Statistical Analysis related to the Inverse problem

Se Hee Kim and Lingsong Zhang

SAMSI/CRSC Undergraduate Workshop

Review of the Inverse problem

- What we get
  - Displacement y_i and time t_i
  - Spring model

$$\frac{d^2 y(t)}{dt^2} + C\frac{dy(t)}{dt} + Ky(t) = 0$$

- Target:

  Estimate C and K based on the observed y_i

Review of the Inverse Problem (cont.)

- Estimation procedure
  - Minimize the cost function

$$J(C,K) = \frac{1}{2}\sum_{i=1}^{N}(y_i - y(t_i, C, K))^2$$

  - Guess initial values of *C* and *K*
  - Using optimization method and differential equations to find the values of *C* and *K* which minimize the above cost function.
  - `inv_beam.m`

Underlying Statistical Models

- y_i has measurement error
  - The above model can be viewed as a regression model

$$y_i = y(t_i, C, K) + \varepsilon_i$$

where ε_i are iid (independent identically distributed) from $N(0, \sigma^2)$.

  - Estimating C and K leads to the procedure mentioned earlier (will show later)

Nonlinear regression

- Linear
  - Linear is for the parameter(s)

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Nonliner
  - $y(t_i, C, K)$ are defined by an ODE, it is not from a simple linear function of C and K
  - A regression model is called nonlinear, if the derivatives of the model with respect to the model parameters depend on one or more parameters

Nonlinear regression

- A regression model is not necessarily nonlinear if the graphed regression trend is curved
  - Example: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$
  - Take derivatives of *y* with respect to the parameters β_0, β_1 and β_2:
    $\partial y / \partial \beta_0 = 1, \partial y / \partial \beta_1 = x, \partial y / \partial \beta_2 = x^2$
  - None of these derivatives depend on a model parameters, thus the model is linear.

Nonlinear regression

- The general form of a nonlinear regression model is

$$y = \eta(x, \beta) + \varepsilon$$

- Where x is a vector of explanatory variables, β is a vector of unknown parameters and ε is a N(0, σ^2) error term
- To estimate unknown parameters,

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \eta(x_i, \beta))^2$$

Statistical problems

- How to evaluate the estimation of C and K?
  - Estimation of σ.
  - Variation of the $\widehat{C}$ and $\widehat{K}$.
- Are those assumptions correct?
  - Measurement errors are truly from iid Normal distribution?
  - Are there better models?

Estimation of σ

- If all model assumptions hold
  - we have $\widehat{\varepsilon}_i = y_i - y(t_i, \widehat{C}, \widehat{K})$

    Use these to estimate σ

$$\widehat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \widehat{\varepsilon}_i^2$$

Our data set gives

$$\widehat{\sigma} = 1.0407 \times 10^{-5}$$

YOURS?

`estimateofsigma.m`

Evaluate the Estimation of *C* and *K*

- Method 1- repeating experiments
  - Independent experiments under the same conditions
    - How many experiments required?

      we will use 8 sets of data to illustrate this method

  - Use simple univariate statistics of *C*'s and *K*'s to evaluate the performance of estimation.

