REGAL: EXPERT SYSTEM FOR MULTIPLE LINEAR REGRESSION ANALYSIS*

BÉLA SIPOS – TIBOR KISS

REGAL – Expert System for Multiple Linear Regression Analysis, fulfils the task to transfer the expertise of the given area to the user, and it takes the model building easier. This paper analyses distributed lag models with the help of REGAL.

There are frequently stochastic type relationships in the area of the social and natural sciences. The spread of computers and computer sciences needs the software that is able to quantify these relationships. Regression analysis fulfils an outstanding role in this area. A computerised scientific method lays double claim: computer knowledge and the familiarity of the given scientific area, namely statistics in this case. REGAL also needs both types of knowledge, however, it strives to minimise them in the following way:

a) The program guides users throughout the program in order to facilitate its handling, and it always informs them what to do in given situation.

b) REGAL facilitates model building with constant help that can be general and specific knowledge as well, such as the explanation of the current model. The basic knowledge of the statistical methods allows – for the orientation of the user amongst the possibilities – to build up a correct regression model. The program unifies double expertise: the scientifically established, published knowledge of experts of mathematical statistics, and the expertise of those teachers of Janus Pannonius University – JPU, Faculty of Economics – who are familiar with the theoretical and practical side of this area.

The program is unique within this area, apart from those that were described above:

a) It examines the conditions of the regression system step by step; it shows the results of the statistical tests, then it warns the user to modify the model in the case of significant difference and it provides a modified estimation if it is necessary.

b) It provides probability levels for the better solution of the given problem where it is necessary and possible.

Data input part of the program has its own spreadsheet part, but at the same time it allows for managing other program data outputs (e.g. Dbase III, Lotus l-2-3, EXCEL), because it stores data in text files.

REGAL is applicable for analysing regression and time series models. The ordinary least square method (OLS) is used for parameter estimation. Transformation of data is

* The article was written in the framework of T 014722 and T 023058 research program of Hungarian Scientific Research Fund (OTKA).

also available. Following from the transformation possibilities, REGAL is able to estimate linear and linearisable models, such as half logarithmic, polynomial, lagged, *S* shaped, seasonal ones. A CESTRANS auxiliary program allows for CES function estimation in two steps with the help of REGAL. Average and marginal values are also available for performance evaluation. Lagged connections can also be modelled, where the reason and the consequence are split in time. Some types of qualitative data can also be used, e.g. in the case of wage regression. Detailed theoretical basis and examples are provided in the book of *Sipos, B.*: Company Forecasting.[1]

Model building

The first step in model building process is the specification that means the selection of dependent variable (*Y*) and independent variables (*X*), and the correct form of the function (*f*). However, the assumption of the distribution of the random variable can also be considered as specification. In this stage the model builder uses a priori information first of all. She or he applies pre-studies and knowledge which are in relationship with the studied phenomenon.

In this stage essential elements have to be handled, and database has an outstanding role. The quality and structure of database can influence the real specification. It is important to consider that the statement of the specification is always only a hypothesis that need further control and examination. Important stage in model building is the parameter estimation. This is a statistical part of it. It is important from the point of view of parameter estimation how to describe the regression relationship: with one or more variables; with one or more equations. A question that frequently emerges is: how does the model fit to different statistical conditions, determined by the OLS or other procedures? Finally it should be stated, that the results could only be hypothetical results. These hypotheses have to be controlled in the statistical control stage. This stage would influence the specification and the parameter estimation stage as well. The model builder decides whether to accept the model with a certain confidence, or not. He or she could change the model or the parameter estimation procedure. These decisions would need the repetition of the whole estimation procedure.

Regression models can be used basically for two types of tasks:

a) analysis of events,
b) forecasting.

That is: estimation of the value of parameters or the dependent variable. In the first case, with the so-called basis examinations, the objective is the analysis of the relationships between variables. In the second case – especially in the case of forecasting – it is also an important information, what elements are essential to influence significantly the process in the future.

In the verification stage the model is compared with reality. This is not merely a technical process. Economic, professional control is needed to decide whether variables

[1] *Sipos, B.*: Vállalati prognosztika. Janus Pannonius Egyetemi Kiadó. Pécs. 1995. 225 p. This book, together with the REGAL software, got the award of 'The most significant publication of the year' in 1996, given by the rector of JPU.

are correctly chosen, and the process is properly modelled or not. During verification the model builder can frequently get information that influence the earlier stages, such as specification.

