Journal of the Korean Data & Information Science Society
2015, **26**(6), 1367–1375

http://dx.doi.org/10.7465/jkdi.2015.26.6.1367
한국데이터정보과학회지

선형성장모형에 대한 ROC 곡선과 AUC

홍종선[1] · 양대순[2]

[12]성균관대학교 통계학과

접수 2015년 9월 10일, 수정 2015년 10월 16일, 게재확정 2015년 11월 9일

요 약

경시적자료의 분석으로 선형성장모형을 고려한다. 시간효과를 고려하는 모형과 임의효과를 추가하는 모형 그리고 가변수가 추가된 모형을 설정한다. 본 연구는 정규분포로 가정한 다양한 자료를 생성하고, 다양한 선형성장모형에 대하여 binormal ROC 곡선과 AUC 통계량을 여러 시점에서 구하여 비교 분석하였다. 공분산의 크기가 증가할수록 그리고 시간이 경과할수록 ROC 곡선은 다른 형태로 나타나며 AUC 값은 서서히 증가한다. 반대로 공분산이 작아질수록 시간이 경과함에 따라 AUC의 증가폭이 커진다. 임의효과모형에서 공분산이 양인 경우에 시간이 경과할수록 임의효과모형의 분산이 증가하며 AUC의 증가량은 시간효과모형의 AUC의 증가량보다 작다. 그리고 시간효과모형의 AUC의 증가량보다 임의효과모형의 증가량이 더 크다는 것을 탐색하였다.

주요용어: 가변수, 경시적자료, 시간효과, 임의효과.

1. 서론

경시적자료 (longitudinal data)는 시간에 따라 반복적으로 관찰된 자료를 의미한다. 경제, 스포츠, 환경, 의학 등 여러 분야에서 시간의 흐름에 따른 변화를 관찰하려는 연구와 주변화 변량효과모형 등 다양한 모형을 적용하려는 연구가 활발하게 진행되고 있다 (Jeon 등, 2014). 특히 의학분야에서 특정질병이 검출되는 시간까지 다양한 시간에서의 모습을 수집한 경시적자료의 분석으로 선형성장모형 (linear growth model)을 고려하였다 (Etzioni 등, 1999). Bloom (1992)는 어린시절의 당뇨병이 아이들의성장에 미치는 연구에 대해서 선형성장모형을 사용하였고, Park (2008)은 학생들의 사교육 참여경험이 학업성취도 수준 및 향상에 영향을 미치는지를 선형성장모형을 이용하여 분석하였다. 이처럼 다양한 분야에서 선형성장모형이 사용되고 있다.

두 종류의 스코어 확률변수 (예를 들어 good/bad, default/non-default) Y^d와 $Y^{\bar{d}}$의 분포를 정규분포인 $Y^d \sim N(\mu_d, \sigma_d^2)$와 $Y^{\bar{d}} \sim N(\mu_{\bar{d}}, \sigma_{\bar{d}}^2)$를 따른다고 가정할 때, binormal ROC (Receiver Operating Characteristic) 곡선과 AUC (the area under the ROC curve) 통계량은 다음과 같이 정의한다.

$$\begin{aligned} ROC(u) &= S(a + bS^{-1}(u)) \\ AUC &= S(\frac{a}{\sqrt{1+b^2}}), \end{aligned} \tag{1.1}$$

여기서 $a = (\mu_{\bar{d}} - \mu_d)/\sigma_d$, $b = \sigma_{\bar{d}}/\sigma_d$, $-\infty < u < \infty$, $S(\cdot)$는 생존함수 (survival function)을 말하며 1-$F(\cdot)$로 정의한다 (상세한 내용은 Pepe (1997, 1998, 2000) 참고).

[1] 교신저자: (110-745) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과, 교수.
E-mail: cshong@skku.ac.kr
[2] (110-745) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과, 대학원생.

Etzioni 등 (1999)은 시간효과 (time effect)만을 고려하는 선형성장모형과 모형의 계수에 임의효과 (random effect)를 추가하는 선형성장모형을 고려하여, 특정시간마다 두 모형을 비교하였다. ROC 곡선과 AUC 통계량 그리고 이를 추정한 Tosteson (1998), Hong 등 (2012, 2013) 그리고 Hong과 Jung (2014)의 연구를 바탕으로 선형성장모형의 판별력을 측정하는 척도로써 ROC 곡선과 AUC 통계량을 계산하여 두 종류의 선형성장모형을 비교하였다. Etzioni 등 (1999)의 연구는 실증예제 분석에 중점을 두었는데 본 연구는 정규분포로 가정한 다양한 자료를 모의실험을 이용하여 생성하고 얻은 ROC 곡선과 AUC 통계량을 탐색적으로 비교 분석한다. 단순선형모형부터 가변수가 들어간 모형까지 다양한 선형성장모형을 고려하고, 여러 시점에서의 ROC 곡선과 AUC 통계량을 구하고 비교 분석한다.

본 논문의 구성은 다음과 같다. 2절에서는 선형성장모형과 임의효과를 적용한 선형성장모형을 설명하고, 임의효과가 가지는 장점에 대해서 설명한다. 그리고 보다 일반적인 가변수를 포함한 선형식을 추가로 논의한다. 3절에서는 2절에서 설명한 다양한 종류의 선형성장모형에서의 분포를 가정하고 각 분포를 따르는 자료를 모의실험을 통하여 생성한다. 생성된 자료에 대하여 ROC 곡선과 AUC 통계량을 구하고 모형들과 분포에 따라 변화량을 살펴본다. 각 모수의 분포를 다양하게 가정하고, 각 모형이 시간의 흐름에 따라 어떻게 변하게 되는지, 그리고 ROC 곡선과 AUC통계량에 어떻게 차이가 나는지를 탐색하여 비교 분석한다. 마지막으로 4절에서 결론을 유도한다.

