Distributed-Lag Structural Equation Modelling with the R package dlsem

Alessandro Magrini
Dep. Statistics, Informatics, Applications
University of Florence, Italy

`dlsem` version 1.2.1 – July 5, 2016.

Contents

1 Introduction

Package `dlsem` implements estimation and path analysis functionalities for structural equation models with second-order polynomial lag shapes. In this vignette, the theory on distributed-lag structural equation modelling is presented in Section 2, then the practical use of `dlsem` is illustrated through worked examples in Section 3. Concluding remarks are pointed out in Section 4.

To cite `dlsem` in publications, please use:

> A. Magrini, F. Bartolini, A. Coli, and B. Pacini (2016). Distributed-Lag Structural Equation Modelling: An Application to Impact Assessment of Research Activity on European Agriculture. *Proceedings of the 48th Meeting of the Italian Statistical Society*, 8-10 June 2016, Salerno, IT.

2 Theory

Distributed-lag structural equation modelling was firstly formalised by [7] as a combination of structural equation modelling (for example, [4]) and distributed-lag linear regression [1]. In this chapter, theory on structural equation modelling and distributed-lag linear regression is briefly reported before presenting distributed-lag structural equation modelling.

2.1 Structural equation modelling

Structural equation modelling (SEM) has a long history starting with the contribution of Wright [9]. The main idea behind SEM is to perform a quantitative assessment of dependence relationships among a set of variables. The basic feature of SEM is a directed acyclic graph (DAG). In a DAG, variables are represented by nodes and directed edges may connect pairs of variables without creating directed cycles (See Figure 1). If a variable receives an edge from another variable, the latter is called *parent* of the former. A DAG encodes a factorization of the joint probability distribution:

$$p(V_1, \ldots, V_m) = \prod_{j=1}^{m} p(V_j \mid \Pi_j) \tag{1}$$

where Π_j is the set of parents of variable V_j. As such, if some pairs of variables are not connected by an edge, the DAG implies a set of conditional independence statements [5]. The DAG may eventually have a causal interpretation. If this is the case, edges represent direct causal relationships. SEM is implemented by simultaneously applying linear regression models:

$$\begin{cases} V_1 = f_1(\Pi_1) \\ \ldots \\ V_j = f_j(\Pi_j) \\ \ldots \\ V_m = f_m(\Pi_m) \end{cases} \tag{2}$$

where $V_j = f_j(\Pi_j)$ is the equation describing the linear regression model where V_j is the response variable and its parents in the DAG are the covariates. A comprehensive review of SEM can be found in [4].

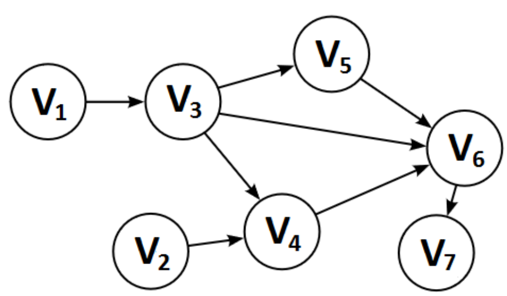

Figure 1: A directed acyclic graph.

An important utility of SEM is path analysis, that is the decomposition of the causal effect of any variable on another. Path analysis can be performed according to trace rules developed by [9] (see also [8]):

- the causal effect associated to each edge in the DAG is represented by the coefficient of the variable originating the edge in the regression model of the variable receiving the edge;
- the causal effect associated to a directed path is represented by the product of the causal effects associated to each edge in the path;
- the causal effect of a variable on another is represented by the sum of the causal effects associated to each directed path connecting the two variables.

For instance, consider variables V_3 and V_6 in the DAG displayed in Figure 1. The directed paths connecting the two variables are (V_3, V_6), (V_3, V_4, V_6) and (V_3, V_5, V_6). The causal effect associated to the first path is the coefficient of V_3 in the regression model of V_6, say $\beta_{6|3}$. The causal effect associated to the second path is the coefficient of V_3 in the regression model of V_4, multiplied by the coefficient of V_4 in the regression model of V_6, say $\beta_{4|3} \cdot \beta_{6|4}$. The causal effect associated to the third path is the coefficient of V_3 in the regression model of V_5 multiplied by the coefficient of V_5 in the regression model of V_6, say $\beta_{5|3} \cdot \beta_{6|5}$. The overal causal effect of V_3 on V_6 is the sum of the causal effect associated to each of the three paths above, say $\beta_{6|3} + \beta_{4|3} \cdot \beta_{6|4} + \beta_{5|3} \cdot \beta_{6|5}$.

Often, the causal effect of a variable on another is termed *overall* causal effect, the causal effect associated to a directed path made by a single edge is called *direct* effect, while the causal effects associated to the other directed paths are denoted as *indirect* effects. In the example above, $\beta_{6|3}$ represents the direct effect of V_3 on V_6, while causal effects $\beta_{4|3} \cdot \beta_{6|4}$ and $\beta_{5|3} \cdot \beta_{6|5}$ represent the indirect effects of V_3 on V_6, and $\beta_{6|3} + \beta_{4|3} \cdot \beta_{6|4} + \beta_{5|3} \cdot \beta_{6|5}$ is the overall causal effect of V_3 on V_6.

