

Compartmental Models in SAS: Application to Model Epidemics

Ka Chun Chong and Benny Chung-Ying Zee

Clinical for Clinical Research and Biostatistics, JC School of Public Health and Primary Care
Clinical Trials and Biostatistics Laboratory, Shenzhen Research Institute
The Chinese University of Hong Kong, China

ABSTRACT

Compartment modeling is useful to quantify the spread of elements in a dynamic system. Apart from the Pharmacokinetics-Pharmacodynamics analysis, epidemic modeling is another broad application of compartmental models. In this paper, we demonstrate how to use PROC MODEL and arrays in DATA steps to generate and fit the epidemic models such as the Kermack and McKendrick model and SEIR model. Practical application is demonstrated for the 2009 pandemic A/H1N1.

INTRODUCTION

Compartmental models offer a framework of how elements (material, information, energy, etc.) transport between different compartments in a dynamic system. They have been widely applied to environment, economical, system biology, engineering, and medical studies. In a Pharmacokinetics-Pharmacodynamics study, the elements refer to the chemicals such as drug concentration and hormones. On the other hand, the elements usually refer to subgroups of population in epidemiological studies. The compartment models are able to describe the disease system under the constraints of interventions from the biological, political, and epidemiological data.

Epidemic compartmental models have been used to study transmission mechanism of infectious diseases for a long time. Hamer (1906) has developed one of the earliest epidemic models in 1906. Ross (1916) adopted the method in a time series model and called it mass action principle in 1916. Until 1927, Kermack and McKendrick (1927) developed a famous Susceptible-Infectious-Recovered (SIR) model and it still works as a principal for various extensions of epidemic models for nowadays.

Epidemic models are useful to determine the transmission dynamics of an infectious disease and effectiveness of interventions. Sometimes clinical trial design is impractical for assessing the effectiveness of some interventions, such as face masks and isolation, because of ethical considerations relating to epidemics in general. Epidemic models thus have been adopted to evaluate various interventions; isolation, quarantine, antiviral drugs, school closures, vaccinations, and face masks, among others.

A large amount of frameworks of epidemic models are from deterministic and stochastic structures. The deterministic structures are relatively easy to build and thus more common. On the other hand, stochastic models are able to capture uncertainties along with the time series of disease propagation, especially when the number of infected individuals is small or the chance event is important in the transmission dynamics.

Several investigators have introduced using SAS to draw epidemic compartmental models (Gallop 2009; Grafe 2014). In this paper, we demonstrate how to use SAS PROC MODEL and arrays in DATA steps to perform compartmental modeling analysis in epidemics.

KERMACK AND MCKENDRICK SIR MODEL

Kermack and McKendrick (1927) [60] SIR model is one of the earliest mathematical models in the history of epidemic model. The model categorizes population into Susceptible, Infectious, and Recovered. Susceptible individuals in S-stage have chance to be infected and progress to Infection I-stage until recovery to R-stage. The flow is shown in figure 1.

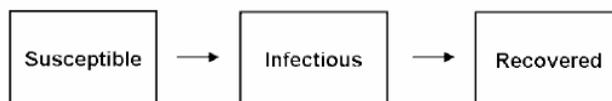


Figure 1. Flow of SIR model

For a mathematical convention, we denote S, I and R as the subpopulations in each compartment for time t. The total population size N, is equal to S+I+R for any time and N=S for time zero. The deterministic SIR model can be written as the following system of nonlinear differential equations:

$$\begin{aligned}\frac{dS}{dt} &= -\beta SI \\ \frac{dI}{dt} &= \beta SI - \gamma I \\ \frac{dR}{dt} &= \gamma I\end{aligned}$$

where β is the transmission rate and γ is the recovery rate. As the infectious period is assumed exponential distributed, we denote $1/\gamma$ as the average infectious period. By linearizing the system [30], the basic reproductive numbers (R_0), defined as the average number of secondary infections produced by a typical infected individual in a wholly susceptible population, is equal to $\beta N/\gamma$. R_0 is a key parameter of disease transmission and is used to determining required control measures. In order to prevent epidemics, reproduction numbers should be maintained smaller than 1. The more control measures and interventions should be introduced if the quantity R_0 is large.

