Depending on the Correlation Coefficient? (and Other Pitfalls in Curve Fitting)

ASP Workshop, August 5, 2020

William Rogers
UCOR

A highly idiosyncratic and personal approach based on 45+ years of curve fitting

William J. Rogers – Sample Management Office, Oak Ridge, TN

Disclaimer

- This guy is so old he had to write his own curve fit routines in FORTRAN on punch cards
  - And derive the equations from scratch
    - That's old!

The question which drove today's talk

- Is there any need to modify the QSM?

Ancient quote regarding curve fitting

- Six adjustable parameters!!!!! No researcher is sufficiently honest to be trusted with 6 adjustable parameters!!!

“Plot Your Data”

- Prof Al Van Hook – U of Tennessee – 1975 “The brain learns in a different way (about your data from feedback) as the muscles move across the paper.”

- *And so, In Conclusion: At a minimum, plot your data and plot your curves (the points in between) on top of the data.*

We understand

- Some of what is going on in data crunching is proprietary – hidden away in the manufacturer's software and they play their cards pretty close.
  - Example: ICP/MS purports to be a "forced through zero" fit but is actually has a little jiggery-pokery going on first.
- Sometimes you encounter a "spline fit" which is actually an interpolation technique. The r-squared values for that are a series of 1.00000000
    - But that's good, right?

Clarification – Linear Regression

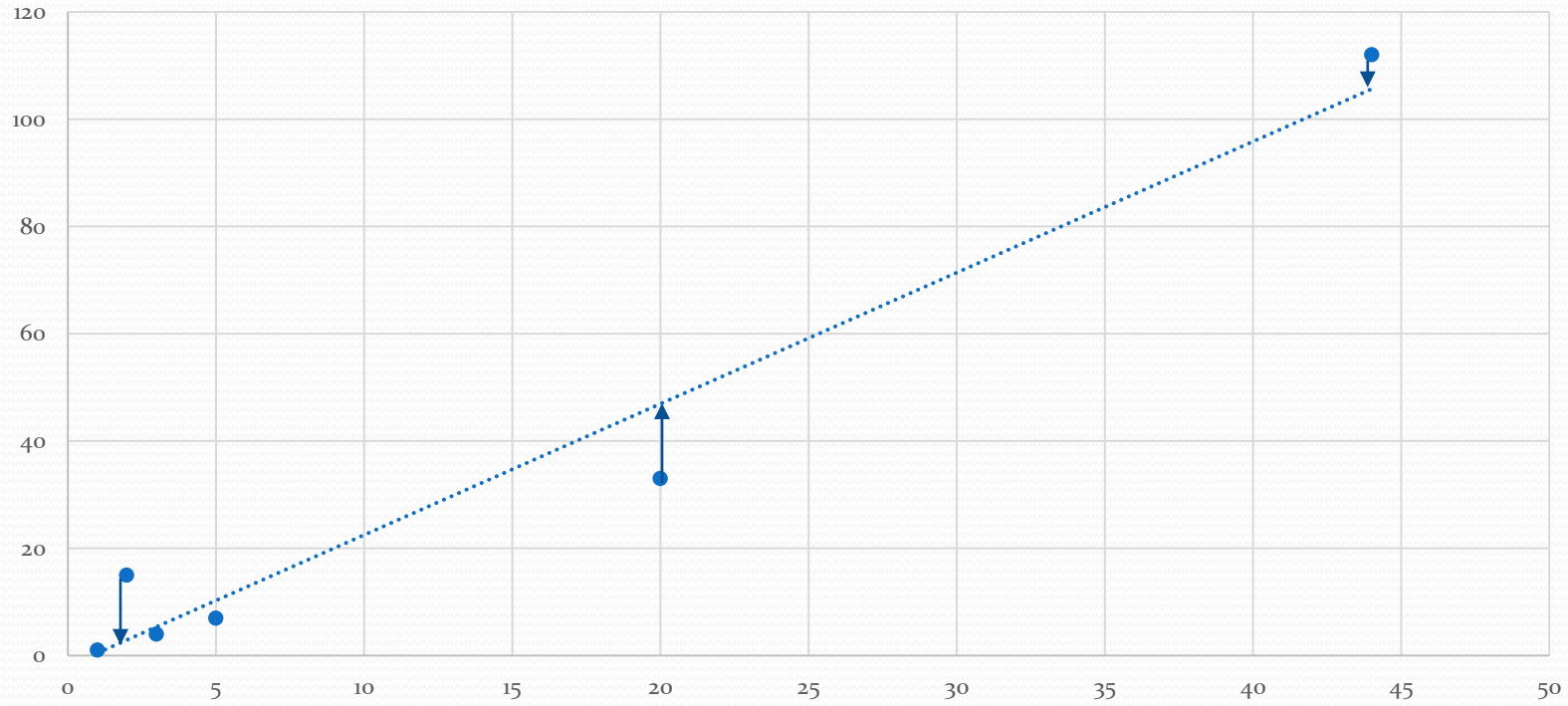
- In statistics, linear regression is a linear approach to modeling the relationship between a scalar response and one or more explanatory variables. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. - wikipedia
- “linear regression” does not mean the form you fit to must be $y = a + bx$