2.2 Distributed-lag linear regression

Distributed-lag linear regression is an extension of the classic linear model including lagged instances of one or more quantitative covariates:

$$y_t = \beta_0 + \sum_{j=1}^{J} \sum_{l=0}^{L_j} \beta_{j,l}\ x_{j,t-l} + \epsilon_t \qquad \epsilon_t \sim \mathrm{N}(0, \sigma^2) \tag{3}$$

where y_t is the value of the response variable at time t and $x_{j,t-l}$ is the value of the j-th covariate at l time lags before t. The set $(\beta_{j,0}, \beta_{j,1}, \ldots, \beta_{j,L_j})$ is denoted as the *lag shape* of the j-th covariate and represents its effect on the response variable at different time lags. Estimation of a distributed-lag linear regression model using ordinary least squares is inefficient because lagged instances of the same covariate are typically highly correlated. Also, the lag shape of a covariate is completely unrestricted, thus problems of interpretation may arise.

Second-order polynomial lag shapes can be used to solve these drawbacks. They include the endpoint-constrained quadratic lag shape:

$$\beta_{j,l} = \begin{cases} \theta_j \left[-\frac{4}{(b_j - a_j + 2)^2} l^2 + \frac{4(a_j + b_j)}{(b_j - a_j + 2)^2} l - \frac{4(a_j - 1)(b_j + 1)}{(b_j - a_j + 2)^2} \right] & a_j \leq l \leq b_j \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

and the quadratic decreasing lag shape:

$$\beta_{j,l} = \begin{cases} \theta_j \frac{l^2 - 2b_j l + b_j^2}{(b_j - a_j)^2} & a_j \leq l \leq b_j \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

(see Figure 2). The endpoint-constrained quadratic lag shape is zero for a lag $l \leq a_j - 1$ or $l \geq b_j + 1$, and symmetric with mode equal to θ_j at $(a_j + b_j)/2$. The quadratic decreasing lag shape decreases from value θ_j at lag a_j to value 0 at lag b_j according to a quadratic function. We refer to a_j as the *gestation lag*, and to $b_j - a_j$ as the *lag width*.

A distributed-lag linear regression model with second-order polynomial lag shapes is linear in parameters θ_j $(j = 1, \ldots, J)$, provided that parameters a_j and b_j $(j = 1, \ldots, J)$ are known. Thus, one can fit several models by applying ordinary least squares where the value of a_j and b_j is varied within a grid of values, and then select the model with the best fit to data. See [1] (Chapter 6) for further details on distributed-lag linear regression.

Note that neither the response variable nor the covariates must contain a trend in order to obtain unbiased estimates [3]. A reasonable procedure is to sequentially apply differentiation to all variables until the Dickey-Fuller test rejects the hypothesis of unit root for all of them.

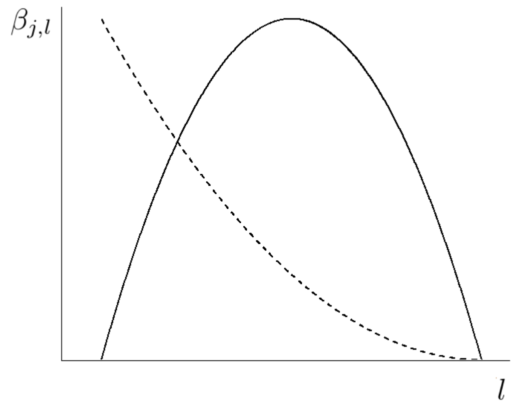

Figure 2: Second-order polynomial lag shapes: endpoint-constrained quadratic lag shape (straight line), quadrating decreasing lag shape (dotted line).

2.3 Distributed-lag structural equation modelling

Distributed-lag structural equation modelling (DLSEM) is an extension of SEM, where variables are related by distributed-lag linear regression models [7]. DLSEM can be used to perform path analysis at different time lags by extending tracing rules reported in Subsection 2.1 (see the box below).

Tracing rules for DLSEM

- The causal effect associated to each edge in the DAG at lag k is represented by the coefficient at lag k of the variable originating the edge in the regression model of the variable receiving the edge.
- The causal effect associated to a directed path at lag k is computed as follows:
  1. denote the number of edges in the path as p;
  2. enumerate all the possible p-uples of lags, one lag for each of the p edges, such that their sum is equal to k;
  3. for each p-uple of lags:
     - for each lag in the p-uple, compute the coefficient associated to the corresponding edge at that lag;
     - compute the product of all these coefficients;
  4. sum all these products.
- The causal effect of a variable on another at lag k is represented by the sum of the causal effects at lag k associated to each directed path connecting the two variables.

A causal effect evaluated at a single lag is denoted as *instantaneous* causal effect. The *cumulative* causal effect at a prespecified lag, say k, is obtained by summing all the instantaneous causal effects sat each lag up to k.

3 Distributed-lag structural equation modelling with dlsem

The practical use of package `dlsem` is illustrated by an application to impact assessment of research activity on European Agriculture. It is widely acknowledged that research activity is effective in increasing productivity, however, it is also expected to improve profitability and consumer surplus independently from productivity. The goal is to test if the influence through time of research activity on profitability and consumer surplus is direct and/or mediated by productivity. In this section, the use of package `dlsem` is illustrated to address this research question by fitting a distributed-lag structural equation model with DAG shown in Figure 3 to dataset agres.

