Modelling Change: Incorporating Dynamic Components into Data Analysis

Jim Ramsay and Theo Koulis
Department of Psychology

McGill University

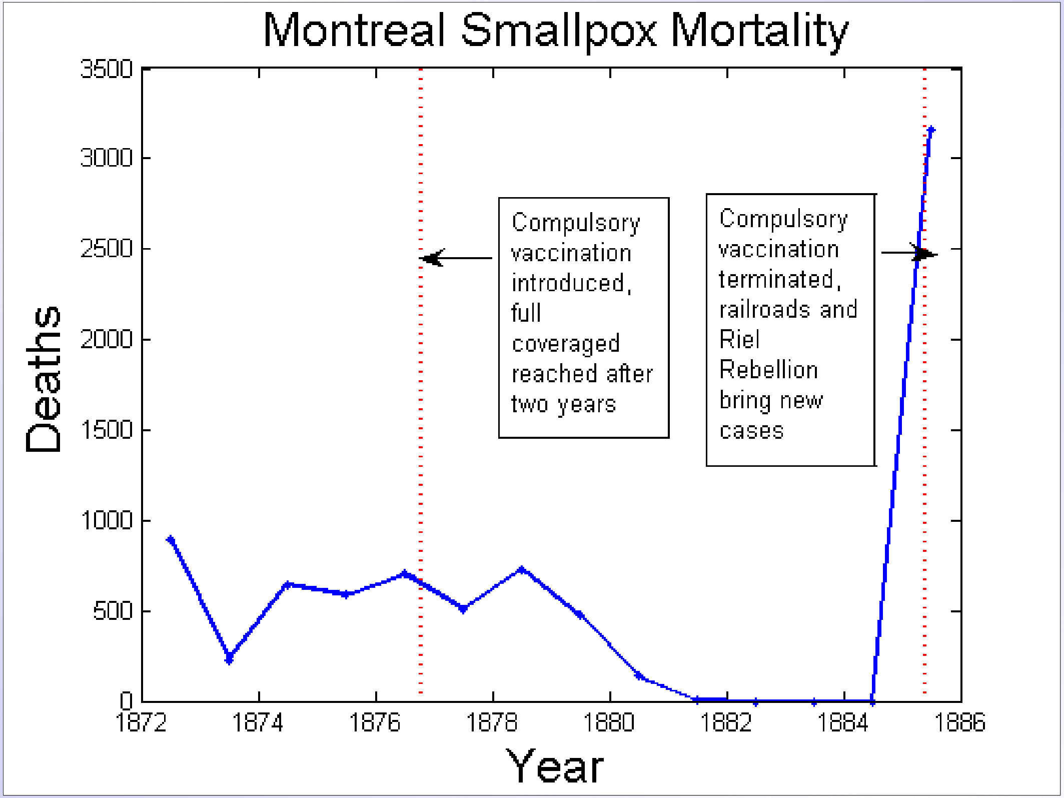
1. Introduction: Input/Output Systems

- We often collect data on units over time.
- There is an *output measure* $y_i(t)$ that reflects the status of a unit i at time t.
- There are also *input measures* $z_{ij}(t), j = 1, \ldots, p$ that indicate the status of various variables thought to affect the output measure.
- We want to study how the status of these units responds to changes in the input variables.
- We especially want to know how a *change* in an input determines the *change* in output.

Examples

- How is driving performance affected by a couple drinks?
- How are golf scores affected by the purchase of a new set of clubs?
- How is pain intensity affected by a dose of morphine?
- How does tumour size respond to radiotherapy?
- How does a couple's social life respond to the birth of a child?
- How does mortality or the incidence of asthma change with an increase in ozone, particulate matter, or other airborne pollutants?

2. Montreal Smallpox Mortality



Inputs: Vaccination coverage, infection from outside the city

3. Functional Regression Analysis

- This sounds like a regression analysis problem that varies over time t.