PROC MODEL TO SIMULATE AN EPIDEMIC

SAS PROC MODEL is useful to simulate an epidemic. We firstly create a dataset DINIT that contains the initial conditions of variables:

TIME – Day of the epidemic

S – Number of susceptible individuals

I – Number of Infectious individuals

R – Number of recovered individuals

We assume the population size (N) is 1000 and only one infectious individual seeds the epidemic. A period of 90 days is set to be observed.

```
data dinit;
  s = 1000;
  i = 1;
  r = 0;
  do time = 1 to 90;
    output;
  end;
run;
```

In the SAS PROC MODEL, the PARMS statement is used to declare the parameter values. We additionally create the GAMMA and BETA parameters for modeling convenience. DERT function is used to draw derivatives of compartments S, I, and R with their corresponding equations. The compartments required to be solved are declared after the statement SOLVE. The option OUT= specifies the output dataset that contains the solved values of compartments by time. The following example is an influenza epidemic with a $R_0=1.4$ (R_0) assuming the infectious period (INF) as 3 days:

```
proc model data = dinit;

  /* Parameter settings */
  parms N 1000 R0 1.4 inf 3;
  gamma = 1/inf;
  beta = R0*gamma/N;

  /* Differential equations */
  dert.s = -1*beta*s*i;
  dert.i = beta*s*i-gamma*i;
  dert.r = gamma*i;

  /* Solve the equations */
```

```

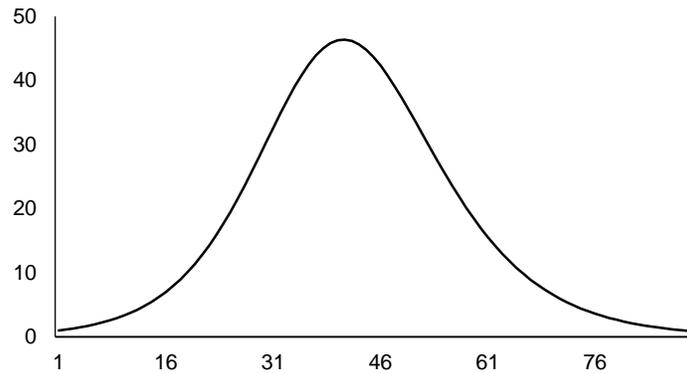
    solve s i r / out = simepi;
run;

```

The output dataset SIMEPI excluding unnecessary variables are as follow:

time	s	i	r
1	1000.00	1.0000	0.000
2	999.50	1.1425	0.357
3	998.93	1.3050	0.764
4	998.28	1.4901	1.229
5	997.54	1.7010	1.760
6	996.69	1.9411	2.366
7	995.73	2.2140	3.058
8	994.63	2.5241	3.846
9	993.38	2.8761	4.745
10	991.96	3.2751	5.769
...

This is the prevalence I by time:



USING ARRAYS TO SIMULATE AN EPIDEMIC

Apart from using SAS PROC MODEL, we can also use arrays in DATA steps to generate an epidemic. The SAS code is as follow:

```

data arrepi (keep=t s i r);

    /* Parameter settings */
    N = 1000;
    R0 = 1.4;
    inf = 3;
    gamma = 1/inf;
    beta = R0*gamma/N;

    array s_arr(90);
    array i_arr(90);
    array r_arr(90);

    do t = 1 to 90;

```

```

/* Initial conditions */
if t = 1 then do;
  s_arr(1) = 1000;
  i_arr(1) = 1;
  r_arr(1) = 0;
end;
else do;
  s_arr(t) = s_arr(t-1)-beta*s_arr(t-1)*i_arr(t-1);
  i_arr(t) = i_arr(t-1)+beta*s_arr(t-1)*i_arr(t-1)-gamma*i_arr(t-1);
  r_arr(t) = r_arr(t-1)+gamma*i_arr(t-1);
end;

/* output the compartments */
s = s_arr(t);
i = i_arr(t);
r = r_arr(t);
output;
end;
run;

```

In the DATA step, the parameter settings is the same as the previous example of PROC MODEL. The arrays of S, I, and R are declared as S_ARR, I_ARR, and R_ARR with a period of 90 days. The initial conditions (S(0)=N, I(0)=1, R(0)=0) are set for day one. And then for each element in the arrays, we store the discrete changes using DO LOOP function. The final values of compartments are output and kept in the dataset ARREPI. The results are the same as the previous example of PROC MODEL.