2. 선형성장모형

2.1. 선형성장모형과 가정

선형성장모형은 시간 변수 t에 선형인 형태를 지니는 모형으로서, 시간의 변화만큼 동일한 양이 늘어나는 모형으로 시간이 증가함에 따라 변하는지를 파악할 수 있는 경시적 모형이다. 일반적인 선형성장모형과 가정은 다음과 같다.

$$\begin{aligned} Y_{ij}^{d} &= \beta_{0i}^{d} + \beta_{1i}^{d} t_{ij} + \epsilon_{ij}^{d}, i = 1, 2, \ldots, n_j, j = 1, 2, \ldots, N_d \\ Y_{ij}^{\bar{d}} &= \beta_{0i}^{\bar{d}} + \epsilon_{ij}^{\bar{d}}, \qquad\quad i = 1, 2, \ldots, n_j, j = 1, 2, \ldots, N_{\bar{d}}, \end{aligned} \tag{2.1}$$

여기서 $\epsilon_{ij}^{d} \sim N(0, \sigma_d^2), \epsilon^{\bar{d}_{ij}} \sim N(0, \sigma_{\bar{d}}^2)$를 의미한다.

다음으로 시간효과모형과 임의효과모형의 두 종류의 모형을 고려할 수 있다. 두 모형을 구분하는 방법은 시간불변의 개별효과가 독립변수들과 관련이 되어 있는지를 살펴보며 관련이 있는 경우 시간효과모형을 쓰고, 관련이 없을 경우 임의효과모형을 선택하게 된다. 임의효과모형이 유효한 경우라도 시간효과모형에 의해 산출된 계수는 여전히 일치추정량을 제공하기 때문에 연구자들은 시간불변의 특정 요소가 독립변수들과 관련되어 있는지에 대한 확실한 정보가 없을경우 임의효과모형보다 시간효과모형을 선호한다 (Pearson, 1996; Johnston, 2000). 시간효과모형의 가장 큰 장점은 개인마다 개별특성효과를 구분하여 계수를 추정한다는 데에 있다. 하지만 개별효과를 반영하는 가변수를 생성하는 과정에서 많은 자유도를 소모하게 되어 결과적으로 독립변수들에 대한 계수값 추정이 상대적으로 정확성을 잃게 되는 단점이 발생한다. 임의효과모형의 경우 시간효과모형처럼 가변수를 설정하는 과정에서 계수값 추정에 정확성이 떨어지는 위험이 적다. 시간불변의 개별효과가 독립변수와 관련이 없어야 하기 때문에 실제 분석에서 이를 만족시키기엔 어려움이 따른다 (Ashenfelter 등, 2003).

이러한 두 모형 중 우선 시간효과모형은 식 (2.1)과 같은 가정을 따르는 모형이다. 시간효과모형의 t시점에서의 ROC 곡선과 AUC 통계량은 식 (1.1)의 a와 b가 식 (2.2)로 대체된다.

$$a(t) = \frac{\beta_0^{\bar{d}} - \beta_0^{d} - \beta_1^{d} t}{\sqrt{\sigma_{\bar{d}}^2}}, \quad b(t) = \sqrt{\frac{\sigma_{\bar{d}}^2}{\sigma_d^2}}. \tag{2.2}$$

모수들이 분포를 따른다고 가정하는 임의효과모형에서의 모수들은 다음과 같이 표현한다.

$$\begin{aligned}\beta_{0i}^{d} &= \beta_{0}^{d} + \delta_{0i}^{d}, \beta_{1i}^{d} = \beta_{1}^{d} + \delta_{1i}^{d} \\ \beta_{0i}^{\bar{d}} &= \beta_{0}^{\bar{d}} + \delta_{0i}^{\bar{d}},\end{aligned} \tag{2.3}$$

여기서 $(\delta_{0i}^{d} \;\; \delta_{1i}^{d}) \sim N[\,\underline{0}, V^{d}]$. 따라서 임의효과모형에서의 확률변수 Y^{d}와 $Y^{\bar{d}}$를 다음과 같은 모형식으로 정의한다.

$$\begin{aligned}Y_{ij}^{d} &= \beta_{0}^{d} + \beta_{1}^{d} + \delta_{0i}^{d} + \delta_{1i}^{d} + \epsilon_{ij}^{d} \\ Y_{ij}^{\bar{d}} &= \beta_{0}^{\bar{d}} \qquad\qquad + \delta_{0i}^{\bar{d}} + \epsilon_{ij}^{\bar{d}},\end{aligned} \tag{2.4}$$

여기서 $\delta_{0i}^{d} + \delta_{1i}^{d} + \epsilon_{ij}^{d}$, $\delta_{0i}^{\bar{d}} + \epsilon_{ij}^{\bar{d}}$는 임의효과이다. t시점에서의 ROC 곡선과 AUC 통계량은 식 (1.1)의 a와 b가 식 (2.5)로 변환된다.