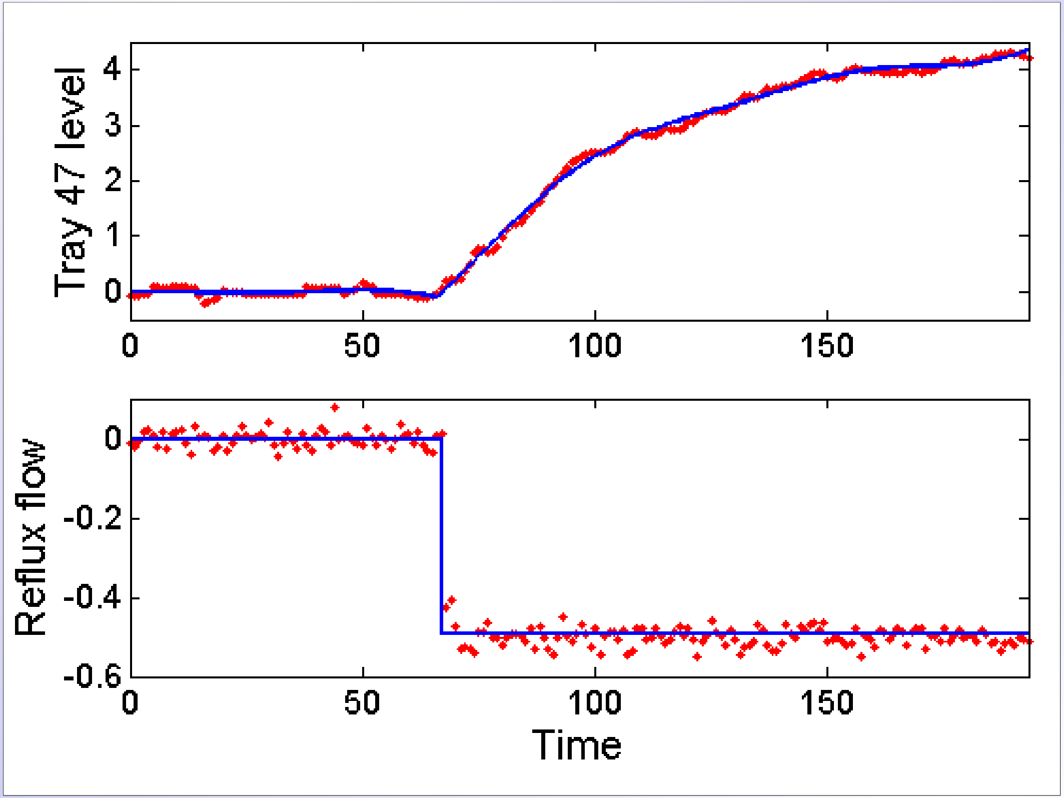
$$y_i(t) = \beta_0(t) + \beta_1(t)z_{i1}(t) + \ldots + \beta_p(t)z_{ip}(t) + \epsilon_i(t)$$

- The regression coefficients $\beta_j(t)$ are now functions of time.
- Software for estimating these regression coefficient functions is readily available. See Ramsay and Silverman (2005) *Functional Data Analysis*, Springer, and the website www.functionaldata.org.
- The model is also a variant of the generalized additive or GAM model.

4. The oil refinery data

- This is a simple input/output system in an oil refinery in Corpus Christi, Texas.
- A fluid, called reflux, flows into a tray in a distillation column in an oil refinery.
- The input variable $z(t)$ is the flow rate.
- The level of fluid in the tray is the output variable $y(t)$.

Refinery output $y(t)$ (top panel) and input $z(t)$ (bottom panel)

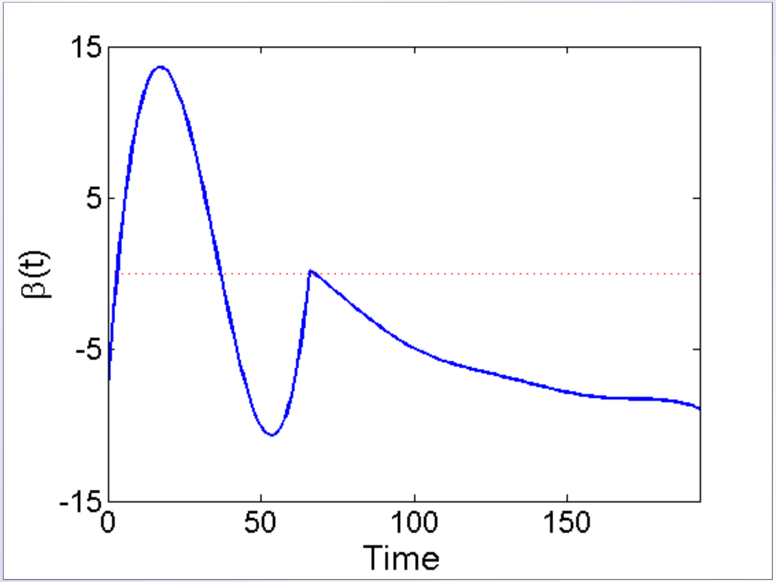


Variation on two time scales

- Over the longer scale, tray level changes from an initial level to a final level.
- But we are also interested in how rapidly the change takes place; that is, short-scale variation.

