한국도로학회 논문집
제14권 제5호 2012년 10월
pp. 189 ~ 199

교통사고모형 개발에서의 함수식 도출 방법론에 관한 연구

Methodology for Determining Functional Forms in Developing Statistical Collision Models

백 종 대 Baek, Jong Dae | 정회원 · 한국건설기술연구원 도로연구실 수석연구원 (E-mail : jdbaek@kict.re.kr)
험머 조셉 Hummer, Joseph | Wayne State Univ., Civil and Environmental Engineering, Chair (E-mail : joseph.hummer@wayne.edu)

ABSTRACT

PURPOSES : The purpose of this study is to propose a new methodology for developing statistical collision models and to show the validation results of the methodology.

METHODS : A new modeling method of introducing variables into the model one by one in a multiplicative form is suggested. A method for choosing explanatory variables to be introduced into the model is explained. A method for determining functional forms for each explanatory variable is introduced as well as a parameter estimating procedure. A model selection method is also dealt with. Finally, the validation results is provided to demonstrate the efficacy of the final models developed using the method suggested in this study.

RESULTS : According to the results of the validation for the total and injury collisions, the predictive powers of the models developed using the method suggested in this study were better than those of generalized linear models for the same data.

CONCLUSIONS : Using the methodology suggested in this study, we could develop better statistical collision models having better predictive powers. This was because the methodology enabled us to find the relationships between dependant variable and each explanatory variable individually and to find the functional forms for the relationships which can be more likely non-linear.

Keywords
collision model, multivariate, non-linear, functional form, model selection, parameter estimation, model expansion

Main Author : Baek, Jong Dae, Senior Researcher
Highway Research Division, SOC Research Institute,
Korea Institute of Construction Technology, 283 Goyangdae-Ro,
Ilsanseo-Gu, Goyang-Si, Gyeonggi-Do, 411-712, Korea
Tel : +82.31.910.0754 Fax : +82.31.910.0161
email : jdbaek@kict.re.kr

International Journal of Highway Engineering
http://www.ksre.or.kr/
ISSN 1738-7159 (Print)
ISSN 2287-3678 (Online)

1. 서론

1.1. 연구의 배경

어떤 물질이나 현상들의 상태나 원인 등을 규명하기 위해 가장 널리 사용되는 방법은 실험이나 관찰이다. 도로에서 일어나는 교통사고들은 사고가 일어날 당시의 도로 요소, 차량 요소, 환경 요소(날씨나 조명 등), 교통류 요소, 인적 요소 등 많은 인자들이 서로 복잡한 상호작용들을 일으킨 결과라고 볼 수 있다. 그러나 이들 인자들 중 인적 요소나 날씨 요소 등과 같은 몇몇 인자들은 이들 인자들과 교통사고 간의 실제적인 관련성을 규명하기 위해 그 정도나 상태 등을 측정하기가 어렵고, 또한 그런 실험조건들을 원하는 정도나 상황으로 임의로 선택하거나 조절하여 실험을 수행하기가 대단히 어렵다. 가상 도로주행 시뮬레이터를 이용하여 어느 정도의 실험조건을 임의 선택 및 조절한 연구들이 수행되어 왔지만 아직 현실성 표현에 적지 않은 한계를 지니고 있

어 그 실험결과의 신빙성이 인정받기에는 이른 단계라는 인식이 지배적이다.

지금까지 교통 안전성과 그에 영향을 미치는 여러 요인들 간의 관계를 규명하기 위해 가장 보편적으로 사용되어지며 유일하게 현실적인 방법은 과거자료에 기반을 둔 통계적 모형의 개발이다. 다시 말해 통계적 교통사고 모형의 개발은 어떤 인자들이 얼마나 교통사고 발생에 영향을 미치는지를 관련 과거 자료와 통계적 분석 기법을 이용하여 밝혀내려는 대안적인 노력이라 볼 수 있다. 이 방법은 여러 인자들이 갖고 있는 교통 안전성에 대한 영향이 과거 자료에 반영되어 있다는 기본적인 가정을 토대로 한다.

통계적 교통사고모형의 개발에 있어 개발된 모형의 성능을 좌우하는 두 가지 핵심은 자료의 건실성과 분석 방법의 적절성이다. 같은 자료에 대해서 어떤 분석 방법, 즉 어떤 모형식을 사용하느냐에 따라 그 결과는 충분히 달라질 수 있다. 보다 극적으로 표현하자면 적절치 못한 방법을 이용하게 되면 도로의 안전성에 미칠 수 있는 여러 도로설계 요인들의 영향을 충분히 발견치 못하거나 심지어 잘 못 해석할 수도 있다. 설명력 높은 교통사고모형의 개발을 위한 적절한 방법론에 대한 연구가 끊임없이 이뤄져야 하는 이유가 여기에 있다.

개발하고자 하는 적절한 모형의 종류를 선택하기 위해서는 그 모형의 기반이 되는 자료의 특성을 잘 살펴야 한다. 일반적으로 교통사고모형의 대상이 되는 자료의 가장 두드러지는 특성들은 다음과 같다.

- 하나의 종속변수와 여러 개의 독립변수들을 갖는다.
- 종속변수는 항상 음수가 아닌 값이며 음이항 분포를 따른다고 가정된다.
- 독립변수들은 이항변수와 연속변수가 모두 포함될 수 있다.
- 종속변수와 각각의 독립변수간의 관계는 선형이나 비선형 모두가 될 수 있다.

이들 중 앞의 두 특성들을 고려할 수 있는 모형 개발 방법에는 일반선형모형(Generalized Linear Model, GLM) 방법이 있다. 이들 GLM 모형들 중 다변수(multivariate) 자료를 위해서는 로그선형 회귀모형(Loglinear Regression Model)이 가장 널리 사용되고 있다.

그러나 GLM 모형들은 도로상의 교통사고를 모형화하는데 한계가 있다. GLM 모형을 이용하면 종속변수와 각각의 독립변수들 간에 존재할 수 있는 다양한 형태의 비선형 관계들을 모형화할 수 없고 단지 로그선형 관계만을 다룰 수 있을 뿐이다. 또 하나의 한계는 GLM 모형으로는 종속변수와 개개의 독립변수들 간에 있을 수 있는 다양한 형태의 관계들(선형이든 비선형이든)을 따로 모형화할 수 없다는 것이다. 각각의 독립변수들이 각자 자기만의 종속변수와의 관계를 어떤 형태로든 독립적으로 가질 수 있다는 사실은 쉽게 받아들일 수 있다는 사실이다.

이러한 개개의 독립변수들과 종속변수가 가질 수 있는 관계의 개별성을 다루기 위해 Hauer(2004)는 각각의 독립변수들을 전체 모형에 하나씩 차례로 그들 각자에 맞는 함수식을 이용하여 추가해 나가며 모형을 개발해 나가는 방법을 제안하였다.

