모델링; 문제를 바라보는 다양한 시각

한 준 희
농림축산검역본부 역학조사과
drhosaker@korea.kr

문 운 경
농림축산검역본부
역학조사과/수의연구관
vetmoonok@korea.kr

1. 들어가는 글

영화 "World War Z"와 "Contagion"은 치명적인 전염병이 인간에게 나타나 매우 빠른 속도로 퍼지고, 이에 의해 나타나는 인류의 반응과 이를 극복하기 위해 여러 방안을 모색해 나가는 과정을 그린다는 공통점을 갖고 있다. 그리고, 국제적으로 저명한 과학자들이 모여 현재 통산 몇 명이 감염되었고 앞으로 얼마나 더 감염이 진행될지 예상하는 장면은, 이러한 영화들에서 공통적으로 나타나는 부분이다. 그런데 여기서 이상한 점이 하나 있다. 그 저명한 과학자들이 감염된 동네에 가서 하나하나 감염자 수를 세었을 리 없고, 좀비가 되어버렸거나 혹은 죽은 사람들에게 전화를 걸어서 동족(?)의 수가 몇이나 되냐고 물었을 리도 없는데 과연 그들은 그 수를 어떻게 알았을까? 잠시 주제를 돌려보자. 볼펜을 생산하는 공장에서 생산라인을 더 늘리려고 하는데, 얼마나 늘려야 수지타산에 맞는지 확인하고 싶다고 한다면, 이를 알아볼 수 있는 방법은 무엇이 있을까? 직접 공장 시설을 지어 라인을 늘려 운영해보면 확실히 알 수 있겠지만, 만약 5개의 새로운 라인을 건설하고 2년이 흐르고 나서야 수지타산에 맞지 않는다는 것을 '확실히' 알았다면, 정말 어처구니가 없는 일이라고 할 수 있겠다. 위와 같이, 현실에서의 상황이나 사건을 이해하고 문제를 해결해야 하는데, 직접 모든 경우를 선행하는 것이 여의치 않을 때 적용할 수 있는 기술을 우리는 모델링(Modeling)이라고 한다.

2. 모델링?

모델링이란 무엇인가? '모델링이란 물리적인 과정에 대한 평가와 이해를 돕기 위해 설계된, 그 과정에 대한 복제(representation)이다… 좀더 구체적으로, 모델링이란 정량적인 수학적(quantitative mathematical) 용어를 이용하여 사건을 복제하는 것이며, 이를 통해 그 사건에 대한 예측을 할 수 있다' [1]. 모델링을 보고 TV 광고 속, '원빈'이나 '캔디스 스와네포엘'과 같은 아름다운 모델이 떠올랐다면 유감이긴 하여도, 위에서 나오는 모델링의 정의는 조금 어려워서 무슨 말인지 잘 모르겠다. 쉽게 설명하면, 모델링이란 모델을 만드는 작업이고, 그 모델이란 현실에 있는 특정 체계의 구성이나 작용방식을 재현한 것이라고 할 수 있다. 그리고 구성이나 방식을 재현하는데 있어 수학이라는 다소 난해한 도구를 이용하는 것이다. 가령, 도입부에 나온 것처럼, 사람들이 좀비로 변해가는데, 얼마나 빠른 속도로 변해가는지를 저명한 과학자들이 파악하기 위해서는, 우선 '사람-좀비 변화모델'이라는 것을 만들고, 감염된 사람들이 변해가는 속도 등을 수학적으로 고려하여, 이를 통해 언제쯤이면 얼마나 변해있을지를 예측할 수 있다는 말이다. 혹은 공장 시설을 얼만큼 늘려야 향후 몇 년간의 수익이 개선되는지를 알고자 할 때, 현재의 공장 생산과 수입에 관한 모델을 만들고, 생산 및 수입과 관련된 현재의 상황을 수식으로 표현한 뒤, 거기에 '설비의 추가'라는 변수가 주어지면 어떠한 결과가 나타날 지를, 굳이 직접 지어보지 않고도, 예측할 수 있다는 뜻이다. 이러한 모델링은 비단 질병 전파와 같은 의학 혹은 엔지니어링 분야뿐만 아니라, 금융, 군사분야 등 다양한 영역에서 적용되고 있다. 신문을 통해 남북한 군사의 모의전투를 통하여 둘의 전투력을 비교하거나 스텔스 전투기 한 대가 몇 대의 일반 전투기를 상대할 수 있다는 내용의 기사를 종종 접해보았을 것인데, 이러한 내용 또한, (직접 전투가 일어난 결과지만 사실 우리가 모르고 있는 것이 아니라면) 모델링을 통한 결과이다.