$$a(t) = \frac{\beta_{0}^{\bar{d}} - \beta_{1}^{d} t_{ij} - \beta_{0}^{d}}{\sqrt{\sigma_{d}^{2} + (1\ t) V^{d} (1\ t)'}}, \quad b(t) = \sqrt{\frac{\sigma_{\bar{d}}^{2} + V^{\bar{d}}}{\sigma_{d}^{2} + (1t) V^{d} (1\ t)'}}. \tag{2.5}$$

임의효과모형에서의 식 (2.5)를 살펴보면 시간효과모형과 다르게 분산이 시간변수 t에 의존하고 있음을 파악할 수 있다. 또한 식 (2.4)를 통해 두 모형의 오차의 분산을 비교해보면 시간효과모형의 ϵ_{ij}^{d}의 정보를 임의효과모형에서는 $\delta_{0i}^{d} + \delta_{1i}^{d} t_{ij} + \epsilon_{ij}^{d}$로 구성되어 있기 때문에 임의효과모형에서의 오차의 분산이 시간효과모형의 오차의 분산보다 작다는 장점이 있다. 임의효과모형의 분산이 시간변수 t에 의존하고 있음을 알 수 있고, 또한 β_{0}^{d}와 β_{1}^{d}간 공분산을 통해 임의효과모형의 의 분산을 예상할 수 있다. 따라서 본 논문에서는 다양한 분포의 분산을 고려하면서 ROC 곡선과 AUC 통계량을 통해 비교한다.

2.2. 선형성장모형의 확장

질적변수인 가변수 Z를 모형 (2.1)에 추가한 모형과 시간변수와 가변수의 교호작용이 추가된 모형 (2.6)과 (2.7)을 고려한다.

$$\begin{aligned}Y_{ij}^{d} &= \beta_{0i}^{d} + \beta_{1i}^{d} t_{ij} + \beta_{2i}^{d} Z + \epsilon_{ij}^{d} \\ Y_{ij}^{\bar{d}} &= \beta_{0i}^{\bar{d}} \qquad\quad + \beta_{2i}^{\bar{d}} Z + \epsilon_{ij}^{\bar{d}},\end{aligned} \tag{2.6}$$

$$\begin{aligned}Y_{ij}^{d} &= \beta_{0i}^{d} + \beta_{1i}^{d} t_{ij} + \beta_{2i}^{d} Z + \beta_{3i}^{d} Z t_{ij} + \epsilon_{ij}^{d} \\ Y_{ij}^{\bar{d}} &= \beta_{0i}^{\bar{d}} \qquad\quad + \beta_{2i}^{\bar{d}} Z \qquad\qquad + \epsilon_{ij}^{\bar{d}}.\end{aligned} \tag{2.7}$$

시간효과모형이 경우에 두 집단의 분포는 (2.8)과 같으며

$$\begin{aligned}Y_{ij}^{d} &\sim N(\beta_{0}^{d} + \beta_{1}^{d} t_{ij} + \beta_{2}^{d} + \beta_{3}^{d} t_{ij}, \sigma_{d}^{2}) \\ Y_{ij}^{\bar{d}} &\sim N(\beta_{0}^{\bar{d}} + \beta_{2}^{\bar{d}}, \sigma_{\bar{d}}^{2}).\end{aligned} \tag{2.8}$$

여기에서 AUC는 식 (2.9)를 이용하여 구한다.

$$a(t) = \frac{\beta_{0}^{\bar{d}} + \beta_{2}^{\bar{d}} - \beta_{0}^{d} - \beta_{1}^{d} t - \beta_{2}^{d} - \beta_{3}^{d} t}{\sqrt{\sigma_{d}^{2}}}, \quad b(t) = \sqrt{\frac{\sigma_{\bar{d}}^{2}}{\sigma_{d}^{2}}}. \tag{2.9}$$

임의효과모형의 분포는 (2.10)과 (2.11) 같이 표현되고

$$\begin{pmatrix}\beta_{0i}^d\\ \beta_{1i}^d\\ \beta_{2i}^d\\ \beta_{3i}^d\end{pmatrix} \sim N\left[\begin{pmatrix}\beta_0^d\\ \beta_1^d\\ \beta_2^d\\ \beta_3^d\end{pmatrix}, \begin{pmatrix}V^d & \underline{0}\\ \underline{0} & W^d\end{pmatrix}\right], \epsilon_{ij}^d \sim N(0,\sigma_d^2) \tag{2.10}$$

$$\begin{pmatrix}\beta_{0i}^{\bar{d}}\\ \beta_{2i}^{\bar{d}}\end{pmatrix} \sim N\left[\begin{pmatrix}\beta_0^{\bar{d}}\\ \beta_2^{\bar{d}}\end{pmatrix}, V^{\bar{d}}\right], \qquad \epsilon_{ij}^{\bar{d}} \sim N(0,\sigma_{\bar{d}}^2). \tag{2.11}$$

임의효과모형의 AUC는 (2.12)를 이용하여 구한다.

$$a(t)=\frac{\beta_0^{\bar{d}}+\beta_2^{\bar{d}}-\beta_0^d-\beta_1^d t-\beta_2^d-\beta_3^d t}{\sqrt{\sigma_d^2+(1\ t\ z\ tz)\begin{pmatrix}V^d & \underline{0}\\ \underline{0} & W^d\end{pmatrix}(1\ t\ z\ tz)'}},\ b(t)=\sqrt{\frac{\sigma_{\bar{d}}^2+(1\ z)V^{\bar{d}}(1\ z)'}{\sigma_d^2+(1\ t\ z\ tz)\begin{pmatrix}V^d & \underline{0}\\ \underline{0} & W^d\end{pmatrix}(1\ t\ z\ tz)'}}. \tag{2.12}$$

