PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 47 2009
TAKSONOMIA 16
Klasyfikacja i analiza danych – teoria i zastosowania

Mirosława Sztemberg-Lewandowska
Uniwersytet Ekonomiczny we Wrocławiu

MODELE RÓWNAŃ STRUKTURALNYCH Z WYKORZYSTANIEM ŚRODOWISKA R

1. Wstęp

Modele równań strukturalnych (*Structural Equation Modeling* – SEM), inaczej nazywane LISREL (por. np. [Sagan 1996; Mueller 1996]), są ogólną, głównie liniową, wielowymiarową techniką statystyczną. Jest ona bardziej konfirmacyjna niż eksploracyjna, czyli wykorzystuje się ją do sprawdzania dopasowania określonego modelu do danych, a nie do budowania pasującego modelu.

SEM jest bardziej ogólną metodą w stosunku do regresji wielorakiej, analizy ścieżkowej, konfirmacyjnej analizy czynnikowej, analizy szeregów czasowych i analizy kowariancji. Procedury te zatem mogą być uważane za szczególne przypadki modeli równań strukturalnych.

W SEM oprócz obserwowalnych zmiennych dopuszcza się także ukryte konstrukty, czyli zmienne nieobserwowalne (por. np. [Bollen i Long 1993; Hair i in. 1998; Mueller 1996]. Zależności między ukrytymi konstruktami opisuje model strukturalny. Najczęściej są to zależności liniowe, chociaż dopuszcza się także związki nieliniowe. Oprócz wykrycia zależności między ukrytymi konstruktami badacze dążą także do oszacowania błędu pomiaru. Natomiast zależności między zmiennymi obserwowalnymi a konstruktami opisuje model pomiarowy.

W artykule przedstawiono wybrane pakiety i funkcje programu **R** służące do estymacji modeli równań strukturalnych. Funkcje te zobrazowano na przykładach przytoczonych z literatury. Celem artykułu jest opisanie oraz porównanie wymienionych funkcji.

2. Pakiety i funkcje programu R

W tabeli 1 przedstawiono wybrane pakiety i funkcje programu **R** wykorzystywane w SEM.

Tabela 1. Wybrane pakiety i funkcje programu **R** wykorzystywane w SEM

Lp.	Problem	Wybrane pakiety i funkcje programu **R**
1	Metody SEM	Pakiet `sem` (funkcje: `sem, tsls`) Pakiet `systemfit` (funkcja: `systemfit`) Pakiet `pls` (funkcja: `mvr`)
2	Alternatywne parametry modelu	Pakiet `sem` (funkcja: `mod.indices`)
3	Wyznaczanie reszt	Pakiet `sem` (funkcje: `residuals.sem, standardized.residuals, standardized.coefficients`) Pakiet `stats` (funkcja: `resiuduals`)
4	Graficzna prezentacja modelu	Pakiet `sem` (funkcja: `path.diagram` (dodatkowo wymagany jest program *Graphviz*))
5	Ocena dopasowania modelu	Pakiet `stats` (funkcje: `glm, lm, extractAIC`) Pakiet `sem` (funkcje: `sem, tsls`) Pakiet `pls` (funkcje: `MSEP, RMSEP`)

Źródło: opracowanie własne na podstawie dokumentacji programu **R**.

W SEM ważnym problemem jest wybór metody estymacji parametrów modeli równań strukturalnych. W tabeli 2 przedstawiono funkcje programu **R** oraz dostępne w nich metody.

Tabela 2. Metody SEM oraz odpowiadające im funkcje w środowisku **R**

Metoda	Funkcja
Metoda największej wiarygodności	sem
2SLS (*Two-Stage Least Squares*)	`tsls`
OLS (*Ordinary Least Squares*), WLS (*Weighted Least Squares*), W2SLS (*Weighted Two-Stage Least Squares*), 3SLS (*Three-Stage Least Squares*), SUR (*Seemingly Unrelated Regressions*), 2SLS (*Two-Stage Least Squares*)	`systemfit`
PLS (*Partial Least Squares*)	`mvr`

Źródło: opracowanie własne na podstawie dokumentacji programu **R**.

