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

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

Cléber da Costa Figueiredo[1] 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.

1. INTRODUÇÃO

A primeira questão que precisa ser respondida antes de se iniciar a modelagem de dados com o uso de regressão é se esses dados são provenientes de uma seção transversal, ou seja, se são dados que foram observados no mesmo instante de tempo, ou se são dados coletados ao longo do tempo.

Quando se define um corte, ou seção, transversal como dados coletados no mesmo instante de tempo, não há a necessidade dos dados terem sido coletados todos de uma só vez, em um único dia. O que essa definição quer esclarecer é que, uma vez que um elemento amostral é observado, uma única observação dele fará parte da amostra.

Dessa maneira, em um corte transversal, os dados podem ser coletados em um mês, uma semana ou até mesmo um dia, porém cada elemento é observado uma única vez. Além disso, as técnicas usuais de regressão linear, aprendidas em cursos iniciais de estatística ou econometria (Gujarati, 2006; Sweeney, Williams, & Anderson, 2013; Wooldridge, 2011), aplicam-se a esse tipo de estudo e não serão tratadas neste texto, pois o conhecimento dessas técnicas será o ponto de partida para as análises que serão feitas aqui.

Por outro lado, em administração, é comum que o pesquisador colete informações de um mesmo elemento amostral ao longo do tempo. São inúmeros os estudos que procuram avaliar a influência da variação do PIB de um país sobre alguma outra variável, ou a quantidade de veículos automotores importados por um determinado país e o impacto disso em alguma variável ao longo de um determinado período, ou a variação do custo de vida de um país e a relação que isso

[1] Contato do autor - Email: cfigueiredo@espm.br

DOI: 10.18568/1980-4865.13342-50

tem com alguma outra variável durante um ano qualquer, ou, por fim, para não ser exaustivo, o caminho percorrido, em pontos, por alguns índices de bolsa de valores, tais como o índice da New York Stock Exchange (NYSE), o índice comercializado pela B3 (Ibovespa), ou o índice da Tokyo Stock Exchange (TSE/TYO) e suas inter-relações. Dentro desse contexto de análise longitudinal, as variáveis carregam um índice t que se refere ao tempo em que a observação ocorreu, sendo indicadas por y_t, x_t, ou z_t, por exemplo. Na literatura especializada, essas variáveis são identificadas como séries temporais.

De uma forma simplista, podem-se dividir os estudos em quatro grandes grupos:

a) quando o interesse de predição recai em utilizar apenas as observações passadas de y_t como preditoras. Matematicamente, trata-se de prever y_{t+1} com os valores passados de y_t por meio de um modelo do tipo $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_k y_{t-k} + \varepsilon_t$.

b) quando o interesse de predição de y_t está ligado a uma possível relação com outra série x_t. O resultado do modelo final pode ser semelhante ao de uma regressão simples do tipo $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$.

c) quando o interesse de predição de y_t está ligado a uma possível relação com várias séries x_t, z_t, etc. O resultado do modelo final pode ser semelhante ao de uma regressão múltipla $y_t = \beta_0 + \beta_1 x_t + \beta_2 z_t + \varepsilon_t$.

d) quando, no tempo t, y_t é composto por vários elementos j distintos que serão observados longitudinalmente. Como se uma seção transversal j fosse observada ao longo do tempo t. Ou ainda, como se j séries fossem avaliadas conjuntamente, ao longo do tempo. Assim, a notação passa a ser y_{tj}, e um possível modelo "simples" seria representado por $y_{tj} = \beta_0 + \beta_1 x_{tj} + \varepsilon_{tj}$ ou, de forma múltipla, por $y_{tj} = \beta_0 + \beta_1 x_{tj} + \beta_2 z_{tj} + \varepsilon_{tj}$.

O caso (a) é conhecido como um modelo autorregressivo. Os casos (b) e (c) podem ser tratados como regressões lineares e os coeficientes podem ser obtidos via o método de mínimos quadrados ordinários (MQO), desde que as amostras sejam grandes e algumas propriedades que serão discutidas adiante sejam válidas. Já, o último caso é conhecido como dados em painel e não será tratado neste texto. Nos quatro casos, o erro aleatório do modelo, ε, será suposto normal e com variabilidade constante ao longo do tempo.