본 연구에서는 이 방법을 근간으로 하여 모형식과 각 독립변수의 함수식 선정방법, 계수값 추정 방법, 모형 선택 방법, 모형 확장 기법을 새로 개발 적용하여 새로운 통계적 교통사고모형 구축 방법을 제시하고자 하였다.

1.2. 연구목적

이 논문의 목적은 보다 설명력 높은 통계적 교통사고 모형을 개발하기 위한 새로운 방법론을 제시하는 것이다. 이를 위해 교통사고에 영향을 미친다고 생각되는 독립변수들을 선정하는 방법론과 이들 독립변수들이 각각 종속변수에 대해 갖는 개개의 관계를 표현하는 함수식의 도출방법론, 계수값의 추정 방법론, 모형의 선택 방법론, 그리고 모형의 확장 방법론을 개발하여 제시하고, 이 새로운 방법론을 검증하기 위해 실제 자료를 이용하여 적용한 결과를 제시하고자 하였다.

2. 선행연구 고찰

2.1. 로그선형 회귀(loglinear regression) 모형

통계적 교통사고모형의 종속변수가 되는 교통사고건수의 가장 두드러지는 특징 중 하나는 항상 음의 값을 갖지 않는다는 것이다. 이런 이유로 지금까지 가장 일반적으로 사용되어지고 있는 통계적 사고모형 개발 방법은 전술한 바와 같이 로그선형 회귀식을 이용하는 것이다. 즉, 교통사고 건수가 포아송(Poisson)이나 음이항(negative binomial) 분포를 따른다고 가정한 로그선형 회귀식 기법이 교통사고 자료 분석에 적합하다고 알려져 있다 (Vogt and Bared, 1998). 보편적으로 널리

사용되고 있는 로그선형 회귀모형 기법에 대해서는 이 논문에서 기술하지 않아도 될 것이다.

2.2. 비선형(nonlinear) 모형 기법

Hauer는 교통사고와 연관된 거의 모든 현상들은 교통사고와 비선형의 관계를 보인다는 사실에 입각하여 교통사고모형식에 포함되는 각 독립변수들과 종속변수 간의 함수식을 따로 찾아야 된다고 제안하였다. 그는 각 독립변수들에 대한 함수식들이 곱셈 또는 덧셈으로 연결되는 다음과 같은 모형식을 제안하였다:

$$y = (\text{스케일 계수}) \times [(\text{도로구간 길이}) \times (\text{곱셈 부분}) + (\text{덧셈 부분})] \quad (1)$$

여기서, (곱셈 부분) = f(AADT)×g(길어깨 폭)×…

(덧셈 부분) = h(접근로 수)+i(교량 수) +…

Eq. (1)에서, y는 방향별 도로구간에서 발생하는 연간 교통사고 건수의 기대값이며, f( ), g( ), h( ), i( ) 는 각각 독립변수들에 해당하는 함수식을 나타낸다. Hauer는 모형식에 독립변수를 하나씩 추가해 가면서 그 독립변수에 가장 적합한 함수식을 찾고, 새로이 독립변수가 추가될 때마다 모형에 속해 있는 모든 계수값들을 다시 추정하는 방식을 제안하였다. 이 방식은 교통사고건수가 음이항분포를 따른다고 가정하는 것에는 음이항 모형과 동일하지만 음이항 모형식과 같이 모든 변수들을 하나의 함수식에 동시에 넣어 계수값들을 추정하는 것이 아니라는 점에서 차이가 있다.

모형에 새로 추가할 하나의 독립변수에 대한 함수식을 찾기 위해 Eq. (2)에서와 같이 실제 관측된 교통사고건수와 추가할 변수가 아직 추가되지 않은 모형을 이용해 예측한 교통사고 건수 간의 관계를 이용하였다.

$$R(\text{추가변수}) = \frac{(\text{실제 관측된 교통사고건수})}{(\text{변수추가 전 모형으로 예측된 교통사고건수})} \quad (2)$$

즉, 추가할 변수의 변수값들을 어느 정도 일정한 값들로 구분하여 몇 개의 단계로 나누고, 전체 자료를 이렇게 구분된 각 단계에 따라 구분한 다음, 각 단계별로 Eq. (2)를 이용하여 R값을 구한다. 이들 R값들의 추세를 살펴 추가하려 하는 변수를 모형에 추가할지의 여부와 추가할 시 취할 함수식을 결정한다. 각 변수의 함수식은 그 함수식이 가질 수 있는 포물선 특성이나 변곡특성 등을 반영할 수 있도록 선택될 수 있다. 이렇게 함으로써 새 변수 추가 전 모형이 실제 교통사고 건수를 잘 설명하지 못하고 있는 부분을 보완할 수 있는 새로운 함수식을 모형에 추가하여 개선된 모형을 만들 수 있다.

3. 교통사고모형 구축 방법론

3.1. 교통사고모형 개발 절차

통계적 교통사고모형의 개발은 적합한 모형개발기법과 자료를 잘 설명해 줄 수 있는 함수식(functional form)을 찾는 것, 그리고 그 함수식에 포함된 각 독립변수들의 계수값 추정이라 할 수 있다. Fig. 1은 이 연구에서 제시하는 전반적인 교통사고모형 개발 절차를 나타내고 있다. 전체적인 흐름은 하나의 독립변수를 이용한 초기 모형을 개발하고 이를 바탕으로 다른 독립변수들을 하나씩 추가하며 모형식을 확장해 나가는 것이다. 하나의 독립변수를 추가할 때마다 그 변수에 적합할 것으로 판단되는 후보 함수식들을 이용해 모형들을 만들고 계수값들을 추정한 다음 그들 중 가장 적합한 모형식 하나를 선택하게 되는 과정을 되풀이한다.

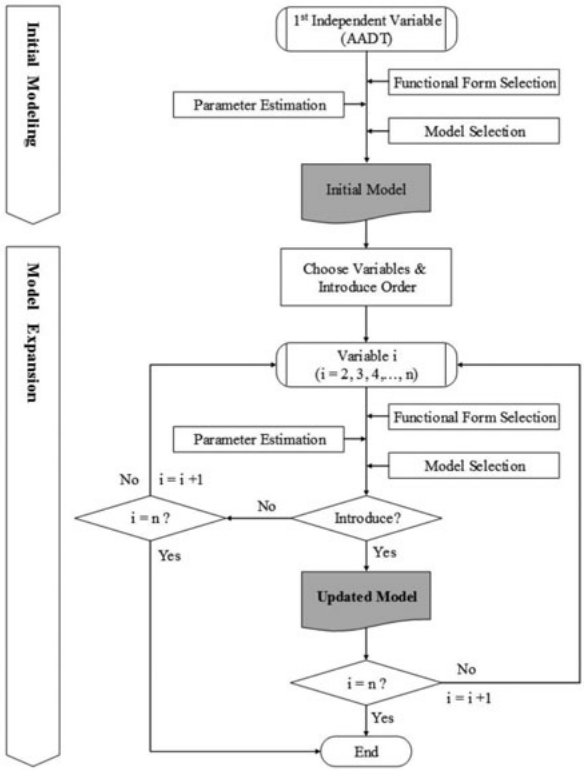


Fig. 1 Overview of Collision Modeling Methodology

이 연구에서 제시하는 전체 모형식은 다음과 같은 곱셈식 형태이다.