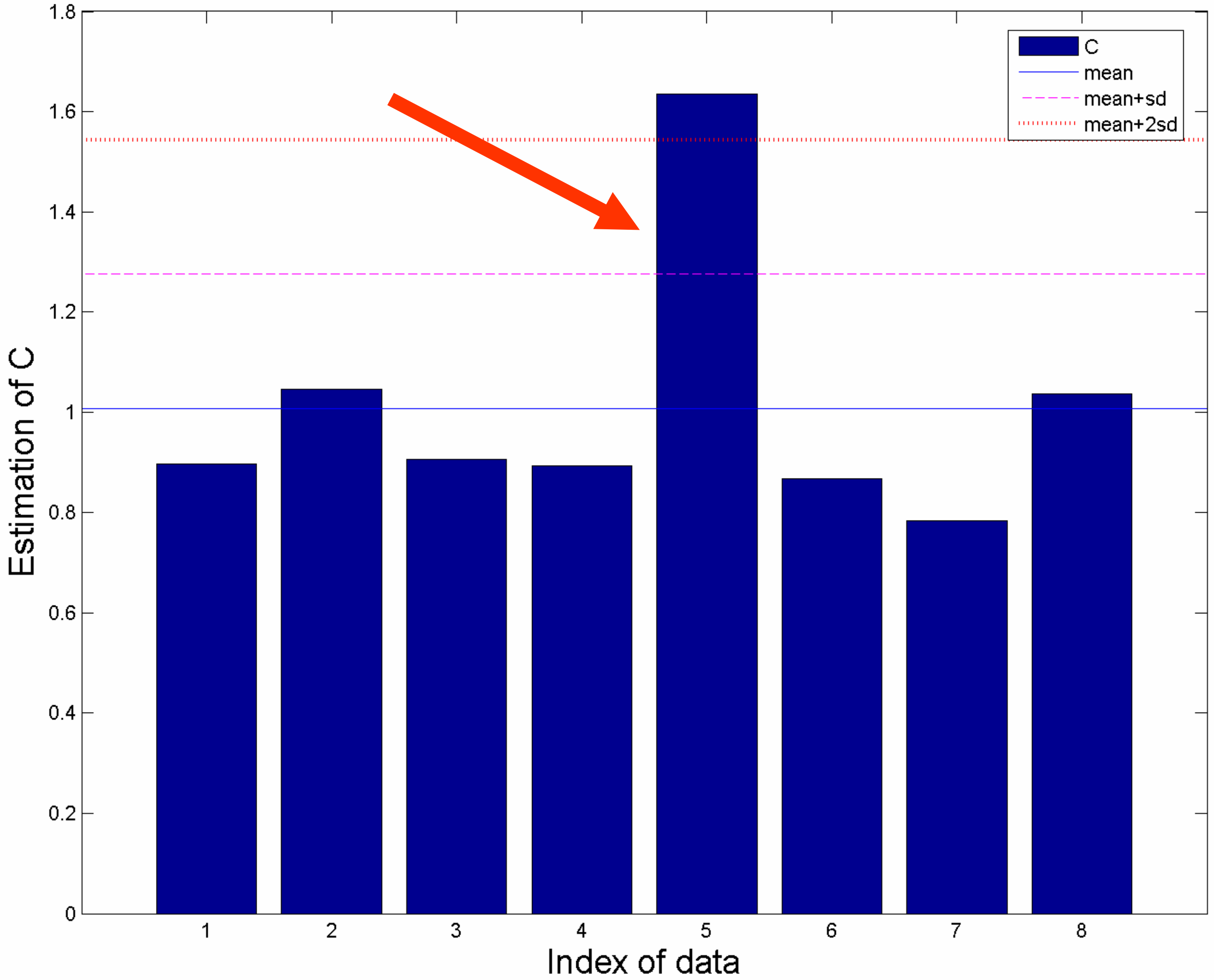
Examples

- 8 sets of data from the same conditions
  - Get the estimations of *C* and *K*
  - Variations of *C* and *K*