3. 시뮬레이션

2절에서 논의한 시간효과모형과 임의효과모형에 대하여 ROC 곡선과 AUC 통계량을 비교해본다. 시간효과모형의 Y^d분산은 $Var(Y_t^d)=\sigma_d^2$이며, 임의효과모형의 분산은 $Var(Y_r^d)=\sigma_d^2+(1\ t)V^d(1\ t)'$이다. 따라서 $t=0$인 경우 임의효과모형의 분산 $(\sigma_d^2+Var(\beta_0^d))$이 항상 시간효과모형의 분산 (σ_d^2)보다 크다. 하지만 시간이 지날수록 공분산의 값에 따라 임의효과모형의 분산이 증가 또는 감소할 수도 있는 경우가 발생한다. 따라서 시간이 지날수록 다음과 같은 세 가지 경우로 나누어 비교한다.

$$Case1: Var(Y_r^d) > Var(Y_t^d)\text{이며, 양의 공분산에 의해 } Var(Y_r^d)\text{이 증가}$$
$$Case2: Var(Y_r^d) > Var(Y_t^d)\text{이며, 음의 공분산에 의해 } Var(Y_r^d)\text{이 감소}$$
$$Case3: Var(Y_r^d) < Var(Y_t^d)\text{이며, 음의 공분산에 의해 } Var(Y_r^d)\text{이 감소}$$

우선, Case 1에 적합한 모수를 다음과 같이 설정하고,

$$\beta_0^d=4,\ \beta_1^d=0.15,\ \sigma_d^2=3.8,\ V^d=\begin{pmatrix}1 & 0.05\\ 0.05 & 0.01\end{pmatrix},$$
$$\beta_0^{\bar{d}}=3,\ \sigma_{\bar{d}}^2=1.8,\ V^{\bar{d}}=0.7$$

Case 2와 Case 3에는 $V^d=\begin{pmatrix}1 & -0.1\\ -0.1 & 0.01\end{pmatrix}$와 $V^d=\begin{pmatrix}1 & -0.2\\ -0.2 & 0.01\end{pmatrix}$로 설정하였고, 모수 추정을 통해 구한 t 시점에 대한 ROC곡선과 AUC는 Table 3.1과 같다. 시간효과모형에서는 식 (2.2)에서 $a(t)$와 $b(t)$식에 사용되지 않기 때문에 $V^{\bar{d}}$ 값은 생략한다.

Table 3.1을 살펴보면 임의효과모형의 경우 우선 두 모수의 공분산의 값이 양수인 경우 (Case1) 시간이 지날수록 임의효과모형의 분산인 $Var(Y_r^d)$가 증가하게 된다. 이런 경우 시간이 지날수록 $a(t)$의 감소폭이 다른 경우보다 작으며 $b(t)$의 값 또한 다른 경우와 다르게 감소하는 것을 볼 수 있다. 또한 AUC 통계량 또한 다른 두 개의 경우보다 증가의 폭이 작음을 파악할 수 있다. 반대로 공분산의 값이 작아질수록 시간이 지남에 따라 $a(t)$의 값의 감소폭이 커지며 $b(t)$의 값이 증가하는 것을 볼 수 있고, 또한 AUC 통계량의 증가폭이 커짐을 볼 수 있다. 시간효과모형의 경우 분산의 값의 변화가 없기 때문에 $a(t)$값만 작아지며 $b(t)$의 값은 고정인 것을 탐색할 수 있다.

Table 3.1 ROC curve and AUC

	Random-effect		Time-effect	
Case 1	ROC	AUC	ROC	AUC
$t=0$	$S(-0.456+0.722\ S^{-1}(u))$	0.6444	$S(-0.598+0.802\ S^{-1}(u))$	0.6795
$t=2$	$S(-0.579+0.704\ S^{-1}(u))$	0.6821	$S(-0.777+0.802\ S^{-1}(u))$	0.7278
$t=4$	$S(-0.691+0.683\ S^{-1}(u))$	0.7159	$S(-0.956+0.802\ S^{-1}(u))$	0.7722
$t=8$	$S(-0.882+0.634\ S^{-1}(u))$	0.7719	$S(-1.315+0.802\ S^{-1}(u))$	0.8475
Case 2	ROC	AUC	ROC	AUC
$t=0$	$S(-0.456+0.722\ S^{-1}(u))$	0.6444	$S(-0.598+0.802\ S^{-1}(u))$	0.6795
$t=2$	$S(-0.617+0.750\ S^{-1}(u))$	0.6892	$S(-0.777+0.802\ S^{-1}(u))$	0.7278
$t=4$	$S(-0.784+0.775\ S^{-1}(u))$	0.7324	$S(-0.956+0.802\ S^{-1}(u))$	0.7722
$t=8$	$S(-1.123+0.807\ S^{-1}(u))$	0.8089	$S(-1.315+0.802\ S^{-1}(u))$	0.8475
Case 3	ROC	AUC	ROC	AUC
$t=0$	$S(-0.456+0.722\ S^{-1}(u))$	0.6444	$S(-0.598+0.802\ S^{-1}(u))$	0.6795
$t=2$	$S(-0.655+0.797\ S^{-1}(u))$	0.6958	$S(-0.777+0.802\ S^{-1}(u))$	0.7278
$t=4$	$S(-0.848+0.838\ S^{-1}(u))$	0.7421	$S(-0.956+0.802\ S^{-1}(u))$	0.7722
$t=8$	$S(-1.262+0.907\ S^{-1}(u))$	0.8250	$S(-1.315+0.802\ S^{-1}(u))$	0.8475

