

Chapter 2

Two-Stage Least Squares

1 Summary

1.1 Background

Both path analysis and multistage least squares are adequate for simultaneously assessing both direct and indirect predictors. This makes interpretation less easy. Also, path analysis does not provide overall p-values.

1.2 Objective

To assess whether one method performs better than the other.

1.3 Methods

A real clinical data example of patients' non-compliance in a drug efficacy study was used. Data analysis was performed using SPSS statistical software.

1.4 Result

A two-path statistic of 0.46 is a lot better than the single-path statistic of 0.19 with an increase of 60.0 %. However, the result is expressed in a standardized way and without overall p-value, and so it is not easy to interpret. The p-value of two-stage least squares was (much) better than that of the simple linear regression p-value of non-compliance versus therapeutic efficacy (0.02 versus 0.10).

1.5 Conclusions

1. Path analysis and multistage least squares are linear regression methods that are adequate for simultaneous assessment of direct and indirect effects of clinical predictors.
2. Multistage least squares is easier to interpret, because it provides unstandardized regression coefficients and an overall p-value.
3. Multistage least squares is at risk of overestimating the precision of the outcome.

2 Introduction

Multistage regression, otherwise called multistage least squares, was invented by Philip G. Wright, professor of economics at Harvard University, Cambridge MA, 1928 [1]. It uses instrumental variables for improved estimation of problematic (i.e., somewhat uncertain) predictors [2]. It is an alternative to traditional path analysis which assumes that predictor variables not only produce direct effects on the outcome but also indirect effects through affecting concomitant predictor variables. With path analysis usual regression coefficients can not be applied, because they have the same unit as the outcome variable. Instead, standardized regression coefficients have to be used. This makes interpretation less easy. Also, path analysis does not provide overall p-values. In the current chapter two stage least squares is explained as an alternative to traditional path analysis, using a real data example.

3 Example Path Analysis

Patients' non-compliance is a factor notoriously affecting the estimation of drug efficacy. An example is given of a simple evaluation study that assesses the effect of non-compliance (pills not used) on the outcome, efficacy of a novel laxative, with numbers of stools in a month as efficacy estimator (the y-variable). The data are in the first three columns of Table 2.1.

Table 2.2 gives the results of two linear regressions assessing the effects of counseling and non-compliance on therapeutic efficacy (upper table), and the effect of non-compliance on counseling (lower table). With $p = 0.10$ as cut-off p-value for statistical significance all of the effects are statistically significant. Non-compliance is a significant predictor of counseling, and at the same time a significant predictor of therapeutic efficacy. This would mean that non-compliance works two ways: it predicts therapeutic efficacy *directly* and *indirectly* through counseling. However, the indirect way is not taken into account in the usual single step linear regression. An adequate approach for assessing both ways simultaneously is path statistics.

Table 2.1 Example of a study of the effects of counseling and non-compliance on the efficacy of a novel laxative drug (*pt* patient no.)

Pt	Instrumental variable (z)	Problematic predictor (x)	Outcome (y)	
	Frequency counseling	Pills not used (non-compliance)	Efficacy estimator of new laxative (stools/month)	Improved values of problematic predictor
1.	8	25	24	27.68
2.	13	30	30	30.98
3.	15	25	25	32.30
4.	14	31	35	29.00
5.	9	36	39	28.34
6.	10	33	30	29.00
7.	8	22	27	27.68
8.	5	18	14	25.70
9.	13	14	39	30.98
10.	15	30	42	32.30
11.	11	36	41	29.66
12.	11	30	38	29.66
13.	12	27	39	30.32
14.	10	38	37	29.00
15.	15	40	47	32.30
16.	13	31	30	30.98
17.	12	25	36	30.32
18.	4	24	12	25.04
19.	10	27	26	29.00
20.	8	20	20	27.68
21.	16	35	43	32.96
22.	15	29	31	32.30
23.	14	32	40	31.64
24.	7	30	31	27.02
25.	12	40	36	30.32
26.	6	31	21	26.36
27.	19	41	44	34.94
28.	5	26	11	25.70
29.	8	24	27	27.68
30.	9	30	24	28.34
31.	15	20	40	32.30
32.	7	31	32	27.02
33.	6	29	10	26.36
34.	14	43	37	31.64
35.	7	30	19	27.02

