UNIVERSITY OF CAMBRIDGE

HUNTER research group

Model comparisons and robust estimators for equilibria in supramolecular chemistry

Hunter Group | Daisy Jia | 21st August 2024

Goal

- Find the equilibrium constant of a specific chemical interaction.

$$mH + nG \leftrightarrow^{K_{mn}} H_mG_n$$

$$K_{mn} = \frac{[H_mG_n]}{[H]^m[G]^n}$$

where $[X]$ represents the concentration of chemical X. We can also say the chemical H_mG_n has an equilibrium constant K_{mn}.

Intuitively, this number represents how ‘strong’ the interaction is.

Outline

I. Background introduction: P4 to P9.

II. The 3 problems:
   a) Model misspecification error: P10 to P25.
   b) Errors in X: P26 to P27.
   c) Outliers in data: P28 to P30.

Background Introduction

Experiment

- Denote host solution by H and guest solution by G. We first prepare a solution only consisting of H in a container, and we add G into H drop by drop. These two chemicals will interact and form new molecules, causing the color of the solution to change.

- We would like to model the relation between color changes and drops of G added to the container. In order to quantify color changes, we use light absorbances.

Linear model

- From chemistry, we have a law called Beer Lambert law, which states that the light absorbance and concentrations of all chemicals in our solution have a linear relation. This allows the usage of linear models:

$$Y = X\beta + \epsilon$$

- where Y is light absorbance, X is concentrations, β is molar absorptivity and ϵ is the measurement error in Y.

How one collects data

- Suppose we add G into H for 51 times. At each addition of G, we measure the light absorbance of the whole solution at, say, 301 wavelengths from 200 to 500. Therefore, we end up having 51×301 datapoints for dependent variables (i.e. light absorbances). This is essentially a linear model with multiple outputs. Our independent variables are concentrations of chemicals existing in the whole solution.

- Instead of finding molar absorptivity β, we are more interested in finding the equilibrium constants K of existing chemicals. Reasons in the next page.

Warning: compute X rather than measure X

- Different than normal applications of statistical techniques …
- Why can’t we measure X …
- As a chemist, one would be able to identify significant interactions but may not be able to realize those reactions with equilibrium constants that are too small. Thus, one must assume an m: n model first before any modelling goes on.

What is an m: n model

- In an m: n model, we assume all possible chemicals are:

$$H, G, HG, \ldots, HG_n, H_2G, \ldots, H_2G_n, \ldots, H_mG, \ldots, H_mG_n$$

with equilibrium constants K_{11} (for HG), K_{12} (for HG_2), … and K_{mn} (for H_mG_n).

- Below unless specified, I use 1:1 model and 1:2 data for illustration.

① 1:1 model: only H, G and HG exist.

② 1:2 data: H, G, HG and HG_2 exist.

Problem 1:
Model Misspecification Error

Problem 1: model misspecification error

- Model misspecification error refers to the problem of incorrectly specifying a model. In statistics, a common goal is to search for a correctly specified model, and then perform parametric approaches. However, in out case, the model is inevitably wrong:

‘Our model can only include finitely many predictors, while there could be infinitely many possible chemicals existing. Even though the number of existing chemicals is finite, we cannot realize their existence if they have negligible equilibrium constants.’

- Question: how do we design an approach that returns $\widehat{K}_{mn}$ as close to true K_{mn} as possible?

Current approach: OLS

Current approach

Suppose we identified all significant chemicals by eyes to be H, G and HG (a 1:1 model), but the data was $1:2$ with K_1 and K_2.

I. Guess K_1.

II. Compute $X \coloneqq [H, G, HG]$ using the set of equations:

$$[H] + [HG] = H_t$$
$$[G] + [HG] = G_t$$
$$K_1 = \frac{[HG]}{[H][G]}$$

III. Fit a linear model by $\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$ and compute the RMSE.

RMSE is a function of K_1. Then one iteratively searches for K1 that minimizes RMSE(k).

Problem with the current approach

- The accuracy of the described approach heavily depends on the assumption that the true K_1 minimizes RMSE. This is not necessarily true when K_2 is not so small compared to K_1. This is because OLS tends to drag K_1 from its true value to compensate light absorbances caused by K_2.