Computations, parameter estimation of the model

After data input, a simple or stepwise regression can be asked for. In the case of simple regression, parameter estimation will be executed with all the selected variables, while in the case of stepwise regression only those variables are included step by step, where the variable can improve the explanation power of the model significantly (at 5 per cent significance level). This is evaluated by partial *F*-statistics, and the R^2 coefficient measures the models' effectiveness. Partial *F*-statistics and the referred probability levels are displayed by REGAL, and it recommends to include a variable where this probability value is the smallest and is smaller than 0.05 (5 %).

Results can be examined and analysed with REGAL globally, and in details as well. The summary screen after the computations diplays the global understanding of the estimated model. Multicollinearity, autocorrelation, homoscedasticity test, the multiple *R* square and test of normality can be seen in one window. If the users see OK in all the five test results, and the estimated partial coefficients are significant (on the basis of the calculated *t*-statistics value) then the regression model can be used for both analytic and forecasting purposes, because it fulfils the theoretical conditions of a regression model.

If any of the evaluation is BAD or questionable, then the model can be modified in the appropriate section of the program. E.g. in the case of significant autocorrelation, the user can ask for a special estimation, eliminating autocorrelation problems. These characteristics of the program will be used later in the discussion of distributed lag models.

Results of the current model can be seen globally, but also in details in the 'detailed analysis' part of the program. Results can be saved and printed from this menu item.

Estimation of Distributed Lag Models

In real economic life, it frequently appears that the effect of an economic event (or process) can not be recognised immediately, only later, with a certain lag in time. Typical example is the investment and the production (or sale) relationship between import and export prices. One group of models serves description or analysis of economic events, consequently it attempts to describe a given (or assumed) state (e.g. optimisation problem that looks for optimal production program with fixed prices and costs, and resources and demand). These models are called static models. Another group of models investigates the underlying motions in an economy (e.g. factors of economic growth). It answers the question: how to influence the actual state of events by the state of its own events or of other events. These models are dynamic models. Dynamic relationship means to investigate relationships of variables in different points of time. Static models can only model stable end state. Static relationships exist only for a short period of time. In the opinion of the classical Newton physics, the motion is continuos, consequently it can be described by a continuos function, and the influential factors take effect immediately, without lag. This tool is not effective in the case of economic processes, because

1. observations are connected to discrete and fixed time points. The result of measurement can be different in this way, if daily, weekly, monthly, quarterly, or yearly data are used; *2.* reaction time is not infinitely small, but finite, or even very long.

The modern physicists (Plank, Einstein, Heisenberg, etc.) re-evaluated the idea of time and motion. They mean that neither the motion, nor the time is continuos.

The reasons for lag (where the effect and reaction are split in time) are the following.

1. Recognition lag: The observation, registration, summation process need time (e.g. we do not buy durable articles frequently).

2. Decision lag: making and executing decisions need-time.

3. Technological lag (production time is the reason for it).

4. Lag because of inertia of processes.

5. Speculative lag.

6. Other reasons, e.g. organisational lag, slowness of bureaucracy, etc.

In econometrics *Alt, F. L.*, *Fisher, I.*, and others investigated the question in the second half of 30s' by separating the short and long term effect of independent variables, and estimating the long term effect. These models are not proved theoretically, therefore they are called naiv models. The research of distributed lag (DL) models took place in the 50s'. Among the researchers are *Koyck, L. M., Almon, S., Solow, R., Nerlove, M.*

Simple and compound models can model lag. In the simple model an event depends on the earlier state of another event. However, this earlier state should be a determined length of time. In the compound case, the investigated event (dependent variable) depends on the much earlier events of other events (independent variables). E.g.:

$$y_t = f(x_{t-1}, x_{t-2}, \dots, x_{t-k}) \qquad /1/$$

where

k – is the biggest assumed lag that can be infinite as well.
If $x = y$, then there is an autoregressive scheme.

The basic equation of compound lag model is (in linear case):

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + \beta_k X_{t-k} + \varepsilon$$

where:

Y_t – is the dependent variable (explained variable),
X_{t-i} – is the value of independent variable, in (t-i) period of time, where i – 0,1,2,...k is the lag,
ε – is error term, non-autocorrelated, with normal distribution, zero mean, constant variance,
α, β_i – are regression coefficients.

Parameters of this function are difficult to estimate in practice because of the necessary existence of multicollinearity and autocorrelation. On the basis of a priori knowledge and economic facts, limitations should be introduced: it should be assumed that there is a certain regularity with which independent variables influence dependent variable. This is the concept of distributed lag, when a limitation is applied to regression coefficients, e.g. it is assumed that these coefficients (βs) are decreasing by geometric progression (this is the traditional distributed lag model); or βs will increase at the beginning, and decrease afterwards.