그렇기 때문에 이렇게 모델링을 사용함으로써 얻을 수 있는 장점은 다음과 같다. 1) 수학적 모델링을 통하여 현상이나 체계에 대한 더 나은 이해를 얻을 수 있고, 나아가 대상 체계에 원하는 변화가 가능한지를 시험할 수 있다; 2) 모델을 사용하면 시간의 제약을 받지 않는다. 즉, 시뮬레이션을 통하여 몇 년의 결과값을 몇 분만에 얻어낼 수 있다; 3) 현재의 체계에 속한 여러 변수 중에서 어느 것이 결과에 가장 큰 영향을 미치는지 파악할 수 있고, 그러한 변수들 사이의 관계에 대해서도 알 수 있다. 그렇다면, 이렇게 일상에서 유용하게 사용될 수 있는 모델은 어떠한 과정을 통해 만들어지는 것일까?

3. 모델 설정 과정

일단 모델을 설정하는 과정을 살펴보기에 앞서, 우리는 위의 모델링에 관한 설명에서 "복제"와 "정량적인 수학"에 주목할 필요가 있다. 왜냐하면, 결국 모델링이라는 도구는, 1) 모델의 설계자(모델러)가 어떠한 모델 골격으로 현상을 복제하는가, 그리고 2) 그 복제된 상황 사이에 적용되는 변수의 값이 얼마인가에 따라 결정된다고 할 수 있기 때문이다. 모델링이 현실의 특정 구조를 파악하고 예측하는데 쓰이는 도구이기 때문에, 모델이 제대로 된 예측을 하기 위해선 대상으로 하는 현실 체계와 그 골격이 매우 비슷하여야 하고 이를 구성하는 요소들을 반드시 포함하고 있어야 한다는 말이 된다. 가령, '사람–좀비 변화모델'의 골격 안에는 잠재 감염기(Latent period)나 감염력, 감염 경로 등이 있어야 하고, 볼펜 공장 모델에는 재료비, 인건비, 생산에 소요되는 시간과 같은 요소들이 반드시 포함되어야 하며, 그렇지 못한 경우 이는 불완전하거나 전혀 쓸모 없는 모델이 된다는 의미이다. 따라서, 관련 요소들을 고려하면 고려할수록 현실에 가까운 모델이 될 수 있고, 복잡 다단한 현실을 더욱 자세하게 반영할 수 있다. 이러한 일들이 예전에는 불가능했지만 요즘에는 슈퍼컴퓨터의 개발로 어느 정도는 가능해지고 있다고 한다. 하지만, 위에서처럼 관련 요소들을 많이 고려할수록, 안타깝게도 모델의 골격은 더 복잡해지기 때문에, 너무 복잡해서 모델 구조를 이해하기 어렵다거나 해당 모델을 통해 시뮬레이션을 진행하기 난해해지게 되므로, 이는 결국 적합한 모델이라 할 수 없게 된다. 그렇다고 해서, 모델을 최대한 단순하게 만들어서 특정 상황을 표현한다면, 앞서 말한 바와 같이 현실을 제대로 표현하는데 한계점을 들어낼 것이다. 모든 것은 가능한 가장 단순하게 만들어야 하지만, 그렇다고 해서 너무 지나치게 단순해서는 안 된다는 저명한 과학자 아인슈타인의 말처럼("Everything should be as simple as possible, but no simpler"), 모델을 처음에 설계할 때에는 현상을 반영하는 변수를 어디까지, 어떻게 선택하고 설정할 것인가에 대한 문제가 현상을 분석하는데 중요하다고 할 수 있다. 따라서 똑같은 상황에 대해서 다른 모델러가 다른 기준으로 분석을 한다면, 역설적으로, 결과값이 조금 차이가 나거나 어쩌면 아예 다른 값이 나올 수 있다.