Bardzo przydatna jest funkcja `mod.indices` dostępna w pakiecie `sem`. Funkcja ta podaje propozycje zmodyfikowanych parametrów w modelu, które zwykle prowadzą do modelu lepiej dopasowanego do danych.

Niestety program **R** nie pozwala na graficzną prezentację modelu SEM. Jednak na stronie http://www.graphviz.org/ dostępny jest program *Graphviz,* który umożliwia taką prezentację. Pomocna jest tutaj funkcja `path.diagram` zwracająca ciąg poleceń, które wywołane w programie *Graphviz* umożliwiają graficzną prezentację modelu.

3. Przykłady

W celu zobrazowania przedstawionych funkcji przytoczono badanie gospodarki Stanów Zjednoczonych przeprowadzone przez Kleina w 1950 r. Dane zawierają obserwacje od 1920 r. do roku 1941 i są dostępne w pakiecie `systemfit` pod nazwą KleinI. W badaniu wykorzystano zmienne: `Year` – rok (**Y**), `consump` – konsumpcja (**C**), `corpProf` – zyski spółek kapitałowych (**P**), `corpProfLag` – zyski spółek kapitałowych w poprzednim roku (**P.lag**), `privWage` – prywatny fundusz płac (**Wp**), `invest` – inwestycje (**I**), `capitalLag` – kapitał w poprzednim roku (**K.lag**), `gnp` – produkt narodowy brutto (**X**), `gnpLag` – produkt narodowy brutto w poprzednim roku (**X.lag**), `govWage` – rządowy fundusz płac (**Wg**), `govExp` – wydatki publiczne (**G**), `taxes` – podatki (**T**), `wages` – suma prywatnego i rządowego funduszu płac (**W**), `trend` – ile lat minęło od 1931 r. (**A**).

Model KleinI jest opisany następującymi równaniami:

$$C_t = \gamma_{10} + \gamma_{11} P_t + \gamma_{12} P.lag_t + \beta_{11} W_t + \varsigma_{1t}$$

$$I_t = \gamma_{20} + \gamma_{21} P_t + \gamma_{22} P.lag_t + \beta_{21} K.lag_t + \varsigma_{2t}$$

$$Wp_t = \gamma_{30} + \gamma_{31} A_t + \beta_{31} X_t + \beta_{32} X.lag_t + \varsigma_{3t}$$

$$X_t = C_t + I_t + G_t$$

$$P_t = X_t - T_t - Wp_t$$

$$K_t = K.lag_t + I_t$$

Trzy ostatnie równania są tożsamościami, więc ich parametry są znane i nie wymagają estymacji. Model ten opisano symbolicznie:

```
r.1 <- C ~ P+ P.lag+W
r.2 <- I ~ P+ P.lag+K.lag
r.3 <- Wp ~ A+ X+X.lag
instruments <- ~ G+T+Wg+A+K.lag+P.lag+X.lag.
```

W opisie występują zmienne instrumentalne, tj. zmienne, które są nieskorelowane z błędami w równaniach strukturalnych. W przykładzie są to zmienne egzogeniczne.

W wyniku składni poleceń:

```
>tsls1 <- tsls(r.1, instruments)
>summary(tsls1)
```

przeprowadzono estymację pierwszego równania podwójną metodą najmniejszych kwadratów:

```
2SLS Estimates
Model Formula: consump ~ corpProf + corpProfLag + wages
```

```
Instruments: ~govExp + taxes + govWage + trend + capi-
talLag + corpProfLag + gnpLag
Residuals:
 Min.      1st Qu.   Median    Mean     3rd Qu.  Max.
-1,89e+00 -6,16e-01 -2,46e-01 -6,61e-11 8,85e-01 2,00e+00
            Estimate Std, Error  t value Pr(>|t|)
(Intercept) 16,55476 1,46798    11,2772  2,587e-09
corpProf    0,01730  0,13120     0,1319  8,966e-01
corpProfLag 0,21623  0,11922     1,8137  8,741e-02
wages       0,81018  0,04474    18,1107  1,505e-12
Residual standard error: 1,1357 on 17 degrees of freedom.
```