$$y = f(\text{교통량}) \times g(\text{길어깨 유형}) \times h(\text{접근로 밀도}) \times \cdots \quad (3)$$

Eq. (3)에서 y는 방향별 도로구간에서 단위길이당 발생할 것으로 기대되는 연간 교통사고 건수의 예측값이다. 교통사고모형은 대상 도로구간이 중앙분리대로 분리되어 방향별로 접근로 밀도 같은 기하구조적인 특성이 다를 경우 방향별로 별도로 다루는 것이 좋다. Eq. (3)에서 f( ), g( ), h( ) 등은 괄호 안에 있는 변수들을 위한 함수식이며 각각 다른 형태를 가질 수 있으나 모두 곱셈으로 연결된다. 각 함수식들이 곱셈으로 연결되는 이유는 이 변수들의 특성이 대상 도로구간에 대해 일정하기 때문이다. 하나의 접근로는 대상 도로구간에 대해서 연속적인 특성이 아니라 어느 지점에만 접속되어 있기 때문에 지점 요인이지만, 접근로 수가 접근로 밀도(단위 길이당 접근로 수)로 계산되어 표현된다면 이는 대상 구간에 대해 동일한 값을 갖게 되는 연속적 요인이 된다. 따라서 접근로수를 고려하는 변수에 대한 함수식도 곱셈으로 연결될 수 있다.

Fig. 1에서 볼 수 있는 바와 같이, 교통량을 나타내는 변수는 항상 모형식에 가장 먼저 투입되어 초기 모형을 구성하는데, 이는 전체 모형에 고려되는 기타 다른 독립변수들의 값들을 결정하는데 근간이 되는 주요 독립변수이기 때문이다(Hauer et al., 2004). 초기 모형을 개발하기 위해 교통량 변수에 대한 절절한 함수식을 찾아야 한다.

3.2. 후보 함수식 선정

앞서 기술한 바와 같이 새로운 독립변수가 모형에 추가될 때마다 이 변수와 종속변수간의 모든 가능한 관계형태, 즉 선형이나 비선형 (볼록 포물선, 오목 포물선, 또는 변곡점이 있는 곡선 등) 관계를 잘 나타낼 수 있는 적절한 함수식을 찾아내는 단계를 거쳐야 한다. 이렇게 함으로써 모든 독립변수들에 대한 함수식은 GLM 방식에서처럼 획일화될 필요가 없어진다.

Hauer(1997)가 제시한 Integrate-Differentiate (ID) 방법을 사용하여 각 독립변수의 함수 후보식을 선택할 수 있다. 이 방법은 대상 독립변수에 대한 실증적분함수(Empirical Integral Function, EIF)를 만들어 이의 함수식을 찾고 이를 다시 미분하여 원래의 함수식을 찾는 것이다. 예로 Fig. 2는 교통량(AADT)과 교통사고건수와의 관계에 대한 한 예를 그래프로 나타낸 것이다. 이 경우 그래프만 보고 이 관계가 선형인지 비선형인지 구별해 내기가 쉽지 않다.

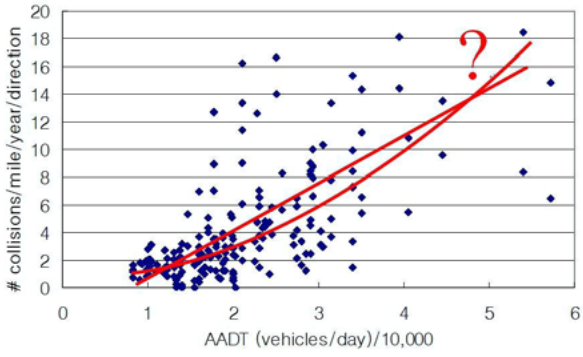


Fig. 2 A Scatter Plot Depicting the Relationship between Number of Collisions and AADT

실증적분함수를 만들기 위해서는 먼저 자료를 대상 독립변수에 따라 정렬하고 각 도로구간마다 bin폭과 bin 높이를 결정한다. Fig. 3은 대상 독립변수가 교통량(AADT/10,000)일 경우에 대해 실증적분함수를 만드는 개념을 나타낸 것이다. Fig. 3에서 볼 수 있듯이 i번째 bin의 폭을 결정하기 위한 i번째 bin의 왼쪽 경계값은 (i-1)번째 bin의 변수값까지 거리의 절반인 지점이며 오른쪽 경계값도 같은 방법으로 (i+1)번째 bin의 변수값까지 거리의 절반인 지점이 된다. bin 높이는 그 bin에 속한 모든 도로구간들이 갖는 교통사고건수의 평균값이다. 따라서 i번째 bin의 오른쪽 경계에서의 실증적분함수값은 그 경계에서의 누적 bin 면적값, 즉 처음부터 i번째 bin 면적까지의 총 합이 된다.

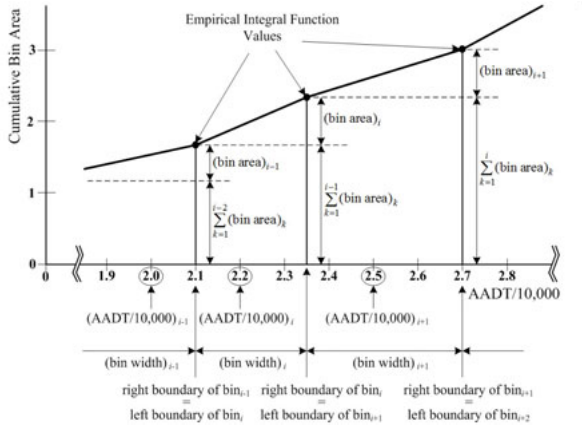


Fig. 3 Conceptual Diagram of Empirical Integral Function(EIF)

Fig. 4는 Fig. 2에 나타난 자료를 가지고 만든 실증적분함수를 나타내고 있다. Fig. 4에서는 어떤 형태의 패턴을 발견해 낼 수 있다.

