韓國港灣學會 第12卷 第1號, pp.55~64. 1998.

論 文

# Directed Graphical Approach for Economic Modeling: A Revision of Crandall's Occupant Death Model

*J. W. Roh**

## Directed Graph를 이용한 경제 모형의 접근
## - Crandall의 탑승자 사망 모형에 관한 수정-

노 재 확

**Key Words** : Artificial intelligence (인공지능), Directed Graph Algorithm(DGA), Occupant Death (탑승자 사망), Latent variable (잠재변수), Nonstationarity (비정상성)

## Abstract

Directed graphic algorithm was applied to an empirical analysis of traffic occupant fatalities based on a model by Crandall. In this paper, Crandall's data on U.S. traffic fatalities for the period 1947-1981 are focused and extended to include 1982-1993. Based on the 1947-1981 annual data, the directed graph algorithms reveal that occupant traffic deaths are directly caused by income, vehicle miles, and safety devices. Vehicle mileage is caused by income and rural driving. The estimation is conducted using three stage least squares regression. Those results show a difference between the traditional regression methodology and causal graphical analysis. It is also found that forecasts from the directed graph based model outperform forecasts from the regression-based models, in terms of mean squared forecasts error. Furthermore, it is demonstrates that there exists some latent variables between all explanatory variables and occupant deaths.

## 1. Introduction

This study analyzes U.S. traffic fatalities using directed graphs. Directed graphs, as presented in Spirtes, Glymour, and Scheines (SGS, 1993), are an alternative to regression-based procedures for specifying 'causal structure' on observational data. The technology is applied to data on variables previously

* 정회원, 부산대학교 강사

studied by Crandall and Graham (1984). Of the two models presented by Crandall and Graham (1984), it was chosen Crandalls model for comparison since we want to concentrate on the factors causing occupant deaths. Out-of-sample forecasts from Crandall's model are compared with those generated with directed graphs. Forecasts of traffic fatalities based on Crandall's model are dominated in mean squared error by those based on the directed graphs.

The paper is presented in six sections. i ) it is offered a brief review of the literature on traffic fatalities and replicate the results in Crandall (1984), ii ) review of directed graphs, iii) presentation of results from the application of directed graphs to the variables defined in Crandall (1984), iv) an analysis of out-of-sample forecasts of traffic fatalities from Crandall's model and from the directed graphs model. v ) results from the application of directed graphs under the assumption of the existence of latent variables. The sixth section concludes the paper.

## 2. Crandalls Model

The concept of offsetting behavior in traffic fatalities was first introduced by Peltzman (1975). Since Peltzman published his findings additional research has been performed in this area, using extended data set or new variables. Crandall and Graham (1984), keeping Peltzmans tradition, contributed a system of equations model of occupant death in which the net effect of safety regulation on death rate may be estimated consistently by applying ordinary least squares to the reduced form, death-rate equation: DR= h (K, A, H, V, R, Y, P).[1] They also suggested a more elaborate simultaneous equation model which tested K, A, and V as endogenous variables.

Crandall and Graham differed, however, over the appropriate measure of traffic fatalities, the use of a crashworthiness index, and a vector of attributes describing highway design. Crandall used the number of occupant deaths as a dependent variable, while Graham used the occupant death rate. Crandall used log transformed data and found safety regulation, rural driving, truck driving, income, trend, weight of cars, and vehicle miles to be significant while Graham used unlogged data. Of the two models presented by Crandall and Graham (1984), Crandalls model will be studied further in the following section.

## 3. Review of Directed Graphs

### 3.1 Directed Graphs

SGS (1993) introduced a number of algorithms to reveal causal structures[2]. The PC algorithm, which is the one we use in this paper, and the most basic algorithm, is designed for the case of causal sufficiency, which assumines we have all necessary variables in our observational set. More advanced refinements of the PC algorithm include the Modified PC algorithm,[3] the Causal Inference

1) DR is the relevant highway death rate per mile; K is a measure of driving by high risk youth; A is per capita alcohol consumption; H is a vector of attributes describing highway design; V is an index of the average weight of the vehicle fleet; R is a proxy for the degree of crashworthiness required by federal regulation; Y is the value of a drivers time (his earned income); and P is an index of the cost of an accident.

