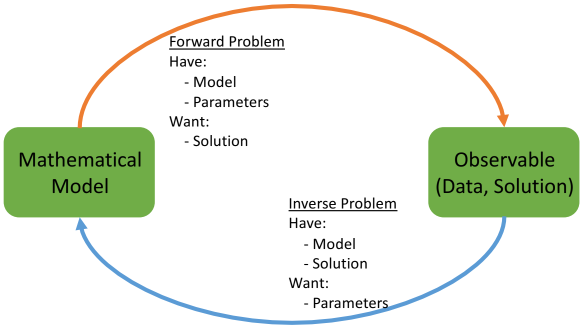
A Tutorial on Parameter Estimation

Introduction

In this post, we will consider how to implement the inverse problem in order to obtain the parameters in our model that best fits our data. This will be broken into several smaller steps: the discussion of statistical error models, formulation of an error function to minimize, how to check that your statistical error model is correct, and an example.

The first question is *what is an inverse problem?*



A *forward* problem is when we have a mathematical model and the associated parameters and we want to generate a solution (orange arrow). An *inverse* problem is when we have a mathematical model and a solution (or observable) and we want to find the parameters that generated that solution (blue arrow).

Notation throughout the rest of this post is as follow:
$y' = f(t;\theta)$: your mathematical ODE model

Although we are only considering one ODE model here, we note the process is similar for systems of ODEs as well as PDEs.

Statistical Error Models

When we create a mathematical model describing a biological or physical phenomenon, there is often error associated with not only the model formulation, but there is inherent error in the data collection process, or “observation error”. We model this noisy data as:

$$Y_j = y(t_j;\theta_0) + h_j * \mathcal{E}_j$$

where Y_j is the solution to the statistical error model at time t_j, $y(t_j;\theta_0)$ is the solution to your mathematical model at time t_j with the ‘true’ parameters θ_0, h_j determines the scaling of the observation error, and $\mathcal{E}_j$ is observation error, assumed to be i.i.d., i.e., $\mathbb{E}[\mathcal{E}_j] = 0$.

So what does h_j mean? Well, let's consider two different examples.

Example 1

Imagine your data comes from a heat rod experiment. A temperature source is applied to one end of a rod and you are measuring the heat along the rod at specific time points and specific locations. The thermometer being used is accurate to 1°C. What is the error if the true temperature is:

1. 10°C?
2. 100°C?

In both cases, your measurement should be accurate to within 1°C! This is an example of *constant* error, in which the error is **independent** from the quantity of interest. In this case:

$$Y_j = y(t_j; \theta_0) + h_j * \mathcal{E}_j$$

where $h_j = 1$

Example 2

You are modeling the growth of cancer cells in a petri dish. You are tasked with counting the number of cells in the petri dish on each day of the experiment. What do you imagine the error might be on:

1. Day 1, when there are 100 cells?
2. Day 3, when there are 1000 cells?

I would hazard a guess that you might miss about 5 cells on day 1, but you are likely to miss or over-count by more than 5 cells on day 3, by maybe 50 cells. This is an example of *proportional* error, in which the error **depends** on the quantity of interest. In our example, the error is approximately 10% of the population. In this case:

$$Y_j = y(t_j; \theta_0) + h_j * \mathcal{E}_j$$

where $h_j = y(t_j; \theta_0)^{\gamma}$. Note that the GLS framework simplifies to OLS if $\gamma = 0$.

Error Function to Minimize

Now we have a mathematical model and statistical model. In order to find the estimated parameters, $\hat{\theta}$, we need to minimize the difference between the data and the simulated solution. You may recall that one way to do this is to minimize the sum of squared errors between data and solution:

$$\hat{\theta} = \underset{\theta \in \Omega}{\operatorname{argmin}} [Y_j - y(t_j; \theta)]^2$$

where Ω represents the parameter space. This is the ***ordinary least squares (OLS)*** approach and is appropriate for *constant* statistical error models.

Why is it not appropriate when the errors are proportional?

Let's look back at our cancer cell example and calculate the sum-of-squared error from day 1 and day 3:

1. On day 1, $Y_j = 105$ and $y(t_j; \theta) = 100$, then the sum of squared error would be 25
2. On day 3, $Y_j = 1050$ and $y(t_j; \theta) = 1000$, then the sum of squared error would be 2500

If you use a sum of squared error approach, your parameter estimation routine is going to consider day 3 more *important* than day 1! So how can we fix this?

