ICIC Express Letters
Part B: Applications
Volume **9**, Number **5**, May **2018**

ICIC International ©2018 ISSN 2185-2766
pp. **461–468**

CAUSE AND EFFECT RELATIONSHIP ANALYSIS USING CHAIN INDEPENDENT GRAPH

Eri Domoto[1], Antonio Oliveira Nzinga Rene[2] and Koji Okuhara[3]

[1]Department of Media Business
Faculty of Economics
Hiroshima University of Economics
5-37-1, Gion, Asaminami-Ku, Hiroshima City, Hiroshima 731-0192, Japan
er-domo@hue.ac.jp

[2]Department of Informatics
Faculty of Engineering
Kindai University
1 Takaya Umenobe, Higashi-Hiroshima City, Hiroshima 739-2116, Japan
nzingar3@hiro.kindai.ac.jp

[3]Graduate School of Engineering
Toyama Prefectural University
5180 Kurokawa, Imizu-shi, Toyama 939-0398, Japan
okuhara@pu-toyama.ac.jp

Received November 2017; accepted February 2018

Abstract. *In this paper, we analyze student data using graphical modeling. One of the ultimate goals of college students is the employment placement. Therefore, we studied factors affecting employment, and through an independent chain graph the inference of causality is considered.*
Keywords: Graphical modeling, Data mining, Education data analysis

1. **Introduction.** Graphical modeling [1-4] is a method used for modeling dependency relationship of various random variables through graphs. In graphical modeling, most statistical models such as regression analysis, factor analysis, SEM, signal detection theory, hidden Markov model, and path analysis, can be expressed uniformly under this model. In this paper, using a chain independent graph [5-8] reasoning about causality, we analyze factors influencing employment based on student data which is considered as numerical example. As factors, we considered high school deviation value, entrance examination type, Grade Point Average (GPA), scholarship, project club. To each of the five departments, and the whole we performed a graphical modeling and expressed the result. Thus, we could understand which factors affect the probability of employment placement. Moreover, through graphs one can visualize the degree of influence of such factors.

2. **Outline of Graphical Modeling.** Graphical modeling graphs are not simply visual representations, but graphs in graph theory, which is a discrete mathematical field. This graph consists of several vertices and lines and arrows connecting them and describes some relations between vertices by lines and arrows. When considering a statistical model corresponding to a graph obtained by taking variables as vertices, the model is called a graphical model.

2.1. **Analysis of multidimensional quantitative data.** The correlation matrix is expressed as $R = (r_{ij})$, and its inverse matrix is expressed as $R^{-1} = (r^{ij})$. Divide off-diagonal elements of the inverse matrix by the square root of the corresponding two diagonal elements and normalize and minus: $r_{ij\cdot\mathrm{rest}} = \frac{-r^{ij}}{\sqrt{r^{ii}}\cdot\sqrt{r^{jj}}}$.

The left side obtained by this calculation is called partial correlation coefficient when the remaining variables i and j are given. rest of $r_{ij\cdot\mathrm{rest}}$ means "remaining". If this value is 0, variable i and variable j are uncorrelated when the values of the remaining variables are fixed. We derive this partial correlation coefficient for all pairs of variables and summarize them in matrix form called partial correlation matrix. A graph created based on whether or not the partial correlation coefficient is 0 is called an independent graph.

2.2. **Analysis of multidimensional qualitative data.** In the case of quantitative data, the means to search for entanglement of three or more variables was in the partial correlation matrix. In the case of qualitative data, there are no direct statistical statistics.

First, the probability that a respondent will enter the α_i, β_j, γ_k, δ_l ($i = 1, 2, 3$; $j = 1, 2$; $k = 1, 2$; $l = 1, 2$) of the cell using p_{ijkl} is as follows.

$$\sum\nolimits_i \sum\nolimits_j \sum\nolimits_k \sum\nolimits_l p_{ijkl} = 1$$

For the logarithm $\log p_{ijkl}$ of p_{ijkl}, consider a model similar to the structural model of no repetitive quaternion variance analysis.