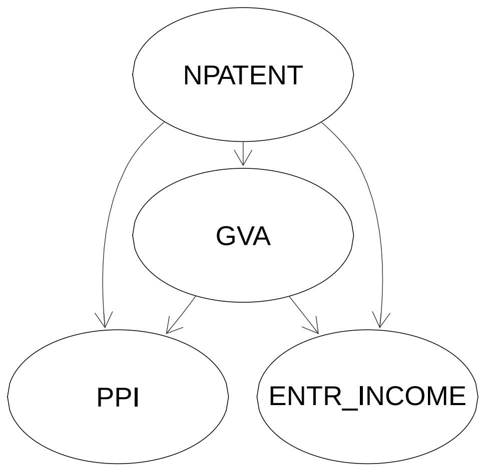

Figure 3: The hypothesized DAG for impact assessment of research activity on European Agriculture. 'RES': research activity. 'PROD': productivity. 'PROFIT': profitability. 'C_SURPL': consumer surplus.

Dataset agres can be loaded using the command:

```
> data(agres)
> summary(agres)
    COUNTRY          YEAR           GDP              FARM_SIZE
 AT     : 22   Min.   :1991   Min.   :  85220   Min.   :0.01820
 BE     : 22   1st Qu.:1996   1st Qu.: 218183   1st Qu.:0.03370
 DE     : 22   Median :2002   Median : 356676   Median :0.05104
 DK     : 22   Mean   :2002   Mean   : 879657   Mean   :0.06222
 EL     : 22   3rd Qu.:2007   3rd Qu.:1678138   3rd Qu.:0.07544
 ES     : 22   Max.   :2012   Max.   :3158590   Max.   :0.21481
 (Other):176
    NPATENT            GVA              PPI              ENTR_INCOME
 Min.   :  0.04   Min.   :  968   Min.   : 60.36   Min.   : 18.75
 1st Qu.:  7.75   1st Qu.: 3593   1st Qu.: 97.14   1st Qu.: 70.70
 Median : 24.18   Median : 6782   Median :102.07   Median : 87.80
 Mean   : 55.27   Mean   :13471   Mean   :105.52   Mean   : 91.85
 3rd Qu.: 71.73   3rd Qu.:21024   3rd Qu.:111.12   3rd Qu.:107.44
 Max.   :472.09   Max.   :41048   Max.   :191.60   Max.   :229.36
 NA's   :1                         NA's   :9         NA's   :8
```

It contains data for 10 European countries (Austria, Germany, Spain, Finland, France, Ireland, Italy, Netherlands, Sweden, United Kingdom) in the period 1990-2010 from the EUROSTAT database (`http://ec.europa.eu/eurostat/data/database`). Variable NPATENT representing the number of Agriculture-related patent applications will be used as a proxy of research activity in Agriculture. Variable GVA representing the gross value added of Agriculture will be used as a proxy of agricultural productivity. Variable ENTR_INCOME representing the net entrepreneurial income index will be used as a proxy of profitability. Variable PPI representing the price index of agricultural products will be used as a proxy of consumer surplus.

The first step is to specify the model code containing the hypothesized DAG and the lag shapes. The model code must be a list of formulas, one for each regression model. In each formula, the response must be a quantitative variable and operators quec( ) and qdec( ) can be employed to specify an endpoint-constrained quadratic or a quadratic decreasing lag shape for any quantitative variables, respectively. Each of these operators has three arguments: the name of the variable to which the lag shape is applied, the minimum lag with a non-zero coefficient (a_j), and the maximum lag with a non-zero coefficient (b_j). The application of one of these two operators is not mandatory for all variables, although the group factor and context variables must not be specified in the model code (see below). The regression model for variables with no parents besides the group factor and context variables can be omitted from the model code. In this illustration, we assume an endpoint-constrained quadratic lag shape between 0 and 15 time lags for all variables:

```
> mycode <- list(
+   GVA~quec(NPATENT,0,15),
+   PPI~quec(NPATENT,0,15)+quec(GVA,0,15),
+   ENTR_INCOME~quec(NPATENT,0,15)+quec(GVA,0,15)
+   )
```

The second step is to specify control options. Control options must be a named list containing one or more among several components. The key component is adapt, a named vector of logical values where each value must refer to one response variable and indicates if, for each lag shape in the regression model of that variable, the minimum and the maximum lag with a non-zero coefficient must be adapted (selected on the basis of the best fit to data), instead of employing the ones specified in the model code. If adaption is requested for a regression model, three further components are taken into account: max.gestation, min.width and sign. Each of these three components is a named list, where each component of the list must refer to one response variable and contain a named vector, including the maximum gestation lag, the minimum lag width and the sign (either '+' or '-') of the coefficients of one or more covariates. Here, we choose to perform adaptation of lag shapes for all regression models with the following constraints: (i) maximum gestation lag of 3 years, (ii) minimum lag width of 5 years, (iii) all coefficients with positive sign, excepting the ones in the regression model of the price index of agricultural products, as consumer surplus improves with the decreasing of prices:

```
> mycontrol <- list(
+   adapt=c(GVA=T,PPI=T,ENTR_INCOME=T),
+   max.gestation=list(GVA=c(NPATENT=3),PPI=c(NPATENT=3,GVA=3),
+     ENTR_INCOME=c(NPATENT=3,GVA=3)),
+   min.width=list(GVA=c(NPATENT=5),PPI=c(NPATENT=5,GVA=5),
+     ENTR_INCOME=c(NPATENT=5,GVA=5)),
+   sign=list(GVA=c(NPATENT="+"),PPI=c(NPATENT="-",GVA="-"),
+     ENTR_INCOME=c(NPATENT="+",GVA="+"))
+   )
```