Se na regressão linear, é preciso que a linearidade seja válida para que tudo funcione bem, com séries temporais, a principal característica para que haja funcionalidade do método de mínimos quadrados é haver estacionariedade. Os coeficientes apresentados nos modelos dos exemplos (a), (b) e (c) serão facilmente aproximados por MQO, desde que as amostras sejam grandes e as séries estacionárias.

2. O CASO AUTORREGRESSIVO

Não é possível entender a modelagem de dados ao longo do tempo, sem antes entender o conceito de autocorrelação. A autocorrelação é um coeficiente de correlação que mede a intensidade da relação da série consigo mesma. De fato, a medida tem a mesma métrica da correlação usual, contudo, ao invés de ser calculada entre duas variáveis, é calculada entre a série y_t e y_{t-1}, ou entre y_t e y_{t-2}, ou y_t e y_{t-3}, e assim por diante.

A distância temporal estabelecida entre as duas séries é chamada de defasagem. Uma autocorrelação é de defasagem 1, quando se calcula a correlação entre y_t e y_{t-1}; de defasagem 2, quando se calcula a correlação entre y_t e y_{t-2}; de defasagem *k*, quando se calcula a correlação entre y_t e y_{t-k}.

Como exemplo, considere o Índice Nacional de Preços ao Consumidor Amplo – IPCA (IBGE, 2018) – no período de janeiro de 2008 a dezembro de 2014. É um índice de inflação composto por itens com preços administrados pelo Governo Federal Brasileiro e outros itens com preços livres.

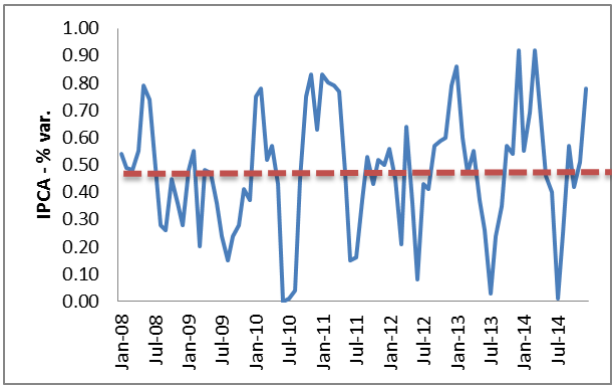

Fig. 1 Variação percentual do IPCA com eixo central imaginário constante.

A Figura 1 mostra a variação percentual do índice. É fácil observar que, aparentemente, esse índice varia em torno de um eixo central, constante. Além disso, a variabilidade em torno desse eixo central imaginário também é constante para todo t. Esse fenômeno é chamado de estacionariedade, que é o pré-requisito para que as técnicas de regressão usuais possam ser aplicadas dentro do contexto de dados ao longo do tempo.

	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 Como construir as defasagens no Excel.

No Excel, é preciso construir as defasagens, conforme mostra a Figura 2. Observe que a cada defasagem, uma observação é perdida no final da série. A autocorrelação de defasagem 1 igual a 0.59 é obtida, por meio do cálculo da correlação das coluna B e C (entre y_t e y_{t-1}) que aparecem na Figura 2.

A autocorrelação de defasagem 2 igual a 0.24 é obtida por meio do calculo da correlação das colunas B e D (entre y_t e y_{t-2}), e assim por diante.

Usualmente, constrói-se um gráfico de barras com as k primeiras autocorrelações e observa-se seu decaimento.

Se esse decaimento for suave, há indícios de que o processo seja autorregressivo, ou seja, o tempo atual carrega bastante informação dos tempos passados (Bueno, 2011).

Para saber quantos valores passados são relevantes, utiliza-se a chamada autocorrelação parcial que nada mais é do que o coeficiente angular estimado, $\hat{b}_k$, referente à equação de regressão linear estimada por $\hat{y}_t = b_0 + b_1 y_{t-1} + b_2 y_{t-2} + \cdots + b_k y_{t-k}$.

