Internext
Revista Eletrônica de Negócios Internacionais
São Paulo, v.13, n. 3, p. 42-50, sep/dec. 2018 | e-ISSN: 1980-4865 | http://internext.espm.br ESPM

Application of Regression Modeling to Data Observed Over Time

Cléber da Costa Figueiredo[1] e Aldy Fernandes da Silva

Higher School of Advertising and Marketing - ESPM, São Paulo, Brazil
Commerce School-Foundation Álvares Penteado - FECAP, São Paulo, Brazil

ARTICLE DETAILS

Article history:
Received: February 08 2018
Accepted July 12 2018
Available online September 01th 2018

Double Blind Review System

Scientifc Editor
Ilan Avrichir

Keywords:
longitudinal data
Stationarity
Autoregressive models
Granger causality
Lag

ABSTRACT

The central idea of this text is to guide researchers through the application of regression modeling when the data under analysis are observed over time. In general, there are no doubts regarding the application of this modeling in cross sections. However, when there is dependence on the data over time, s ome care needs to be taken for the results to be reliable and have the same interpretation of the coefficients obtained using the least squares method. The text begins with a presentation of the concept of autocorrelation and partial autocorrelation to identify and apply autoregressive modeling. Following this approach, the Augmented Dickey-Fuller test for detecting stationarity is presented, an essential condition for the estimators of ordinary least squares to be consistent. The Granger causality test is also presented and an example of regression applied to the series of the Cost of Living Index and the National Price Index for General Consumers. All the examples are presented with the help of Microsoft Excel to universalize the technique.

1. INTRODUCTION

The first question that needs to be answered before the data modeling is begun with the use of regression is whether these data are from a cross section, i.e., if the data were observed at the same moment in time, or whether they were collected over time.

When a cross section is defined as data collected at the same moment in time, there is no need for the data to have been collected all at once on a single day. What this definitions means is that since a sample element is observed, a single observation of it will be part of the sample.

Thus, in a cross sectional cut, the data can be collected in a month, a week or even on the same day. However, each element is observed only once. Furthermore, the usual linear regression techniques, learned in early statistics or econometrics courses (Gujarati, 2006; Sweeney, Williams, & Anderson, 2013; Wooldridge, 2011), apply to this type of study and will not be addressed in this text, because knowledge of these techniques will be the starting point for the analysis that will be conducted here.

On the other hand, in management it is common for a researcher to collect information from the same sample element over time. Numerous studies seek to evaluate the influence of the variation in a country's GDP on some other variable, or the number of motor vehicles imported by a certain country and the impact of this on a given variable over a period of time, or the variation of the cost of living in a country and its relationship with some other variable during a certain year or, finally, although not exhaustive, the path followed, in points, by some stock market indices, such as the index of the New York Stock Exchange (NYSE), the B3 Brazil Stock Exchange and Over-the-Counter Market, or the Tokyo Stock Exchange (TSE/TYO) and their interrelations.

[1] Corresponding author - Email: cfigueiredo@espm.br

DOI: 10.18568/1980-4865.13342-50

In this longitudinal analysis context, the variables load an index t that refers to the time in which the observation occurred, indicated by y_t, x_t, or z_t, for instance. In the specialized literature, these variables are identified as time series.

Put simply, studies can be divided into four major groups:

a) when the interest in prediction lies in using only the past observations of y_t as predictors. Mathematically, this means predicting y_{t+1} with past values of y_t through a model of $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_k y_{t-k} + \varepsilon_t$.

b) when the interest in prediction of y_t is linked to a possible relationship with another series x_t. The result of the final model may be similar to that of a simple regression such as $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$.

c) when the interest in prediction of y_t is linked to a possible relationship with various series x_t, z_t, etc. The result of the final model may be similar to that of a multiple regression $y_t = \beta_0 + \beta_1 x_t + \beta_2 z_t + \varepsilon_t$.

d) when, in time t, y_t is composed of various distinct elements j that will be observed longitudinally, as if a cross section j were observed over time t, or if j series were evaluated together, over time. Thus, the notation becomes y_{tj}, and a possible "simple" model would be represented by $y_{tj} = \beta_0 + \beta_1 x_{tj} + \varepsilon_{tj}$ or in multiple form by $y_{tj} = \beta_0 + \beta_1 x_{tj} + \beta_2 z_{tj} + \varepsilon_{tj}$.