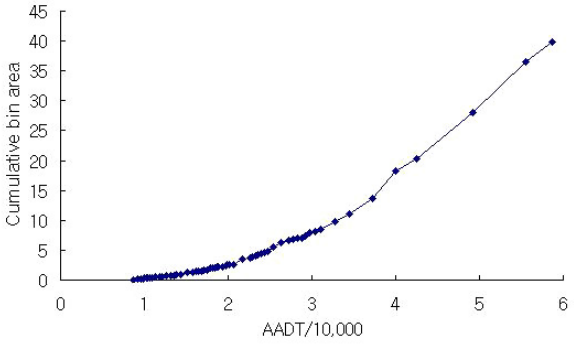
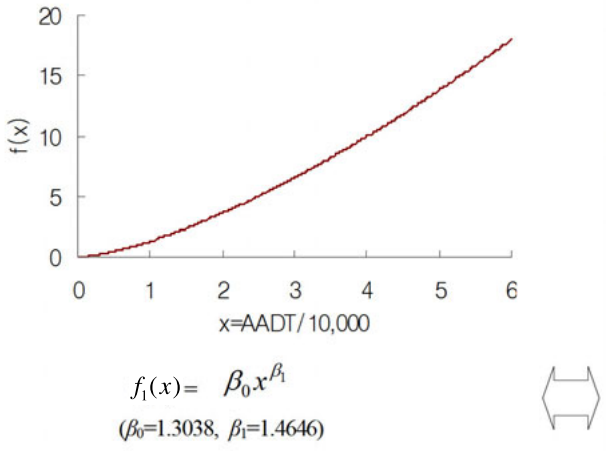


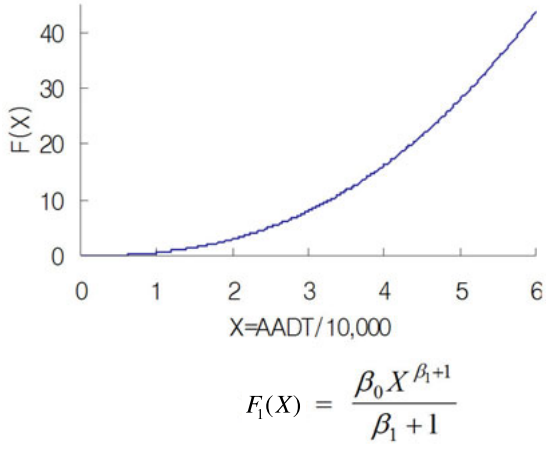
Fig. 4 The Empirical Integral Function(EIF) for the Data in Fig. 2

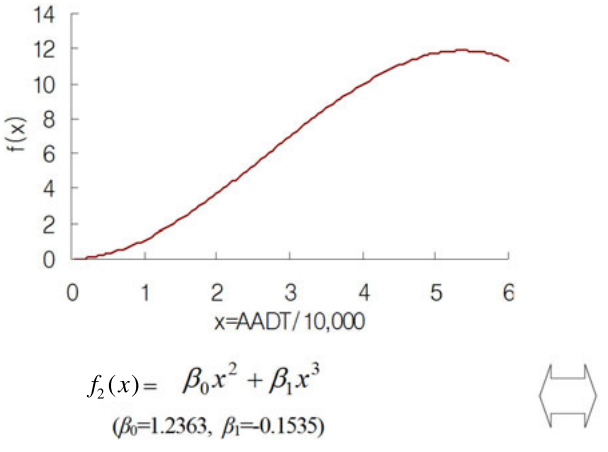
이제 교통사고건수 y와 교통량(AADT) x간의 관계를 나타내는 어떤 한 함수 $y=f(x)$가 있다고 가정하면 이 함수의 적분함수 $F(X)$는 함수 $f(x)$의 아래 면적, 즉 $x=0$에서 $x=X$까지의 $f(x)$의 적분값이 된다. bin 면적들의 합으로 나타내지는 실증적분함수인 $F_E(X)$는 $F(X)$의 추정값이 된다. 따라서 실증적분함수의 함수식을 찾아냄으로써 $F(X)$의 함수식을 추정할 수 있다. 함수 $f(x)$의 함수식은 $F(X)$의 미분함수로 볼 수 있다.

이제 관건은 실증적분함수인 $F_E(X)$의 적절한 미분함수를 구하는 것이다. 그러나 문제는 여러 다른 적분함수들이 비슷한 유형의 곡선을 갖는다는 것이다. 다시 말해, 많은 함수들이 Fig. 4에 나타난 실증적분함수의 미분함수 후보가 될 수 있다는 것이다. 예로, 교통사고건수와 AADT/10,000간의 관계를 나타낼 두 후보함수 $f_1(x)=1.3038x^{1.4646}$과 $f_2(x)=1.2363x^2-0.1535x^3$의 적분함수들이 Fig. 5에서 보는 바와 같이 매우 유사한 곡선 형태를 보일 수 있다.

따라서, Fig. 4에 나타난 실증적분함수와 비슷한 적분함수를 갖는 여러 후보함수들을 조사해야 한다. 가능한 후보함수들로는 교통공학분야에서 비교적 잘 알려진 다음과 같은 일곱 개의 함수들을 생각해 볼 수 있다.







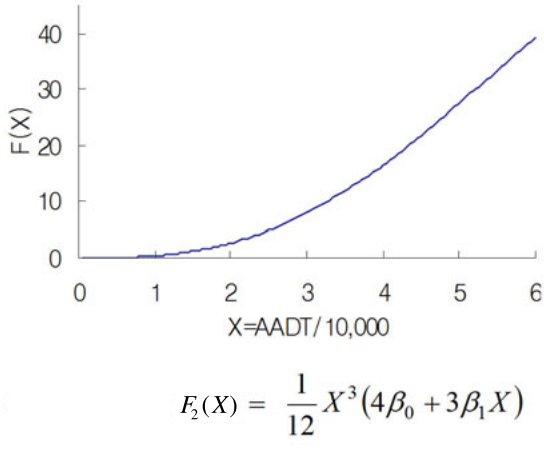


Fig. 5 Example of Two Different Functions Having Similar Cumulative Functions

$$f_1(x)=\beta_0+\beta_1 x \quad \leftrightarrow F_1(X)=\frac{1}{2}X(2\beta_0+\beta_1 X) \tag{4}$$

$$f_2(x)=\beta_0 x^{\beta_1} \quad \leftrightarrow F_2(X)=\frac{\beta_0 X^{\beta_1+1}}{\beta_1+1} \tag{5}$$

$$f_3(x)=\beta_0 x+\beta_1 x^2 \quad \leftrightarrow F_3(X)=\frac{1}{6}x^2(3\beta_0+2\beta_1 x) \tag{6}$$

$$f_4(x)=\beta_0 x^2+\beta_1 x^3 \quad \leftrightarrow F_4(X)=\frac{1}{12}x^3(4\beta_0+3\beta_1 x) \tag{7}$$

$$f_5(x)=\beta_0 x e^{\beta_1 x} \quad \leftrightarrow F_5(X)=\frac{\beta_0 e^{\beta_1 x}(\beta_1 x-1)}{\beta_1^2} \tag{8}$$

$$f_6(x)=\beta_0 x+e^{\beta_1 x} \quad \leftrightarrow F_6(X)=\frac{\beta_0 X^2}{2}+\frac{e^{\beta_1 x}}{\beta_1} \tag{9}$$

$$f_7(x)=x^{\beta_0}e^{\beta_1 x} \quad \leftrightarrow F_7(X)=-X^{\beta_0+1}E_{-\beta_0}(-\beta_1 x) \tag{10}$$
$$E_n(x)=\int_1^\infty \frac{e^{-xt}}{t^n}dt$$

이들 후보함수들 중 최적의 한 함수를 선택하기 위해, 즉 모형 선택을 위해서 각 후보 모형들의 계수값들을 추정한 후에 적합도 판정지수(fitting information)를 비교한다.

