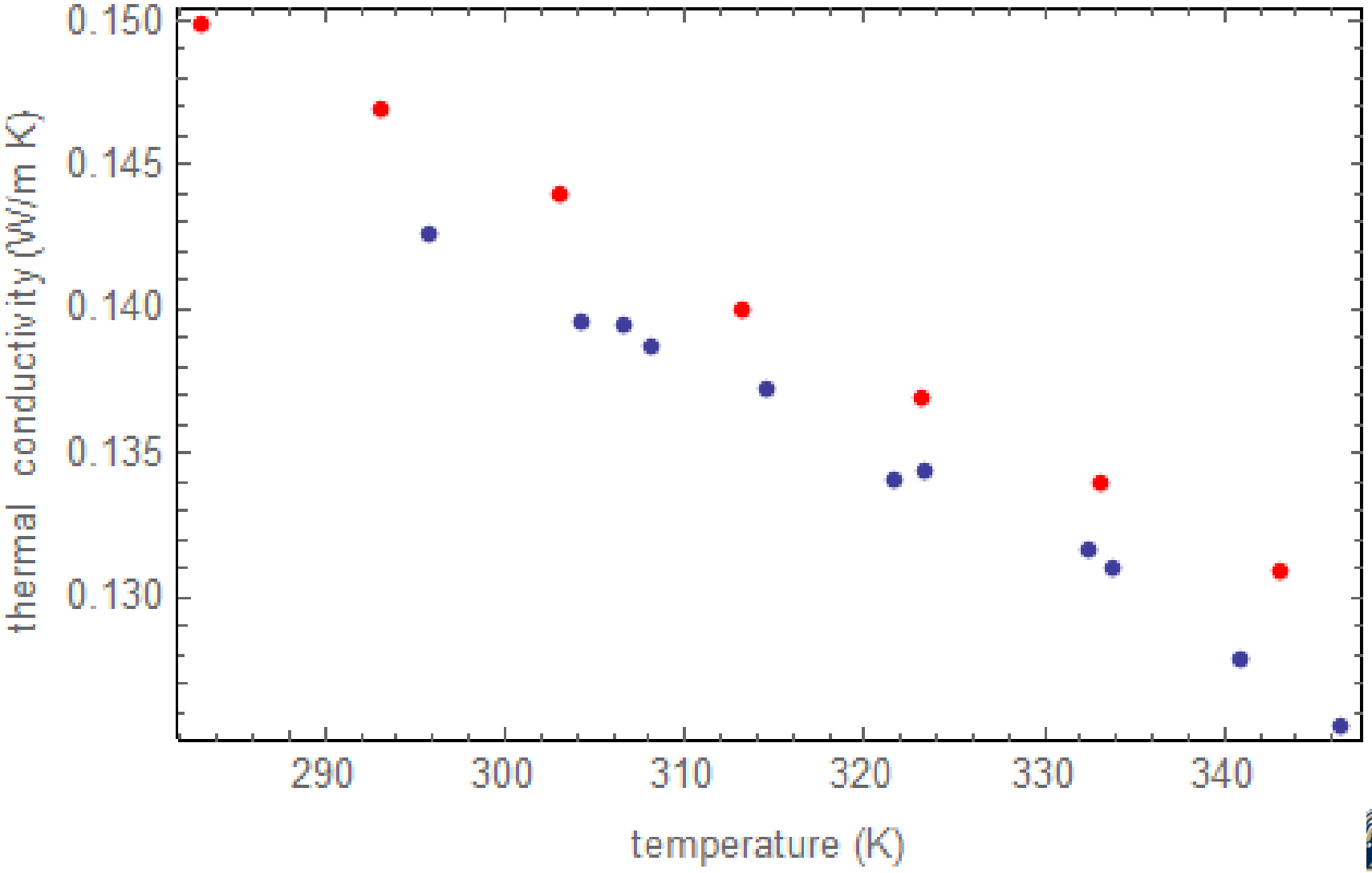
Practical Statistical Analysis

Objectives:

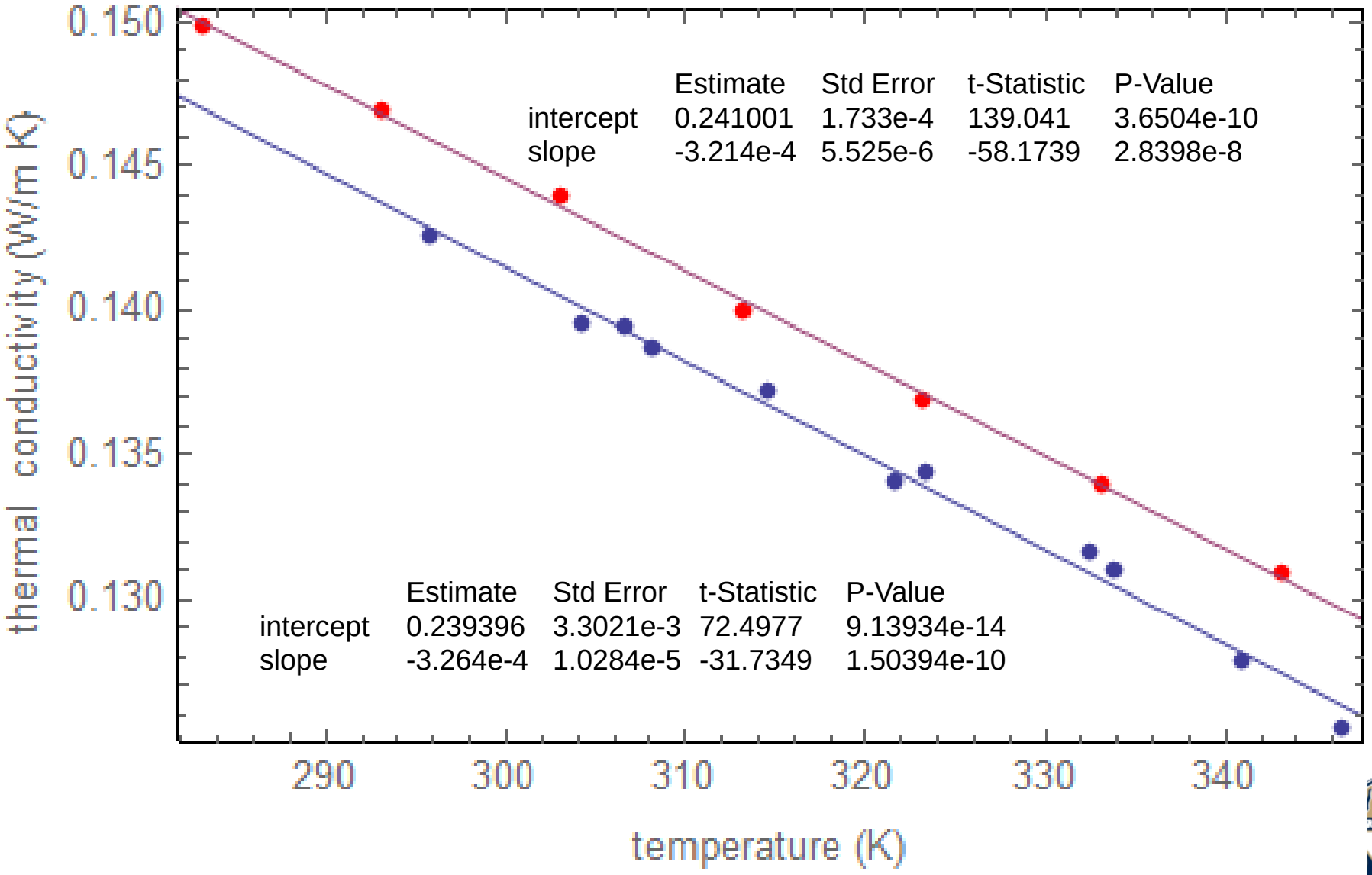
Conceptually understand the following for both linear and nonlinear models:

1. **Best fit to model parameters**
2. **Experimental error and its estimate**
3. **Prediction confidence bands**
4. **Single-point confidence bands**
5. **Parameter confidence regions**
6. **Experimental design**

Two typical datasets



Straight-line Regression



ANOVA and Conf. Int. Tables

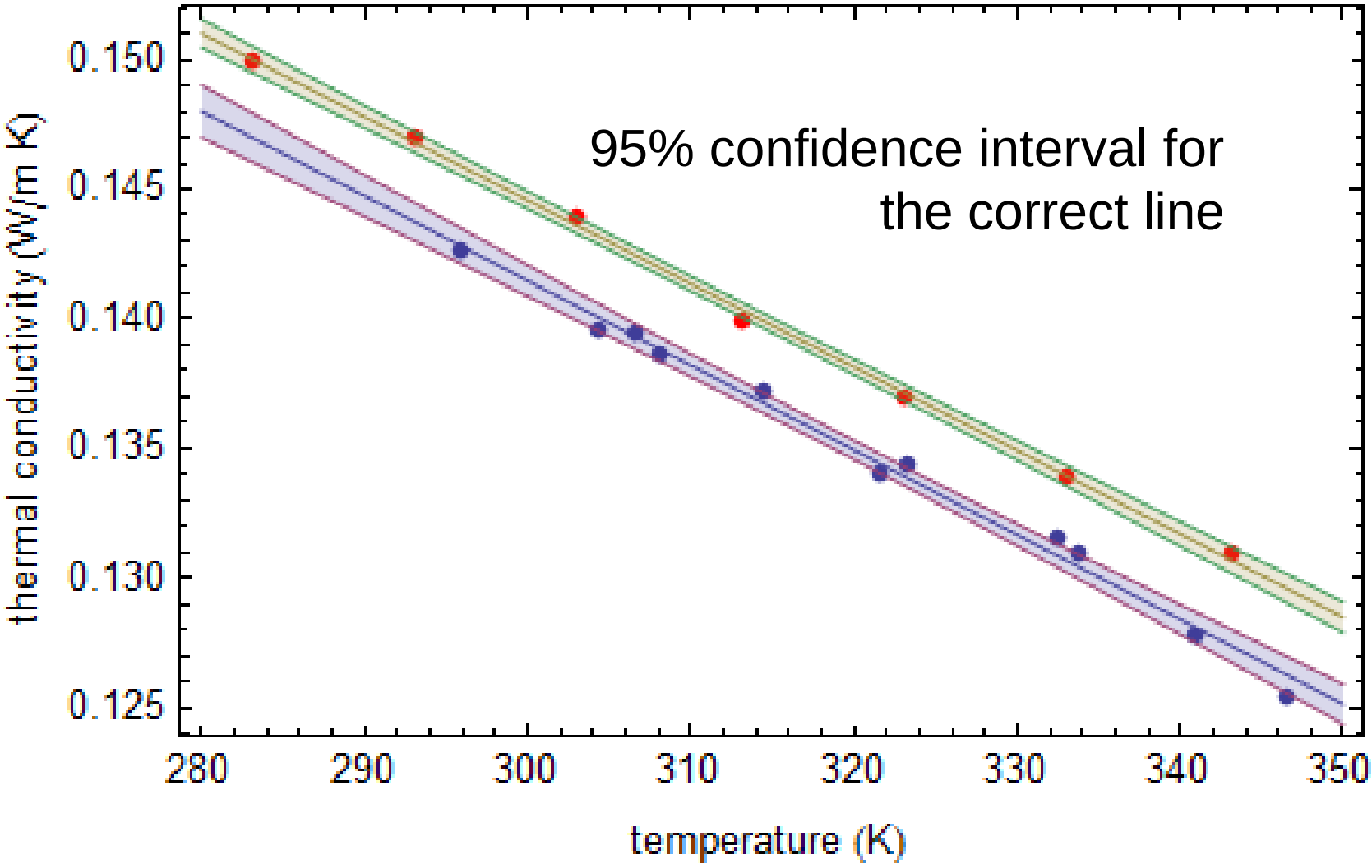
	DF	SS	MS	F-Statistic	P-Value
x	1	2.8515e-4	2.8515e-4	1007.1	1.50394e-10
Error	9	2.5483e-6	2.83142e-7		
Total	10	2.8770e-4			

	DF	SS	MS	F-Statistic	P-Value
x	1	2.89221e-4	2.89221e-4	3384.21	2.8398e-8
Error	5	4.27311e-7	8.54621e-8		
Total	6	2.89649e-4			

	Estimate	Standard Error	Confidence Interval
intercept	0.239396	0.00330212	{0.231926,0.246866}
slope	-3.26366e-4	1.02841e-5	{-3.49631e-4,-3.03102e-4}

	Estimate	Standard Error	Confidence Interval
intercept	0.241001	0.00173331	{0.236545,0.245457}
slope	-3.2139e-4	5.52469e-6	{-3.35595e-4,-3.07191e-4}

Prediction Bands



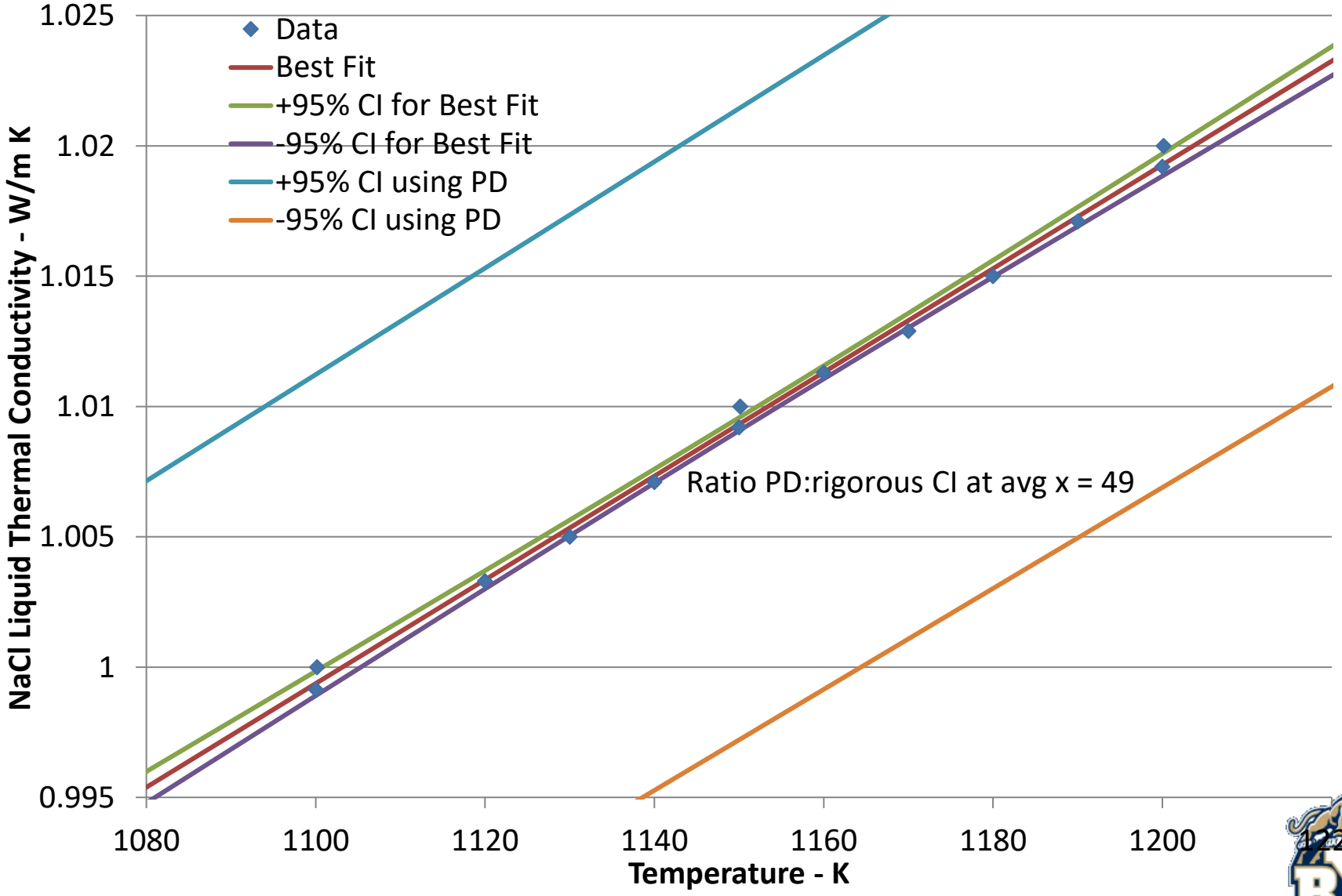
Propagation of Error

$$\sigma_f^2 = \sum_i \left(\left. \frac{\partial f}{\partial \beta_i} \right|_{\beta_i} \right)^2 \sigma_{\beta_i}^2 =$$

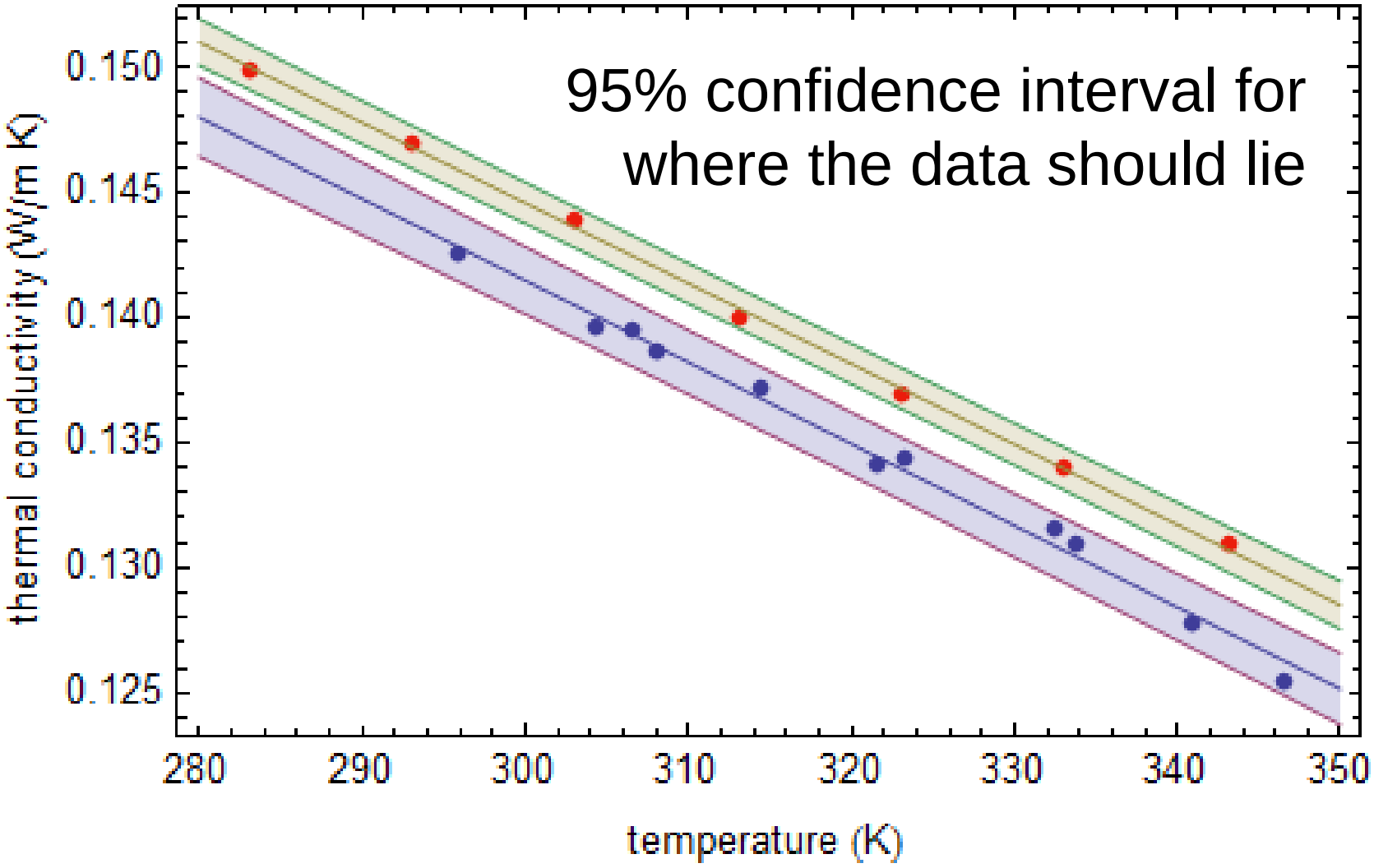
$$\left(\left. \frac{\partial (a + bx)}{\partial a} \right|_b \right)^2 \sigma_a^2 + \left(\left. \frac{\partial (a + bx)}{\partial b} \right|_a \right)^2 \sigma_b^2 =$$