Aqui, utilizar o Excel se torna trabalhoso, porque para o cálculo da autocorrelação parcial de ordem 1, é preciso encontrar a "regressão simples" entre y_t e y_{t-1}; para o cálculo da autocorrelação parcial de ordem 2, é preciso encontrar a regressão múltipla entre y_t, y_{t-1} e y_{t-2}. Para a terceira ordem, outra regressão múltipla entre y_t e três defasagens suas (y_{t-1}, y_{t-2} e y_{t-3}). Assim por diante.

Como a autocorrelação parcial é quem define a quantidade de defasagens relevantes no modelo (Bueno, 2011; Morettin, 2008; Morettin & Tolói, 2006), então é preciso utilizar a seguinte regra de significância: para ser significativa, uma autocorrelação parcial precisa ser superior a $\pm 2/\sqrt{n}$ com significância de 5%.

No caso da série do IPCA, com *n* = 84, então os limites críticos, inferior e superior, são -0.22 e 0.22, respectivamente.

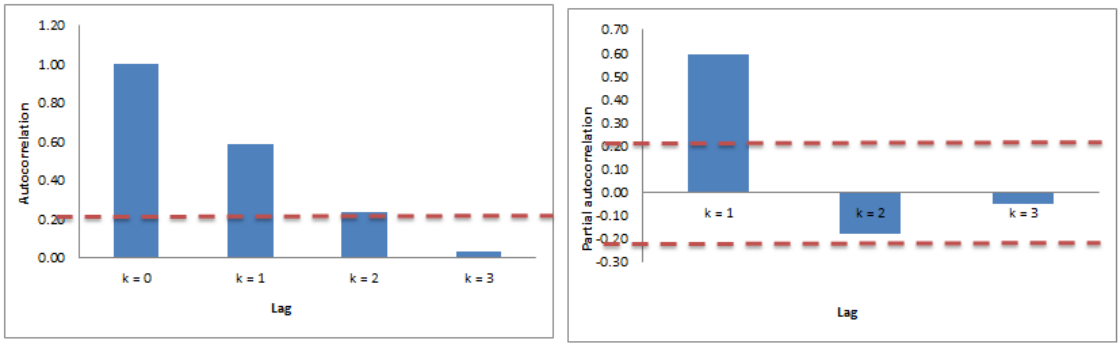

Fig. 3 Correlogramas da função de autocorrelação e da função de autocorrelação parcial.

A Figura 3 mostra, exatamente, como funciona a identificação de um modelo autorregressivo. Suas autocorrelações decaem suavemente (gráfico de barras à esquerda – Fig. 3), indicando a dependência com o passado e a única dependência relevante se refere ao tempo com uma defasagem, uma vez que apenas a primeira autocorrelação parcial (k = 1) é significativa (gráfico de barras à direita – Fig. 3).

Dessa maneira, o modelo autorregressivo, estimado para a variação do IPCA, via regressão, é:

$$\hat{y}_t = 0.19 + 0.60\, y_{t-1}$$

indicando que, em média, a variação da inflação no instante t é 0.60 da variação do instante $t - 1$. O coeficiente linear, 0.19, indica que, mesmo que a variação do mês passado (instante $t - 1$) fosse zero, no instante atual, t, poderia haver uma variação esperada de 0.19.

3. A ESTACIONARIEDADE

Tudo que foi discutido na seção anterior, só será válido quando houver estacionariedade. Com essa propriedade, os estimadores de MQO passam a ser consistentes. Uma série é dita estacionária quando:

- $E(y_t) = \mu$ (a série deve variar em torno de um eixo central constante); e
- $E[(y_t - \mu)(y_{t-k} - \mu)] = \gamma_k$ (a variabilidade em torno desse eixo central imaginário também deve ser constante), para todo t.

