Chapter 34
Two-Stage Least Squares (35 Patients)

General Purpose

The two stage least squares method assumes that the independent variable (x-variable) is problematic, meaning that it is somewhat uncertain. An additional variable can be argued to provide relevant information about the problematic variable, and is, therefore, called instrumental variable, and included in the analysis.

Primary Scientific Question

Non-compliance is a predictor of drug efficacy. Counseling causes improvement of patients' compliance and, therefore, indirectly improves the outcome drug efficacy.

$$\text{y} = \text{outcome variable}\left(\text{drug efficacy}\right)$$
$$\text{x} = \text{problematic variable}\left(\text{non} - \text{compliance}\right)$$
$$\text{z} = \text{instrumental variable}\left(\text{counseling}\right)$$

With two stage least squares the underneath stages are assessed.

$$1^{\text{st}}\ \text{stage}$$
$$\text{x} = \text{intercept} + \text{regression coefficient times z}$$

This chapter was previously published in "Machine learning in medicine-cookbook 2" as Chap. 10, 2014.

T.J. Cleophas, A.H. Zwinderman, *Machine Learning in Medicine - a Complete Overview*, DOI 10.1007/978-3-319-15195-3_34

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:

$$1^{st}\,\text{stage}$$
$$x_{\text{improved}} = \text{intercept} + \text{regression coefficient times}\,8 = 27.68$$
$$2^{nd}\,\text{stage}$$
$$y = \text{intercept} + \text{regression coefficient times}\,x_{\text{improved}}$$

Example

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, the efficacy of a novel laxative with numbers of stools per month as efficacy estimator (the y-variable). The data of the first 10 of the 35 patients are in the table below. The entire data file is in extras.springer.com, and is entitled "twostageleastsquares".

Patient no	Instrumental variable (z)	Problematic predictor (x)	Outcome (y)
	Frequency counseling	Pills not used (non-compliance)	Efficacy estimator of new laxative (stools/month)
1.	8	25	24
2.	13	30	30
3.	15	25	25
4.	14	31	35
5.	9	36	39
6.	10	33	30
7.	8	22	27
8.	5	18	14
9.	13	14	39
10.	15	30	42

SPSS version 19 and up can be used for analysis. It uses the term explanatory variable for the problematic variable. Start by opening the data file.

Command:

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

Model description

		Type of variable
Equation 1	y	Dependent
	x	Ppredictor
	z	Instrumental

ANOVA

		Sum of squares	df	Mean square	F	Sig.
Equation 1	Regression	1408,040	1	1408,040	4,429	,043
	Residual	10490,322	33	317,889		
	Total	11898,362	34			

Coefficients

		Unstandardized coefficients				
		B	Std. Error	Beta	t	Sig.
Equation 1	(Constant)	−49,778	37,634		−1,323	,195
	x	2,675	1,271	1,753	2,105	,043

The result is shown above. The non-compliance adjusted for counseling is a statistically significant predictor of laxative efficacy with p = 0.043. This p-value has been automatically been adjusted for multiple testing. When we test the model without the help of the instrumental variable counseling the p-value is larger and the effect is no more statistically significant as shown underneath.

Command:

Analyze….Regression….Linear….Dependent: therapeutic efficacy ….Independent: non-compliance….OK.

ANOVA[a]

Model		Sum of squares	df	Mean square	F	Sig.
1	Regression	334,482	1	334,482	3,479	,071[b]
	Residual	3172,489	33	96,136		
	Total	3506,971	34			

[a] Dependent variable: drug efficacy
[b] Predictors: (Constant), non-compliance

Coefficients[a]

		Unstandardized coefficients		Standardized coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	15,266	7,637		1,999	,054
	Non-compliance	,471	,253	,309	1,865	,071

[a] Dependent variable: drug efficacy

Conclusion

Two stage least squares with counseling as instrumental variable was more sensitive than simple linear regression with laxative efficacy as outcome and non-compliance as predictor. We should add that two stage least squares is at risk of overestimating the precision of the outcome, if the analysis is not adequately adjusted for multiple testing. However, in SPSS automatic adjustment for the purpose has been performed. The example is the simplest version of the procedure. And, multiple explanatory and instrumental variables can be included in the models.

Note

More background theoretical and mathematical information of two stage least squares analyses is given in Machine learning in medicine part two, Two-stage least squares, pp 9–15, Springer Heidelberg Germany, 2013, from the same authors.