2) These algorithms are described in detail in Spirtes, et al. (1993, pp.117-118).

3) Ibid., (pp.166-167)

Algorithm[4] and the Fast Causal Inference algorithm[5] Those refinements are designed to detect latent variables in the no causal sufficiency case.

The directed graph is determined by these algorithms. A directed graph is an ordered triple <V,M,E> where V is a non-empty set of vertices (variables), M is a non-empty set of marks (symbols attached to the end of undirected edges), and E is a set of ordered pairs. Each member of E is called an edge: (i) undirected edges (e.g., A ---- B); (ii) directed edges (e.g., B ----> C); (iii) both or bi-directed edges (C <----> D); non-directed edges (o ---- o) and partially directed edges (o ---->). Bi-directional edges, non-directed edges, and partially directed edges are possible only when latent variables exists. A directed acyclic graph is a directed graph that contains no directed cyclic paths (an acyclic graph contains no vertex more than once). Only acyclic graphs are used in the paper.

### 3.2 PC Algorithm

The PC algorithm[6] is an ordered set of commands which begins with a general unrestricted set of relationships among variables and proceeds step-wise to remove edges between variables and to direct "causal flow." Edge removal and direction of causal flow are based on independence or conditional independence (or lack thereof) as represented by zero correlation or partial correlation. Briefly, one forms the complete undirected graph C on the vertex set V. The complete undirected graph shows an undirected edge between every variable of the system (every variable in V). Edges between variables are removed based on zero correlation or partial correlation. The conditioning variable on removed edges between two variables is called the sepset of the variables whose edge has been removed (for vanishing zero order conditioning information the sepset is the empty set). Direct edges between triples X ---- Y ---- Z as X ---->Y< ---- Z if Y is not in the sepset of X and Z.[7] If X ---->Y, Y and Z are adjacent, X and Z are not adjacent, and there is no arrowhead at Y, then orient Y ---- Z as Y ---->Z. If there is a directed path from X to Y, and an edge between X and Y, then orient the edge X ---- Y as X ---->Y. This process continues until no more edges can be oriented.

### 3.3 FCI Algorithm.

Without causal sufficiency means that we do not have enough variables V to include all the common causes of every pair of variables. The PC algorithm is modified to allow for the possibility that a third unknown variable, or set of variables, may be responsible for the observed correlations between variables.

The first stage of the FCI is just like the first stage of the PC algorithm. We initialize a partially oriented inducing path graph[8] to

4) Ibid., (pp. 183-184)

5) Ibid., (pp.188-189)

6) The algorithm is described in detail in Spirtes, et al. (1993, pp. 117-118).

7) Partial correlation X and Z given Y is tested to determine the sepset.

8) A partially oriented inducing path graph for directed acyclic graph G with inducing path graph G over O is intended to represent the adjacencies in G and some of the orientations of G

complete undirected graph, and then we remove the edge between X and Z if they are d-separated[9] given subsets of vertices adjacent to X or Z in a partially oriented path graph. The result is essentially the graph constructed by the PC algorithm given data faithful to the directed acyclic graph. In the fourth step of the FCI, to remove the edges which were connected due to latent variables, and are not really correlated, all possible d-separations are tested. During this procedure, all edges uncorrectly generated by latent variables are removed and only the correct set of adjacencies are survive. As a last step, the algorithm unorients all edges o---o and reorients them using a collider and definite discriminating path.[10]

## 4. An Application of Directed Graphs to Crandalls Model

### 4.1 Replication of Crandalls Model

Crandalls original (1984) paper used the number of occupant deaths as a dependant variable. Table 1 summarizes my attempt to replicate Crandalls results. He used a log transformed model which showed R2 (safety regulation), rural, truck, income, trend, weight, and mile to be significantly different from zero. In my replication (Table 1), I get the same sign for six of these seven variables, and find the same significance for four out of seven. Alcohol is significantly different from zero, while it is not significant in Grahams model. Likewise, limited access road is significantly different from zero in my replication, but not significant in his model. In Crandalls model, truck and trend were significantly different from zero, but they were not in my model. The low D-W value and high R2 cast suspicion on the possibility of the spuriousness of the regression (see Granger and Newbold (1974)).