Mas como testar a presença de estacionariedade? Dickey e Fuller (1979) desenvolveram um método para verificar a presença de estacionariedade. É o chamado Teste de Dickey-Fuller Aumentado. Para trazê-lo para o Excel, imagine o modelo autorregressivo com uma defasagem, $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$, desenvolvido na seção anterior. Como a homoscedasticidade já é suposta para o erro aleatório, ε_t, basta verificar se o coeficiente angular, β_1, está contido entre os valores –1 e 1, exclusive.

Geometricamente, é dizer que os valores do coeficiente angular precisam estar contidos em um círculo de raio unitário, contudo, não podem nem estar no centro (valor = 0) do círculo, nem estar sobre a circunferência dele.

Na literatura, $\Delta y_t = y_t - y_{t-1}$ é chamada de diferenciação de primeira ordem. Dessa maneira, se:

- $\Delta y_t = y_t - y_{t-1}$, então, por substituição:
- $\Delta y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t - y_{t-1}$;
- $\Delta y_t = \beta_0 + (\beta_1 - 1) y_{t-1} + \varepsilon_t$, quando se coloca y_{t-1} em evidência.

Assim, se for pensado que $(\beta_1 - 1)$, então basta testar as hipóteses: H_0: = 0 contra H_1: < 0 para verificar a presença ou ausência de estacionariedade.

A rejeição de H_0 é favorável à estacionariedade. O chamado "teste de Dickey-Fuller Aumentado" ainda considera L defasagens de Δy_t, somadas ao resultado anterior. A quantidade de defasagens de Δy_t, que são acrescentadas na hora de realizar o teste, pode ser obtida por meio do valor inteiro da expressão $\sqrt[3]{n-1}$.

Desse modo, no Excel, o teste de Dickey-Fuller Aumentado é realizado ao se encontrar a regressão estimada para o modelo:

$$\Delta y_t = \beta_0 + \rho y_{t-1} + \gamma_1 \Delta y_{t-1} + \cdots + \gamma_L \Delta y_{t-L} + \varepsilon_t. \quad \text{(Equação01)}$$

Se a estatística t observada para o coeficiente for menor do que –2.9, então se rejeita H_0, concluindo que a série é estacionária.

	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 Como preparar o Excel para o teste de Dickey-Fuller Aumentado.

No exemplo das variações do IPCA, tem-se *n* = 84, ou seja, deve-se considerar L = 4, no modelo especificado pela Equação 01. A Figura 4 mostra como criar as defasagens da variável original, y_t, bem como, da sua diferença Δy_t. Após, ter preparado a planilha, obtém-se uma estimativa para o modelo do teste. A solução aparece na Tabela 1. Observe que a estatística t observada referente ao coeficiente da variável y_{t-1} é –5.49, menor que o valor crítico, –2.9, que indica que a série de variações percentuais do IPCA pode ser considerada estacionária.

Tab. 1 Resultados do Teste de Dickey-Fuller Aumentado

	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

Em geral, séries de variações percentuais são estacionárias. Contudo, quando uma série não for estacionária, será preciso trabalhar com sua diferenciação, Δy_t. Se após o processo de diferenciação, a série se tornar estacionária, diz-se que a série original é integrada de primeira ordem.

4. O CASO DA REGRESSÃO ENTRE DUAS SÉRIES: x_t E y_t.

Dadas duas séries temporais x_t e y_t, a primeira tarefa é testar a presença de estacionariedade em ambas as séries. Dessa forma, seja x_t a série da variação percentual do índice do custo de vida (ICV), medido pelo DIEESE (2018), durante o período de janeiro de 2008 a dezembro de 2014, e y_t a variação do IPCA, já trabalhada anteriormente.