A functional regression model

$$\texttt{Tray}(t) = \beta(t)\texttt{Reflux}(t) + \epsilon(t)$$



But what does this tell us?

Adding the derivative $D\texttt{Tray}(t)$ to the output

- Can we also model the *rate of change* in the output, as reflected by the first derivative $D\texttt{Tray}(t)$?
- Suppose that we model a mixture of the *rate of change* in the output and the the output itself.
- We'll use constants for the regression functions in the hope of keeping things simple.

$$D\texttt{Tray}(t) + \gamma\texttt{Tray}(t) = \beta\texttt{Reflux}(t) + \epsilon(t)$$

- Coefficient γ controls the relative emphasis on fitting the derivative of the output versus fitting the output itself.
- We estimate $\gamma = 0.02$ and $\beta = -0.20$.

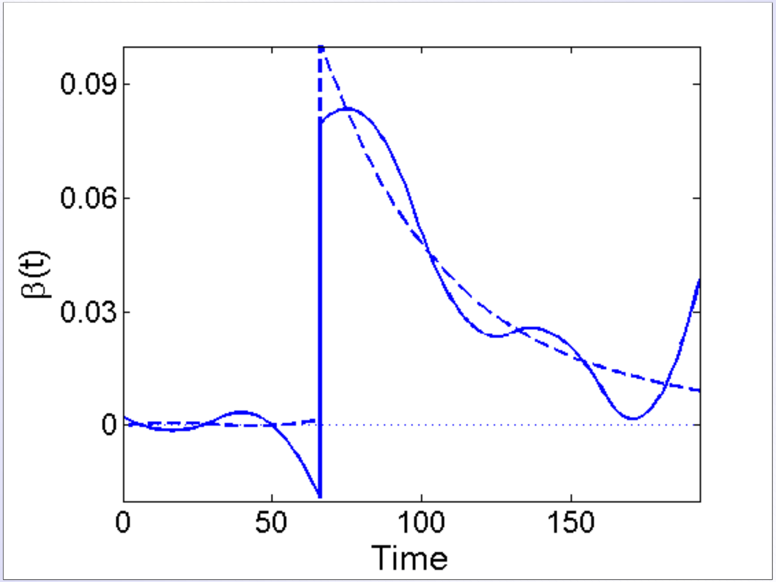
Expressing the model as a Differential Equation

- Models involving derivatives are called *differential equations*.
- They are usually expressed in this rearrangement of our model:

$$D\texttt{Tray}(t) = -\gamma\texttt{Tray}(t) + \beta\texttt{Reflux}(t) + \epsilon(t)$$

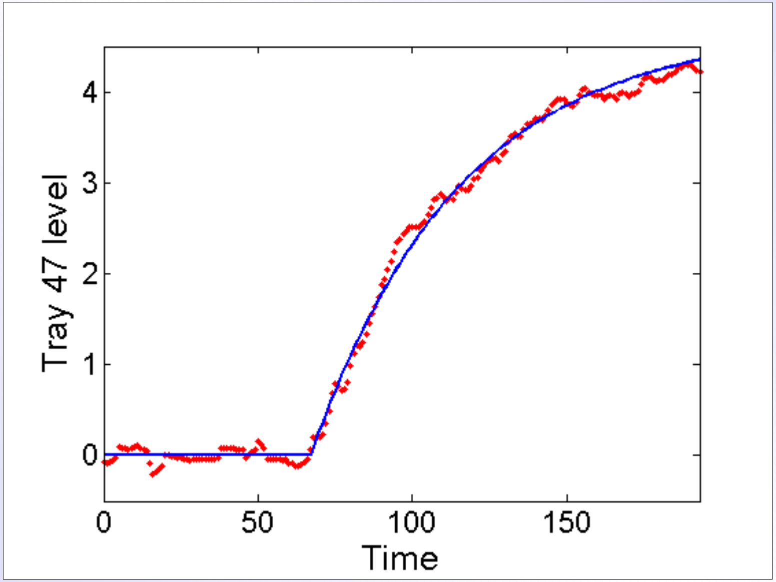
- Input $\texttt{Reflux}(t)$ is called a *forcing function*.

