Predictive Performance Model Validation Methods with Uncertainty Quantification

Edward L. Boone

Virginia Commonwealth University

November 12, 2012

Members of Working Group

SAMSI Working Group on Uncertainty Quantification and Model Validation

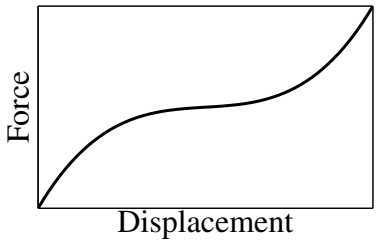
- Fabrizio Ruggeri [CNR IMATI].
- Serge Prudhomme [UT Austin].
- Yifang Li [NCSU].
- Jan Hannig [UNC].
- Sujit Ghosh [NCSU].
- **Edward L. Boone** [VCU].
- Maarten Arnst [ULg].

Outline

1. The Single Degree of Freedom Oscillator
2. Using the Bootstrap to Estimate the Model.
3. Using the Bootstrap for Simple Uncertainty Quantification.
4. The SEIR Model.
5. Limitations of the Bootstrap approach.
6. Conclusion.

Single degree-of-freedom oscillator

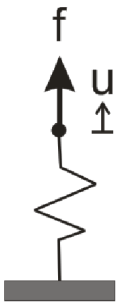
Data generating model



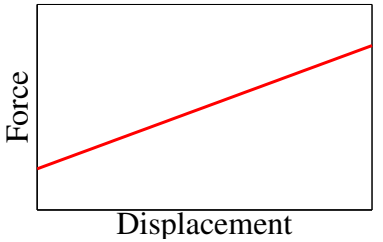
$$\tilde{m}\ddot{u} + \tilde{c}\dot{u} + \tilde{k}_{\mathsf{e}}u + \tilde{k}_{\mathsf{n}}u^3 = f$$

$$u = u_0 \quad \text{at } t = 0$$

$$\dot{u} = v_0 \quad \text{at } t = 0$$



Simulation model



$$m\ddot{u} + c\dot{u} + ku = f$$

$$u = u_0 \quad \text{at } t = 0$$

$$\dot{u} = v_0 \quad \text{at } t = 0$$

Generating the External Force

For our process we chose to explore the following pattern of external force.

$$\boxed{\begin{aligned}\tilde{f}(t) &= f(t) + f_{\text{noise}}(t) \\ f(t) &= f_1 \sin(\omega_1 t) + f_2 \sin(\omega_2 t)\end{aligned}}$$

where:

- Time interval: $T = 2{,}000$
- Amplitude: $f_1 = 1.0$
- Amplitude: $f_2 = 0.5$
- Noise: $|f_{\text{noise}}(t)| \ll f_1 + f_2 \; \forall t \in (0, T)$
- Frequencies: $\omega_1 = 2\pi n_1/T$, $\omega_2 = 2\pi n_2/T$

Sources of error

Suppose we viewed our process in the following manner:

$$u(t) = \mu(f, t) + \varepsilon_T$$

Here ε_T is the total error associated with the model which consists of the following:

$$\varepsilon_T = \varepsilon_C + \varepsilon_M + \varepsilon_S$$

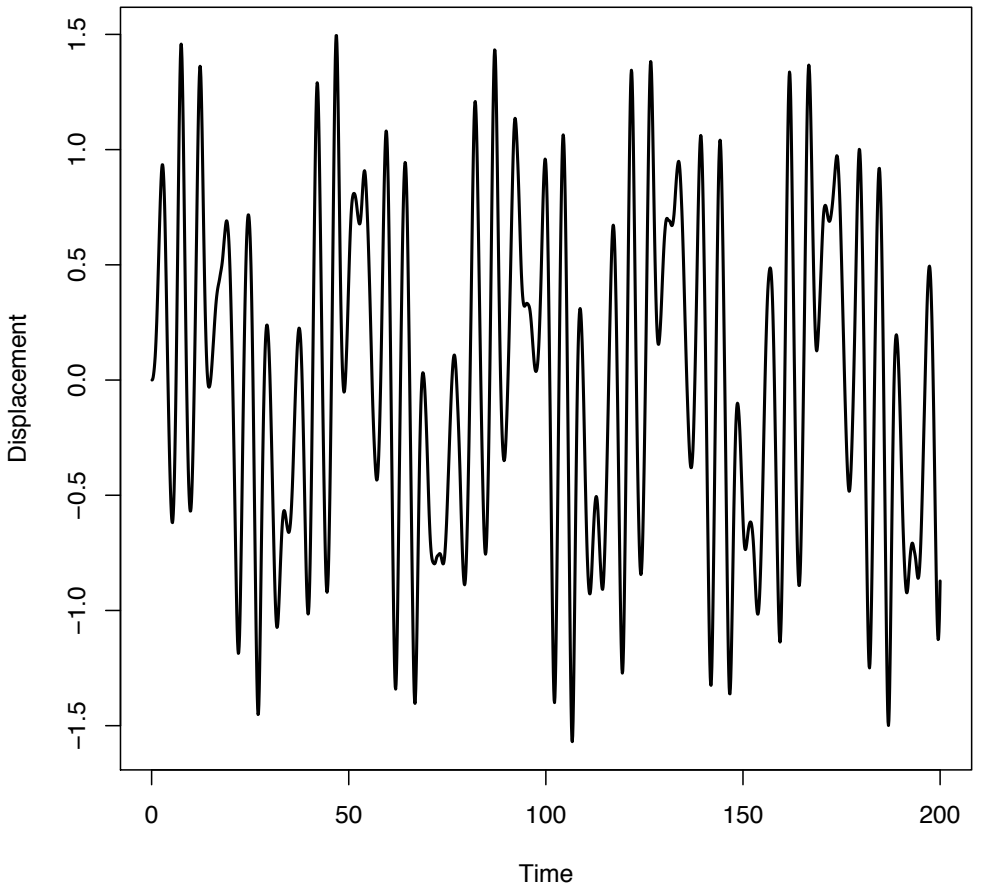
where

- ε_C is the error associated with coefficient estimation.
- ε_M is the error associated with model misspecification.
- ε_S is the error associated with the system.

We wish to understand each of these errors for model validation.

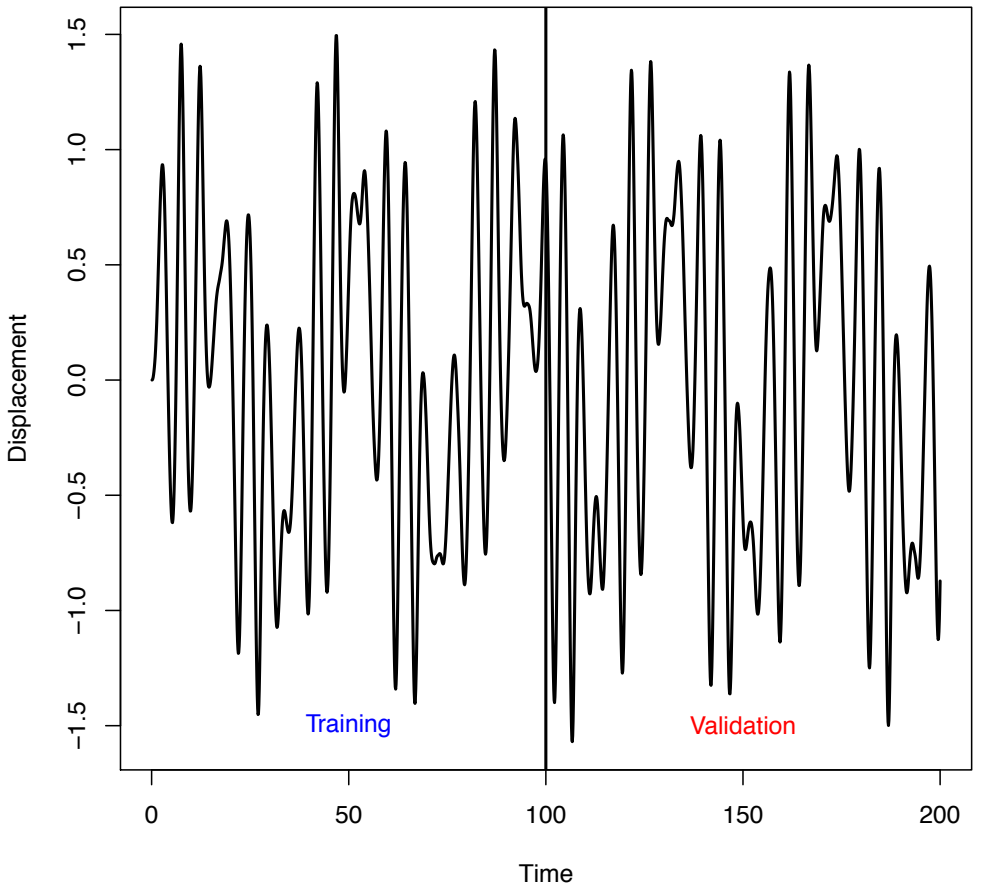
Partition the Data

We will use the predictive performance to understand the impacts of each of the types of errors. Hence we will need a training and validation set.



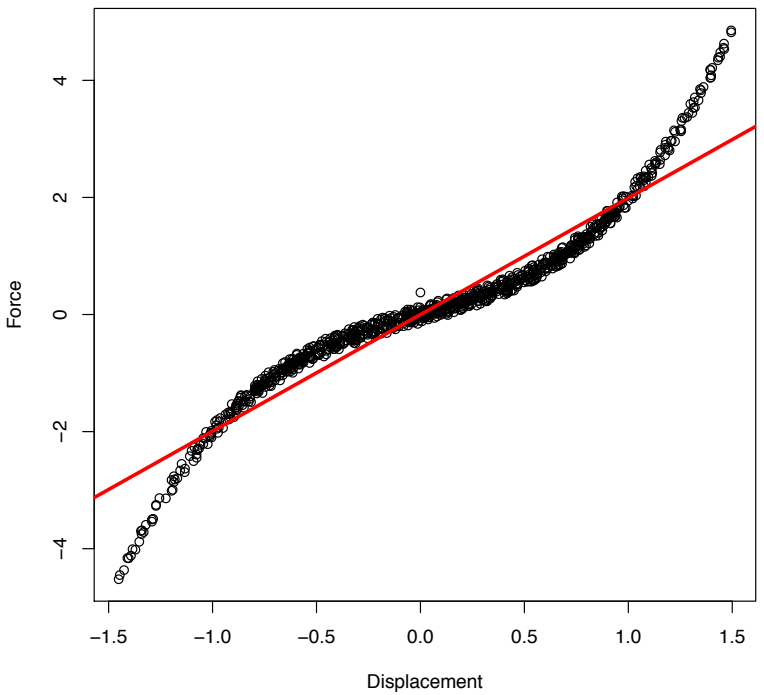
Partition the Data

We will use the predictive performance to understand the impacts of each of the types of errors. Hence we will need a training and validation set.