Zatem oszacowany model jest następujący (standardowe błędy szacunku estymatorów parametrów zamieszczono w nawiasach):

$$\underset{(1,47)}{\text{consump}} = \underset{(1,47)}{16,55} + \underset{(0,13)}{0,02}\,\text{corpProf} + \underset{(0,12)}{0,22}\,\text{corpProfLag} + \underset{(0,04)}{0,81}\,\text{wages} + \hat{e}_t .$$

Estymację pierwszego równania metodą częściowych najmniejszych kwadratów można otrzymać za pomocą poleceń:

```
>mvr1 <- mvr(r.1, data = KleinI)
>summary(mvr1)
>RMSEP(mvr1)
>R2(mvr1)
>coef(mvr1).
```

W wyniku zastosowania tej procedury otrzymano następujące wyniki:

```
Fit method: kernelpls
Number of components considered: 3
TRAINING: % variance explained
        1 comps  2 comps  3 comps
X       80,75    95,73    100,0
consump 96,90    98,09    98,1
    (Intercept)  1 comps  2 comps  3 comps
R2  0,00         0,97     0,98     0,98
coef(mvr1)
             consump
corpProf     0,19
corpProfLag  0,09
wages        0,80.
```

Oszacowane parametry modelu (`coef(mvr1)`) są zbliżone do poprzedniej estymacji. Dodatkowo wyznaczono współczynnik determinacji R^2, który przyjmuje wartości z przedziału [0; 1] i wskazuje, jaka część zmienności zmiennej objaśnianej została wyjaśniona przez zbudowany model. Dla równania pierwszego R^2 jest bliskie jedynki dla wszystkich parametrów z wyjątkiem stałej.

Wykorzystując funkcję `systemfit`, może estymować jednocześnie kilka równań:

```
>sf <- systemfit(list(r.1, r.2, r.3), data = KleinI,
method = "WLS")
>summary(sf).
```

W wyniku tych poleceń otrzymano estymację trzech równań metodą ważonych najmniejszych kwadratów:

```
method: WLS
         N   DF  SSR       detRCov  OLS-R2    McElroy-R2
system  63   51  45,2069   0,37084  0,977268  0,991302
    N  DF SSR      MSE       RMSE      R2        Adj R2
eq1 21 17 17,8794 1,051732 1,025540 0,981008 0,977657
eq2 21 17 17,3227 1,018982 1,009447 0,931348 0,919233
eq3 21 17 10,0047 0,588515 0,767147 0,987414 0,985193
The covariance matrix of the residuals used for estimation
    eq1      eq2      eq3
eq1 1,05173 0,00000 0,000000
eq2 0,00000 1,01898 0,000000
eq3 0,00000 0,00000 0,588515
The covariance matrix of the residuals
     eq1       eq2        eq3
eq1  1,0517323 0,0611432 -0,470419
eq2  0,0611432 1,0189825  0,149681
eq3 -0,4704191 0,1496807  0,588515
The correlations of the residuals
     eq1       eq2        eq3
eq1  1,0000000 0,0590626 -0,597935
eq2  0,0590626 1,0000000  0,193288
eq3 -0,5979346 0,1932875  1,000000
```

`WLS estimates for 'eq1' (equation 1)` (zamieszczono tylko estymację dla równania 1)

```
Model Formula: consump ~ corpProf + corpProfLag + wages
            Estimate   Std, Error t value  Pr(>|t|)
(Intercept) 16,2366003 1,3026983  12,46382 5,6208e-10 ***
corpProf    0,1929344  0,0912102  2,11527  0,049474 *
corpProfLag 0,0898849  0,0906479  0,99158  0,335306
wages       0,7962187  0,0399439  19,93342 3,1597e-13 ***
Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.'
0,1 ' ' 1
Residual standard error: 1,02554 on 17 degrees of freedom
Number of observations: 21 Degrees of Freedom: 17
SSR: 17,879449 MSE: 1,051732 Root MSE: 1,02554
Multiple R-Squared: 0,981008 Adjusted R-Squared: 0,977657.
```

Estymacja metodą ważonych najmniejszych kwadratów daje podobne wyniki jak poprzednie metody.