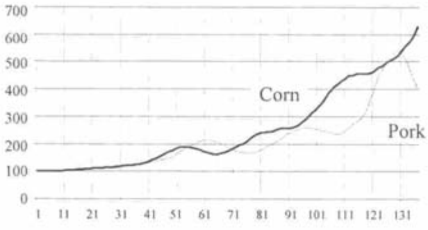
The most important Distributed Lag Models

The most important models and processes are demonstrated with the help of the following example. Databases are:

X – crumbled corn, monthly average price from 1985 VIII. 1. to 1997. XII. 31, 149 monthly values,
Y – pork, monthly average price on animal markets, for the same time period.[2]

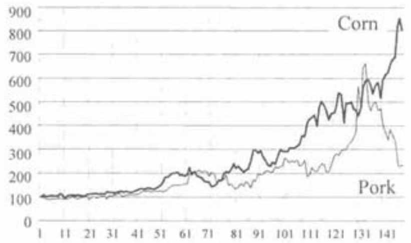
The lagged relation can be seen well from the figures below: corn prices (price movement, peaks) are followed by pork price movement. It means that pig breeders have increased cost, and will raise their prices. The reason for this lagged connection is technological, in this case it is the breeding period.

Figure 1. Monthly process of corn and pork moving average values
(137 monthly data)



January 1986 – 31 July 1997

F igure 2. Moving average prices of corn and pork
(149 monthly data)



August 1985 – 31 December 1997

Source: The corresponding volumes of Statisztikai havi közlemények (Central Statistical Office. Budapest.).

Simple Lag Models

Simple Lag Models are those where an event depends on its own (or other events') earlier values, but the length of this period can be determined.

$$Y_t = f_{(X_{t-i})} \qquad /2/$$

where

Y – is the dependent variable,
X – is the independent variable,
t – is the period of time (e.g. year, quarter, etc.),
i – 1, 2, 3…k is the lag.

[2] 12 element moving averages are calculated. They were divided by the first data and multiplied by 100 in order to make the two data series comparable.

In linear case:

$$Y_t = \alpha + \beta X_{t-i} + \varepsilon_i \qquad (i=1, 2, 3, \ldots j) \qquad /3/$$

If i=2 :

$$Y_t = \alpha + \beta X_{t-2} + \varepsilon_t \qquad /4/$$

For i = 1, 2, 3, 4, and 5, equations are created where the names are: LIN1, LIN2, LIN3, LIN4 and LIN5 respectively. Testing procedure is running for all the models. Theoretical conditions are tested: whether the model is significant (R^2 , F-statistics), or not, when the estimated regression coefficients are significantly different from zero; independent variables are linearly independent, therefore the probability of multicollinearity is below the limit; there is no significant autocorrelation; the model is homoscedastic, the variance of residuum is constant. If there are more models fulfilling these conditions, the model where the fit is better is accepted: the R^2 value is larger. LIN0 means zero lag. Results are summarised in Table 1.

Table 1

Summary results of Simple Lag Models

Statistics	LIN0 (i=0)	LIN1 (i=1)	*LIN2 (i=2)*	LIN3 (i=3)	LIN4 (i=4)	LIN5 (i=5)
R^2	0.62	0.65	0.68	0.70	0.71	0.71
	OK	OK	OK	OK	OK	OK
D–W-statistics	0.2	0.2	0.21	0.27	0.26	0.26
	BAD	BAD	BAD	BAD	BAD	BAD
Homoscedasticity	h=2.96	h=2.92	h=2.74	h=2.56	h=2.53	h=2.51
	p=1	p=1	p=1	p=1	p=1	p=1
	OK	OK	OK	OK	OK	OK
α	10.3	8.8	7.3	6.47	6.95	7.57
t-statistics	1.5	1.345	1.16	1.07	1.86	1.26
	BAD	BAD	BAD	BAD	BAD	BAD
β_{t-i}*	7.4	7.6	7.7.	7.87	7.9	7.93
t-statistics	15.5	16.6	17.7	18.7	18.3	18.8
	OK	OK	OK	OK	OK	OK

In the case of significant autocorrelation estimations, the ordinary least square method can be misleading. If the origin of autocorrelation can be found in the non-explained part of the regression model, iterative parameter estimation is reasonable instead of the modification of the model. A transformation with the help of the autocorrelation coefficient results in reduced autocorrelation value. This procedure can be described in the following way:

$$y_t - \bar{\rho} y_{t-1} = \beta_0 (1-\bar{\rho}) + \beta_1 (x_{1t} - \bar{\rho} x_{1t-1}) + \ldots + \beta_k (x_{kt} - \bar{\rho} x_{kt-1}) \qquad /5/$$

This procedure is known as Cochrane-Orcutt (CORC) iteration procedure.[3]

[3] See in *Hajdú, O. – Herman, S. – Pintér, J. – Rédey, K.*: Ökonometriai alapok. Tankönyvkiadó. Budapest. 1987. 59–61. p.