Essentially, path analysis assumes two effects, (1) the effect of non-compliance on efficacy, and (2) the effect of non-compliance through counselling on efficacy. These two effects can be, simply, added up, and can be used to cover the joint effects of non-compliance on the efficacy. If we want to add up the effects of both variables,

Table 2.2 The results of two linear regressions assessing (upper table) the effects of counseling and non-compliance on therapeutic efficacy, and (lower table) the effect of non-compliance on counseling

Coefficients ^a					
Model	Unstandardized coefficients		Standardized coefficients		
	B	Std. error	Beta	t	Sig.
1 (Constant)	2.270	4.823		.471	.641
Counseling	1.876	.290	.721	6,469	.000
Non-compliance	.285	.167	.190	1,705	.098
Coefficients ^b					
Model	Unstandardized coefficients		Standardized coefficients		
	B	Std. Error	Beta	t	Sig.
1 (Constant)	4.228	2.800		1.510	.141
Non-compliance	.220	.093	.382	2,373	.024

^aDependent variable: ther eff
^bDependent variable: counseling

their units will have to be the same. Standardizing the data is the solution. For that purpose both the values and their variance are divided by their own variance.

Values (variance of values)

Values (variance of values)

variance

variance

The “values/variance” terms are called the standardized values. They have a variance of 1, which is very convenient for the calculations. In the analysis all regressions are performed on standardized values, giving rise to standardized regression coefficients, which can be simply added up, since they now have the same unit, while variances (equaling 1) no longer have to be taken into account. The standardized regression coefficients are calculated by many software programs. SPSS statistical software [3] routinely reports the standardized regression coefficients together with the usual regression coefficients (Table 2.2).

With simple linear regression the standardized regression coefficient is equal to the r-value. A path diagram must be constructed with arrows for indicating supposed causal effect paths. Figure 2.1 summarizes the supposed effects for our example: (a) non-compliance causes an effect on efficacy, and (b) non-compliance causes an additional effect on efficacy though counseling. The standardized regression coefficients are added to the arrows (Fig. 2.1). Then, they are added up according to:

$0.19 + 0.38 \times 0.72 = 0.46$

This result is expressed as the path statistic, and equals the sum of the standardized regression coefficients. Its magnitude is sometimes interpreted similarly to the overall r-square value of a multiple regression, but this is not entirely correct.

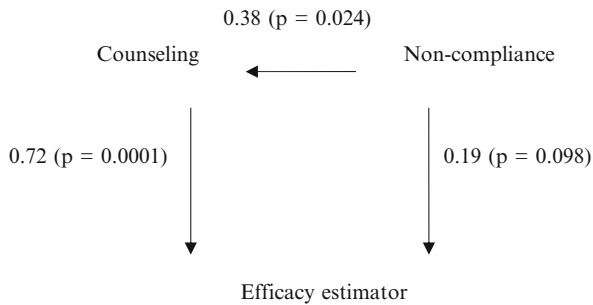


Fig. 2.1 Path diagram of study assessing the direct and indirect effects of non-compliance on efficacy of a new laxative

Unlike r-square values, standardized regression coefficients and their sums can be somewhat larger than 1.0. Also negative indirect factors are sometimes produced, reducing the magnitude of the add-up sums. Yet, the interpretation is pretty much the same. The larger the result of your path statistic, the better the independent variable predicts the dependent one. And, so, the two-path statistic of 0.46 is a lot better than the single-path statistic of 0.19 with an increase of 60.0 %.

4 Multistage Least Squares Method

Instead of path analysis, multistage least square is another possibility for the analysis of the above study. Also this method requires both a significant correlation between the predictors and the outcome in the multiple regression, and a significant correlation between the predictors. The simplest version, called 2 stage least squares (2SLS), is available in the regression module of SPSS [3], and is adequate for most data files. Its theoretical basis is slightly different from that of path analysis. Path analysis assumes causal pathways. The 2SLS method assumes that the independent variable (x-variable), here traditionally called the exogenous variable, is problematic. Problematic means that it is somewhat uncertain. If an additional variable can be argued to provide additional information about a problematic variable, then it may be worthwhile to include it in the analysis.