Zastosowane funkcje pozwalają na estymację modeli, które zawierają tylko zmienne obserwowalne, więc można je stosować głównie do regresji wielorakiej. Inaczej jest z funkcją `sem`, która estymuje równania strukturalne ze zmiennymi obserwowalnymi i latentnymi (ukrytymi). W celu zobrazowania tej funkcji przytoczono badanie przeprowadzone przez Duncana, Hallera i Portesa w 1968 r., dotyczące wpływu rówieśników na aspiracje uczniów szkoły średniej. Zmienne obserwowalne objęte badaniem to:

RSES (FSES) – status socjoekonomiczny respondenta (najlepszego przyjaciela),
RParAsp (FParAsp) – aspiracje rodziców respondenta (najlepszego przyjaciela),
RIQ (FIQ) – iloraz inteligencji respondenta (najlepszego przyjaciela),
REdAsp (FEdAsp) – plany edukacyjne respondenta (najlepszego przyjaciela),
ROccAsp (FOccAsp) – plany zawodowe respondenta (najlepszego przyjaciela).

Zmienne latentne:

RGenAsp (FGenAsp) – ogólne aspiracje respondenta (najlepszego przyjaciela).

Funkcja `sem` wymaga, aby zależności między zmiennymi były opisane w języku RAM, który symbolicznie koduje diagram ścieżkowy. Zależności między badanymi zmiennymi w opisie RAM:

```
model.DHP <- matrix(c(
'RParAsp -> RGenAsp', 'gam11', NA,
'RIQ -> RGenAsp', 'gam12', NA,
'RSES -> RGenAsp', 'gam13', NA,
'FSES -> RGenAsp', 'gam14', NA,
'RSES -> FGenAsp', 'gam23', NA,
'FSES -> FGenAsp', 'gam24', NA,
'FIQ -> FGenAsp', 'gam25', NA,
'FParAsp -> FGenAsp', 'gam26', NA,
'FGenAsp -> RGenAsp', 'beta12', NA,
'RGenAsp -> FGenAsp', 'beta21', NA,
'RGenAsp -> ROccAsp', NA, 1,
'RGenAsp -> REdAsp', 'lam21', NA,
'FGenAsp -> FOccAsp', NA, 1,
'FGenAsp -> FEdAsp', 'lam42', NA,
'RGenAsp <-> RGenAsp', 'ps11', NA,
'FGenAsp <-> FGenAsp', 'ps22', NA,
'RGenAsp <-> FGenAsp', 'ps12', NA,
'ROccAsp <-> ROccAsp', 'theta1', NA,
'REdAsp <-> REdAsp', 'theta2', NA,
'FOccAsp <-> FOccAsp', 'theta3', NA,
'FEdAsp <-> FEdAsp', 'theta4', NA),
ncol=3, byrow=TRUE).
```

Estymację modelu DHP (Duncana, Hallera i Portesa) metodą największej wiarygodności można otrzymać za pomocą poleceń:

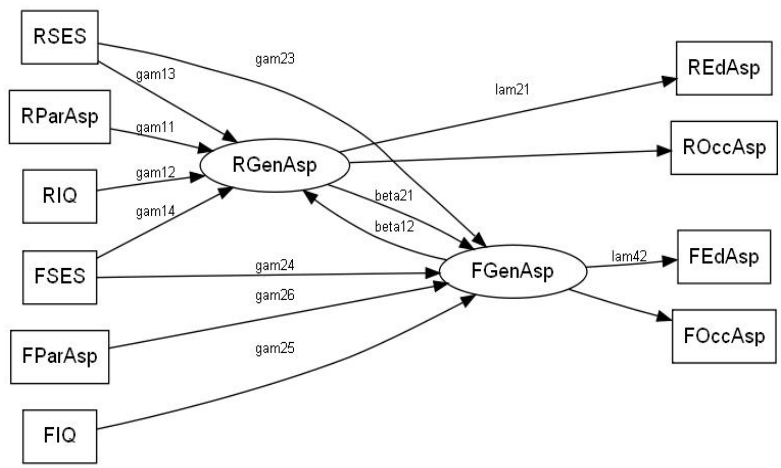
```
>sem.DHP <- sem(model.DHP, kor.DHP, 329,
fixed.x=c('RParAsp', 'RIQ', 'RSES', 'FSES', 'FIQ',
'FParAsp'))
>summary(sem.DHP).
```

W składni funkcji sem wskazano ustalone zmienne egzogeniczne, których wariancja i kowariancja nie jest estymowana (wynosi 0).

Wyniki estymacji za pomocą funkcji sem:

```
Model Chisquare = 26,70 Df = 15 Pr(>Chisq) = 0,03
Chisquare (null model) = 872 Df = 45
Goodness-of-fit index = 0,98
Adjusted goodness-of-fit index = 0,94
RMSEA index = 0,05 90% CI: (0,01, 0,08)
Bentler-Bonnett NFI = 0,97
Tucker-Lewis NNFI = 0,96
Bentler CFI = 0,98
SRMR = 0,02
BIC = -60,24
Normalized Residuals
  Min.  1st Qu. Median Mean 3rd Qu. Max.
 -0,80 -0,12    0,00  -0,01 0,04    1,57
Parameter Estimates
       Estimate Std Error z value  Pr(>|z|)
gam11  0,16     0,04      4,19     2,8019e-05 RGenAsp <--- RParAsp
gam12  0,25     0,04      5,60     2,1428e-08 RGenAsp <--- RIQ
gam13  0,22     0,04      5,02     5,0730e-07 RGenAsp <--- RSES
gam14  0,07     0,05      1,43     1,5350e-01 RGenAsp <--- FSES
gam23  0,06     0,05      1,20     2,3158e-01 FGenAsp <--- RSES
gam24  0,23     0,04      5,14     2,6938e-07 FGenAsp <--- FSES
gam25  0,35     0,04      7,83     4,6629e-15 FGenAsp <--- FIQ
gam26  0,16     0,04      3,97     7,0224e-05 FGenAsp <--- FParAsp
beta12 0,18     0,10      1,91     5,5506e-02 RGenAsp <--- FGenAsp
beta21 0,23     0,12      1,97     4,9255e-02 FGenAsp <--- RGenAsp
lam21  1,06     0,09      11,55    0,0000e+00 REdAsp <--- RGenAsp
lam42  0,93     0,07      13,07    0,0000e+00 FEdAsp <--- FGenAsp
ps11   0,28     0,05      6,07     1,2999e-09 RGenAsp <--> RGenAsp
ps22   0,26     0,04      5,87     4,2067e-09 FGenAsp <--> FGenAsp
ps12  -0,02     0,05     -0,44     6,6168e-01 FGenAsp <--> RGenAsp
theta1 0,41     0,05      7,89     2,8866e-15 ROccAsp <--> ROccAsp
theta2 0,34     0,05      6,30     2,9003e-10 REdAsp <--> REdAsp
theta3 0,31     0,05      6,67     2,5800e-11 FOccAsp <--> FOccAsp
theta4 0,40     0,05      8,66     0,0000e+00 FEdAsp <--> FEdAsp.
```

Obliczone miary dopasowania modelu do danych: GFI (`Goodness-of-fit index`), `AGFI (Adjusted goodness-of-fit index)`, `Bentler-Bonnett NFI`, `Tucker-Lewis NNFI`, `Bentler CFI` bliskie 1 oraz `RMSEA` – pierwiastek średniokwadratowego błędu przybliżenia równy 0,5 wskazują na dobre dopasowanie otrzymanego modelu do danych. W kolumnie `Estimate` podano oszacowane parametry modelu, natomiast w kolumnie `Std Error` ich błędy standardowe.