While β_k (i=1,2,...k) regression coefficients can be directly evaluated, β_0 parameter has to be transformed in order to fit to the original model.

$$\beta_0(1-\hat{\rho})=\alpha_0$$
$$\beta_0=\alpha_0/(1-\hat{\rho})$$

There is significant autocorrelation in all of the models. Applied CORC methods were inefficient, because the value of the *D–W*-statistics has not improved properly, and the explanation power of the model decreased at the same time. Therefore these models are not used for forecasting purposes. However, it can be determined that a 3-period lag seems to be the most efficient.

Naiv Distributed Lag Models

The model, elaborated by *Fisher, I.*,[4] is called naiv distributed lag model. Equation is based on a short cut theory: the effect of X variable on the Y variable is the largest one in the first period, and decreasing afterwards. Selection procedure is the same as before.

Fisher's equations:

1. Fisher 1 (*F*1)

$$Y_t=\alpha_0+\beta_1(2X_t+X_{t-1})+\varepsilon_t \qquad /6/$$

Let:

$$F_1=(2X_t+X_{t\text{-}1})$$
$$Y_t=\alpha_0+\beta_1F_1+\varepsilon_t$$

2. Fisher 2 (*F*2)

$$Y_t=\alpha_0+\beta_2(3X_t+2X_{t\text{-}1}+X_{t\text{-}2}) \qquad /7/$$

Let:

$$F_2=(3X_t+2X_{t\text{-}1}+X_{t\text{-}2})$$
$$Y_t=\alpha_0+\beta_2\,\mathrm{F}_2+\varepsilon_t$$

3. Fisher 3 (*F*3)

$$Y_t=\alpha_0+\beta_3(4X_t+3X_{t\text{-}1}+2X_{t\text{-}2}+X_{t\text{-}3}) \qquad /8/$$

Let:

$$F_3=(4X_t+3X_{t\text{-}1}+2X_{t\text{-}2}+X_{t\text{-}3})$$
$$Y_t=\alpha_0+\beta_3F_3+\varepsilon_t$$

Results can be transformed back in all cases into the original variables. Parameter estimation of α_0 is neglectible.

[4] *Fisher, I.*: Note on a short-cut method for calculating distributed lags. *Bulletin de l'Institut International de Statistique*. 1937. No. 3.

The results of model estimations are presented in Table 2.

Table 2

Results of Fisher's models

Name	F1	F2	F3
R^2	0.64	0.65	0.72
	OK	OK	OK
D–W-statistics	0.14	0.12	0.12
	BAD	BAD	BAD
Homoscedasticity	h=2.98	h=2.97	h=2.69
	p=1	p=1	p=1
	OK	OK	OK
α_0	8.34	7.45	3.94
t-statistics	1.2	1.1	0.65
	BAD	BAD	BAD
β_t	2.53	1.28	0.81
t-statistics	16.1	16.44	19.2
	OK	OK	OK

Autocorrelation is significant in all the models; the filtering procedure has not been efficient, therefore these models are not used later.

Alt-method[5]

In stage 1 Y_t value is explained by X_t value; in stage 2 by X_{t-1}, and so on, until the model is relevant.

1. Alt-method in stage 1: Alt1

$$Y_t = \alpha_0 + \beta_0 X_t + \varepsilon_t \qquad /9/$$

2. Alt-method in stage 2: Alt2

$$Y_t = \alpha_0 + \beta_0 X_t + \beta_1 X_{t-1} + \varepsilon_t \qquad /10/$$

Results of calculations are depicted in Table 3. They prove, that Alt methods are not appropriate in this case, because R^2 values do not differ significantly from zero. Alt's first method is the same as LIN0 which was shown before, therefore only the results of Alt2 method are:

R^2	0.65	Multicollinearity x^2	162	β_1	1.3
	OK		p=1.0	t-statistics	0.78
D–W-statistics	0.16		BAD		BAD
	BAD	α_0	8.07	β_2	6.34
Homoscedasticity	h=2.95	t-statistics	1.2	t-statistics	3.82
	p=1		BAD		OK
	OK				

[5] *Alt, F.L.*: Distributed Lags. *Econometrica*. 1942. No. 10.