Once model code and control options are specified, the structural model can be fitted using the command dlsem( ). The user can indicate a group factor to argument group and one or more context variables to argument context. By providing the group factor, one intercept for each level of the group factor will be estimated in each regression model. By providing context variables, they will be included as covariates in each regression model in order to eliminate spurious effects due to differences between the levels of the group factor. Furthermore, the user can decide to perform any number of the following operations:

- differentiation until the hypothesis of unit root is rejected by the Dickey-Fuller test for all the quantitative variables (by setting argument uniroot.check to TRUE);

- imputation of missing values for quantitative variables using the Expectation-Maximization algorithm [2] (by setting argument imputation to TRUE);
- apply the logarithmic transformation to all quantitative variables in order to interpret each regression coefficient as an elasticity (by setting argument log to TRUE).

Here we consider the country as group factor, provide gross domestic product and average farm size as context variables, allow differentiation until stationarity, imputation of missing values and logarithmic transformation for all quantitative variables:

```
> mod0 <- dlsem(mycode,group="COUNTRY",context=c("GDP","FARM_SIZE"),
+   data=agres,control=mycontrol,uniroot.check=T,imputation=T,log=T)
Checking stationarity...
 Order 1 differentiation performed
Starting EM...
EM iteration 1. Log-likelihood: 1394.8399
EM iteration 2. Log-likelihood: 1395.3395
EM iteration 3. Log-likelihood: 1395.4016
EM iteration 4. Log-likelihood: 1395.4073
EM iteration 5. Log-likelihood: 1395.4067
EM converged after 4 iterations. Log-likelihood: 1395.4067
Start estimation...
 Estimating regression model 1/4 (NPATENT)
 Estimating regression model 2/4 (GVA)
 Estimating regression model 3/4 (PPI)
 Estimating regression model 4/4 (ENTR_INCOME)
Estimation completed
```

After fitting the structural model, the user can display the DAG including only the edges associated to statistically significant estimates of regression coefficients, coloured according to the sign of the estimates (green for positive, red for negative):

```
> plot(mod0)
```

The result is shown in Figure 4.

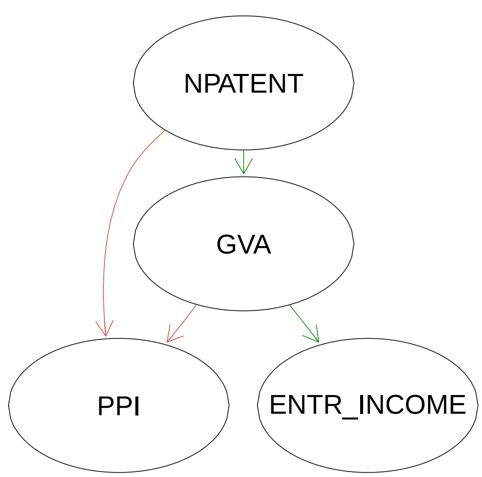

Figure 4: The DAG showing only the edges associated to statistically significant estimates of regression coefficients. Positive regression coefficients are shown in green. Negative regression coefficients are shown in red.

We see that all edges are associated to statistically significant coefficients, excepting the one linking research activity to profitability. This provides evidence that the effect of research activity on consumer surplus is both direct and mediated by productivity, and the effect of research activity on profitability is only mediated by productivity.