또한, 모델링에는 수학이 반드시 필요하다. 이는 잠시만 생각해보면 너무나도 당연한 내용이다. 현실이 변해가는 양상을 얼마나, 어떻게 라는 기준으로 표현하려면 숫자를 이용하지 않고는 딱히 다른 방법이 떠오르지 않을 것이다. 예를 들어, 돼지에게 치명적인 질병 A가 있다고 생각해보자. 모돈 100두 규모의 돼지 농장에 질병 A가 발생했다면, 질병에 감수성이 있는 개체와 질병에 걸린 개체, 그리고 질병에 의해 폐사하는 개체의 수는 시간에 따라 달라질 것이고, 이 값은 충분히 수치화할 수 있을 것이다. 그리고 이렇게 수치화 된 값을 통하여, 개체들이 변화해가는 어느 비율을 산출할 수 있을 것이고, 이 비율들이 '질병 A 모델'에 필요한 변수나 값들이 되어 모델을 완성할 수 있게 되는 것이다. 그럼 여기서 또 의문점이 생긴다. 위의 사례에서 얻은 비율이나 값이 항상 같다고 말할 수 있을까? 다시 말해, 위의 모돈 100두 규모농장에서 질병 A에 의한 첫날 폐사축 수가 12마리였다고 가정했을 때, 그렇다면 우리는 "전국 모든 모돈 100두 규모농장은 질병 A에 의해 첫날 폐사 두수가 12마리이다." 라고 말할 수 있을까? 값이 어느 정도 비슷하게 나올지도 모르겠지만, 답은 "그럴 리가 없잖아!" 이다. 모델의 변수를 설정할 때 참고하게 되는 데이터가 어떻게 되느냐에 따라서 모델의 변수들이 달라지고, 마찬가지로 모델의 결과도 다르게 나타나게 된다. 따라서, 모델링을 이해하기 위해서는 '모델러의 설계'와 '정량화된 변수' 라는 큰 두 축을 염두하여야 하고, 이들에 따라 동일한 현상을 해석한 두 모델이 전혀 다른 결과값을 낼 수 있다는 것을 알아야 하겠다. 그렇다면 이제, 모델을 설계하기 위한 과정을 살펴보도록 하자.

그림 1. (2)모델 개발과 사용 단계에 대한 모식도.

그림 1. 을 통하여 모델 개발 및 평가의 단계를 간단하게 도식화하였다. 현실의 특정 문제를 인식하고 이와 관련된 모델을 설계하기 위해 문제와 관련된 변수를 설정하는, 즉 사실관계를 이해하는 것에 대한 이야기는 이미 위에서 언급을 했다. 그 다음으로 모델링 기법을 선택하는 단계가 있는데 이를 설명하기 위해서 전에 사용하였던 돼지질병 A를 다시 생각해보도록 하겠다. 종전의 가정에서 모돈 100두 농장(이제부터 편의상 '가 농장' 이라 부르겠다)의 질병 A 감염에 따른 첫날 폐사축이 12마리였다. 그렇다면 가 농장에 질병 A가 발생하기 전의 시점으로 시간을 거슬러 올라간 뒤에, 앞서의 가정과 똑같은 시점에서 가 농장에 A가 다시 발생한다는 불가능한 가정을 해보자. 그렇다면 과연 이번에도 앞에서와 같은 두수인 12마리가 감염 첫날 폐사할까? 무슨 평행우주 같은 이야기를 하고 있냐고 물을 사람이 있겠지만, 여기서 말하고 싶은 점은, 군집의 감염패턴이 일정한 대표값에 의해 나타날 수도 있지만, 반대로 어느 확률분포 범위 안에서 랜덤하게 나타날 수도 있다는 것이다. 내용이 몹시 복잡하게 느껴진다면 또 한번 유감이지만, 간단하게 말하면 현실의 사건이 발생하는 양상은 특정 수치에