Reversed V Lag Models

Weights increase for a certain time period, and decrease afterwards. This type of distribution of weights is called reversed *V* lag models.[6]

One method of estimating this type of reversed *V* lag models is elaborated by *Solow, R. M.*, and is called Pascal distributed lag model.[7] Koyck worked out the second method.[8] These types have common theoretical roots.

Equation /2/, which is the basic equation of distributed lag models, can be rewritten as:

$$Y_t = \alpha + \beta(w_0 X_t + w_1 X_{t-1} + w_2 X_{t-2} + \ldots + w_k X_{t-k} + \ldots) + \varepsilon_t \qquad /11/$$

where w_i are relative weights, a type of probability; consequently in equation /11/ the weights multiplied by the β result in the expected value (i=0,1,2,3,.....∞).

w_i weights (i=1,2,...k) in the case of Pascal distribution (negative binomial distribution) are the following:

$$\begin{matrix} 0 \le w_i \le 1 \\ \beta\sum_{i=0}^{\infty} w_i = 1 \end{matrix} \qquad w_i = \binom{r+i-1}{i}(1-\lambda)^r \lambda^i = \frac{(r+i-1)!}{i!(r-1)!}(1-\lambda)^r \lambda^i \qquad /12/$$

Considering, that n elements' k class combination is:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

where:

r – is the order of distribution; positive integer value: 1, 2, 3 …,
i – is the lag, 0, 1, 2, …k,
λ – is the parameter, estimated from the model.

For example, in the case of i=2:

$$w_2 = \binom{r+2-1}{2}(1-\lambda)^r \lambda^2 = \frac{(r+2-1)!}{2!(r-1)!}(1-\lambda)^r \lambda^2 = \frac{(r+1)r}{2!}(1-\lambda)^r \lambda^2 \qquad /13/$$

Equation /11/ can be rewritten as:

$$Y_t = \alpha + \beta(1-\lambda)^r \left\{ X_t + r\lambda X_{t-1} + \frac{r(r+1)}{2!}\lambda^2 X_{t-2} + \ldots + \frac{r(r+1)(r+2)\ldots(r+k-1)}{k!}\lambda^k X_{t-k} + \ldots \right\} + \varepsilon_t =$$
$$= \alpha + \beta(1-\lambda)^r \sum_{k=0}^{\infty} \binom{r+i-1}{i} \lambda^i X_{t-i} + \varepsilon_t$$

[6] *Vető, Istvánné*: A dinamika vizsgálata autoregresszív és osztott késleltetésű modellekkel. Központi Statisztikai Hivatal. Budapest. 1980. 47 p.

[7] *Solow, R. M.*: On a Family of Lag Distributions. *Econometrica*. 1960. No 2. 393–413. p.

[8] *Koyck, L. M.*: Distributed Lags and Investment Analyses. North-Holland Publishing Co. Amsterdam. 1954. 100 p.

If r=1 then

$$w_i = \binom{r+i-1}{i}(1-\lambda)^r \lambda = (1-\lambda)\lambda^i$$

$$(1-\lambda)\beta = \beta_i$$

the Pascal distribution is reduced to geometric lag distribution. If the value of r increases, then the turning point of the increasing period occurs at larger lag period. Geometric lag model is, using Equation /2/:

$$Y_t = \alpha + \beta_0(X_t + \lambda X_{t-1} + \lambda^2 X_{t-2} + \ldots + \lambda^k X_{t-k} + \ldots) + \varepsilon_t$$

Multiplied by β_0, and solved the equation of (Y_t -λY_{t-1}), Koyck's first method is derived:

$$Y_t = \alpha + \beta_0 X_t + \beta_0\lambda X_{t-1} + \beta_0\lambda^2 X_{t-2} + \ldots + \beta_0\lambda^k X_{t-k} + \ldots + \varepsilon_t \qquad /14/$$

$$\lambda Y_{t-1} = \lambda\alpha + \beta_0\lambda X_{t-1} + \beta_0\lambda^2 X_{t-2} + \beta_0\lambda^3 X_{t-3} + \ldots + \beta_0\lambda^k X_{t-k} + \ldots + \varepsilon_{t-1}$$

consequently:

$$Y_t - \lambda Y_{t-1} = \alpha(1-\lambda) + \beta_0 X_t + (\varepsilon_t - \lambda\varepsilon_{t-1}) \qquad /15/$$

because the other members will be eliminated. It can be seen, that in /15/ the value of the error term depends on the previous error term, and it has the autocorrelation problem in itself. Traditional *D–W*-statistics is not appropriate for testing this autocorrelation,[9] consequently the calculated *D–W*-statistics is only an informative test. However, this problem in practice is not so serious, especially if the sample size is big.[10]