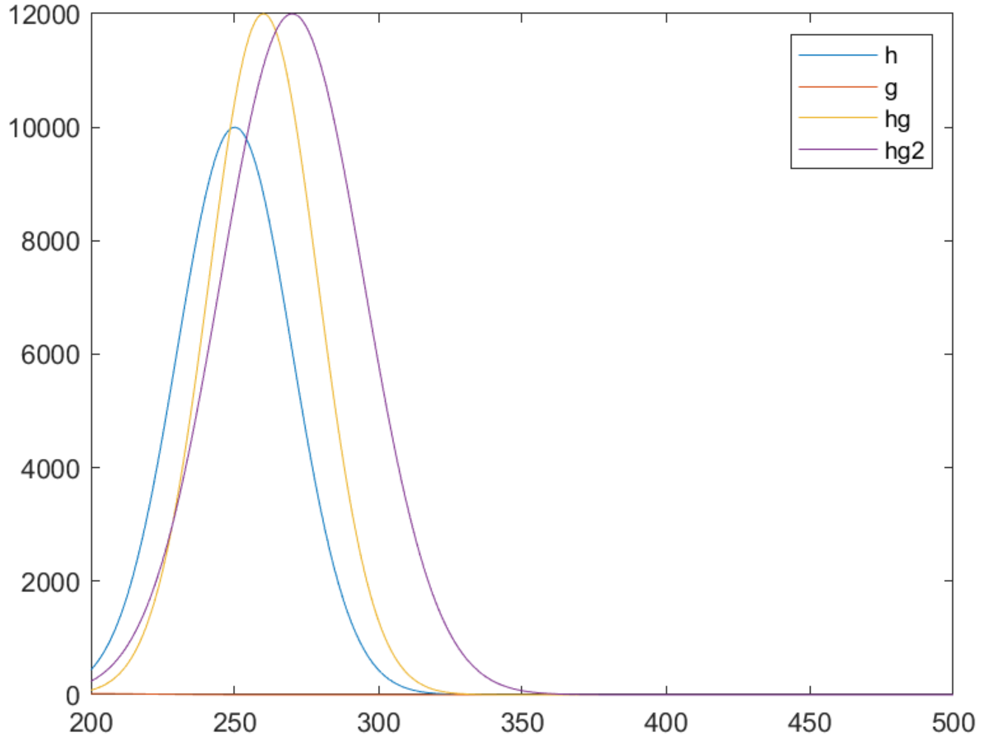
K_1 and K_2 values	**OLS K_1**	**RMSE**
$K_1 = 4100, K_2 = 100$	2743.4	0.0312
$K_1 = 4020, K_2 = 20$	3851.2	0.0287
$K_1 = 4001, K_2 = 1$	4000.1	0.0284