### 4.2 The Nonstationarity of Crandalls Model

When Crandall performed his study, the concept of nonstationarity of time series data was relatively unknown to the traffic field. As a result, researchers generally used levels data in statistical modeling and did not consider the possibility of a spurious regression. Tests of nonstationarity suggest that the data are not mean stationary.

Nonstationarity was tested using the Augmented Dickey-Fuller test. All test values (Table 2) indicate that the data are nonstationary. If there are no cointegrating relationships between or among these nonstationarity variables, then the analysis shoud be carried out in first difference. We have not studied these data for cointegration due to the short span of our data (1947-1981). One should also be concerned that inferences based on such nonstationary data will be misleading, calling into question Crandall's results given in Table 1. This may also explain the high R-squared value. Below we consider the use of directed graphs on Crandall's data under a difference transformation.

9) For example, in XYZ, X and Z are d-separated given Y. For more detail, see SGS (1993) , p. 71.

10) The path has a triangle relationship which can help to distinguish direction. For detail see SGS p.181.

Table 1. The Comparison of Replication to Crandalls Estimates.

| | Original Paper | | Replication | | |
|---|---|---|---|---|---|
| Variable | Coefficient | | Coefficient | t-Statistic | |
| R2 | 2.331 | * | 4.448091 | 7.78963C | * |
| YOUTH | 0.7325 | | -0.127786 | -0.35983C | |
| ALCO | 0.2941 | | 1.450538 | 3.710685 | * |
| RURAL | 0.7506 | * | 0.882133 | 2.696192 | * |
| LAC | -0.06 | | -0.192828 | -4.775678 | * |
| TRUCK | 0.389 | * | 0.178662 | 1.861946 | |
| INCOME | 1.062 | * | 0.681215 | 2.43988E | * |
| TREND | -0.025 | * | 0.058154 | 0.842792 | |
| WEIGHT | -2.86 | * | -3.420881 | -3.321916 | * |
| MILE | 1.367 | * | 0.598594 | 1.926253 | |
| R-squared | 0.979 | - | 0.9817 | - | - |
| Adjusted R2 | NA | - | 0.974 | - | - |
| F-statistic | NA | - | 129.44 | - | - |
| Durbin-Watson | NA | - | 1.466 | - | - |
| S.E. of regression | NA | - | 0.033 | - | - |

* shows significance. In the original paper, t-values were not reported. N.A means not reported