Rys. 1. Graficzna prezentacja modelu `DHP`
Źródło: opracowanie na podstawie programu **R** i *Graphviz*.

Następnie zastosowano funkcję `path.diagram do` modelu `DHP`:

```
> path.diagram(sem.DHP, min.rank='RIQ, RSES, RParAsp, FParAsp, FSES, FIQ', max.rank='ROccAsp, REdAsp, FEdAsp, FOccAsp').
```

W wyniku zastosowania tej funkcji otrzymano ciąg poleceń, które wywołane w programie *Graphviz* umożliwią graficzną prezentację modelu `DHP` (rys. 1).

4. Podsumowanie

W artykule przedstawiono cztery podstawowe funkcje, które służą do estymacji parametrów modeli równań strukturalnych: `tsls`, `mvr`, `systemfit`, `sem`. Funkcje te różnią się przede wszystkim metodami estymacji oraz modelami, jakie estymują. Pierwsza funkcja `tsls` wykorzystuje podwójną metodę najmniejszych kwadratów do modeli, które zawierają tylko zmienne obserwowalne. Druga funkcja `mvr` wykorzystuje metodę częściowych najmniejszych kwadratów do modeli zawierających tylko zmienne obserwowalne. Trzecia funkcja `systemfit` zawiera

metody: klasyczną najmniejszych kwadratów, ważoną najmniejszych kwadratów, podwójną najmniejszych kwadratów, ważoną podwójną najmniejszych kwadratów i potrójną najmniejszych kwadratów. Może być wykorzystywana do modeli zawierających tylko zmienne obserwowalne, jednak w odróżnieniu od poprzednich funkcji może równocześnie estymować kilka równań. Zatem funkcje `tsls`, `mvr`, `systemfit` służą do estymacji modeli zawierających tylko zmienne obserwowalne, np. do modeli regresyjnych. Czwarta funkcja `sem` wykorzystuje metodę największej wiarygodności dla modeli zawierających zmienne obserwowalne i latentne. Jako jedyna może być zastosowana do modeli ze zmiennymi ukrytymi, a zatem np. do konfirmacyjnej analizy czynnikowej.

Literatura

Bollen K.A., Long J.S. (1993), *Testing structural equation models,* Newbury Park, CA: Sage.

Dalgaard P. (2002), *Introductory statistics with R*, Springer-Verlag, New York.

Everitt B.S., Hothorn T. (2006), *A handbook of statistical analyses using R*, Chapman & Hall, London.

Hair J.F., Anderson R.E., Tatham R.L., Black W.C. (1998), *Multivariate data analysis with readings*, Englewood Cliffs, Prentice-Hall.

Mueller R.O. (1996), *Basic principles of structural equation modeling, an introduction to LISREL and EQS,* Springer, New York.

Sagan A. (1996), *Metoda LISREL w badaniach marketingowych cech i struktur latentnych,* [w:] *Metody badań marketingowych*, X konferencja Katedry Marketingu i Handlu Uczelni i Wydziałów Ekonomicznych, Kraków, s. 44-57.

STRUCTURAL EQUATION MODELS WITH R

Summary

Structural Equation Models (*SEM*) are multi-equation regression models. These structural equations are meant to represent causal relationship among the variables in the model.

Structural equation modeling is a very general, powerful multivariate analysis technique that includes the specialized versions of a number of other analysis methods as special cases. Factor analysis, path analysis and regression represent special cases of SEM.

In this article, the author presents packages and functions in **R** system, which computed SEM. The aim of the article is o characterize these functions and illustrate them with the use of empirical examples.