Table 3.1을 바탕으로 Case1, Case2, Case3에 대한 각 시점에 대한 AUC를 비교하기 위하여 Figure 3.1에 표현하였다. Figure 3.1을 살펴보면 Case 1에서 임의효과모형에서 공분산이 양의 값을 가지며, 의 분산이 커질수록 시간이 흐름에 따라 AUC의 증가량은 시간효과모형의 AUC통계량의 증가량 보다 작다. Case 2의 경우도 두 모형에서의 AUC의 증가량이 유사하다. 마지막으로 Case 3의 경우에 시간효과모형의 AUC의 증가량보다 임의효과모형의 증가량이 더 크다는 것을 파악할 수 있다.

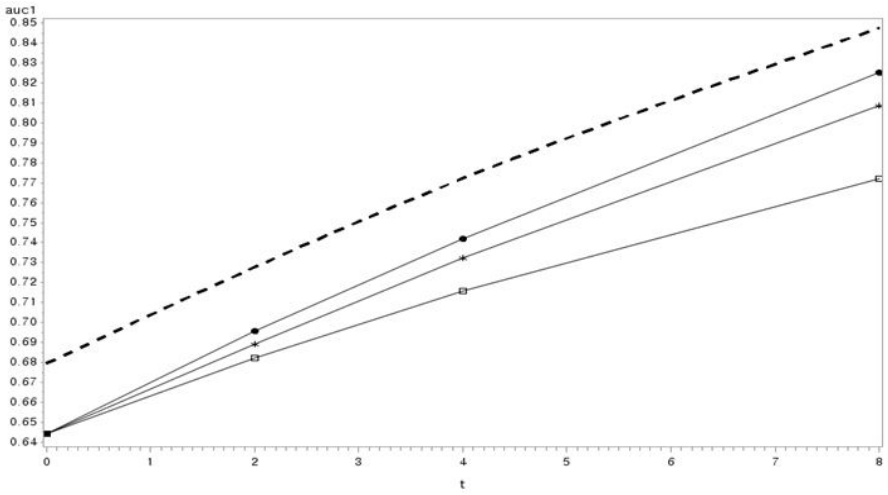

Figure 3.1 AUC for Case 1 to Case 3
(spline: time-effect model, square : Case 1, star : Case 2, dot : Case3)

다음으로는 2.2절에서 논의한 가변수가 추가된 모형 (2.6)과 (2.7)에서 시간효과모형과 임의효과모형에 대하여 살펴본다. 시간효과모형의 Y^d분산은 $Var(Y_t^d)=\sigma_d^2$로 동일하지만, 임의효과모형의 분산은 $Var(Y_r^d)=\sigma_d^2+(1\ t\ z\ tz)\begin{pmatrix} V^d & \underline{0} \\ \underline{0} & W^d \end{pmatrix}(1\ t\ z\ tz)'$이다. 여기서 가변수의 계수와 아닌 변수간의 상관관계는 없다고 가정하였다. 앞서 공분산을 통해 세 가지 경우로 나누어 비교한 경우와 유사하게 가변수간의

공분산의 계수를 적용하지 않았을 때와 하였을 때의 값이 양수인지 음수인지에 따라 변동을 탐색한다.

$$Case4 : Cov(\beta_0^d, \beta_1^d) > 0,\ Cov(\beta_2^d, \beta_3^d) = 0\text{인 경우}$$
$$Case5 : Cov(\beta_0^d, \beta_1^d) < 0,\ Cov(\beta_2^d, \beta_3^d) = 0\text{인 경우}$$
$$Case6 : Cov(\beta_0^d, \beta_1^d) > 0,\ Cov(\beta_2^d, \beta_3^d) > 0\text{인 경우}$$
$$Case7 : Cov(\beta_0^d, \beta_1^d) > 0,\ Cov(\beta_2^d, \beta_3^d) < 0\text{인 경우}$$
$$Case8 : Cov(\beta_0^d, \beta_1^d) < 0,\ Cov(\beta_2^d, \beta_3^d) < 0\text{인 경우}$$

우선, Case 4부터 Case 8까지를 만족하기 위화여 모수를 다음과 같이 설정한다.

$$\text{Case 4 : } \beta_0^d = 5,\ \beta_1^d = 0.15,\ \beta_2^d = 3,\ \beta_3^d = 0.1,\ \sigma_d^2 = 3.8,\ \begin{pmatrix} V^d & \underline{0} \\ \underline{0} & W^d \end{pmatrix} = \begin{pmatrix} 1 & 0.05 & 0 & 0 \\ 0.05 & 0.01 & 0 & 0 \\ 0 & 0 & 0.01 & 0 \\ 0 & 0 & 0 & 0.01 \end{pmatrix}$$

$$\beta_0^{\bar{d}} = 5,\ \beta_1^{\bar{d}} = 2,\ \sigma_{\bar{d}}^2 = 1.8,\ V^{\bar{d}} = \begin{pmatrix} 1.7 & 0.1 \\ 0.1 & 0.3 \end{pmatrix}$$