STOCHASTIC SIR MODEL

In the deterministic models, the incidence (βSI) represent the average number of infections by time. When the stochastic variation is adopted by assuming the incidence follows a Poisson distribution i.e. *Poisson*(βSI), we can use a similar SAS procedure to simulate an stochastic curve. In the previous example, we can simply replace the equations of arrays as:

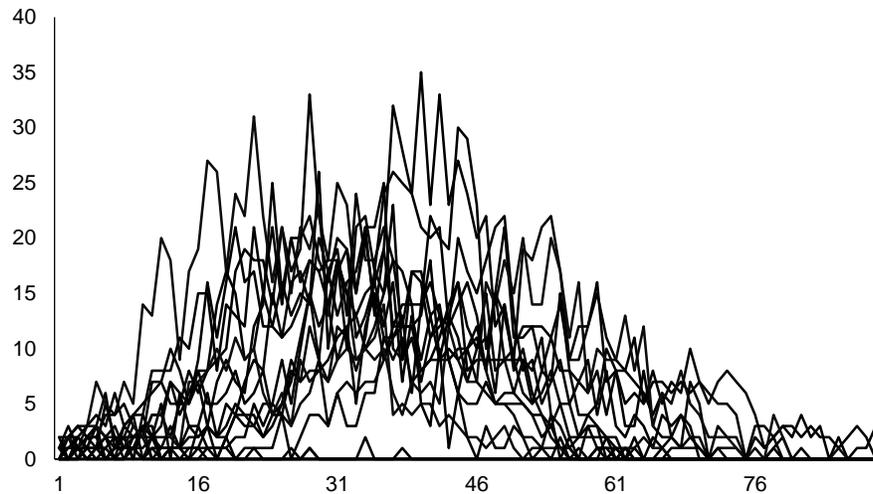
```

do t = 1 to 90;

/* Initial conditions */
if t=1 then do;
  s_arr(1) = 1000;
  i_arr(1) = 1;
  r_arr(1) = 0;
end;
else do;
  poicase = ranpoi(0, beta*s_arr(t-1)*i_arr(t-1));
  s_arr(t) = s_arr(t-1)-poicase;
  i_arr(t) = i_arr(t-1)+poicase-gamma*i_arr(t-1);
  r_arr(t) = r_arr(t-1)+gamma*i_arr(t-1);
end;
end;

```

RANPOI function is used in the variable POICASE to generate a Poisson incidence. 20 simulated incidence curves is as follow:



FIT INCIDENCE DATA TO SIR MODEL

Apart from simulation, sometimes we have to fit the surveillance data to obtain an estimate of R_0 in order to determine the required control measures for an epidemic. We employ the 2009 pandemic influenza A/H1N1 in La Gloria, Veracruz as an example (Fraser *et al.* 2009). Suppose we adapt the surveillance of reported cases from 15 February, 2009 to 13 April, 2009 and convert it to a SAS data file named FLUDATA. Variable CASE is the incidence data by time. Similar PROC MODEL steps are used. Suppose the population size is 2155, the SAS code is as follow:

```
proc model data = fludata;

    /* Parameters of interest */
    parms R0 1.4 i0 2;
    bounds 1 <= R0 <= 3;

    /* Fixed values */
    N = 2155;
    inf = 3;

    /* Differential equations */
    gamma = 1/inf;
    beta = R0*gamma/N;
    if time = 1 then do;
        s = N;
        i = i0;
        r = 0;
    end;
    else do;
        dert.s = -1*beta*s*i;
        dert.i = beta*s*i-gamma*i;
        dert.r = gamma*i;
    end;
    case = beta*s*i;

    /* Fit the data */
    fit case / outpredict out = epipred;
run;
```

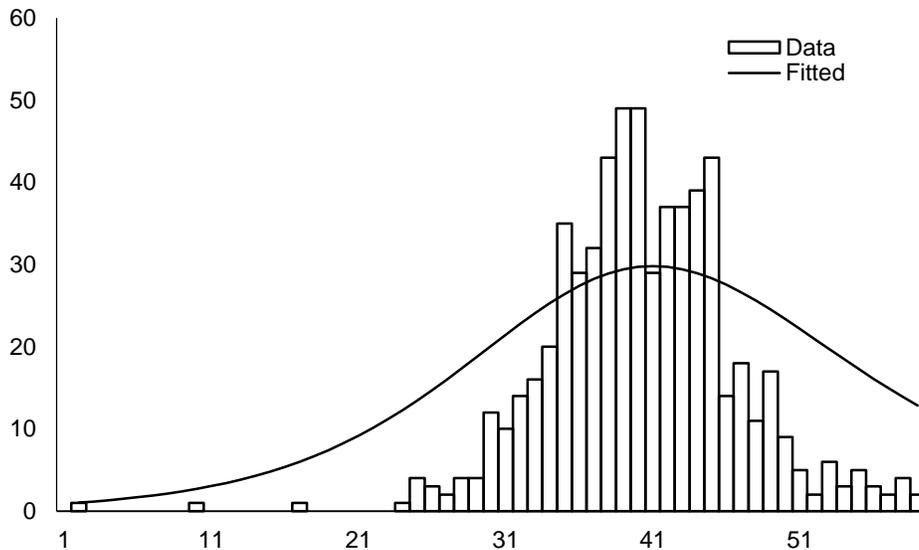
In the SAS PROC MODEL, the parameters of interest are declared in the PARMS statement i.e $I(0)$ and R_0 with their initial values 2 and 1.4 respectively. BOUND statement is used to impose boundary constraints to specified parameters. In the differential equations, we specify CASE as the model generated values. By using the FIT statement, the estimates can be obtained by the nonlinear ordinary least square (OLS) fitting method. Option