Case (a) is known as an autoregressive model. Cases (b) and (c) may be treated as linear regressions, and the coefficients may be obtained using the ordinary least squares (OLS) method, as long as the samples are large and some properties that will be discussed later are valid. The last case is known as panel data, and will not be addressed in this text. In the four cases, the random error of the model ε will be assumed as normal and with constant variability over time.

If in linear regression it is necessary for the linearity to be valid for all to function well, with time series, the main characteristic for the least squares method to function is stationarity. The coefficients presented in the models of examples (a), (b) and (c) will be easily approximated by OLS, providing the samples are large and the series stationary.

2. THE AUTOREGRESSIVE CASE

It is not possible to understand the modeling of data over time without first understanding the concept of autocorrelation. Autocorrelation is a coefficient of correlation that measures the intensity of the relationship of the series with itself. Indeed, the measurement has the same usual metric of correlation. However, instead of being calculated between two variables, it is calculated between the series y_t and y_{t-1}, or between y_t and y_{t-2}, or y_t and y_{t-3}, and so forth.

The time distance established between the two series is known as a lag. An autocorrelation is of Lag 1 when the correlation between y_t and y_{t-1} is calculated; Lag 2 when the correlation between y_t and y_{t-2} is calculated; Lag k, when the correlation between y_t and y_{t-k} is calculated

For example, consider the National Price Index for General Consumers (IPCA) (IBGE, 2018), from January 2008 to December 2014. This is an inflation index made up of items with prices administered by the Brazilian Federal Government and other items with unregulated prices.

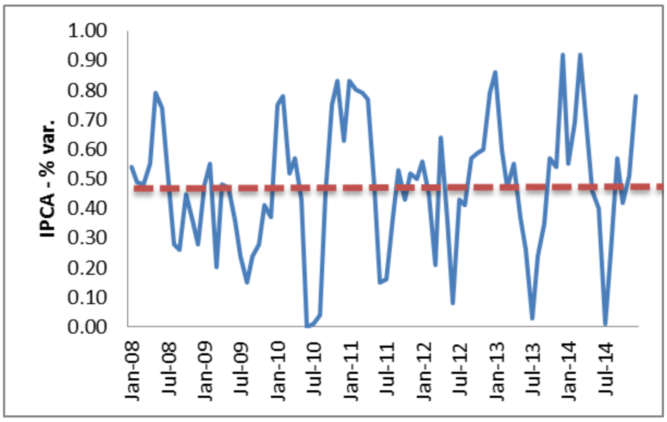
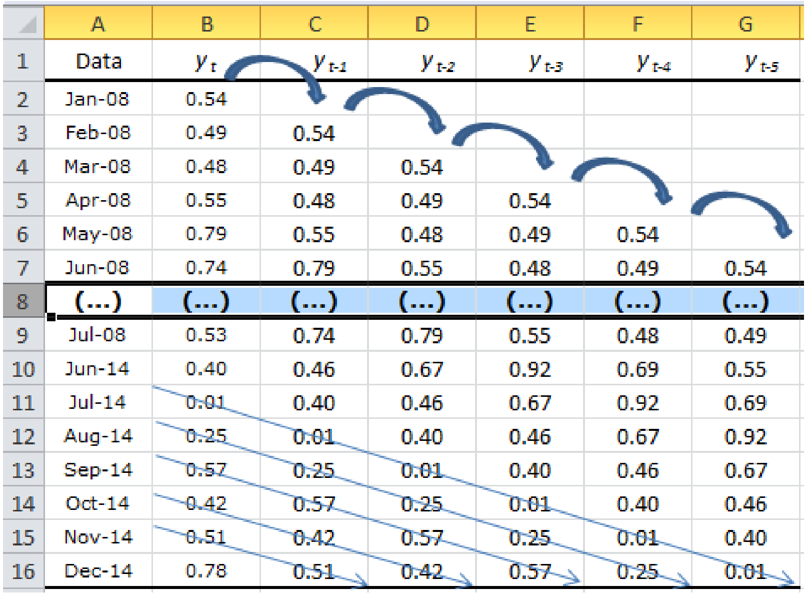


Fig. 1 Percentage variation of the IPCA with a constant imaginary central axis.

Figure 1 shows the percentage variation of the index. It is easily observed that this index apparently varies around a constant central axis. Furthermore, the variability around this imaginary central axis is also constant throughout t. This phenomenon is known as stationarity, which is a prerequisite for usual regression techniques to be applied within the context of data over time.