Understand what is happening mathematically

- The equation chosen must realistically match the response of your device
  - Gamma energy-efficiency curves are hard
  - Nitrate-nitrite on a colorimeter is a fairly straight line up to a point
- You are minimizing the "sum" of the distances between the data and the fitted line.
  - It should not be zero – too good to be true
    - Be aware of degrees of freedom
    - Try to fit a parabola to 3 points and see what happens (example to follow)

Understand what is happening mathematically

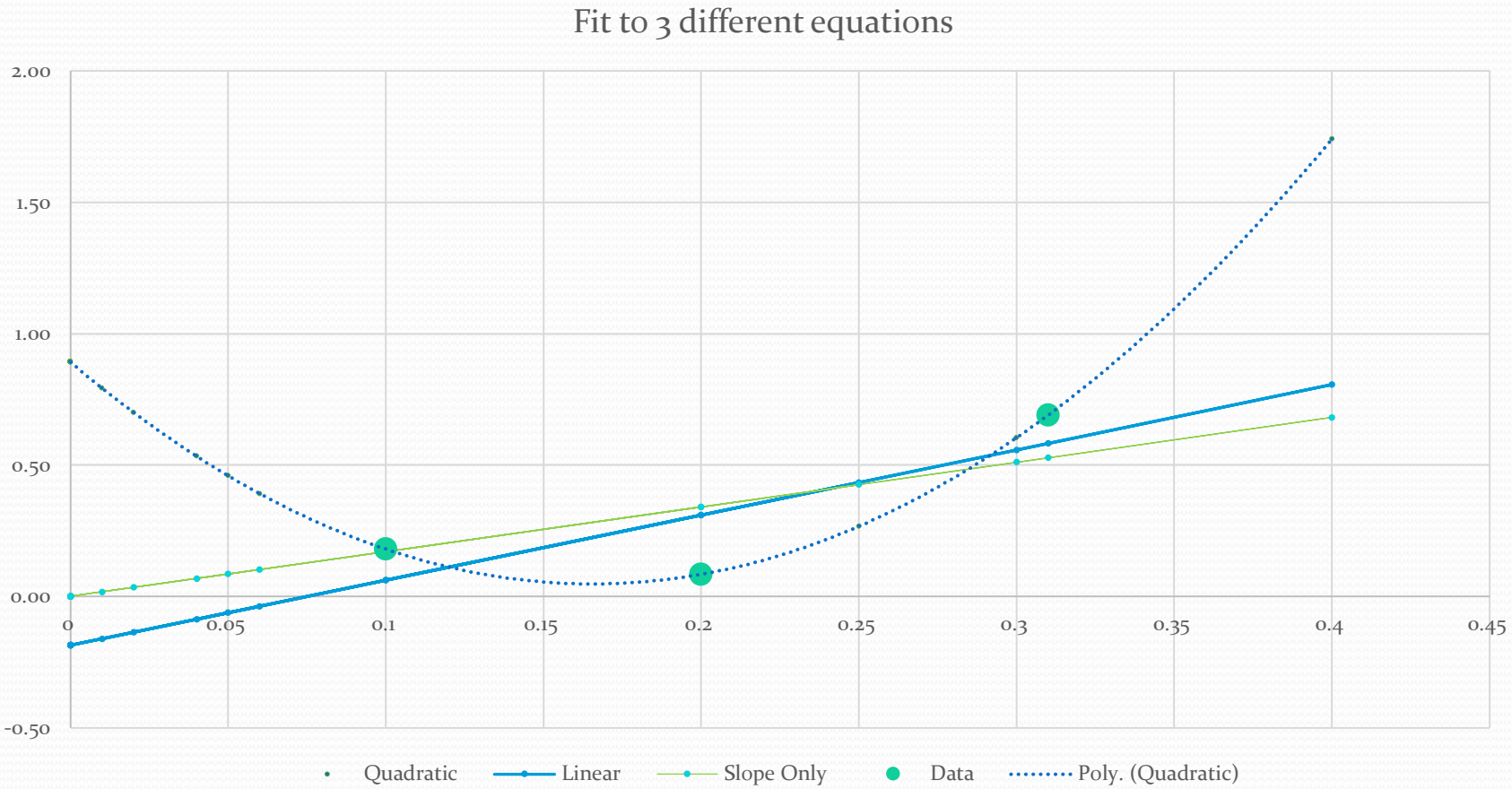
Minimize the sum of the absolute differences of the “drop lines”