Simple Estimation

Use OLS to estimate c and k



True model

$$\tilde{m}\ddot{u} + \tilde{c}\dot{u} + \tilde{k}_{\mathsf{e}}u + \tilde{k}_{\mathsf{n}}u^3 = f$$

$$u = u_0 \quad \text{at } t = 0$$

$$\dot{u} = v_0 \quad \text{at } t = 0$$

Estimated Model

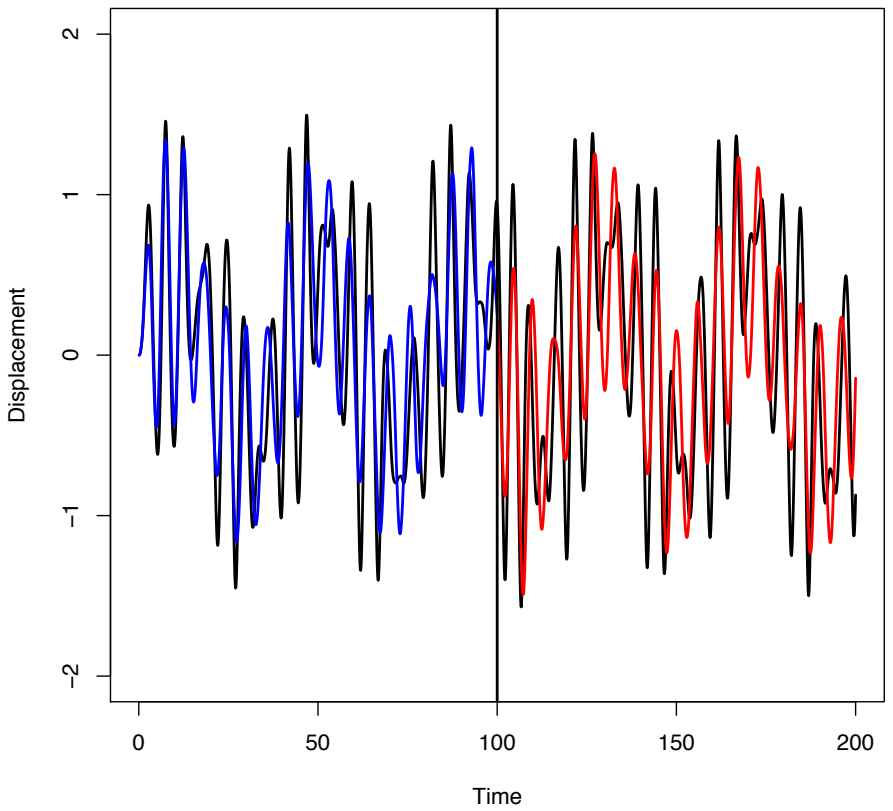
$$m\ddot{u} + \hat{c}\dot{u} + \hat{k}u + \varepsilon_M = f$$

$$u = u_0 \quad \text{at } t = 0$$

$$\dot{u} = v_0 \quad \text{at } t = 0$$

Simple Prediction

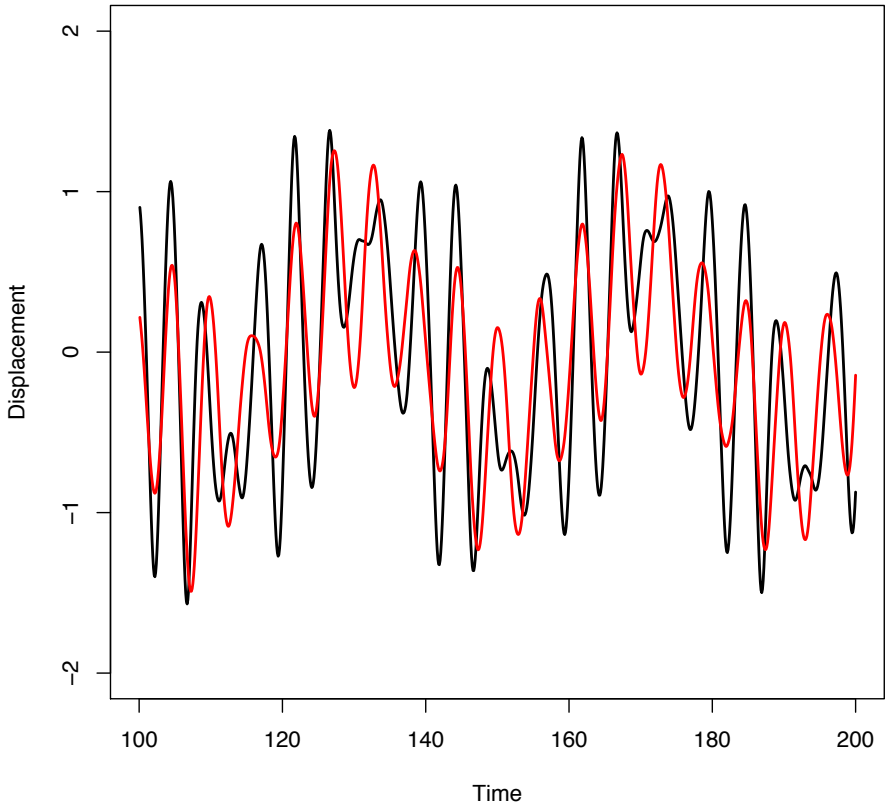
Using the training and validation set we can generate some simple point estimate predictions.



However this does not address uncertainty in predictions.

Simple Prediction

Using the training and validation set we can generate some simple point estimate predictions.



However this does not address uncertainty in predictions.

Simple Bootstrap Approach

We will want to estimate the total prediction error ε_T.

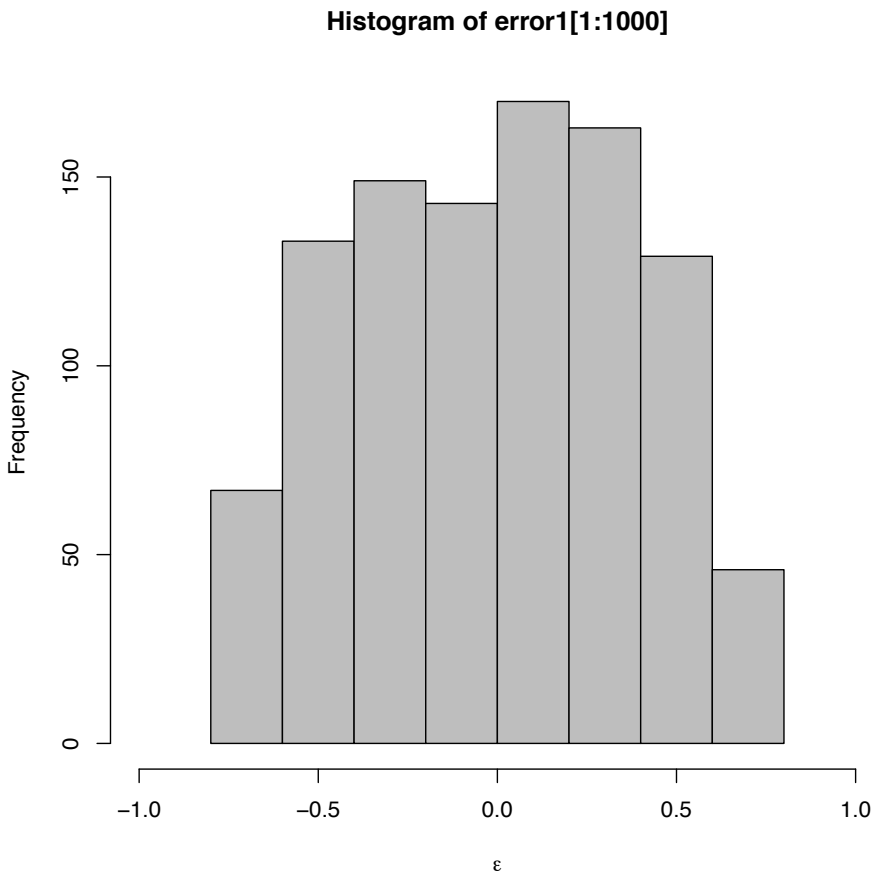
$$u(t) = \mu(f, t) + \varepsilon_T$$

We could use bootstrapped errors to sample ε based training data.

- Use the predictions as $\hat{\mu}(f, t)$.
- Sample $\hat{\varepsilon_T}$ from errors on displacement.
- Calculate $Var[\hat{\varepsilon}]$.

Simple Bootstrap Approach

Using the training set the errors, ε_T have the following distribution.



Simple Bootstrap Prediction Errors

We will want to estimate the total prediction error ε_T.

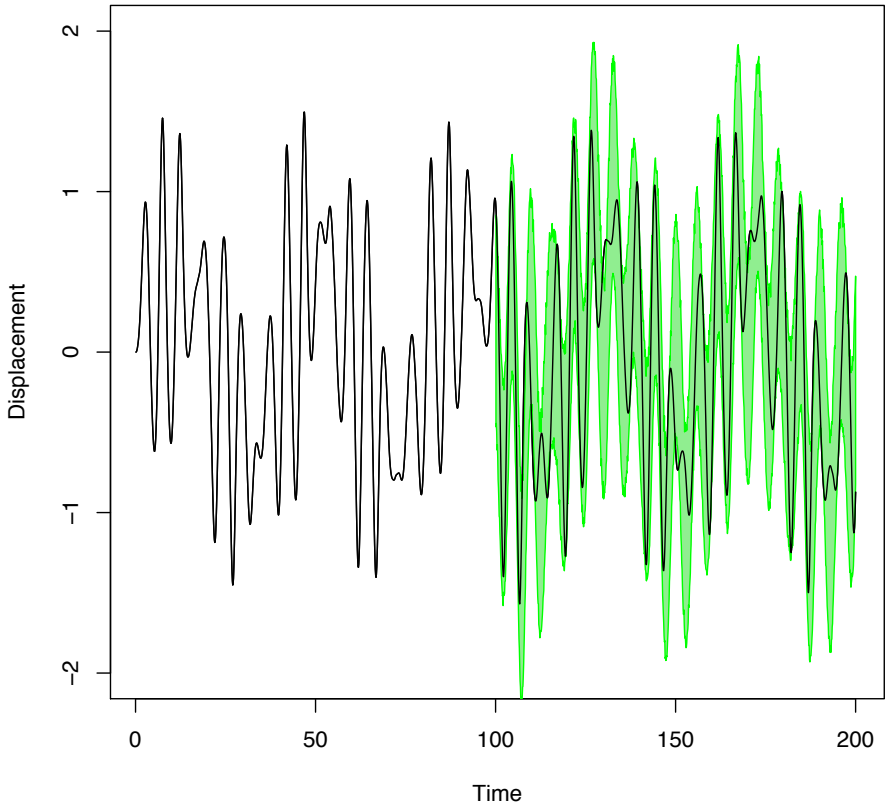
$$u(t) = \mu(f, t) + \varepsilon_T$$

We could use bootstrapped errors to sample ε based training data.