	A	B	C	D	E	F	G
1	Data	y_t	y_{t-1}	y_{t-2}	y_{t-3}	y_{t-4}	y_{t-5}
2	Jan-08	0.54					
3	Feb-08	0.49	0.54				
4	Mar-08	0.48	0.49	0.54			
5	Apr-08	0.55	0.48	0.49	0.54		
6	May-08	0.79	0.55	0.48	0.49	0.54	
7	Jun-08	0.74	0.79	0.55	0.48	0.49	0.54
8	(...)	(...)	(...)	(...)	(...)	(...)	(...)
9	Jul-08	0.53	0.74	0.79	0.55	0.48	0.49
10	Jun-14	0.40	0.46	0.67	0.92	0.69	0.55
11	Jul-14	0.01	0.40	0.46	0.67	0.92	0.69
12	Aug-14	0.25	0.01	0.40	0.46	0.67	0.92
13	Sep-14	0.57	0.25	0.01	0.40	0.46	0.67
14	Oct-14	0.42	0.57	0.25	0.01	0.40	0.46
15	Nov-14	0.51	0.42	0.57	0.25	0.01	0.40
16	Dec-14	0.78	0.51	0.42	0.57	0.25	0.01

Fig. 2 How to construct the lags in Excel.

In Excel, it is necessary to construct the lags, as shown in Figure 2. Observe that at each lag, an observation is lost at the end of the series. The Lag 1 autocorrelation equal to 0.59 is obtained by calculating the correlation of columns B and C (between y_t and y_{t-1}) that appear in Figure 2.

The Lag 2 autocorrelation equal to 0.24 is obtained by calculating the correlation of columns B and D (between y_t and y_{t-2}), and so forth.

Usually, a bar chart is constructed with the first k correlations and their decay is observed. If this decay is slow, there are signs that the process is autoregressive, i.e., the current time carries considerable information from past times (Bueno, 2011).

To know how many past values are relevant, so-called partial autocorrelation is used, which is nothing more than the estimated angular coefficient, $\hat{b}_k$, regarding the linear regression equation estimated by $\hat{y}_t = b_0 + b_1 y_{t-1} + b_2 y_{t-2} + \cdots + b_k y_{t-k}$.

Here, using Excel becomes laborious because to calculate the first-order partial autocorrelation it is necessary to find the "simple regression" between y_t and y_{t-1}; to calculate the second-order partial autocorrelation, it is necessary to find the multiple regression between y_t, y_{t-1} and y_{t-2}.

For the third order, another multiple regression between y_t and three of its lags (y_{t-1}, y_{t-2} and y_{t-3}) are required, and so forth.

As the partial autocorrelation is what defines the number of relevant lags in the model (Bueno, 2011; Morettin, 2008; Morettin & Tolói, 2006), it is necessary to use the following rule of significance to be significant.

A partial autocorrelation needs to be higher than $\pm 2/\sqrt{n}$ with significance of 5%. In the case of the series of the IPCA, with n = 84, the critical lower and upper limits are -0.22 and 0.22, respectively.

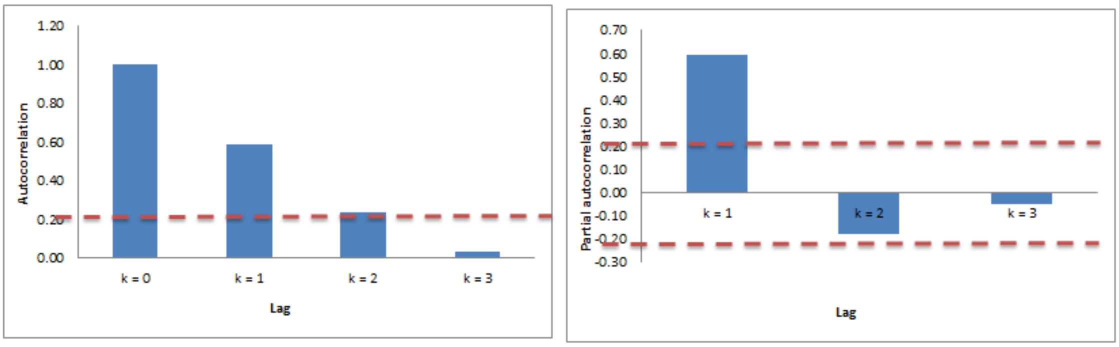


Fig. 3 Correlograms of the autocorrelation function and the partial autocorrelation function.

Figure 3 shows exactly how the identification of an autoregressive model functions. Its autocorrelations decay slowly (left-hand bar chart, Fig. 3), indicating dependence on the past and the only relevant dependence is at the time with a lag, as only the first partial autocorrelation (*k* = 1) is significant (right-hand bar chart, Fig. 3).