$$\sigma_a^2 + x^2 \sigma_b^2$$

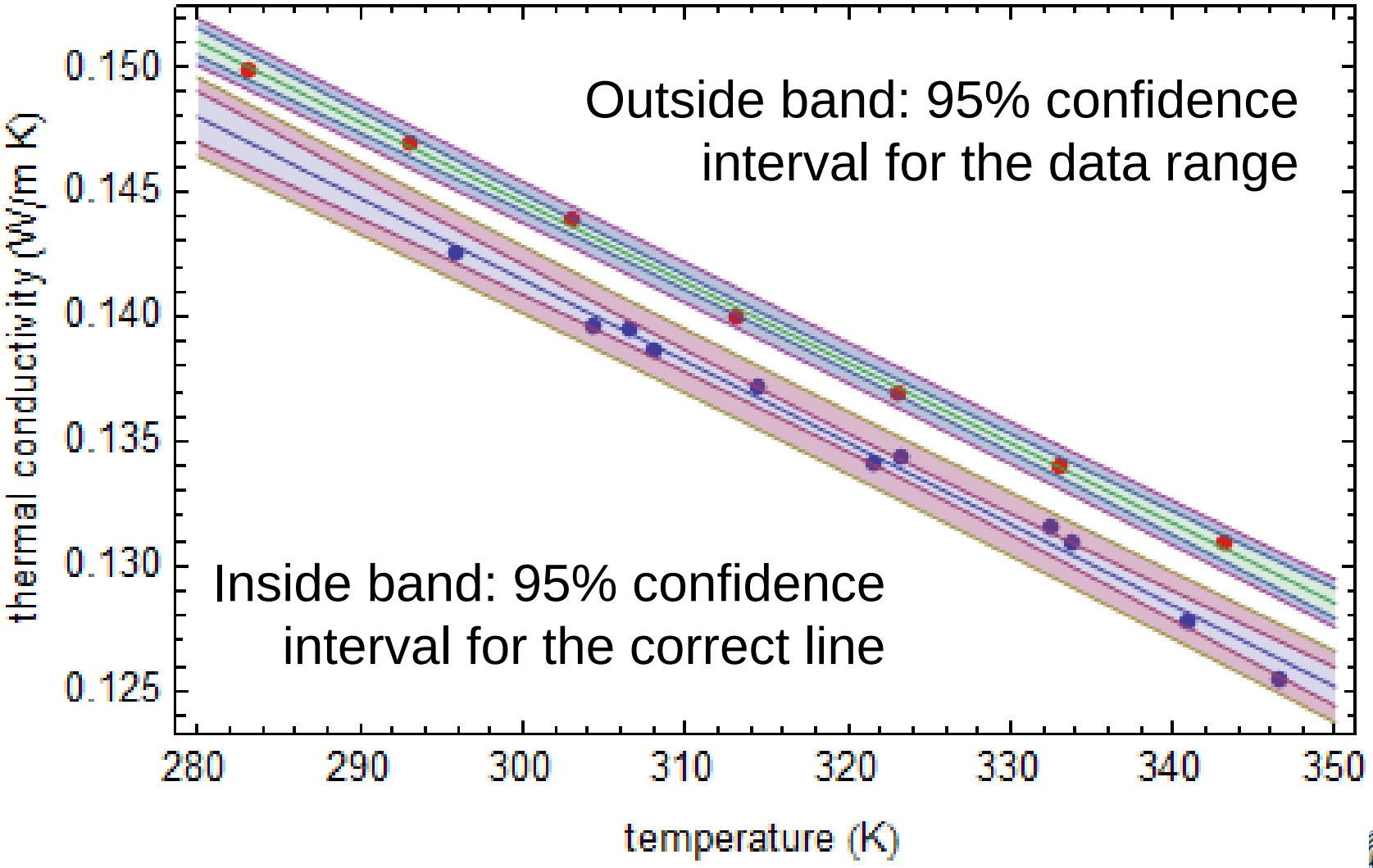
Rigorous vs. Propagation of Error



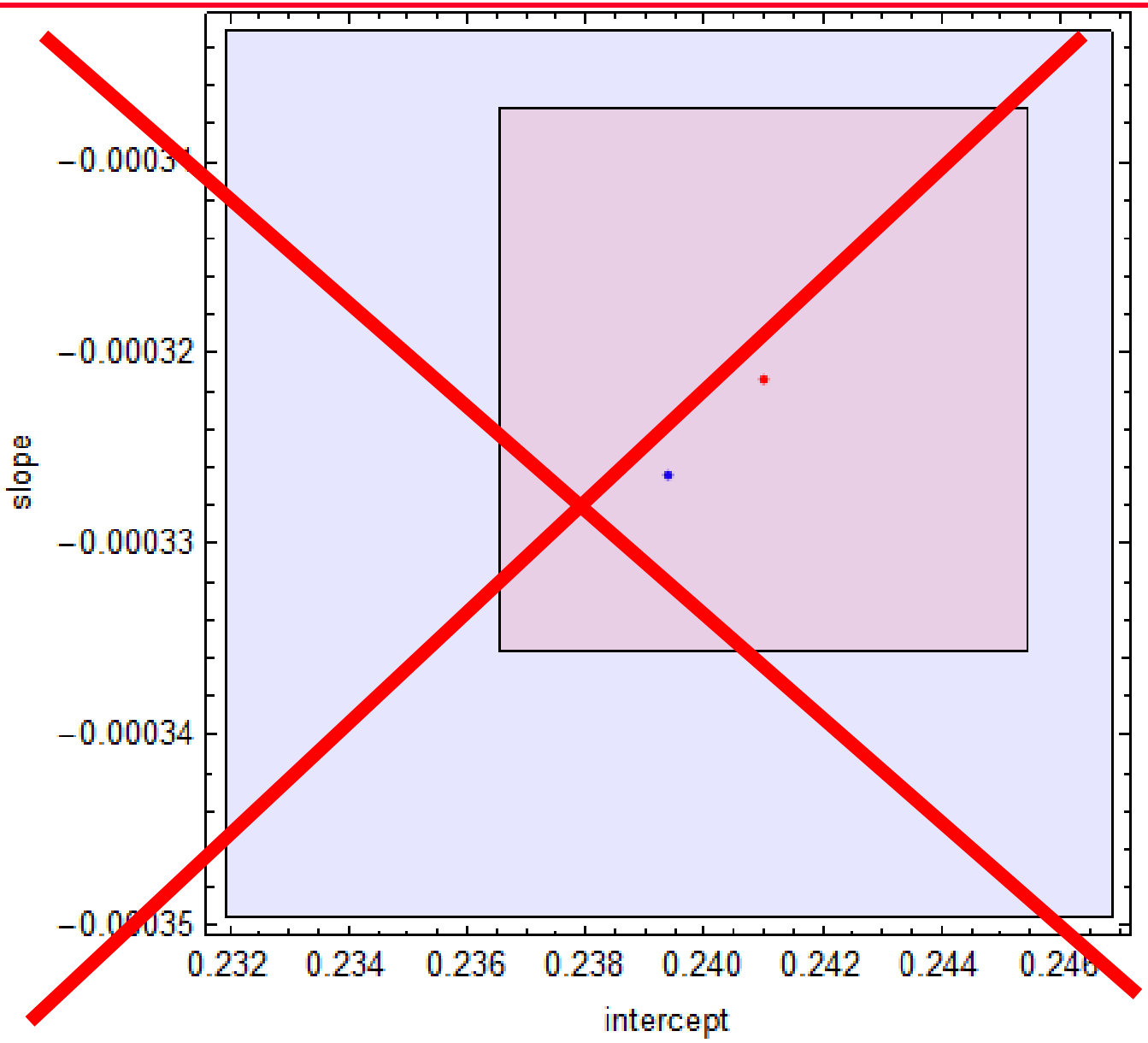
Single-Point Prediction Bands



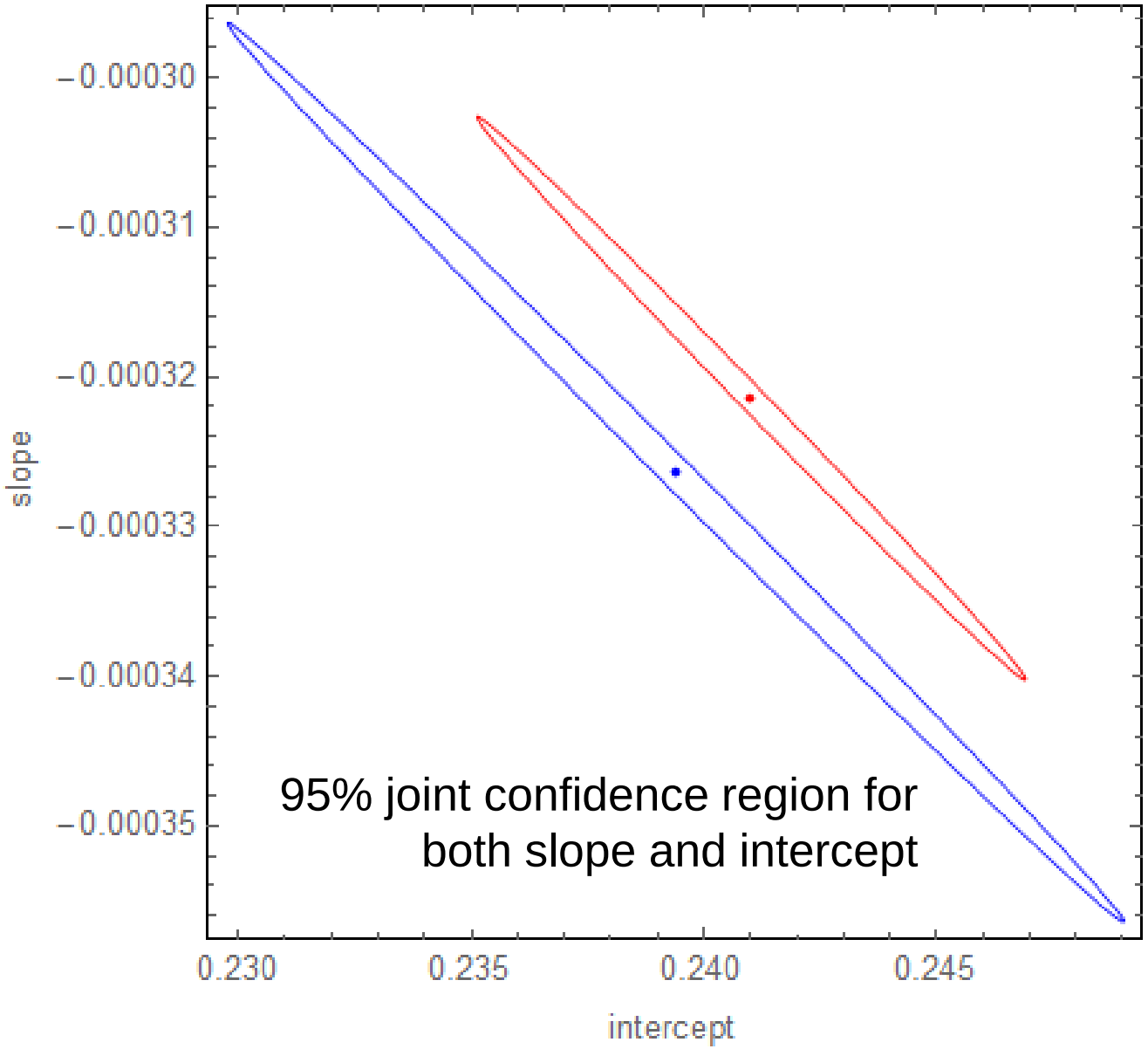
Point and Mean Prediction Bands



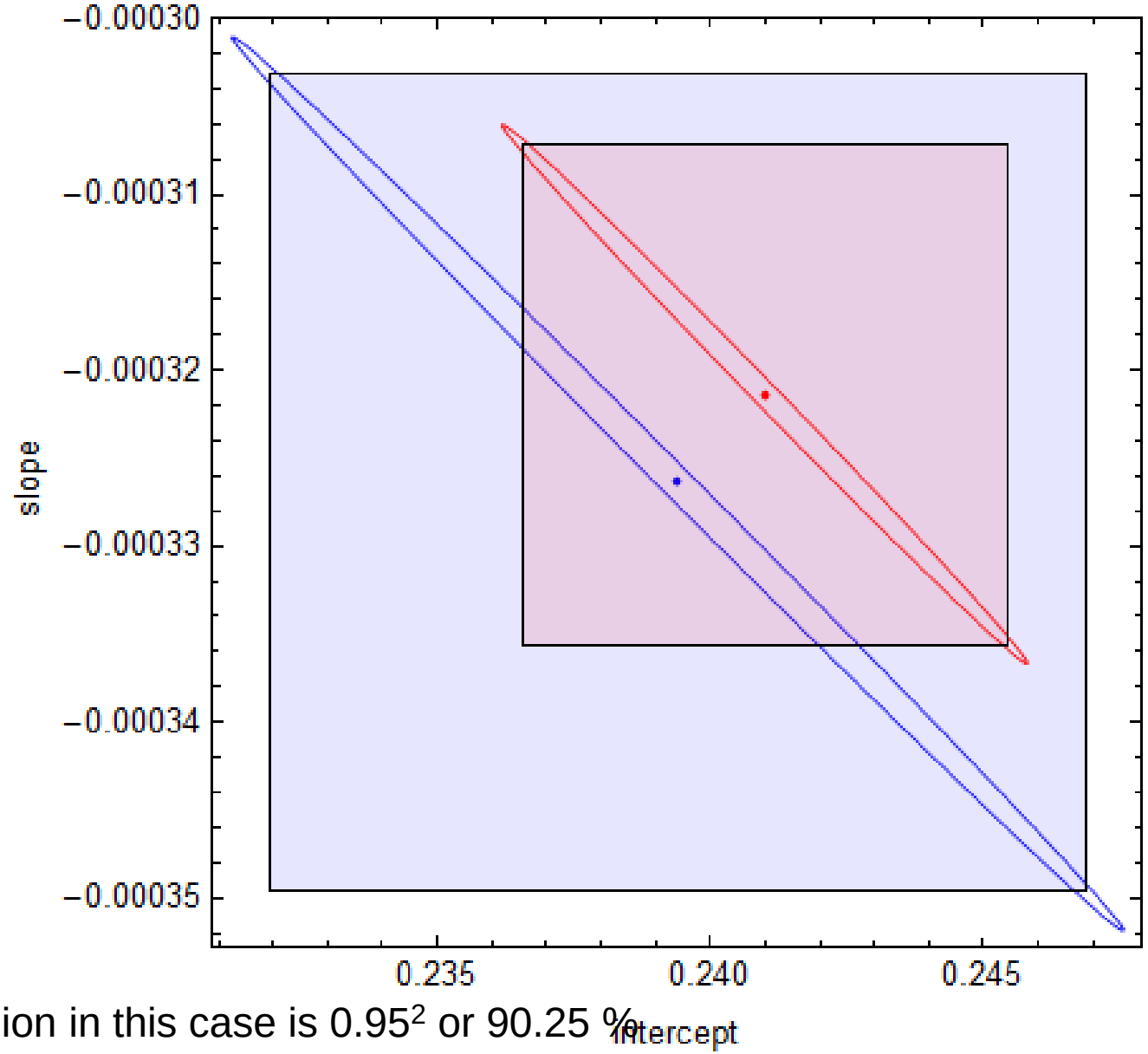
Confidence Interval Ranges



Joint Confidence Region

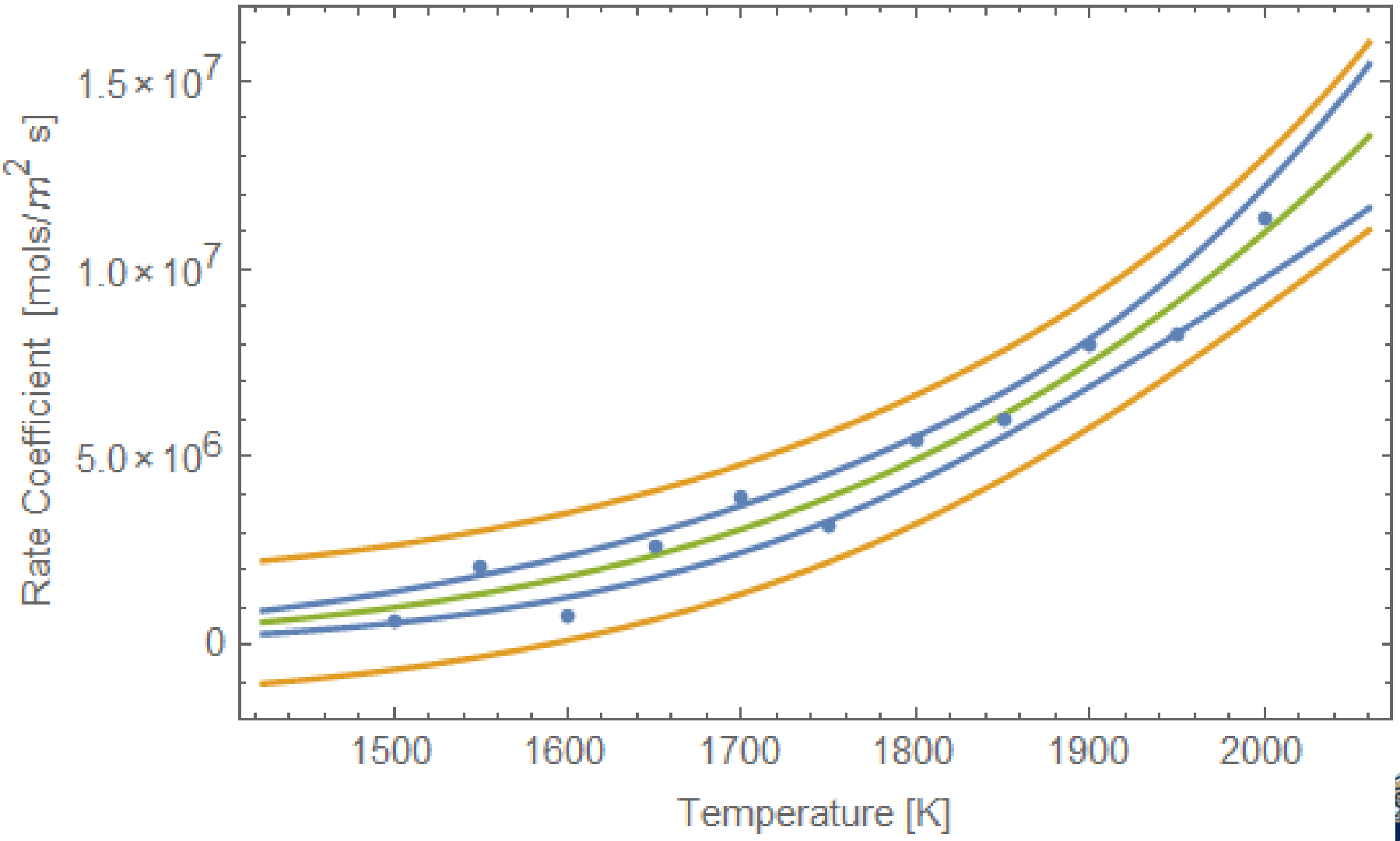


Regions and Intervals Compared

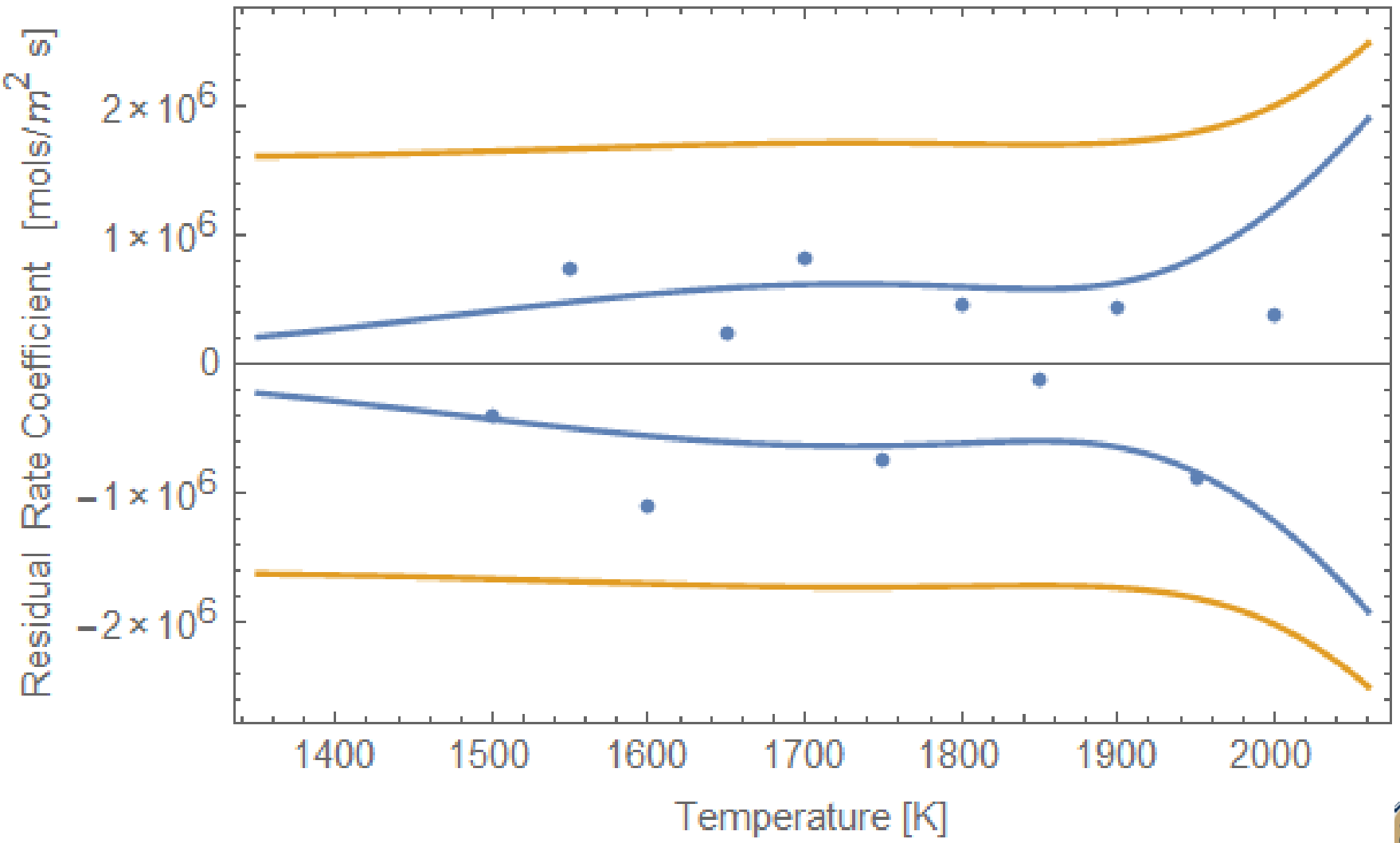


Joint region in this case is 0.95^2 or 90.25 %

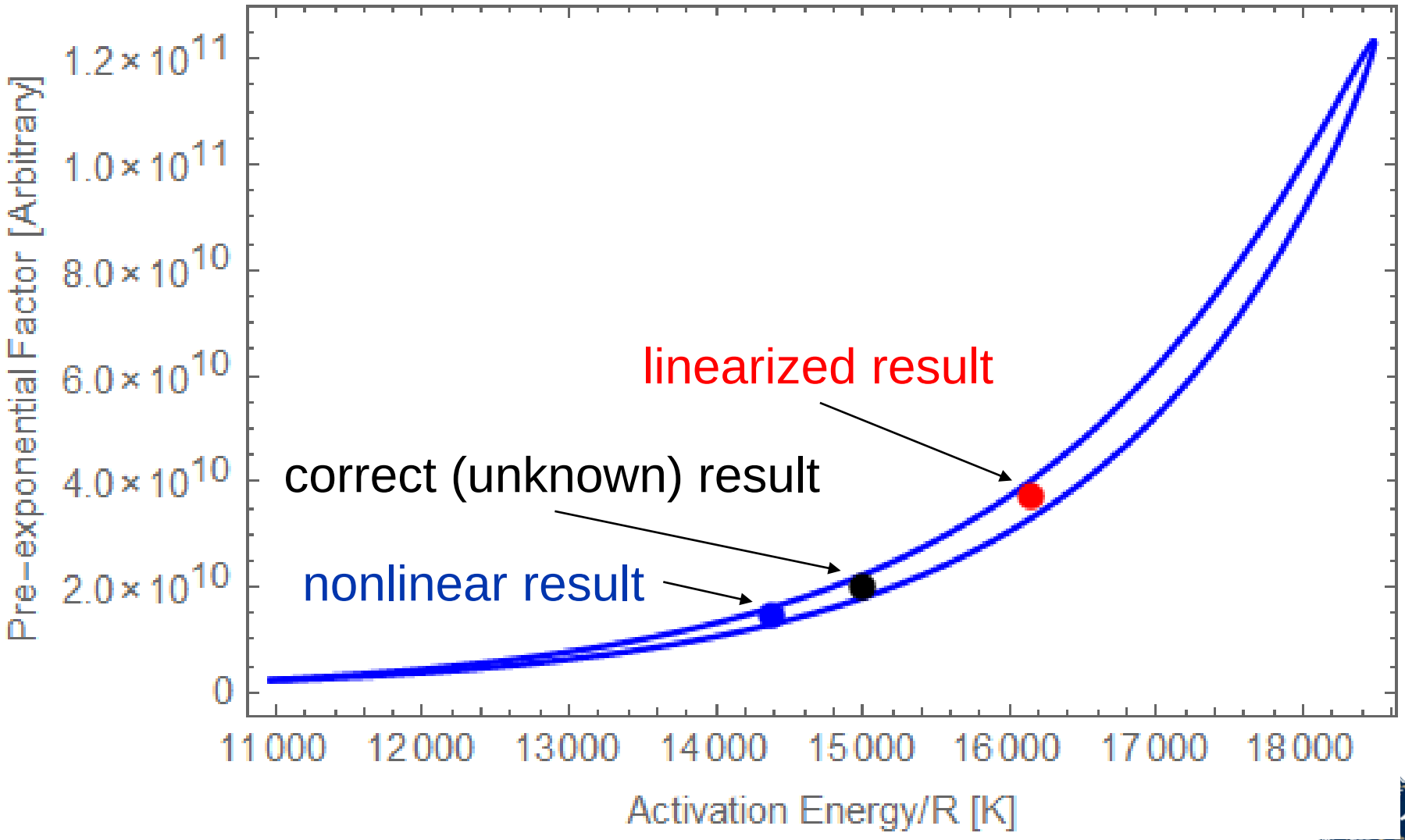
Single-Point Prediction Band



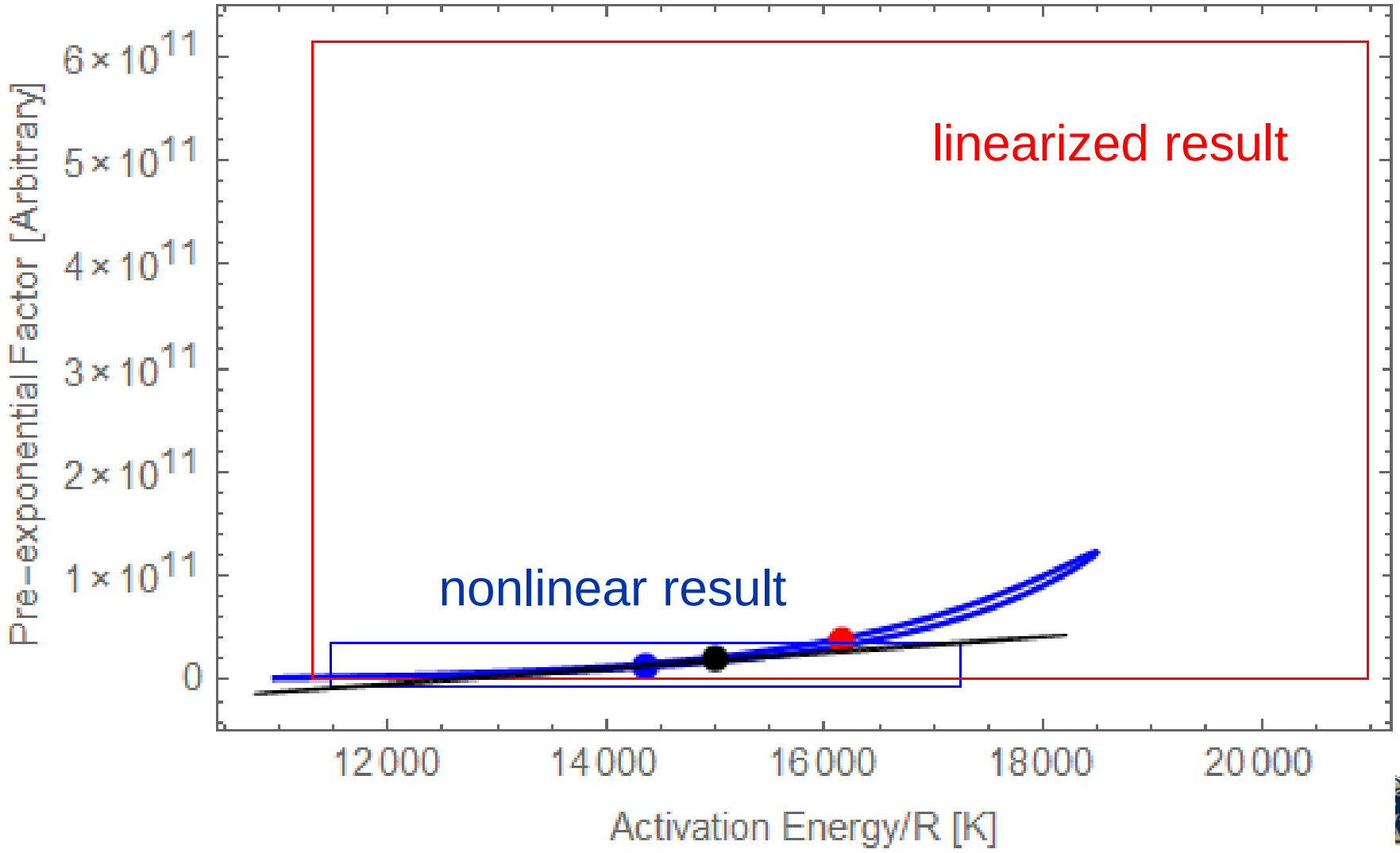
Residual SP Prediction Band (95%)