- Use the predictions as $\hat{\mu}(f, t)$.
- Sample $\hat{\varepsilon_T}$ from errors on displacement.
- Generate predictions $\hat{u}_{t,i} = \hat{\mu}(f, t) + \hat{\varepsilon_T}$ for the validation set where i represents the i^{th} bootstrap sample.
- Repeat 200 times.
- Take the 0.025 and 0.975 quantiles as prediction bounds.

Simple Bootstrap Prediction Errors

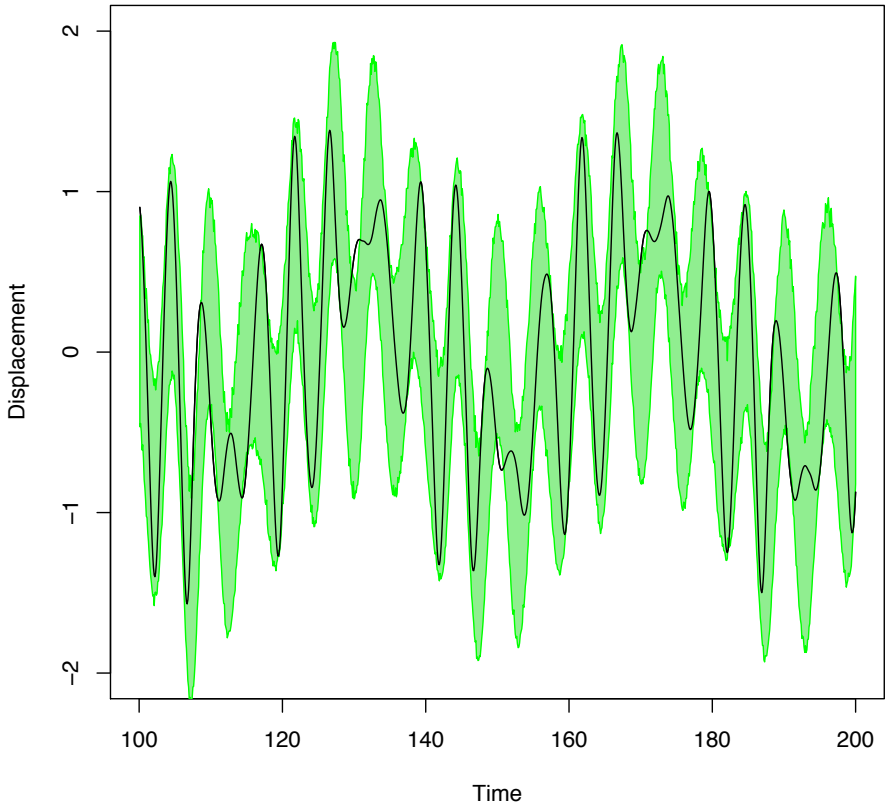
This simple approach results in the following:



Notice this envelope works well for the predictions.

Simple Bootstrap Prediction Errors

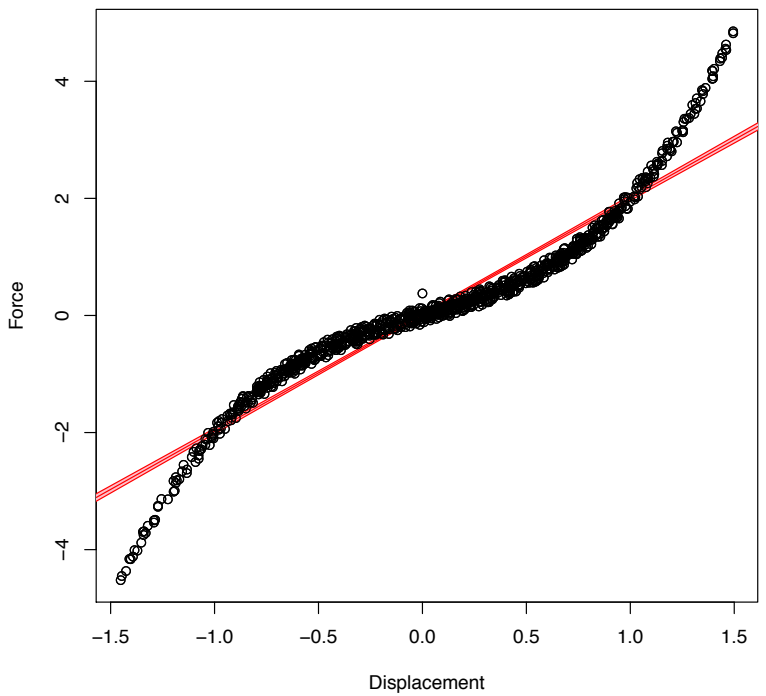
This simple approach results in the following:



Notice this envelope works well for the predictions.

Incorporating Coefficient Error

Use OLS to estimate c and k



True model

$$\tilde{m}\ddot{u} + \tilde{c}\dot{u} + \tilde{k}_{\mathsf{e}}u + \tilde{k}_{\mathsf{n}}u^3 = f$$

$$u = u_0 \quad \text{at } t = 0$$

$$\dot{u} = v_0 \quad \text{at } t = 0$$

Estimated Model

$$m\ddot{u} + \hat{c}\dot{u} + \hat{k}u + \varepsilon_M = f$$

$$u = u_0 \quad \text{at } t = 0$$

$$\dot{u} = v_0 \quad \text{at } t = 0$$

Note that $\hat{c}$ and $\hat{k}$ corresponds to coefficient error ε_C once propagated through the DE.

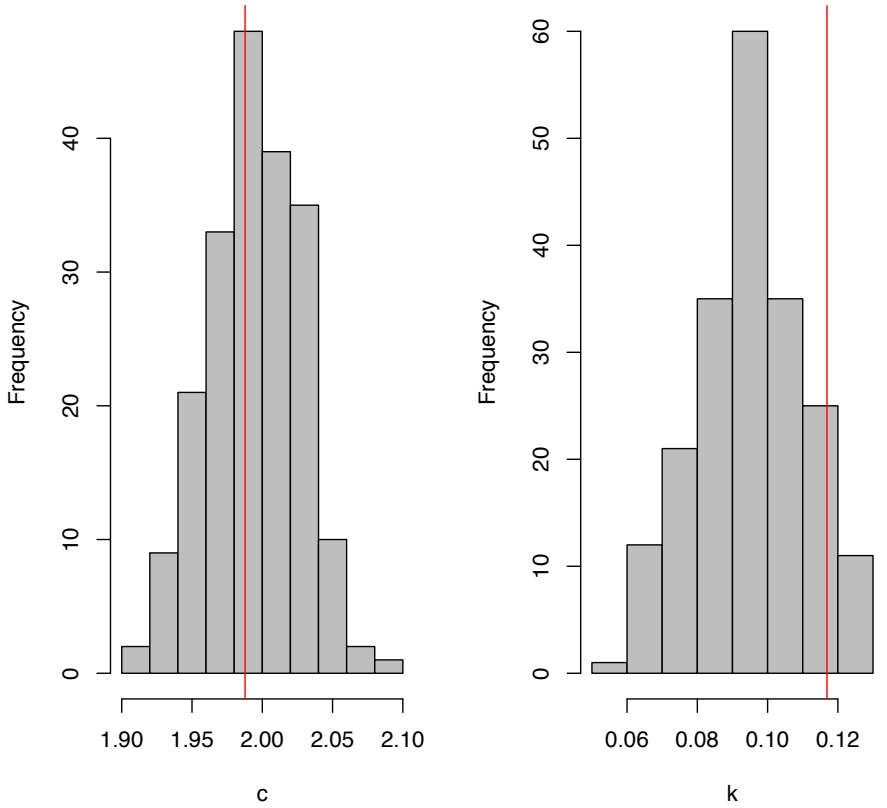
Bootstrap coefficients

We can use bootstrapping to estimate the error associate with the coefficients.

- Sample 1,000 observations from the training set.
- Generate $\hat{c}$ and $\hat{k}$.
- Solve the differential equation using these estimates
- Repeat 200 times.
- Calculate $Var[\hat{\varepsilon_C}]$

Bootstrap Coefficients

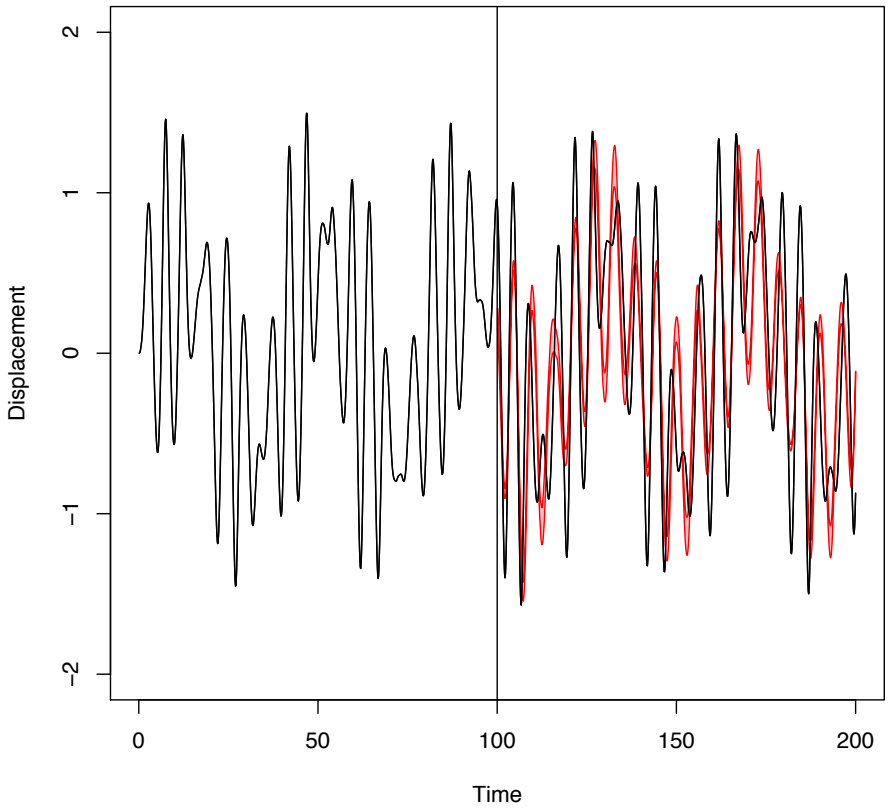
Here are histograms of the coefficients with red line at full training data value.



