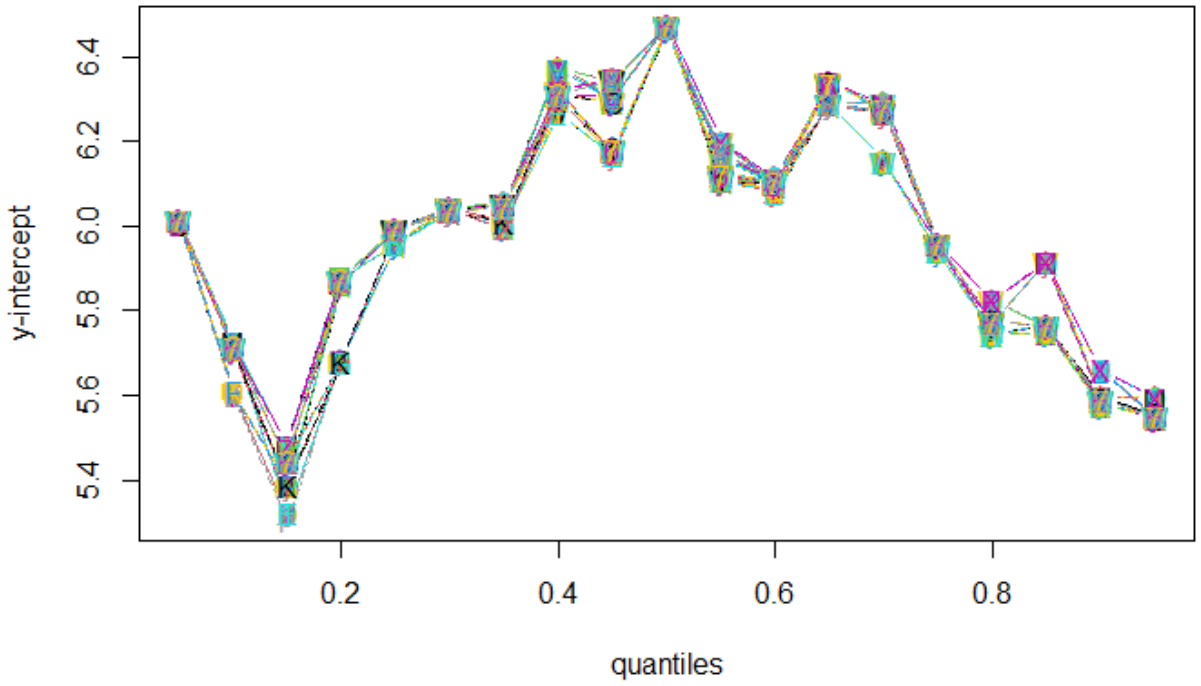


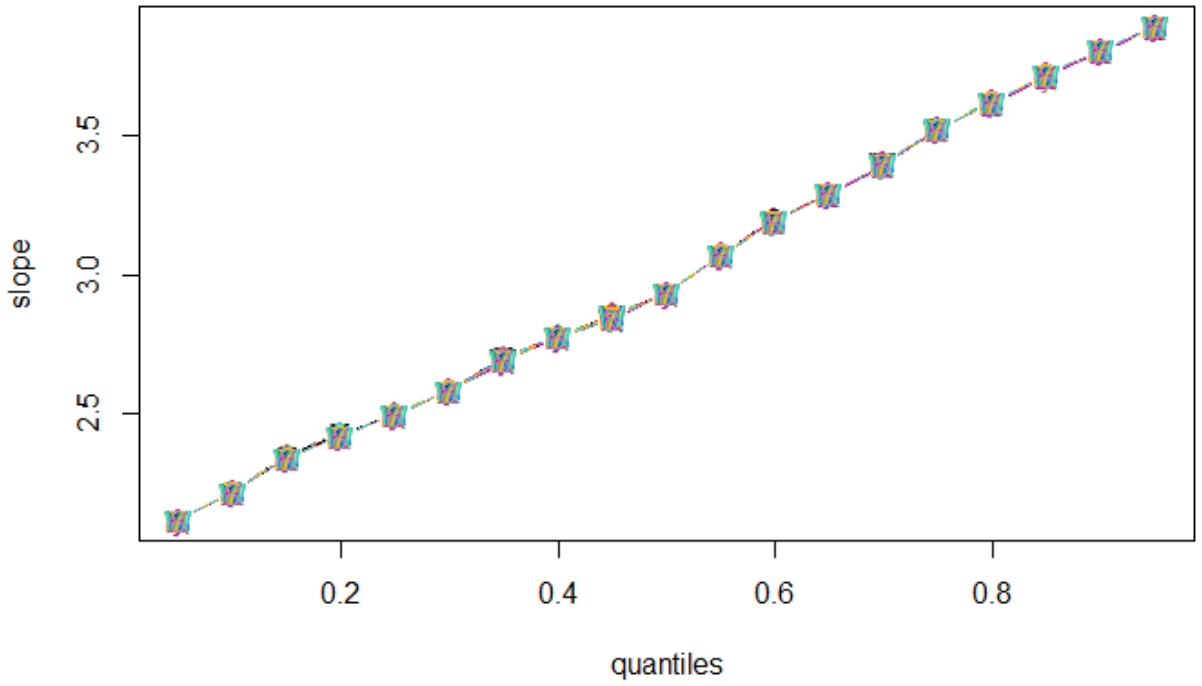
N=100, uniformly-distributed





N=1000, uniformly-distributed





Notes and cautions

- Finding an influential point is not prescriptive – doesn't say whether you should/shouldn't use QR, or that you must drop/keep the offending point. It simply allows you to peer inside the black box a little bit.
- Points singled out by the jackknife method can help to better understand potential cost drivers.

- CAUTION! For small n, regression may not be uniquely defined.

Conclusions and next steps

- **Lessons Learned:**
  - Found that PA&E datasets may be too small for quantile regression.

- **Path Forward:**
  - Identify influential points and find other ways to mitigate their impact without tossing them out.
  - How to handle bias in model (if even necessary)?