** Crandall (1984) included the following variables and data sources His death data comes from the National Safety Council, Accidents Facts. R2 is a proxy for safety regulation, which is the weighted proportion of miles driven in each calendar year by cars of model year 1986 and later. The proxy is generated using registration data for cars of different model years. Registration data comes from R.L. Polk Company and Motor Vehicle Facts and Figures, U.S. Motor Manufacturers Association, Detroit, Michigan, annual. Annual estimates of the average weight of car come from Automotive News Market Data Book, various years and Ward's Automotive Yearbook, various years. I could not find specific weight data in those two books. In Ward's Automotive Yearbook, I found the consumption weight of materials in the automotive industry and a couple of years estimation of typical passenger car weight. The Automotive News Market Data Book does not provide consistent data for car weight either. I assumed that his calculation comes from the average weight of material consumption of the industry. I get those data from Motor Vehicle Facts and Figures, U.S. Motor Manufacturers Association, Detroit, Michigan, annual. Very surprisingly, my replication of his results improves substantially when this variable is added to the model. Rural is the proportion of total vehicle miles driven on rural roads and highways. Lac is used to capture the relatively safe, modern design of rural and urban Interstate highways. Limit access (Lac) is the proportion of total miles driven on limit-access highways. Crandall states that this proportion has been increasing in the post-war period due to the completion of the interstate highways. His data source is Highway Statistics. Again, I could not find this data. For example, Peltzman (1975) used primary limited access roads for 1956-1964 and estimated before 1964. I followed Peltzmans methodology. For young driver, he used national estimates of the proportion of licensed drivers under the age of 25. These estimates, published by the National Safety Council, are available only since 1957. For earlier years, estimates are predicted from a regression of licensed-driver data on population data. Alcohol is per capita consumption of alcohol in gallons per year per working age adult. It is assumed that the prevalence of drunken driving is roughly proportional to the amount of alcohol consumed per person. His data source for alcohol consumption per person is the Center for Alcohol Studies, Rutgers University, and the National Research Council, Washington D.C. The Center for Alcohol Studies, Rutgers University, provided me the data for 19471993. Earned income is the same as Peltzmans. Earned income is defined as disposable income minus transfers, interest income, dividends, and rental income. The measurement is expressed in thousands of 1972 dollars. His data source is the Survey of Current Business, U.S. Bureau of Labor Statistics, Washington D.C. His cost variable is the ratio of the automobile maintenance and repair component of the CPI to the CPI for all items. The source is Survey of Current Business, various years. Ironically, Graham criticized Peltzmans use of a time trend for omitted variables that have a gradual beneficial effect on highway safety (1982), but he also included a time trend (1984). Truck mileage and total vehicle mileage comes from Highway statistics.

### 4.3 Graphical Analysis on Crandalls Model

The following results are generated in the causal sufficiency case, which implies that every common cause is in the set of variables studied. Differenced data (noted as Δ) and a 10% significance level are used in the directed graph analysis, and the result is reported in Figure 1. Directed graphs need some background knowledge to determine the directions of causality.[11] Figure 1 gives the causally sufficient directed graphs found under 10% levels of significance on first differenced data.

The edges between Δdriver, Δalcohol , Δrural, Δtruck mileage, Δweight of car, and Δlimit access driving are removed by zero order partial correlation. For example, the edge between occupant death and alcohol consumption is removed since the partial correlation is 0.2490 with probability of 0.1227 which means that if the population partial correlation equals zero then the probability of observing a sample partial correlation with absolute value greater than 0.2490 is 12 %. Since the significance level is 10%, this hypothesis is accepted and as a result the edge is removed. After considering first and second order partial correlations, the adjacent relationships are decided. To decide the direction of causality, the algorithm uses the concept of a Sepset.[12] The direction is decided towards a variable if the Sepset does not exist. The final result is reported in panel (d) in Figure 1.

ΔOccupant death is caused by Δvehicle mileage, Δincome, and Δsafety device. ΔIncome causes Δalcohol consumption, Δvehicle mileage, and Δ occupant deaths. Undirected edges between Δ limited access roads and Δsafety device (R2) in the DAG (directed acyclic graph) imply that the algorithm cannot completely determine the causal direction with present background knowledge, so more background knowledge is required. ΔVehicle weight, Δyoung driver ratio, and Δtruck mileage are not causally connected to Δoccupant death.

This model uses only six of the ten variables used in Crandalls work. In particular, we here, use only three variables to specify the number of occupant deaths.