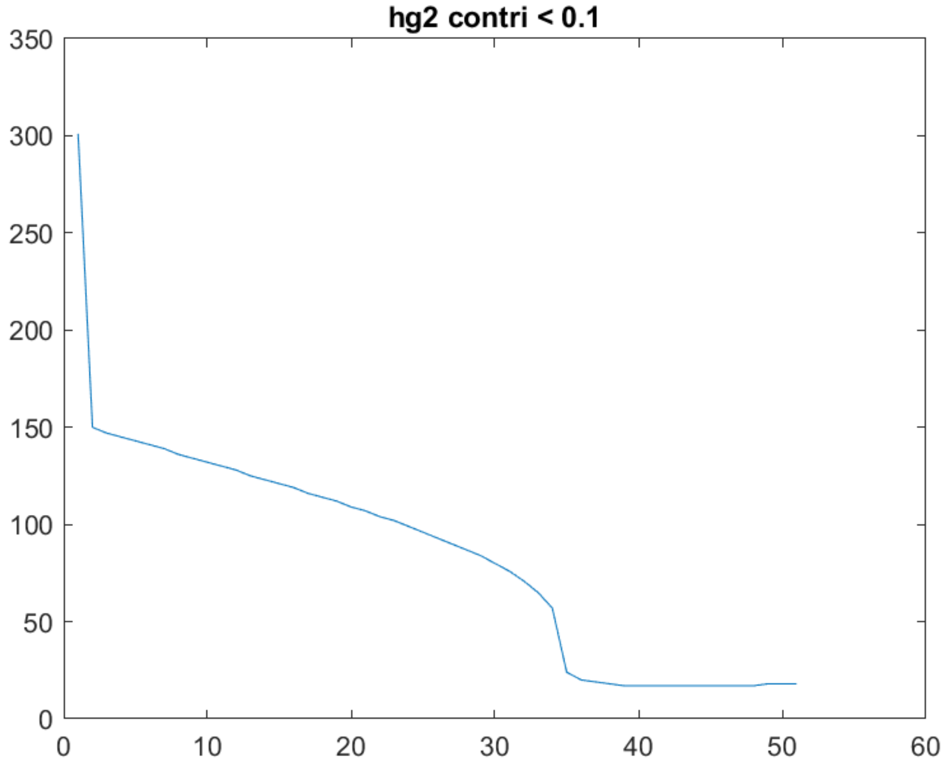
New approach: PE approach

PE approach: error sources

I. Model errors:

Induced by the contribution to light absorbances from HG_2.





PE approach: error sources

II. Independent variables errors:

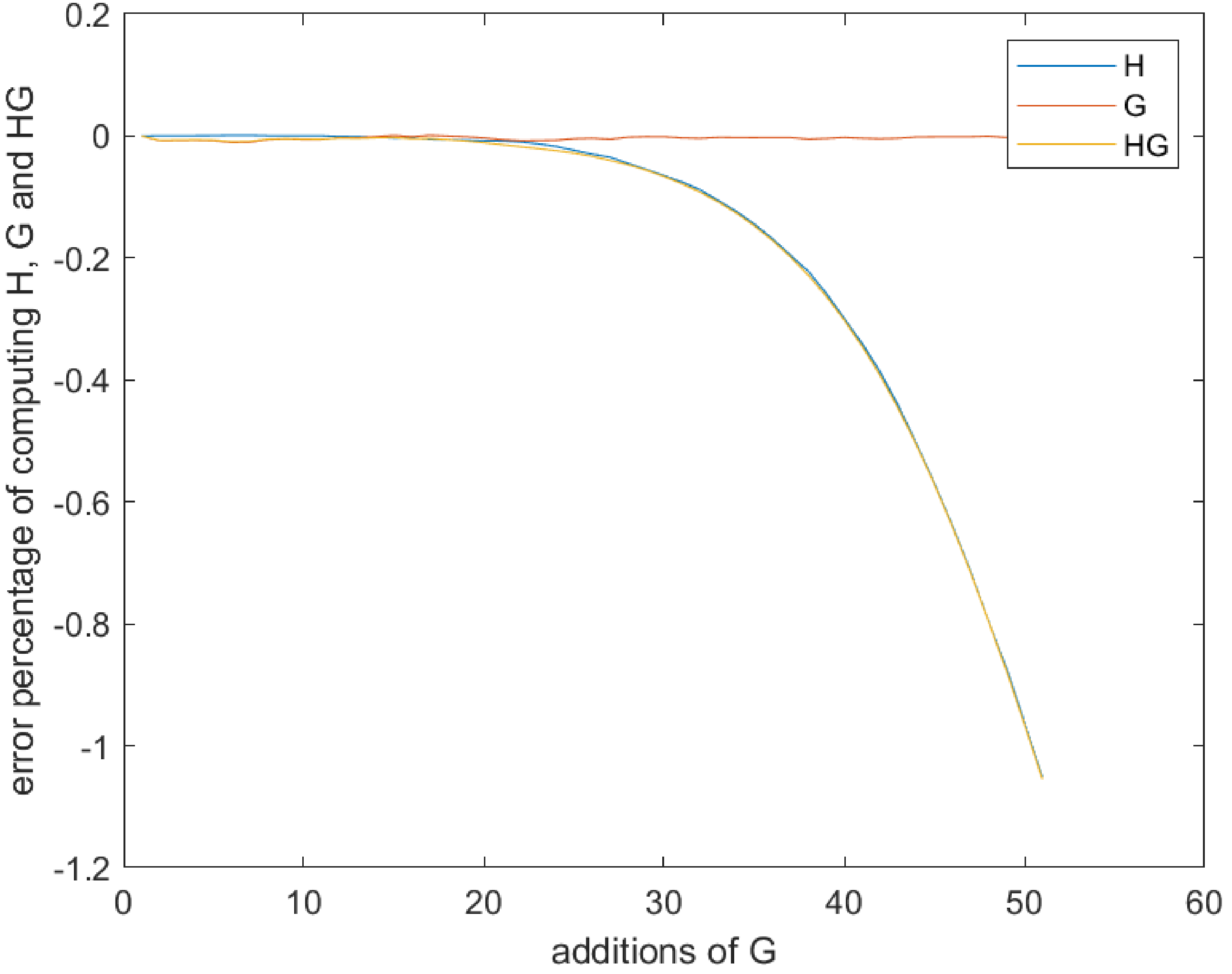
Induced by additions of G. This often occurs even if we used true K_1 to fit models if K_2 is not too small compared to K_1. Algebraically,