The derivative $D\texttt{Tray}(t)$ and its estimate



The solid line is the derivative estimated from the data, and the dashed line is the model's fit to this derivative.

The fit to the data

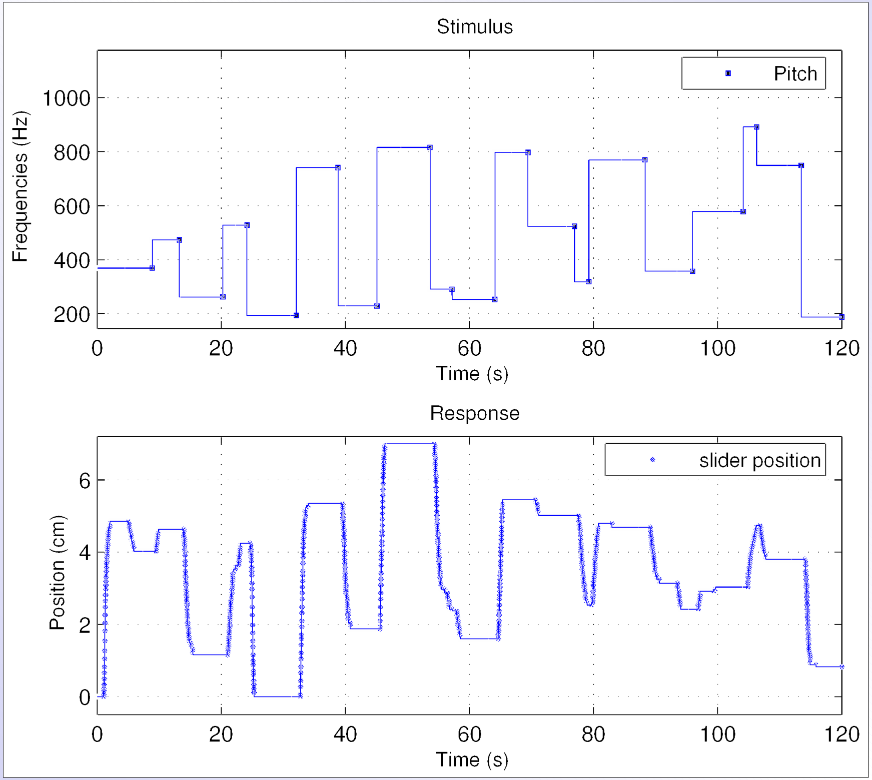


This seems impressive given only two parameters.

5. A Psychoacoustics Experiment

- This is an input/output system in music cognition.
- Subjects are asked to follow a series of sequential pitches.
- Subjects adjust a slider on a computer input device (potentiometer).
- If the pitch increases $\longrightarrow$ slider position is increased.
- If the pitch decreases $\longrightarrow$ slider position is decreased.

Input $z(t)$ (top panel) and slider output $y(t)$ (bottom panel)