Based on 200 bootstrap samples of 1,000 each.

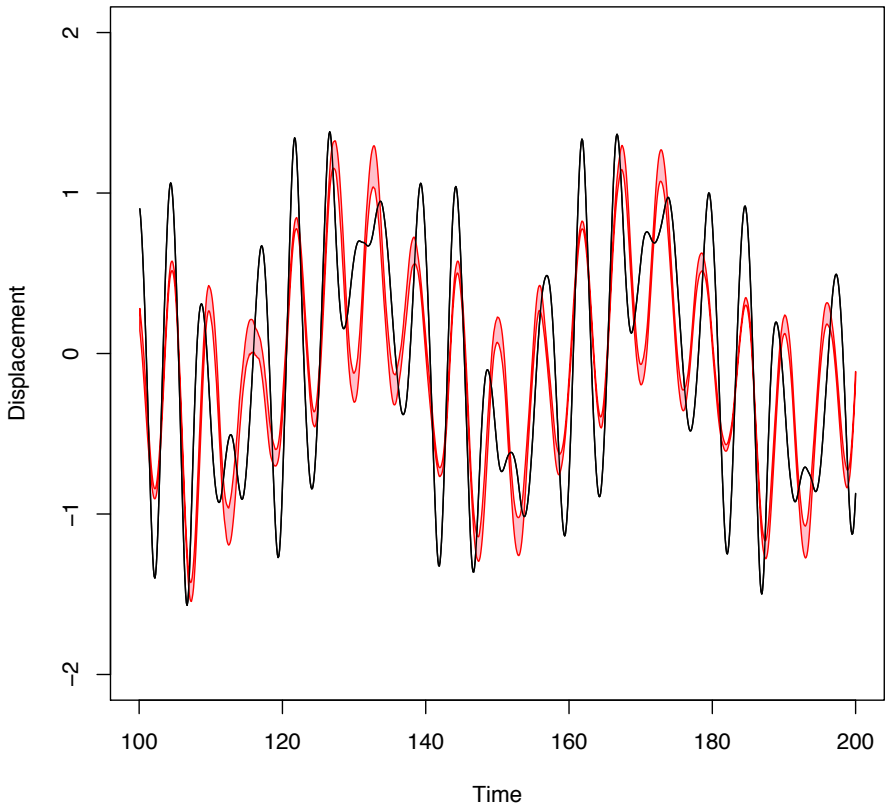
Bootstrap Coefficients

We can generate prediction intervals that are associated with the estimation error in the coefficients.



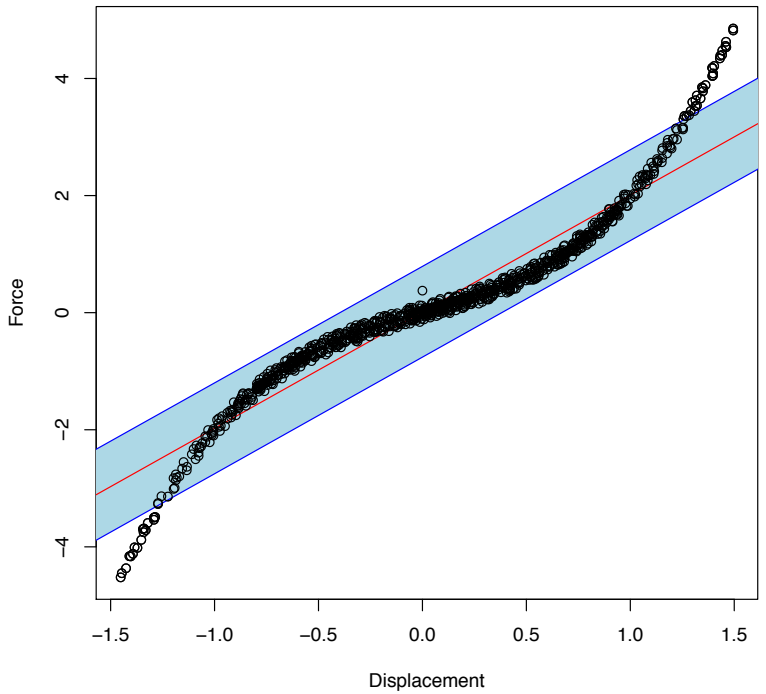
Bootstrap Coefficients

We can generate prediction intervals that are associated with the estimation error in the coefficients.



Incorporating Model Error

Use OLS to estimate c and k



True model

$$\tilde{m}\ddot{u} + \tilde{c}\dot{u} + \tilde{k}_{\mathsf{e}}u + \tilde{k}_{\mathsf{n}}u^3 = f$$

$$u = u_0 \quad \text{at } t = 0$$

$$\dot{u} = v_0 \quad \text{at } t = 0$$

Estimated Model

$$m\ddot{u} + \hat{c}\dot{u} + \hat{k}u + \varepsilon_M = f$$

$$u = u_0 \quad \text{at } t = 0$$

$$\dot{u} = v_0 \quad \text{at } t = 0$$

Note that $\tilde{k}_{\mathsf{n}}u^3$ corresponds to model error ε_M once propagated through the DE.

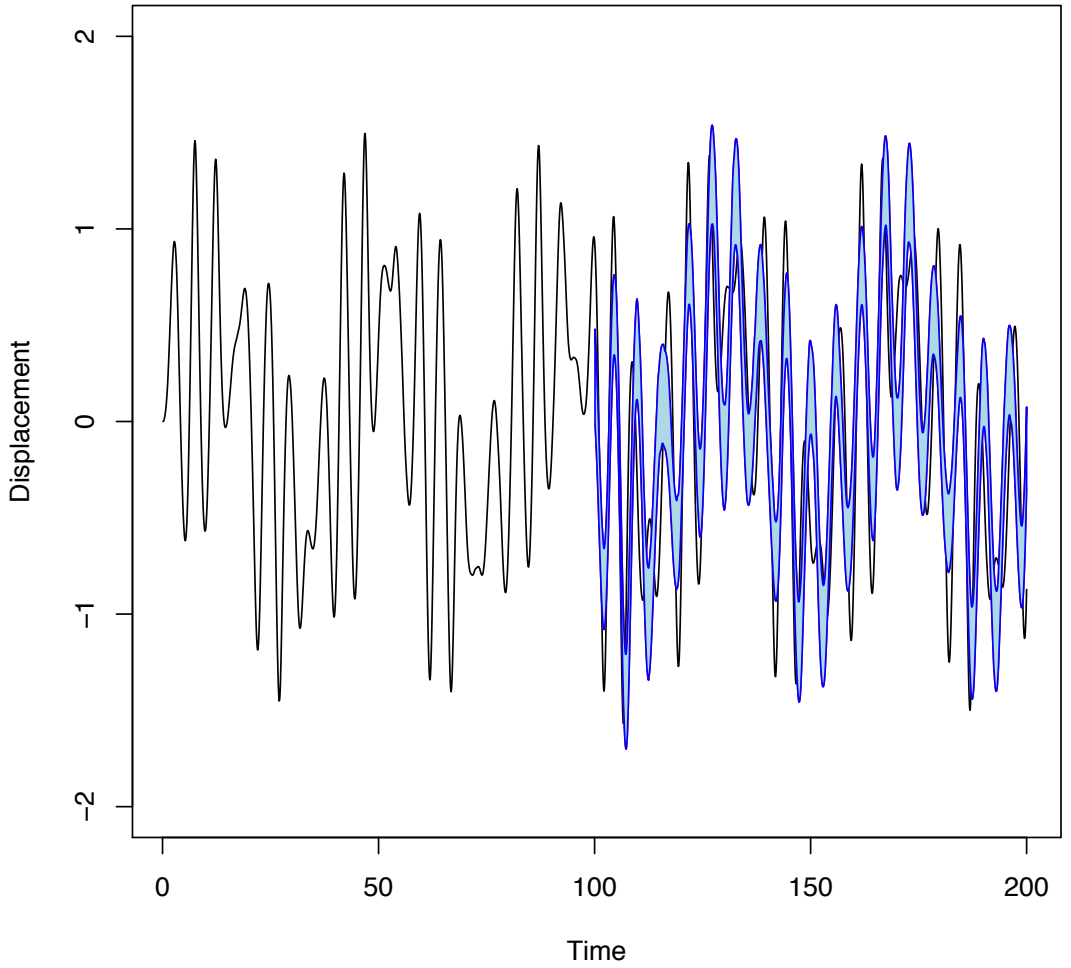
Incorporating Model Error

We can still use bootstrapping to incorporate coefficient error but to also add the error associated with using a line to estimate a nonlinear system.

- Sample 1,000 observations from the training set.
- Generate $\hat{c}$ and $\hat{k}$.
- Find the residuals.
- Sample 1,000 observations from the residuals.
- Add the sampled residuals in the differential equation.
- Solve the differential equation.
- Repeat 200 times.