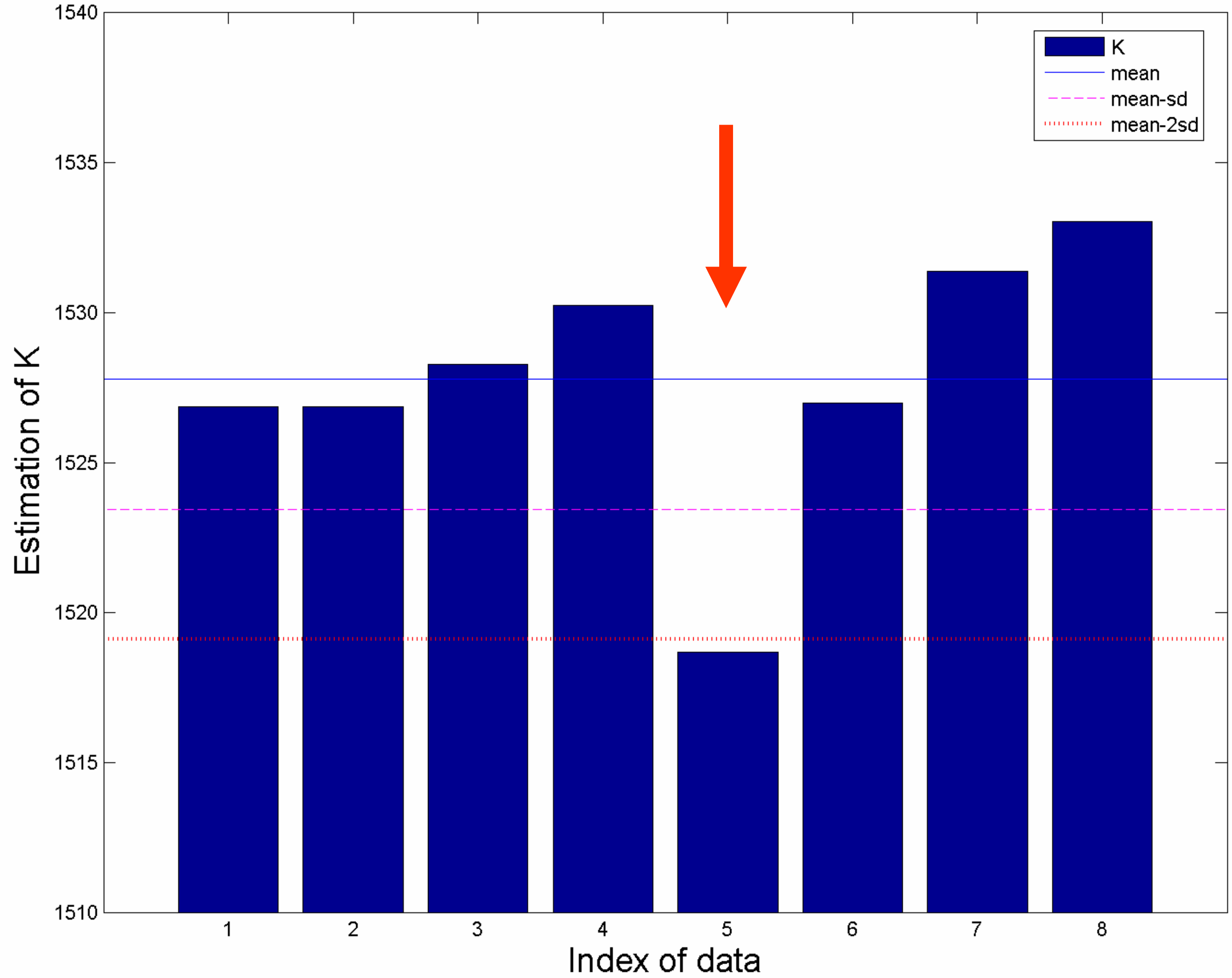
    sd(*C*)= .2675

    sd(*K*)=4.3437

C	*K*
0.89710	1526.9
1.04680	1526.8
0.90619	1528.3
0.89321	1530.2
1.63440	1518.7
0.86776	1527.0
0.78324	1531.4
1.03730	1533.0
1.00820	1527.8



1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Estimation of C
1
2
3
4
5
6
7
8
Index of data
C
mean
mean+sd
mean+2sd



1540
1535
1530
1525
1520
1515
1510
Estimation of K
1
2
3
4
5
6
7
8
Index of data
K
mean
mean-sd
mean-2sd

Examples

- 8 sets of data from the same conditions
  - Get the estimations of C and K
  - Variations of C and K

    sd(C)= .2675

    sd(K)=4.3437

After removing the "outlier"

sd(C)=.0937, sd(K)=2.4828

C	K
0.89710	1526.9
1.04680	1526.8
0.90619	1528.3
0.89321	1530.2
1.63440	1518.7
0.86776	1527.0
0.78324	1531.4
1.03730	1533.0
1.00820	1527.8

Exercise

- To get the standard deviations from your 10 data sets (`std`)
  - If you do not have the record of estimations,
    - Use `inv_beam_all.m` to get `Cvec`, `Kvec`,

      (it will take a long time, do it during the break)
  - If you have the record of the estimations
    - Input them into `Cvec` and `Kvec`,
- Bar-chart?
  - Try `barCvec.m` and `barKvec.m` (need to adjust some numbers)

Evaluate the Estimation (cont.)

- Method 2 - Using the nonlinear regression model to evaluate the estimation of *C* and *K*.
  - How to get the standard deviation (or variance) of $\widehat{C}$ and $\widehat{K}$?