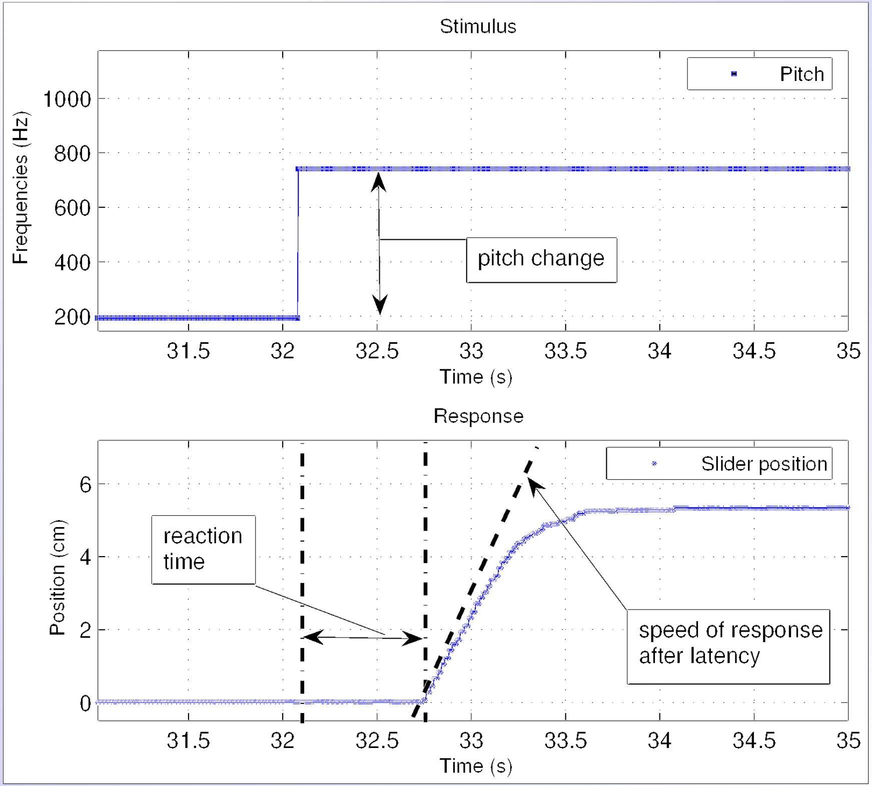


Features of the Slider Data

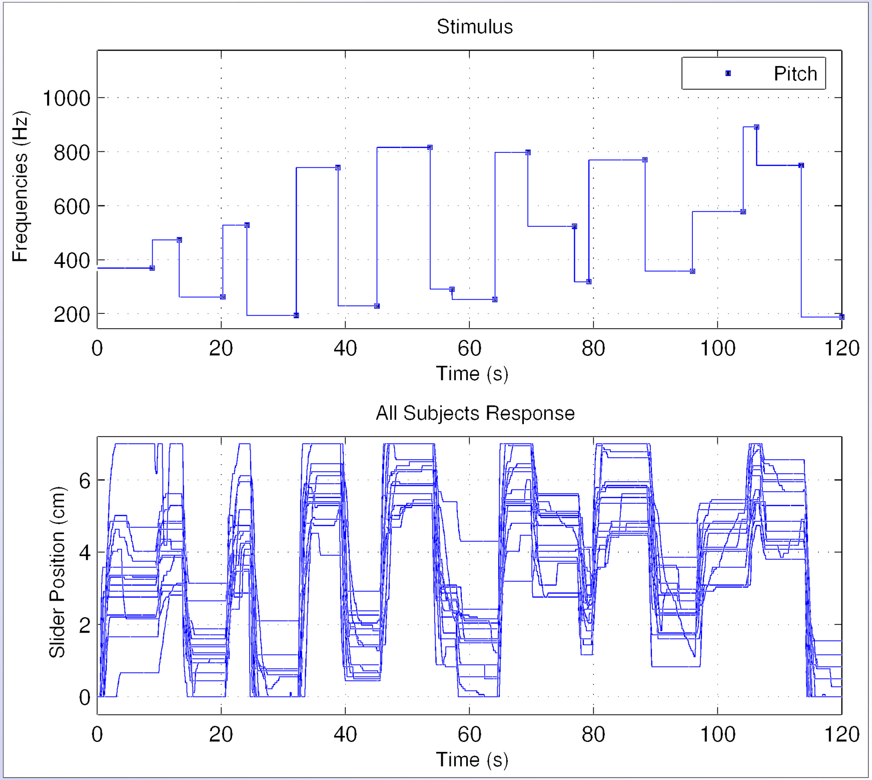
Psychologists are interested in 3 features of the data.

- Reaction Time: the latency between the onset of a fixed stimulus and the response to it.
- Response Speed: a measure of how fast a subject implements the response to the stimulus.
- Gain: the amount of "energy" required to get to a steady state. It is the ratio of "output to input".

Features: Example from Data



Common stimulus (top panel) and output $y(t)$ (bottom panel: all subjects)



Primary Goal

- The input variable $z(t)$ is the pitch.
- The position of the slider is the output variable $y(t)$.
- There is a lot of variation across subjects.
- Both inter-subject and intra-subject variation.
- Our goal is to quantify this variation to facilitate comparisons (Calibration)

Using Derivatives: A First Approach

- A simple 3-parameter model:

$$D\texttt{Slider}(t) = -\gamma\texttt{Slider}(t) + \beta\texttt{Pitch}(t-\delta) + \epsilon(t)$$

- $\texttt{Slider}(t)$ is the output and $\texttt{Pitch}(t)$ is the forcing function
- Parameters: γ, β, δ
- How do the parameters correspond to the features of the data?
- The parameter δ is the reaction time.
- What about the other two?