Additional thanks to Victoria Walter and Andrew Campo

Backup

- In order to make inferences from OLS, you must make certain assumptions about your data (independence, heteroscedasticity, etc.)
- QR makes no such assumptions. In one sense this is very freeing, because it allows you to explore different datasets where the assumptions of OLS inference do not apply. On the other hand, you probably have your own set of preconceived assumptions when you make inferences (linearity, additivity, continuity), and without the well-developed infrastructure of OLS you must check these things for yourself.

Approach

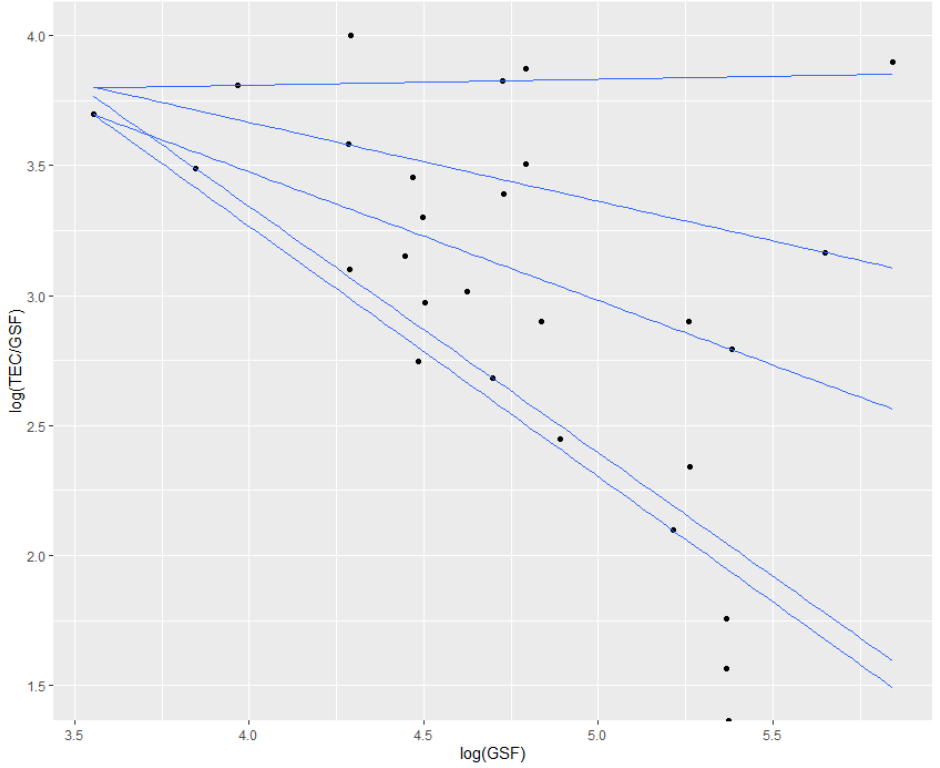
- *Use linear quantile regression to determine model coefficients for capital acquisitions cost and schedule data (TEC/GSF, OPC % of TPC, and CD-1 to CD-4 duration) and D&D cost data.*