The user can also request the summary of model fitting:

```
> summary(mod0)
```

```
$NPATENT

Call:
"NPATENT ~ COUNTRY+GDP+FARM_SIZE"

Residuals:
    Min      1Q  Median      3Q     Max 
-3.6255 -0.2156  0.0172  0.2146  3.8613 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
factor(COUNTRY)AT -0.032677   0.153155  -0.213    0.831
factor(COUNTRY)BE -0.051062   0.153745  -0.332    0.740
factor(COUNTRY)DE -0.029085   0.153605  -0.189    0.850
factor(COUNTRY)DK -0.028752   0.153359  -0.187    0.851
factor(COUNTRY)EL  0.008343   0.151246   0.055    0.956
factor(COUNTRY)ES -0.033427   0.154347  -0.217    0.829
factor(COUNTRY)FI -0.012809   0.155401  -0.082    0.934
factor(COUNTRY)FR -0.061953   0.152594  -0.406    0.685
factor(COUNTRY)IE -0.080913   0.167465  -0.483    0.629
factor(COUNTRY)IT  0.001801   0.151346   0.012    0.991
factor(COUNTRY)NL -0.063467   0.153436  -0.414    0.679
factor(COUNTRY)PT  0.028596   0.151790   0.188    0.851
factor(COUNTRY)SE -0.093923   0.154125  -0.609    0.543
factor(COUNTRY)UK -0.102351   0.154367  -0.663    0.508
GDP                2.060751   1.586265   1.299    0.195
FARM_SIZE          0.049937   0.562659   0.089    0.929

Residual standard error: 0.686 on 278 degrees of freedom
  (14 observations deleted due to missingness)
Multiple R-squared:  0.008403,	Adjusted R-squared:  -0.04867 
F-statistic: 0.1472 on 16 and 278 DF,  p-value: 1


$GVA

Call:
"GVA ~ COUNTRY+quec(NPATENT,1,15)+GDP+FARM_SIZE"

Residuals:
      Min        1Q    Median        3Q       Max 
-0.298977 -0.034302  0.000572  0.041155  0.257996 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)  
factor(COUNTRY)AT   -7.015e-02  5.340e-02  -1.314   0.1935  
factor(COUNTRY)BE   -6.750e-02  4.757e-02  -1.419   0.1605  
factor(COUNTRY)DE   -2.994e-02  4.272e-02  -0.701   0.4858  
factor(COUNTRY)DK   -2.912e-02  3.948e-02  -0.737   0.4634  
factor(COUNTRY)EL   -1.265e-01  6.798e-02  -1.860   0.0672 .
factor(COUNTRY)ES   -1.297e-01  6.765e-02  -1.917   0.0595 .
factor(COUNTRY)FI   -3.056e-02  4.268e-02  -0.716   0.4765  
factor(COUNTRY)FR   -1.918e-02  3.789e-02  -0.506   0.6145  
factor(COUNTRY)IE   -8.036e-02  4.204e-02  -1.912   0.0602 .
```

```
factor(COUNTRY)IT   -5.455e-02  4.506e-02  -1.210   0.2303
factor(COUNTRY)NL   -2.338e-02  3.948e-02  -0.592   0.5557
factor(COUNTRY)PT   -1.879e-01  9.417e-02  -1.995   0.0501 .
factor(COUNTRY)SE   -4.723e-02  3.990e-02  -1.184   0.2406
factor(COUNTRY)UK    4.418e-05  3.602e-02   0.001   0.9990
theta0_quec.NPATENT  1.015e-01  4.750e-02   2.137   0.0362 *
GDP                  2.555e-01  3.358e-01   0.761   0.4494
FARM_SIZE            1.438e-01  1.372e-01   1.048   0.2982
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08788 on 67 degrees of freedom
  (224 observations deleted due to missingness)
Multiple R-squared:  0.1184,        Adjusted R-squared:  -0.1052
F-statistic: 0.5295 on 17 and 67 DF,  p-value: 0.9282


$PPI

Call:
"PPI ~ COUNTRY+quec(NPATENT,0,13)+quec(GVA,0,14)+GDP+FARM_SIZE"

Residuals:
      Min        1Q    Median        3Q       Max
-0.167506 -0.036284 -0.000584  0.045151  0.132116

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
factor(COUNTRY)AT    0.09617    0.02959   3.250  0.00169 **
factor(COUNTRY)BE    0.07313    0.02898   2.524  0.01360 *
factor(COUNTRY)DE    0.05803    0.02540   2.285  0.02499 *
factor(COUNTRY)DK    0.08315    0.02528   3.290  0.00149 **
factor(COUNTRY)EL    0.08880    0.03868   2.296  0.02432 *
factor(COUNTRY)ES    0.09384    0.03102   3.025  0.00334 **
factor(COUNTRY)FI    0.07735    0.02576   3.002  0.00357 **
factor(COUNTRY)FR    0.06450    0.02533   2.546  0.01281 *
factor(COUNTRY)IE   -0.01945    0.04607  -0.422  0.67398
factor(COUNTRY)IT    0.08050    0.02574   3.128  0.00245 **
factor(COUNTRY)NL    0.03607    0.02562   1.408  0.16303
factor(COUNTRY)PT    0.13945    0.04462   3.125  0.00248 **
factor(COUNTRY)SE    0.05435    0.02754   1.973  0.05193 .
factor(COUNTRY)UK    0.07131    0.02428   2.938  0.00432 **
theta0_quec.NPATENT -0.07098    0.02161  -3.285  0.00152 **
theta0_quec.GVA     -0.17540    0.07322  -2.395  0.01893 *
GDP                  2.04719    0.22917   8.933 1.19e-13 ***
FARM_SIZE            0.14364    0.09643   1.490  0.14027
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06331 on 80 degrees of freedom
  (210 observations deleted due to missingness)
Multiple R-squared:  0.641,        Adjusted R-squared:  0.5602
F-statistic: 7.934 on 18 and 80 DF,  p-value: 1.936e-11


$ENTR_INCOME

Call:
"ENTR_INCOME ~ COUNTRY+NPATENT+quec(GVA,1,14)+GDP+FARM_SIZE"
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.95895 -0.11610  0.00384  0.14344  0.58251

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
factor(COUNTRY)AT -0.03373    0.11244  -0.300  0.76497
factor(COUNTRY)BE -0.18861    0.12298  -1.534  0.12907
factor(COUNTRY)DE -0.09090    0.11230  -0.809  0.42067
factor(COUNTRY)DK -0.36121    0.11263  -3.207  0.00193 **
factor(COUNTRY)EL  0.09771    0.12763   0.766  0.44619
factor(COUNTRY)ES -0.10773    0.11720  -0.919  0.36074
factor(COUNTRY)FI -0.03752    0.10962  -0.342  0.73302
factor(COUNTRY)FR -0.10659    0.11392  -0.936  0.35225
factor(COUNTRY)IE  0.23779    0.16357   1.454  0.14994
factor(COUNTRY)IT  0.00781    0.10894   0.072  0.94303
factor(COUNTRY)NL -0.07936    0.11425  -0.695  0.48929
factor(COUNTRY)PT  0.03953    0.11007   0.359  0.72045
factor(COUNTRY)SE -0.07024    0.12249  -0.573  0.56795
factor(COUNTRY)UK -0.10613    0.10999  -0.965  0.33751
NPATENT           -0.03030    0.05813  -0.521  0.60360
theta0_quec.GVA    0.65386    0.30247   2.162  0.03363 *
GDP               -2.53157    0.98964  -2.558  0.01241 *
FARM_SIZE         -1.29697    0.43587  -2.976  0.00387 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2864 on 80 degrees of freedom
  (210 observations deleted due to missingness)
Multiple R-squared:  0.2795,	Adjusted R-squared:  0.1174
F-statistic: 1.724 on 18 and 80 DF,  p-value: 0.05181
```