맞추어 결정되어 있다고 볼 수도 있고, 반대로 확률론적으로 나타날 수도 있다는 것이다. 여기서 전자의 경우를 결정론적(deterministic) 모델, 후자를 확률론적(stochastic) 모델이라고 부른다. 결정론적 모델에서는 변수들을 특정한 대표값으로 계산하기 때문에, 결과도 하나의 값으로 이미 결정되어 있지만, 확률론적 모델에서는 변수들이 어느 범위 안에서 다양한 값을 취할 수 있기 때문에 결과 역시 범위의 값으로 나타나게 되어 있다. 두 방법 모두 나름의 장점과 단점을 갖고 있지만, 실제로 모델을 설정할 때 두 접근법 모두에서 완전히 자유롭게 만들기란 사실 어렵다. 따라서, 두 방법을 적절히 사용하여 효율적인 모델을 만드는 것이 중요하다고 할 수 있겠다. 이렇게 기법까지 설정을 하였으면, 기법에 맞추어 필요한 변수들을 구체화할 수 있고 궁극적으로 모델을 설정하게 된다. 그리고 설정된 모델을 시뮬레이션을 통하여 결과를 도출하여보고, 그 결과를 현실의 문제와 비교하여 평가한 뒤 최적화하여, 비슷한 미래상황에 대한 예측을 가능하게 하는 것이다. 이렇게 말로만 설명하니 조금 애매모호하다. 따라서, 좀더 쉬운 설명을 위해 가 농장의 불쌍한 돼지들을 마지막으로 불러와야겠다.

4. 모델링의 예

어느 양돈 수의사가 '데일리벳'을 통하여 '이웃나라에서 질병 A가 발생하여 양돈산업에 금전적인 손해를 입히고 있다'는 뉴스를 접하고, '컨설팅 고객 중 하나인 가 농장에 질병 A가 발생할 경우 얼마나 피해가 나타날 지 분석을 해봐야겠다(모델 설정의 첫 단계인 문제 인식)'는 생각이 불현듯 들었다. 그래서 몇몇 논문을 찾아보니, A 질병에 대한 돼지들의 상태를 [감수성 단계(Susceptible) → 잠재감염 단계(Exposed) → 감염 단계(Infectious) → 폐사 단계(Removed)]로 구분할 수 있고 각각 잠재 감염 단계는 1일, 감염 단계는 5일 정도로 볼 수 있다는 결과를 얻었다. 가 농장의 모돈이 100마리 규모였으니 전체 두수를 1,000마리로 예상하고, 질병 A의 감염시작은 1,000마리 중 한 마리에서 발생하며, 원하는 연구기간 동안 농장 내 돼지의 추가적인 유입이나 유출은 없다고 가정하자(문제에 대한 사실관계 이해 및 모델 변수들의 구체화). 이제 우리의 '질병 A 모델'은 단위 시간당 단계에서 다음단계로 넘어가는 수식만 알면 완성이 된다(모델의 설정)! 질병이 발생한 후, 시간 t에서 감수성 단계, 잠재감염 단계, 감염 단계 그리고 폐사 단계에 해당하는 개체 수를 각각 S_t, E_t, I_t, R_t로 표현한다면, 수식을 이용하여 다음 단위시간의 단계별 개체 수를 그림 2와 같이 표현할 수 있다.

$$S_{t+1} = S_t - (\lambda_t \times S_t)$$
$$E_{t+1} = E_t + (\lambda_t \times S_t) - (f \times E_t)$$
$$I_{t+1} = I_t + (f \times E_t) - (\gamma \times R_t)$$
$$R_{t+1} = R_t + (\gamma \times R_t)$$

그림 2. 각 단계의 기준 시간 별 총 개체 수.