3.3. 계수값 추정

계수값 추정은 항상 최적화 기법, 즉 잔차의 제곱값을 최소화하거나 로그 우도(log likelihood)값을 최대화하는 과정에 기초한다. 계수값 추정을 위한 프로그램을 선정할 때에는 모형식이 비선형이 될 수도 있으므로 비선형 모형을 다룰 수 있는 것으로 해야 한다. 또한 교통사고건수의 분포특성으로 알려진 음이항 분포를 고려할 수 있는 것이어야 한다. 이들 조건을 만족하는 것으로 통계프로그램인 SAS® 안에 있는 NLMIXED 프로시저가 있다.

NLMIXED 프로시저는 계수값 추정을 위해 로그우도 최대화 기법(maximum log likelihood estimation method)을 사용한다. 음이항 분포를 가정했을 때 NLMIXED가 사용하는 우도함수는 다음과 같은 형태를 갖는다.

$$L(y,\mu,k)=\sum_i \log(f(y_i,\mu_i,k)) \tag{11}$$

$$\begin{aligned} l_i &= \log(f(y_i,\mu_i,k)) \\ &= y_i\log(k\mu_i)-(y_i+\frac{1}{k})\log(1+k\mu_i) \\ &\quad +\log\left(\frac{\Gamma(y_i+1/k)}{\Gamma(y_i+1)\Gamma(1/k)}\right) \end{aligned} \tag{12}$$

여기서, $L()$ = 우도함수(log likelihood function)

l = 도로구간 i의 우도(log likelihood)

y_i = 도로구간 i의 사고건수

μ_i = 도로구간 i의 추정된 평균 사고건수

k = 과분산계수(dispersion parameter)

3.4. 초기 계수값 선정

계수값 추정에 있어 인지해야 할 한 가지 중요한 점은 NLMIXED 프로시저를 비롯한 최적화 알고리즘을 이용해 추정되는 계수값은 모형 개발자가 설정하는 계수의 초기값에 따라 그 추정 결과가 달라질 수 있다는 것이다. 아직까지 일반적인 비선형 모형의 경우 모든 초기값에 대해 항상 글로벌 최적해를 찾을 수 있다고 보장되는 알고리즘은 없다. 다시 말해, NLMIXED 프로시저가 사용하고 있는 최적화 알고리즘은 설정된 초기값에 따라 지역적 최적해를 찾고 알고리즘이 종결될 수도 있다. 예로 Fig. 6은 어떤 비선형 함수에 대해 초기값에 따라 최적화 알고리즘이 글로벌 최적해를 찾을 수도 있고 지역적인 최적해를 찾을 수도 있음을 나타내주고 있다. 따라서, 일반적으로 적절한 초기값을 찾아 대입해 주는 것이 매우 중요하다. 참고로 NLMIXED 프로시저는 계수들의 초기값이 별도로 주어지지 않을 경우 기본값인 "1"을 사용한다.

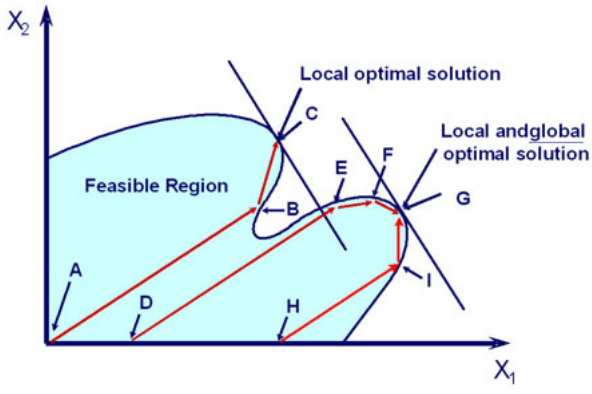
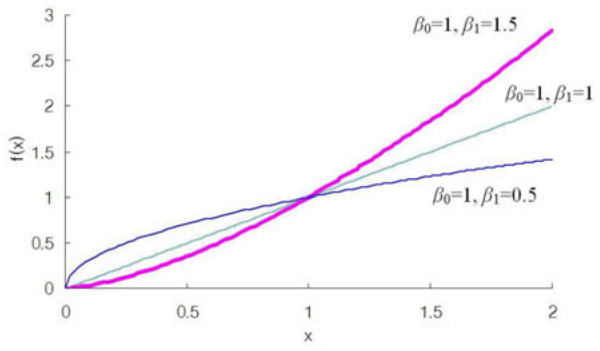


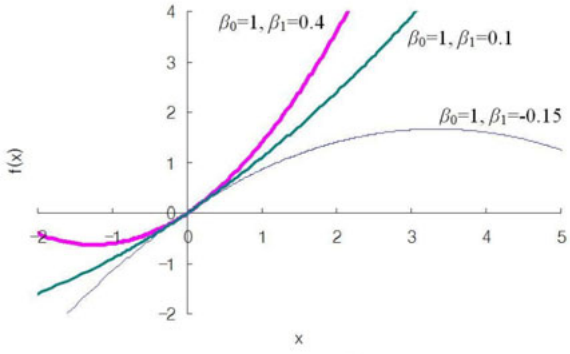
Fig. 6 Local vs. Global Optimum Solutions
(Source: Ragsdale, 2004)

교통사고모형 개발에 있어 예측되는 최적값에 최대한 가깝다고 생각되는 초기 계수값을 결정하기 위해 모형을 개발하고자 하는 자는 모형에 포함된 여러 함수식의 곡선(또는 직선) 형태를 잘 알고 있어야 할 필요가 있다. Fig. 7은 여러 함수별로 함수 내 계수값들의 변화에 따라 함수의 모양이 어떻게 변화하는가를 보여주고 있다. 이 그래프들을 이용하면 개발하고자 하는 모형 관계를 잘 표현할 수 있는 함수식의 초기 계수값들을 찾는데 도움이 된다.



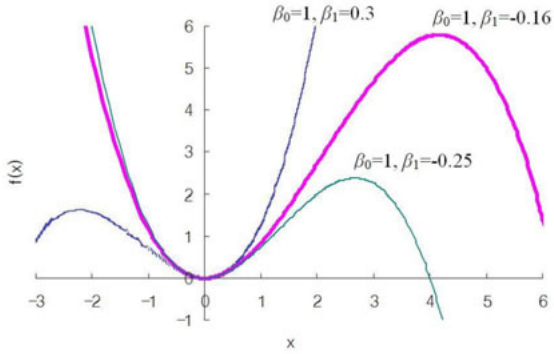
$f(x) = \beta_0 x^{\beta_1}$

(a)