The summary of model fitting returns estimates of parameters θ_j $(j = 1, \ldots, J)$. Instead, the command edgeCoeff( ) can be used to obtain estimates and confidence intervals of regression coefficients at the relevant time lags $\beta_{j,l}$ $(j = 1, \ldots, J;\ l = 0, 1, \ldots)$:

```
> edgeCoeff(mod0)
$`0`
                          2.5%         50%        97.5%
GVA~NPATENT          0.00000000  0.00000000  0.000000000
PPI~NPATENT         -0.02820862 -0.01766653 -0.007124428
PPI~GVA             -0.07474607 -0.04111016 -0.007474252
ENTR_INCOME~NPATENT  0.00000000  0.00000000  0.000000000
ENTR_INCOME~GVA      0.00000000  0.00000000  0.000000000

$`1`
                           2.5%         50%       97.5%
GVA~NPATENT          0.001971873  0.02379354  0.04561522
PPI~NPATENT         -0.052387442 -0.03280926 -0.01323108
PPI~GVA             -0.139526002 -0.07673897 -0.01395194
ENTR_INCOME~NPATENT  0.000000000  0.00000000  0.00000000
ENTR_INCOME~GVA      0.015190326  0.16273946  0.31028860

$`2`
                          2.5%         50%       97.5%
GVA~NPATENT          0.00368083  0.04441462  0.08514840
PPI~NPATENT         -0.07253646 -0.04542821 -0.01831996
PPI~GVA             -0.19433979 -0.10688642 -0.01943306
ENTR_INCOME~NPATENT  0.00000000  0.00000000  0.00000000
```

```
ENTR_INCOME~GVA      0.02821061  0.30223043  0.57625026

$`3`
                          2.5%         50%       97.5%
GVA~NPATENT          0.00512687  0.06186322  0.11859956
PPI~NPATENT         -0.08865567 -0.05552337 -0.02239106
PPI~GVA             -0.23918743 -0.13155252 -0.02391761
ENTR_INCOME~NPATENT  0.00000000  0.00000000  0.00000000
ENTR_INCOME~GVA      0.03906084  0.41847291  0.79788498

$`4`
                           2.5%         50%       97.5%
GVA~NPATENT          0.006309994  0.07613934  0.14596869
PPI~NPATENT         -0.100745081 -0.06309473 -0.02544439
PPI~GVA             -0.274068932 -0.15073726 -0.02740559
ENTR_INCOME~NPATENT  0.000000000  0.00000000  0.00000000
ENTR_INCOME~GVA      0.047741025  0.51146689  0.97519275

$`5`
                           2.5%         50%       97.5%
GVA~NPATENT          0.007230202  0.08724300  0.16725579
PPI~NPATENT         -0.108804688 -0.06814231 -0.02747994
PPI~GVA             -0.298984290 -0.16444065 -0.02989701
ENTR_INCOME~NPATENT  0.000000000  0.00000000  0.00000000
ENTR_INCOME~GVA      0.054251164  0.58121237  1.10817358

$`6`
                           2.5%         50%       97.5%
GVA~NPATENT          0.007887493  0.09517418  0.18246087
PPI~NPATENT         -0.112834491 -0.07066610 -0.02849771
PPI~GVA             -0.313933504 -0.17266268 -0.03139186
ENTR_INCOME~NPATENT  0.000000000  0.00000000  0.00000000
ENTR_INCOME~GVA      0.058591258  0.62770936  1.19682747

$`7`
                           2.5%         50%       97.5%
GVA~NPATENT          0.008281867  0.09993289  0.19158391
PPI~NPATENT         -0.112834491 -0.07066610 -0.02849771
PPI~GVA             -0.318916576 -0.17540336 -0.03189014
ENTR_INCOME~NPATENT  0.000000000  0.00000000  0.00000000
ENTR_INCOME~GVA      0.060761304  0.65095786  1.24115441

$`8`
                           2.5%         50%       97.5%
GVA~NPATENT          0.008413325  0.10151913  0.19462492
PPI~NPATENT         -0.108804688 -0.06814231 -0.02747994
PPI~GVA             -0.313933504 -0.17266268 -0.03139186
ENTR_INCOME~NPATENT  0.000000000  0.00000000  0.00000000
ENTR_INCOME~GVA      0.060761304  0.65095786  1.24115441

$`9`
                           2.5%         50%       97.5%
GVA~NPATENT          0.008281867  0.09993289  0.19158391
PPI~NPATENT         -0.100745081 -0.06309473 -0.02544439
PPI~GVA             -0.298984290 -0.16444065 -0.02989701
ENTR_INCOME~NPATENT  0.000000000  0.00000000  0.00000000
ENTR_INCOME~GVA      0.058591258  0.62770936  1.19682747

$`10`
```

```
                          2.5%        50%      97.5%
GVA~NPATENT          0.007887493  0.09517418  0.18246087
PPI~NPATENT         -0.088655672 -0.05552337 -0.02239106
PPI~GVA             -0.274068932 -0.15073726 -0.02740559
ENTR_INCOME~NPATENT  0.000000000  0.00000000  0.00000000
ENTR_INCOME~GVA      0.054251164  0.58121237  1.10817358

$`11`
                          2.5%        50%      97.5%
GVA~NPATENT          0.007230202  0.08724300  0.16725579
PPI~NPATENT         -0.072536459 -0.04542821 -0.01831996
PPI~GVA             -0.239187432 -0.13155252 -0.02391761
ENTR_INCOME~NPATENT  0.000000000  0.00000000  0.00000000
ENTR_INCOME~GVA      0.047741025  0.51146689  0.97519275

$`12`
                          2.5%        50%      97.5%
GVA~NPATENT          0.006309994  0.07613934  0.14596869
PPI~NPATENT         -0.052387442 -0.03280926 -0.01323108
PPI~GVA             -0.194339788 -0.10688642 -0.01943306
ENTR_INCOME~NPATENT  0.000000000  0.00000000  0.00000000
ENTR_INCOME~GVA      0.039060838  0.41847291  0.79788498

$`13`
                         2.5%        50%       97.5%
GVA~NPATENT          0.00512687  0.06186322  0.118599563
PPI~NPATENT         -0.02820862 -0.01766653 -0.007124428
PPI~GVA             -0.13952600 -0.07673897 -0.013951937
ENTR_INCOME~NPATENT  0.00000000  0.00000000  0.000000000
ENTR_INCOME~GVA      0.02821061  0.30223043  0.576250263

$`14`
                         2.5%        50%       97.5%
GVA~NPATENT          0.00368083  0.04441462  0.085148405
PPI~NPATENT          0.00000000  0.00000000  0.000000000
PPI~GVA             -0.07474607 -0.04111016 -0.007474252
ENTR_INCOME~NPATENT  0.00000000  0.00000000  0.000000000
ENTR_INCOME~GVA      0.01519033  0.16273946  0.310288603

$`15`
                         2.5%       50%     97.5%
GVA~NPATENT         0.001971873 0.02379354 0.04561522
PPI~NPATENT         0.000000000 0.00000000 0.00000000
PPI~GVA             0.000000000 0.00000000 0.00000000
ENTR_INCOME~NPATENT 0.000000000 0.00000000 0.00000000
ENTR_INCOME~GVA     0.000000000 0.00000000 0.00000000
```