    Use the covariance matrix of $(\widehat{C}, \widehat{K})$

Recall: Linear Regression

- Simple Linear Model:
  - The model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

we estimate the covariance matrix of β_0 and β_1 using $\mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = (X'X)^{-1}\widehat{\sigma}^2$

where

$$X = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} = \begin{pmatrix} \frac{\partial Y_1}{\partial \beta_0} & \frac{\partial Y_1}{\partial \beta_1} \\ \frac{\partial Y_2}{\partial \beta_0} & \frac{\partial Y_2}{\partial \beta_1} \\ \vdots & \vdots \\ \frac{\partial Y_n}{\partial \beta_0} & \frac{\partial Y_n}{\partial \beta_1} \end{pmatrix}$$

Apply to the inverse problem

Our model $y_i = y(t_i, C, K) + \varepsilon_i$

we will have similar result

$$\mathsf{Cov}(\widehat{C}, \widehat{K}) = (X'X)^{-1}\widehat{\sigma}^2$$

where

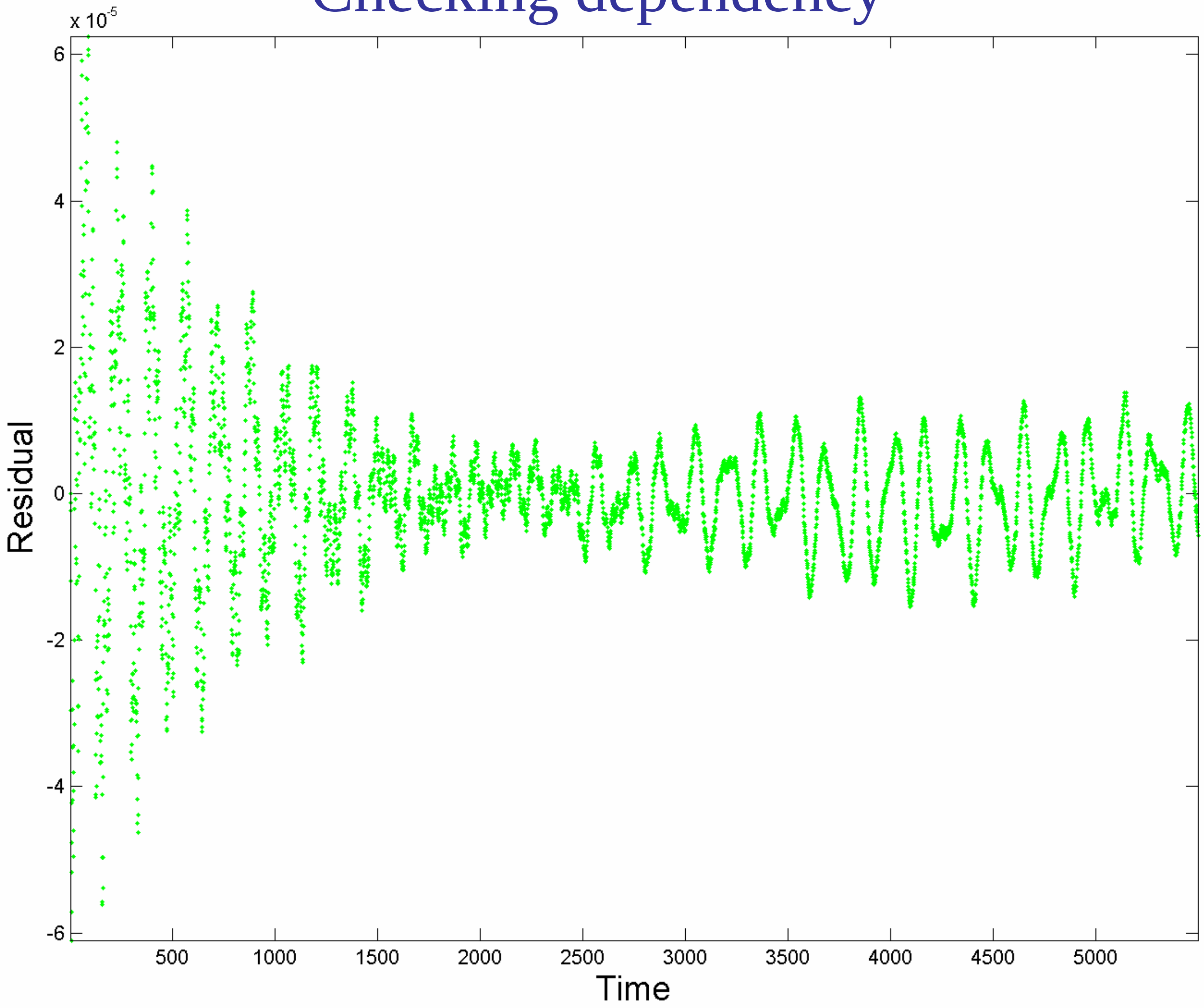
$$X = \begin{pmatrix} \frac{\partial y(t_1)}{\partial C} & \frac{\partial y(t_1)}{\partial K} \\ \frac{\partial y(t_2)}{\partial C} & \frac{\partial y(t_2)}{\partial K} \\ \vdots & \vdots \\ \frac{\partial y(t_n)}{\partial C} & \frac{\partial y(t_n)}{\partial K} \end{pmatrix}$$

the standard errors of $\widehat{C}$ and $\widehat{K}$ are square roots of the diagonal elements of $\mathsf{Cov}(\widehat{C}, \widehat{K})$