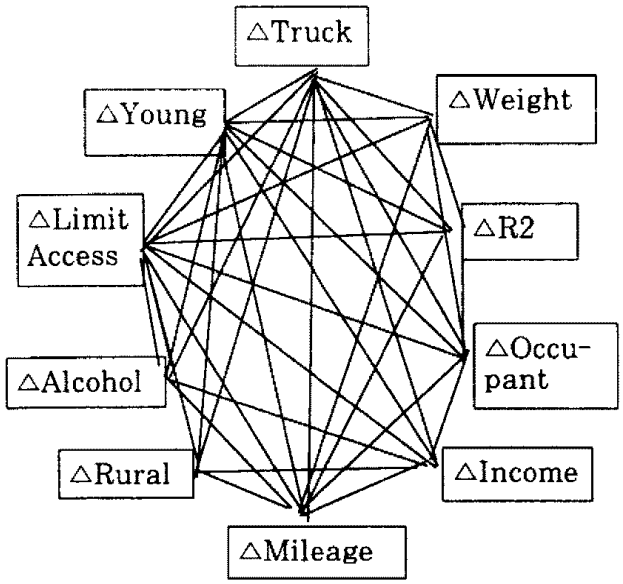
Table 2. Augmented Dickey-Fuller Test for Nonstationarity on Replicated Crandalls Data, 1947-1981.

| Series | ADF (No Trend ) (1) | ADF (Trend ) (2) |
|---|---|---|
| Occupant Death | -2.11 | -1.65 |
| R2 | -1.97 | -3.25 |
| Truck | - .07 | - .89 |
| Mile | .44 | - .98 |
| Youth | - .33 | -2.46 |
| Alcohol | .11 | -1.84 |
| Laccess | -1.67 | - .59 |
| Income | -1.16 | -2.31 |
| Weight | .76 | -2.53 |
| Rural | - .14 | -3.56 |

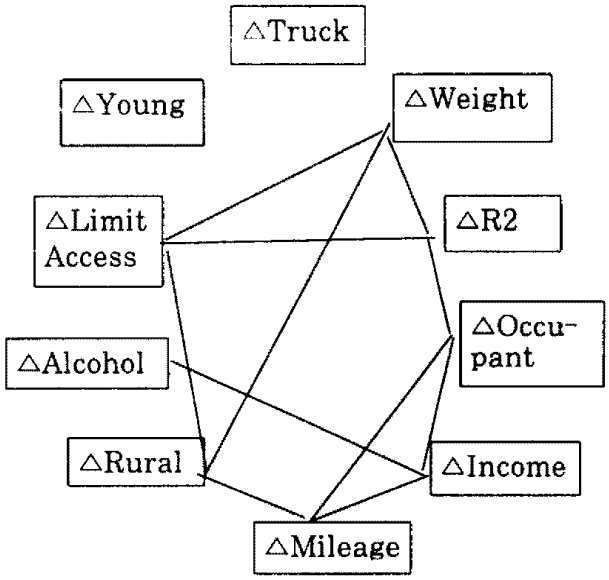
Note : (1) Critical Value at 5% is -2. 89. Reject null of nonstationarity for calculated values less than this critical value. (2) Critical Value at 5% is -3.61. Reject null of nonstationarity for calculated values less than this critical value.

11) We used restrictions such as the assumption that income is exogenous to all explanatory variables and dependent variables. Other variables are only exogenous to independent variables. We try to minimally restrict the relations among variables.

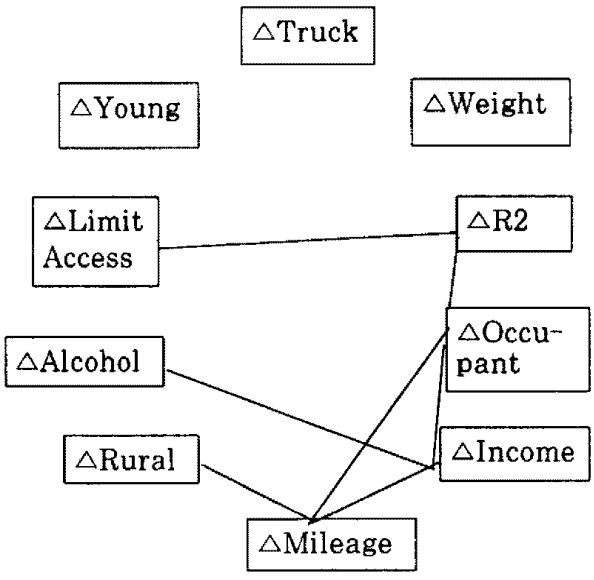
12) see the previous section or SGS (1993), p.72