The variable counseling in the above example may, indeed, cause improvement of patients' compliance and, thus, indirectly improve the outcome variable. In the 2SLS model counseling will be used as an instrumental variable, for the purpose of reducing the uncertainty of the problematic variable non-compliance according to the underneath model:

y = outcome variable (drug efficacy)

x = problematic variable (non-compliance)

z = instrumental variable (counseling)

Table 2.3 Results of the 2LS analysis of the data from Table 2.1 (predictor = explanatory = non-compliance)

Model Description

		Type of Variable
Equation 1	VAR00001	dependent
	VAR00003	predictor
	VAR00002	instrumental

MOD_1

Coefficients

		Unstandardized Coefficients		Beta	t	Sig.
		B	Std. Error			
Equation 1	(Constant)	-61,095	37,210		-1,642	,110
	VAR00003	3,113	1,256	2,078	2,478	,019

1st stage

$x = \text{intercept} + \text{regression coefficient times } z$

With the help of the calculated intercept and regression coefficient from the above simple linear regression analysis improved x-values are calculated e.g. for patient 1

$x_{\text{improved}} = \text{intercept} + \text{regression coefficient times } 8 = 27.68$

The right column of the Table 2.1 gives the improved x-values.

2nd stage

$y = \text{intercept} + \text{regression coefficient times improved x-values.}$

SPSS is used for analysis [3].

Command: Analyze . . . Regression . . . 2 Stage Least Squares . . . Dependent: therapeutic efficacy . . . Explanatory: non-compliance . . . Instrumental: counseling . . . OK.

The Table 2.3 shows the results of the 2LS method. As expected the final p-value is (much) smaller (0.019 versus 0.098) than the simple linear regression p-value of the effect of non-compliance on therapeutic efficacy.

5 Discussion

The present chapter uses data of an efficacy study of a new treatment to explain path analysis and multistage least squares. Both are multistep statistical methods that are adequate for simultaneous assessment of direct and indirect effects. They are very successful in economics, but rarely used in clinical research. The current chapter shows that they can be readily applied to clinical efficacy studies, and enable to make better predictions from the data than does usual linear regression.

Some limitations of multistep regression have to be mentioned. First, instrumental variables may be weak predictors. In order to exclude weak predictors multistage regression should always be preceded by the usual multiple regression: only relatively strong predictor variables significantly predicting the outcome variables can be included. Second, also significant correlations between the predictor variables are required. At the same time, however, they must not be too strong. An r -value > 0.85 indicates the presence of collinearity, which is an important validity criterion of multiple regression.

We should add, that multistage least squares has the advantage compared to path analysis that it uses unstandardized b -values and provides an overall p -value. However, it assumes that instrumental variables are uncorrelated with the error terms of the problematic variable. If this is not warranted, precision of the outcome statistics is likely to be somewhat overestimated.

6 Conclusions

1. Path analysis and multistage least squares are linear regression methods that are adequate for simultaneous assessment of direct and indirect effects of clinical predictors.
2. Multistage least squares is easier to interpret, because it provides unstandardized regression coefficients and an overall p -value.
3. Multistage least squares is at risk of overestimating the precision of the outcome.

References

1. Wright PG (1928) The tariff on animal and vegetable oils. MacMillan, New York, Ed by Wright PC
2. Angrist JD, Krueger AB (2001) Instrumental variables and the search for identification: from supply and demand to natural experiments. *J Econ Perspect* 15:69–85
3. SPSS statistical software. www.spps.com. 25 Dec 2012



<http://www.springer.com/978-94-007-6886-4>

Machine Learning in Medicine

Part Two

Cleophas, T.J.; Zwinderman, A.H.

2013, XIV, 231 p.,

ISBN: 978-94-007-6886-4