Path analysis can be performed using the command pathAnal( ). The user must specify one or more starting variables (argument from) and the ending variable (argument to). Optionally, specific time lags on which path analysis should be focused can be provided to argument lag, otherwise all the relevant ones are considered. Also, the user can choose whether instantaneous (argument cumul set to FALSE, the default) or cumulative (argument cumul set to TRUE) coefficients must be returned. Here we perform path analysis from research activity to profitability and consumer surplus at time lags 5, 10, 15, 20 and 25, requesting cumulative coefficients:

```
> pathAnal(mod0,from="NPATENT",to="ENTR_INCOME",lag=c(5,10,15,20,25),cumul=T)
$`NPATENT*GVA*ENTR_INCOME`
         2.5%      50%     97.5%
5  0.02513504 0.113583  0.202031
10 0.51067352 1.137859  1.765044
```

```
15 1.86019646 3.510748  5.161300
20 3.20971941 5.883638  8.557556
25 3.69525789 6.907913 10.120569

$overall
         2.5%      50%     97.5%
5  0.02513504 0.113583  0.202031
10 0.51067352 1.137859  1.765044
15 1.86019646 3.510748  5.161300
20 3.20971941 5.883638  8.557556
25 3.69525789 6.907913 10.120569
```

```
> pathAnal(mod0,from="NPATENT",to="PPI",lag=c(5,10,15,20,25),cumul=T)
```

```
$`NPATENT*GVA*PPI`
          2.5%         50%      97.5%
5  -0.09077204 -0.05435072 -0.0179294
10 -0.59019516 -0.39351888 -0.1968426
15 -1.55081124 -1.08108464 -0.6113580
20 -2.44436338 -1.71655161 -0.9887398
25 -2.84103140 -1.98134075 -1.1216501

$`NPATENT*PPI`
         2.5%        50%      97.5%
5  -0.4513380 -0.2826644 -0.1139909
10 -0.9752124 -0.6107570 -0.2463017
15 -1.1283449 -0.7066610 -0.2849771
20 -1.1283449 -0.7066610 -0.2849771
25 -1.1283449 -0.7066610 -0.2849771

$overall
        2.5%        50%      97.5%
5  -0.542110 -0.3370151 -0.1319203
10 -1.565408 -1.0042759 -0.4431443
15 -2.679156 -1.7877457 -0.8963352
20 -3.572708 -2.4232126 -1.2737170
25 -3.969376 -2.6880018 -1.4066272
```