Thus, the autoregressive model, estimated for the variation of the IPCA, via regression, is: $\hat{y}_t = 0.19 + 0.60\ y_{t-1}$

Indicating that on average the variation of inflation at time t is 0.60 of the variation of time $t - 1$. The linear coefficient, 0.19, indicates that although the variation of the past month (time $t - 1$) was zero, at the current time t, there could be an expected variation of 0.19.

3. STATIONARITY

Everything that was discussed in the previous section will only be valid when there is stationarity. With this property, the OLS estimators become consistent. A series is said to be stationary when:

- $E(y_t) = \mu$ (the series must vary around a constant central axis); and
- $E[(y_t - \mu)(y_{t-k} - \mu)] = \gamma_k$ (the variability around this imaginary central axis must also be constant), for all t.

But how can the presence of stationarity be tested? Dickey and Fuller (1979) developed a method to verify the presence of stationarity called the Augmented Dickey-Fuller test. To bring it to Excel, imagine the autoregressive model with a lag, $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$, developed in the previous section.

As homoscedasticity is already assumed for the random error, ε_t, it is sufficient to verify whether the angular coefficient, β_1, is between the values of –1 and 1, exclusive.

Geometrically, this means that the values of the angular coefficient need to be within a circle of unit radius. However, they cannot be at the center (value = 0) of the circle or on the circumference.

In the literature, $\Delta y_t = y_t - y_{t-1}$ is called first-order differentiation. Thus, if:

- $\Delta y_t = y_t - y_{t-1}$, then, by substitution:
- $\Delta y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t - y_{t-1}$;
- $\Delta y_t = \beta_0 + (\beta_1 - 1)y_{t-1} + \varepsilon_t$□, when y_{t-1} becomes evident.

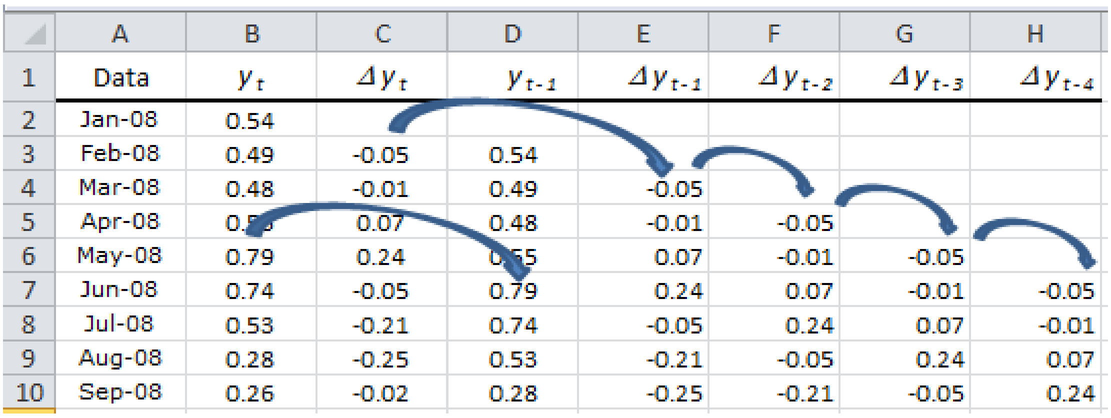
Thus, if it is thought that $(\beta_1 - 1)$□□□□, then it is sufficient to test the hypotheses: H₀:□□ = 0 against H₁: □ < 0 to gauge the presence or absence of stationarity.

The rejection of H₀ favors stationarity. The Augmented Dickey-Fuller test also considers *L* lags of Δy_t, added to the previous result. The number of lags of Δy_t, which are added when the test is performed, can be obtained through the whole value of $\sqrt[3]{n-1}$.

Thus, in Excel, the Augmented Dickey-Fuller test is performed when the estimated regression for the model is found:

$$\Delta y_t = \beta_0 + \rho y_{t-1} + \gamma_1 \Delta y_{t-1} + \cdots + \gamma_L \Delta y_{t-L} + \varepsilon_t. \quad \text{(Equation 01)}$$

If the statistic *t* observed for the coefficient □ is lower than –2.9, then H₀ is rejected, concluding that the series is stationary.