$$\text{Case 5 : } \begin{pmatrix} V^d & \underline{0} \\ \underline{0} & W^d \end{pmatrix} = \begin{pmatrix} 1 & -0.1 & 0 & 0 \\ -0.1 & 0.01 & 0 & 0 \\ 0 & 0 & 0.01 & 0 \\ 0 & 0 & 0 & 0.01 \end{pmatrix},\quad \text{Case 6 : } \begin{pmatrix} V^d & \underline{0} \\ \underline{0} & W^d \end{pmatrix} = \begin{pmatrix} 1 & 0.05 & 0 & 0 \\ 0.05 & 0.01 & 0 & 0 \\ 0 & 0 & 0.01 & 0.1 \\ 0 & 0 & 0.1 & 0.01 \end{pmatrix}$$

$$\text{Case 7 : } \begin{pmatrix} V^d & \underline{0} \\ \underline{0} & W^d \end{pmatrix} = \begin{pmatrix} 1 & 0.05 & 0 & 0 \\ 0.05 & 0.01 & 0 & 0 \\ 0 & 0 & 0.01 & -0.1 \\ 0 & 0 & -0.1 & 0.01 \end{pmatrix},\quad \text{Case 8 : } \begin{pmatrix} V^d & \underline{0} \\ \underline{0} & W^d \end{pmatrix} = \begin{pmatrix} 1 & -0.05 & 0 & 0 \\ -0.05 & 0.01 & 0 & 0 \\ 0 & 0 & 0.01 & -0.1 \\ 0 & 0 & -0.1 & 0.01 \end{pmatrix},$$

Table 3.2 ROC curve and AUC

	Random-effect		Time-effect	
Case 4	ROC	AUC	ROC	AUC
$t=0$	$S(-0.999 + 0.999\ S^{-1}(u))$	0.7601	$S(-1.026 + 1.453\ S^{-1}(\mathrm{u}))$	0.7196
$t=2$	$S(-1.207 + 0.966\ S^{-1}(u))$	0.8074	$S(-1.282 + 1.453\ S^{-1}(\mathrm{u}))$	0.7664
$t=4$	$S(-1.379 + 0.920\ S^{-1}(u))$	0.8450	$S(-1.539 + 1.453\ S^{-1}(\mathrm{u}))$	0.8085
$t=8$	$S(-1.416 + 0.809\ S^{-1}(u))$	0.8645	$S(-1.795 + 1.453\ S^{-1}(\mathrm{u}))$	0.8456
Case 5	ROC	AUC	ROC	AUC
$t=0$	$S(-0.999 + 0.999\ S^{-1}(u))$	0.7601	$S(-1.026 + 1.453\ S^{-1}(\mathrm{u}))$	0.7196
$t=2$	$S(-1.309 + 1.047\ S^{-1}(u))$	0.8170	$S(-1.282 + 1.453\ S^{-1}(\mathrm{u}))$	0.7664
$t=4$	$S(-1.634 + 1.089\ S^{-1}(u))$	0.8654	$S(-1.539 + 1.453\ S^{-1}(\mathrm{u}))$	0.8085
$t=8$	$S(-1.941 + 1.109\ S^{-1}(u))$	0.9032	$S(-1.795 + 1.453\ S^{-1}(\mathrm{u}))$	0.8456
Case 6	ROC	AUC	ROC	AUC
$t=0$	$S(-0.999 + 0.999\ S^{-1}(u))$	0.7601	$S(-1.026 + 1.453\ S^{-1}(\mathrm{u}))$	0.7196
$t=2$	$S(-1.156 + 0.925\ S^{-1}(u))$	0.8019	$S(-1.282 + 1.453\ S^{-1}(\mathrm{u}))$	0.7664
$t=4$	$S(-1.276 + 0.850\ S^{-1}(u))$	0.8344	$S(-1.539 + 1.453\ S^{-1}(\mathrm{u}))$	0.8085
$t=8$	$S(-1.262 + 0.721\ S^{-1}(u))$	0.8470	$S(-1.795 + 1.453\ S^{-1}(\mathrm{u}))$	0.8456
Case 7	ROC	AUC	ROC	AUC
$t=0$	$S(-0.999 + 0.999\ S^{-1}(u))$	0.7601	$S(-1.026 + 1.453\ S^{-1}(\mathrm{u}))$	0.7196
$t=2$	$S(-1.268 + 1,914\ S^{-1}(u))$	0.8133	$S(-1.282 + 1.453\ S^{-1}(\mathrm{u}))$	0.7664
$t=4$	$S(-1.515 + 1.010\ S^{-1}(u))$	0.8568	$S(-1.539 + 1.453\ S^{-1}(\mathrm{u}))$	0.8085
$t=8$	$S(-1.652 + 0.944\ S^{-1}(u))$	0.8852	$S(-1.795 + 1.453\ S^{-1}(\mathrm{u}))$	0.8456
Case 8	ROC	AUC	ROC	AUC
$t=0$	$S(-0.999 + 0.999\ S^{-1}(u))$	0.7601	$S(-1.026 + 1.453\ S^{-1}(\mathrm{u}))$	0.7196
$t=2$	$S(-1.338 + 1.070\ S^{-1}(u))$	0.8195	$S(-1.282 + 1.453\ S^{-1}(\mathrm{u}))$	0.7664
$t=4$	$S(-1.700 + 1.130\ S^{-1}(u))$	0.8694	$S(-1.539 + 1.453\ S^{-1}(\mathrm{u}))$	0.8085
$t=8$	$S(-2.059 + 1.176\ S^{-1}(u))$	0.9088	$S(-1.795 + 1.453\ S^{-1}(\mathrm{u}))$	0.8456