A function is derived with the rearrangement of this equation, that is appropriate for estimation, considering, that

$$\alpha(1-\lambda) = \varpi$$

$$Y_t = \varpi + \beta_0 X_t + \lambda Y_{t-1} \qquad /16/$$

Parameters in /14/ can be estimated with OLS, with the help of the following relationship:

$$\varpi = \alpha(1-\lambda)$$

$$\alpha = \frac{\varpi}{(1-\lambda)}$$

λ value is derived from the model, because it is the regression coefficient of Y_{t-1}, while β_0 is the regression coefficient of X_t. It can be seen, that the quotient of any two neighbouring elements is a constant λ value that can have values between 0 and 1.

[9] *Kiss, T.*: Koyck és Solow modelljének felhasználása a döntéselőkészítésben. *Statisztikai Szemle*. 1985. No. 10. 1001–1011. p.

[10] *Hunyadi, L.*: Megosztott késleletetésű modellek. *Szigma*. 1980. No.1–2. 57–68. p.

Function that is appropriate for analysis and forecasting is:

$$Y_t = \alpha + \beta_0 X_t + \beta_0 \lambda X_{t-1} + \beta_0 \lambda^2 X_{t-2} + \ldots + \beta_0 \lambda^k X_{t-k} + \ldots + \varepsilon_t \qquad /17/$$

Using the sum of β parameters, the accumulated effect is also available. An example for this is: if the independent variable is the sum of the investments in operation, then the accumulated effect is 1, because usually all of the investments will be in operation after a while.

$$\beta_0 (1 + \lambda + \lambda^2 + \lambda^3 + \ldots) = 1$$

Let the sum of the geometric sequence be s, with q_0 starting value (β_0 in this case), and q is the quotient of two neighbouring elements (λ in this case) which is constant.

$$s = q_0 (\frac{q^n - 1}{q - 1}) = \beta_0 \frac{\lambda^n - 1}{\lambda - 1} = \beta_0 \frac{1}{1 - \lambda} = 1 \qquad \beta_0 = 1 - \lambda$$

because

$$\lim_{n \to \infty} \lambda^n = 0$$

where $0<\lambda<1$.

In this example the result of the estimation of Koyck's first model (Pascal distribution, r=1):

$$Y= -0.37+0.135X_t+1Y_{t-1}$$

λ=1, consequently the assumption of a geometric sequence is not proved. The measure of multicollinearity is high; the other tests were adequate.

Koyck's second model [11] differs from the first, because X_t and X_{t-1} variables have arbitrary weights,[12] and the geometric progress starts only afterwards. Because of the basic conception of geometric progression:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 (X_{t-1} + \lambda X_{t-2} + \lambda^2 X_{t-3} + \ldots + \lambda^{k-1} X_{t-k}) + \varepsilon_t$$

is the appropriate equation. The task is the calculation of α, β_0, β_1, and λ parameters. Multiplied by β and rearranged the equation, creating Y_t-λY_{t-1}:

$$\begin{aligned} Y_t &= \alpha + \beta_0 X_t + \beta_t X_{t-1} + \beta_1 \lambda X_{t-2} + \beta_1 \lambda^2 X_{t-3} + \ldots + \beta_1 \lambda^{k-1} X_{t-k} + \varepsilon_t \qquad /18/ \\ Y_{t-1} &= \alpha + \beta_0 X_{t-1} + \beta_1 X_{t-2} + \beta_1 \lambda X_{t-3} + \beta_1 \lambda^2 X_{t-4} + \ldots + \beta_1 \lambda^{k-1} X_{t-k-1} + \varepsilon_{t-1} \\ \lambda Y_{t-1} &= \lambda \alpha + \lambda \beta_0 X_{t-1} + \lambda \beta_1 X_{t-2} + \beta_1 \lambda^2 X_{t-3} + \beta_1 \lambda^3 X_{t-4} + \ldots + \beta_1 \lambda^k X_{t-k-1} + \lambda \varepsilon_{t-1} \\ Y_t - \lambda Y_{t-1} &= \alpha (1 - \lambda) + \beta_0 X_t + (\beta_1 - \lambda \beta_0) X_{t-1} \end{aligned}$$

[11] *Griliches, Z.*: Distributed Lags: A survey. *Econometrica*. 1967. No. 1. 16–49. p.; *Sipos, B.*: Industrial Price Forecasting. IGK. 1982. No. 4. 40–45. p.