OUTPREDICT is specified in the FIT statement to obtain the predicted values in the dataset EIPRED declared in the OUT= option. The estimation summary is showed:

Nonlinear OLS Summary of Residual Errors							
Equation	DF Model	DF Error	SSE	MSE	Root MSE	R-Square	Adj R-Sq
case	2	56	5942.3	106.1	10.3011	0.5228	0.5142

Nonlinear OLS Parameter Estimates				
Parameter	Estimate	Approx Std Err	t Value	Approx Pr > t
R0	1.364329	0.0225	60.61	<.0001
i0	2.28682	0.5988	3.82	0.0003

The above two tables show the modeling fitting results and the parameter estimates. By using the predicted values in the dataset EIPRED, we can plot the fit curve as follow:



SUSCEPTIBLE-EXPOSED-INFECTIOUS-RECOVERED (SEIR) MODEL

Based on the Kermack and McKendrick SIR model, many epidemic models are further developed. SEIR model is another common epidemic model with adding an Exposed (latent) compartment on SIR model. Latent period is defined as the period of time that individuals get infected but not yet infectious. Once susceptible individual get infected, they will refer to the Exposed E-stage and followed by Infectious I-stage. The flow is shown in the following figure 2:

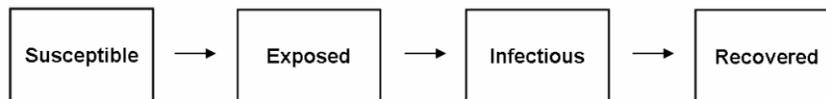


Figure 2. Flow of SEIR model

The latent period is also assumed exponential distributed, so the average latent period is equal to $1/\alpha$. Following similar configuration of SIR model, the system of nonlinear differential equations of SEIR model can be written as:

$$\begin{aligned}\frac{dS}{dt} &= -\beta SI \\ \frac{dE}{dt} &= \beta SI - \alpha E \\ \frac{dI}{dt} &= \alpha E - \gamma I \\ \frac{dR}{dt} &= \gamma I\end{aligned}$$

The formula of the reproduction number in SEIR model is the same as that in SIR model. However, SEIR has a slower growth rate as the susceptible individuals require to pass through the latent class before contributing to the disease transmission process.

SAS CODE

As similar to the previous example, we firstly create a dataset DINIT that contains the SEIR stages with an additional compartment E (number of latent individuals). Assuming the population size is equal to 1000 and only one latent and infectious individuals, the corresponding SAS code is as follow:

```
data dinit;
  s = 1000;
  e = 1;
  i = 1;
  r = 0;
  do time = 1 to 60;
    output;
  end;
run;
```

In the SAS PROC MODEL, additional parameter of ALPHA is created which represents the rate of latent individuals transported to the infectious stage. Compartment E is also developed in the differential equations. Supposed 1 day of latent period (LAT) and $R_0=2$, the SEIR curves can be simulated by solving the compartments. The SAS coded is as follow:

```
proc model data = dinit;

  /* Parameter settings */
  parms N 1000 R0 2 inf 3 lat 1;
  gamma = 1/inf;
  alpha = 1/lat;
  beta = R0*gamma/N;

  /* Differential equations */
  dert.s = -1*beta*s*i;
  dert.e = beta*s*i-alpha*e;
  dert.i = alpha*e-gamma*i;
  dert.r = gamma*i;

  /* Solve the equations */
  solve s e i r / out = simepi;
run;
```