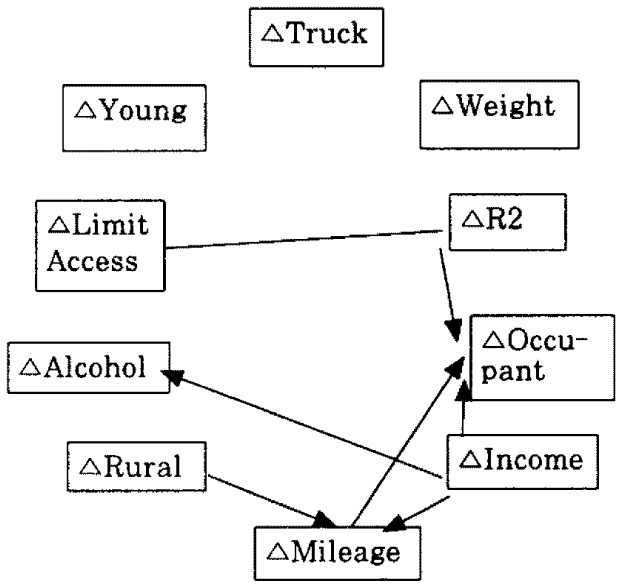

(A) Complete Undirected Graph




(B) Resulting Adjacencies after Zero Order Independence




(C) Resulting Adjacencies after Considering




(D) The Final Result First Order Independence

Fig. 1 Directed Graphs at Alternative Steps in the TETRAD Search of the Causally Sufficient Case of Crandalls Occupant Death Model Based on 1947-1981 Differenced Data and 10 % Significance Level.

### 4.4 Estimation, Forecasting and Comparison

I estimate the parameters underlying the directed graph given in Figure 1 using three stage least squares estimation, to account for possible endogeneity. The signs are as we expected. For example, safety device, which is represented by the danger of cars (R2), and deaths are directly related, as is also the case for income and mile in equation(1). (.) represents t-values:

(1) $\Delta$Occupant death
$= 1.4*\Delta\text{Income} + .04*\Delta\text{Mile} + 2.19\Delta R2$
(5.93) (0.24) (3.68)

(2) $\Delta\text{Mile} = .86*\Delta\text{Income} - .61*\Delta\text{Rural}$
(4.78) (-2.44)

The rate of rural roads has a negative effect on deaths since the rate of urban roads positively affects traffic deaths.

To demonstrate the superiority of the directed graphs over Crandalls specification, forecasts of occupant deaths from both Crandalls model and the DAG based model as compared. For comparison, I use a recursive forecasting methodology which updates the data set and reestimates the parameters every year.

Over the 1982-1993 period, Crandalls model has a root mean squared forecast error of .00046, while the DAG based model has a MSE of .003. In Table 3 I present results of tests of significant differences between these models, using the test given in Ashley, Granger and Schmalensee (1969). The coefficient $\beta o$ tests the difference in bias between the two forecast series and the coefficient $\beta 1$ tests the difference in variance of the forecasts error.

The results show that the new model has a lower MSE than Crandalls model at a .139 marginal significance level. This result implies that the regression based on observational data may convey causal relations that are not genuine. Thus, their forecasts for out-of-sample are fragile. Closeness to the true structural relation is indirectly proven by the smaller MSE.

This fact can provide us additional proof that the regression method may produce incorrect results since regression does not consider the possibility of latent variables and loses forecast power in out-of-sample data.