여기서 단위시간을 1일로 설정한다면, 위의 식은 특정 단계의 개체 수는 1) 전날의 해당 단계 개체 수와, 2) 전 단계에서 해당 단계로 새롭게 넘어온 수를 더한 값에, 3) 해당 단계로부터 다음 단계로 넘어간 수를 제외한 값이라는 뜻이고, 해당 단계로 넘어오고, 넘어가는 비율은 하나의 대표값으로 정해져 있다는 것이다. 따라서, 위 식의 경우, 감수성 단계로부터 폐사 단계까지 넘어가는 각각의 비율은 λ_t, f, γ 로 표현할 수 있고, λ_t를 제외하고 f 와 γ 는 1을 각각 집단 내 평균 잠재감염 단계(1일이므로, f는 1)와 감염 단계(5일이므로, γ 는 0.2)로 나눈 값이라 유추할 수 있다. 앞에서 λ_t를 제외한 이유는, 이 값이 단순히 1을 감수성 개체가 병원체에 접촉되는 평균 기간으로 나눈 값이 아니기 때문이다. 단위 시간 동안 병원체를 배출하는 개체와 감수성 개체가 서로 만나 감염이 되는 사건은 수학적으로 $\beta \cdot S_{(t)} \cdot I_{(t)}$로 표현할 수 있고, 여기서 β 란 전파계수로, 단위 시간 동안 집단 내에서 감수성 단계와 감염 단계의 개체가 서로 만나 감염이 일어나는 유의한 접촉을 이루는 값을 의미한다. 따라서 λ_t 란, 감수성 개체에 감염을 일으킬 유의한 접촉을 이루는 값과 감염 단계의 개체 수를 곱한 값이므로, 쉽게 말하면 질병의 전파력을 의미하는 값이라고 할 수 있다. 그리고 우리의 양돈 수의사가 논문을 찾아본 결과 β 값이 0.00212였다고 가정하자. 자, 이제 질병 A 전파모델이 완성되었다. 위의 자료를 통하여 모델을 시뮬레이션 하면 아래의 그림 2.와 같은 결과를 도출할 수 있다. 참고로 부연설명을 하면, 이와 같이 질병의 전파양상을 나타내는 모델을 '질병전파모델'이라고 하고, 특히 질병에 대한 대상을 감수성, 잠재감염, 감염, 그리고 폐사와 같이 나누는 모델을 SEIR(각 단계의 영문 알파벳 앞 글자) 질병전파모델 이라고 한다. 이는 실제로 역학 분야에서 널리 사용되는 모델이며, 질병의 양상에 따라 다양한 전파모델을 만들 수 있다.

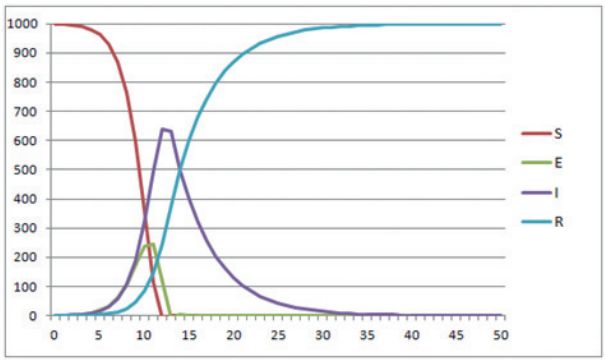

그림 2. 가 농장의 질병 A 전파 SEIR 모델.

이 모델에 따르면, 가 농장에서 질병 A에 감염된 돼지 한 마리가 생길 때, 대략 10일에서 15일 사이부터 폐사 개체 수가 폭발적으로 증가하고, 12일경 이후에는 질병 A에 감수성이 있는 개체가 존재하지 않는다는 것을 알 수 있다. 이 그래프를 보고 그 양돈수의사가 이웃나라와의 발생현황과 비교를 해보니, 가 농장의 시뮬레이션 결과가 이웃나라에 비해 터무니없이 빠르게 진행된다는 것을 알았다. 생각해보니 그 차이의 이유는, 앞서 말한 바와 같이, 이웃나라와 가 농장의 상황이 다르고, 그렇기 때문에 모델에서의 변수도 그 차이가 어느 정도 반영되었어야 하기 때문인 것으로 판단되었다. 따라서 그는 다른, 가령 가 농장과 비슷한 여건의 농장에서 발생했던 질병 A의 사례를 다룬 논문을 참고해 변수를 수정하여 보다 최적화된 모델을 만들 수 있을 것이다.