$$[\hat{H}] + [\widehat{HG}] = H_t$$
$$[\hat{G}] + [\widehat{HG}] = G_t$$
$$\hat{K}_1 = \frac{[\widehat{HG}]}{[\hat{H}][\hat{G}]}$$

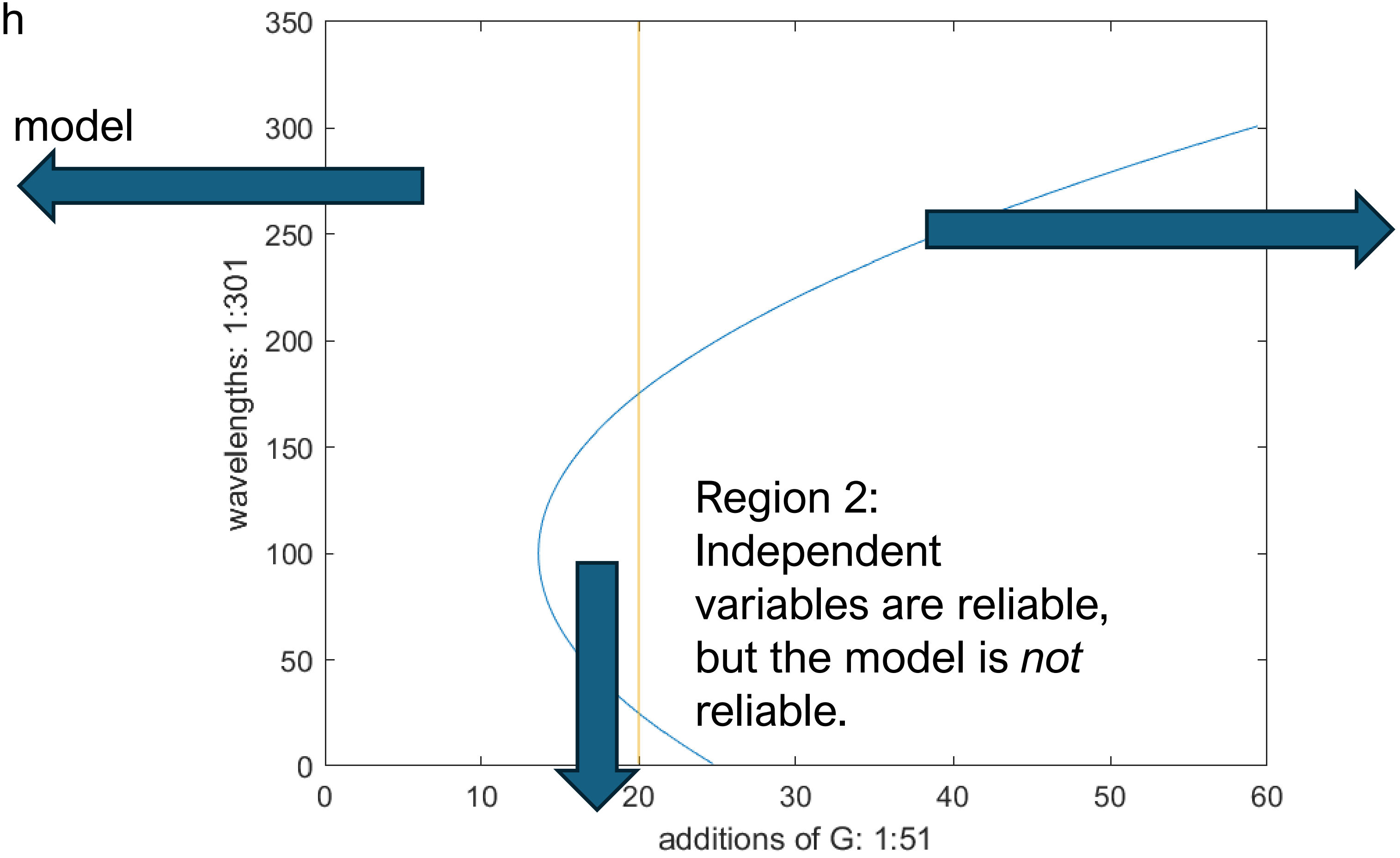
→

$$[H] + [HG] = H_t - [HG_2]$$
$$[G] + [HG] = G_t - [HG_2]$$
$$K_1 = \frac{[HG]}{[H][G]}$$

PE approach: error sources



PE approach: reliability regions



Region 1: both independent variables and model are reliable.