	A	B	C	D	E	F	G	H
1	Data	y_t	Δy_t	y_{t-1}	Δy_{t-1}	Δy_{t-2}	Δy_{t-3}	Δy_{t-4}
2	Jan-08	0.54						
3	Feb-08	0.49	-0.05	0.54				
4	Mar-08	0.48	-0.01	0.49	-0.05			
5	Apr-08	0.55	0.07	0.48	-0.01	-0.05		
6	May-08	0.79	0.24	0.55	0.07	-0.01	-0.05	
7	Jun-08	0.74	-0.05	0.79	0.24	0.07	-0.01	-0.05
8	Jul-08	0.53	-0.21	0.74	-0.05	0.24	0.07	-0.01
9	Aug-08	0.28	-0.25	0.53	-0.21	-0.05	0.24	0.07
10	Sep-08	0.26	-0.02	0.28	-0.25	-0.21	-0.05	0.24

Fig. 4 How to prepare Excel for the augmented Dickey-Fuller test.

In the example of the variations of the IPCA, *n* = 84, in other words, L = 4 should be considered in the model specified by Equation 01. Figure 4 shows how to create the lags of the original variable, y_t, and their difference Δy_t. With the spreadsheet prepared, an estimate for the test model is obtained. The solution is shown in Table 1. Observe that the observed statistic t for the coefficient of the variable y_{t-1} is – 5.49, lower than the critical value, –2.9, which indicates that the series of percentage variations of the IPCA may be considered stationary.

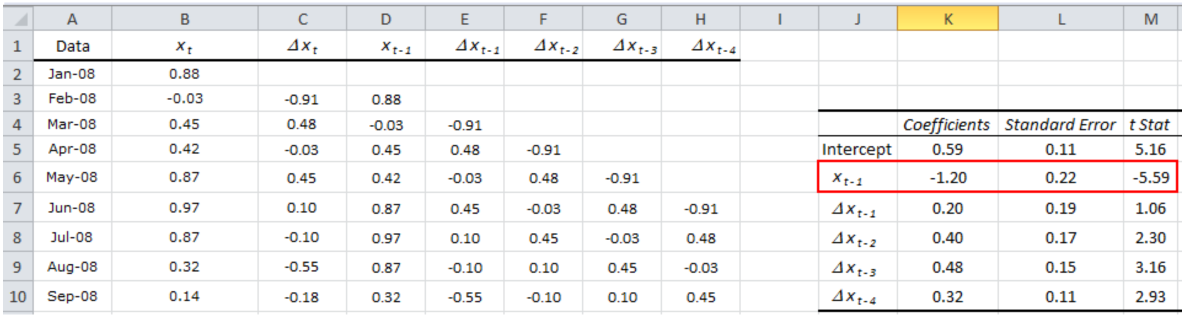
Tab. 1 Results of the augmented Dickey-Fuller test

	Coefficients	*Standard Error*	*t Stat*
Intercept	0.37	0.07	5.26
y_{t-1}	-0.79	0.14	-5.49
Δy_{t-1}	0.40	0.13	3.13
Δy_{t-2}	0.25	0.13	1.93
Δy_{t-3}	0.41	0.11	3.54
Δy_{t-4}	0.16	0.12	1.38

In general, percentage variation series are stationary. However, when a series is not stationary, it will be necessary to work with its differentiation, Δy_t. If, following the differentiation process, the series becomes stationary, it can be said that the original series is integrated in the first order.

4. THE CASE OF REGRESSION BETWEEN TWO SERIES: x_t AND y_t.

Given two time series x_t and y_t, the first task is to test for the presence of stationarity in both series. Thus, let x_t be the percentage variation series of the cost of living index (ICV), measured by the DIEESE (Inter-union Department of Statistics and Socio-economic Studies) (2018), from January 2008 to December 2014, and let y_t be the variation of the IPCA that was addressed above.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Data	x_t	Δx_t	x_{t-1}	Δx_{t-1}	Δx_{t-2}	Δx_{t-3}	Δx_{t-4}					
2	Jan-08	0.88											
3	Feb-08	-0.03	-0.91	0.88									
4	Mar-08	0.45	0.48	-0.03	-0.91						*Coefficients*	*Standard Error*	*t Stat*
5	Apr-08	0.42	-0.03	0.45	0.48	-0.91				Intercept	0.59	0.11	5.16
6	May-08	0.87	0.45	0.42	-0.03	0.48	-0.91			x_{t-1}	-1.20	0.22	-5.59
7	Jun-08	0.97	0.10	0.87	0.45	-0.03	0.48	-0.91		Δx_{t-1}	0.20	0.19	1.06
8	Jul-08	0.87	-0.10	0.97	0.10	0.45	-0.03	0.48		Δx_{t-2}	0.40	0.17	2.30
9	Aug-08	0.32	-0.55	0.87	-0.10	0.10	0.45	-0.03		Δx_{t-3}	0.48	0.15	3.16
10	Sep-08	0.14	-0.18	0.32	-0.55	-0.10	0.10	0.45		Δx_{t-4}	0.32	0.11	2.93

Fig. 5 The augmented Dickey-Fuller test for the cost of living index.