And the equivalent SAS code in arrays:

```
data arrepi (keep=t s e i r);

  /* Parameter settings */
  N = 1000;
  R0 = 2;
  lat = 1;
  inf = 3;
```

```

alpha = 1/lat;
gamma = 1/inf;
beta = R0*gamma/N;

array s_arr(60);
array e_arr(60);
array i_arr(60);
array r_arr(60);

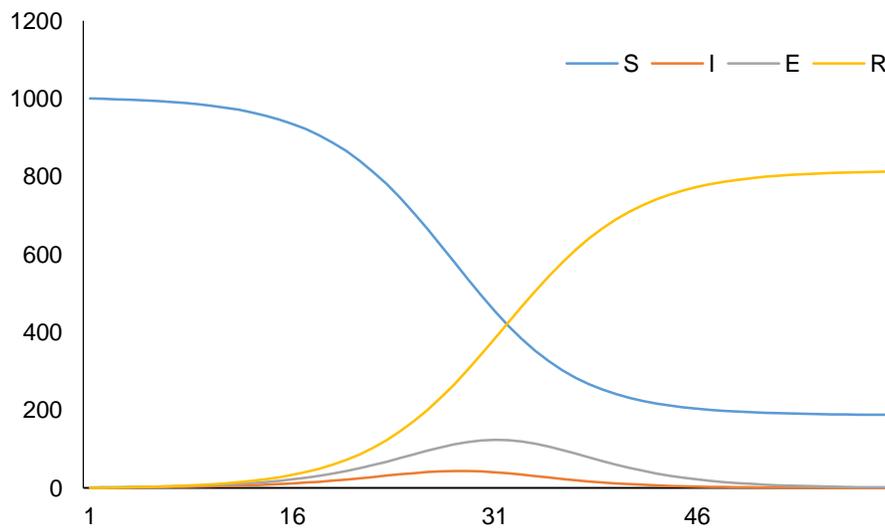
do t = 1 to 60;

    /* Initial conditions */
    if t = 1 then do;
        s_arr(1) = 1000;
        e_arr(1) = 1;
        i_arr(1) = 1;
        r_arr(1) = 0;
    end;
    else do;
        s_arr(t) = s_arr(t-1)-beta*s_arr(t-1)*i_arr(t-1);
        e_arr(t) = e_arr(t-1)+beta*s_arr(t-1)*i_arr(t-1)-alpha*e_arr(t-1);
        i_arr(t) = i_arr(t-1)+alpha*e_arr(t-1)-gamma*i_arr(t-1);
        r_arr(t) = r_arr(t-1)+gamma*i_arr(t-1);
    end;

    /* output the compartments */
    s = s_arr(t);
    e = e_arr(t);
    i = i_arr(t);
    r = r_arr(t);
    output;
end;
run;

```

The SEIR curves (individuals by times) are as follow:



CONCLUSION

Compartment models are useful to describe the individual movements in a system especially for the disease transmission dynamics in an epidemic. In this paper, we proposed using SAS to develop the compartmental models through PROC MODEL and using arrays in DATA steps. We described how to generate the deterministic and stochastic models from the systems of differential equations. Application to the 2009 pandemic A/H1N1 was also demonstrated. The SEIR model was further described in the SAS environment. In summary, the SAS program assists researchers to conduct compartmental modeling for epidemics analysis conveniently.

REFERENCES

- Hamer WH (1906). Epidemic disease in England. *The Lancet*, 1:739
- Kermack WO and McKendrick AG (1927). Contributions to the mathematical theory of epidemics, part I. *Proc. Roy. Soc. London*, 772:700-721
- Ross R (1916). An application of the theory of probabilities to the study of a priori pathometry. *Proc. Roy. Soc. London*, 92:204:230
- Gallop RJ (2009). Modeling General Epidemics: SIR MODEL. The NorthEast SAS Users Group
- Grafe C (2014). Using Arrays for Epidemic Modeling in SAS®. SAS Global Forum. Paper 1624.
- Fraser, et al (2009). Pandemic Potential of a Strain of Influenza A (H1N1): Early Findings. *Science* 324:1557-1561.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Marc Chong
Enterprise: Centre for Clinical Research and Biostatistics, The Chinese University of Hong Kong
Address: Rm502, JC School of Public Health and Primary Care, The Chinese University of Hong Kong, Shatin, NT, Hong Kong, China
Work Phone: +852 2252 8865
Fax: +852 2646 7297
E-mail: marc@cct.cuhk.edu.hk
Web: www.cct.cuhk.edu.hk

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration.

Other brand and product names are trademarks of their respective companies.