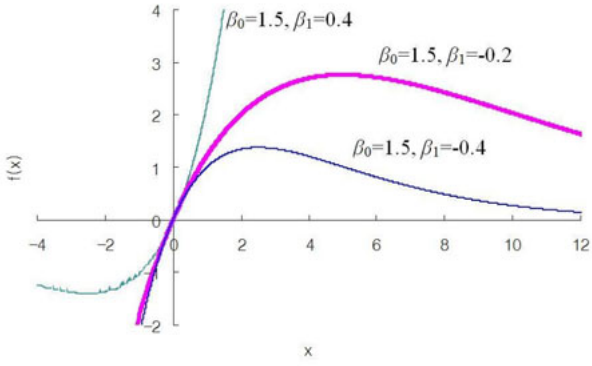
$f(x) = \beta_0 x + \beta_1 x^2$

(b)



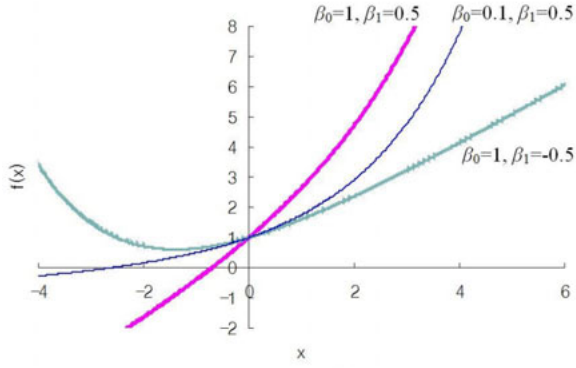
$f(x) = \beta_0 x^2 + \beta_1 x^3$

(c)



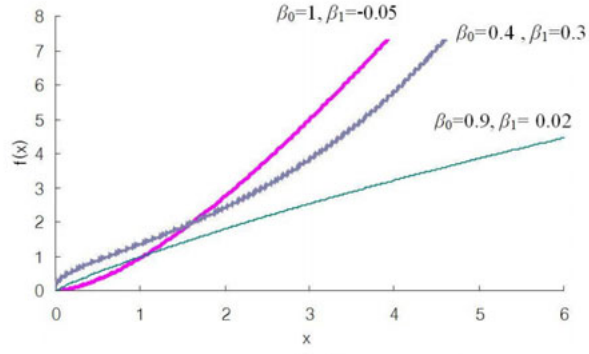
$f(x) = \beta_0 x e^{\beta_1 x}$

(d)



$f(x) = \beta_0 x + e^{\beta_1 x}$

(e)



$f(x) = x^{\beta_0} e^{\beta_1 x}$

(f)

Fig. 7 Example of Change in Shape of Functions According to Parameter Values

3.5. 모형 선택

선택을 위한 비교의 대상이 되는 모형들 간에 서로 내포성(nested)이 있는지에 따라 모형 선택 방법이 다르다. 예로 Eq. (4)에서 Eq. (10)에 나타난 모형들은 서로 내포하지 않는다. 한 독립변수에 대한 함수식을 선정하기 위해 도출된 각 후보 함수식들이 이에 해당한다. 반면, 어떤 모형에 새로 독립변수를 추가하여 확장된 모형을 만들면 확장된 모형은 확장되기 전 모형을 내포한다고 볼 수 있다. 다음에서는 이들 두 경우에 대한 모형 선택 기법들에 대해 설명하였다.

3.5.1. 각 독립변수의 함수식 선정을 위한 모형 선택 (비내포(non-nested) 모형)

NLMIXED 프로시저는 최적화 연산이 이루어지면 그 결과로 "fit statistics" 라는 표를 제공하는데 여기에는 우도값(likelihood)에 근거한 평가지수들, 즉 '-2 log likelihood' 값, 'Akaike information criterion (AIC)', 'a finite-sample corrected version of AIC (AICC)', 'Schwarz's Bayesian information criterion (BIC)' 등이 제시된다. 이들을 이용하여 서로 다른 비내포(non-nested) 모형들의 적합도를 비교하여 최적의 모형을 선택할 수 있다. AIC, AICC, 그리고 BIC를 계산하기 위해 NLMIXED 프로시저가 사용하는 식은 다음과 같다(SAS, 2004):

$$AIC = -2L(\hat{\theta}) + 2p \tag{13}$$

$$AICC = -2L(\hat{\theta}) + 2pn/(n-p-1) \tag{14}$$

$$BIC = -2L(\hat{\theta}) + p\log(s) \tag{15}$$

Eq. (13)부터 Eq. (15)에서 $L()$은 최대화된 로그우도 함수(log likelihood function)를, $\hat{\theta}$는 추정된 계수값들을, p는 계수들의 수를, n은 관측자료수를, 그리고 s는 샘플수를 나타낸다. 이들 평가지수들은 기본적으로 최소화된 '-2 log likelihood value'를 사용하지만 너무 많은 계수를 사용하는데 대한 벌점을 가한다. 결국 가장 작은 AIC, AICC, 또는 BIC 값을 갖는 모형이 가장 적합도가 높은 모형이 된다.

3.5.2. 내포모형(nested Model)에 대한 모형 선택

새 변수가 모형에 추가될 때마다 추가된 이후의 확장된 모형과 추가되기 전의 모형간의 모형 선택이 이루어져야 한다. 이 경우, 우도비 테스트(likelihood ratio test, LRT) 기법이 사용될 수 있다. LRT 기법은 기존 모형(변수가 추가되기 전 모형, nested model)에 대한 '-2 log likelihood' 값에서 새 모형(변수가 추가된 모형)에 대한 '-2 log likelihood' 값을 빼고, 그 결과를 두 모형간의 계수 수의 차이를 자유도 n으로 하는 X_n^2 값과 비교하는 것이다. 이 테스트에서의 귀무가설은 새로 추가된 변수의 모든 계수들 값은 0이며, 따라서 두 모형은 차이가 없다는 것이다.

예를 들어, 두 모형 A 와 B가 있고, 두 개의 계수를 갖고 있는 모형 A가 네 개의 계수(모형 A의 두 계수를 포함하여)를 갖고 있는 모형 B에 내포된 모형이고, 모형 A의 '-2 log likelihood' 값은 450이고 모형 B의 '-2 log likelihood' X_n^2값은 442라고 가정하자. 이 경우, LRT 통계값은 450 - 442 = 8 이고 자유도가 4 - 2 = 2인 X_n^2값은 약 5.99이다. LRT 통계값이 X_n^2값보다 크므로 귀무가설은 기각되고, 따라서 모형 B가 선택된다.

3.6. 모형의 확장

한 개의 독립변수로 구성된 초기 모형을 구축한 뒤, Fig. 1에서 설명된 대로 어떤 독립변수들을 어떤 순서로 추가해 나가며 모형을 확장할 것인지를 결정해야 한다. 모형에 포함될 변수들을 결정하기 위해서는 먼저 그 변수들이 초기 모형에 대해 갖는 영향력을 조사한다. 이를 위해 첫 번째 독립변수인 교통량과 추가될 각 후보 변수들을 두 번째 독립변수로 하는 모형, 즉 y= f(AADT)·g(두 번째 변수) 형태의 모형들을 개발한다. 여기서의 기본 개념은 초기 모형과 각 두 번째 모형들에 대해 LRT 테스트를 시행하여 두 번째 변수들이 로그우도(log likelihood)값의 개선에 미치는 영향력을 조사한다는 것이다.