[12] See Note 9.

After rearrangement:

$$Y_t = \alpha(1-\lambda) + \beta_0 X_t + (\beta_1 - \lambda\beta_0) X_{t-1} + \lambda Y_{t-1}$$

Estimation of Y_t dependent variable with the help of X_t, X_{t-1}, Y_{t-1} independent variables:

$$Y_t = \varpi + \beta_0 X_t + \mu X_{t-1} + \lambda Y_{t-1} \qquad /19/$$

where ϖ, β_0, μ and λ help to calculate β_1 and α values:

$$\beta_1 = \mu + \lambda\beta_0, \quad \alpha = \frac{\varpi}{1-\lambda} \quad \text{and} \quad \beta_i = \lambda \cdot \beta_{i-1}, i = 2,3,4,\ldots$$

The estimated function is:

$$Y = -0.568 - 0.518X_t + 0.757X_{t-1} + 1Y_{t-1}$$

$\lambda=1$, consequently the conclusion is the same as it happened to be in the case of Koyck's first model. The measure of multicollinearity is high; the other tests were adequate.

Koyck suggests a modified estimation to decrease autocorrelation, because Y_{t-1} is the independent variable in the model. The error term has autocorrelation as well. The modified estimation is also available in the REGAL software:

$$Y_t - \rho Y_{t-1} = \varpi(1-\rho) + \beta_1(X_{1t} - \rho X_{1(t-1)}) + \ldots + \beta_k(X_{kt} - \rho X_{k(t-1)}) \qquad /20/$$

where ρ is the first order autocorrelation coefficient (correlation coefficient between ε_t and ε_{t-1} residuum). The original constant parameter can be calculated by:

$$\varpi(1-\rho) = c, \quad \varpi = c/(1-\rho)$$

The average length of lag is:

$$\sum_{i=0}^{k} \beta_i i \Big/ \sum_{i=0}^{k} \beta_i$$

The meaning of $\sum_{i=0}^{k} \beta_i$, that is the sum of parameters is: increasing X variable with 1 unit results in that amount of increase in k periods (e.g. quarters) in the value of Y dependent variable.

Let $r = 2$. In this case the equation of Pascal lag is (Pascal 2):

$$y_t = \frac{\beta(1-\lambda)^2}{(1-\lambda L)^2} x_t + u_t \qquad /21/$$

where

L – is the lag operator,
r – is the order of distribution,
u – is the error term.

It can be seen that in /22/ the coefficients of y_{t-1} and y_{t-2} contain λ and only λ. This fact causes estimation problem,[13] where estimating λ can be calculated in the following.

$$\begin{aligned} y_t(1-\lambda L)^2 &= \beta(1-\lambda)^2 x_t & \qquad y_t - 2\lambda y_{t-1} + \lambda^2 y_{t-2} &= \beta(1-\lambda)^2 x_t \\ y_t(1-2\lambda L+\lambda^2 L^2) &= \beta(1-\lambda)^2 x_t & \qquad y_t &= 2\lambda y_{t-1} - \lambda^2 y_{t-2} + \beta(1-\lambda)^2 x_t \end{aligned} \qquad /22/$$

Write equation /22/ in the following way:

$$y_t = ax_t + by_{t-1} + cy_{t-2}$$

Estimation of coefficients λ can be computed:

$$\lambda = \frac{b \pm \sqrt{b^2-4c}}{2} \qquad /23/$$

where

$0<b<2$,
$-1<c<1$
$1-b-c>0$, and
$b^2=-4c$

The average lag is:

$$\theta = \frac{b+2c}{1-b-c}$$

where $b=\lambda_1+\lambda_2$, $c=-\lambda_1\lambda_2$; λ_1 and λ_2 are two roots for λ. In the example:

$$a=-0.074, \qquad b=0.9101, \qquad c=0.1014$$

Conditions are not fulfilled, because $1-b-c<0$. Therefore it is not worth calculating either the value of λ, or the average lag.

Almon's polynomial Distributed Lag Models

Almon assumed,[14] that weights of regression models, which contain predetermined lag period, follow a polynomial,[15] e.g. weights can be estimated by a linear function:

$$\begin{aligned} \hat{y}_t &= a_0 + b_0x_t + b_1x_{t-1} + \ldots + b_kx_{t-k} \\ b_0 &= d_0 + d_1 \\ b_1 &= d_0 + 2d_1 \\ &\vdots \\ b_k &= d_0 + (k+1)d_1 \end{aligned}$$

[13] See Note 11.