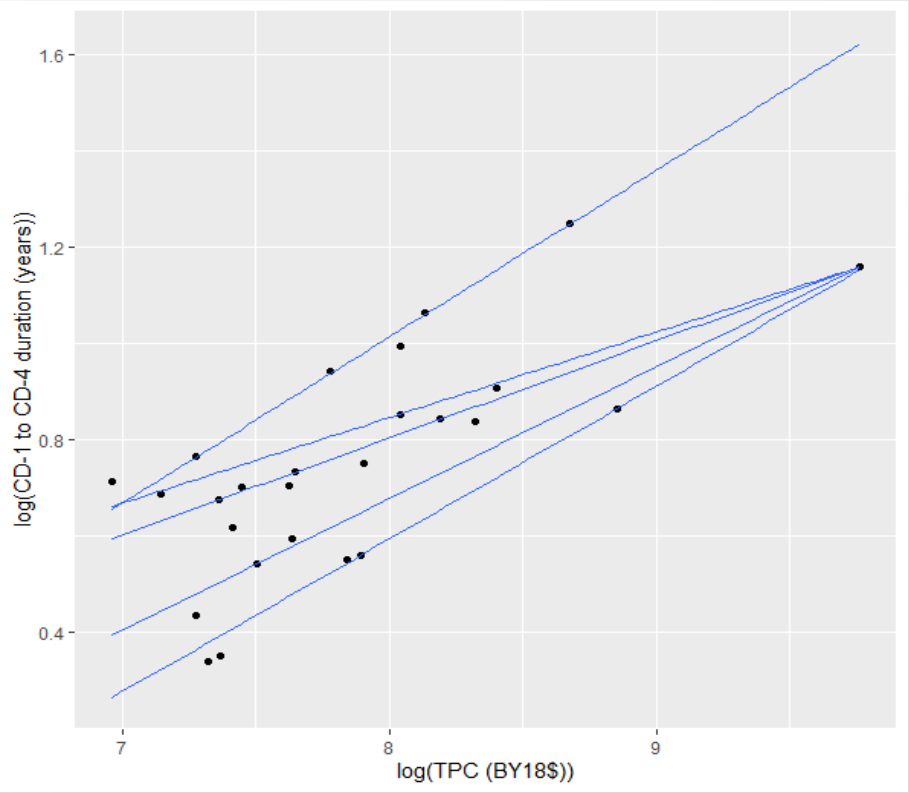
- *Compare QR model parameters to CSPER-C and DICEROLLER model parameters.*

- *Use statistical technique known as bootstrapping to generate confidence intervals around model parameters. Bootstrapping can also help mitigate the impact of "influential points" in small datasets.*

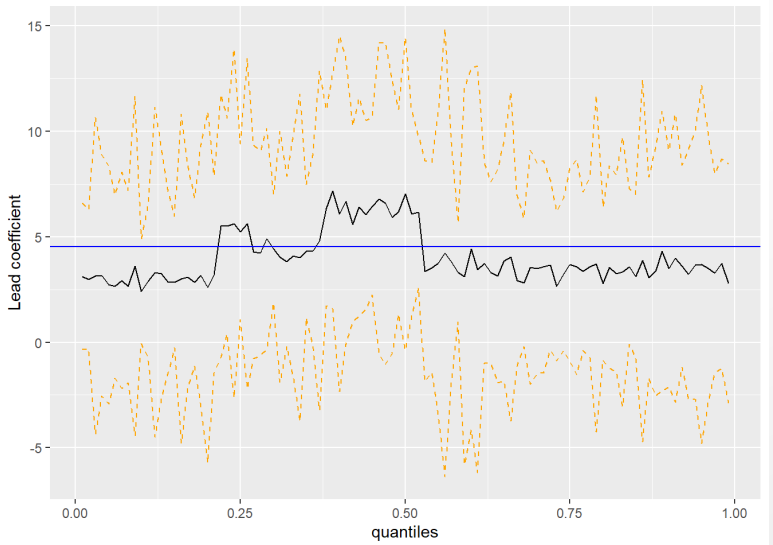
- *Compute 70% prediction intervals using DICEROLLER and QR D&D model and compared results.*