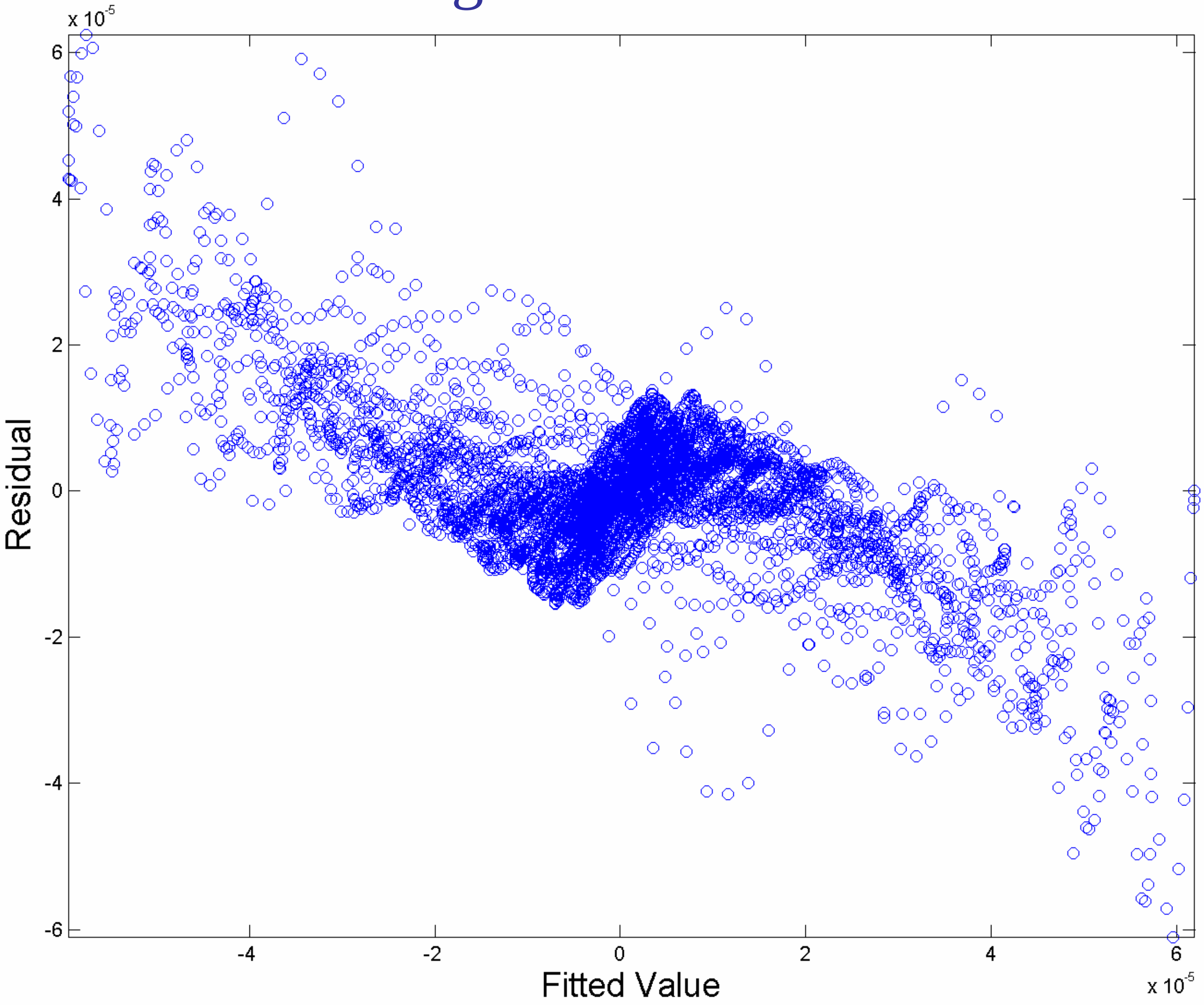
Checking the Assumptions

- Check whether the residuals are iid Normal noise
  - Independent?
    - Residual plot vs Time
  - Variance are constant?
    - Residual plot vs. fitted values
  - Residuals are Normally distributed?
    - Normal Quantile-Quantile plot