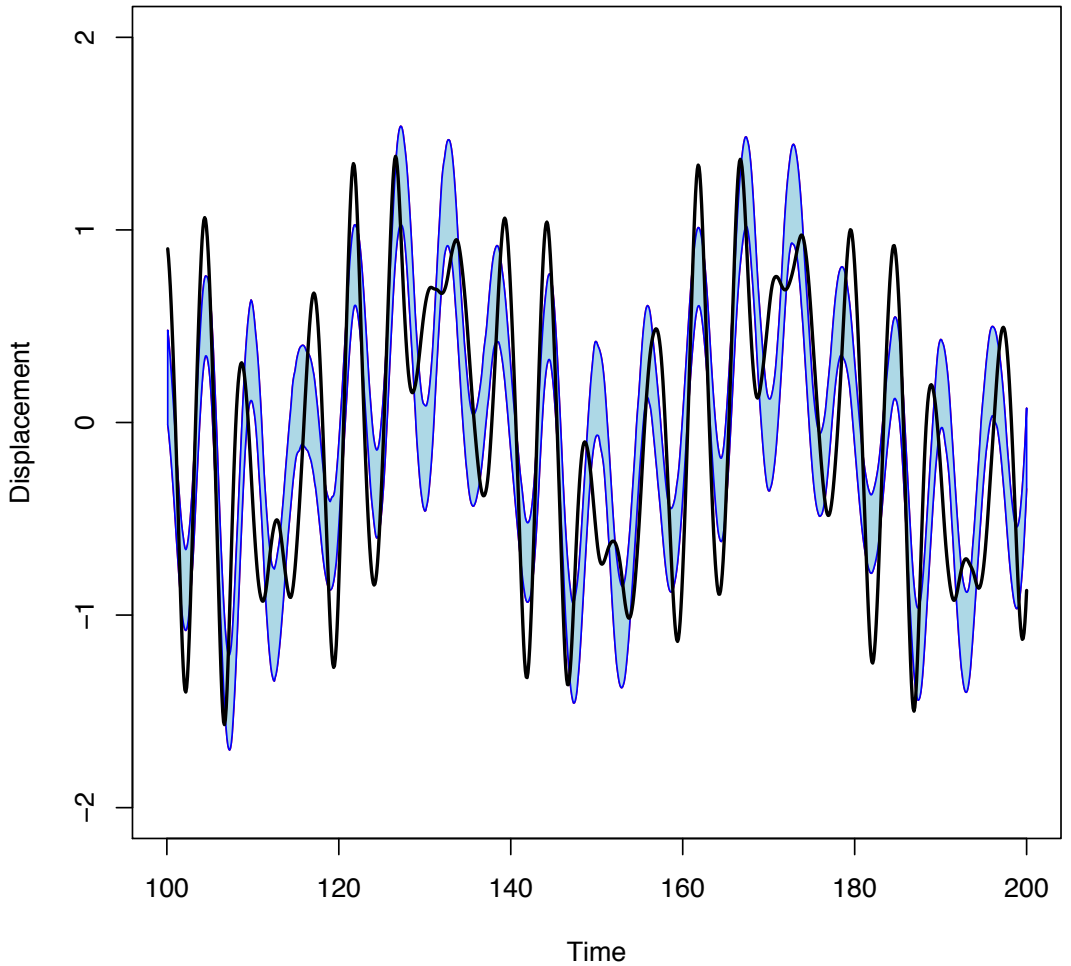
Incorporating Model Error

We can generate prediction intervals that are associated with the estimation error in the coefficients.



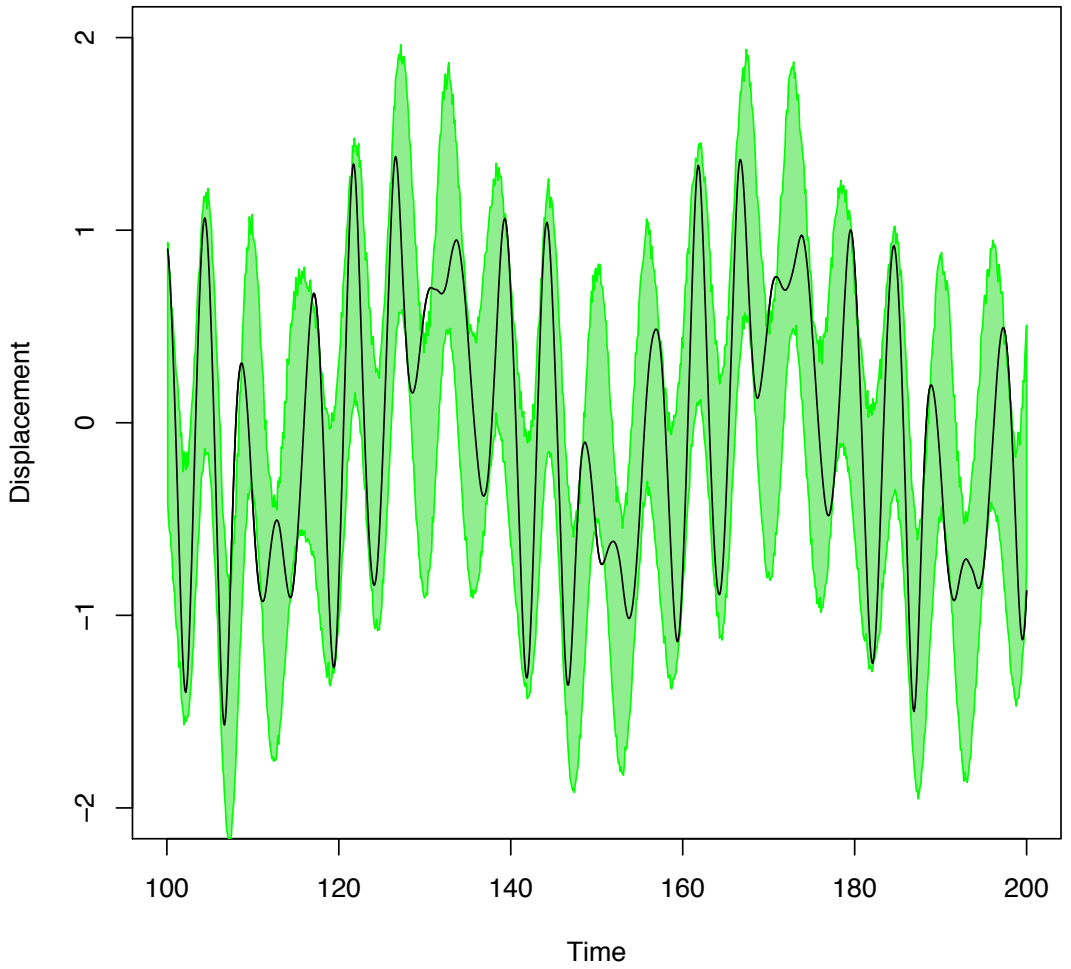
Incorporating Model Error

We can generate prediction intervals that are associated with the estimation error in the coefficients.



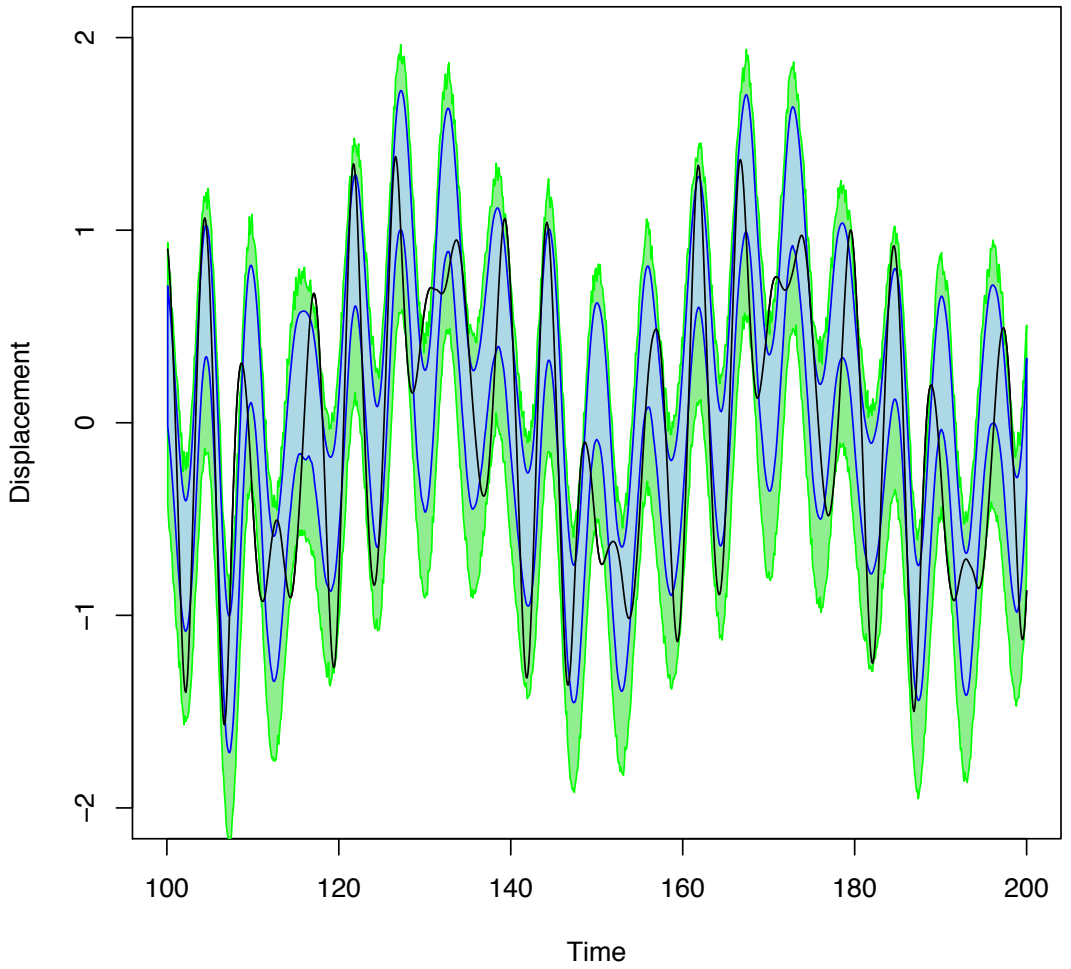
Sources of Errors

Overall Error



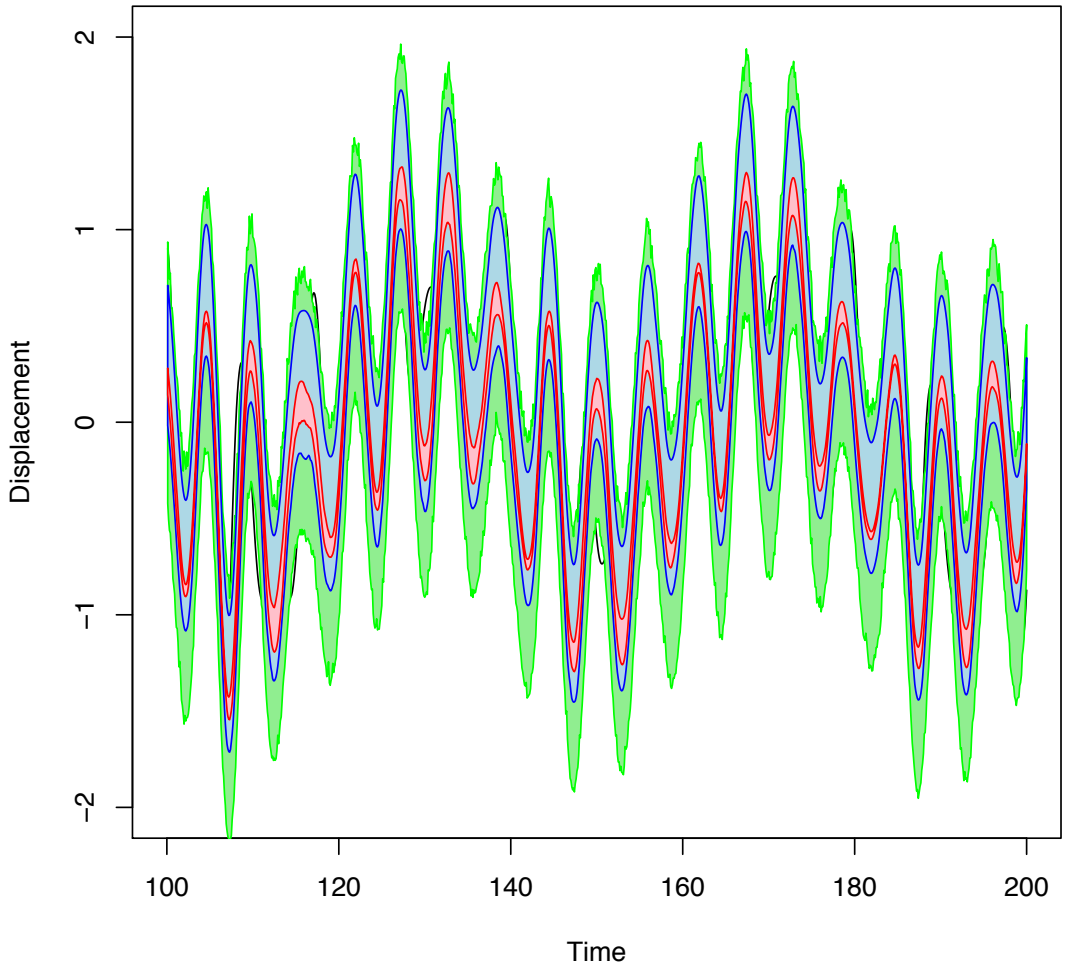
Sources of Errors

Overall and Model Error



Sources of Errors

Overall, Model and Coefficient Error



Sources of Variation

We can use these various approaches to determine what are the sources of variation in our system.