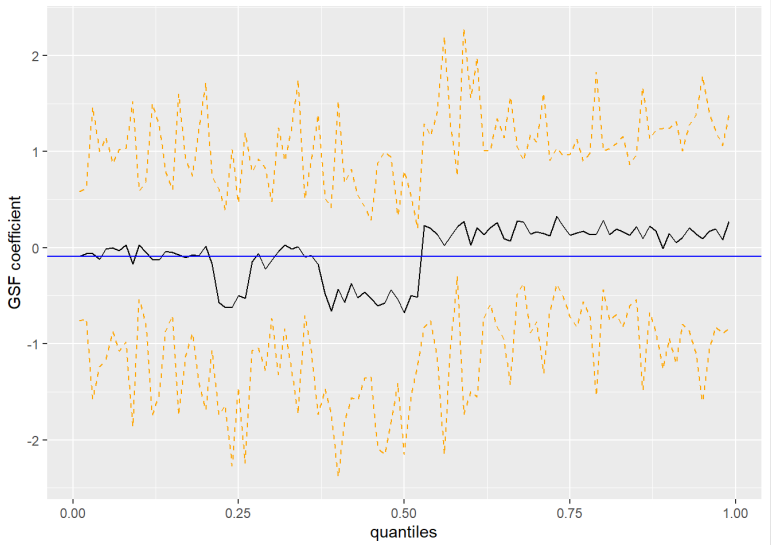
Results – CSPER-C cost and schedule model