Region 2: Independent variables are reliable, but the model is *not* reliable.

Region 3: Both independent variables and model are *not* reliable.

PE approach: assign weights

- Intuitively, we should assign weights to data points:

Regions	Level of weights
Region 1	Large
Region 2	Moderate
Region 3	Small

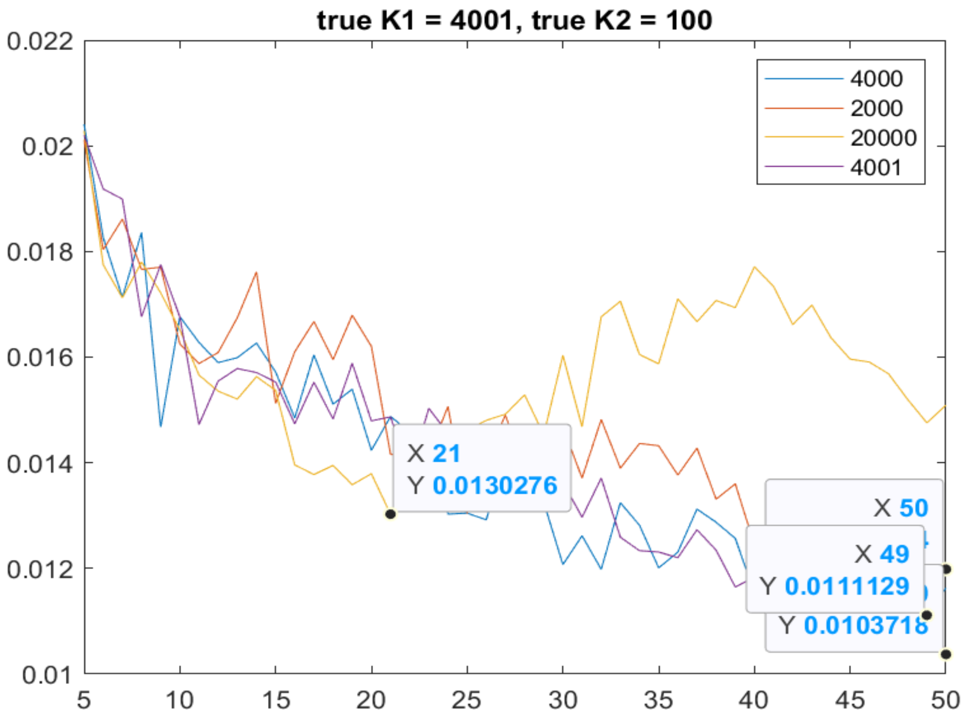
- Question: how do we distinguish regions while we have no information for HG_2?