- Overall variation, SS_T.
- Variation associated with estimation of coefficients, SS_C.
- Variation associated with modeling error/mispecification, SS_M.
- Variation associated with the system error, SS_E
- These should be related in the following manner (similar to ANOVA):

$$SS_T = SS_M + SS_C + SS_E$$

- We can use our above approaches to estimate these quantities.

Estimating the Sources of Variation

We can estimate these using each of the methods mentioned above:

$$SS_T = \frac{1}{n_b} \sum_{i=1}^{n_b} \sum_{t=1}^{n_t} (u_t - \hat{u}_{t,i})^2$$

$$SS_C = \frac{1}{n_b} \sum_{i=1}^{n_b} \sum_{t=1}^{n_t} \left(\hat{u}_{t,i} - \hat{u}_{t,i}^{C}\right)^2$$

$$SS_M + SS_C = \frac{1}{n_b} \sum_{i=1}^{n_b} \sum_{t=1}^{n_t} \left(\hat{u}_{t,i} - \hat{u}_{t,i}^{M+C}\right)^2$$

Hence we can find SS_M and SS_E by subtraction.

$$SS_E = SS_T - SS_M - SS_C$$

Single-degree of freedom oscillator example

For our example problem we can estimate the overall error as:

$$\begin{aligned} SS_T &= 149.76 \\ SS_C &= 1.24 \\ SS_M + SS_C &= 28.21 \\ SS_E &= SS_T - SS_M - SS_C \\ &= 121.55 \end{aligned}$$

This shows that approximately 81% of the variation in the predictions is due to system error and approximately 19% is due to model misspecification and estimation.

Using a Bayesian Framework

True model

$$\tilde{m}\ddot{u} + \tilde{c}\dot{u} + \tilde{k}_{\mathsf{e}}u + \tilde{k}_{\mathsf{n}}u^3 = f$$
$$u = u_0 \quad \text{at } t = 0$$
$$\dot{u} = v_0 \quad \text{at } t = 0$$

Estimated Model

$$m\ddot{u} + \hat{c}\dot{u} + \hat{k}u + \varepsilon_M = f$$
$$u = u_0 \quad \text{at } t = 0$$
$$\dot{u} = v_0 \quad \text{at } t = 0$$

Note that $\tilde{k}_{\mathsf{n}}u^3$ corresponds to model error ε_M once propagated through the DE.

Using a Bayesian Framework

Estimated Model

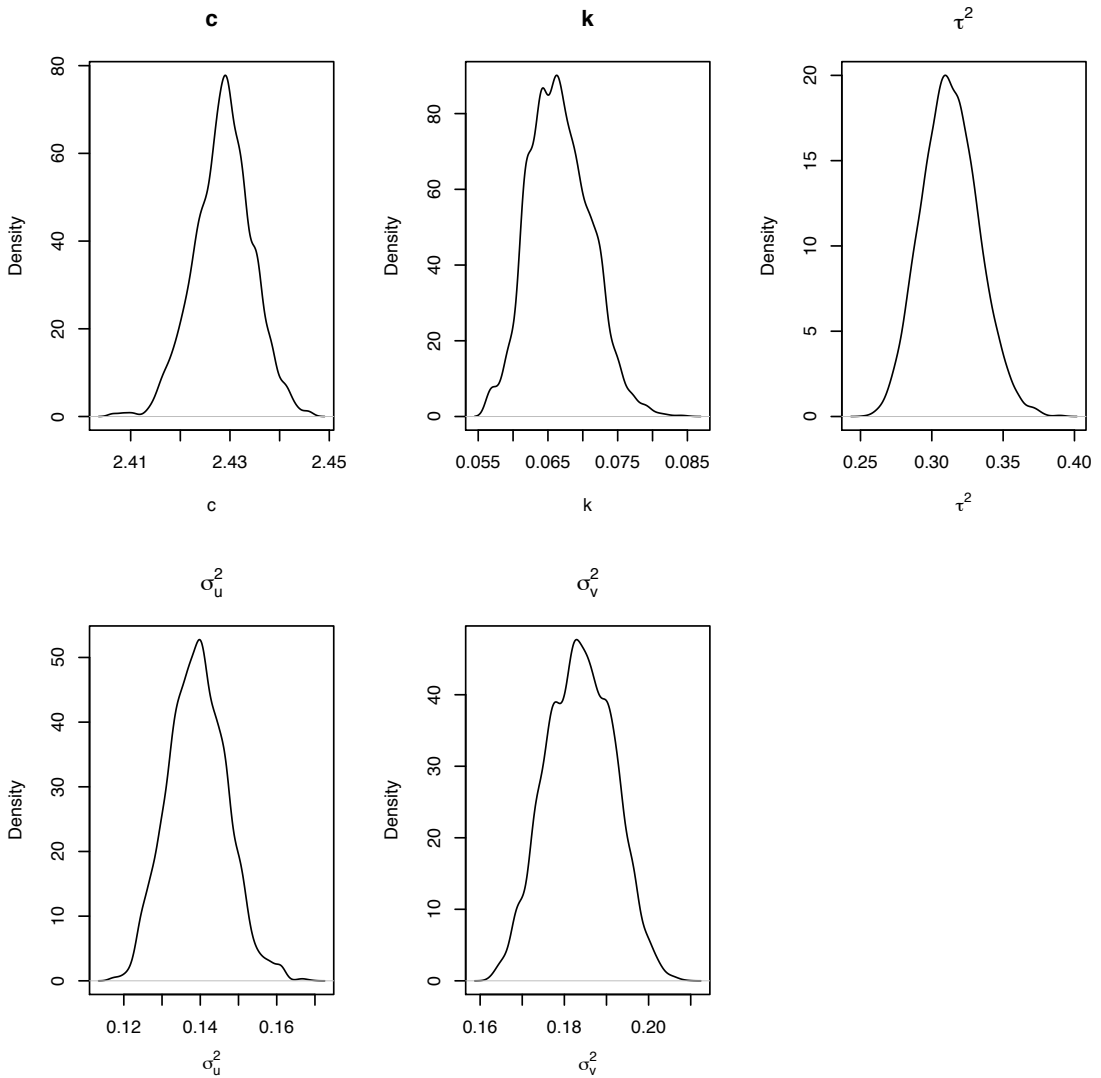
$$m\ddot{u} + \hat{c}\dot{u} + \hat{k}u + \varepsilon_M = f$$
$$u = u_0 \quad \text{at } t = 0$$
$$\dot{u} = v_0 \quad \text{at } t = 0$$

We will put the following prior distributions on the parameters:

$$\hat{c} \sim N(0, 10)$$
$$\hat{k} \sim N(0, 10)$$
$$\varepsilon_M \sim N(0, \tau), \quad \tau^2 \; Inv - \chi^2(1, 1)$$
$$\varepsilon_u \sim N(0, \sigma_u), \quad \sigma_u^2 \sim \chi^2(1)$$
$$\varepsilon_v \sim N(0, \sigma_v), \quad \sigma_v^2 \sim \chi^2(1)$$

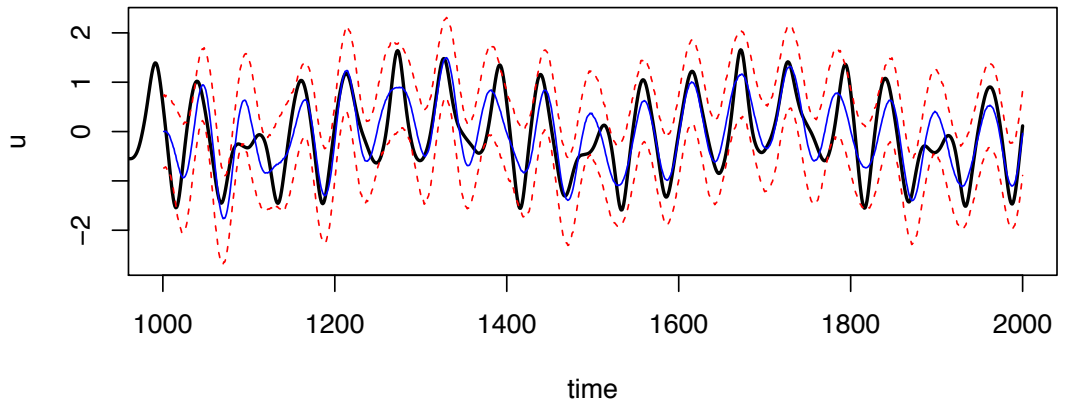
Using a Bayesian Framework

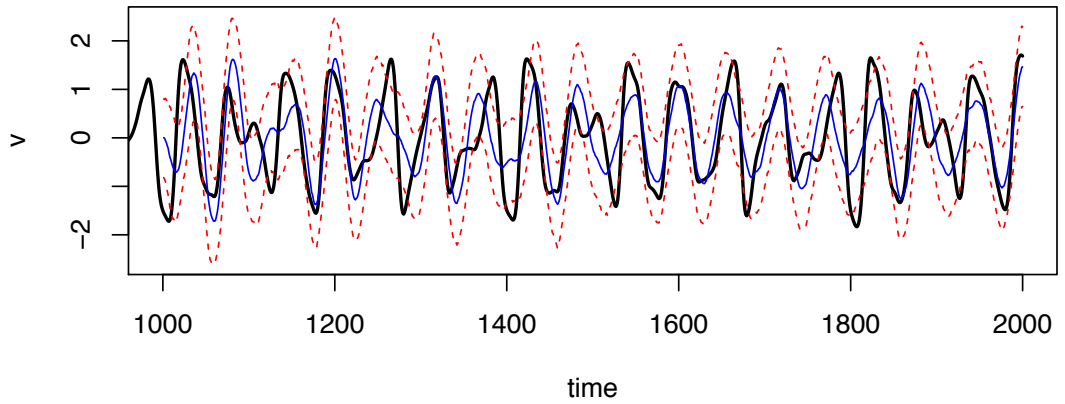
Parameter estimates:



Using a Bayesian Framework

Results





Epidemiological Example

- The bootstrapping approach seems to work well for the simple Single-degree of freedom oscillator example.
- What about if we cannot observe the relationship to the forcing directly?
- Consider the SEIR model from Epidemiology.
- Does this approach work for modelling influenza?