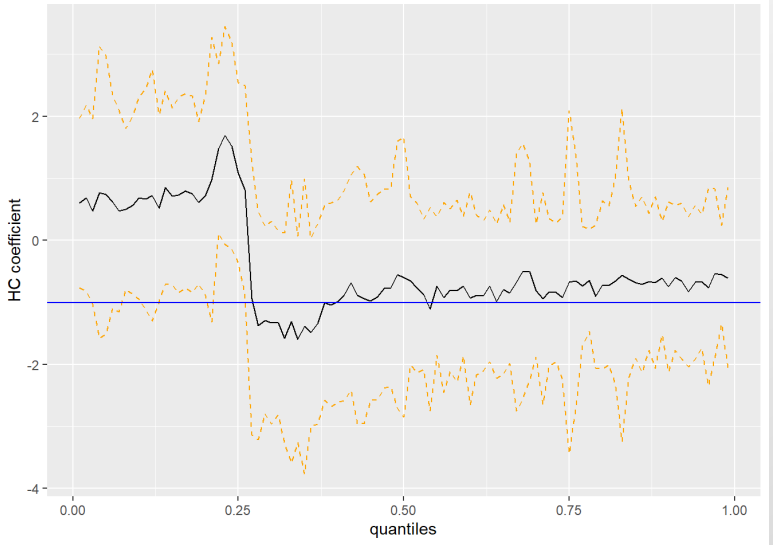


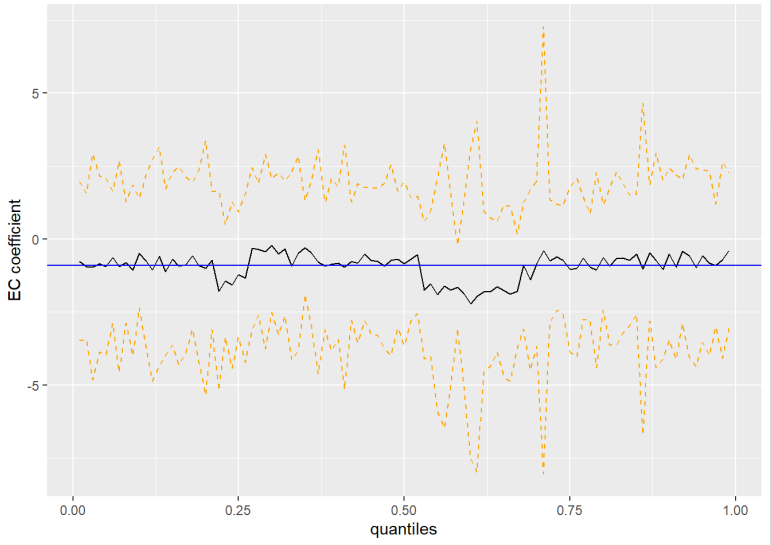


Results – CSPER-C coefficients



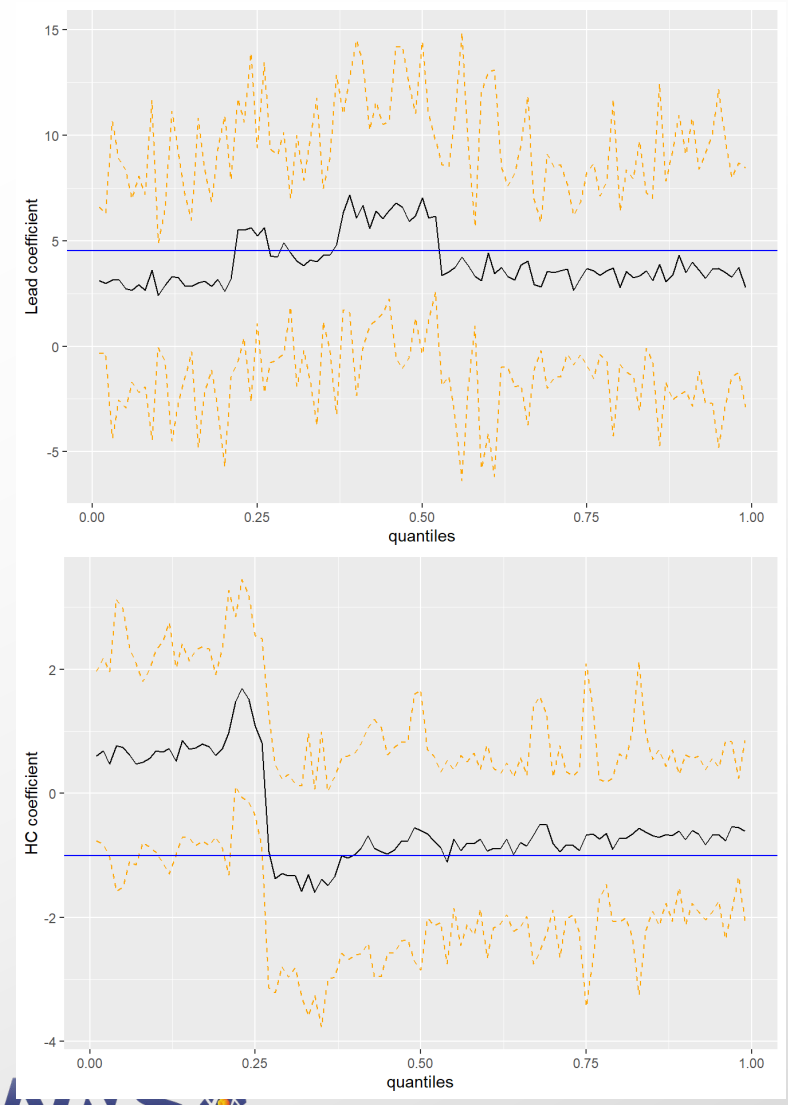


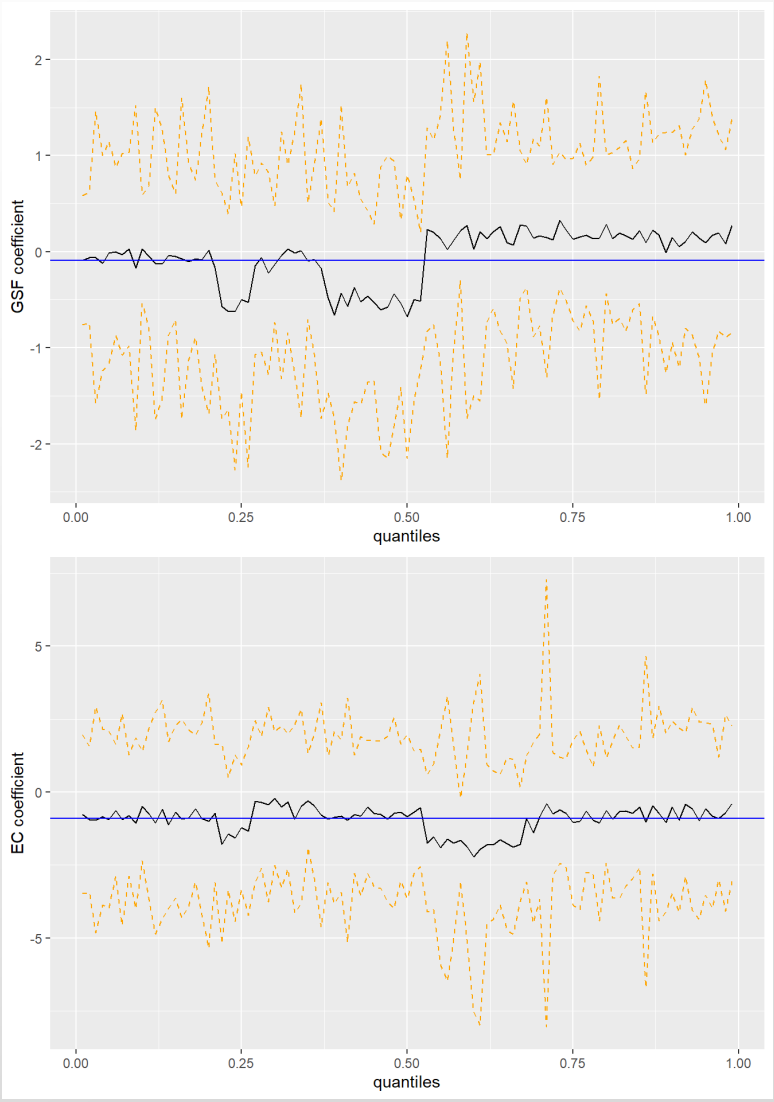


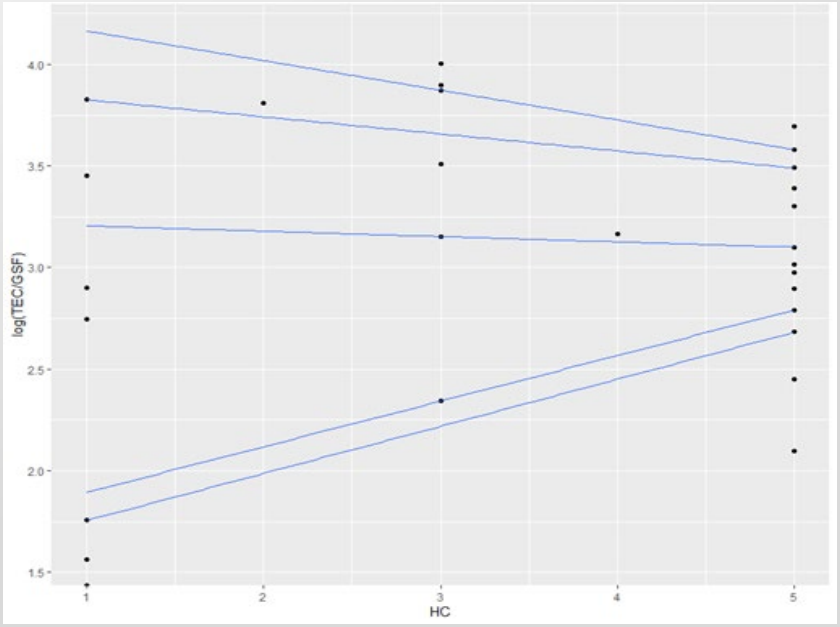


National Nuclear Security Administration

Results – CSPER-C coefficients







Comparing D&D prediction intervals

- Consider a building with the following characteristics:
  - 10,000 GSF
  - Nuclear and asbestos contamination
  - Hardened process facility

	DICEROLLER (OLS)	Quantile Regression
Max (85th %ile) log(TPC)	7.242	7.258
Most likely log(TPC)	6.889 (mean)	6.744 (median)
Min (15th %ile) log(TPC)	6.537	6.488