$$\begin{aligned}\log p_{ijkl} = {} & \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\alpha\delta)_{il} + (\beta\gamma)_{jk} + (\beta\delta)_{jl} \\ & + (\gamma\delta)_{kl} + (\alpha\beta\gamma)_{ijk} + (\alpha\beta\delta)_{ijl} + (\alpha\gamma\delta)_{ikl} + (\beta\gamma\delta)_{jkl} + (\alpha\beta\gamma\delta)_{ijkl}\end{aligned}$$

Such a model is called a log-linear model. Here, as in the case of the structural model of the variance analysis, in all the main effect term, the two-factor interaction term, the three-factor interaction term, and the four-factor interaction term, the constraint is that the sum of the subscripts is 0.

$$\sum\nolimits_i \alpha_i = \sum\nolimits_j \beta_j = \sum\nolimits_k \gamma_k = \sum\nolimits_l \delta_l = 0$$

$$\sum\nolimits_i (\alpha\beta)_{ij} = \sum\nolimits_j (\alpha\beta)_{ij} = \cdots = \sum\nolimits_l (\alpha\beta\gamma\delta)_{ijkl} = 0$$

In the analysis of variance, a model in which some interaction terms and main effect terms are set to 0 is adopted according to the result of the test. This process is repeated, but at this time, for example, when leaving the two-factor interaction term $(\alpha\beta)_{ij}$ in the model without setting it to 0, a policy of leaving the main effect terms α_i and β_j included therein in the model is widely accepted. Otherwise, interpretation of the parameters becomes difficult. More generally, when an interaction term exists, the lower order interaction term and the main effect term included therein are also present, and the model satisfying this is called a hierarchical model. Usually, we often select models only for this hierarchical model. As a further partial class of this hierarchical model, there is a class called graphical model. For graphical modeling, we select models only for this graphical model.

3. **Outline of Chain Independent Graph Modeling.** Graphical modeling has several variations. To perform graphical modeling of independent chain graphs, it is necessary to be able to assume to some extent the order relation of causality between variables. "To some extent" means that variables are grouped into several hierarchies ranging from the causal system to the results, and the causal order is established among the groups, but the order of the causality among the variables in one group is ambiguous. In many cases, it is not an unreasonable requirement to assume this degree before analyzing. The goal of modeling in these situations is to obtain a "chain independent graph". An independent chain graph is a graph where causal relationships between variables belonging to different groups are represented by arrows and conditionally independent relations among variables

in the same group are represented by lines. It is a causal graph between the groups and an independent graph in the group. The procedure for creating an independent chain graph is described as follows. Hereinafter, the variable group is referred to as the first group, the second group, ... from the cause system side.

3.1. **Chain independent graph modeling procedure.**

(1) Graphical modeling is performed only in the first group, and an independent graph is obtained.
(2) Perform graphical modeling with the variables of the first group and the second group. However, the partial correlation between the first groups is not zero. The result is expressed as an independent chain graph connecting variables not having the partial correlation of 0 by lines/arrows.
(3) Replace the part in the first group of the chain independent graph obtained in 2 with the independent graph obtained in 1.
(4) Incorporate the variable group of the resulting series one by one, and perform graphical modeling in which the partial correlation between the variables of the cause system is not set to 0, and obtain the part of the causal system of the obtained chain independent graph at the previous stage. It replaces it with a graph.

3.2. **Evaluation in a chain independent graph.**

(1) Evaluate the goodness of fit of each stage by x^2 test.
(2) Evaluate the fitness of the model as a whole, the deviation degree of each stage as the total deviation degree, the total degree of freedom at each stage as the total degree of freedom, and evaluate by x^2 test.

4. **Numerical Example.**

We analyze student data using graphical modeling. A dataset related to 578 students in 5 departments is used. Details of the data are six categories of high school deviation value, entrance examination classification, scholarship, Grade Point Average (GPA), project club, employment place. High school deviation values were classified as (1) 55 or more, (2) 50 to 55, (3) 45 to 50, (4) 40 to 45, (5) 35 to 40, (6) unknown. The entrance examination classification was classified as (1) General Entrance Exam (Previous Period Entrance Examination), (2) General Entrance Exam (Latter Period Entrance Examination), (3) Entrance Exam for General Public Advertisement Recommended, (4) Special Designated School Recommendation Entrance Examination, (5) Designated School Recommendation Entrance Examination, (6) Designated Club Recommendation Entrance Examination, (7) Entrance Exam for Special Recommended (Sports Field), (8) Entrance Exam for Special Recommended (Specific Qualification Field), (9) Admissions Office Entrance Exams, (10) Entrance Exam for International Students, (11) University Center Examinations (Term 1, 2), (12) Entrance Exam for Working People. GPA was classified as (1) 3.0 or more, (2) 2.5 to 3.0, (3) 2.0 to 2.5, (4) 1.5 to 2.0, (5) 0.0 to 1.5. Scholarships were classified as (1) get, (2) not get. In the project club, we classified it as follows: (1) project affiliation, (2) belonging to the activities of university relations, (3) exercise club, (4) culture club, (5) no affiliation. Here, the project is an organization that tackles what is needed in society. Employers were classified as (1) employment place the university wanted to get a job, (2) public servant, (3) other companies, (4) no employment.

Each department name is A to E, and the number of people is A = 39, B = 48, C = 68, D = 254, E = 329. We conducted graphical modeling for each department and the whole and analyzed which factors were affecting employment. Figures 1 to 5 show the results of graphical modeling for each department. Also, Figure 6 shows the results of graphical modeling for all departments. Here, a chain independent graph is used. Causal order setting was as follows.

First class: High School Deviation Value

Second class: Type of Entrance Examination, GPA, Scholarship, Project Club

Third class: Place of Employment

Path coefficients which are numerical values indicating the magnitude of the causal influence are given to the one-sided arrow representing the causal influence. In addition, the number of correlation stations is given to double-sided arrows indicating that there is a correlation. Arrows are displayed with a line of a thickness proportional to the pass coefficient, double arrows are indicated by a slightly thick line of light color, and the numerical values of the correlation coefficients given to the double arrow are also displayed in bold type in light color. Also, when the correlation coefficient and the pass coefficient are negative, the arrow is indicated by a broken line, and the numerical value is indicated by a minus. A small speech marked "e" represents the residual (variation due to all factors outside the model). The variable with residuals is given a numerical value of "decision coefficient (R^2)" indicating what percentage of variation of the variable is explained by the causal influence in the model. Table 1 shows the goodness-fit-index list. Good-of-fit of each graph is confirmed based on this table.

Figure 1 shows the result of graphical modeling of department A. Goodness-of-fit is as follows. $x^2 = 1.079$, df = 7, p = 99.3%, $x^2/\text{df} = 0.154$, GFI = 0.991, AGFI = 0.972, SRMR = 0.036, AIC = −12.9, RMSEA = 0.000, NFI = 0.949, CFI = 1.000. From these

Table 1. Goodness-of-fit index list

Index	Range	Good range	Bad range
x^2	$x^2 \geqq 0$	Judge by P value	Judge by P value
GFI	GFI $\leqq 1$	0.95 or more	Less than 0.9
AGFI	AGFI $\leqq 1$	0.95 or more	Less than 0.9
SRMR	SRMR $\geqq 0$	Less than 0.05	0.1 or more
AIC	No limit	Relative comparison	Relative comparison
RMSEA	RMSEA $\geqq 0$	Less than 0.05	0.1 or more
NFI	$0 \leqq$ NFI $\leqq 1$	0.95 or more	Less than 0.9
CFI	$0 \leqq$ CFI $\leqq 1$	0.95 or more	Less than 0.9

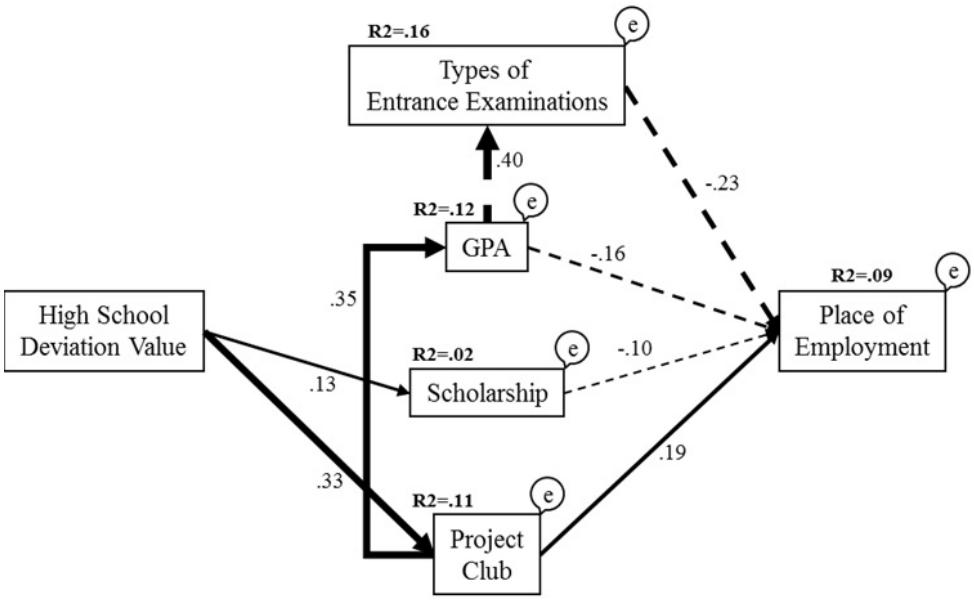

Figure 1. Results of graphical modeling of department A

goodness-of-fit, it is considered that the model's fitness is high. From the figure, the following can be seen.

- High school deviation value is high → It belongs to a project or club
- High school deviation value is high → Receive scholarship
- Belong to a project or club → High GPA
- Belong to a project or club → Place of employment is good

Figure 2 shows the result of graphical modeling of department B. Goodness-of-fit is as follows. $x^2 = 0.696$, df = 6, p = 99.5%, $x^2/\text{df} = 0.116$, GFI = 0.995, AGFI = 0.983, SRMR = 0.023, AIC = −11.3, RMSEA = 0.000, NFI = 0.942, CFI = 1.000. From these goodness-of-fit, it is considered that the model's fitness is high. From the figure, the following can be seen.

- High school deviation value is high → It belongs to a project or club
- High school deviation value is high → Receive scholarship
- Receive scholarship → High GPA
- High school deviation value is high → Place of employment is good
- Differences in entrance examination classification → GPA changes
- Belong to a project or club → Place of employment is good
- High GPA → Place of employment is good

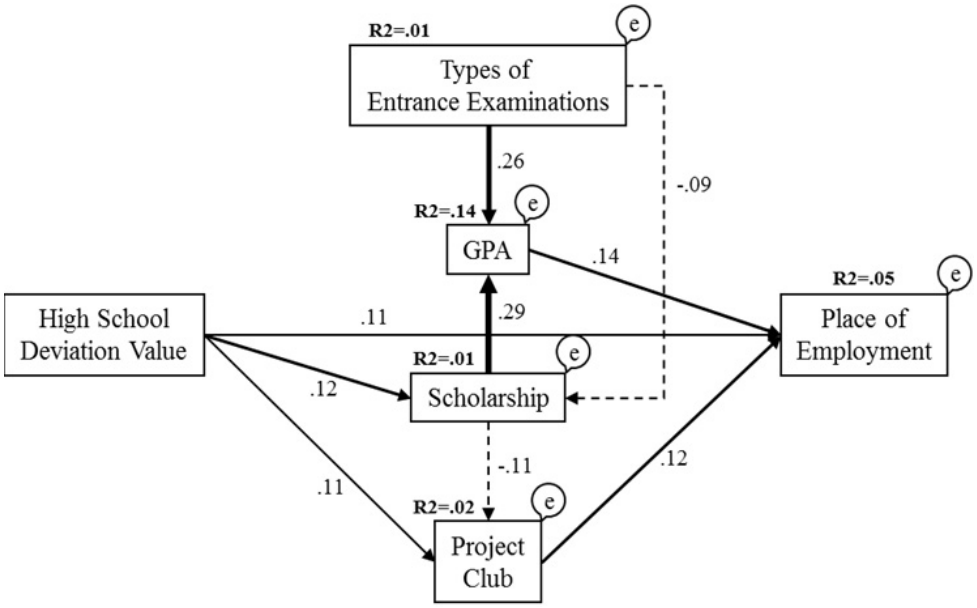

FIGURE 2. Results of graphical modeling for department B

Figure 3 shows the result of graphical modeling of department C. goodness-of-fit is as follows. $x^2 = 1.869$, df = 7, p = 96.7%, $x^2/\text{df} = 0.267$, GFI = 0.990, AGFI = 0.971, SRMR = 0.035, AIC = −12.1, RMSEA = 0.000, NFI = 0.921, CFI = 1.000. From these goodness-of-fit, it is considered that the model's fitness is high. From the figure, the following can be seen.