Joint Confidence Region



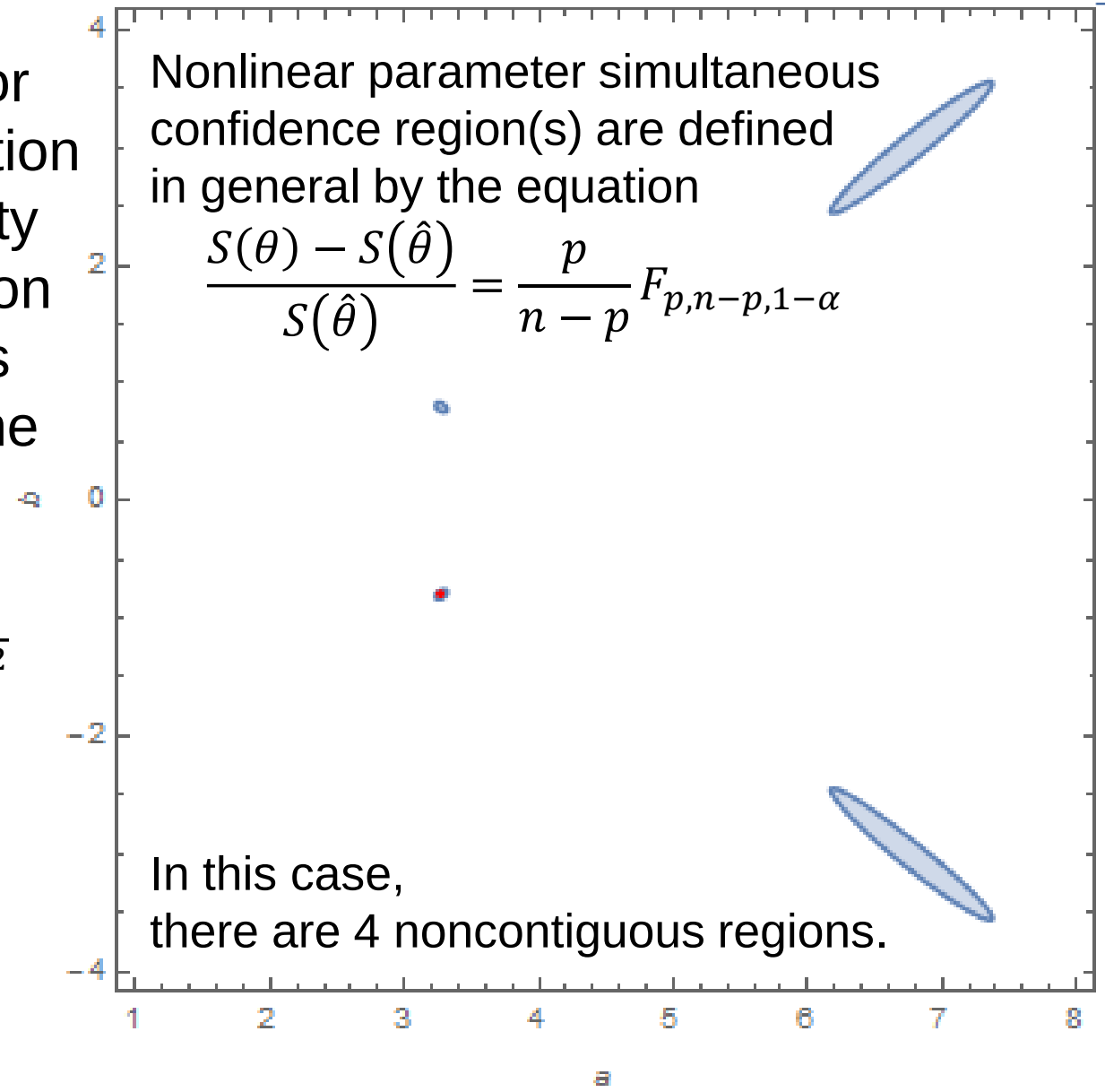
Joint Confidence Regions



Nonlinear SCR More Complex

The Cauchy or Lorentz equation (is a probability density function and describes some laser line widths).

$$y = \frac{1}{1 + \frac{(x-a)^2}{b^2}}$$

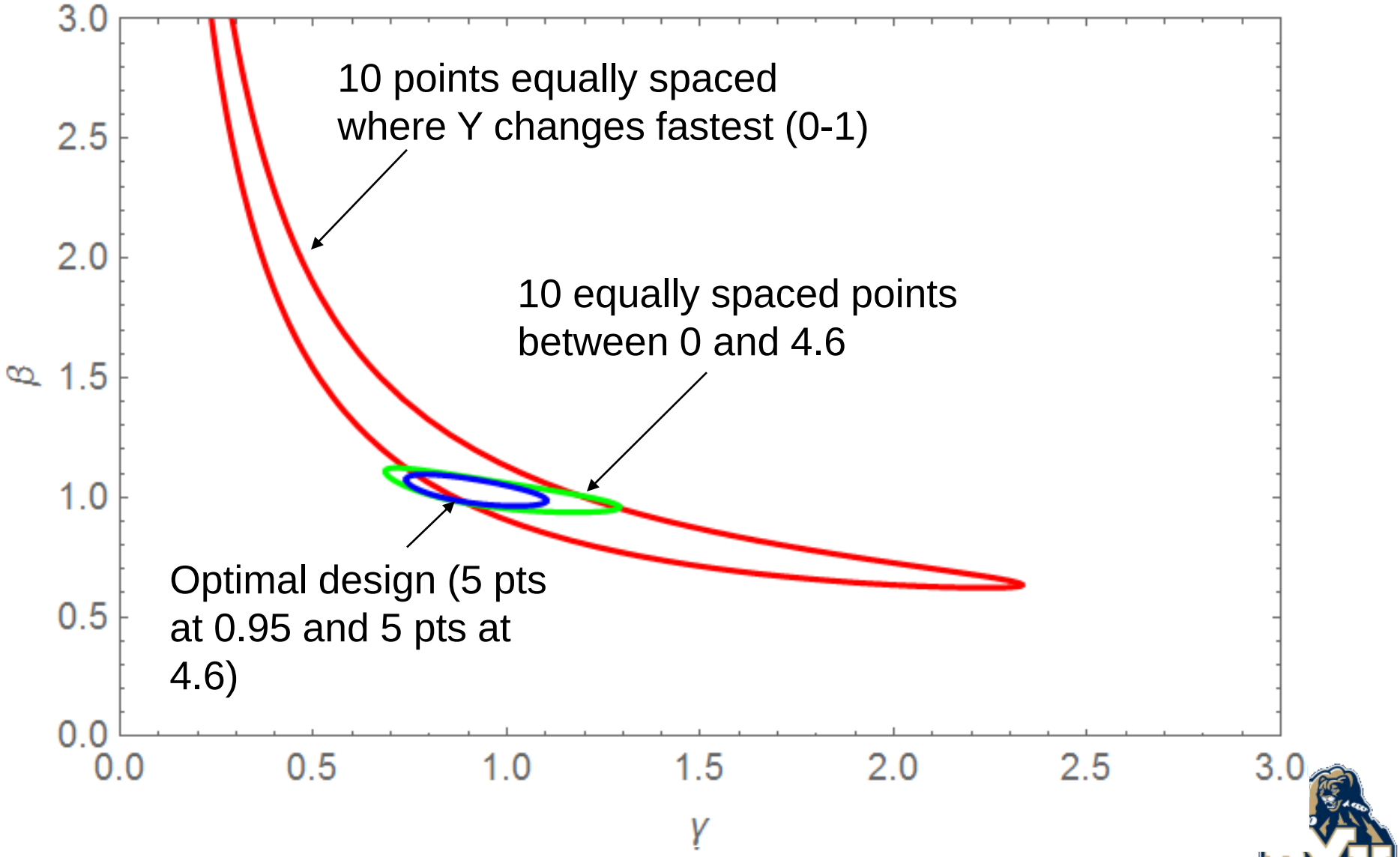


Example

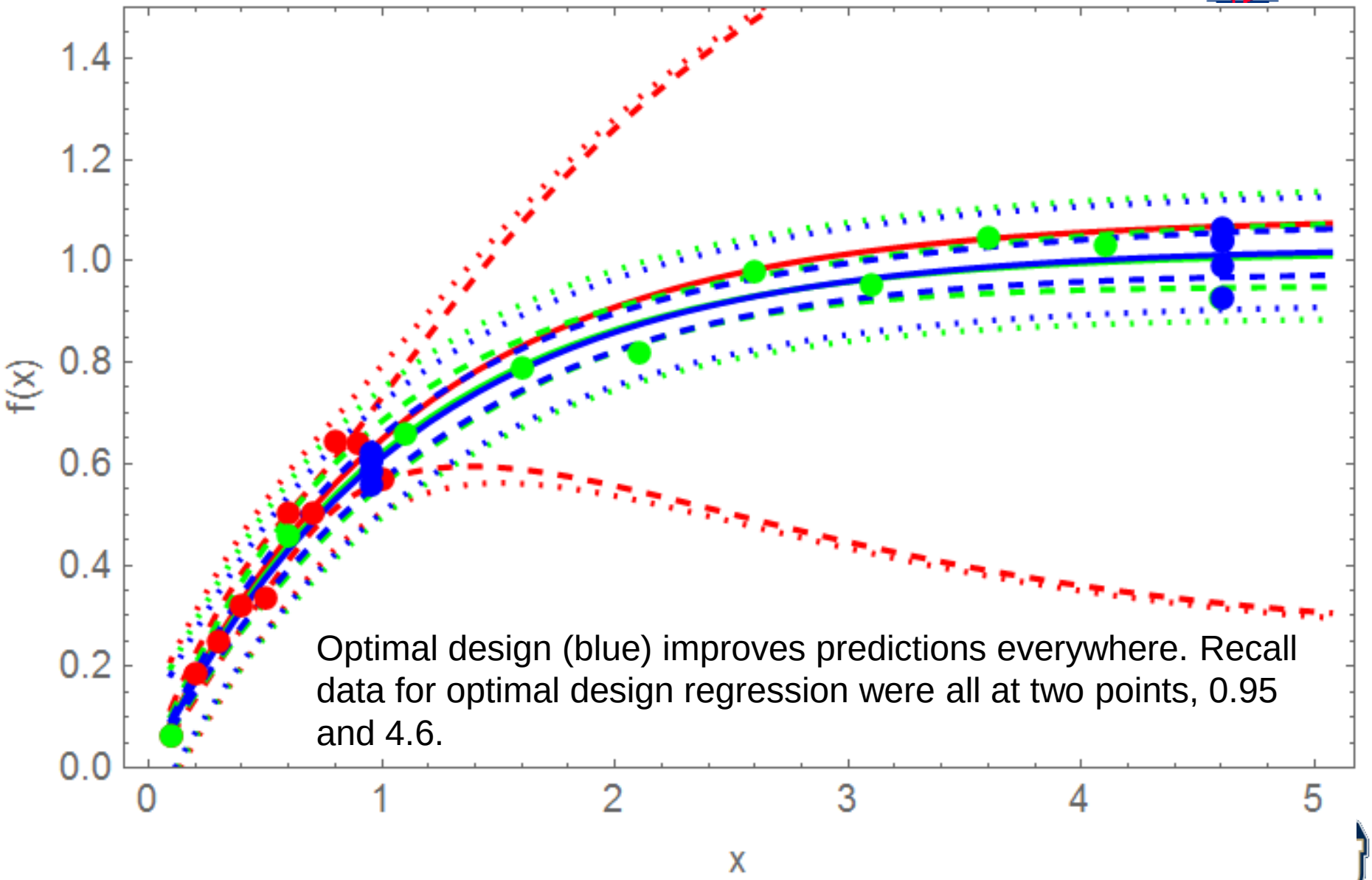
Compare 3 different design, each with the same number of data points and the same errors

Point No.	x_1	y_1	x_2	y_2	x_3	y_3	ϵ_i
1	0.1	0.0638	0.1	0.0638	0.954	0.5833	0.03150
2	0.2	0.1870	0.6	0.4569	0.954	0.6203	-0.00550
3	0.3	0.2495	1.1	0.6574	0.954	0.6049	0.00990
4	0.4	0.3207	1.6	0.7891	0.954	0.6056	0.00920
5	0.5	0.3356	2.1	0.8197	0.954	0.5567	0.05810
6	0.6	0.5040	2.6	0.9786	4.605	1.0428	-0.05280
7	0.7	0.5030	3.1	0.9545	4.605	0.9896	0.00040
8	0.8	0.6421	3.6	1.0461	4.605	1.0634	-0.07340
9	0.9	0.6412	4.1	1.0312	4.605	1.0377	-0.04770
10	1.0	0.5678	4.6	0.9256	4.605	0.9257	0.06430

SCR Results for 3 Cases

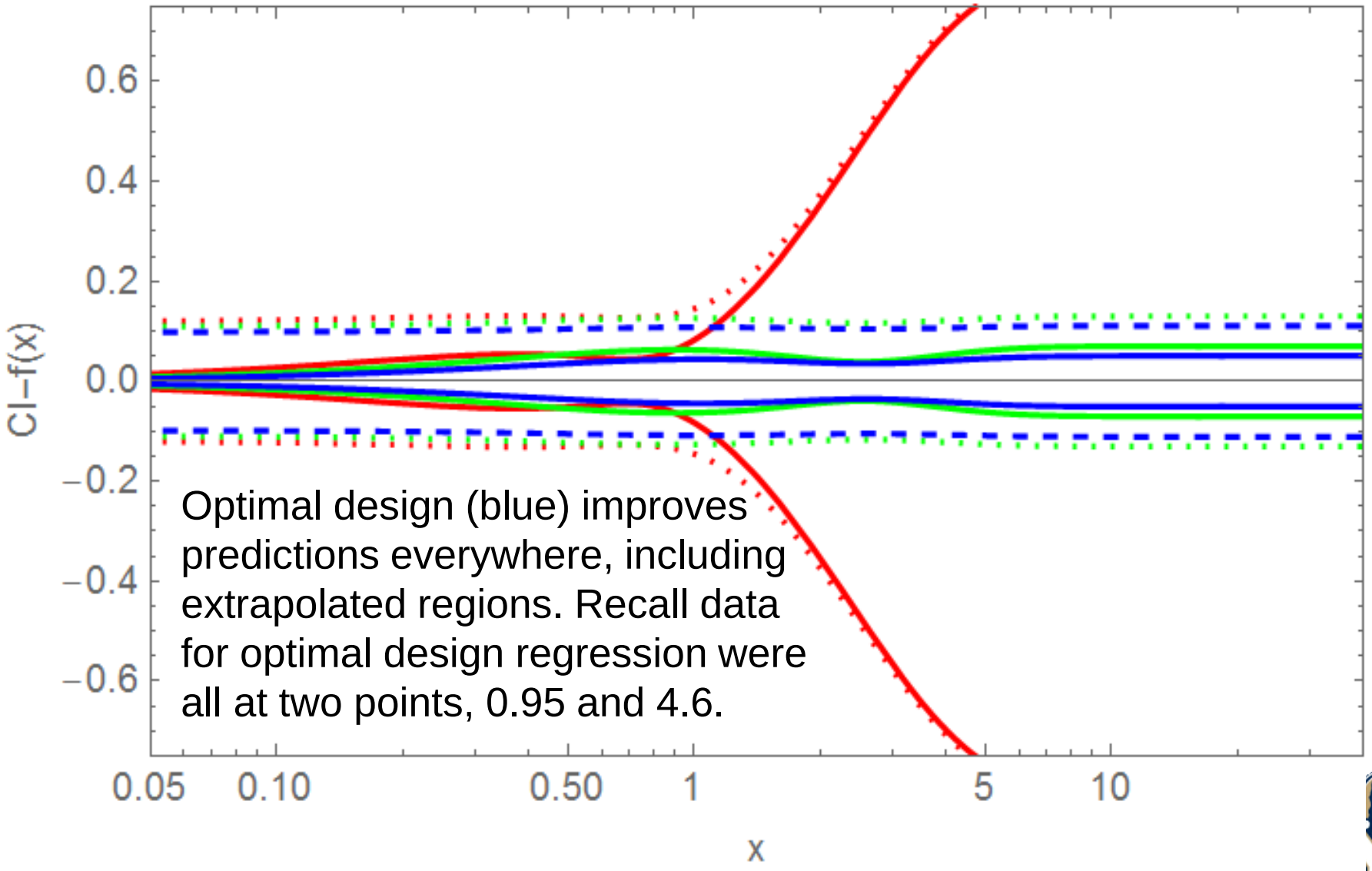


Prediction and SP Conf. Intervals



Optimal design (blue) improves predictions everywhere. Recall data for optimal design regression were all at two points, 0.95 and 4.6.

Improved Prediction & SP Intervals



Optimal design (blue) improves predictions everywhere, including extrapolated regions. Recall data for optimal design regression were all at two points, 0.95 and 4.6.

Linear vs. Nonlinear Models

- **Linear and nonlinear refer to the coefficients, not the forms of the independent variable.**
- **The derivative of a linear model with respect to a parameter does not depend on any parameters.**
- **The derivative of a nonlinear model with respect to a parameter depends on one or more of the parameters.**

Linear vs. Nonlinear Models

Linear Models

$$c_p(T) = a + bT + cT^2 + \cdots$$

$$f(x,y) = a + b\sin x + c\,\mathrm{erf}\,y$$

$$D_A(T) = a\left(\frac{T}{T_0}\right)^n \quad n \text{ and } T_0 \text{ known}$$

$$y(x) = a\log x$$

$$y(x) = \frac{a}{x} + b\sinh x$$

$$s(T,p) = c_p \ln\frac{T}{T_{ref}} - R\ln\frac{p}{p_{ref}}$$

$$h(T) = a\,T + \frac{b}{2}\,T^2 + \frac{c}{3}T^3 + \cdots + h_f^0$$

in general

$$y(x_i) = a\,f_1(x_1) + b\,f_2(x_2) + \cdots$$

Nonlinear Models

$$c_p = a + b(T-c) + d(T-c)^2 + \cdots$$

$$f(x) = a\sin bx$$

$$D_A(T) = a\left(\frac{T}{T_0}\right)^b \quad b \text{ or } T_0 \text{ unknown}$$

$$y = a\log\frac{x}{b}$$

$$k = A\exp\left(-\frac{E}{RT}\right)$$

$$c_p(T) = a + b\left(\frac{\frac{c}{T}}{\sinh\frac{c}{T}}\right)^2 + d\left(\frac{\frac{e}{T}}{\cosh\frac{e}{T}}\right)^2$$

$$h(T) = aT + bc\coth\frac{c}{T} - de\tanh\frac{e}{T} + h_f^0$$

$$\ln\gamma_1 = x_2^2\left[\tau_{21}\left(\frac{G_{21}}{x_1 + x_2G_{21}}\right)^2 + \frac{\tau_{12}G_{12}}{(x_2 + x_1G_{12})^2}\right]$$

no general expression

LS with data, not transformed data

$$S(\theta) = \sum_i \left(y_i - f(x_i;\theta)\right)^2$$

Best estimates of model parameters

$$S(\hat{\theta}) = \min \sum_i \left(y_i - f(x_i;\theta)\right)^2 = \sum_i (y_i - \hat{y}_i)^2$$

Parameter joint confidence region

$$\frac{S(\theta) - S(\hat{\theta})}{S(\hat{\theta})} \leq \frac{p}{n-p} F_{p,n-p,1-\alpha}$$

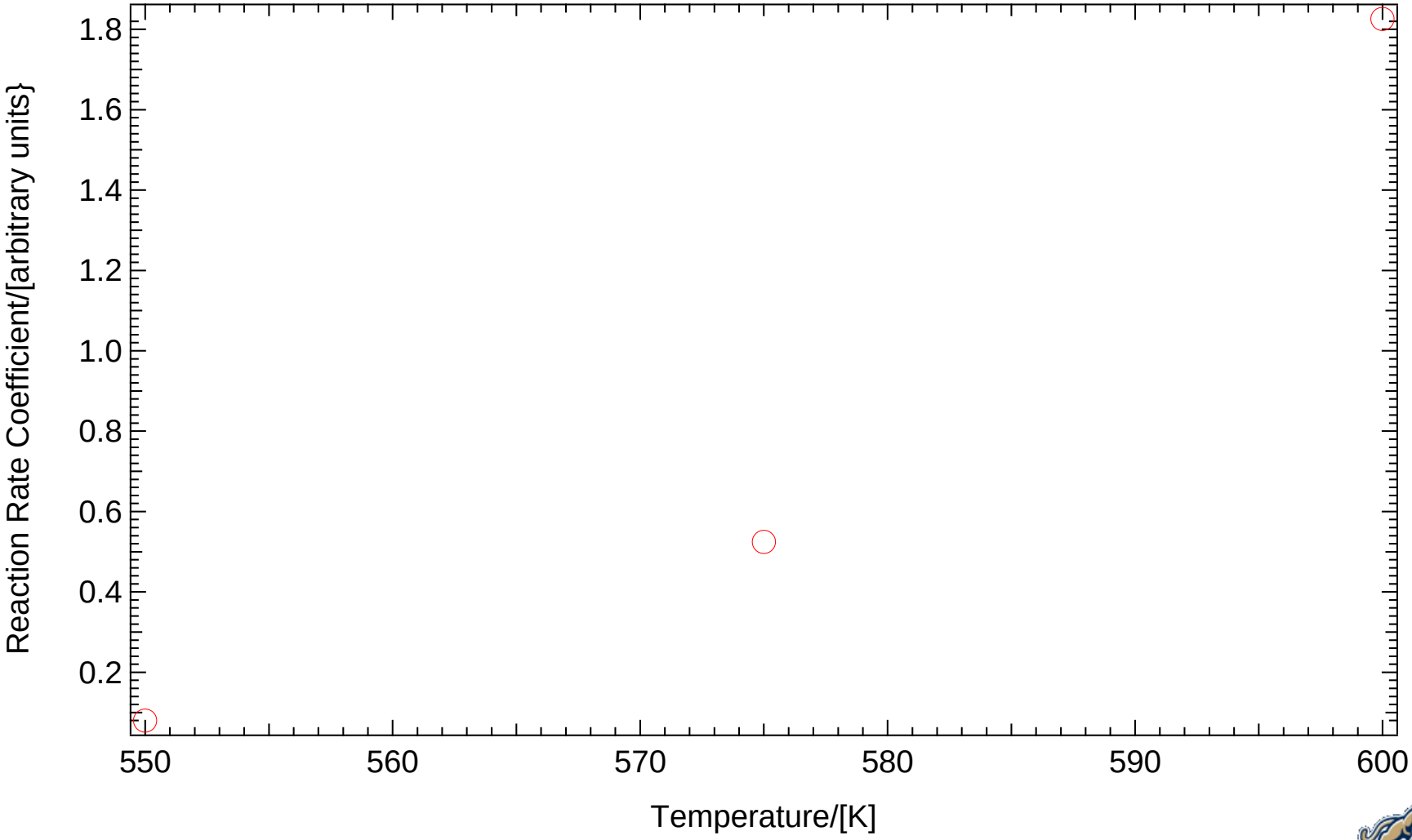
p = number of parameters, n = number of points, F = F-distribution