Defining the Gain and Response Speed

- Consider the differential equation

$$D\texttt{Slider}(t) = -\gamma\texttt{Slider}(t) + \beta\texttt{Pitch}(t-\delta)$$

  with initial condition $Slider(0) = 0$.
- $\texttt{Pitch}(t)$ is a step function:

$$\texttt{Pitch}(t) = \begin{cases} 0 & \text{if } t < 0 \\ P & \text{if } t \geq 0 \end{cases}$$

- P is the change in pitch.

- The solution is

$$\texttt{Slider}(t) = \begin{cases} 0 & \text{if } t < \delta \\ \frac{\beta}{\gamma} P \left(1 - \exp\left\{-\gamma(t-\delta)\right\}\right) & \text{if } t \geq \delta \end{cases}$$

- The slider position starts at 0 and increases to a limiting value:

$$\texttt{Slider}^* = \frac{\beta}{\gamma} P$$

- The ratio

$$G = \frac{\beta}{\gamma} = \frac{\texttt{Slider}^*}{P}$$

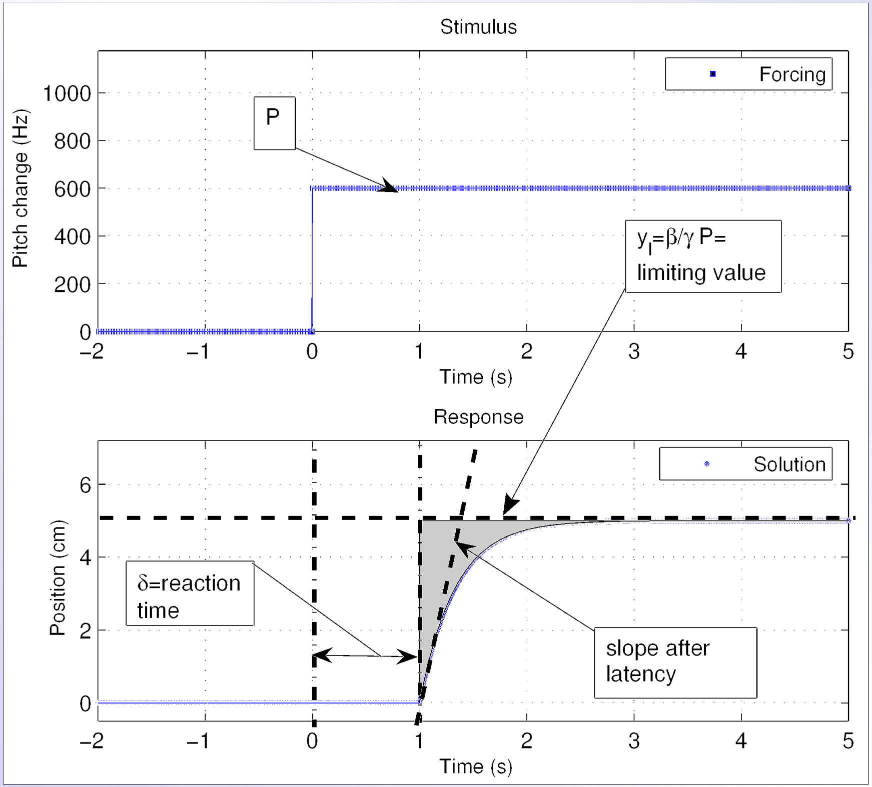
relates the input to the output. We call the ratio G the gain.

- After $\frac{1}{\gamma}$ time units, $\texttt{Slider}(t)$ has reached $2/3$ of the final value $\texttt{Slider}^*$.
- After $\frac{2}{\gamma}$ time units: $7/8$ of the final value.
- After $\frac{4}{\gamma}$ time units: 98% of the final value.
- For this reason, we call the ratio

$$\tau = \frac{1}{\gamma}$$

  the response time constant.
- The parameter γ is called the response speed.