Following the Augmented Dickey-Fuller test, the observed value of t is lower than the critical value –2.9, which favors the hypothesis of stationarity, as shown in Figure 5. Thus, it can be said that the regression techniques can be applied to the series and the estimators of OLS will be consistent, always bearing in mind that the usual tests for the presence of normality and homoscedasticity still need to be verified.

After this comparison, the next step is to determine which series can be seen as regressive and which as a predictor. In the literature, this procedure is known as the Granger causality test (1969).

The method is consistent in using different lags of one series to predict the other. The idea is to determine whether x_t influences y_t, or whether y_t influences x_t, or even whether there is a reciprocal influence between the series (when this occurs, it is said that there is endogeneity in the model). In this latter case, there may be an exogenous series, z_t, that needs to be controlled or there may be a better candidate for the proposal of Granger causality. In this text, the issue of endogeneity will not be discussed.

The hypotheses of the Granger causality test are:

- H0: $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_k y_{t-k} + \varepsilon_t$ (restricted model);
- H1: $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_k y_{t-k} + \alpha_1 x_{t-1} + \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \alpha_k x_{t-k} + \varepsilon_t$ (unrestricted model).

The idea is to determine statistically whether x_t provides more information on future values of y_t than past values of y_t alone. Moreover, the opposite should also be tested:

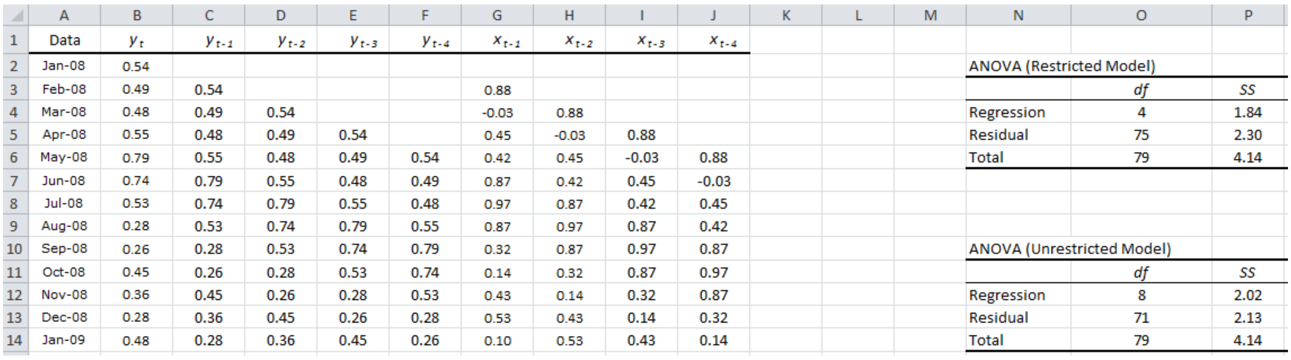
- H0: $x_t = \beta_0 + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \cdots + \beta_k x_{t-k} + \varepsilon_t$ (restricted model);
- H1: $x_t = \beta_0 + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \cdots + x y_{t-k} + \alpha_1 y_{t-1} + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \alpha_k y_{t-k} + \varepsilon_t$ (unrestricted model).

to know whether y_t provides more information on future values of x_t.

In both hypotheses, the test statistic is obtained through the following equation:

$$F = \frac{(SS_{restricted} - SS_{unrestricted})/g}{(SS_{restricted})/(n-p-1)} \quad \text{(Equation 02)}$$