이를 위해 첫째로, 새로 추가할 각 변수에 대한 함수식을 결정해야 한다. 이 경우 두 번째 변수를 위한 함수 g(두 번째 변수)의 종속변수는 초기 모형 구축 시 사용된 y가 아니라, y와 초기 모형에 의해 예측된 사고건수 f(AADT)의 비인 y/f(AADT)가 된다. 이 값은 Eq. (2)에서 설명된 R과 같다.

모든 샘플(예, 분석 대상 도로구간들)에 대한 이 R값들은 바로 초기 모형에 곱해져야 할 값들이 된다. 다시 말해, 만약 어떤 도로구간에 대한 R값이 1보다 크다면, 초기 모형이 그 도로구간에서 실제 일어난(기록된) 교통사고 건수보다 적은 수의 교통사고 건수를 추정한 것을 의미하며, 반대로 R값이 1보다 작다면, 초기 모형이 실제 교통사고 건수보다 많은 교통사고 건수를 추정한 것이 된다. 따라서 기록된 사고수와 모형에 의해 추정된 사고수가 같아지게 하기 위해 R값에 가까운 수만큼 초기 모형에 곱해주어야 한다. 요약하자면, R값과 두 번째 변수간의 관계를 잘 설명해 줄 수 있는 함수식을 찾는다는 것이다.

추가하려는 각 독립변수들의 함수식이 결정되면 이들 함수식이 첨가된 확장된 모형의 모든 계수값들을 재추정하여 모형을 완성한다. 이렇게 구축된 각각의 두 번째 독립변수가 추가된 확장 모형들의 '-2 log likelihood' 값들을 초기 모형의 '-2 log likelihood' 값과 비교한

다. 즉, LRT 기법을 이용하여 초기 모형에 추가될 독립변수들을 결정한다. 초기 모형에 추가될 독립변수의 순서는 그 변수들이 '-2 log likelihood' 값을 감소시킨 정도에 따라 결정된다.

확장된 모형의 계수값 추정에 있어 새로 추가된 계수들의 초기값은 앞서 설명된 방법으로 결정하며, 기존 계수들의 초기값은 Fig. 8에서 볼 수 있는 바와 같이 확장 전의 모형 개발 시에 추정되었던 계수값들을 사용한다.

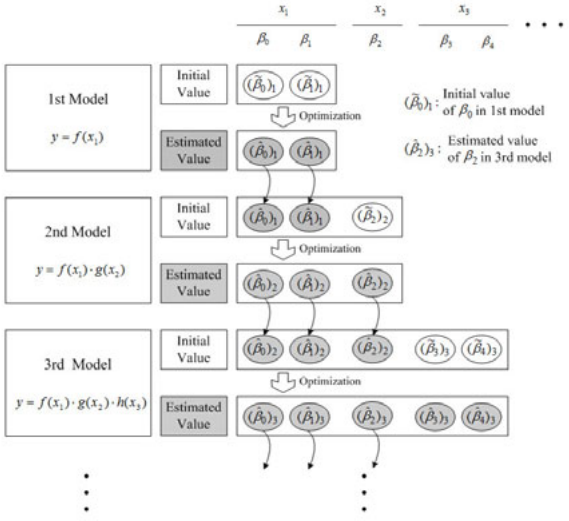


Fig. 8 Concept of Setting Initial Parameter Values During Model Expansion

새 확장 모형이 개발되면 새롭게 추가된 변수가 모형의 설명력에 통계적으로 유의한 수준의 영향을 끼치는지를 판별하기 위해 LRT 기법을 이용한다. 그 결과에 따라 그 변수가 추가될 수도 있고 제거될 수도 있다.

4. 방법론 적용 결과

이 장에서는 이 논문에서 제시한 방법론을 적용하여 미국 North Carolina 주의 4차로 도로 기본구간에 대한 교통사고모형을 개발한 예를 결과 위주로 소개하였다.

4.1. 교통사고모형 개발 개요

4.1.1. 개발 목적

미국 North Carolina 주의 도시근교지역 다차로 도로구간에 대해 여러 구성요소들이 교통안전성에 미치는 영향을 알아보기 위한 것이었다.

4.1.2. 연구 범위

- 대상 도로: North Carolina 주 내에 있는 0.5 마일 이상 길이의 4차로 기본구간
- 제한속도: 시속 45 마일과 시속 55 마일
- 중앙분리대 형태: Non-Traversable Median (NTM)과 Two-Way Left-Turn Lane (TWLTL)
- 교통량 및 사고자료: 2001년~2003년

4.2. 자료 개요

미국 North Carolina 주 교통국(Department of Transportation, DOT)의 데이터베이스를 이용하여 대상이 되는 도로 구간들에 대한 2001년에서 2003년까지의 3년간의 교통량과 사고자료를 구하였으며, 그 밖의 도로구간 길이, 접근로 수, 차로폭, 제한속도, 길어깨 유형 및 폭, 중앙분리대 유형 및 폭 등에 대해서는 현장 조사를 통해 구하였다. 본 교통사고모형 개발 연구에 사용된 자료의 개괄은 Table 1과 같다.

4.3. 개발 결과

4.3.1. 개발 모형

이 연구에서 제시한 방법론을 적용하여 개발된 전체 사고에 대한 최종 통계적 교통사고모형은 Eq. (16)과 같다.

$$\begin{aligned} y &= \beta_0 x_1^{\beta_1}(1+\beta_2 x_2)(1+\beta_3 x_3 e^{\beta_4 x_3})(1+\beta_5 x_4) \\ &= 1.0541 x_1^{1.3411}(1+0.0401 x_2) \\ &\quad (1+16.8186 x_3 e^{-3.5433 x_3})(1-0.1189 x_4) \end{aligned} \tag{16}$$

여기서, y = 추정사고건수/년/mile/년

x_1 = AADT/10,000(대/일)

x_2 = 접근로밀도(접근로/mile)

x_3 = 길어깨폭(ft)

x_4 = 길어깨 유형

(연석 비설치=0, 연석설치=1)

Eq. (16)에서 볼 수 있듯이 전체 사고에 대하여 개발된 교통사고모형에는 총 네 개의 독립변수가 선정되었으며, 각 독립변수에 대한 함수식은 모두 다른 형태를 취하고 있다. 특히 함수식의 형태가 선형과 비선형 모두 포함되어 있음에 주목할 필요가 있다. 이 모형에 따르면 다차로 도로 기본구간의 경우 교통량이 많을수록, 접근