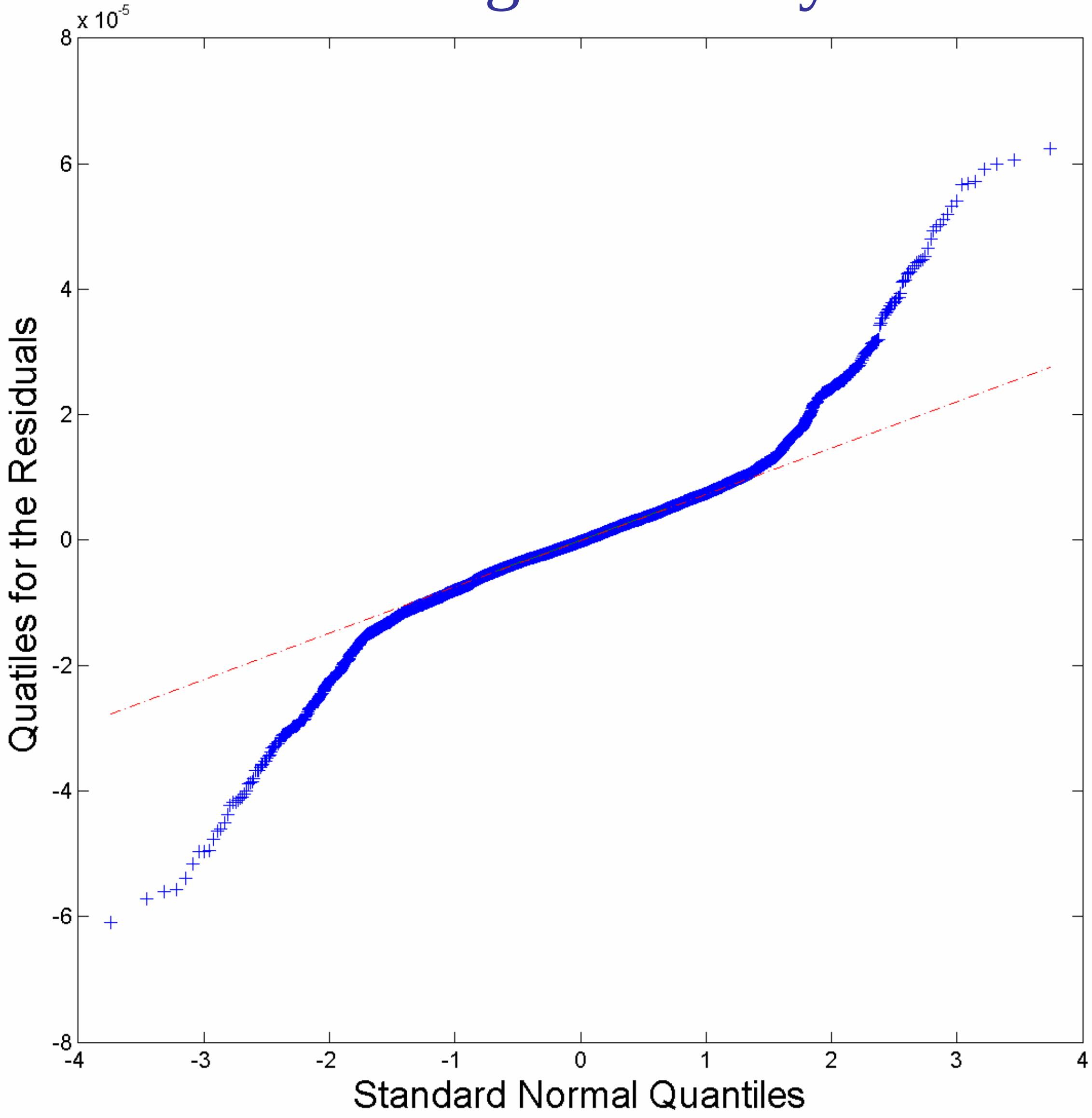
Checking dependency



Checking Constant Variance



Checking Normality



Other models?

- Is the spring model appropriate for our data?
  - Residual shows dependent structure
  - It seems that variances are not equal
  - Normal assumption might not hold
- Other statistical inference method?
  - Same underlying model, but different assumptions
  - Other statistical models to fit the data
- Some alternative physical models for our data?