What is the correlation coefficient?

- It is the “goodness of fit” statistic
- Also called the residual sum of squares (RSS)
- Also called r-squared
- Excel ™ documentation states
  - the RSS “compares estimated and actual y-values, and ranges in value from 0 to 1. If it is 1, there is a perfect correlation in the sample ”

Understand what is happening, mathematically



What does the QSM say?

- A few quotes
  - . . . linear least squares regression for each analyte: r2 ≥ 0.99
  - . . .non-linear least squares regression (quadratic) for each analyte: r2 ≥ 0.99
  - . . .95% confidence limit of the fitted function (curve) over the calibration range to ≤ 10% and ≤ 5% uncertainty for alpha and beta, respectively (MARLAP 18.5.6.1)
  - Best fit of data with correlation coefficient closest to 1.00 and the smallest standard error.

There are other statistics

$$Y = b + m_1 X + m_2 X^2$$

Key for Regression Output.		
m1	m2	b
std err m1	*std err m2*	*std err b*
R2	*Standard error of Y estimate*	
F	Degrees of freedom	
ss regression	ss residual	

Excel ™ LINEST function

Bad Examples

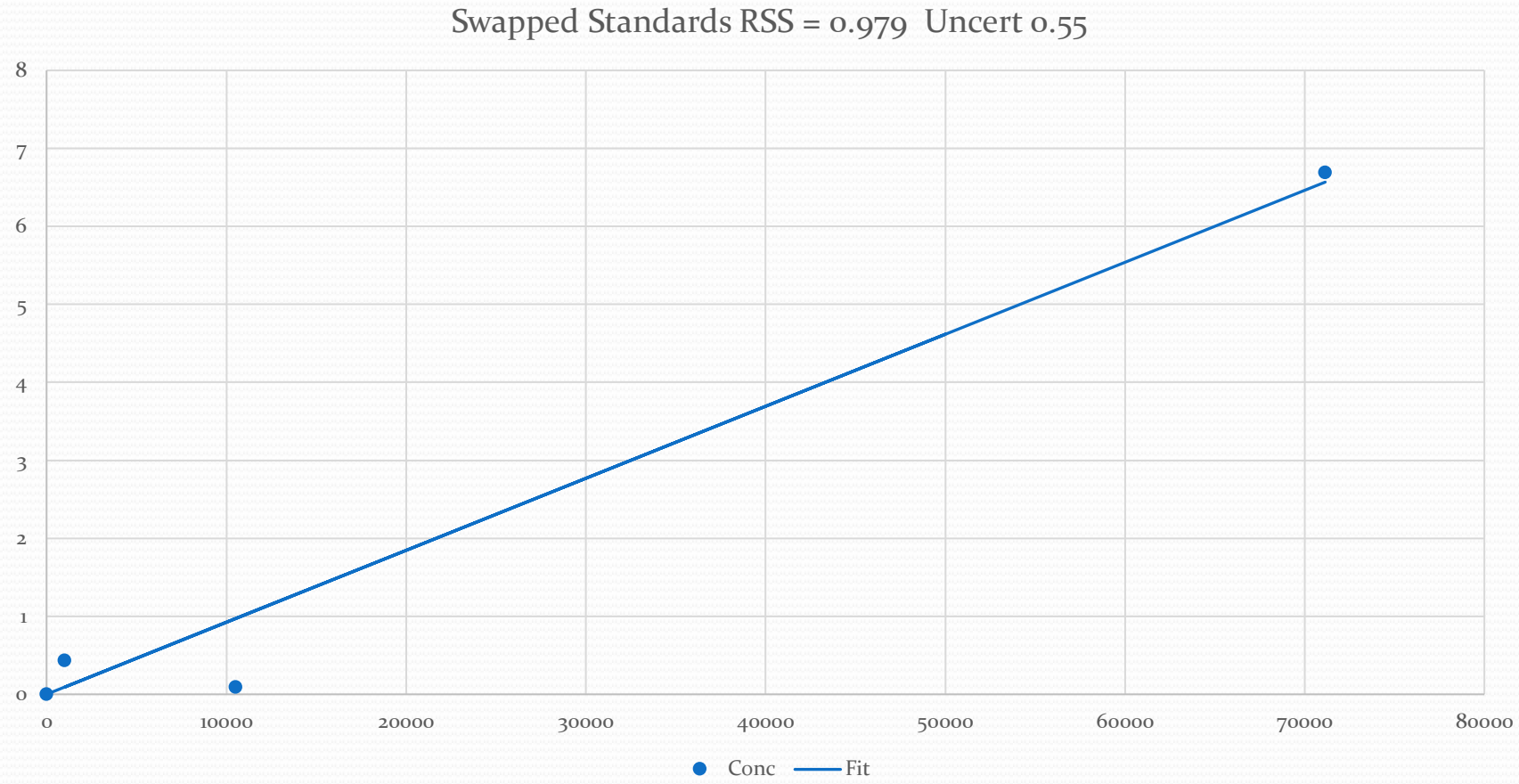
Bad

- Bad mathematical model – does not match device response
  - Government entity made us force through zero on gas chromatography calibration curves
    - Low end ignored blanks signal
- Mis-loading calibration standards – two swapped