The output of path analysis is a list of matrices, each containing estimates and confidence intervals of the coefficient associate to each path connecting the starting variables to the ending variable at the requested time lags. Also, estimates and confidence intervals of the overall coefficient is shown in the component named overall.

Since the logarithmic trasformation was applied to all quantitative variables, coefficients above are interpreted as elasticities, that is, for a 1% of patent applications more, profitability and consumer surplus are expected to grow by 6.6% and 1.8%, respectively, after 25 years.

The estimated lag shape associated to any overall causal effect can be displayed using the command lagPlot( ):

```
> lagPlot(mod0,from="NPATENT",to="ENTR_INCOME")
> lagPlot(mod0,from="NPATENT",to="PPI")
```

The result is shown in Figure 5.

4 Concluding remarks

Package `dlsem` is conceived to perform impact analysis, that is the quantitative assessment of the consequences on a system due to an investment. The illustration proposed in this tutorial applies

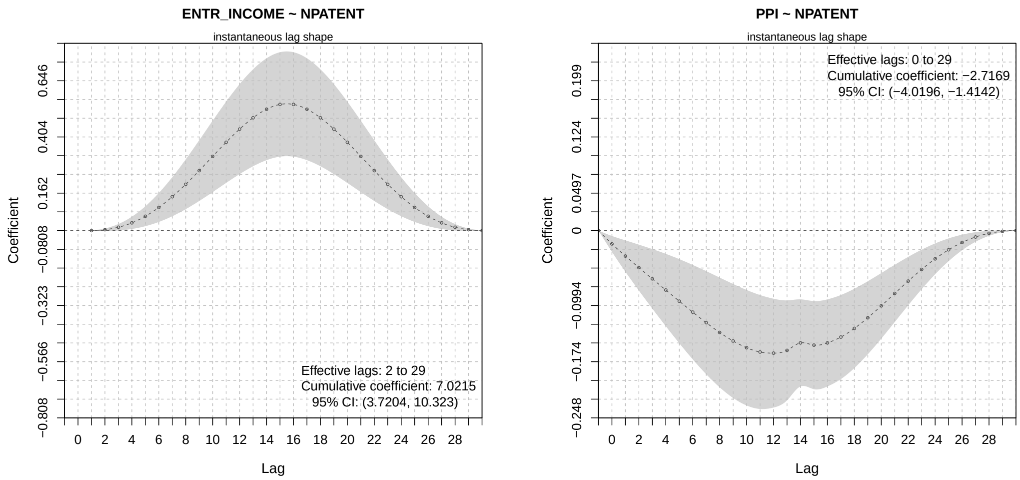

Figure 5: The estimated lag shape associated to the overall causal effect of research activity on profitability and consumer surplus. 95% confidence intervals are shown in grey.

impact analysis to a simplified problem of agricultural economics. The model here proposed can be extended by considering research investment as a parent of research activities, besides a larger number variables to better measure the macroeconomic state of the system.

Future implementation of package `dlsem` will include a graphical user interface allowing the user to exploit any of the implemented functionalities without writing R code.

References

[1] B. H. Baltagi (2008). Econometrics. Springer Verlag, 4th edition, Berlin, DE.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Series B, 39(1): 1-38.

[3] C. W. J. Granger, and P. Newbold (1974). Spurious Regressions in Econometrics. *Journal of Econometrics*, 2(2), 111-120.

[4] R. B. Kline (2010). Principles and Practice of Structural Equation Modelling. Guilford Press, 3rd edition, New York, US-NY.

[5] S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H. G. Leimer (1990). Independence Properties of Directed Markov Fields. *Networks*, 20(5): 491-505.

[6] W. K. Newey, and K. D. West (1978). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3), 703-708.

[7] A. Magrini, F. Bartolini, A. Coli, and B. Pacini (2016). Distributed-Lag Structural Equation Modelling: An Application to Impact Assessment of Research Activity on European Agriculture. *Proceedings of the 48th Meeting of the Italian Statistical Society*, 8-10 June 2016, Salerno, IT.

[8] J. Pearl (2012). The Causal Foundations of Structural Equation Modelling. In: R. H. Hoyle (ed.), Handbook of Structural Equation Modelling, Chapter 5. Guilford Press, New York, US-NY.

[9] S. Wright (1934). The Method of Path Coefficients. *Annals of Mathematical Statistics*, 5(3): 161-215.