Recommendations

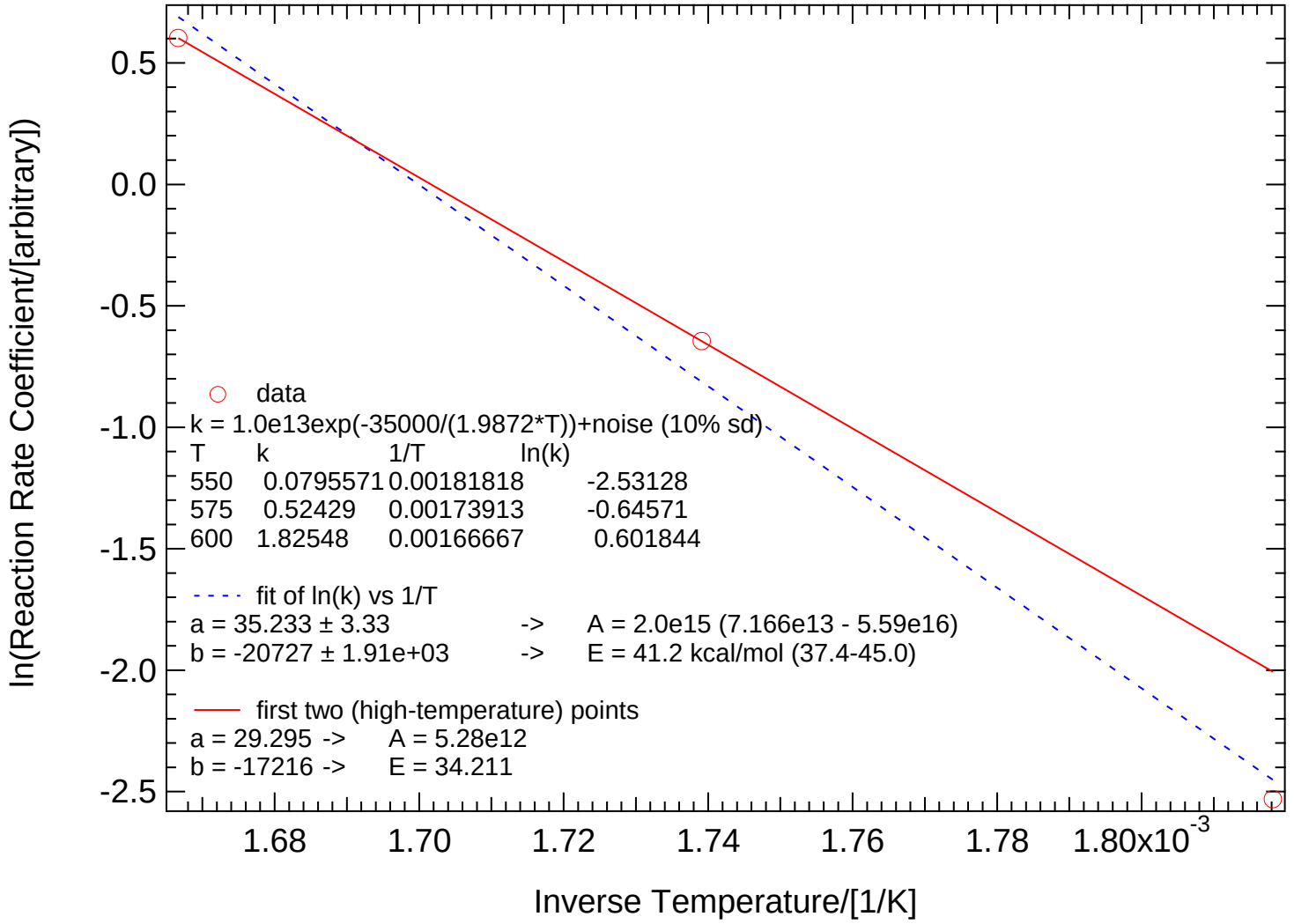
- **Minimize to sum of squares of differences between measurements and model written in term of what you measured.**
- **DO NOT linearize the model, i.e., make it look something like a straight line model.**
- **Confidence intervals for parameters can be misleading.**
- **Joint/simultaneous confidence regions are much more reliable.**
- **Propagation of error formula grossly overestimates error**
- **Compute joint/simultaneous confidence regions from**

$$\frac{S(\theta) - S(\hat{\theta})}{S(\hat{\theta})} \leq \frac{p}{n-p} F_{p,n-p,1-\alpha}$$

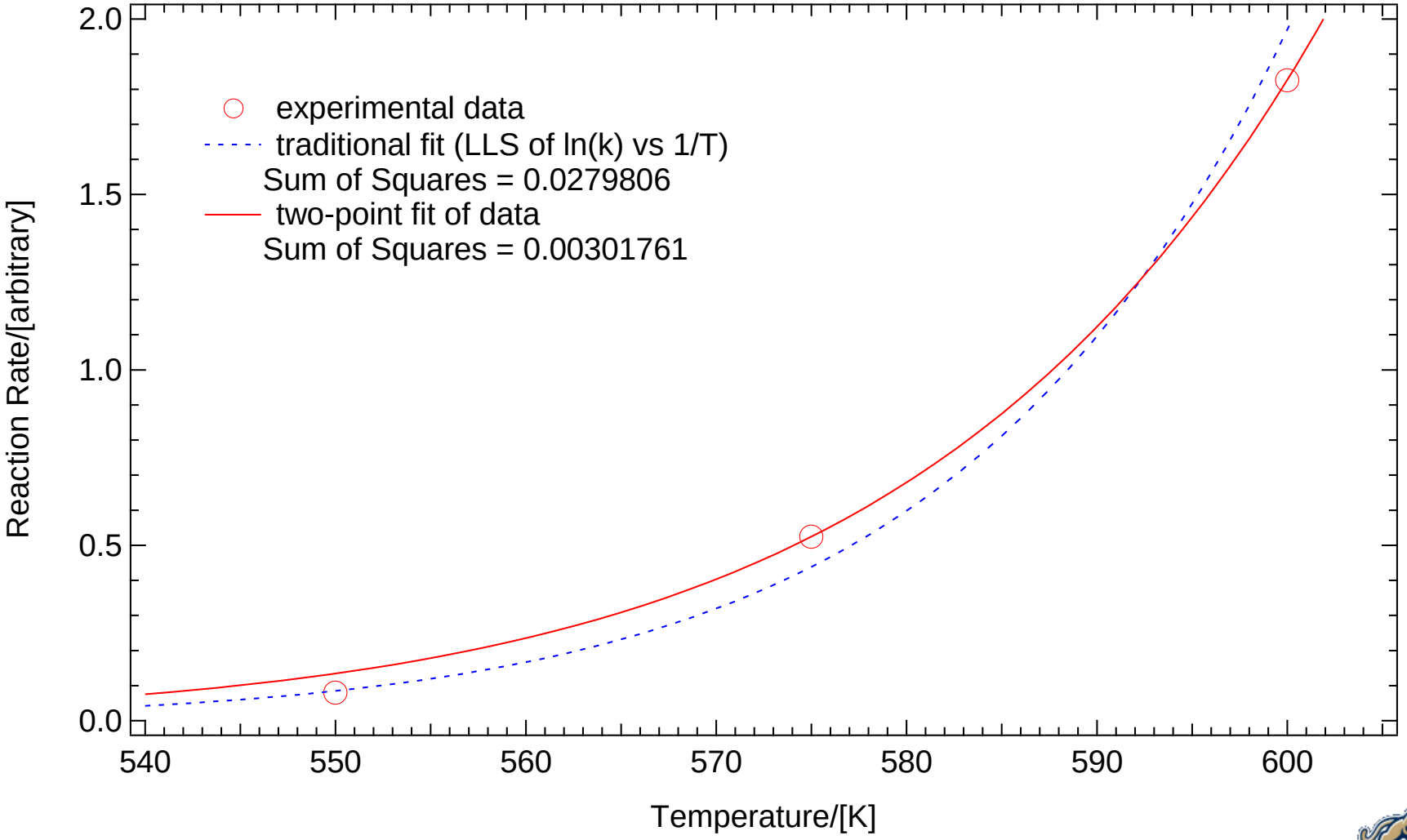
Typical Data



Kinetic Data Analysis

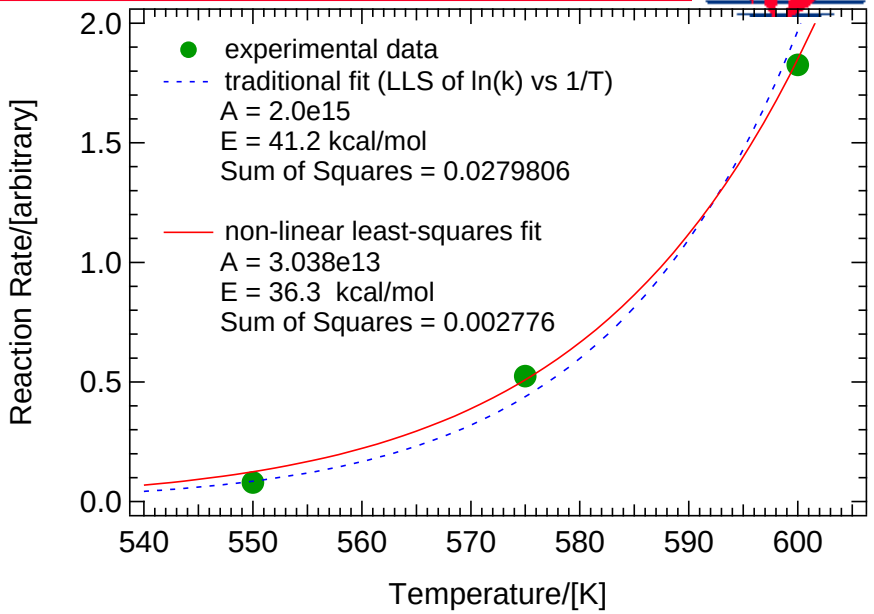


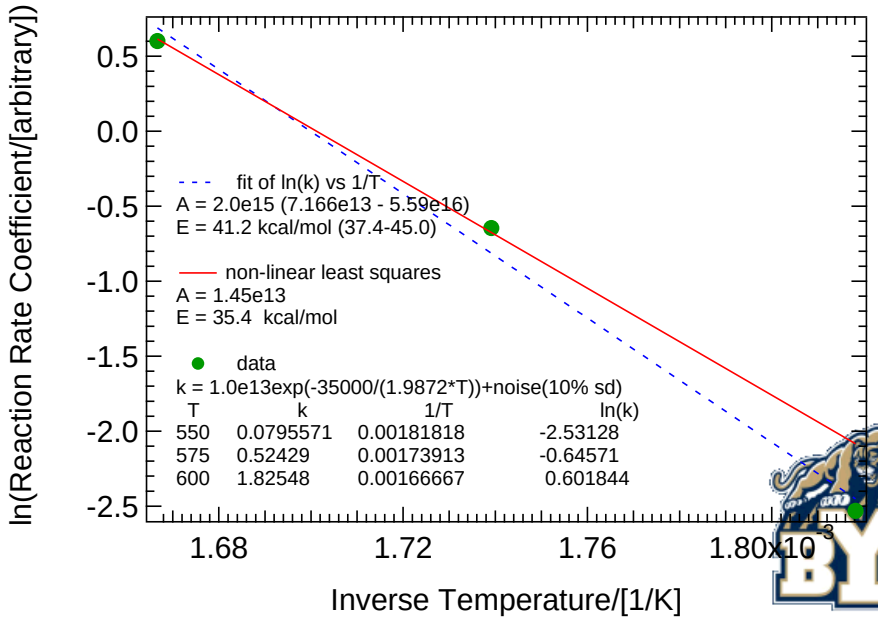
Kinetic Data Analysis



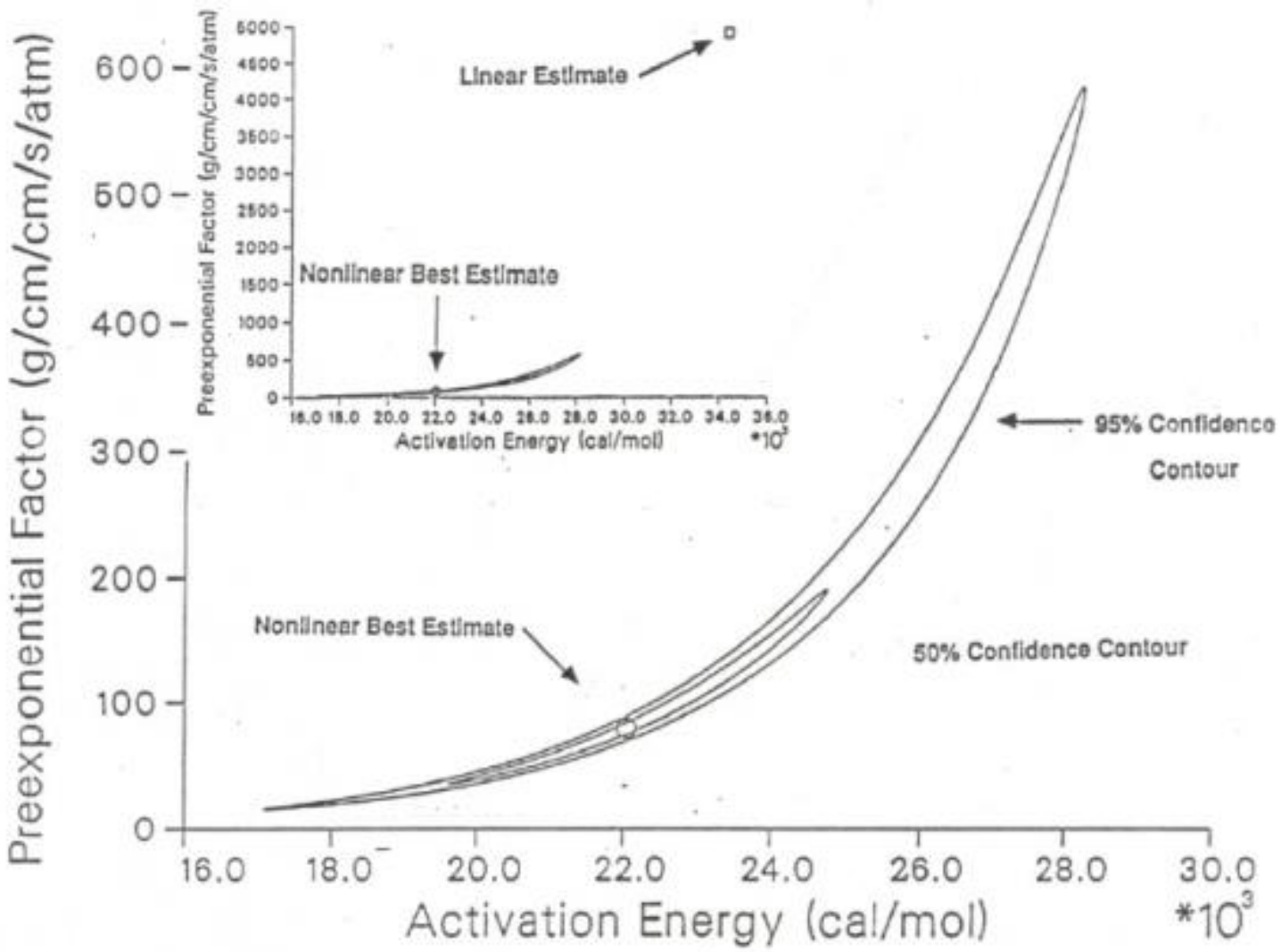
Graphical Summary

- **The linear and non-linear analyses are compared to the original data both as k vs. T and as ln(k) vs. 1/T.**
- **As seen in the upper graph, the linearized analysis fits the low-temperature data well, at the cost of poorer fits of high temperature results. The non-linear analysis does a more uniform job of distributing the lack of fit.**
- **As seen in the lower graph, the linearized analysis evenly distributes errors in log space**

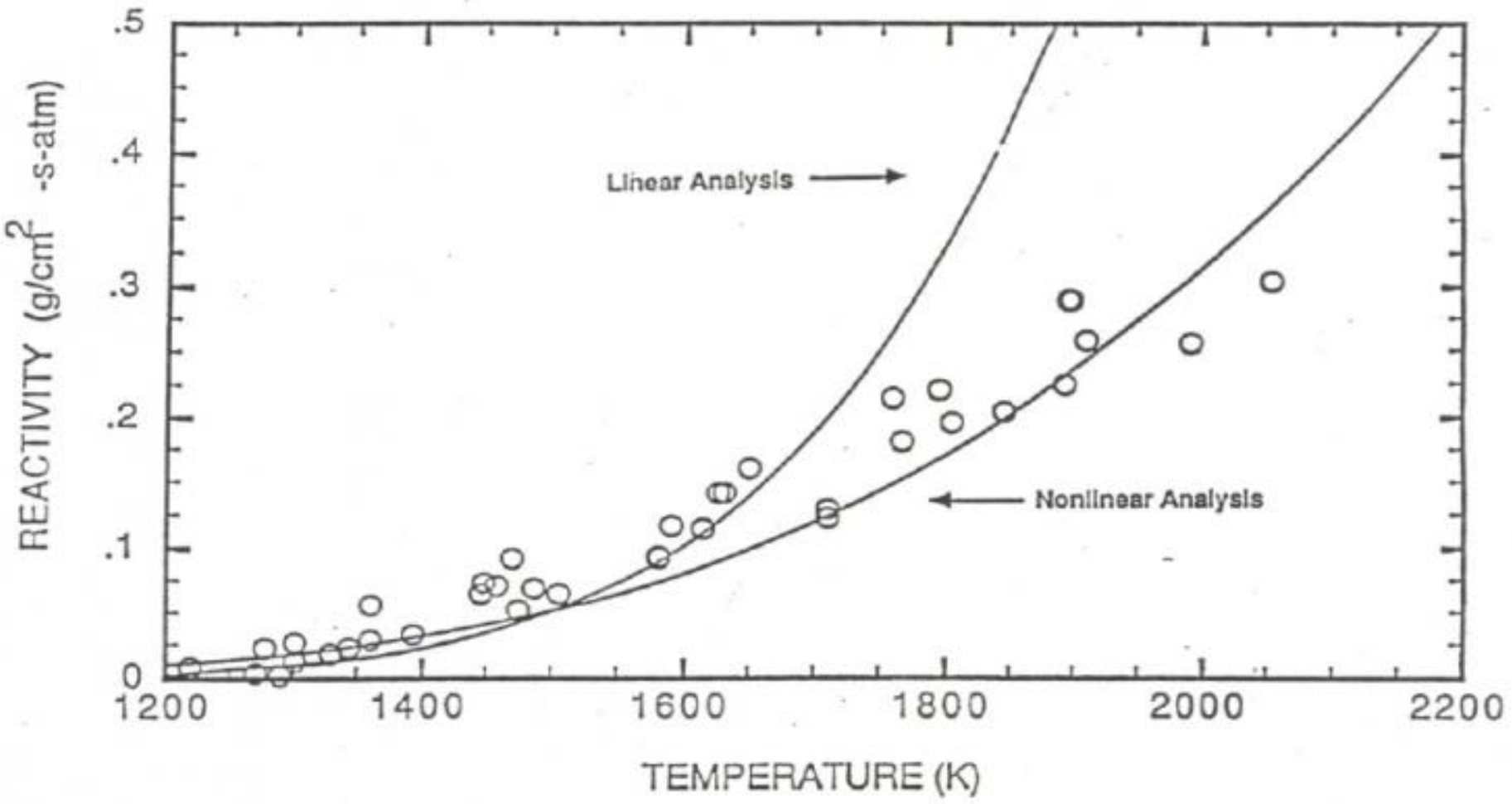




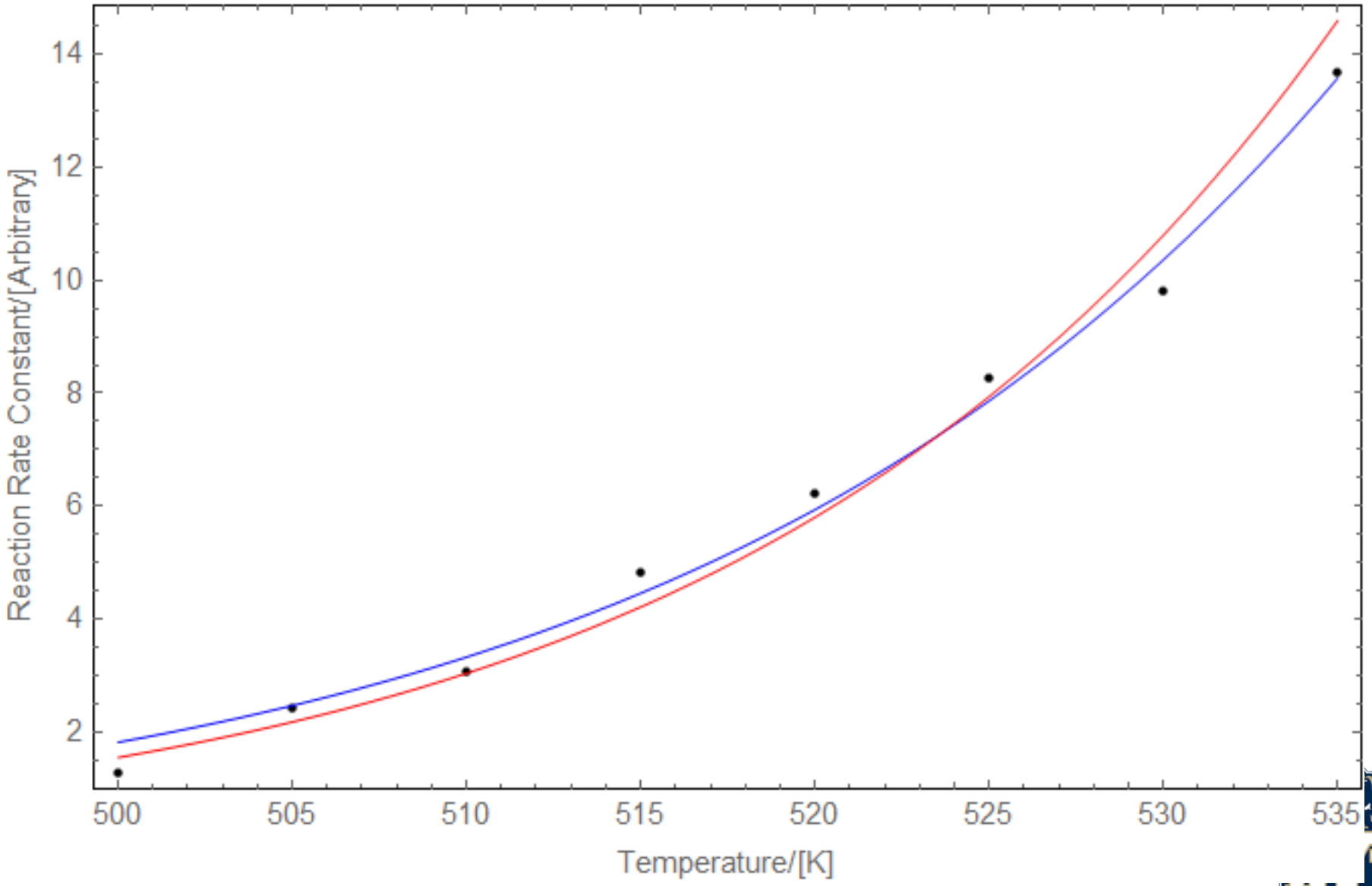
A practical Illustration



Extension


REACTIVITY (g/cm^2 -s-atm)
.5
.4
.3
.2
.1
0
1200
1400
1600
1800
2000
2200
TEMPERATURE (K)
Linear Analysis
Nonlinear Analysis