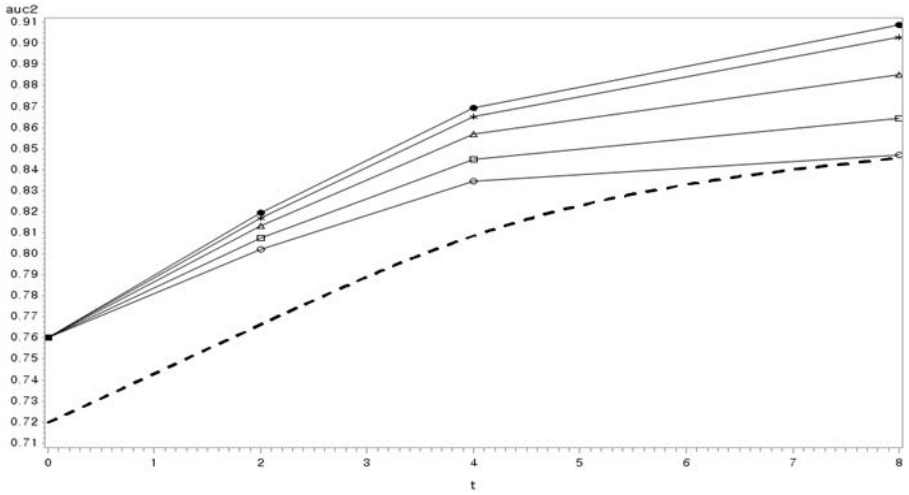

Figure 3.2 AUC for Case 4 to Case 8
(Spline: time-effect model, Triangle : Case 4, Star : Case5, Circle : Case 6, Square : Case 7, dot : Case8)

ROC 곡선과 AUC 통계량의 값을 구하여 Table 3.2에 정리하였다. 시간효과모형에서는 식 (2.9)에서 $a(t)$와 $b(t)$식에 사용되지 않았기 때문에 V^d와 W^d의 값은 생략한다.

Table 3.2를 살펴보면 Table 3.1와 같이 공분산의 값에 의해서 시간이 지날수록 임의효과모형의 분산인 $Var(Y_r^d)$ 의 값이 커질 경우 (Case 6)는 다른 경우와 비교했을 때 $a(t)$의 값의 감소폭이 작고 $b(t)$의 값 또한 다른 경우와 다르게 감소한다. AUC 통계량 또한 다른 경우와 다르게 증가하는 폭이 작아짐을 파악할 수 있다. 이와 반대로 음의 공분산에 의해 $Var(Y_r^d)$의 값이 작아지는 경우 (Case 8)에는 시간이 지날수록 $a(t)$의 값의 감소폭이 크며 $b(t)$값 또한 커지는 것을 볼 수 있다. AUC 통계량 또한 다른 경우보다 증가의 폭이 가장 크다는 것을 발견할 수 있다.

다음으로는 Case 4부터 Case 8의 경우에 AUC를 Figure 3.2에 구현하였다. Figure 3.2를 통하여 임의효과모형에서 Y^d의 분산이 양의 공분산으로 인해 커지고 시간이 경과할수록 임의효과모형의 AUC의 증가량이 감소하며, 음의 공분산으로 인해 임의효과모형에서 Y^d의 분산이 작아지면 임의효과모형에서의 AUC의 증가량이 증가한다. 즉 공분산이 어떤 값을 가지고 있는지 여부에 따라 시간이 경과할수록 AUC의 증가폭에 차이가 있다는 것을 탐색할 수 있다.

이 같은 결과를 이용해, 다양한 분야에서 시간을 두고 두 집단을 비교, 연구하는데 있어 모형설정에 도움이 되리라 생각한다. 본 연구에서는 정규분포를 가정하고 가변수의 계수와 아닌 계수의 상관관계를 고려하지 않았지만 향후 더 연구를 하는데 있어 기초를 마련한 것으로 생각한다.

4. 결론

최근에 다양한 분야에서 시간의 흐름에 따른 변화를 관찰하려는 연구가 활발하여 다양한 시간에서 수집한 경시적자료의 분석으로 선형성장모형을 사용하였다. Etzioni 등 (1999)은 시간효과를 고려하는 전통적인 선형성장모형과 모형의 계수에 임의효과를 추가하는 모형을 고려하고, 모형의 판별력을 측정하기 위하여 ROC 곡선과 AUC 통계량으로 두 종류의 모형을 비교하였다. 실증예제 분석에 중점을 둔 Etzioni 등 (1999)의 연구를 확장하여 본 연구는 정규분포로 가정한 다양한 자료를 모의실험에 이용하

여 생성하고 다양한 선형성장모형에 대하여 여러 시점에서의 ROC 곡선과 AUC 통계량을 구하여 비교 분석하였다. 2.1절에서 설명한 것처럼 모형을 선택할때의 기준은 시간불변의 개별효과와 독립변수간의 관계가 중요하다. 관계가 있거나 모르는 경우에는 선형성장모형과 가변수의 모수를 추가한 시간효과모형이 선호되며, 관계가 없는 경우에는 모수추정의 정확성이 더 높은 임의효과모형이 선호될 것이다.