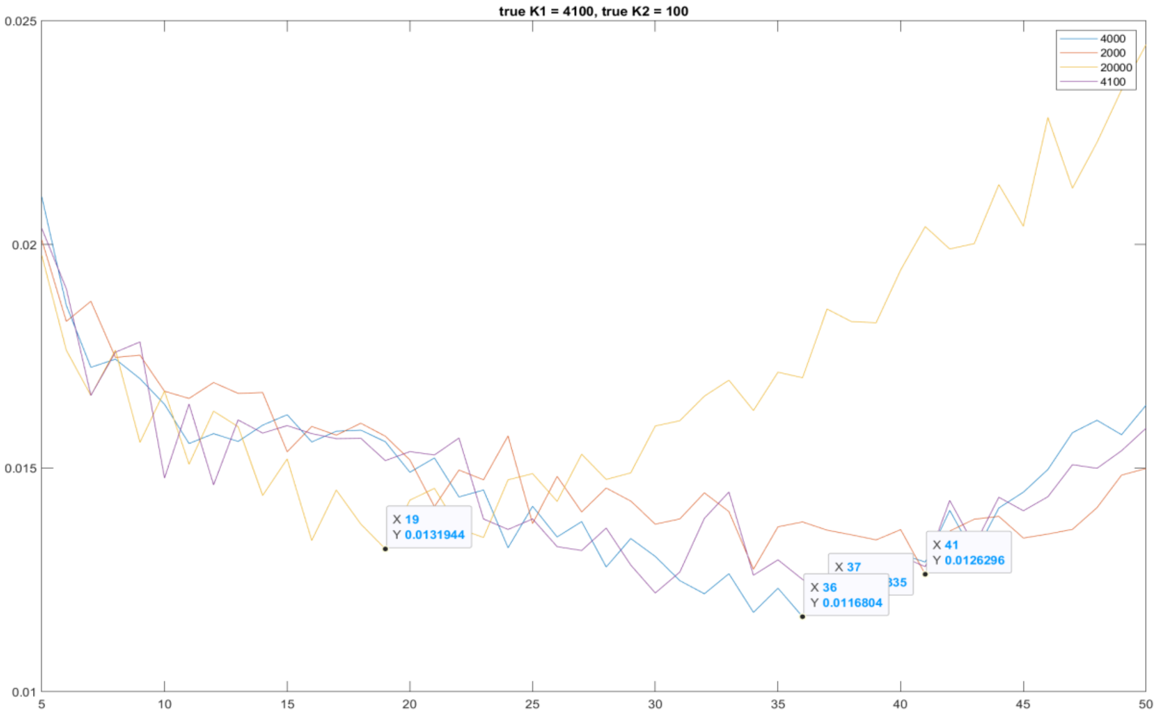
PE approach: assign weights

- Fit a model to the first k samples.
- Use the fitted model to predict the $(k+1)$th sample.
- Track the prediction errors.

(Will later call this the ‘PE approach’).

PE approach: prediction error plot





x axis: additions of G from 1 to 50

y axis: prediction error on the $(x + 1)$th data

PE approach: restrictions and improvements

Restrictions:

I. The effects of not accounting for insignificant but existent chemicals are small until we add G to a certain amount.

II. Measurement errors have smaller variances than influences of HG_2.

III. Have a proper initial guess.

IV. Use a derivative-free minimizer algorithm.

Improvements:

I. Stability: bootstrapping.

PE approach: results

K_1 and K_2 values	OLS K_1 (error percent)	PE K_1 (error percent)
$K_1 = 4100, K_2 = 100$	2743.4 (33.1%)	3837.6 (6.4%)
$K_1 = 4020, K_2 = 20$	3851.2 (4.2%)	3923.5 (2.4%)
$K_1 = 4001, K_2 = 1$	4000.1 (0.02%)	4000.7 (0.007%)

The table below gives results from a dataset where the main issue isn't model misspecification but the presence of many outliers. Although this wasn't our initial focus, our approach should still yield reliable results.

True K_1	OLS K_1	PE K_1
$C \times 10^6$	1.8210×10^6	1.1061×10^6

PE approach: sensitivity to initial guesses and algorithms

- In the example of real data, we tried 3 initial guesses:

I. 1×10^6: this yields $K_1 = 3.3433 \times 10^5$.

II. 1.8210×10^6: this yields $K_1 = 1.1061 \times 10^6$.

III. 1.1061×10^6: this yields $\mathrm{K}_1 = 1.1061 \times 10^6$.

and 2 algorithms:

I. fminsearch (derivative-free): uses the Nelder-Mead simplex algorithm; gives optimal answers at proper guesses.

II. fmincon (gradient-based): stop immediately.

Problem 2: Errors in X

Problem 2: errors in X

- This occurs because …

- Can be mitigated by 1) PE approach and 2) bootstrapping:

I. It accounts for errors in X. If some additions of G are terribly wrong, they are unlikely to give good predictions on the next addition. However, this method also ignores all later additions that might be valuable.

II. Bootstrapping averages random errors in X by resampling.

Problem 3: Outliers in Data

Problem 3: outliers in data

- Such problems happen because Beer lambert law $L = A \times l \times X$ may not hold at some data points due to following reasons:

I. High concentrations …

II. Scattering of light …

III. Path length variability …

Possible approaches

- Wild bootstrap.
- Huber loss.
- Gaussian kernels.
- Spline regression.