Table 3. Parameter Estimates from Test of Equality of Mean Squared Forecast Errors.

| Parameter Estimates | t-stat |
|---|---|
| $\beta o$ = - .017 | -1.08 |
| $\beta 1$ = - .04 | - .27 |

Note : Forecasts for 1982-1993. The tests is defined in Ashley, Grandger and Schmalensee (1969). The parqameter $\beta o$ gives the differences in bias between the two models. The parameter $\beta 1$ is proportional to differences in error variances. We define $\varepsilon 1$=real data-new model forecasts and $\varepsilon 2$=real data -Crandall forecasts, and the estimation is performed with regressand $\varepsilon 1 - \varepsilon 2$ and regressor $(\varepsilon 1 - \varepsilon 2) - E(\varepsilon 1 - \varepsilon 2)$. F(2,10) is 0.5557, showing that the DAG model is jointly better than Crandall's model at 0.319 level of significance.

## 5. No Causal Sufficiency Case

The previous chapter presented the causal sufficiency case of directed graph analysis where DAG (Directed Acyclic Graph) was derived from the PC algorithm. This section provides the no causal sufficiency case, in which the possibility of latent variables is considered. This DAG is derived using FCI algorithm.The results are reported in

Figure 3. Notice that the results support the existence of latent variables, a finding which was not possible before the application of directed graphs. The partial directed edges ( o---> ) between A and B represent the notion that either A is a cause of B, or there is a common cause of A and B, or some combination of these. Bi-directional edges (<---> ) represent the presence of common latent variables.

The possibility of latent variables changes the interpretation of the DAG for the no sufficiency case. Adjacencies are similar to those in the sufficiency case, but edges between alcohol and consumption and mile and occupant death disappears. Those two edges are removed by partial correlation considering occupant deaths. The bi-directional edges between R2 and occupant death reveals that there exists a latent common variable. Further, partially directed edges between income and occupant death also reveal possible latent variables.

The possibility of latent variables biases the estimates generated by regression methods. Furthermore, the meaning of the partial regression coefficient is no longer valid since the existence of latent variables implies that the coefficients do not represent what is theoretically implied as the net effect of a specific variable on dependent variables.

## 6. Conclusion

This paper began with mentioning the nonstationarity of data and has demonstrated that the data series used in the original Crandall study are nonstationary and need to be differenced before credible model estimates can be generated. The nonstationarity of data in traffic fields have often been ignored by many researchers.

A new way of modeling based on structural causal relations Co has been introduced in the area of traffic fatalities. This procedure is applied to the data used in the model given in Crandall. Concentrating on the variables in Crandall's model and assuming causal sufficiency, occupant death is found to be caused by safety device, income, and vehicle mileage. Further, vehicle mileage is a function of income and rural driving. The parameters were estimated underlying the directed graph were estimated using three stage least squares estimation. The system of equations modeling is advocated, but is not supported specifically by Crandall and Graham.

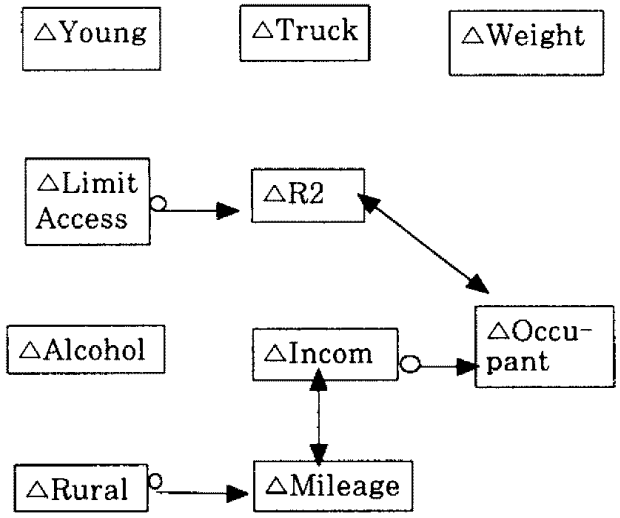



Fig. 2 Causal Graphical Representation of Crandalls Occupant Death Model Based on 1947-1981 Differenced Data and 10 % Significance Level.