본 연구를 통하여 공분산의 크기에 따라 ROC 곡선과 AUC 값에 차이를 보였다. 임의효과모형에서 공분산이 양인 경우에 시간이 지날수록 임의효과모형의 분산이 증가하며 AUC 통계량은 0.6444에서 0.7719까지 증가하였다. 이는 0.6795에서 0.8475까지 증가한 시간효과모형의 AUC 통계량의 증가량보다 작다. 반대로 공분산이 작아질수록 임의효과모형에서 시간이 지남에 따라 AUC 통계량은 0.6444에서 0.8250까지 증가하는 것을 볼 수 있고, 0.6795에서 0.8475까지 증가한 시간효과모형의 AUC 통계량보다 AUC 통계량의 증가폭이 더 크다는 것을 탐색하였다.

가변수의 모수를 추가한 모형에서 임의효과모형의 분산이 양의 공분산으로 인해 커지고 시간이 경과할수록 AUC 통계량은 0.7601에서 0.8470으로 증가한다. 시간효과모형의 AUC 통계량은 0.7196에서 0.8456까지 증가하며 이 경우에는 임의효과모형의 AUC 통계량의 증가폭이 시간효과모형의 증가폭 보다 작은 것을 확인하였다. 그리고 음의 공분산으로 인해 임의효과모형의 분산이 작아지면 AUC 통계량은 0.7601에서 0.9088까지 증가하며 시간효과모형의 AUC 통계량은 0.7196에서 0.8456까지 증가하는 것을 알 수 있고 임의효과모형의 AUC 통계량의 증가폭이 시간효과모형의 증가폭보다 크다는 것을 탐색하였다. 그러므로 공분산의 값에 따라 시간이 경과할수록 AUC 통계량의 증가폭에 차이가 있다는 것을 탐색하였다.

References

Ashenfelter, O., Levine, B. P. and Zimmerman, J. D. (2003). *Statistics and econometrics: Methods and applications*, John Wiley & Sons, New Jersey.

Bloom, L., Persson, L. A. and Dahlquist, G. (1992). A high linear growth is associated with an increased risk of childhood diabetes mellitus. *Diabetologia*, **35**, 528-533.

Etzioni, R., Pepe, M. S., Longton, G., Hu, C. and Goodman, G. (1999). Incorporating the time dimension in receiver operating characteristic curve: A case study of prostate cancer. *Medical Decision Making*, **19**, 242.

Hong, C. S., Kim, G. C. and Jeong, J. A. (2012). Bivariate ROC curve. *Communications of The Korean Statistical Society*, **19**, 277-286.

Hong, C. S., Jung, E. S. and Jung, D. G. (2013). Standard criterion of VUS for ROC surface. *The Korean Journal of Applied Statistics*, **26**, 977-985.

Hong, C. S. and Jung, D. G. (2014). Standard criterion of hypervolume under the ROC manifold. *Journal of the Korean Data & Information Science Society*, **25**, 473-483.

Jeon, J. Y. and Lee, K. B. (2014). Review and discussion of marginalized random effects models. *Journal of the Korean Data & Information Science Society*, **25**, 1263-1272.

Johnston, J. and Dinardo, J. (2000). Econometric methods. *Econometric Theory*, **16**, 139-142

Pearson, J. D., Luderer, A. A., Metter, E. J., Partin, A. W., Chan, D. W., Fozard, J. L. and Carter, H. B. (1996). Longitudinal analysis of series measurements of free and total psa among men with and without prostatic cancer. *Elsevier Science*, **48**, 4-9.

Pepe, M. S. (1997). A regression modelling framework for receiver operating characteristic curves in medical diagnositic testing. *Biometrika*, **84**, 595-608.

Pepe, M. S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics*, **54**, 124-135.

Pepe, M. S. (2000). An interpretation for the ROC curve and inference using glm procedure. *Biometrics*, **56**, 352-359.

Park, H. J., Sang, K. A. and Kang, J. Y. (2008). Effect of private tutoring on middle school students' achievement. *Journal of Educational Evaluation*, **21**, 107-127.

Tosteson, A. N. and Begg, C. B. (1998). A general regression methodology for ROC curve estimation. *Medical Decision Making*, **8**, 204.

Journal of the Korean Data & Information Science Society
2015, **26**(6), 1367–1375

http://dx.doi.org/10.7465/jkdi.2015.26.6.1367
한국데이터정보과학회지

ROC curve and AUC for linear growth models

Chong Sun Hong[1] · Dae Soon Yang[2]

[12]Department of Statistics, Sungkyunkwan University

Received 10 September 2015, revised 16 October 2015, accepted 9 November 2015

Abstract

Consider the linear growth models for longitudinal data analysis. Several kind of linear growth models are selected such as time-effect and random-effect models as well as a dummy variable included model. In this work, simulation data are generated with normality assumption, and both binormal ROC curve and AUC are obtained and compared for various linear growth models. It is found that ROC curves have different shapes and AUC increase slowly, as values of the covariance increase and the time passes for random-effect models. On the other hand, AUC increases very fast as values of covariance decrease. When the covariance has positive value, we explored that the variances of random-effect models increase and the increment of AUC is smaller than that of AUC for time-effect models. And the increment of AUC for time-effect models is larger than the increment for random-effect models.

Keywords: Dummy variable, longitudinal data, random-effect, time-effect.

[1] Corresponding author: Professor, Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea. E-mail: cshong@skku.edu
[2] Graduate student, Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea.