Beyond Single Equation Regression Analysis: Path Analysis and Multi-Stage Regression Analysis

Jeonghoon Ahn

School of Pharmacy, University of Maryland, 100 North Greene Street, 6th Floor, Baltimore MD 21201

Multi-stage regression analysis and path analysis provide important complements to the traditional regression analysis. Although regression (covariance) analysis is a useful and common multivariate analysis methodology in pharmacy and many other sciences, there is a problem of limited measurability that only the direct effects of included independent variables can be captured. Furthermore, the traditional regression analysis might yield biased estimates because of the ignored indirect effects in some cases: the compliance and effectiveness studies; the cost of illness studies; or the patient reported outcomes (PRO) studies. Therefore, multi-stage regression analysis can provide not only a refinement of established conclusions but also a significant improvement to regression analysis. The main purpose of this paper is to highlight the usefulness of multi-stage regression models and path analysis models in a pharmaceutical research setting. This paper can be also used as an introduction to these two models in a research methodology class.

INTRODUCTION

Regression analysis, or covariance analysis, is a popular multivariate analysis methodology in many sciences including pharmacy administration. Since regression analysis is a more general methodology than Analysis of Variance (ANOVA)[1], it has been used for a more complex problem, typically in a multivariate environment(1,2). In other words, the strength of regression analysis is the ability to capture multiple relationships simultaneously, while providing a simple and fast estimation result. For example, an effectiveness study of a new drug may need to consider multiple factors related to the effectiveness, but regression analysis methodology can deal with all these factors simultaneously[2] or by a single regression equation, as long as they are observable.

In cases where the indirect factors play an important role, regression analysis is not suitable. In the example of the effectiveness study above, regression analysis can capture all the direct effects from the included factors, but it cannot deal with any indirect effects coming from causal relationships among the factors included. For example, compliance is a typical latent factor in a RCT of drug effectiveness and there is a causal structure related to the effectiveness and compliance. There can be some demographic factors or any other factors affecting both effectiveness and compliance rate but these factors might appear as "not significant" in the regression analysis of effectiveness once they are included in the regression with compliance rates. Since the variations of these factors are already explained by the compliance rate variable, we get such insignificances. This result implies that there are no significant direct effects of those demographic factors to the effectiveness of the drug but we are not sure about any indirect effects through the compliance rate. In an extreme case, exactly the opposite thing can also happen. Because of the correlation between the compliance rate and some demographic factors, the compliance rate might seem insignificant in the effectiveness regression. In this case, a similar explanation is also possible: the variation of the compliance rate is mostly explained by the included demographic factors.

There can be many different remedies for the aforementioned problem, but we are more interested in the alternatives which can also provide implications of compliance rate. In the first case where compliance rate explains the variations of demographic factors and these indirect factors are not important, regression analysis does not need any other consideration. We can conclude that there is no significant direct effect from these demographic factors to the effectiveness variable. For the second case, if the compliance rate appeared to be insignificant, there are many solutions suggested. One solution is multiplying compliance rate to the effectiveness regression. This is based on the judgment that compliance rate is independent of any factors included in the effectiveness regression. If the compliance rate is not independent to the effectiveness variable, this method is invalid. Another solution is to exclude those demographic factors in the effectiveness regression analysis. This is a typical solution of the regression analysis problem called "multicolinearity." In addition, there is a data dependent modeling method regardless of the first case or the second case. Efron and Feldman used the Lipid Research Clinics Coronary Primary Prevention Trial (LRC-CPPT) data to estimate the dose-response curve adjusted for the differences in the compliance rate between the treatment group and the controlled group(3). Finally, the two proposed models in this paper

---

[1] This is the reason why regression analysis is also called "Analysis of Covariance" (ANCOVA). Note that this term connotes ANOVA since a covariance with itself is a variance. See Neter, Wasserman,, and Whitemore for the details and examples(2).

[2] Mathematically, regression analysis is a projection. Projecting a subject (the effectiveness) on the screen (all the factors considered) is a regression analysis. Therefore, the screen is important to get a correct image of the subject.

Am. J. Pharm. Educ., **66,** 37–42(2002); received 9/21/01; accepted 1/10/02.

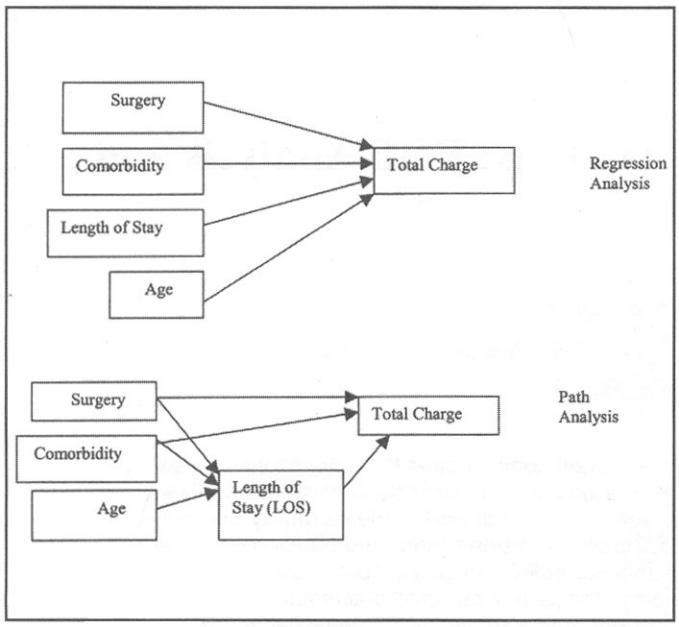

Fig. 1. Regression analysis and path analysis.

can provide solutions for both cases.

This paper is organized as follows: the next two sections (Section 2 and Section 3) each explain path analysis model and multi-stage regression model, respectively; Section 3 continues with the compliance and effectiveness example in terms of path analysis model and multi-stage regression model; and Section 5 concludes with remarks.

PATH ANALYSIS MODEL

Path analysis model has a long history. It started in the 1930s as a method of studying direct and indirect effects of variables while regression analysis model remains as a method of discovering causal relationships(4). Also path analysis model is not a substitute of regression analysis, rather it is a complementary methodology to regression analysis. A set of additional regressions is added to the original regression analysis to trace out indirect effects. Because of this complexity, a path diagram is typically used to display all of the causal relationships.

As we can see from Figure 1, a path analysis separates direct effects and indirect effects through a medium variable while regression analysis considers direct effect only. In regression analysis, Factor A, Factor B, and Outcome A are all independent variables for a dependent variable Outcome B, but path analysis includes an additional regression of Outcome A on Factor A and Factor B. This regression can be simply done as a separate regression analysis since path analysis assumes that each regression is independent of the others.[3] If we found a significant relationship between Outcome A and either Factor A or Factor B, it can be interpreted as an indirect effect to Outcome B through Outcome A (given that Outcome A is significant to Outcome B).

Path analysis further assumes a unit variance for all the variables to allow the comparison of the magnitudes of each variable included. For example, the SAS® Procedure CALIS provides standardized[4] coefficients in the output. Using the standardized path coefficients, we can compare the magnitudes of each factor including both direct and indirect effects. Additional programming help in path analysis can be found in Hatcher's book(5).

Example 1. Health Outcomes Research of a Surgery

(a) Length of Stay = α_S Surgery + α_C Comorbidity + α_A Age + Error 1

(b) Total Charge = β_S Surgery + β_C Comorbidity + β_{LOS} Length of Stay + Error 2

Estimation results by a path analysis[5]:

Length of Stay = 0.5 Surgery + 0.5 Comorbidity + 0.4 Age

Total Charge = 0.3 Surgery + 0.3 Comorbidity + 0.4 Length of Stay

If the above example result is obtained by a path analysis, we know that the direct effect of Surgery on Total Charge is 0.3 but there is also indirect effect of 0.4 x 0.5 = 0.2 that exists, and the total effect on Total Charge is 0.3 + 0.2 = 0.5.

MULTI-STAGE REGRESSION ANALYSIS

Multi-stage regression analysis was originally developed as an estimation method for a widely used modeling scheme in economics called Simultaneous Equation Modeling (SEM). A SEM consists of multiple equations and each equation is related to the others by either endogenous variables or correlated error terms. On the other hand, as mentioned in the previous section, path analysis does not allow any correlation among the error terms. The original modeling insight of SEM arose from economic theory of the markets and equilibrium, which requires the simultaneous determination of economic variables. Even though many econometric models are based on the existing statistical models, SEM is one of the most remarkable developments in econometrics since the early 1940s(6). This model was further developed as "structural equation modeling"[6] in sociology and psychology(7).

The first multi-stage estimation method developed was Two-Stage Least Squares (2SLS) by two independent researchers(8,9). To eliminate the correlation between the error term and a problematic independent variable[7], 2SLS estimates the predicted variables of dependent variables from all the equations in the first stage and substitutes any problematic independent variable with its predicted variable in the second stage estimation. This substitution idea is essentially the same as the Instrumental Variable (IV) method(10). Since the error structure of multi-stage analysis is more general, multi-stage regression results are more robust (*i.e.,* less vulnerable) to possible correlations among the error terms than path analysis results. If we consider error terms as unobserved noise, it

[3] This independence assumption is an important distinction from multi-stage regression.

[4] This is simply converting each independent variable to have a unit variance (divide each variable by its own standard deviation). Note that it is different from "standardization" of regression analysis (STB option of MODEL statement in the procedure REG of SAS®), which makes dependent variable to have a unit variance. Also there is a term called "studentization," which makes residuals to have a unit variance (STUDENT option of OUTPUT statement in the procedure REG of SAS®).

[5] Typical path analysis output is standardized one (see the previous footnote), but we consider raw output here for the sake of comparison with multi-stage regression later.

[6] Structural equation model is a more generalized model from simultaneous equation model since the former allows errors in variable or multiple indicators for a latent construct while the latter does not. For more details on latent variables modeling for path analysis and structural equation models, Bollen is an excellent text(1).

[7] If this independent variable is the dependent variable in another equation, it is also called as endogenous variable. If an independent variable is considered as predetermined or not influenced by any other variable in the model, it is called as exogenous variable.

Table I. An example of three stage regression analysis for variables influencing total charges undergoing hysterectomy for endometrial carcinoma[a]

Variable	Coefficient β	Std Error	P^b
Length of Stay (per day)*	$1,388.61	155.67	0.000
Age (total charge per year of life)	$112.72	16.013	0.000
Age Squared (accounts for non-linear relationships of age and charges)	-$1.22	0.17	0.000
Charlson Comorbidity Index (per each comorbidity point)	$347.51	40.26	0.000
Cardiac complications	$2,949.01	529.45	0.000
Respiratory complications	$24,980.44	2,151.15	0.000
Infection	$2,406.19	565.12	0.000
Pulmonary embolus	$5,158.93	1,876.24	0.006
Anemia due to blood loss	$2,475.56	500.18	0.000
TVH[d] and Lymph Nodes	$4,773.51	1,253.39	0.000
Radical hysterectomy	$2,252.34	745.65	0.003
Lymph node dissection	$1,617.58	260.46	0.000
Nervous system disorder	$3,559.71	960.49	0.000
Chronic respiratory disease	$2,413.73	528.67	0.000

[a]Table 4 from Brooks *et al., Cancer,* Vol. 92, No. 4, 2001, p. 955. Copyright © 2001 American Cancer Society. Reprinted by permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.

[b]$R^2 = 0.7442$; Overall significance ($P = 0.0000$).

[c]Length of stay is from the next regression result in Table II.

[d]TVH= Total Vaginal Hysterectomy.

makes more sense to allow for correlation since those noises are more likely to be correlated for the same individual (observation). For example, a study on patient reported outcomes such as drug abuse and alcohol abuse is more likely to have a correlated structure of the error terms since a patient who hides his drug abuse would be likely to hide his alcohol abuse.

Example 2. Patient Reported Outcome (PRO) study

(a) Drug Abuse = α_1 Alcohol Abuse + α_2 Other DA Factors + Error 1

(b) Alcohol Abuse = β_1 Drug Abuse + β_2 Other AA Factors + Error 2

From Example 2, Error 1 and Error 2 are likely to be correlated since the dependent variables are reported outcomes from the same patients.

The three Stage Least Squares (3SLS) model adds a correction for heteroscedasticity to 2SLS by utilizing Generalized Least Squares (GLS) method.[8] Since 3SLS uses all the information in the system of equations to estimate parameters in each individual equation while 2SLS uses only the information in the specific individual equation to estimate the parameters from the corresponding equation, 3SLS is more efficient than 2SLS.[9] However, 3SLS is also more vulnerable to a specification error since an error from an equation will be transmitted to all the equations(11).

The only difficulty for using multi-stage regression analysis is the identification problem.[10] Since the original purpose of the multi-stage regression model is to estimate SEM, economists developed a method to match the estimated results and the structural model they originally considered. This is similar to the condition needed to solve a system of equations with many unknowns. A greater number of independent equations compared to the number of unknowns is needed to identify all the unknowns. Typically, each equation needs a higher number of unique exogenous variables than the number of endogenous variables included in the equation. SAS® ETS package procedure SYSLIN or procedure MODEL can calculate 2SLS model as long as this identification condition is satisfied. These procedures also provide maximum likelihood estimation options, but it is sometimes difficult to achieve the numerical maximum.

Example 3. (Returning to the Health Outcomes Research in Example 1)

(a) Length of Stay = α_S Surgery + α_C Comorbidity + α_A Age + Error 1

(b) Total Charge = β_S Surgery + β_C Comorbidity + β_{LOS} Length of Stay + Error 2

New Estimation Results by 2SLS:

Predicted (Length of Stay) = 0.5 Surgery + 0.5 Comorbidity + 0.4 Age

Total Charge = 0.2 Surgery + 0.3 Comorbidity + 0.4 Predicted (LOS)

Note that the estimated result of Example 3 is not necessarily same as Example 1, even though Example 3 uses the same variables from Example 1. The above example shows how 2SLS eliminates a possible correlation between Outcome A and the error term of equation (b), Error2. Since the predicted values from the regression are independent of the error term, the predicted variable of Outcome A does not include any randomness, or Error1, which might be correlated with Error2. We can interpret direct and indirect effects similarly to Example 1 (after decomposing the predicted Length of Stay).

An extension of Example 3 can be found in a recent article in *Cancer*(12). The Health Care Utilization Project (HCUP) data from the Agency for Healthcare Research and Quality (AHRQ) is used to compare health service utilization of endometrial cancer patients by the ethnicity and the different

[8] GLS is a method of correcting heteroskedasticity in econometrics. When we do not know the functional form of heteroskedasticity, we can estimate a linear form by OLS and use this form to correct for heteroskedasticity.

[9] In 3SLS information from other equations would be also incorporated into the estimation, while 2SLS does not use such information for estimation.

[10] We are considering a linear system here. Some special non-linear systems may avoid identification process easily, however, this approach will have at least two problems; local minimum and advocacy for the choice of a specific non-linear functional form. For the details, see concluding remarks section.

Table II. An example of three stage regression analysis for variables influencing length of stay (in days) undergoing hysterectomy for endometrial carcinoma[3]

Variable	Coefficient β	Std Error	*Pb*
Age (per year of life)	0.04	0.002	0.000
African American Race	1.19	0.26	0.000
Teaching Hospital	0.60	0.11	0.000
Charlson Comorbidity Index (per each comorbidity point)	0.17	0.01	0.000
Cardiac complications	1.367	0.25	0.000
Hernia/dehiscence	3.33	0.93	0.000
Respiratory Complications	5.60	0.90	0.000
Infectious Complications	2.50	0.19	0.000
Pulmonary complication (respiratory failure)	9.04	0.56	0.000
Anemia due to blood loss	1.19	0.21	0.000
Radical Hysterectomy	0.972	0.33	0.004
TVH and Lymph node dissection	-1.89	0.25	0.000
Circulatory System Disorders	0.51	0.19	0.007
Dementia	3.17	0.58	0.000
Endocrine Disorders	0.42	0.16	0.009
Hypertension	-0.61	0.13	0.000
Nervous System Disorders	1.67	0.428	0.000
Chronic Respiratory Conditions	0.75	0.24	0.002

[a]Table 5 from Brooks *et al., Cancer,* Vol. 92, Mo. 4, 2001, p. 956. Copyright © 2001 American Cancer Society. Reprinted by permission of Wiley-Liss, Inc., a subsidiary of John Wiley & Sons, Inc.
[b]$R^2 = 0.6109$; Overall significance ($P = 0.0000$).
[c]Total abdominal hysterectomy (mean length of stay = 5.19 days) is the reference group in this analysis.

types of hysterectomies. Tables I and II show the estimation results using 3SLS. One of the interesting result in these tables is that the ethnicity (being African American) of patient has an increasing effect on the length of stay but it does not increase total charge additionally. Therefore, there is no direct effect of the ethnicity on the total charge variable but there is an indirect effect of 1.19 days x $1,388.61 ≡ $1,652.45 on total charge.

IMPLICATIONS

The effectiveness of a specific drug might be dependent on several demographic and clinical factors, but regression analysis eliminates the necessity of doing several different comparisons to capture varying degree of effectiveness depending on each factor. More precisely speaking, a regression analysis of effectiveness on a set of observable factors can find each factor's significance conditioning on the given set of observable factors. In statistical phalanx, the variable of interest or the variable to be analyzed is the dependent variable (effectiveness), and factors (demographics) which influence the dependent variable are called independent variables. Hence, regression analysis is especially useful when a dependent variable (such as effectiveness) is correlated with many independent variables (different drug characteristics, demographic information, clinical information, etc.). A regression analysis can generate a result, which summarizes all the pair-wise analyses between the effectiveness variable and each independent variable. To be ideal, each independent variable should represent distinctive characteristic or low correlations with each other. However, we seldom have a real data in this ideal situation. Therefore, path analysis or multi-stage regression analysis can be useful for unraveling the complicated influence structure. This is also related to a complex concept called multicolinearity[11](13).

[11]If there is an exact linear relation exists among independent variables, regression analysis is impossible (exact multicolinearity). If the relationship is near linear, regression result is still best linear unbiased, but explanatory powers of closely related variables are inseparable and some of the variables show insignificance since their explanatory powers are captured by other variable.

Another merit of regression analysis is the statistical prediction on the dependent variable. Once we get the direct effect estimates of each factor (coefficients of each independent variable) on the effectiveness variable, we can predict the average effectiveness varying by a change in each factor. For example, we can predict the effect of dosage on the effectiveness while considering other independent variables (the other factors influencing effectiveness) as well. To get a better prediction, a data with enough variation is essential (we cannot predict a good dosage-response by using one or two dosage levels). A more general model such as path analysis or multi-stage regression model can be helpful for a better prediction since they can model the impacts of each dosage level on some intermediate outcomes with enough variation which may be significant on the effectiveness.

As we can see from Figure 2, dotted arrows cannot be estimated in usual regression analysis. To capture these effects, we need to consider either path analysis or multi-stage regression analysis. Path analysis resolves this issue by running an additional regression of compliance on all the other independent variables(6). On the other hand, two-stage regression analysis introduces an additional regression model of compliance similarly, but further generalizes the model with a general correlation structure between the error terms of two regressions, ε_1 and ε_2:

Example 4. Effectiveness and Compliance

$\text{Effectiveness}_i = \alpha + \beta\ \text{Common IndVars}_i + \varepsilon_1 + \beta_2\ \text{Compliance}_i + \varepsilon_1$

$\text{Compliance}_i = \gamma + \delta_1\ \text{Common IndVarS}_i + \delta,\ \text{Specific Covariates}_i. + \varepsilon_2$

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim \text{MN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma \right)$$

Path Analysis Assumption: ε_1 and $\varepsilon_2$are independent (Σ is a diagonal matrix)

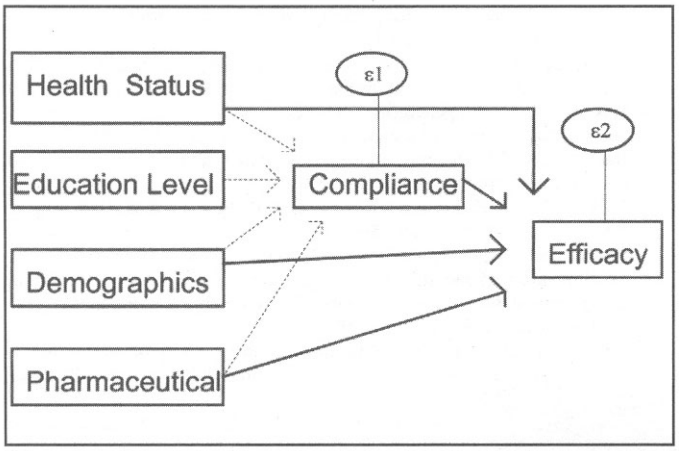

Fig. 2. Effectiveness and compliance.

Multi-Stage Regression Assumption: D is a general covariance matrix.

where *Common IndVars* indicates the Independent Variables such as Health Status, Demographics, Pharmaceuticals, etc. MN stands for multivariate normal distribution. Σ is a 2x2 covariance matrix of error terms. Therefore, path analysis model is a nested version of two-stage regression model with a special covariance structure.[12] Multi-stage regression is more appealing if the measurement error of compliance is related to the measurement error of effectiveness for each individual. In that case, the correlation between two error terms will not be zero.

A path analysis model can add causal relationship structure to the regression analysis and capture both indirect effects and direct effects. Furthermore, as in Example 1, there can be a substantial difference between direct effect and total effect including indirect effect. The selection of path analysis model is more dependent on the usefulness of this information. If a researcher believes it is enough to consider only the direct effect proxies, path analysis does not need to be used. However, a multi-stage regression analysis has some important advantages, which regression analysis or path analysis cannot provides. For instance, the compliance rate is typically found to be a significant factor for the effectiveness variable. However, this high impact of compliance might be an overestimation by ignoring various indirect effects of other factors on effectiveness through compliance. This potential overestimation problem can be solved by identifying the significant indirect effects using a multi-stage regression analysis, which actually separates the overestimated coefficients into two parts, unbiased direct effect and indirect effect. In addition, it is recommended to try a simple multi-stage regression model such as 2SLS and compare the result with the corresponding single equation regression analysis to check for any bias such as overestimation.

A recent article on the comorbidity and cost of illness study highlights the possibility of bias in cost of illness study using regression analysis(14). The article classifies cost of illness estimate biases from two different sources: a bias from controlling for some comorbid condition when the comorbid condition is actually the sequela of the illness (osteoarthritis); and a bias from ignoring a comorbid condition when the illness (osteoarthritis) and the omitted comorbid condition are correlated. For both of these biases, multi-stage regression analysis can provide solutions to the bias. The first type of bias can be controlled by an additional equation capturing the causal relationship between the comorbid condition (sequela) and the illness (osteoarthritis). On the other hand, the second type of bias can be more easily dealt with by simply including omitted comorbid condition. In this case, a multi-stage regression analysis can not only correct for the bias but also further refine the result. Example 5 illustrates a case where the effect of age increases both the illness and the comorbid condition (which is not a sequela of the illness) and the comorbid condition is correlated with the illness. If we omit the comorbid condition from the cost of illness estimation, we get biased result. To solve this bias, we can include comorbid condition to the cost of illness estimation, but the indirect effect of Age can be separated by multi-stage regression analysis.

Example 5. Cost of Illness Study (A comorbid condition is correlated with the cost of illness)

(a) Cost of Illness = β_A Age + ... + Error 1
(Without Comorbid condition, the result is biased)

(b) Cost of Illness = (β_C Comorbid Condition + β_A Age + ... + Error 2
(Bias is corrected but β_A is direct effect only)

(c) Comorbid Condition = β_A Age + ... + Error 1
Cost of Illness = β_C Comorbid Condition + β_A Age + ... + Error 2
(Bias is corrected and both indirect effect and direct effect of Age are captured)

CONCLUDING REMARKS

We have already seen the interpretations of path analysis model and multi-stage regression model from the examples, but an important remark is that these methods also provide simple *t*-test and F-test procedures for individual effect and joint effect, respectively. It is a crucial step to check for the significance of each variable for the sake of valid specification. A misspecified model can generate a serious bias in the estimation of the coefficient of each independent variable.

Regarding the usefulness of path analysis model, even though path analysis cannot correct for the bias like multi-stage regression models, it can still compare the magnitudes of influence from each included factor, whether it is a direct effect factor or an indirect effect factor. This comparison can provide a clue about possible bias in the usual regression analysis. Therefore, it is better to compare with a more general model to check for the correct specification. More complex specification is not always better, but it is one of the most useful ways to check for correct specification.

Only linear models are considered in this paper; however, there is increasing use of non-linear modeling in many fields of pharmaceutical research such as pharmacokinetics (PK) and pharmacodynamics (PD). The characteristics of non-linear models are quite different from the ones of linear models. As mentioned earlier, the identification of parameters in multistage regression models can be simplified by assuming a special non-linear functional form. However, the choice of the special non-linear form should be validated in a theoretical

[12]There are no substantial differences between path analysis and multi-stage regression ananlysis other than covariance structure of error terms (path analysis model has a diagonal covariance matrix), but they are mostly differences in conventional notations such as standardized (normalized) variables used inpath analysis.

[13]This problem is called "local minimum." Since the functional form is nonlinear, the least squares may not be well defined. Therefore, a numerical minimization is needed but it may fail to achieve the real minimum point called "global minimum." Instead, the computer may find a local minimum point called "local minima." Note that a search for the maximum is also the same process with the negative sign.

background. In addition, non-linear models generally require a complex numerical minimization process, which might yield incorrect estimates.[13] Although some data-dependent non-linear approaches, such as Artificial Neural Network (ANN), facilitate the choice of a specific non-linear functional form, it still requires the fragile numerical search of the minimum (maximum) point. A more general explanation on PK/PD modeling using ANN can be found in many recent articles(15-17) Also special software is available for this type of research (NONMEM or Neural Network Toolbox of MATLAB).

References

(1) Bollen, K.A., *Structural Equation with Latent Variables,* John Wiley & Sons, New York NY (1989) pp. 130-131.

(2) Neter, J., Wasserman, W. and Whitemore, G.A., *Applied Statistics,* 4th ed., Prentice Hall, Englewood Cliffs, NJ (1993) pp. 651-684.

(3) Efron, B. and Feldman, D., "Compliance as an explanatory variable in clinical trials," *J. Am. Stat. Assoc.,* **86**(413), 9-17(1991).

(4) Pedhazur, E.J., *Multiple Regression in Behavioral Research – Explanation and Prediction,* 3rd ed., Holt, Harcourt Brace & Company, Orlando FL (1997) pp. 769-770.

(5) Hatcher, L., A Step-by-Step Approach to Using the SAS® System for Factor Analysis and Structural Equation Modeling, SAS Institute, Cary NC (1994).

(6) Hausman, J.A., "Specification and estimation of simultaneous equation models," Chapter 7 in *Handbook of Econometrics,* Vol. 1, (edit., Griliches, Z. and Intriligator, M.D.) North-Holland Publishing Company, Amsterdam The Netherlands (1983) pp. 392-396.

(7) Bollen, K.A. and Long, J.S., *Testing Structural Equation Models,* Sage Publications, Newbury Park CA (1993) p.1.

(8) Theil, H., "Estimation and simultaneous correlation in complete equation systems," The Hague: Centraal Planbureau (1953).

(9) Basmann, R.L., "A generalized classical method of linear estimation of coefficients in a structural equation," *Econometrica,* **25,** 77-83(1957).

(10) Sargan, J.D., "On the estimation of economic relationships by means of instrumental variables," *ibid.,* **26,** 393-415(1958).

(11) Greene, W.H., *Econometric Analysis,* 3rd ed., Prentice Hall, Upper Saddle River NJ (1997) pp.759-761.

(12) Brooks, S.E., Ahn, J., Mullins, C.D., Baquet, C.R. and D'Andrea, A., "Health care cost and utilization project analysis of comorbid illness and complications for patients undergoing hysterectomy for endometrial carcinoma," *Cancer,* **92,** 950-958(2001).

(13) Intriligator, M.D., Bodkin, R.G. and Hsiao, C., *Econometric Models, Techniques, and Applications,* 2nd ed., Prentice Hall, Upper Saddle River NJ (1996) pp. 126-128.

(14) Lee, D.W., Meyer, J.W. and Clouse, J., "Implications of controlling for comorbid conditions in cost-of-illness estimates: a case study of osteoarthritis from a managed care system perspective," *Value Hlth.,* **4,** 329-334(2001).

(15) Chow, H.H., Tolle, K.M., Roe, D.J., Elsberry, V. and Chen, H., "Application of neural networks to population pharmacokinetics data analysis,"*J. Pharm. Sci.,* **86,** 840-845(1997).

(16) Agatonovic-Kustrin, S. and Beresford, R., "Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research," *J. of Pharm. & Biomed. Anal.,* **22,** 717-727(2000).

(17) Takayama, K., Fujikawa, M. and Nagai, T., "Artificial neural network as a novel method to optimize pharmaceutical formulations," *Pharm. Res.,* **16,** 1-6(1999).