- High school deviation value is high → Receive scholarship
- Belong to a project or club → Receive scholarship
- Belong to a project or club → Place of employment is good
- High GPA → It belongs to a project or club
- Belong to a project or club → Place of employment is good
- High GPA → Place of employment is good

Figure 4 shows the result of graphical modeling of department D. goodness-of-fit is as follows. $x^2 = 4.634$, df = 9, p = 86.5%, $x^2/\text{df} = 0.515$, GFI = 0.994, AGFI = 0.986,

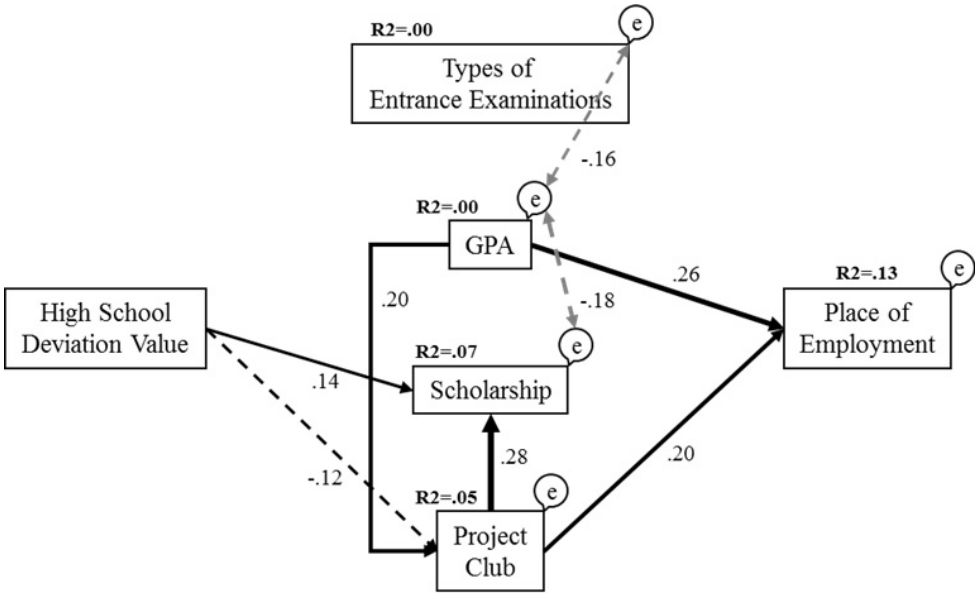

FIGURE 3. Results of graphical modeling for department C

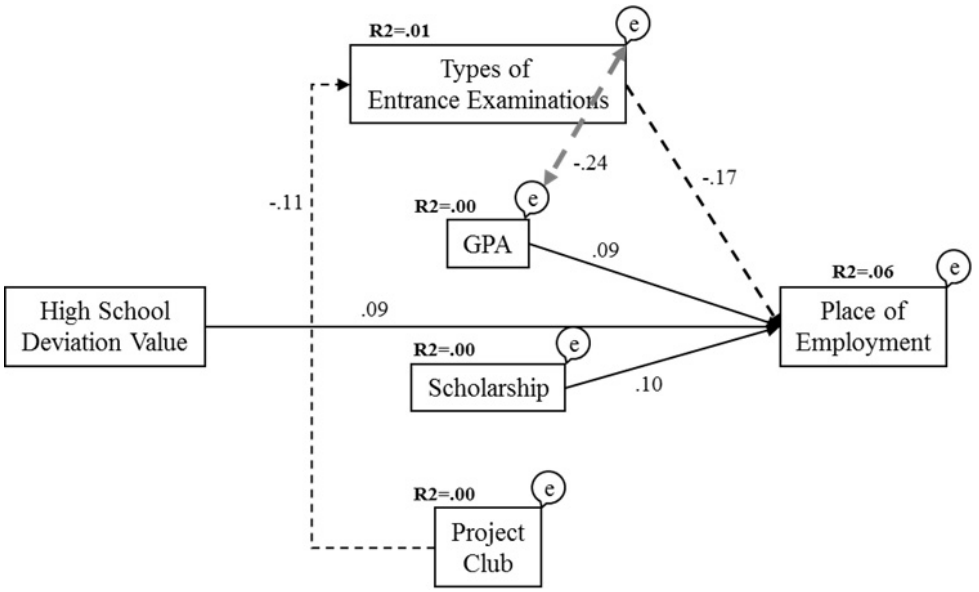

FIGURE 4. Results of graphical modeling for department D

SRMR = 0.030, AIC = −13.4, RMSEA = 0.000, NFI = 0.882, CFI = 1.000. From these goodness-of-fit, it is considered that the model's fitness is high. However, it is judged that the value of NFI is basically 0.9 or more, it is judged that the Goodness-of fit is high, but it is found that NFI = 0.882 is not good a little. From the figure, the following can be seen.

- High school deviation value is high → Place of employment is good
- High GPA → Place of employment is good
- Receive scholarship → Place of employment is good

Figure 5 shows the result of graphical modeling of department E. Goodness-of-fit is as follows. $x^2 = 11.48$, df = 12, p = 45.9%, x^2/df = 0.986, GFI = 0.988, AGFI = 0.980, SRMR = 0.041, AIC = −12.2, RMSEA = 0.000, NFI = 0.612, CFI = 1.000. As you can see from the figure and NFI = 0.612, it is a graph close to the independent model. Since this graph has a low degree of conformance, it can be understood that it is not very helpful.

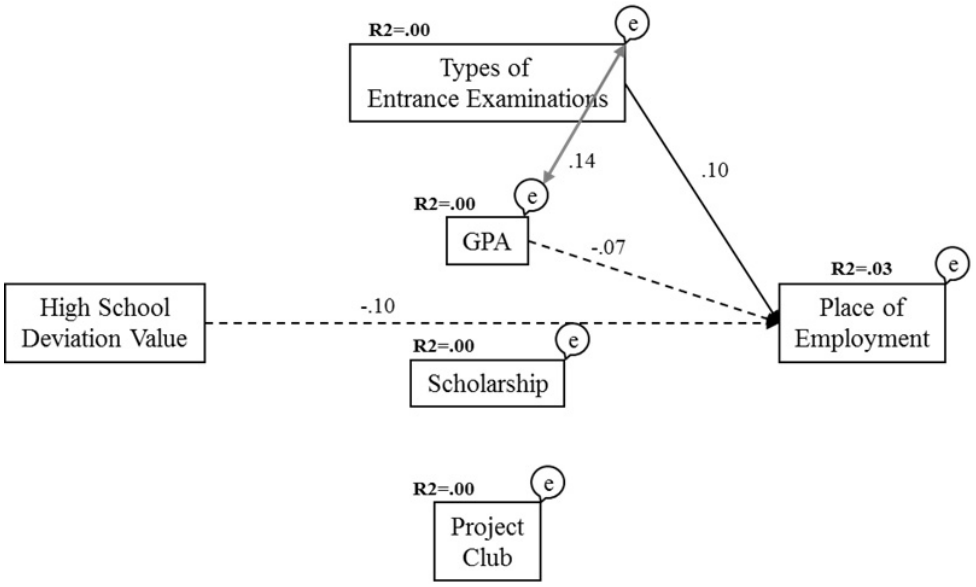

FIGURE 5. Results of graphical modeling for department E

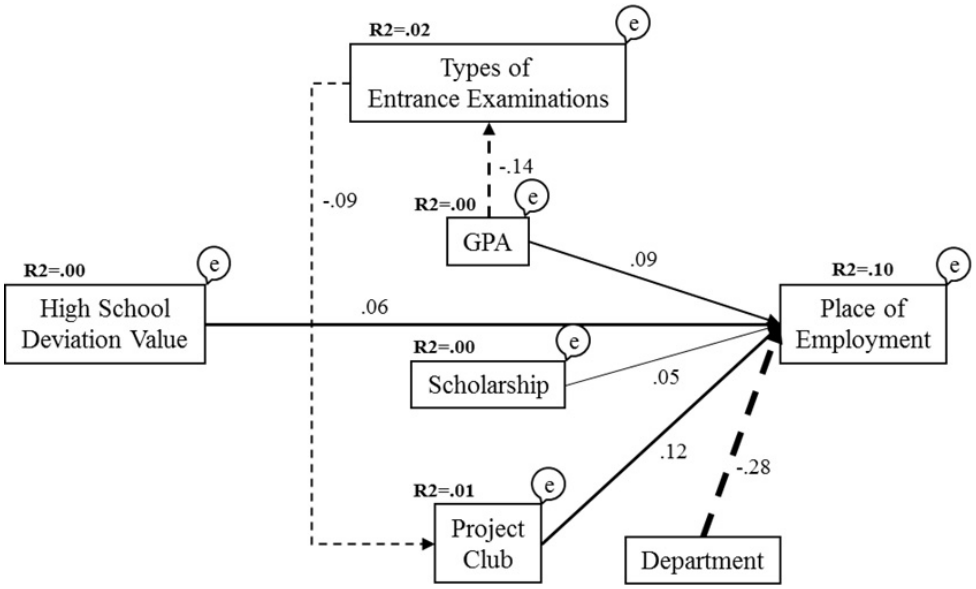

FIGURE 6. Results of graphical modeling for all departments

Figure 6 shows the result of graphical modeling of all departments. Goodness-of-fit is as follows. $x^2 = 16.19$, df = 14, p = 30.2%, $x^2/\text{df} = 1.156$, GFI = 0.94, AGFI = 0.987, SRMR = 0.029, AIC = −11.8, RMSEA = 0.0146, NFI = 0.861, CFI = 0.977. From these goodness-of-fit, it is considered that the model's fitness is high. Although it is NFI = 0.861, we will consider it because it has a large number of samples. From the figure, the following can be seen.

- High school deviation value is high → Place of employment is good
- High GPA → Good job placement
- Receive scholarship → Place of employment is good
- Belong to a project or club → Place of employment is good

5. **Conclusion.** In this paper, an analysis was performed using a chain independent graph among graphical modeling. In the numerical example, we analyzed factors influencing employment on the student data. It turned out that the project or club, GPA

had a big influence on employment. Then it turned out that the scholarship, high school deviation value affect employment. Moreover, it was found that the high school deviation value has a big influence on scholarship. In addition, it turned out that it was influenced as follows. The high school deviation value → GPA, project or club, project or club → GPA, scholarship, type of entrance examination → GPA, scholarship → GPA, GPA → project or club.

Acknowledgment. This work is supported by JSPS KAKENHI Grant Number 253503 09.

REFERENCES

[1] M. Tsubaki, Y. Matsuda, Y. Doi and Y. Noguchi, A study on the relations between elements to improve necessary abilities required to function as a member of the society of university students of science and technology, *Kodo Keiryogaku (The Japanese Journal of Behaviormetrics)*, vol.39, no.1, pp.11-32, 2012.

[2] M. Miyagawa, *Graphical Modeling*, Asakura Publishing Co., Ltd., 1997.

[3] N. Shoji and T. Kojima, A study on the negative image as use prevention factor of the public library, *Journal of Architecture and Planning (Transactions of AIJ)*, vol.77, no.681, pp.829-836, 2012.

[4] K. Hino, R. Manabe and T. Kojima, Development of an information provision system about crime situation through WebGIS and analysis of citizens' attitude: A study of WebGIS system supporting citizens' action against crime, *Journal of Architecture and Planning (Transactions of AIJ)*, vol.70, no.597, pp.135-140, 2005.

[5] S. Kubota and A. Shinozaki, What expanded U.S. imports of services?: Graphical modeling analysis of personal network and income level, *InfoCom Review*, vol.67, pp.34-43, 2016.

[6] Y. Horimoto, H. Maruyama and K. Kurosawa, Analysis of personality traits influencing clinical education: Addressing issues prior to clinical training, *Journal of Exercise Physiology*, vol.26, no.4, pp.541-547, 2011.

[7] T. Kojima, N. Wakabayashi and K. Hirate, Exploratory modeling for causality in hierarchical structure of evaluation: Causality analysis on environmental evaluation Part 2, *Journal of Architecture and Planning (Transactions of AIJ)*, vol.67, no.556, pp.77-82, 2002.

[8] T. Kojima and M. Yamamoto, *Covariance Structure Analysis*, Ohmsha, 2013.