[14] *Mundruczó, Gy.*: Alkalmazott regressziószámítás. Akadémiai Kiadó. Budapest. 1981. 171–178. p.

[15] *Kiss, T.*: Almon osztott késleltetésű modelljének felhasználása a döntéselőkészítésben. *Statisztikai Szemle*. 1986. No. 2. 161–175. p.

If the original b_0, $b_1 \ldots b_k$ parameters are replaced by their estimations, then the following equation is created:

$$Y_t = \alpha + d_0 \sum_{i=0}^{k} X_{t-1} + d_1 \sum_{i=0}^{k} (i+1) X_{t-i}$$

Let:

$$\begin{aligned} D_0 &= \sum_{i=0}^{k} X_{t-i} \\ D_1 &= \sum_{i=0}^{k} (i+1) X_{t-i} \\ Y &= \alpha + d_0 D_0 + d_1 D_1 + \varepsilon_i \end{aligned} \qquad /24/$$

Let $k = 2$

$$D_0 = \sum_{i=0}^{2} X_{t-i} = X_t + X_{t-1} + X_{t-2}$$

$$D_1 = \sum_{i=0}^{k} (i+1) X_{t-i} = X_t + 2X_{t-1} + 3X_{t-2}$$

The estimation can be calculated in this way; Y is the dependent variable, D_0 and D_1 are independent variables, while parameters of α, d_0 and d_1 are estimated by the OLS.

In the example:

$$Y = 6.51 - 3.49 D_0 + 3.049 D_1$$

The model has significant autocorrelation and multicollinearity, therefore it can not be used for analytic or forecasting purposes.

Results of the models described above (Koyck 1 and 2, and Almon) are shown in Table 3.

Autocorrelation values of the two Koyck models are acceptable and, because of the sample size, in spite of the discussed one, bias is acceptable. Multicollinearity values are significantly large, that is a warning against analytic purposes. However, models can be used for forecasting purposes assuming that the relationship between independent variables remain unchanged.

These models have a drawback, that is X_t value should be forecasted in a way that would take an additional error in the extrapolation of Y value. In this case it is assumed that the last actual value will be repeated in the following period (X_t). Extrapolation is performed within the model, in other words the last period's value will be extrapolated (providing that last period's data is not known), which allows to check the validity of extrapolation.

The used X value is 14.2, that is the data of the 148th month (the month before the last month). This value is X_{t-1}, for extrapolating Y_t value for the 149th month. The used Y_{t-1} value is 318.8 that is also the value in month 148. Extrapolated values are depicted in Table 4.

Table 3

Results of Koyck's and Almon's models

Name	Koyck 1	Koyck 2	Almon
R^2	0,9798	0.979	0.686
	OK	OK	OK
D–W-statistics	2.15	2.04	0.14
	OK	OK	BAD
Homoscedasticity	2.129	2.16	2.85
	p=0.828	p=0.878	p=1.00
	OK	OK	OK
Multikollinearity x^2	61.7	228.16	350
	p=1.00	p=1.00	p=1.00
	BAD	BAD	BAD
α_0	-0.38	-0.56	6.05
t-statistics	-0.22	-0.34	1.02
	BAD	BAD	BAD
β_1	0.135	-0.51	-3.5
t-statistics	0.73	-1.26	-1.42
	BAD	BAD	BAD
β_2	1.00	0.757	3.05
t-statistics	49.56	1.783	2.05
	OK	OK	OK
β_3		0.994	
t-statistics		47.4	
		OK	

Table 4

Extrapolation with Koyck's models

	Month	X_t	Y_{t-1} value	Y_t value
Extrapolated *Y* – Koyck 1	149	14.20*	318.8	321.81
Extrapolated *Y* – Koyck 2	149	14.20*	318.8	319.8

* Accepted, as estimated X_t value from previous period.

The actual *Y* value in month 149 is 300 which means a big difference from the extrapolated value. Koyck's second model provided the better estimation, but the lower value of the confidence interval (5 per cent significant level) is 311.17, which is far higher than the actual value.

In the case of Koyck's models the estimation of an expected geometric lag was not successful, because the value of λ was 1 in both cases. That can also be the reason for the fact, that the extrapolation was not successful. Obviously, a sudden change in *Y* – that the model can not follow – may also be the reason.

As a summary it can be ascertained that the usage of distributed lag models for explaining the connection between the prices of corn and pork was not successful, because none of the models provided appropriate results.