5. 마치며

수학자인 [2]Emilia Vynnycky는, "모델은 현실의 단순화이기에 이를 통해 우리는 데이터의 패턴을 이해할 수 있으며, 운이 좋으면 이러한 패턴을 설명하는 기본적인 통찰을 깨달을 수 있다"고 말했다. 패턴을 이해한다는 것은 현실의 문제를 분석하고 파악한다는 뜻이며, 기본적인 통찰을 깨닫는다는 것은 모델을 통해 앞으로를 예측할 수 있다는 뜻이라고 유추해볼 수 있다. 그렇다면, 왜 그녀는 '운이 좋으면' 이라는 표현을 굳이 사용하였을까? 이를 이해하기 위해, 모델링 기법을 선택하는 단계로 다시 돌아가보자. 결정론적인 모델을 선택할 것이냐 확률론적인 모델을 선택할 것이냐는 문제에서, 모델의 대상인 문제의 특성을 반영하여야 하는데, 그 문제가 위에서 언급한대로 언제나 결정론적인 경과를 따른다고 할 수는 없지만, 반대로 항상 확률론적으로 진행된다고 말할 수도 없다. 그렇기 때문에 모델을 결정론적으로 그려나갈지 아니면 확률론적으로 풀어나갈지는 이를 관찰하는 사람의 선택에 몫이다. 그리고 이 선택은 관찰자의 목적에 따라 달라진다고 할 수 있으며, 그러한 이유로, 모델이 말이 되느냐 안되느냐에 대한 논란은 충분히 발생할 수 있게 된다. 또한, 모델에 적용되는 다양한 변수들은 앞서 언급한 바와 같이 상황에 따라 다르기 때문에, 가령 질병전파 모델의 경우 결국 하나의 모델로 언제 어디서건 특정질병의 전파를 설명한다는 것은 거의 불가능에 가깝다. 하지만, 그 모델을 통하여 해당 질병의 특징과 성격을 파악할 수 있기 때문에, 미래의 질병 발생에 있어 이를 이해하고 예측하는데 도움이 될 수 있는 도구로 사용할 수 있는 것이다.

100명의 사람을 한 곳에 모아놓고 모두 동전 하나를 던지게 한다고 생각해보자. 첫 번째 시도를 통하여 48명만이 앞면이 나왔었고, 그 다음의 시도에서는 57명, 그리고 그 다음에는 39명이 앞면이 나왔다고 가정하고, 이를 통해 모델을 설정하여 분석하였더니, 위와 같은 시도를 하면 43명에서 56명이 앞면이 나올 가능성이 크다는 결과가 나왔다고 하자. 그렇다면 다음의 시도에서는 반드시 43명에서 56명 사이의 사람이 앞면이 나올까? 모두가 예상하듯, 그 답은 "그럴 수도 있고 아닐 수도 있다"이다. 바꾸어 말하면, 현재 문제의 양상이 이러하였고 이를 모델로 해석하여 다음과 같은 결과가 나왔으니, 다음에 문제가 발생하면 반드시 이러할 것이다 라고 말하는 것은 몹시 어렵다는 말이고, 이는 단추 몇 개만 누르면 질병이 퍼져나가는 그래프와 지도가 그려지는 영화의 한 장면이 말 그대로 픽션일 뿐이라는 이야기이기도 하다. 하지만 반대로 생각해보자. 질병과 관련된 예측이나 정보 없이 무방비상태로 이를 맞이하는 것과 다소 거칠지만 그런대로의 예측값을 알고 있는 상태로 미리 준비하고 대응하는 것 중 어느 것이 더 나은 선택일까? 당연히 후자이다. 이렇게 모델링, 그리고 역학은 우리에게 문제에 맞서 준비하고 대응할 수 있는 다양한 도구이자 무기를 제공한다. 비록 이러한 분석이 현시대의 기술로는 아직은 무리가 따르고 모델을 통한 문제의 예측이 미흡하며, 혹은 모델을 이용하는 사람들의 해석에 빈틈이 있을지라도, 모델은 항상 문제 해결을 위한 하나의 길을 제시한다는 점에서 모델링은 여전히 내일을 예측하는 최선의 방침들 중 하나라 말할 수 있을 것이다.

참고문헌

(1) Thrusfield M, 2007, Veterinary Epidemiology, 3rd ed, Wiley-Blackwell.
(2) Vynnycky E, White RG, 2010, An Introduction to Infectious Disease Modelling, Oxford University Press.
(3) Taylor HM, Karlin S, 1998, An Introduction to Stochastic Modeling, 3rd ed, Academic Press.