For the estimation and forecasts, only three variables for occupant deaths out of the ten original variables were used in Crandall's model. This makes the difference between the directed graph model and Crandall's model. It is demonstrated that the directed graph model results in superior out-of-sample forecasts when compared to forecasts

from Crandall's original model. Better forecasts to original data implies that modeling based on the directed graph is closer to the true model of occupant death.

Furthermore, it has been shown that latent variables may exist between both safety device and income and occupant death. This result makes inference based on regression analysis further suspect. The DAG of nosufficiency case demonstrated that we need try to find more direct data for occupant death and that the inferences based on the regression model will be misleading.

## 7. 요 약

본 논문은 Crandall의 탑승자 사망에 관한 모형에 Directed Graph를 응용한 것으로써 데이터는 Crandall이 사용한 미국의 1947-1981 기간의 탑승자 사망 데이터를 1993년까지 확장한 것을 사용하였다. Directed Graph Algorithm방법은 최근에 컴퓨터과학 분야에서 발전된 것을 원용한 것이다. 먼저 1947-1981 기간의 데이터를 기초로 하여 회귀분석을 통한 분석 대신에 Directed Graph Algorithm을 이용한 결과, 회귀분석을 이용했던 Crandall의 결과와는 달리 탑승자 사망은 소득수준, 자동차의 운행거리, 자동차의 안전장치 수준에 의하여 직접적으로 결정이 되는 것으로 나타났다. 자동차의 운행거리는 소득수준과 시내주행에 대한 교외주행 의 비 에 의해서 결정되는 것으로 나타났다. 이런 결과에 근거하여 3SLS (three stage least squares regression)를 이용하여 추정하고, 이러한 추정에 근거하여 1982-1993 기간을 예측했으며, Crandall의 원래의 모형의 예측력과 비교를 하였다. 예측 결과 본 모형이 MSE( mean squared error)를 기준으로한 예측력에서 훨씬 뛰어난 결과를 보였다. 더욱 중요한 것은 본고에서는 Crandall이 사용한 변수간에 기존의 계량적 방법으로는 색출이 불가능했던 잠재변수 (Latent variable)가 존재함을 구체적으로 보임으로써 회귀분석을 통한 모형화는 진정한 변수간의 관계를 반영치 못함을 보인 것이다.

## REFERENCE

1) Ashley, R.; Granger C.W.J. and Schmalensee, R. Advertising and Aggregate Consumption an Analysis of Causality. *Econometrica*, 1969, *21*, pp. 243-247.
2) Bureau of Economic Analysis, U.S. Department of Commerce. *Survey of current business*, Washington, DC: U.S. Government Printing Office, various years.
3) Crandall, R., and Graham, J. D. Automobile Safety Regulation and Offsetting Behavior: Some New Empirical Estimates. *American Economic Review*, 1984, *74*, pp.328-331.
4) Dickey, David and Fuller, Wayne. Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root. *Econometrica*, 1981, *49*, pp.1241-69.
5) Federal Highway Administration, U.S. Department of Transportation, *Highway Statistics*, Washington, DC: U.S. Government Printing Office, various years.
6) Granger, C. W. J., and Newbold, Paul. Spurious Regressions in Econometrics. *Journal of Econometrics*, 1974, *2*, pp. 111-120.
7) National Institute of Health, U.S. Department of Health and Human Services, *Surveillance Report #35*, December 1995.
8) National Safety Council, Accidents Facts, Itasca, IL, various years.
9) Peltzman, S. The effects of automobile safety regulation. *Journal of Political Economy*, 1975, *83*, pp.677-725.
10) _____. The effects of automobile safety regulation. *Accid. Anal. Prev.*, 1976, *8*, pp.139-142.
11) Spirtes, Glymour, and Scheines, *Causation, Prediction and Search*, New-York : Springer-Verlag, 1993.