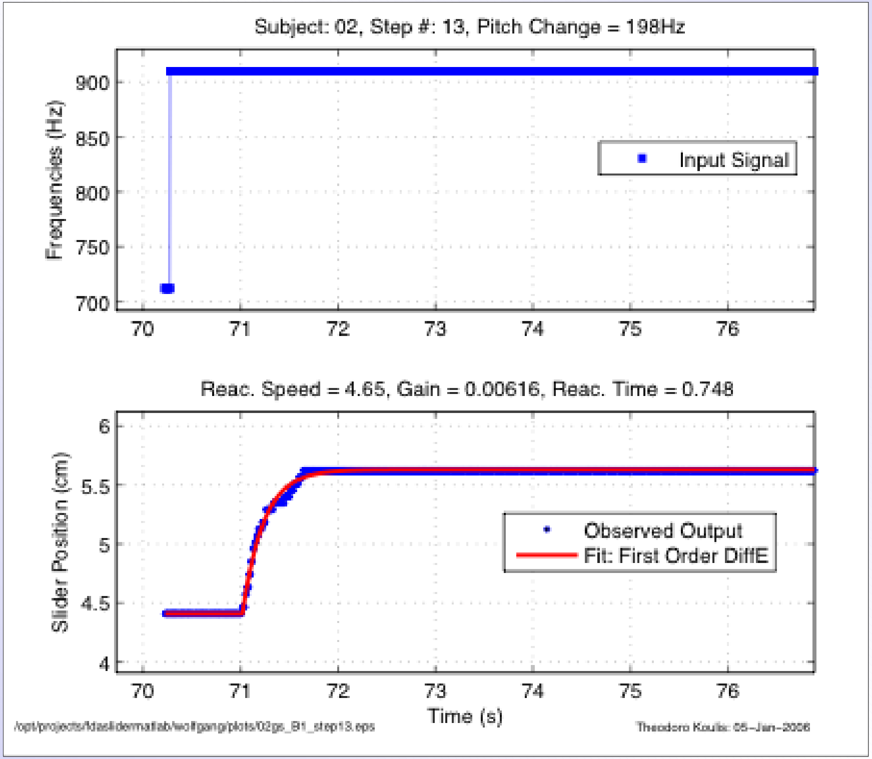
Example: $P = 600$, $\delta = 1$, $\gamma = 3$, $\beta = 0.025$



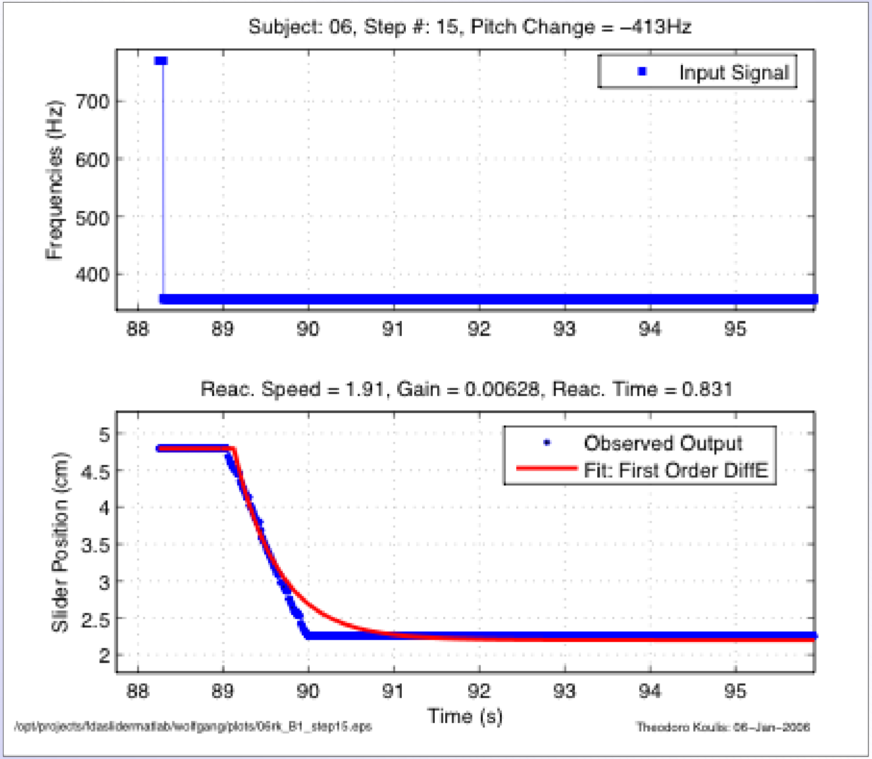
Fitting the Data

- The model does a good job at capturing the shape of the data curves.
- For most cases, the model seems adequate.
- For a few cases, the model does not fit well.
- Even so, we want to keep the simple model to make interpretation easy.