SEIR Model

The Susceptible, Exposed, Infected, Recovered Model.

$$\begin{aligned}
\frac{\partial S}{\partial t} &= \mu N - \phi S - \beta \frac{SI}{N} \\
\frac{\partial E}{\partial t} &= \beta \frac{SI}{N} - (\phi + \sigma) E \\
\frac{\partial I}{\partial t} &= \sigma E - (\phi + \gamma) I \\
\frac{\partial R}{\partial t} &= \gamma I - \phi R
\end{aligned}$$

This is subject to $N = S + I + E + R$. Here μ is the birth rate, ϕ is the mortality rate, β is infection rate, σ length of exposed state, γ is the length of infected state.

SEIR Model

For influenza we need to change the model to allow for reinfection. Hence the model becomes:

$$\begin{aligned}
\frac{\partial S}{\partial t} &= \mu N - \phi S - \beta \frac{SI}{N} + \psi R \\
\frac{\partial E}{\partial t} &= \beta \frac{SI}{N} - (\phi + \sigma) E \\
\frac{\partial I}{\partial t} &= \sigma E - (\phi + \gamma) I \\
\frac{\partial R}{\partial t} &= \gamma I - (\phi + \psi) R
\end{aligned}$$

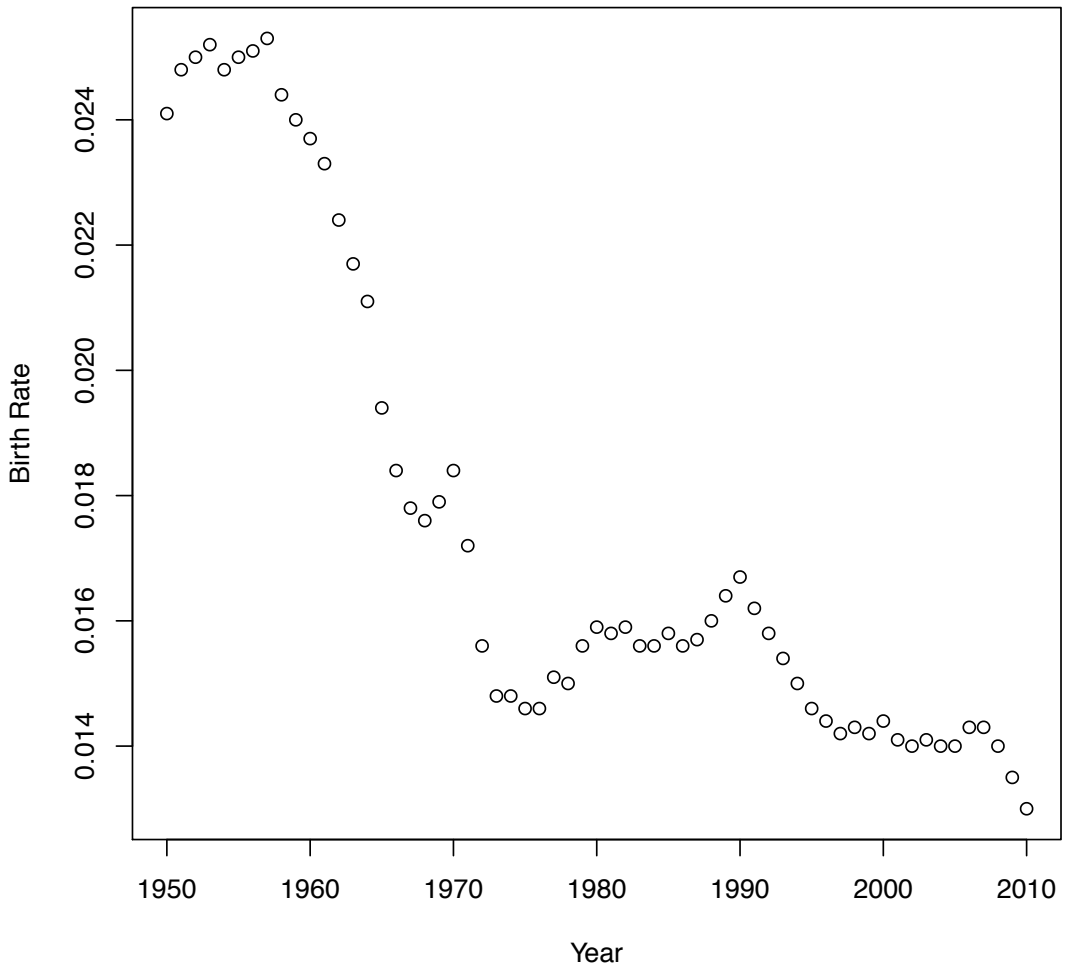
Here ψ represents the length of time one is immune.

SEIR Model

- One problem is that some items are observable and some are not observable.
- We often want to know the unobservables.
- We want to incorporate uncertainty into the unobservables.
- Hence there is a lot of uncertainty associated with the parameters in the model.

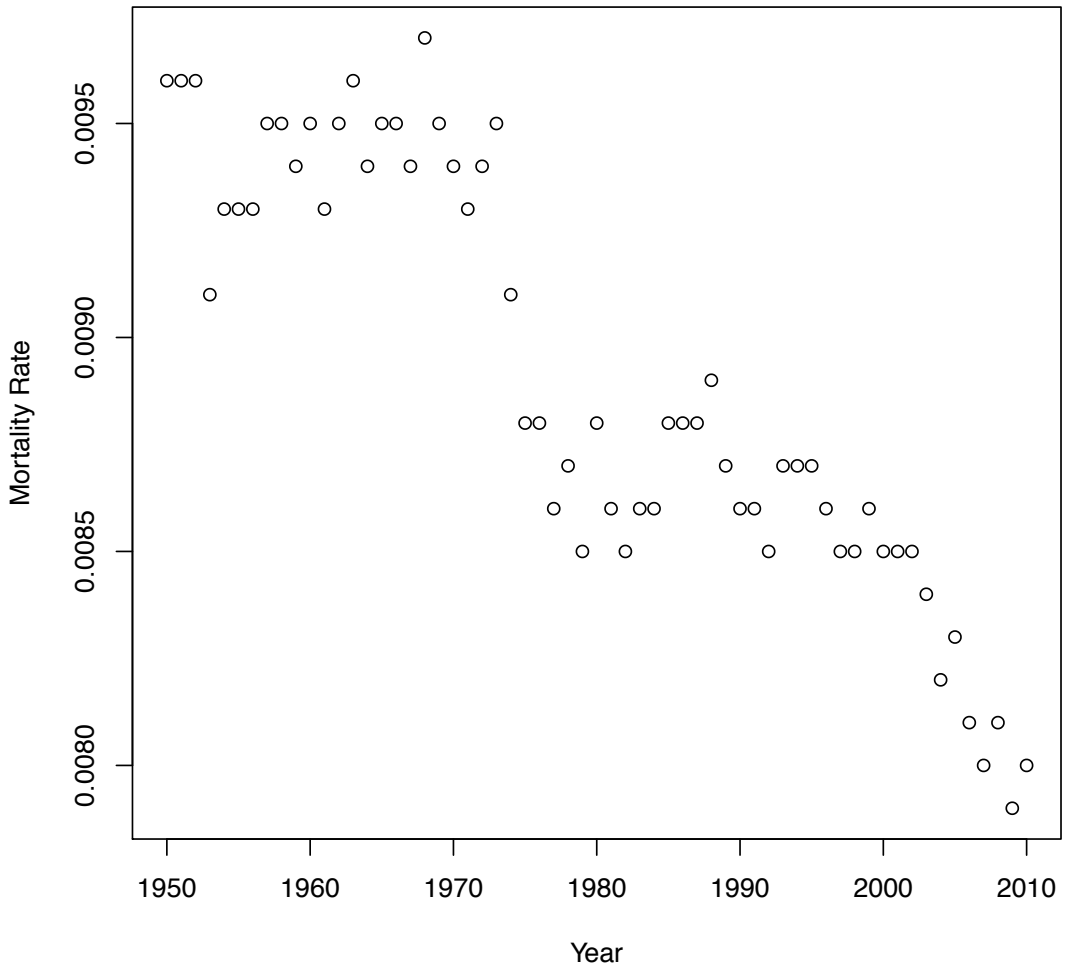
SEIR Model

Annual Birth Rates



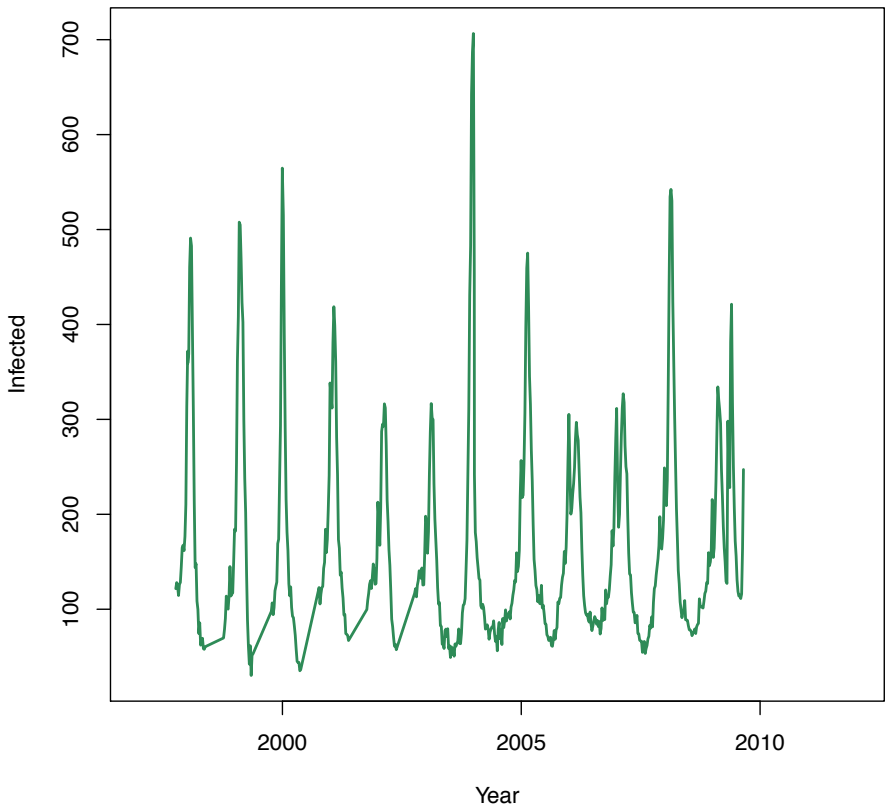
SEIR Model

Annual Mortality Rates



SEIR Model

Weekly Number of Infections



This is infections per 10,000 people from the CDC.