Table 1. Descriptive Statistics of the Data

Group		Statistic	# of directional segments	Total length (mile)	Average access point density (access points/mile)	Average AADT (vehicles/day)	# of total collisions	# of injury collisions
Sites with curbs	45 mph	NTM*	26	24.34	5.54	25,404	293	99
		TWLTL	28	26.60	13.51	21,851	399	180
		Subtotal	54	50.94	9.68	23,562	692	279
	55 mph	NTM*	8	11.32	2.82	27,660	185	79
		TWLTL	30	38.04	10.22	17,552	316	139
		Subtotal	38	49.36	8.66	19,680	501	218
	Subtotal		92	100.30	9.26	21,958	1,193	497
Sites w/o curbs	45 mph	NTM*	30	19.12	9.72	25,367	337	124
		TWLTL	16	11.67	17.07	22,708	211	72
		Subtotal	46	30.80	12.28	24,442	548	196
	55 mph	NTM*	35	39.59	4.84	17,968	368	136
		TWLTL	26	21.16	5.90	14,977	165	56
		Subtotal	61	60.75	5.29	16,693	533	192
	Subtotal		107	91.55	8.29	20,024	1081	388
Grand total			199	191.85	8.74	20,918	2,274	885

*NTM : Non-traversable median

Table 2. Validation Results

Model		# independent variables	# parameters	-2 log likelihood	BIC
Total Collision	Model_this study (four independent variables)	4	7	832.7	869.8
	Model_GLM (four independent variables)	4	6	844.5	876.3
	Model_GLM (eight independent variables)	8	10	841.2	894.1
Injury Collision	Model_this study (three independent variables)	3	6	567.1	598.9
	Model_GLM (three independent variables)	3	5	579.7	606.1
	Model_GLM (eight independent variables)	8	10	576.6	629.6

로밀도가 높을수록, 길어깨폭이 좁아질수록, 그리고 길가에 연석이 설치되지 않았을 경우 전체 사고수가 더 높은 것으로 나타났다.

마찬가지로 부상 사고에 대해 이 연구에서 제시한 방법론을 적용하여 교통사고모형을 개발한 결과는 Eq. (17)과 같다.

$$y = \beta_0(1+\beta_1 x_1)(1+\beta_2 x_2)(1+\beta_3 x_3 e^{\beta_4 x_3})$$
$$= -0.2244(1-2.7264x_1)$$
$$(1+0.0434x_2)(1+53.5534x_3 e^{-4.9734x_3}) \quad (17)$$

여기서, y = 추정사고건수/년/mile/년

x_1 = AADT/10,000(대/일)

x_2 = 접근로밀도(접근로/mile)

x_3 = 길어깨폭(ft)

Eq. (17)에서 볼 수 있듯이 부상사고만을 대상으로 하여 개발된 교통사고모형에는 총 세 개의 독립변수가 포함되었으며, 마찬가지로 각 독립변수가 갖는 함수식의 형태는 선형과 비선형을 모두 보였다. 전체 사고를 대상으로 했을 때와는 달리 부상 사고를 대상으로 하였을 때는 노견에 설치된 연석의 설치 유무가 통계적으로 유의한 수준의 영향을 미치지 않는 것으로 나타났다.

4.4. 타당성 검토

이 연구에서 제시한 개발 방법론을 적용하여 개발된 최종 교통사고모형의 타당성을 검토하기 위해 일반적으로 사용되고 있는 Generalized Linear Model(GLM) 기법을 이용하여 개발한 모형과 그 설명력들을 비교하였다. 이를 위해 각 모형의 '-2 log likelihood' 값 및 BIC 값을 구하였으며 그 결과를 Table 2에 나타내었다.

Table 2에서 보는 바와 같이 전체 사고와 부상사고에 대한 모형 모두 본 연구에서 제시한 개발 방법론을 적용하여 개발된 모형이 기존 GLM 모형에 비해 '-2 log likelihood' 값 및 BIC 값이 더 작게 나와 설명력이 우수한 것으로 나타났다.

5. 결론

통계적 교통사고모형은 지금까지 특정 조건들을 갖고 있는 도로 구간 또는 교차로 등에 대해 그 조건들이 교통사고 수에 얼마만큼의 영향을 미치는가를 예측하기 위해 사용되는 유일한 현실적 도구이다. 이 도구는 과거 자료에 전적으로 의지하여 개발되므로 그 자료를 분석하여 관계들을 발견해 내는 방법론에 따라 이 도구의 성능이 결정된다.

현재 통계적 교통사고모형 개발에 가장 보편적으로 사용되고 있는 방법은 GLM기법 중 하나인 로그선형회귀모형(Generalized log-linear regression model)인데, 이 방법은 종속변수와 독립변수들 간의 관계가 선형이 아닌 비선형일 수도 있음을 고려치 못하는 것과 종속변수와 각 독립변수들 간의 관계를 개별적으로 고려치 못한다는 것이다.

이 연구에서는 Hauer가 제안한 모형 개발 기법을 토대로 기존의 GLM 방법과 다른 교통사고모형 개발 방법론을 제시하였다. 최초 하나의 독립변수를 이용해 초기 모형식을 만들고 나머지 독립변수들을 하나씩 추가해가며 초기 모형식을 확장해 나간다는 점이 가장 두드러지는 차이점이다. 이 연구에서는 포함 대상으로 고려된 독립변수들을 추가하는 순서를 정하는 방법과, 추가되는 각 독립변수들에 대한 함수식을 결정하는 방법을 제시하였다. 함수식에 포함된 계수값의 추정법과 모형의 확장 방법에 대해서도 소개하였다.

이 연구에서 제시한 교통사고모형 개발 방법론의 특징은 개개의 독립변수가 교통사고건수에 대해 가질 수 있는 관계의 개별성을 고려할 수 있다는 것과, 그 관계의 다양한 형태(선형이 될 수도, 비선형이 될 수도 있음)를 보다 가깝게 찾아낼 수 있다는 것이다.

이 연구에서 제시한 방법론을 미국 North Carolina 주 자료를 이용하여 적용한 사례를 소개하였다. 그 결과 GLM 방식을 이용하였을 때보다 개발된 모형의 설명력이 더 우수한 것으로 나타났다.

감사의 글

본 연구는 North Carolina Department of Transportation의 자료를 이용하여 수행되었습니다.

References

Hauer, E., 2004, Statistical Road Safety Modeling. Transportation Research Record: Journal of the Transportation Research Board, No. 1897, *TRB, National Research Council*, pp. 81-87.

Hauer, E., F.M. Council and Y. Mohammedshah,, 2004, Safety Models for Urban Four-Lane Undivided Road Segments. Transportation Research Record: Journal of the Transportation Research Board, No. 1897, *TRB, National Research Council*, pp. 96-105.

Hauer, E., and J. Bamfo., 1997, Two Tools for Finding What Function Links the Dependent Variable to Explanatory Variables, Proc., *ICTCT 97 (International Cooperation on Theories and Concepts in Traffic Safety)*.

SAS Institute Inc., 2004, *SAS/STATÆ 9.1 User's Guide*, SAS Institute Inc.

Vogt, A., and Bared, J.G., 1998, Accident Models for Two-Lane Rural Roads: Segments and Intersections. *Final Report, FHWA Report No. FHWA-RD-98-133*, Federal Highway Administration.

(**접수일** : 2012. 9. 7 / **심사일** : 2012. 9. 9 / **심사완료일** : 2012. 9. 17)