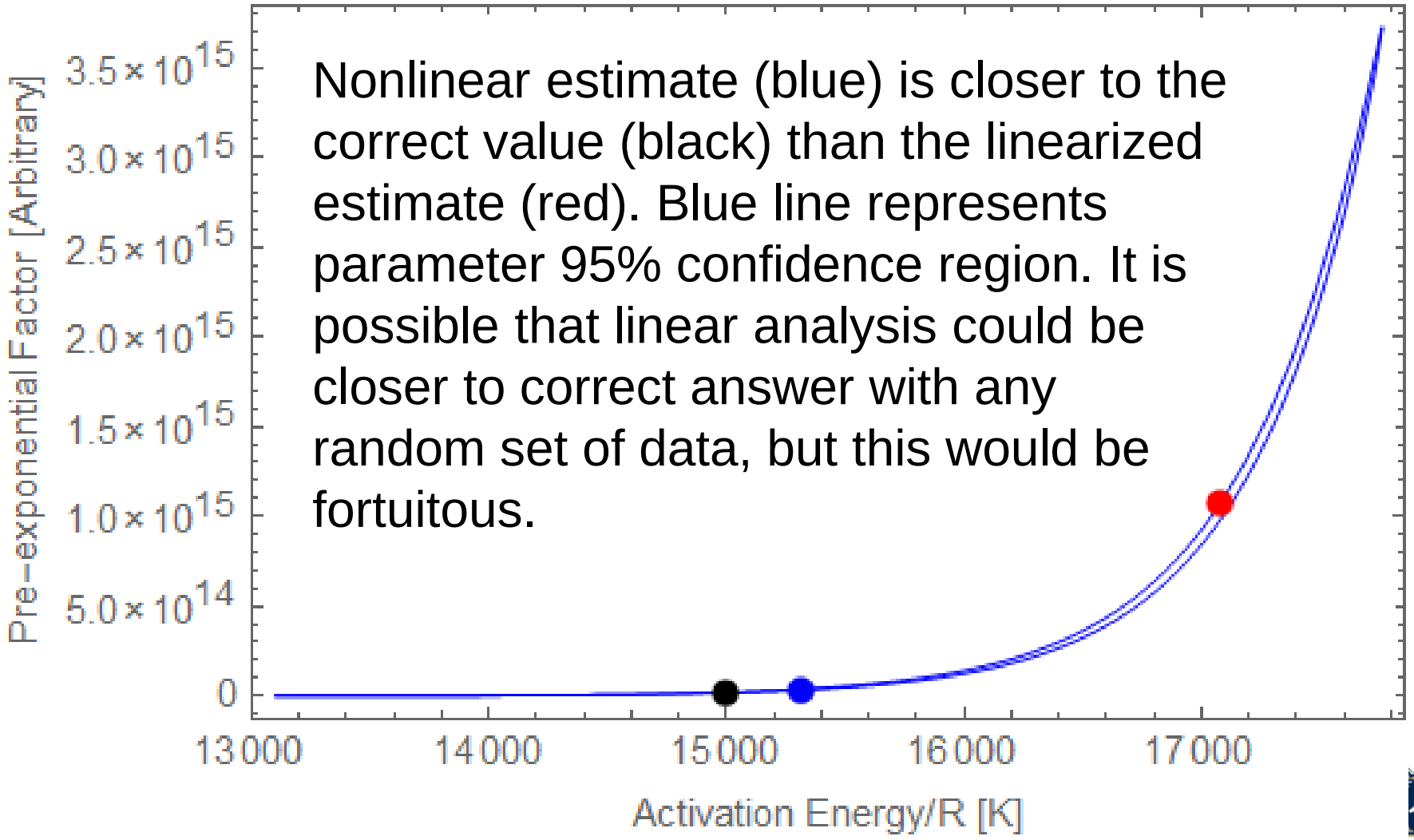
Typical Data



Parameter Estimates

- **Best estimate of parameters for a given set of data.**
- **Linear Equations**
  - **Explicit equations**
    - **Requires no initial guess**
    - **Depends only on measured values of dependent and independent variables**
    - **Does not depend on values of any other parameters**
- **Nonlinear Equations**
  - **Implicit equations**
    - **Requires initial guess**
    - **Convergence often difficult**
    - **Depends on data and on parameters**

Parameter Estimates



Nonlinear estimate (blue) is closer to the correct value (black) than the linearized estimate (red). Blue line represents parameter 95% confidence region. It is possible that linear analysis could be closer to correct answer with any random set of data, but this would be fortuitous.

For Parameter Estimates

- **In all cases, linear and nonlinear, fit what you measure, or more specifically the data that have normally distributed errors, rather than some transformation of this. Any nonlinear transformation (something other than adding or multiplying by a constant) changes the error distribution and invalidates much of the statistical theory behind the analysis.**
- **Standard packages are widely available for linear equations.**
- **Nonlinear analyses should be done on raw data (or data with normally distributed errors) and will require iteration, which Excel and other programs can handle.**

Experimental Error

- **The inherent error in the data figures prominently into almost all analyses except the best estimates of the parameters.**
- **The classical assumptions are that these errors are additive, normally distributed with a mean of 0 and constant variance (independent of the value of the dependent and independent variables and of time), and independent.**
- **None of these assumptions is always true.**
  - **Errors could be multiplicative with mean of 1, but this is equivalent to additive with mean 0 but proportional to the predicted value. Could have other forms.**
  - **Errors may be non-normally distributed, but the Central Limit Theorem provides reasonably strong motivation for them being normally distributed in many cases.**
  - **Errors often depend on each other, especially when data come from computerized acquisition systems at high rates.**

Experimental Error

- The most robust method of estimating experimental error is by replicating measurements at every value of the independent variable(s) and each combination of independent variable(s). This is commonly not practical. $s(\hat{y}_i)_e^2 = \frac{\sum_{i=0}^{r}(y_i - \hat{y}_i)^2}{r-1}$ where r is number of replicates.
- The mean difference between the predicted and measured values provides and estimate of experimental error even with no data replicates if the model is statistically correct (capable of describing every actual trend in the data) and the error assumptions are correct (independent, constant variance, additive) according to $s_e^2 = \frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}{n-p}$ where $n$is number of data points and p is number of parameters.
- In the previous example, the nonlinear and linear error estimates are 1.07 and 1.09, respectively. That is, the models estimate that the standard deviation in the normal distribution that describes the errors is about 1.08.

Population vs Sample Statistics

There are two common definitions of a standard deviation that sometimes lead to confusion.

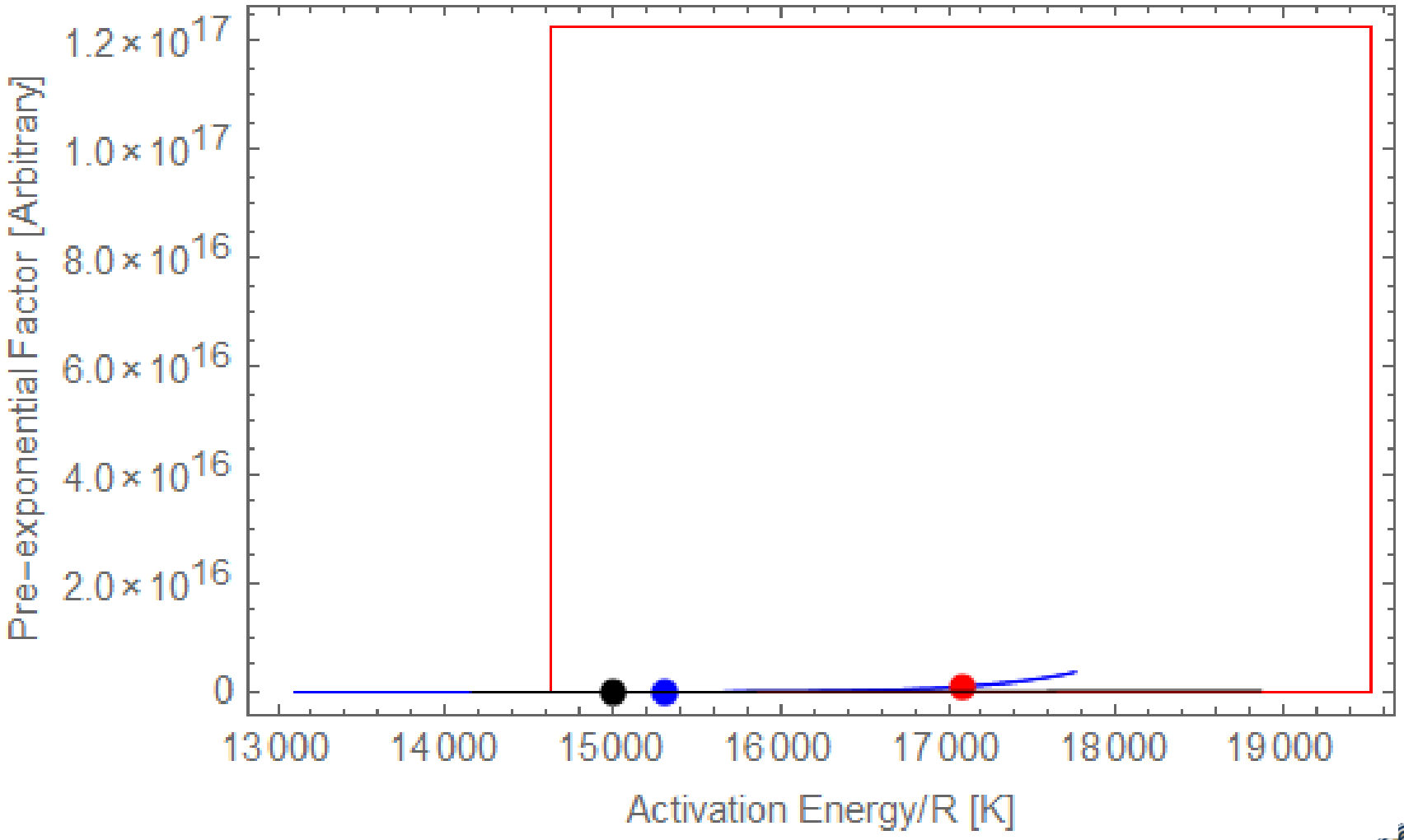
$$s^2 = \frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}{n-1}$$

is the sample estimate. That is, it is the estimated standard deviation based on a sample set of data drawn from a usually much larger population when the mean is also based on this sample set of data. Excel functions STDEV() and STDEV.S() return this value for a list of data.

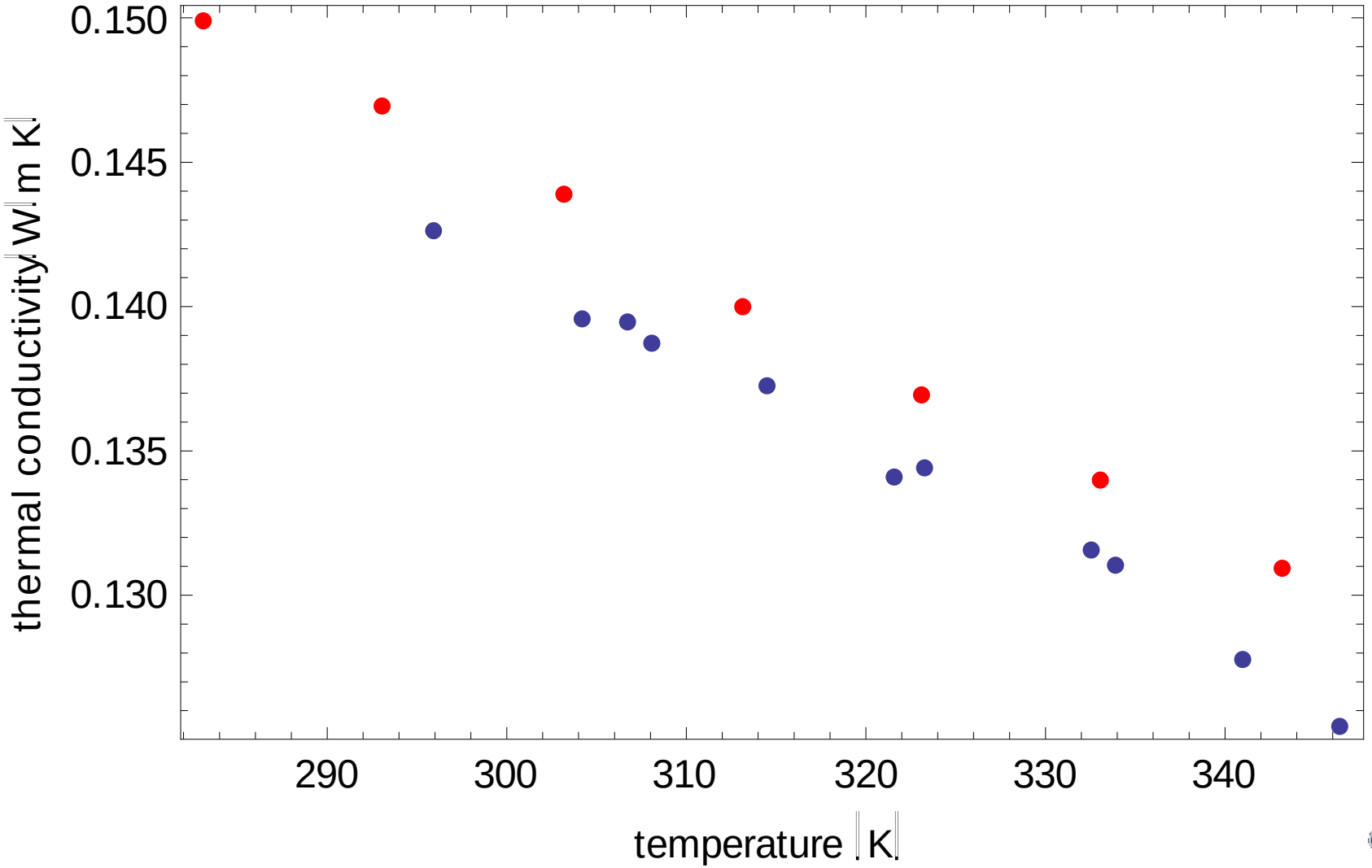
$$s^2 = \frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)^2}{n}$$

is the population estimate. That is, it is the standard deviation based on the entire population or based on a sample when the mean is known or estimated from some independent source. Excel functions STDEVP() and STDEV.P() return this value for a list of data.

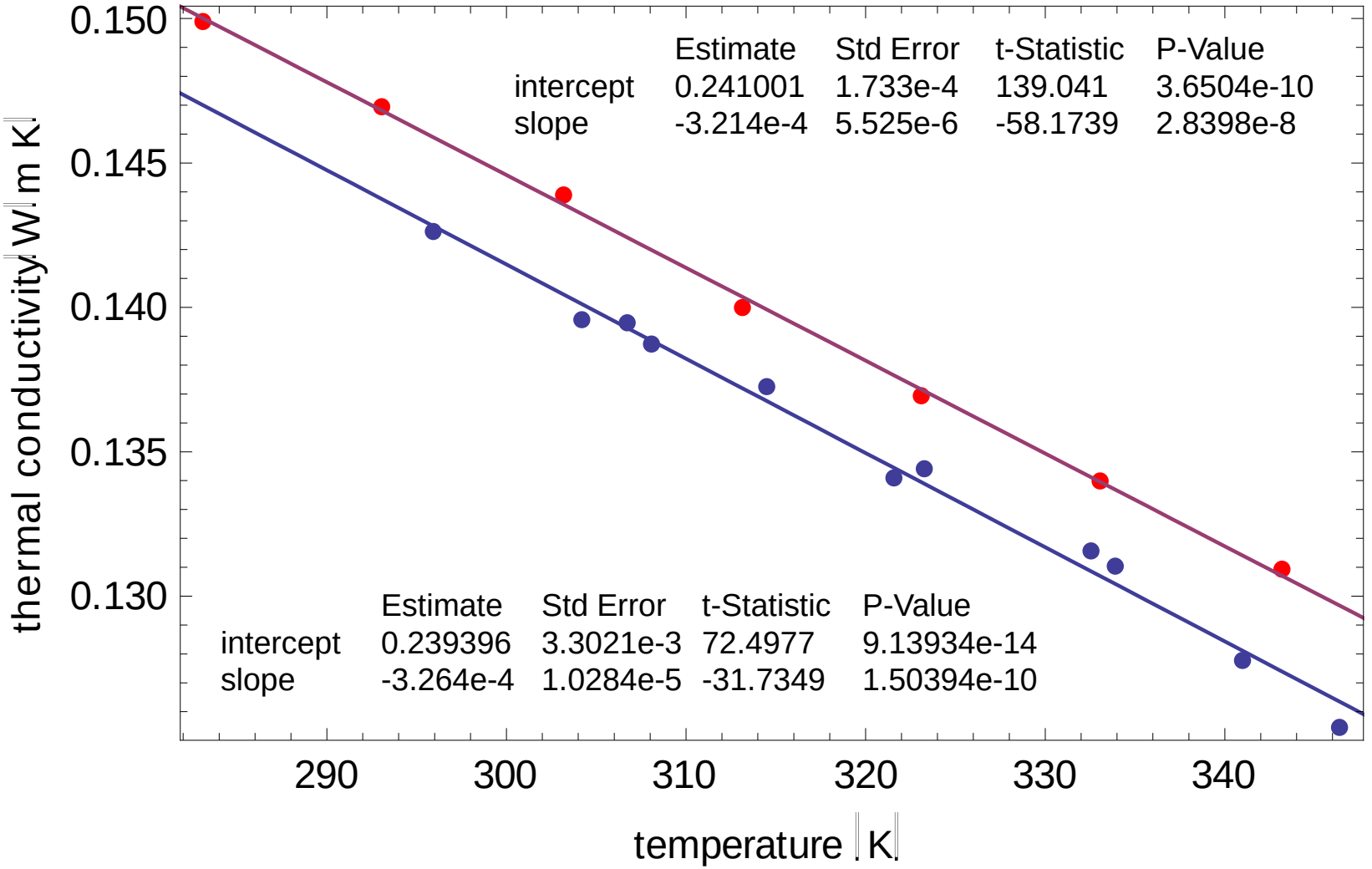
Confidence Intervals



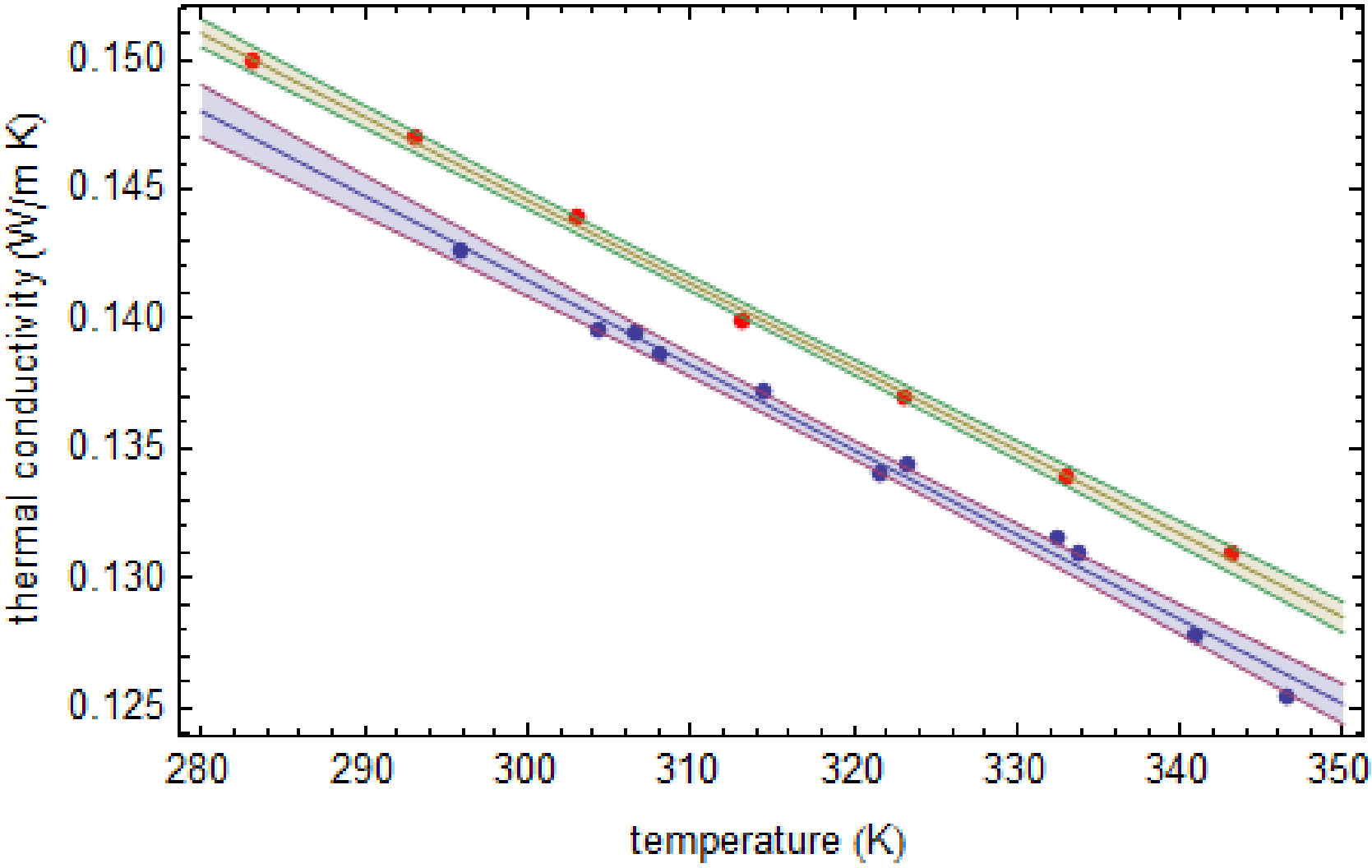
Two typical datasets



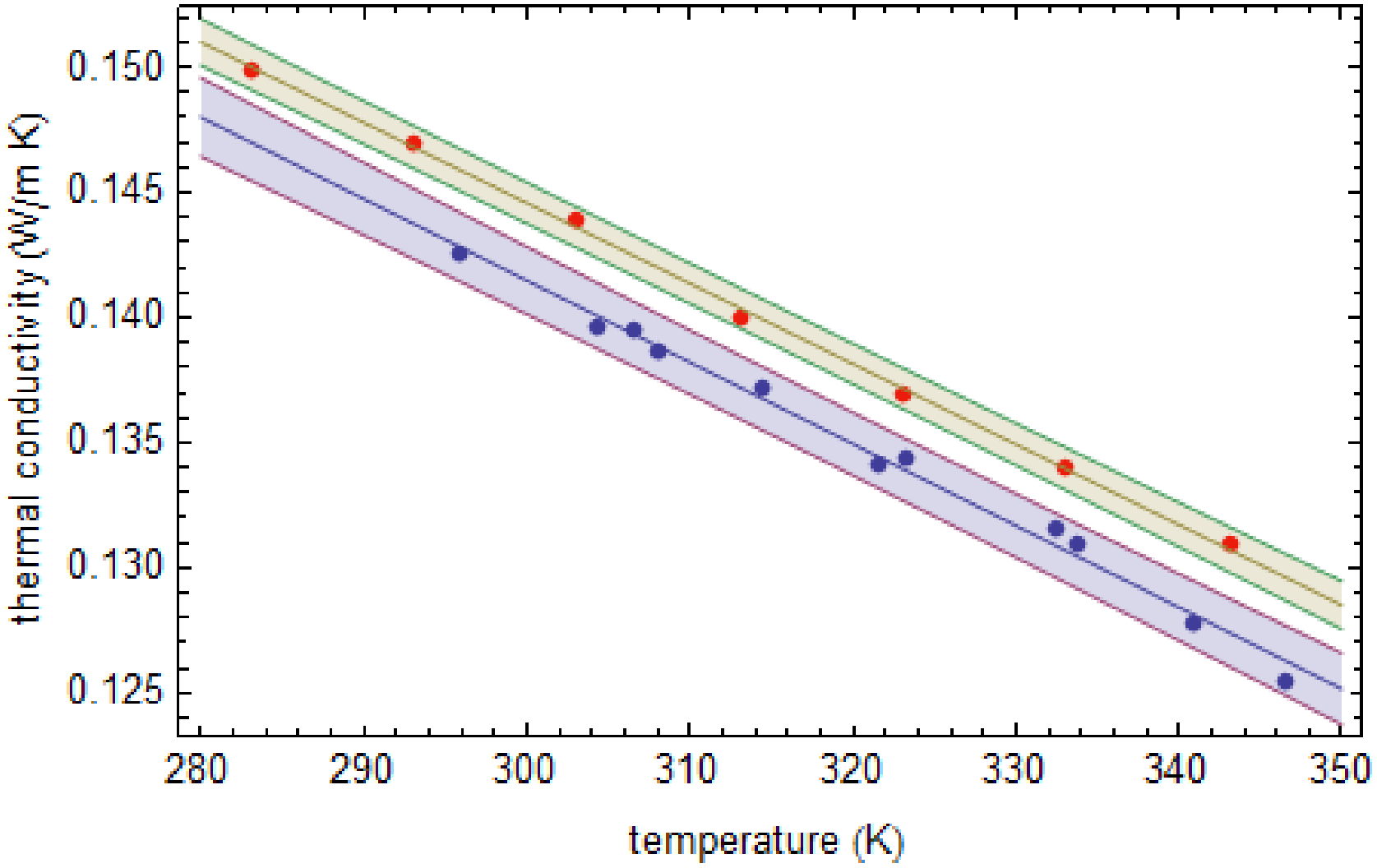
Straight-line Regression



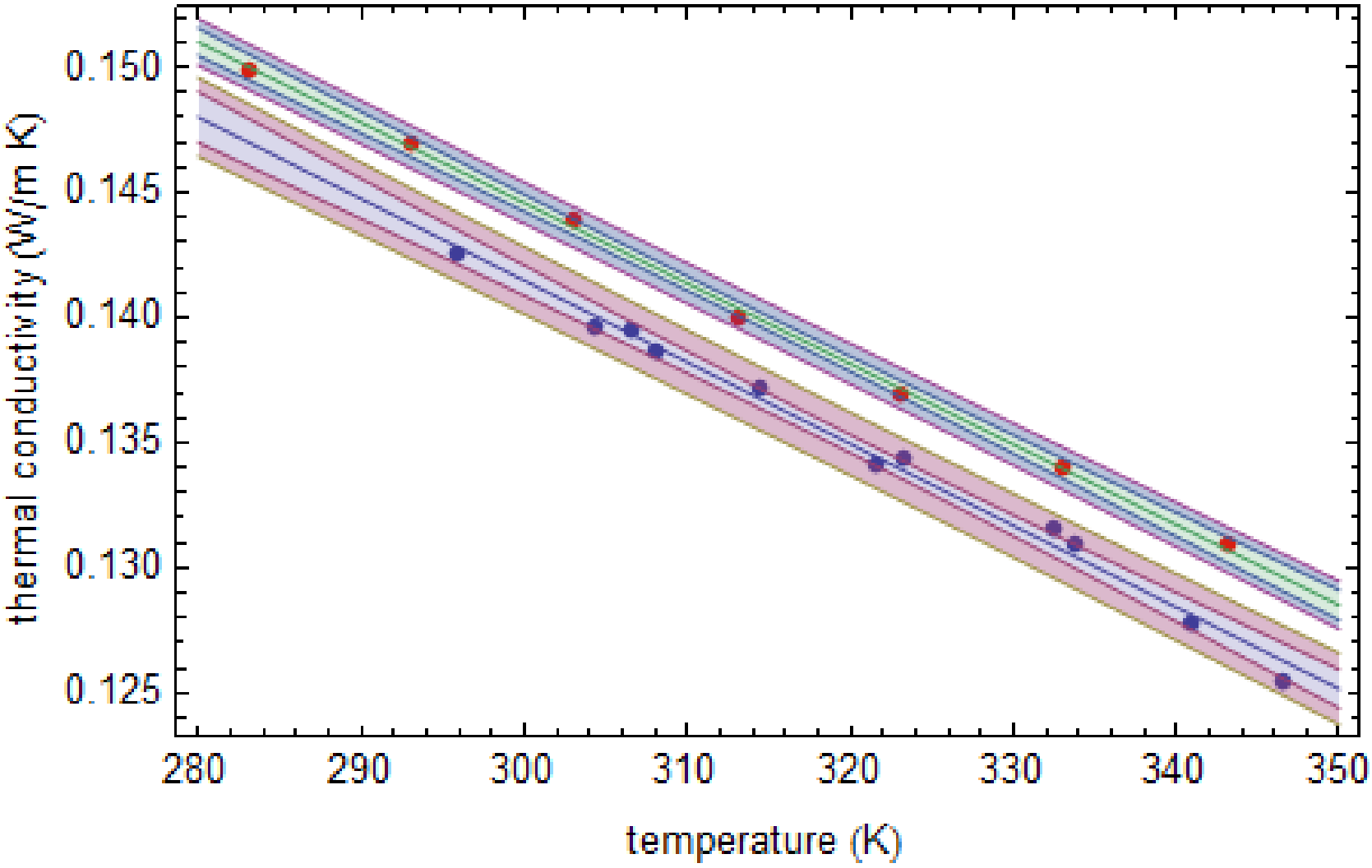
Mean Prediction Bands



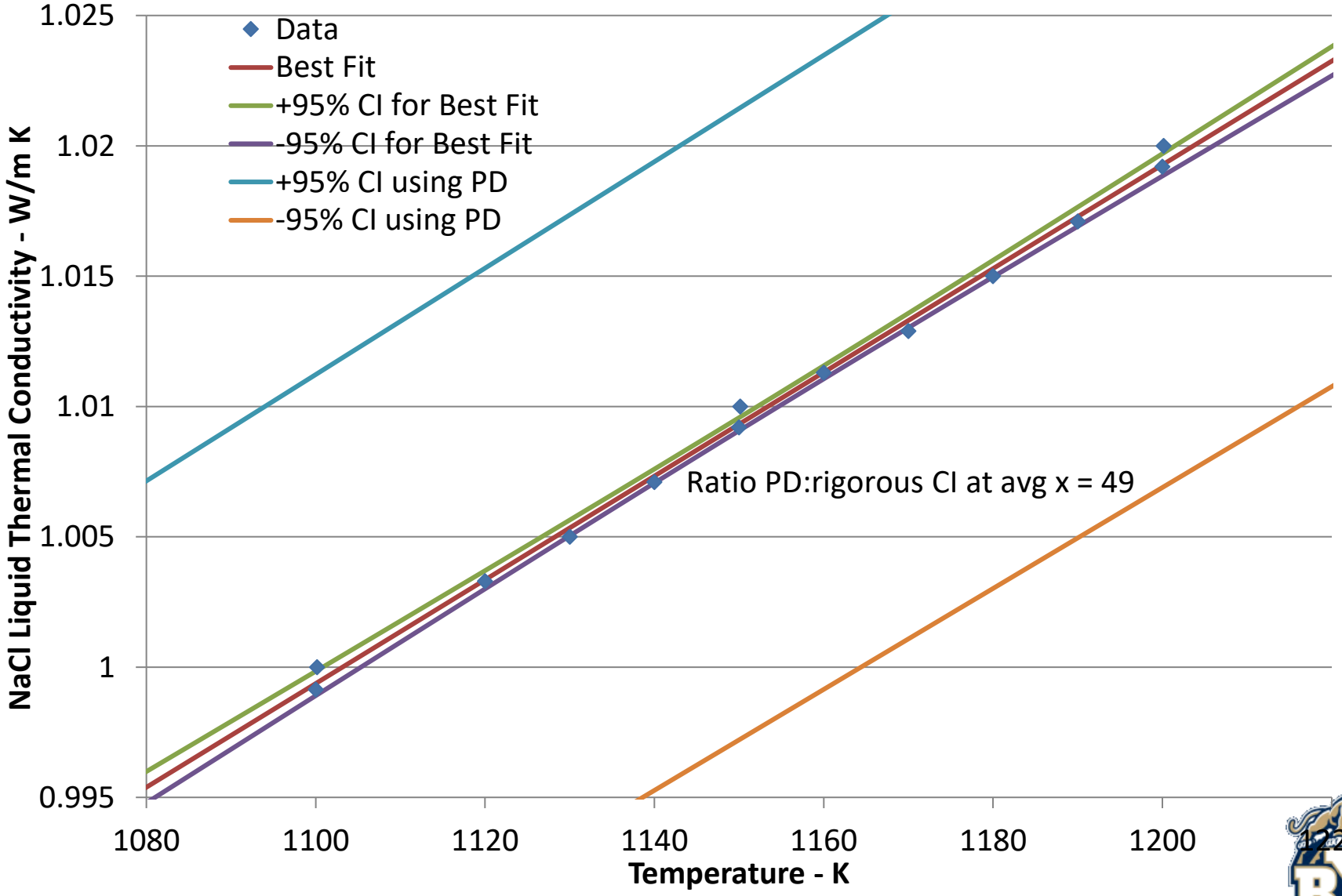
Single-Point Prediction Bands



Point and Mean Prediction Bands



Rigorous vs. Propagation of Error



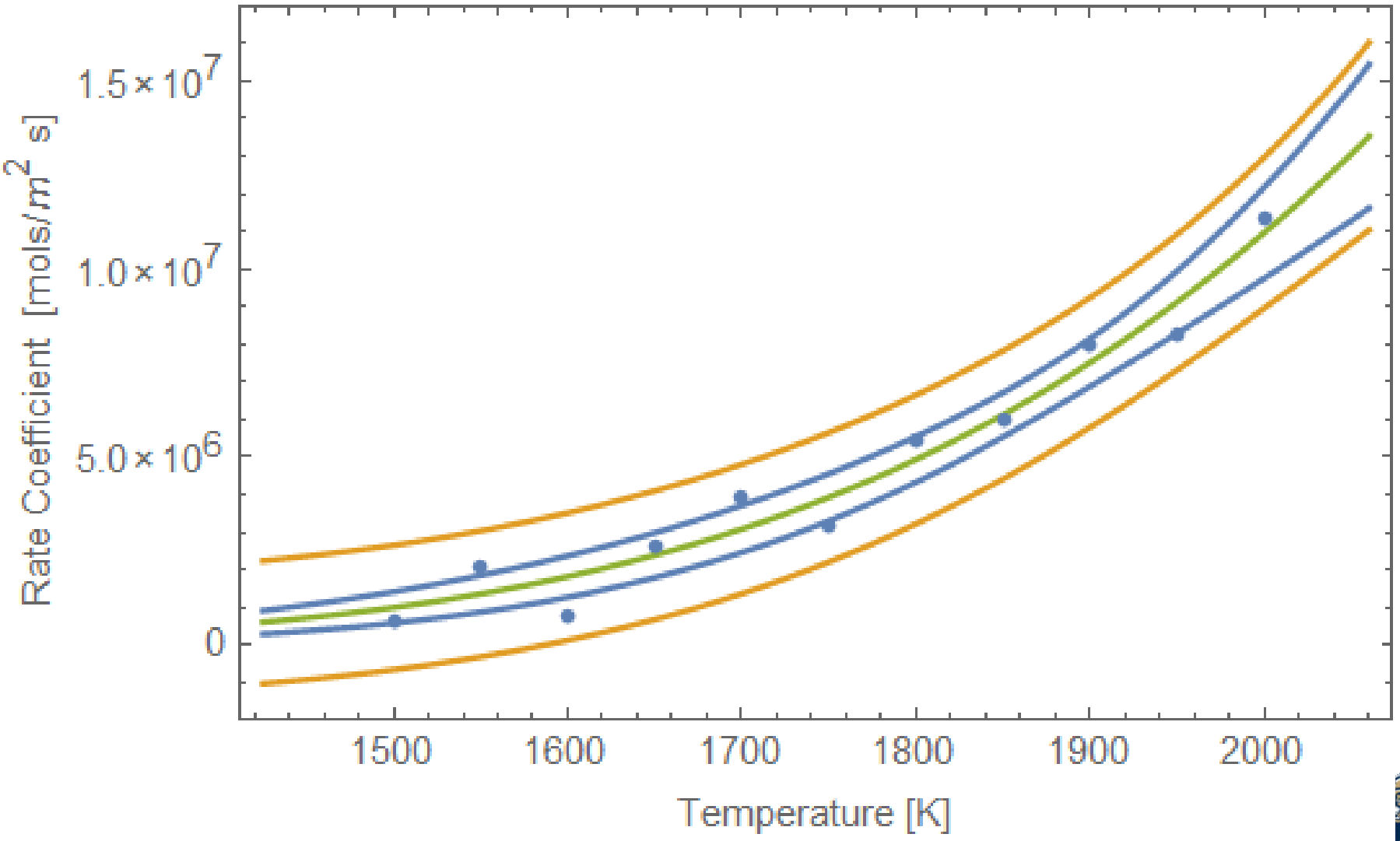
Propagation of Error

$$\sigma_f^2 = \sum_i \left(\left. \frac{\partial f}{\partial \beta_i} \right|_{\beta_i} \right)^2 \sigma_{\beta_i}^2 =$$

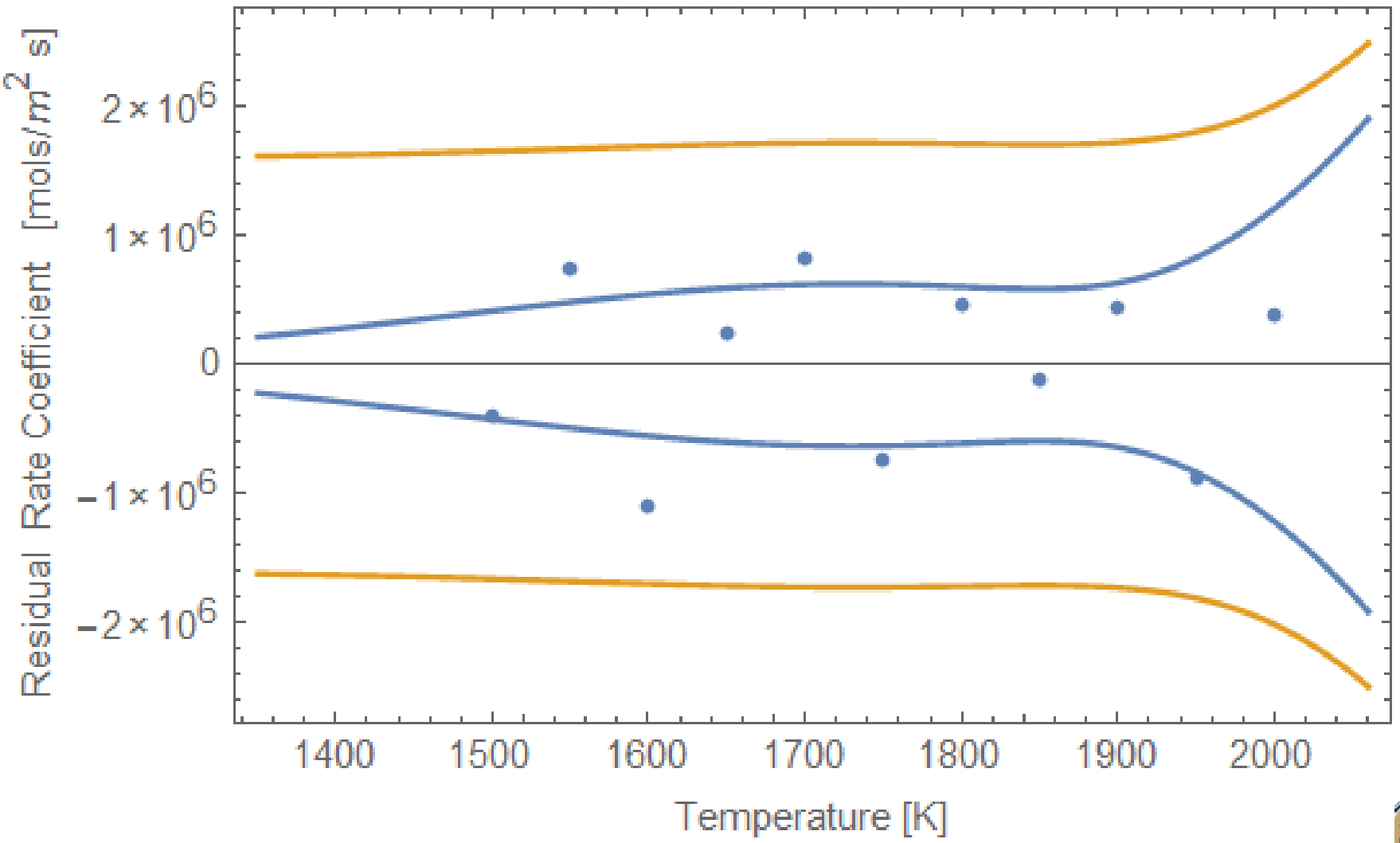
$$\left(\left. \frac{\partial (a + bx)}{\partial a} \right|_b \right)^2 \sigma_a^2 + \left(\left. \frac{\partial (a + bx)}{\partial b} \right|_a \right)^2 \sigma_b^2 =$$

$$\sigma_a^2 + x^2 \sigma_b^2$$

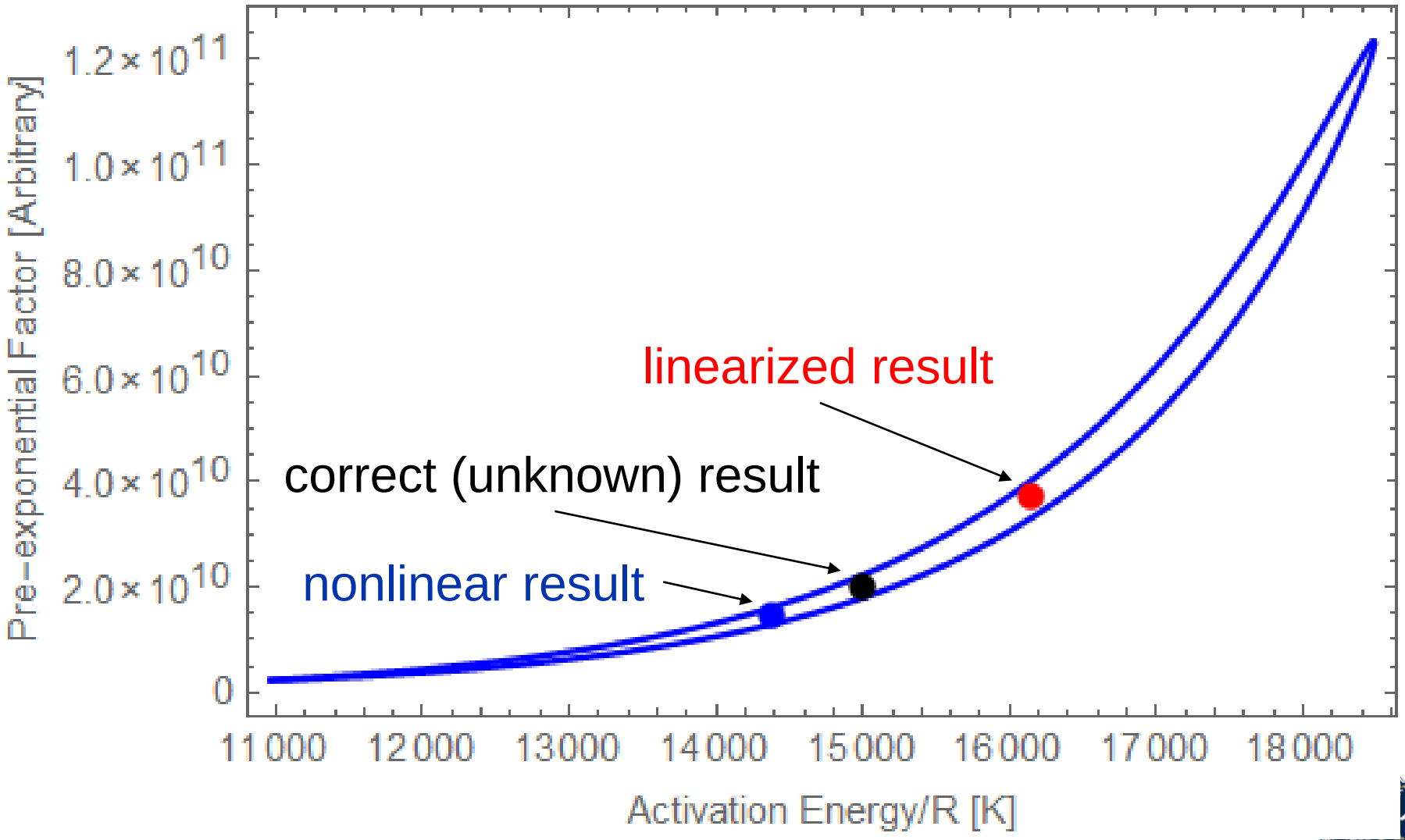
Single-Point Prediction Band



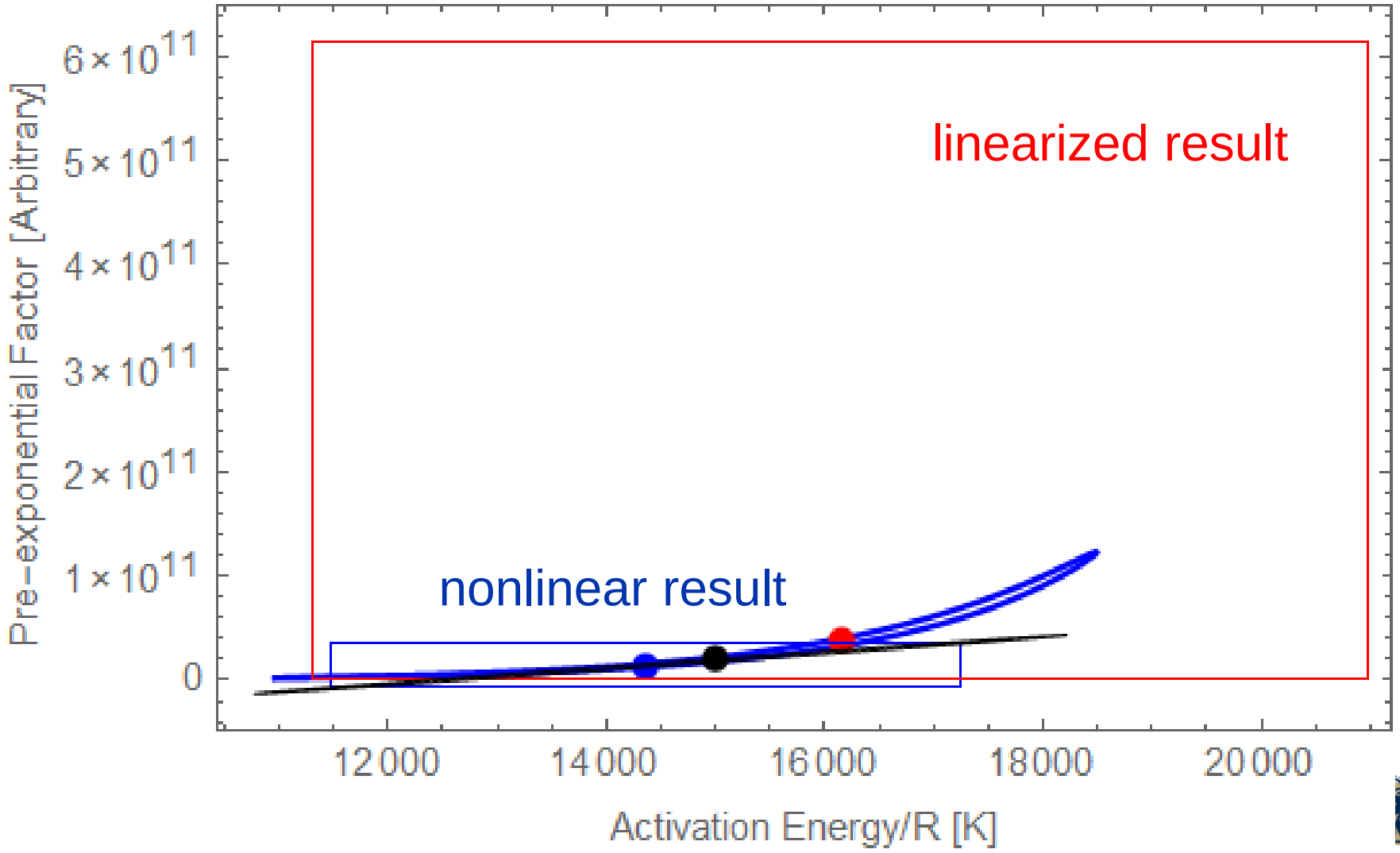
Residual SP Prediction Band (95%)



Confidence Region



Joint Confidence Regions



Prediction Band Characteristics

- **Straight Line**
  - **Form hyperbolae with waist (minimum) at $\bar{x}$.**
  - **Band range at a given x increases monotonically and without bound as $x \to \pm\infty$.**
  - **Mean Prediction Band**
    - **Individual points commonly lie outside the range**
    - **The range of the mean prediction band goes to 0 as $n \to \infty$.**
  - **Single-point Prediction Band**
    - **Individual points rarely lie outside the band (5% of the time for a 95% band).**
    - **The range of the single-point prediction band is finite as $n \to \infty$.**
    - **Becomes the mean prediction band as the number of additional points approaches ∞.**

Prediction Band Characteristics

- **Linear Equations (not necessarily straight line)**
  - **Waist or multiple waists (minima) in range of data.**
  - **Band range at a given x increases non-monotonically and without bound as $x \to \pm\infty$.**
  - **Mean Prediction Band**
    - **Individual points commonly lie outside the range**
    - **The range of the mean prediction band goes to 0 as $n \to \infty$.**
  - **Single-point Prediction Band**
    - **Individual points rarely lie outside the band (5% of the time for a 95% band).**
    - **The range of the single-point prediction band is finite as $n \to \infty$.**
    - **Becomes the mean prediction band as the number of additional points approaches ∞.**

Nonlinear Equations

- **Nonlinear Equations**
  - **Minimum waist can and commonly does occur outside range of measured data.**
  - **Band range at a given x increases non-monotonically and frequently is bounded on at least one side as $x \to \pm\infty$.**
  - **Mean Prediction Band**
    - **Individual points commonly lie outside the range**
    - **The range of the mean prediction band goes to 0 as $n \to \infty$.**
  - **Single-point Prediction Band**
    - **Individual points rarely lie outside the band (5% of the time for a 95% band).**
    - **The range of the single-point prediction band is finite as $n \to \infty$.**
    - **Becomes the mean prediction band as the number of additional points approaches ∞.**

Parameter Characteristics

- **Linear models**
  - **Parameters are explicit functions of data – do not depend on themselves.**
  - **Parameters require no iteration to compute.**
  - **Normal equations are independent of parameters.**
- **Nonlinear models**
  - **Parameters depend on themselves – need an estimate to begin iterative computation**
  - **Parameters generally determined by converging and optimization problem, not by explicit computation.**
  - **Optimization problem commonly quite difficult to converge.**

ANOVA and Conf. Int. Tables

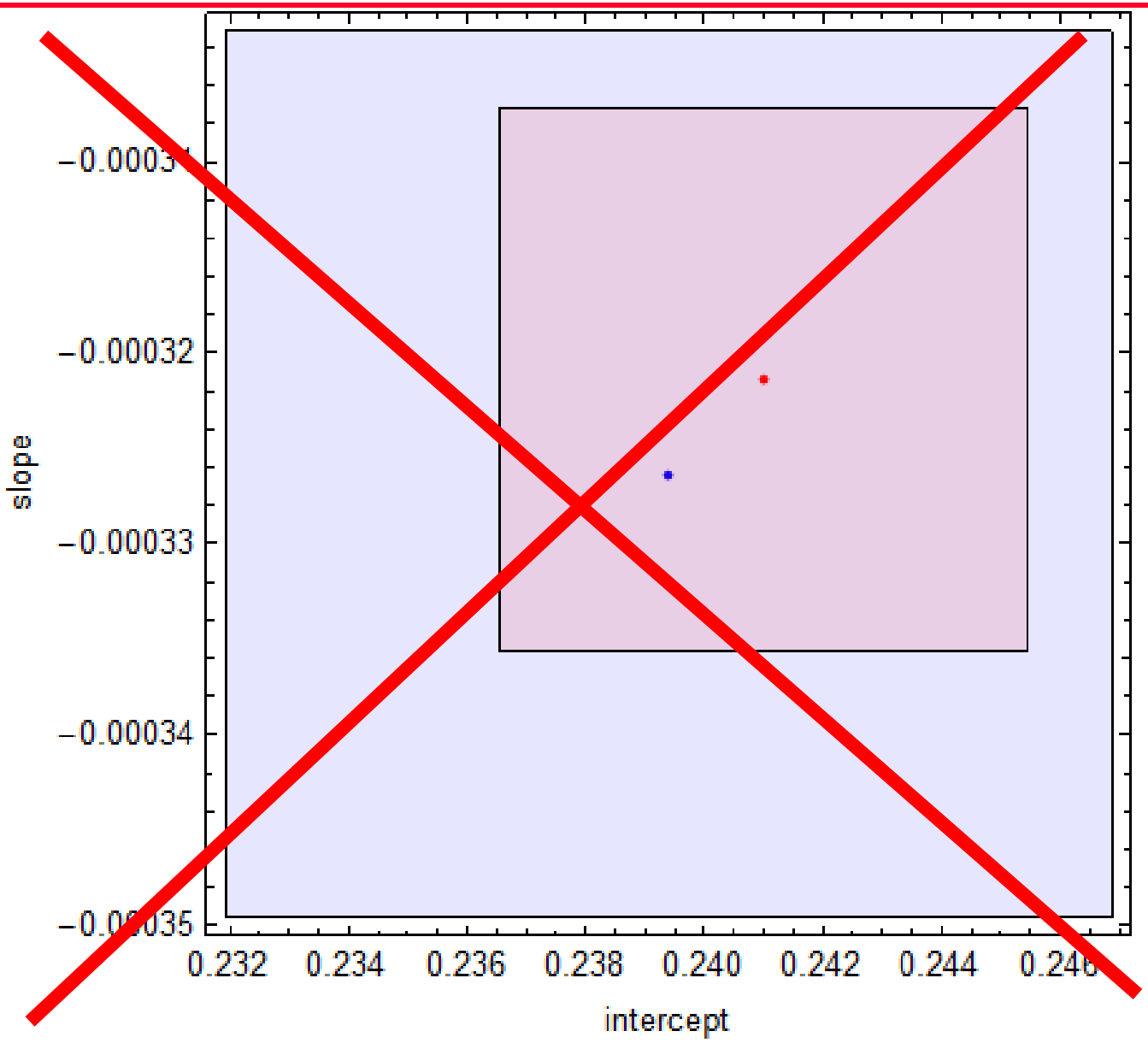
	DF	SS	MS	F-Statistic	P-Value
x	1	2.8515e-4	2.8515e-4	1007.1	1.50394e-10
Error	9	2.5483e-6	2.83142e-7		
Total	10	2.8770e-4			

	DF	SS	MS	F-Statistic	P-Value
x	1	2.89221e-4	2.89221e-4	3384.21	2.8398e-8
Error	5	4.27311e-7	8.54621e-8		
Total	6	2.89649e-4			

	Estimate	Standard Error	Confidence Interval
intercept	0.239396	0.00330212	{0.231926,0.246866}
slope	-3.26366e-4	1.02841e-5	{-3.49631e-4,-3.03102e-4}

	Estimate	Standard Error	Confidence Interval
intercept	0.241001	0.00173331	{0.236545,0.245457}
slope	-3.2139e-4	5.52469e-6	{-3.35595e-4,-3.07191e-4}

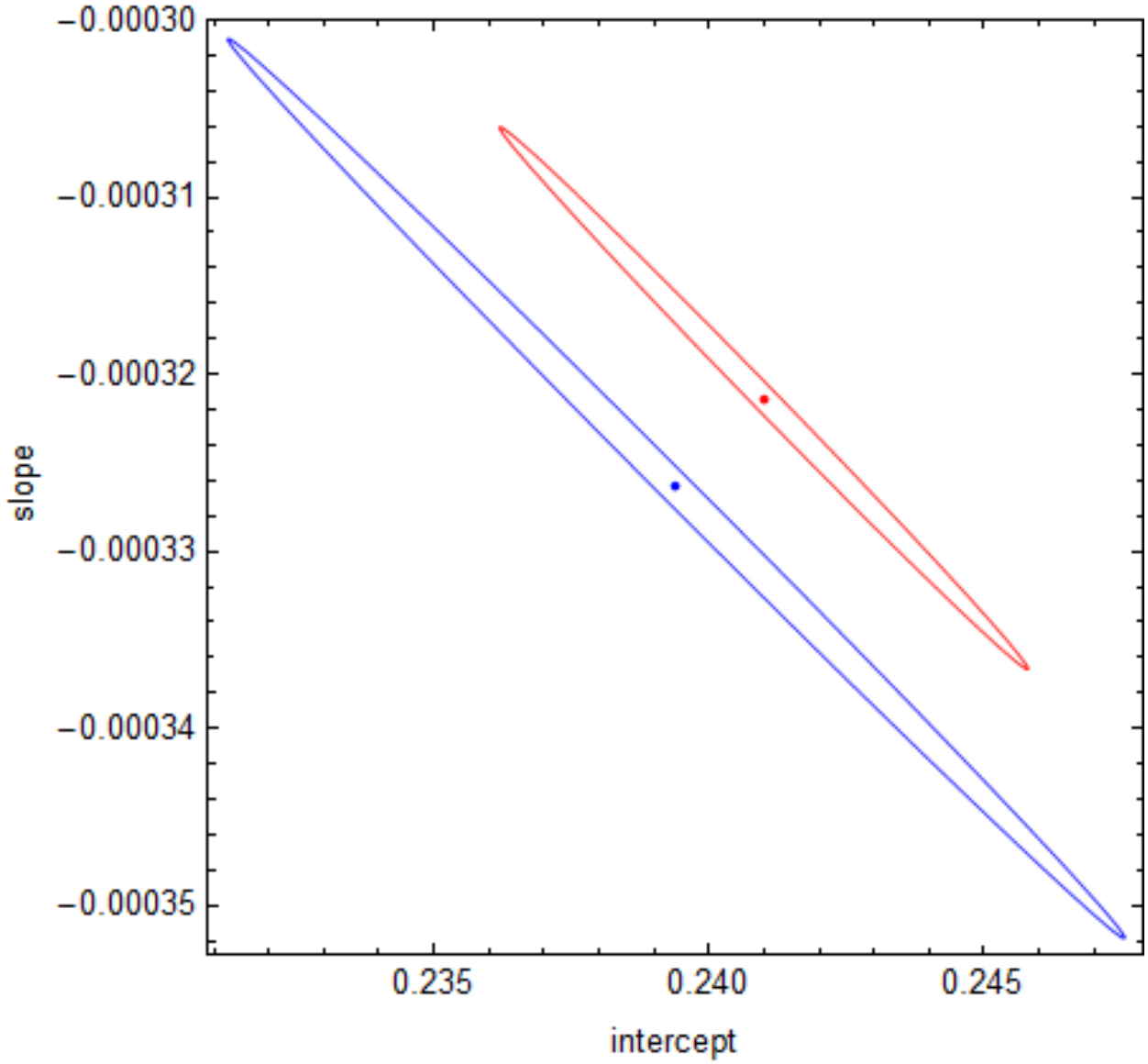
Confidence Interval Ranges



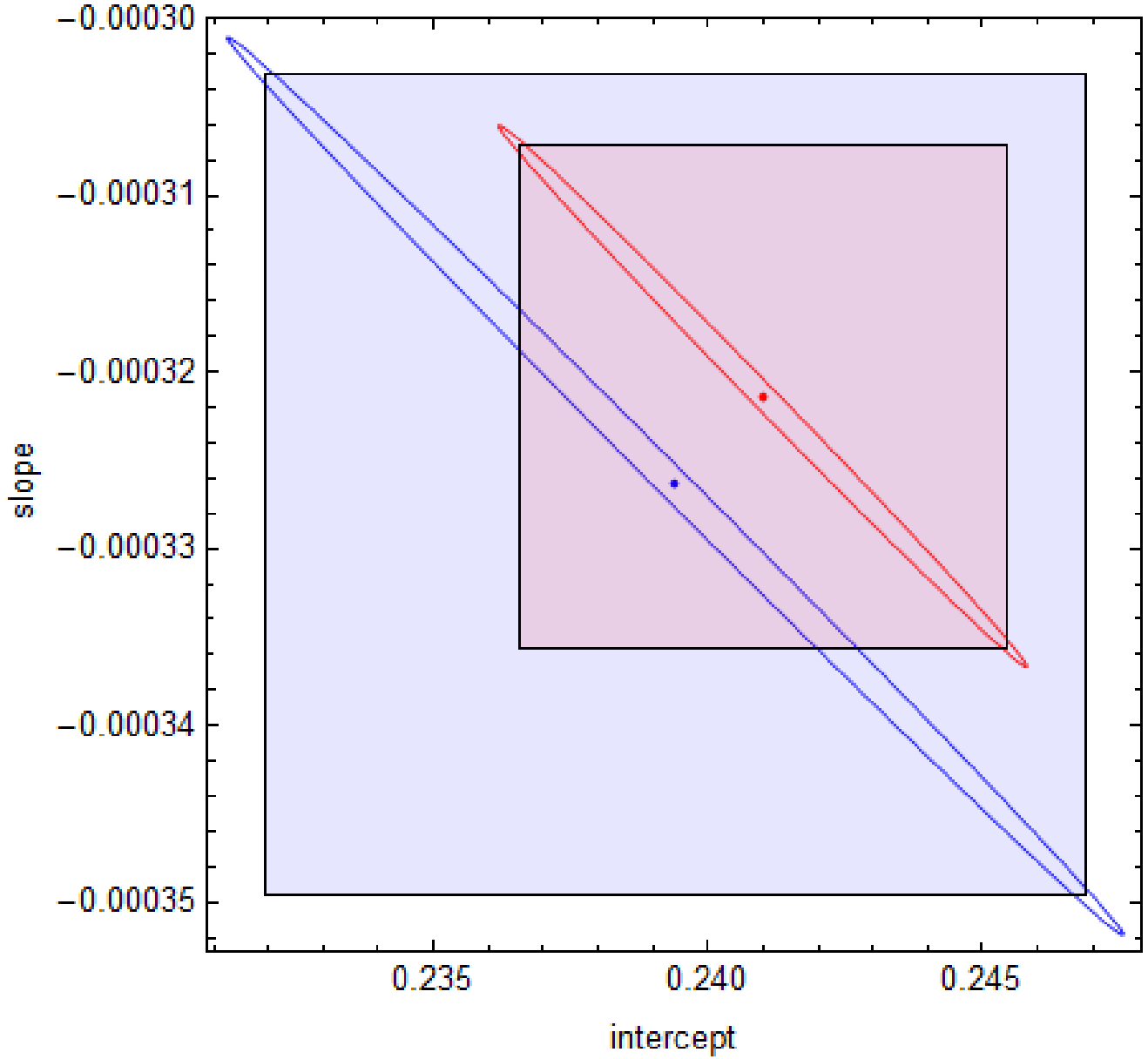
Some Interpretation Traps

- **It would be easy, but incorrect, to conclude**
  - **That reasonable estimates of the line can, within 95% probability, be computed by any combination of parameters within the 95% confidence interval for each parameter**
  - **That experiments that overlap represent the same experimental results, within 95% confidence**
  - **That parameters with completely overlapped confidence intervals represent essentially indistinguishable results**
- **If you consider the first graph in this case, all of these conclusions seem intuitively incorrect, but experimenters commonly draw these types of conclusions.**
- **Joint or simultaneous confidence regions (SCR) address these problems**

Simultaneous Confidence Region



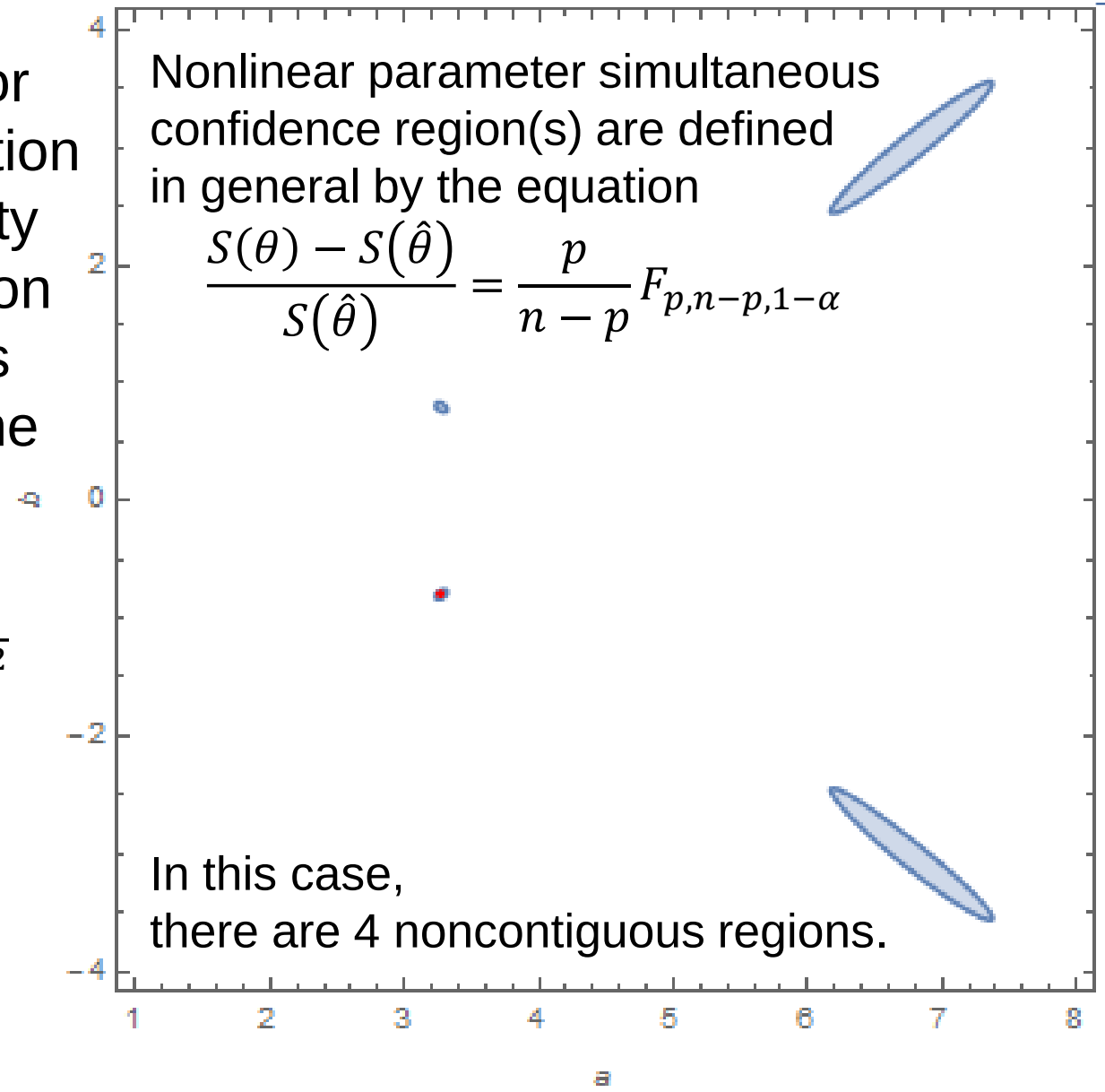
Regions and Intervals Compared



Nonlinear SCR More Complex

The Cauchy or Lorentz equation (is a probability density function and describes some laser line widths).

$$y = \frac{1}{1 + \frac{(x-a)^2}{b^2}}$$



Sim. or Joint Conf. Regions

Region defined by

$$\frac{S(\theta) - S(\hat{\theta})}{S(\hat{\theta})} \leq \frac{p}{n-p} F_{p,n-p,1-\alpha}$$

$S(\theta)$ is the sum square errors as a function of the parameters, represented by the vector θ. This is a function that depends on the parameter values.

$S(\hat{\theta})$ is $S(\theta)$ evaluated at the optimum parameters, represented by $\hat{\theta}$. This is a number, not a function.

p is the number of parameters (a number).

n is the number of data points (a number).

$F_{p,n-p,1-\alpha}$ is the critical value of the F distribution with p and $n-p$ degrees of freedom and at confidence level α (a number)

Simultaneous Confidence Regions

- **Overlapping confidence intervals is a poor test for difference in data sets.**
- **Data that may appear to be similar based on confidence intervals may in fact be quite different and certainly distinct from one another.**
- **Parameters in the lonely corners of interval unions are exceptionally poor estimates.**

Confidence Region Characteristics

- **Linear Equations**
  - **Always form p-dimensional ellipsoids, where p is the number of parameters.**
  - **Parameters are generally linearly correlated (ellipsoid axes are not aligned with parameter axes).**
    - **Not correlated for straight line if $\bar{x} = 0$.**
    - **Constant and quadratic parameters always correlated for quadratic**
  - **Parameter uncertainty is usually much smaller than confidence interval at given values of other parameters.**
  - **Parameter uncertainty range slightly exceeds conf. interval range.**

Confidence Region Characteristics

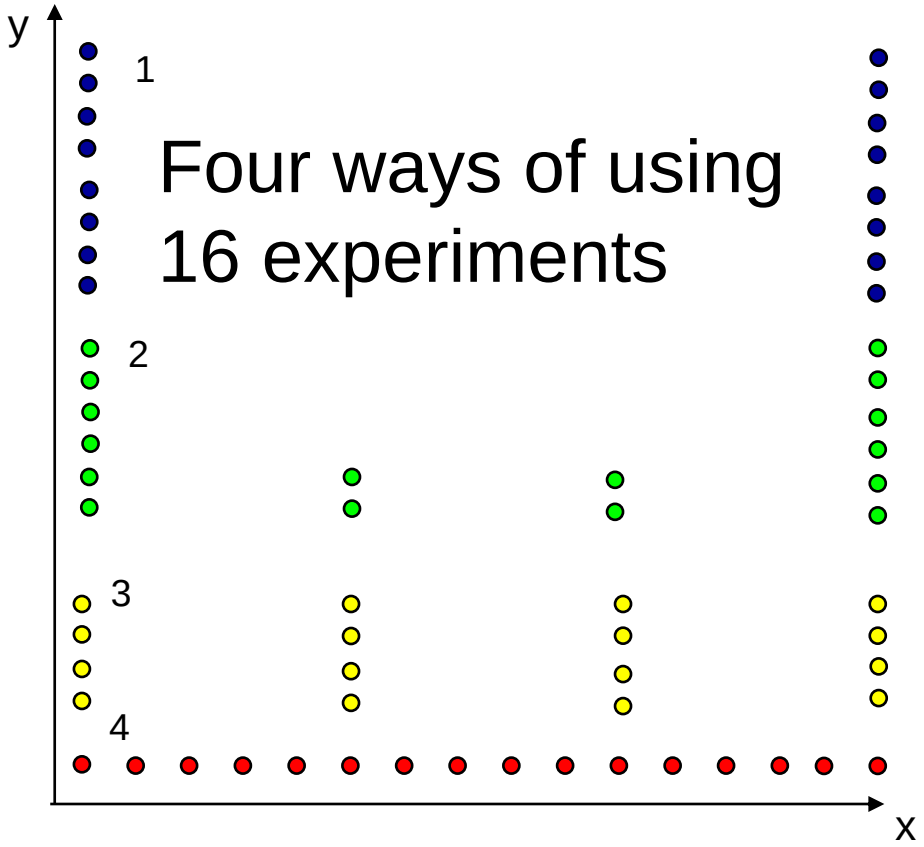
- **Nonlinear Equations**
  - **Regions assume many shapes and may not be contiguous.**
  - **Parameters generally correlated, but not linearly correlated.**
  - **Parameter uncertainty usually much smaller than confidence interval at given values of other parameters.**
  - **Parameter uncertainty range exceeds conf. interval range and may not be bounded.**

Experimental Design

- **In this context, experimental design means selecting the conditions that will maximize the accuracy of your model for a fixed number of experiments.**
- **There are many experimental designs, depending on whether you want to maximize accuracy of prediction band, all parameters simultaneously, a subset of the parameters, etc.**
- **The d-optimal design maximizes the accuracy of all parameters and is quite close to best designs for other criteria. It is, therefore, by far the most widely used.**

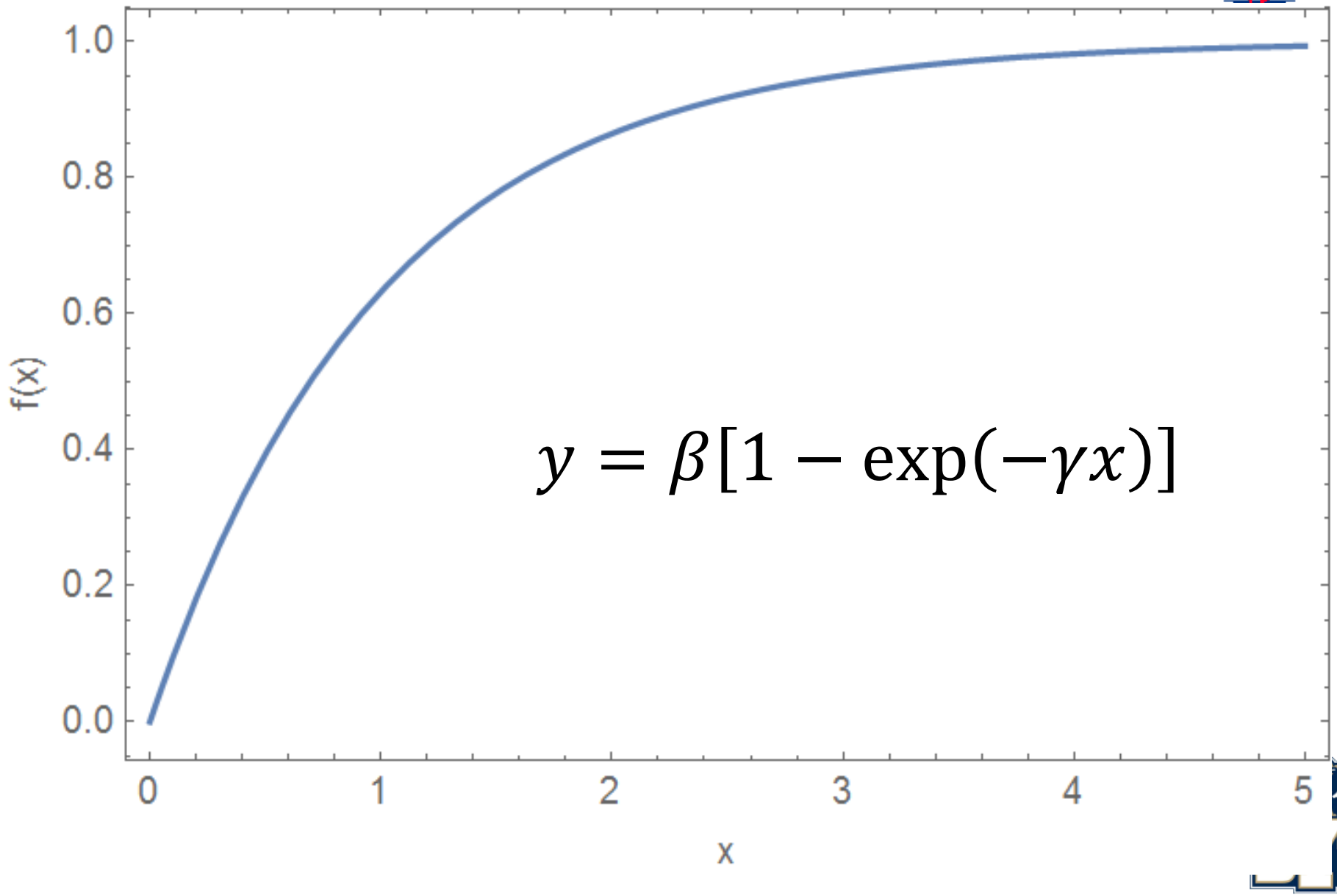
Fit vs. Parameter Precision

Generally there is a compromise between minimizing parameter variance and validating the model.



Design	1	2	3	4
Lack of fit df	0	2	2	14
Pure error df	14	12	12	0
$\frac{sd(b_1)}{\sigma}$	0.25	0.28	0.34	0.41
P sites	2	4	4	16

Typical (nonlinear) Application

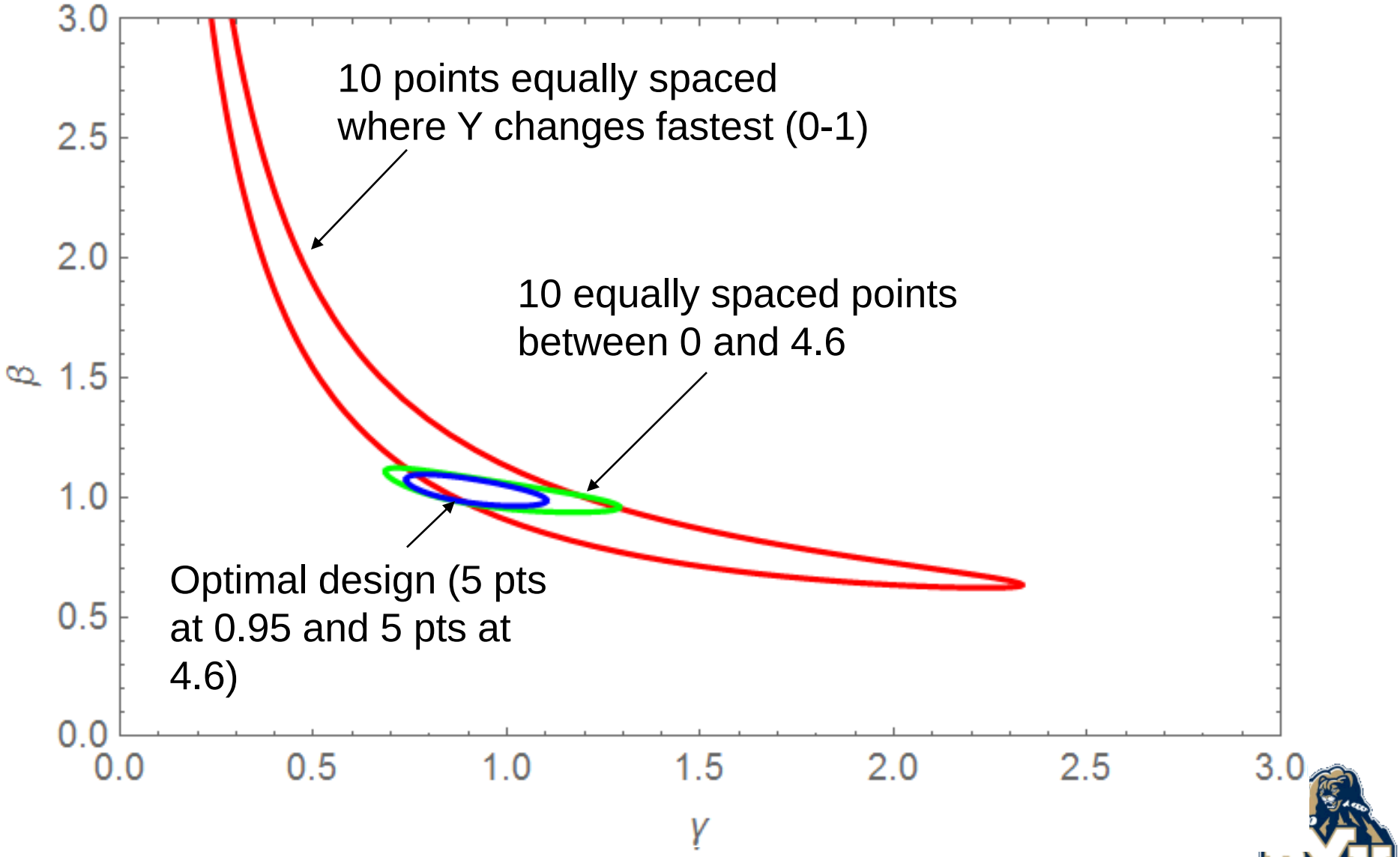


Example

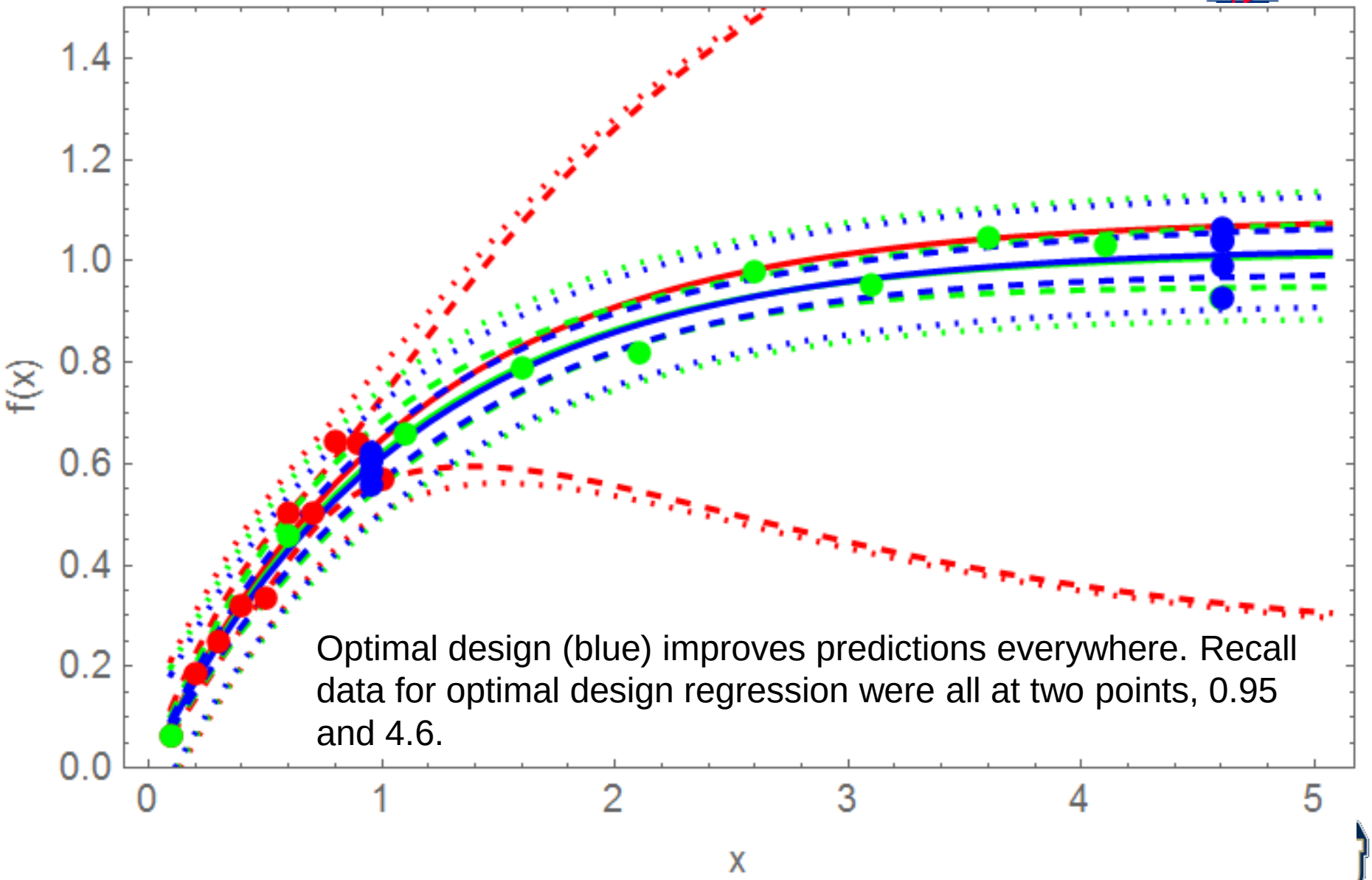
Compare 3 different design, each with the same number of data points and the same errors

Point No.	x_1	y_1	x_2	y_2	x_3	y_3	ϵ_i
1	0.1	0.0638	0.1	0.0638	0.954	0.5833	0.03150
2	0.2	0.1870	0.6	0.4569	0.954	0.6203	-0.00550
3	0.3	0.2495	1.1	0.6574	0.954	0.6049	0.00990
4	0.4	0.3207	1.6	0.7891	0.954	0.6056	0.00920
5	0.5	0.3356	2.1	0.8197	0.954	0.5567	0.05810
6	0.6	0.5040	2.6	0.9786	4.605	1.0428	-0.05280
7	0.7	0.5030	3.1	0.9545	4.605	0.9896	0.00040
8	0.8	0.6421	3.6	1.0461	4.605	1.0634	-0.07340
9	0.9	0.6412	4.1	1.0312	4.605	1.0377	-0.04770
10	1.0	0.5678	4.6	0.9256	4.605	0.9257	0.06430

SCR Results for 3 Cases

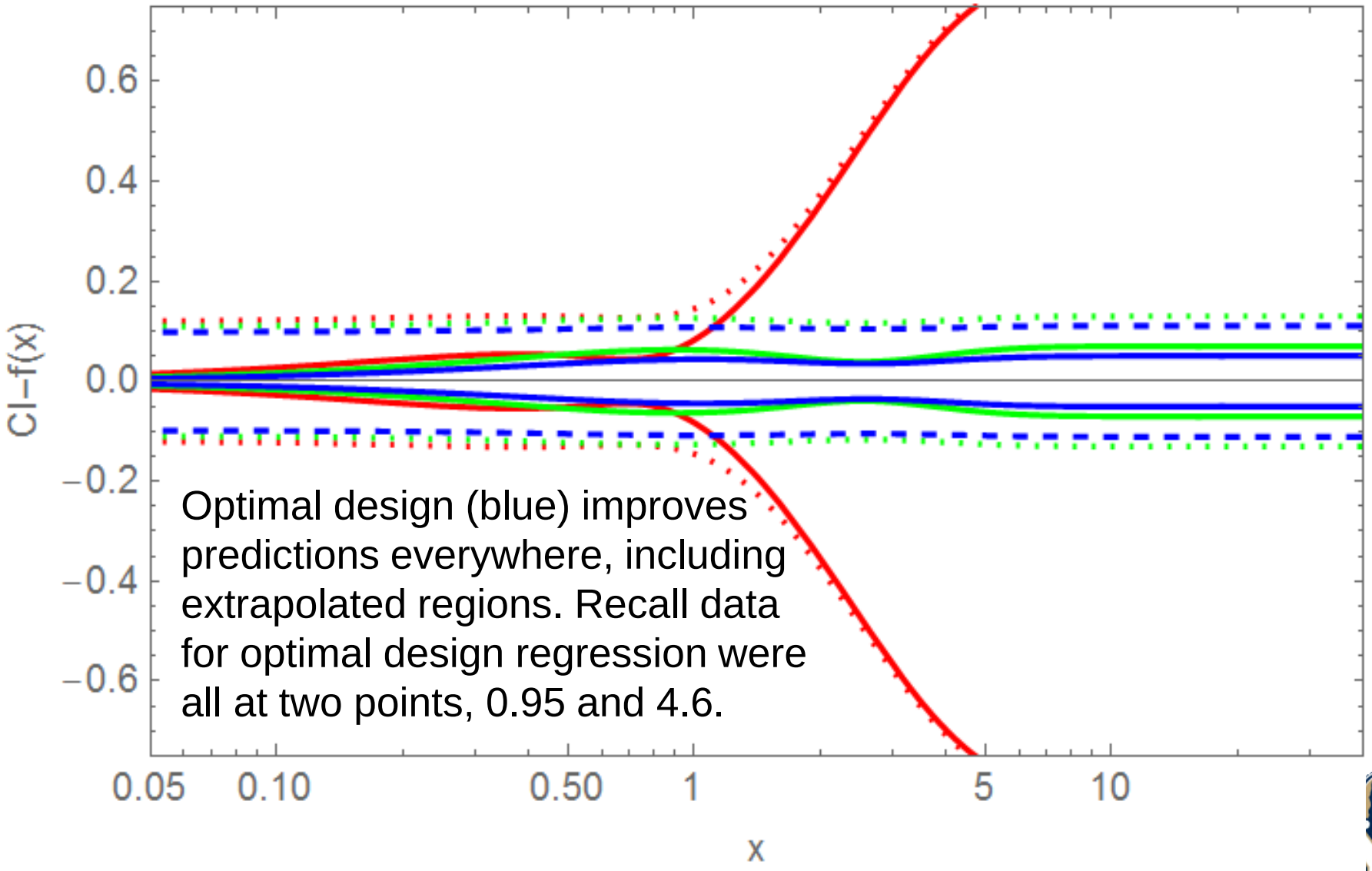


Prediction and SP Conf. Intervals



Optimal design (blue) improves predictions everywhere. Recall data for optimal design regression were all at two points, 0.95 and 4.6.

Improved Prediction & SP Intervals



Optimal design (blue) improves predictions everywhere, including extrapolated regions. Recall data for optimal design regression were all at two points, 0.95 and 4.6.

Linear Design Summary

- **For linear systems with a possible experimental range from x_1 to x_2**
  - **Straight line – equal number of points at each of two extreme points**
  - **Quadratic – extreme points plus middle**
  - **Cubic – extreme points plus points that are located at**

(equally spaced would be at)

$$\frac{(x_1 + x_2)}{2} \pm \frac{\sqrt{5}}{10}(x_2 - x_1) \qquad \frac{(x_1 + x_2)}{2} \pm \frac{1}{6}(x_2 - x_1)$$

  - **In general, the optimal points are at the maxima of**

$$\prod_i^{p-1} \prod_{j=i+1}^{p} (x_i - x_j)^2$$

  - **Generally, a few points should be added between two of these points to assure goodness of fit.**

Nonlinear Experimental Design

Optimal points (minimum parameter and prediction uncertainty) are at the extrema (positive or negative) of the Jacobian matrix determinant with respect to the parameters, that is, for a function $f(x)$, optimal points are at the maxima of the determinant of

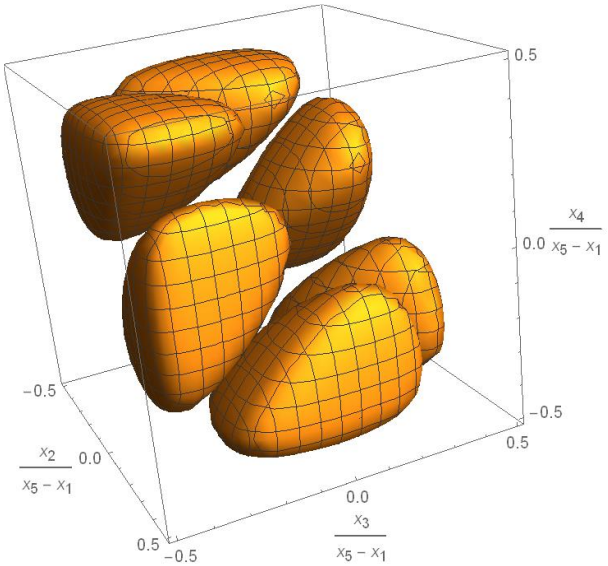
$$F = \begin{bmatrix} \left.\frac{\partial f}{\partial p_1}\right|_{x_1} & \left.\frac{\partial f}{\partial p_2}\right|_{x_1} & \cdots & \left.\frac{\partial f}{\partial p_p}\right|_{x_1} \\ \left.\frac{\partial f}{\partial p_1}\right|_{x_2} & \left.\frac{\partial f}{\partial p_2}\right|_{x_2} & \cdots & \left.\frac{\partial f}{\partial p_p}\right|_{x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \left.\frac{\partial f}{\partial p_1}\right|_{x_p} & \left.\frac{\partial f}{\partial p_2}\right|_{x_p} & \cdots & \left.\frac{\partial f}{\partial p_p}\right|_{x_p} \end{bmatrix}$$

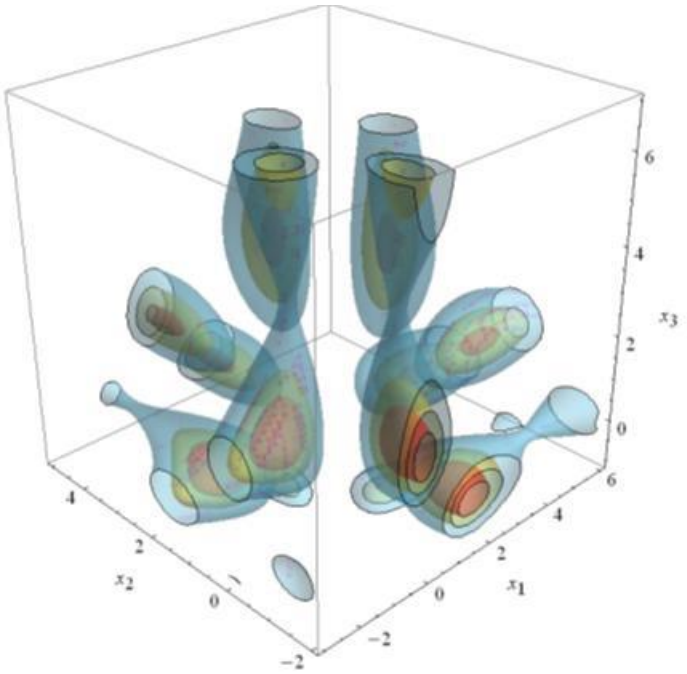
where p is the number of parameters in the model and x_i are the optimal design points

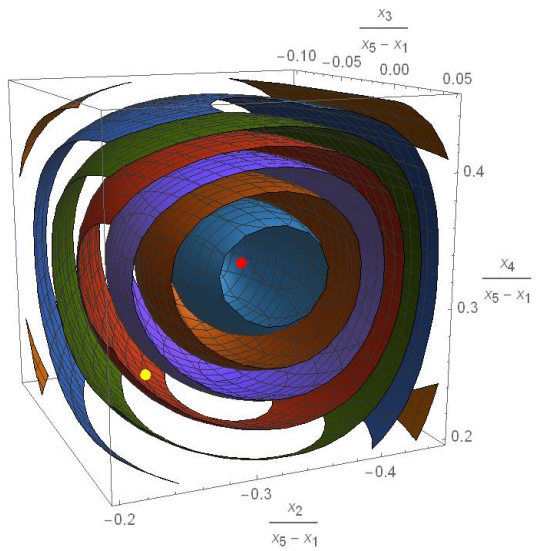
Nonlinear Design Summary

- **Nonlinear design depends on the value of parameters. Determining the parameters is usually the objective of the design. A bit of a circular (iterative) process. Start with reasonable estimates.**
- **The design can be stated as finding the maxima of $F'F$ or the extrema of F. The latter is simpler math.**
- **In many but not all cases, one or more extrema will be at the highest or lowest achievable value of the independent variable.**
- **There will be at least $p - 1$ inflection points, located at $x_i = x_j$. These are optimally poor (useless), not optimally good design points.**
- **Frequently, the extrema can only be found numerically or graphically, not analytically.**

Analytical and Graphical Solutions







Conclusions

- **Statistics is the primary means of inductive logic in the technical world. With proper statistics, we can move from specific results to general statements with known accuracy ranges in the general statements.**
- **Many aspects of linear statistics are commonly misunderstood or misinterpreted.**
- **Nonlinear statistics is a generalization of linear statistics (becomes identical as the model becomes more linear) but most of the results and the math are more complex.**
- **Statistics is highly useful in both designing and analyzing experiments.**