We use a ***generalized least squares (GLS)*** approach in which we *weight* the sum of squared errors with our model solution. Thus, our new parameter estimate is:

$$\hat{\theta} = \underset{\theta \in \Omega}{\operatorname{argmin}} \, w_j \left(Y_j - y(t_j; \theta)\right)^2$$

In practice, this optimization process boils down into an iterative process:

1. Set the initial weights to $w_j = 1$
2. Estimate the parameters to minimize the weighted error
3. Set the weights to $w_j = y(t_j; \theta)^{-2\gamma}$ (Note: set w_j to 0 if $y(t_j; \theta)$ is small)
4. Repeat steps 2-3 until parameters converge

Checking Statistical Error Model

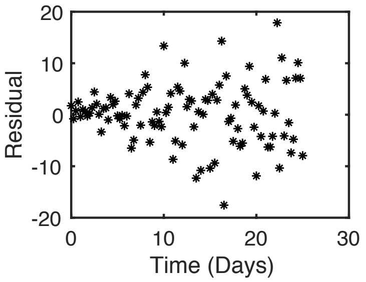
How do you know that the statistical error model you chose was correct? You can examine the error *residuals*. If these residuals appear i.i.d., then your error model is correct. Residuals are given by:

$$r_j = \frac{Y_j - y(t_j; \theta)}{y(t_j; \theta)^\gamma}$$

Note that in the case of ordinary least squares (OLS), this simplifies to $r_j = Y_j - y(t_j; \theta)$, since $\gamma = 0$.

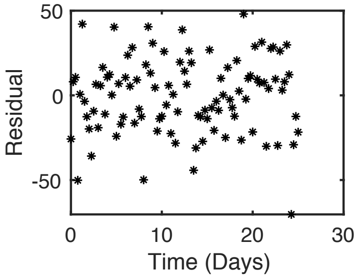
Example 1

Here is a residual plot for a system that used an OLS framework but had an underlying proportional error. As can be seen, the residuals do not appear i.i.d., rather they seem to have a dependence on time (the residuals are growing with time). This indicates that the incorrect statistical error model was chosen. In particular, when the errors 'fan' outward, γ should be increased.



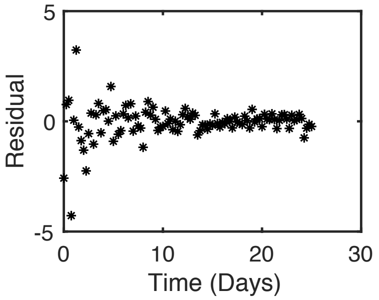
Example 2

In this example, a GLS framework was used for a system that does have proportional error. When examining the residual error plots, it does appear that the residuals are i.i.d., and do not depend on time or the model value.



Example 3

In this example, a GLS framework was used when the system had constant errors. Like example 1, the residual errors do not appear i.i.d., but in this case they seem to decrease as time increases. In this case the residuals 'fan' inward, and this indicates that the value of γ is too large.



So how to determine the correct value of γ?

1. Make an assumption based on your data collection procedure (do you suspect it is population-dependent? Set $\gamma = 1$)
2. Run the OLS/GLS framework to estimate your parameters
3. Examine the residuals
    a. If they look i.i.d., done!
    b. If not, change γ: If residuals 'fan out', increase γ, if residuals 'fan in', decrease γ.

A Practical Example

Consider the logistic growth equation $y' = ry\left(1 - \frac{y}{K}\right)$. In our practical example we will

1. Generate two synthetic datasets with $r = 0.5$ and $K = 200$. We will have 101 collection points between time 0 and time 30, and we will use 10% proportional error for dataset1 and a constant error of 20 for dataset 2.
2. Estimate r and Kusing the OLS framework for the constant error dataset
3. Estimate r and Kusing the GLS framework for the proportional error dataset
4. Plot the residuals/modified residuals to ensure our choice of γ was correct.

See parameter_tutorial.m for code

Your Task

To test your understanding of the code:

1. Estimate r and Kusing the GLS framework for the constant error dataset
2. Estimate r and Kusing the OLS framework for the proportional error dataset
3. Plot the residuals/modified residuals and see how the choice of γ was incorrect.