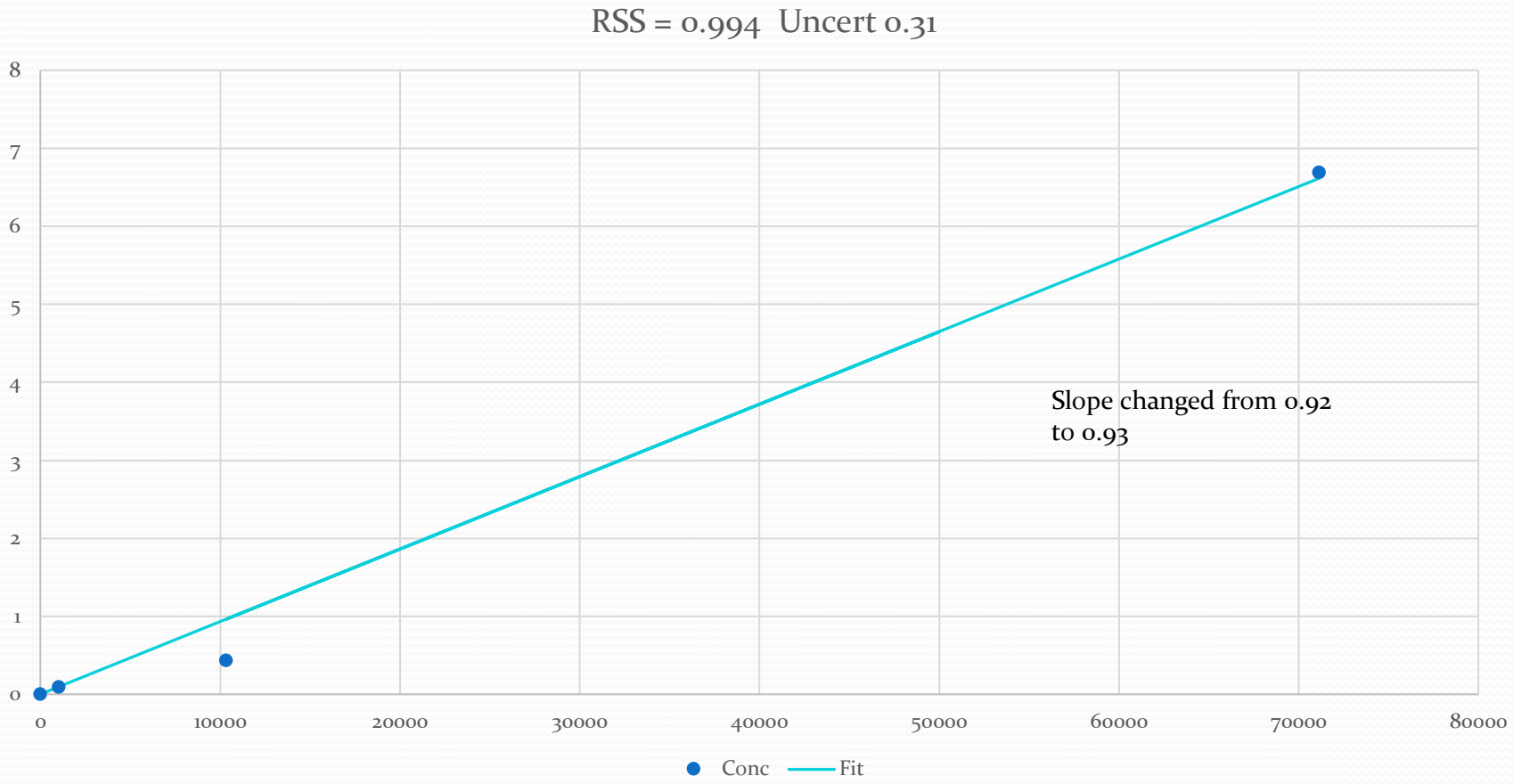
Bad

- Attempting too large a calibration range – uncertainty at high end dominates the fit
- Unrealistic notion of uncertainty in the data set
- Being unaware of what can happen when extrapolating outside the bounds of your fit
  - War story: low energy gamma