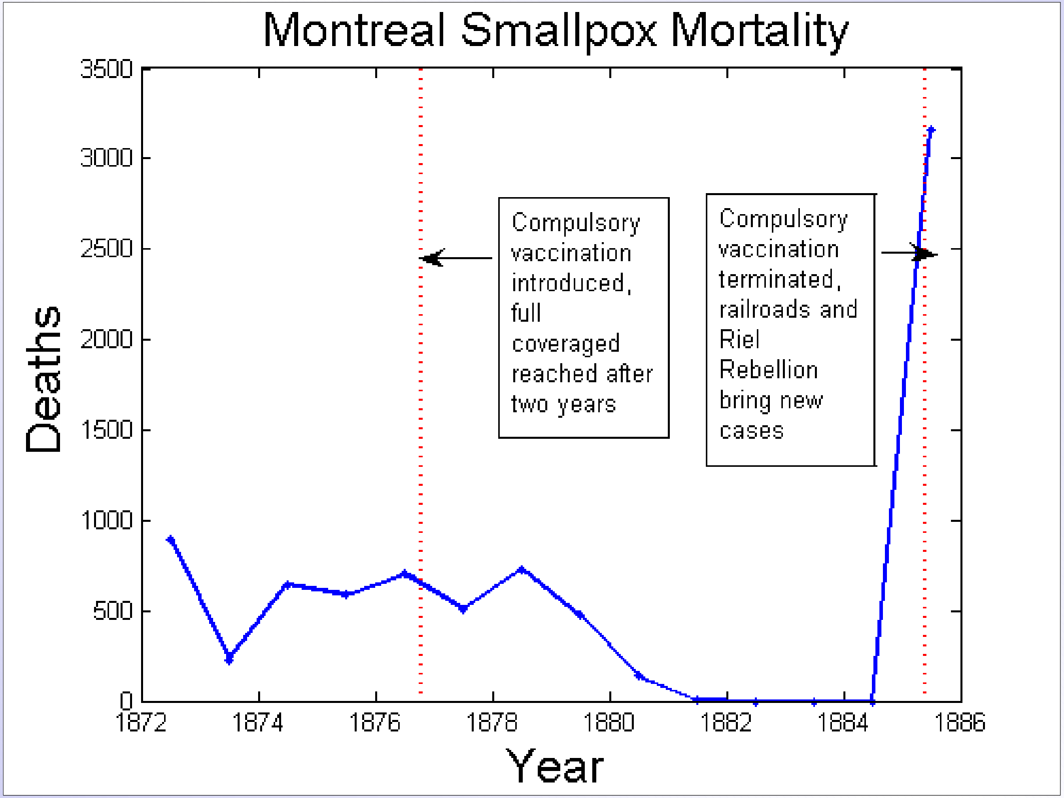
Fitting the Data: Example 1



Fitting the Data: Example 2



6. Montreal Smallpox Again



- There is a delay δ of about two years before enough vaccination coverage is reached to be effective.
- Then the disease all but disappears in two years, suggesting a time constant $\tau = 6$ months.
- The epidemic in 1885 goes from just detectable in April to full force in October, suggesting no delay and a time constant of $\tau = 1.5$ months.
- Once the epidemic was obvious to all, full vaccination coverage was almost immediate, and the disease was under control by the end of the year.
- What's most exciting about the smallpox data is the *rate of change* or *dynamics* of the system.

7. How do I go about modelling change?

Consider that there are three basic features of how a system responds to a change in input:

- How quickly does the change take place? ($4/\gamma = 4\tau$ time units)
- How much change happens? ($\beta/\gamma = \beta\tau$ output units per input unit)
- How long before the change begins? (δ time units)

There are other things to model, too, but these are the big three.
More exotic characteristics of how the output responds to a change in input might require the use of higher order derivatives, such as $D^2y(t)$ and etc.

Extending the basic regression equation

- Set down a time-varying regression equation, with the output $y(t)$ on the left side and various inputs $z_j(t)$ on the right. Some of the inputs can, of course, be constant.
- Each input is multiplied by its regression coefficient function $\beta_j(t)$, which, of course, can be constant if desired.
- Now consider replacing the output $y(t)$ by a mixture or linear combination of $y(t)$ with one or more of its derivatives, $Dy(t), D^2y(t)$ and etc. $y(t)$ and other lower-order derivatives are multiplied by weight functions $\gamma(t)$.
- Add delay parameters as required.

How do I actually fit data with a dynamic model?

- Visit the website www.functionaldata.org to find software in R, S and Matlab along with worked examples.
- Consider buying *Functional Data Analysis*; all the analyses illustrated in the book are also available on the website.