In which *SS* are the sums of the squared residuals, obtained directly from the ANOVA regression tables; *g* is the number of coefficients that were zeroed (restrictions) in H0; *n* is the sample size effectively used in the regressions, already discounting the observations lost through lags; and *p* is the number of coefficients of the lagged variables in H1.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Data	y_t	y_{t-1}	y_{t-2}	y_{t-3}	y_{t-4}	x_{t-1}	x_{t-2}	x_{t-3}	x_{t-4}						
2	Jan-08	0.54												ANOVA (Restricted Model)		
3	Feb-08	0.49	0.54				0.88								*df*	*SS*
4	Mar-08	0.48	0.49	0.54			-0.03	0.88						Regression	4	1.84
5	Apr-08	0.55	0.48	0.49	0.54		0.45	-0.03	0.88					Residual	75	2.30
6	May-08	0.79	0.55	0.48	0.49	0.54	0.42	0.45	-0.03	0.88				Total	79	4.14
7	Jun-08	0.74	0.79	0.55	0.48	0.49	0.87	0.42	0.45	-0.03						
8	Jul-08	0.53	0.74	0.79	0.55	0.48	0.97	0.87	0.42	0.45						
9	Aug-08	0.28	0.53	0.74	0.79	0.55	0.87	0.97	0.87	0.42						
10	Sep-08	0.26	0.28	0.53	0.74	0.79	0.32	0.87	0.97	0.87				ANOVA (Unrestricted Model)		
11	Oct-08	0.45	0.26	0.28	0.53	0.74	0.14	0.32	0.87	0.97					*df*	*SS*
12	Nov-08	0.36	0.45	0.26	0.28	0.53	0.43	0.14	0.32	0.87				Regression	8	2.02
13	Dec-08	0.28	0.36	0.45	0.26	0.28	0.53	0.43	0.14	0.32				Residual	71	2.13
14	Jan-09	0.48	0.28	0.36	0.45	0.26	0.10	0.53	0.43	0.14				Total	79	4.14

Fig. 6 Preparation of the Excel spreadsheet for the application of the Granger causality test.

When Equation 02 is applied to the first set of hypotheses in which y_t = IPCA *and* x_t = ICV, *g* = 4; *n* = 80; *p* = 8 and the numerical expression of Equation 02, we have:

$$F = \frac{(2.30 - 2.13)/4}{(2.13)/(80 - 8 - 1)} = 1.47.$$

Thus, the p-value = 0.22 associated with this statistic is easily obtained using the distribution of the F statistic in Excel (*=1 – F.DIST(F = 1.47, g = 4, n – p – 1 = 71, 1)*). The conclusion of the test is that there is no evidence for rejecting the hypothesis that the IPCA only carries information about itself and the inclusion of lags in the ICV does not bring any predictive benefit to the IPCA.

On the other hand, the relationship in an inverse sense generates the following result:

$$F = \frac{(11.92 - 9.49)/4}{(9.49)/(80 - 8 - 1)} = 4.55.$$

Leading to p=value = 0.0025, favoring H_1. This result means that the IPCA can be used as a predictor of the ICV. In other words, the IPCA has a predictive characteristic when it comes to the ICV. In the literature, it is a common finding that “the IPCA, in the sense of Granger, causes the ICV”. Therefore, the ICV must be regressed on the IPCA to determine the impact of one on the other. Only now can the series be evaluated as in usual linear regression.

Tab. 2 Results of the regression of the ICV on the IPCA

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.04	0.09	0.48	0.63	-0.13	0.21
IPCA	0.95	0.16	5.77	0.00	0.62	1.28

Table 2 can be interpreted as a usual simple regression. Here, the relationship is arrived at through ICV = 0.95IPCA + 0.04, which indicates that at each increase in a unit of variation of the National Price Index for General Consumers, there is an increase in the cost of living index in the order of 0.95 with reliability of 95%. This means that a positive variation in inflation has the power to increase the variation of the cost of living index.

As the intercept is not significant, it can be removed from the regression equation, reevaluating a model that only takes the slope of the line into consideration: $x_t = \beta_1 y_t + \varepsilon_t$; or uses centralization $(x_t = \beta_0 + \beta_1(y_t - \bar{y}))$, in an attempt to seek a practical meaning for the intercept when the values of the cost of living series draw close to their average values.

Here, other lags of the variation of the ICV and the IPCA may be added to the model and the marginal p-values of these lags evaluated to determine how many of them should be used.

Furthermore, usual normality tests (Shapiro & Wilk, 1965) and homoscedasticity (Breusch & Pagan, 1979) are required to verify the usual regression assumptions. In addition, the test created by Ljung-Box (1978) must be performed to verify whether the residuals are not autocorrelated after the modeling. The idea is that the residuals behave like a white noise.

If an autocorrelation structure is detected in the residuals, it can be incorporated into a range of solutions described by Bueno (2011) or Morettin and Tolói (2006).

4. REFERENCES

Breusch, T. S., & Pagan, A. R. (1979). A Simple test for heteroskedasticity and random coefficient variation. Econometrica, 47(5), 1287–1294.