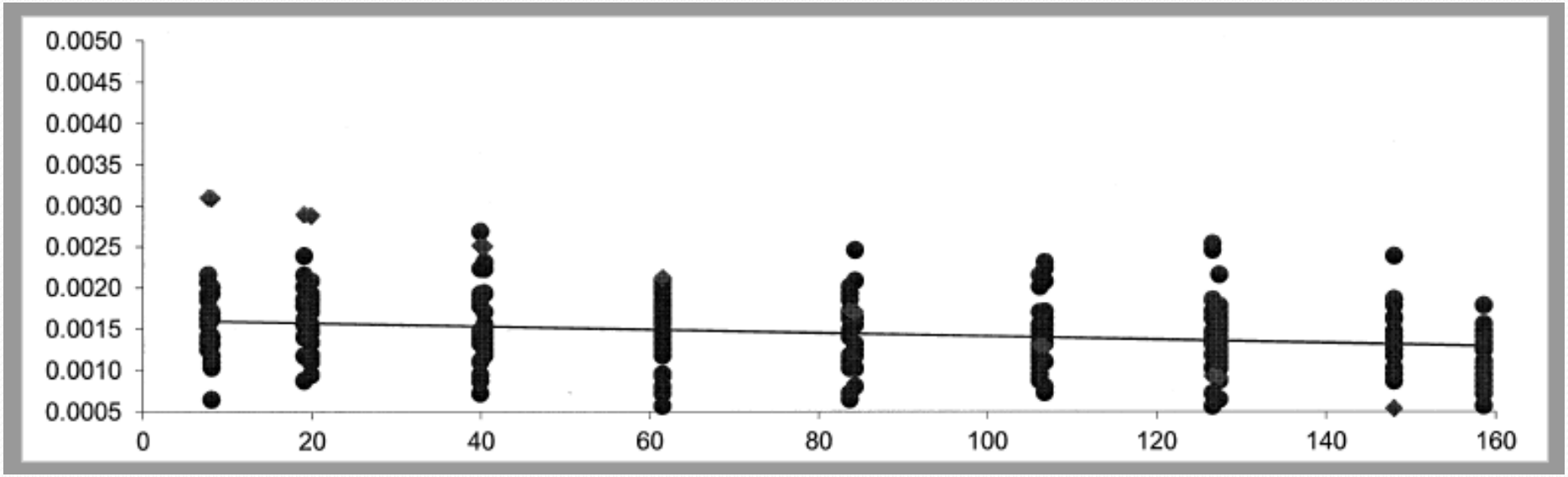
Swapped Standards



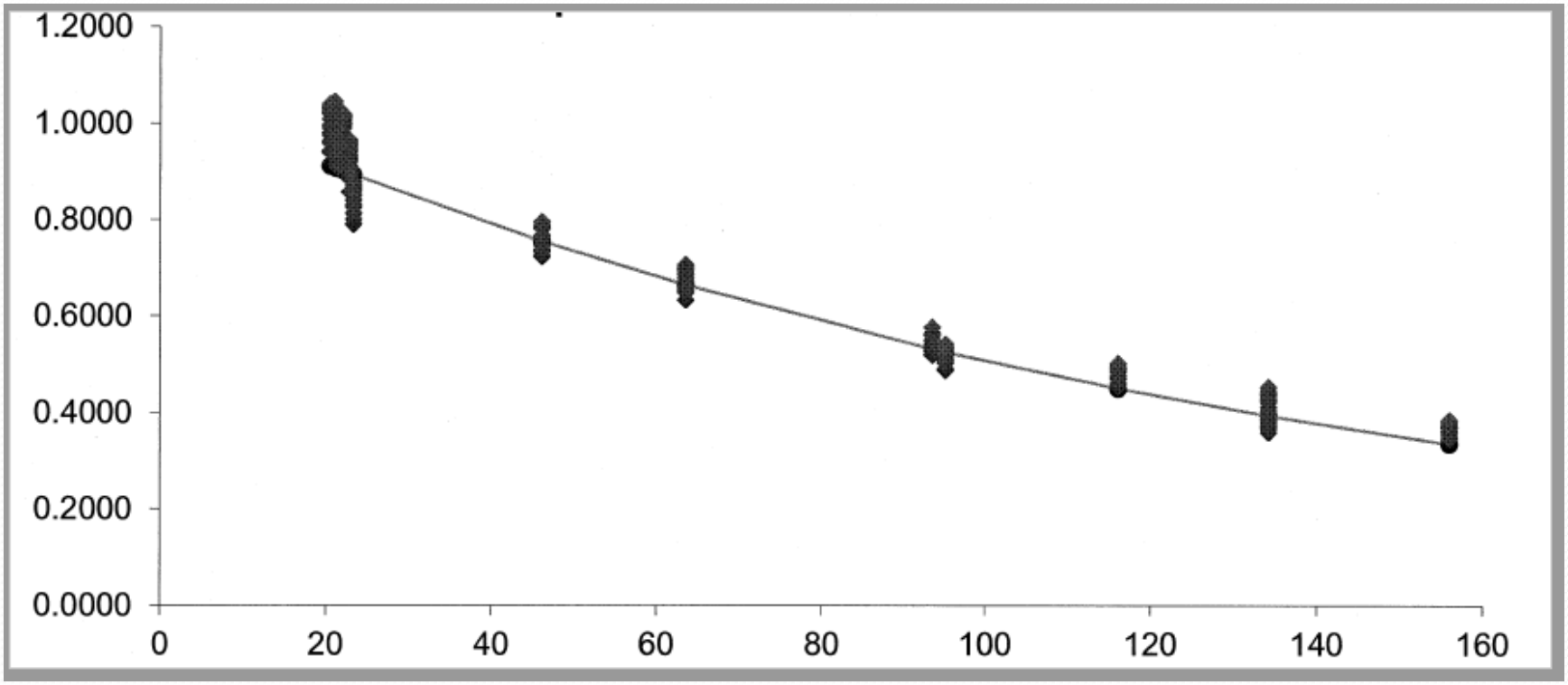
Swapped Standards



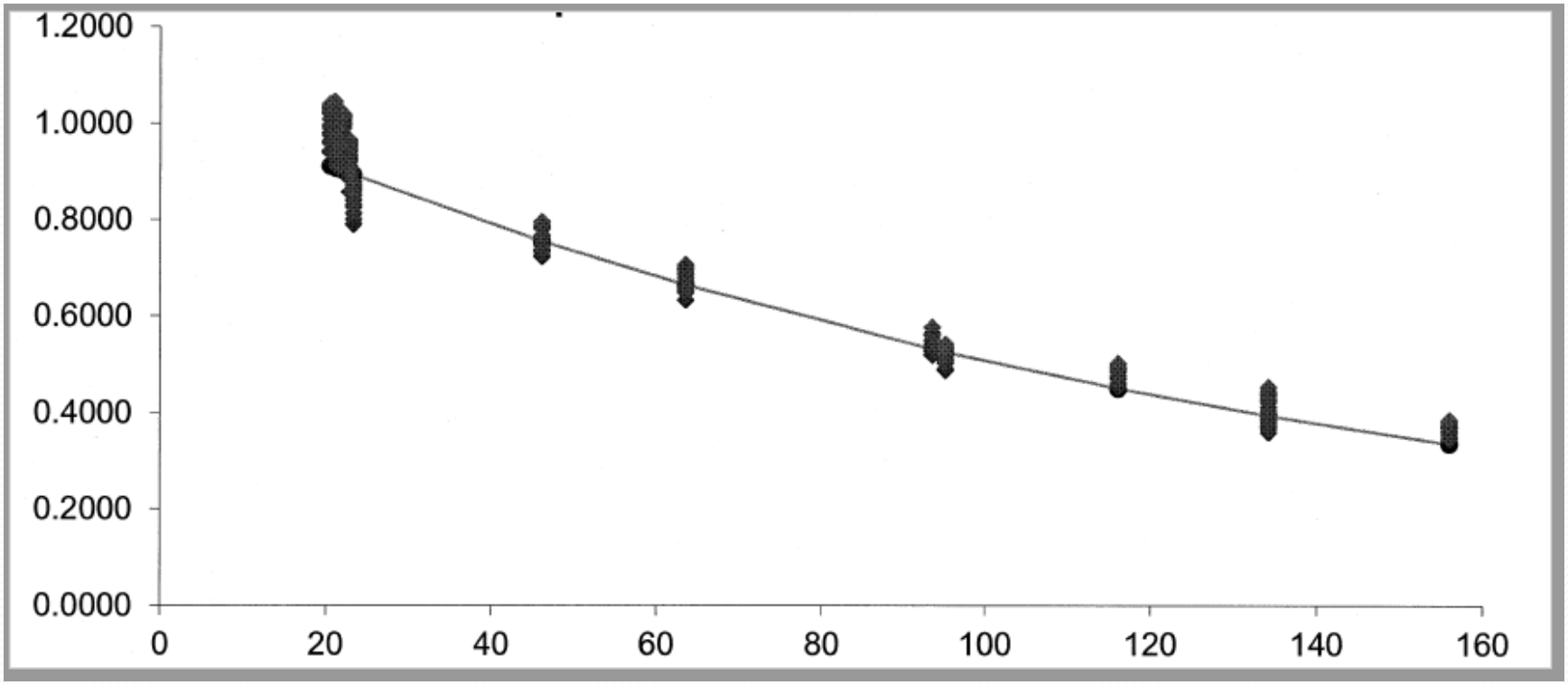
Unrealistic data set



Better

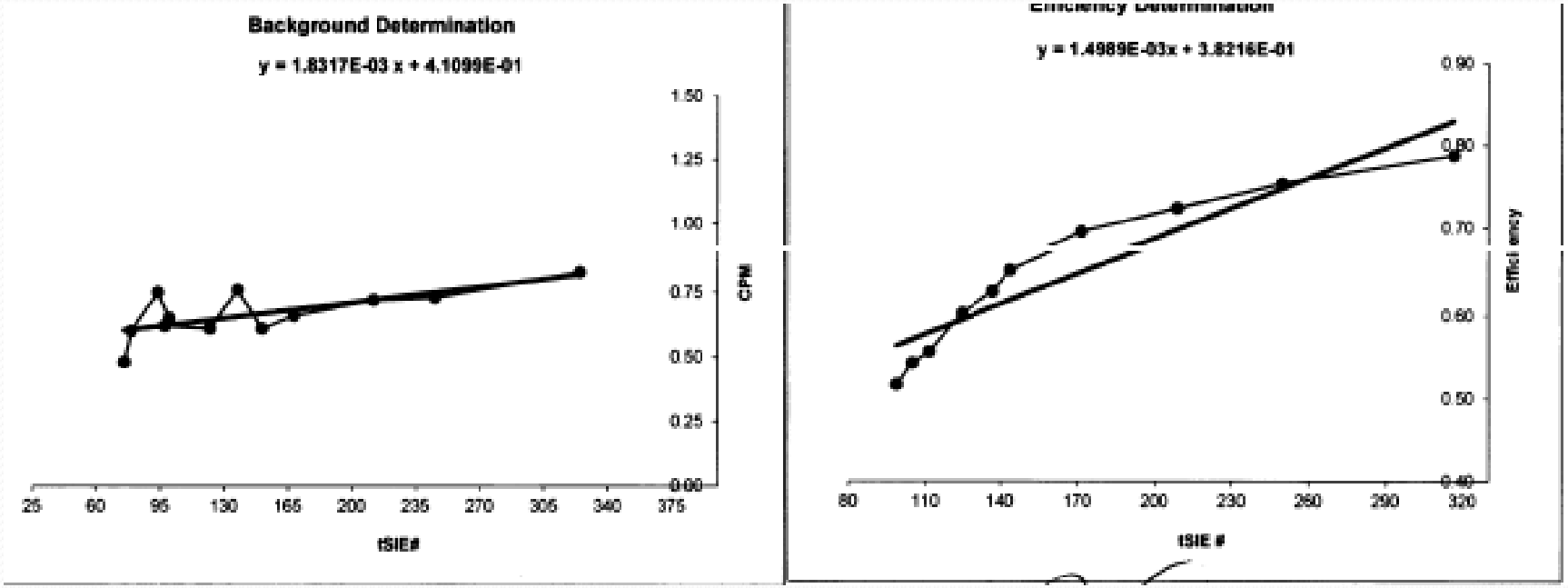

1.2000
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
0
20
40
60
80
100
120
140
160

Better



You can actually guess at what the 95% error envelope might be

Could be better



Desirable Behavior

Desirable

- Plot your data and the curve between points

Desirable

- Examine your calibration range for reasonableness

Desirable

- Try multiple models (forced through zero, linear, quadratic) to see if the RSS changes
- See if the uncertainty in the fit parameters changes with various models

Desirable

- If you can get your hands on it, evaluate the statistic “standard error of the Y estimate” in light of your lowest and highest calibration points

Desirable

- Realistically evaluate the uncertainty “band” from eyeballing your curve

Will anything change in the QSM

RSS > 0.99 has been around a long time

- RSS > 0.99 is fairly useful
- "Std Err Y est" is not available without sophisticated programming changes
  - But you could try your own fits via Excel™
- If you are operating in a quality arena which allows RSS > 0.95 – not very useful at all

Ideally: "visual plots of data and fitted curves must be made available to technical staff and clients"

Realistically, no change.

Questions?