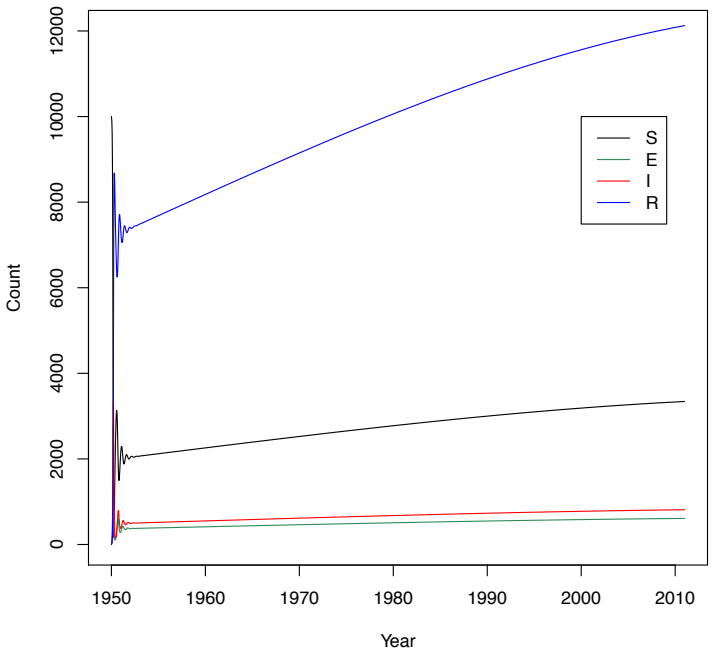
SEIR Model

Notice the following:

- Birth Rates are not constant.
- Mortality Rates are not constant.
- Infections are not in a steady state (i.e. constant)
- Most models assume Birth Rates and Mortality Rates are constant or at least equal. This is not true in this case.
- We cannot observe transmission rates. We do know that transmission rates go up in winter.
- Exposure to infection times are difficult to observe.
- We can observe how long the infection state lasts.
- Difficult to know how long immunity lasts.

SEIR Model

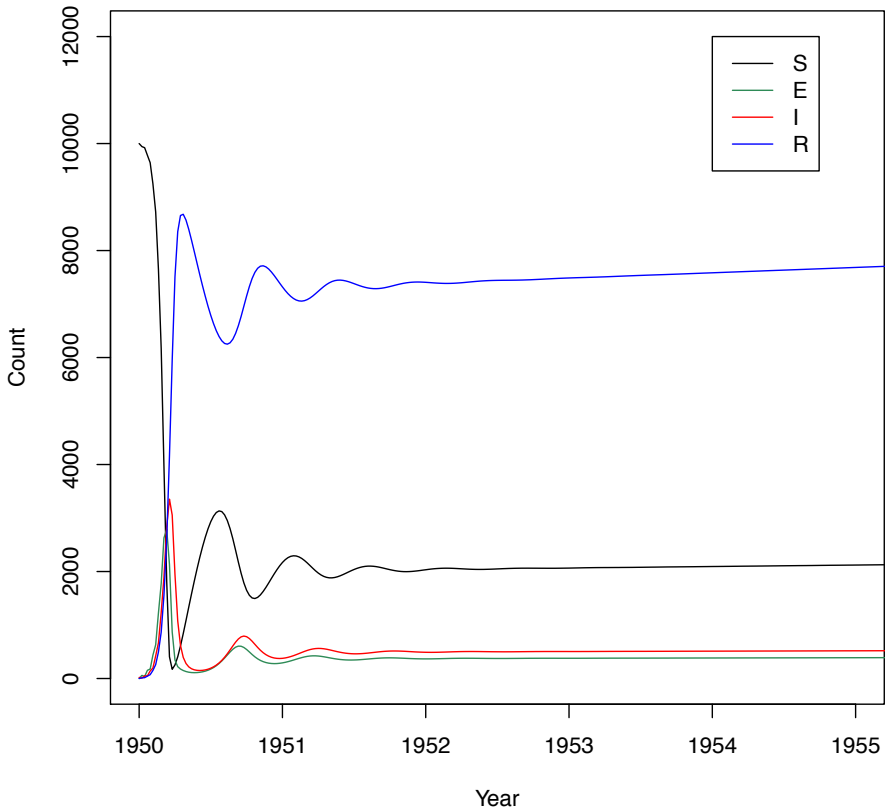
Using regression to estimate the Birth and Mortality Rates, $\beta = 25$, $\sigma_1 = 365/9$, $\gamma_1 = 365/12$, $\psi = 365/180$, and $N_{1950} = 10,000$, then the SEIR model produces:



This reaches steady state?

SEIR Model

A closer look produces:



Only oscillations early in the process. Hence we will need a forcing mechanism.

SEIR Model

A possible forcing mechanism could be:

$$\beta_t = (a - m)\cos\left(2\pi(t + kdt)\right) + m + \epsilon$$

- However this forcing process is not directly observable.
- We can look at the process to see where peaks may be.
- The ϵ allows for incorporation of uncertainty.
- However bootstrapping is out since we cannot directly observe this.

SEIR Model

For influenza we need to change the model to allow for reinfection. Hence the model becomes:

$$\begin{aligned}
\frac{\partial S}{\partial t} &= \mu_t N - \phi_t S - \beta_t \frac{SI}{N} + \psi R \\
\frac{\partial E}{\partial t} &= \beta_t \frac{SI}{N} - (\phi_t + \sigma) E \\
\frac{\partial I}{\partial t} &= \sigma E - (\phi_t + \gamma) I \\
\frac{\partial R}{\partial t} &= \gamma I - (\phi + \psi) R
\end{aligned}$$

Here ψ represents the length of time one is immune.

SEIR Model

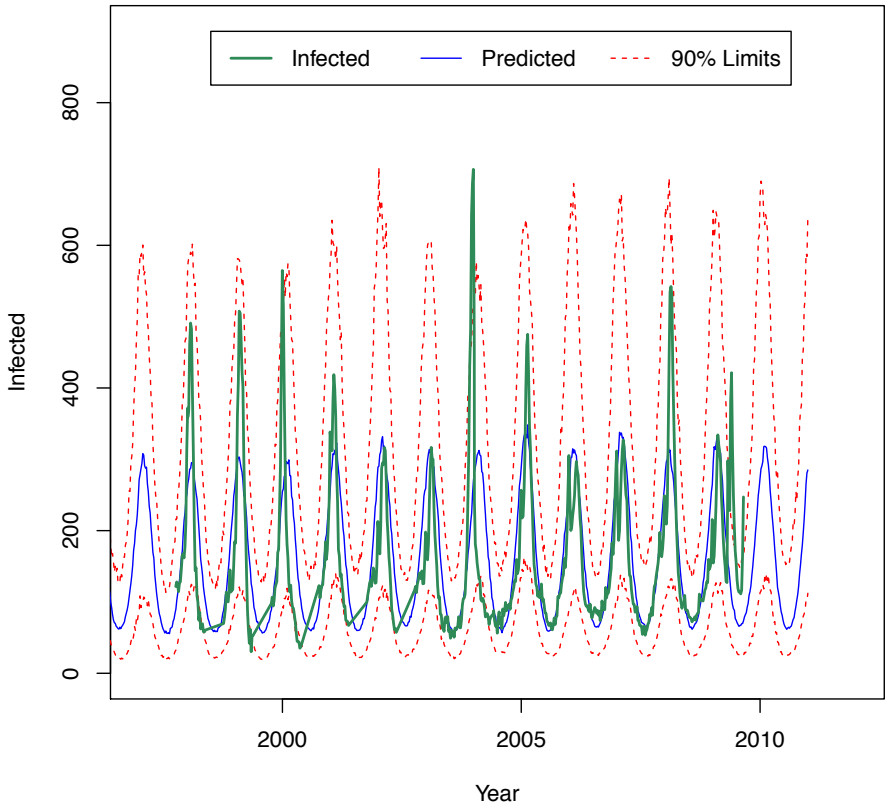
Consider the following set up:

$$
\begin{aligned}
\mu_t &= \rho_0 + \rho_1 t + \epsilon_\mu \\
\phi_t &= \alpha_0 + \alpha_1 t + \epsilon_\alpha \\
\beta_t &= 9\left(cos(2\pi(t + 8dt)) + 4\right) + \epsilon_\beta \\
\gamma &= 365/D_1, \\
\sigma &= 365/D_2, \\
\psi &= 365/D_3,
\end{aligned}
$$

- Here ρ_0, ρ_1, α_0 and α_1 are estimated via bootstrap, ϵ_μ and ϵ_α are bootstrap errors and $\epsilon_\beta \sim N(0, 10)$.
- This incorporates the uncertainty associated with μ, ϕ and β.
- $D_1 \sim N(12, 1/3)$, $D_2 \sim N(9, 1/3)$ and $D_3 \sim N(180, 10)$.

SEIR Model

This produces.



This appears to work well.

SEIR Model

- In the above approach we were able to incorporate uncertainty about the birth and mortality rates as well as the impact of the forcing.
- The above approach does not incorporate any uncertainty associated with γ, σ, nor ψ.
- It also does not incorporate any uncertainty associated with the parameters in the seasonal forcing.
- We hope to instead use a fully Bayesian approach for this problem as bootstrapping approach is not feasible.
- This is what I am working on now.

Conclusions

- The bootstrap approach can be used for uncertainty quantification.
- This approach can be extended to other attributes of the model such as velocity and acceleration in a multivariate fashion.
- It requires that all forcings be observable.
- Another approach that should be explored is building a Bayesian approach to handle the parameter estimation under uncertainty. It also allows for the incorporation of prior information.
- More study needs to be done in order to determine a good metric that incorporates fit, prediction error and uncertainty calibration.
- The Working Group offered a great opportunity for collaboration between experts in statistics and engineering

Conclusions

Thank you!

Questions ???