Bueno, R. L. S. (2011). Econometria. 2ª ed. São Paulo: Cengage Learning, 341 p.

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, 74(366), 427–431.

DIEESE (2018). Índice do custo de vida. Available at: https://www.dieese.org.br/analiseicv/icv.html. Accessed on 15 July 2018.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 37(3), 424–438.

Gujarati, D. (2006). Econometria básica. Rio de Janeiro: Elsevier, 812 p.

IBGE (2018). Índice Nacional de Preços ao Consumidor Amplo. Available at: https://ww2.ibge.gov.br/home/estatistica/indicadores/precos/inpc_ipca/defaultinpc.shtm. Accessed on 15 July 2018.

Ljung, G. M., & Box, G. E. P. (1978). On a measure of a lack of fit in time series models. Biometrika, 65(2), 297–303.

Morettin, P. A. (2008). Econometria financeira: um curso de séries temporais financeiras. São Paulo: Blucher, 319 p.

Morettin, P. A., & Tolói, C. M. C. (2006). Análise de séries temporais. São Paulo: Edgar Blücher, 535 p.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3–4), 591–611.

Sweeney, D. J., Williams, T. A., & Anderson, D. R. (2013). Estatística aplicada à administração e economia. 3ª ed. São Paulo: Cengage Learning, 692 p.

Wooldridge, J. M. (2011). Introdução à econometria: uma abordagem moderna. 4ª ed. São Paulo: Cengage Learning, 701 p.

- **Cléber da Costa Figueiredo** PhD degrees in statistics from the University of São Paulo – USP, São Paulo, (Brazil). He teaches multivariate statistics, classical and Bayesian inference, and regression and econometric models at the School of Business Administration of the Getúlio Vargas Foundation (EAESP-FGV) and at Higher School of Advertising and Marketing (ESPM). E-mail: cfigueiredo@espm.br Orcid id: http://orcid.org/0000-0002-1632-6625

- **Aldy Fernandes da Silva** PhD in Engineering from the University of São Paulo - USP, São Paulo, (Brazil). Teacher Researcher of the Master's Degree in Accounting Sciences of Commerce School-Foundation Álvares Penteado - FECAP. E-mail: aldy@fecap.br Orcid id: http://orcid.org/0000-0003-3686-9288

Aplicação da Modelagem de Regressão em Dados Observados ao Longo do Tempo

Cléber da Costa Figueiredo e Aldy Fernandes da Silva

Escola Superior de Propaganda e Marketing – ESPM, São Paulo, Brasil
Fundação Escola de Comércio Álvares Penteado – FECAP, São Paulo, Brasil

DETALHES DO ARTIGO

Histórico do Artigo*:*
Recebido: 08 de fevereiro de 2018
Aceito: 12 de julho de 2018
Disponível online: 01 de set. de 2018

Sistema de revisão "Double blind review"

Editor Científico
Ilan Avrichir

Palavras-chaves:
Dados longitudinais
Estacionariedade
Modelos autorregressivos
Causalidade de granger
Defasagem

RESUMO

A ideia central deste texto é orientar o pesquisador a aplicar a modelagem de regressão quando os dados em análise foram observados ao longo do tempo. Em geral, não há dúvidas da aplicação dessa modelagem em seções transversais. Contudo, quando há dependência dos dados ao longo do tempo, alguns cuidados precisam ser tomados para que os resultados sejam confiáveis e valham as mesmas interpretações dos coeficientes obtidos via o método de mínimos quadrados. O texto inicia com a apresentação do conceito de autocorrelação e de autocorrelação parcial, a fim de identificar e aplicar a modelagem autorregressiva. Após essa abordagem, é apresentado o teste de Dickey-Fuller Aumentado para a detecção de estacionariedade, condição essencial para que os estimadores de mínimos quadrados ordinários sejam consistentes. Também é apresentado o teste de causalidade de Granger e um exemplo de regressão aplicado às séries do Índice de Custo de Vida e Índice de Preços ao Consumidor Amplo. Todos os exemplos foram apresentados com a ajuda do Microsoft Excel, a fim de universalizar a técnica.

For access this article:

Figueiredo, C., & Silva, A. (2018). Aplicação da Modelagem de Regressão em Dados Observados ao Longo do Tempo. *Internext, 13*(3), 42-50. doi:http://dx.doi.org/10.18568/1980-4865.13342-50

For access this article: http://dx.doi.org/10.18568/1980-4865.13342-50