	A	B	C	D	E	F	G	H
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			
5	Apr-08	0.42	-0.03	0.45	0.48	-0.91		
6	May-08	0.87	0.45	0.42	-0.03	0.48	-0.91	
7	Jun-08	0.97	0.10	0.87	0.45	-0.03	0.48	-0.91
8	Jul-08	0.87	-0.10	0.97	0.10	0.45	-0.03	0.48
9	Aug-08	0.32	-0.55	0.87	-0.10	0.10	0.45	-0.03
10	Sep-08	0.14	-0.18	0.32	-0.55	-0.10	0.10	0.45

	Coefficients	*Standard Error*	*t Stat*
Intercept	0.59	0.11	5.16
x_{t-1}	-1.20	0.22	-5.59
Δx_{t-1}	0.20	0.19	1.06
Δx_{t-2}	0.40	0.17	2.30
Δx_{t-3}	0.48	0.15	3.16
Δx_{t-4}	0.32	0.11	2.93

Fig. 5 O teste de Dickey-Fuller Aumentado para a série o índice do custo de vida.

Após a realização do teste de Dickey-Fuller Aumentado, obtém-se que o valor de t observado é menor do que o valor crítico –2.9, o que é favorável à hipótese de estacionariedade, conforme aparece na Figura 5. Dessa forma, pode-se afirmar que as técnicas de regressão poderão ser aplicadas às séries e os estimadores de MQO serão consistentes, sempre lembrando que os testes usuais de presença de normalidade e de homoscedasticidade ainda precisam ser averiguados.

Após essa aferição, o próximo passo é saber qual série pode ser vista como regressora e qual pode ser vista como preditora. Na literatura, esse procedimento é chamado de teste de causalidade de Granger (1969).

O método consiste em usar diferentes defasagens de uma das séries para predizer a outra. A ideia é desvendar se x_t influencia y_t, ou se y_t influencia x_t, ou ainda, se há influência recíproca entre as séries (quando isso ocorre, diz-se que há endogeneidade no modelo). Nesse último caso, é provável que exista uma série exógena, z_t, que precisa ser controlada ou que possa ser um melhor candidato para a proposta de causalidade de Granger. Neste texto, não será discutida a questão da endogeneidade.

As hipóteses do teste de causalidade de Granger são:

- H_0: $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_k y_{t-k} + \varepsilon_t$ (modelo restrito);
- H_1: $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_k y_{t-k} + \alpha_1 x_{t-1} + \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \alpha_k x_{t-k} + \varepsilon_t$ (modelo irrestrito).

A ideia é determinar estatisticamente se x_t fornece mais informações sobre valores futuros de y_t do que os valores passados de y_t sozinhos. Além disso, também se deve testar o contrário:

- H_0: $x_t = \beta_0 + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \cdots + \beta_k x_{t-k} + \varepsilon_t$ (modelo restrito);
- H_1: $x_t = \beta_0 + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \cdots + xy_{t-k} + \alpha_1 y_{t-1} + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \alpha_k y_{t-k} + \varepsilon_t$ (modelo irrestrito).

para saber se y_t fornece mais informação sobre valores futuros de x_t.

Em ambas as hipóteses, a estatística de teste é obtida por meio do quociente:

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

em que *SS* são as somas de quadrados dos resíduos, obtidas diretamente das tabelas de ANOVA de regressão; *g* é o número de coeficientes que foram zerados (restrições) em H_0; *n* é o tamanho da amostra efetivamente utilizado nas regressões, já descontadas as observações perdidas por defasagens; e *p* a quantidade de coeficientes das variáveis defasadas em H_1.

	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 Preparo da planilha do Excel para a aplicação do teste de causalidade de Granger.

Ao aplicar a Equação 02 ao primeiro conjunto de hipóteses em que y_t = IPCA *e* x_t = ICV, tem-se que *g* = 4; *n* = 80; *p* = 8 e expressão numérica da Equação 02, fica:

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

Desse modo, o valor-p = 0.22 associado a essa estatística é facilmente obtido, utilizando a distribuição da estatística F no Excel (*=1 – F.DIST(F = 1.47, g = 4, n – p – 1 = 71, 1)*). A conclusão do teste é

que não há evidências para se rejeitar a hipótese de que o IPCA só carregue informação de si mesmo e a inclusão de defasagens do ICV não traz nenhum benefício preditivo ao IPCA.

Por outro lado, a relação em sentido inverso, gera o seguinte resultado:

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

Que leva ao valor-p = 0.0025, favorável à H_1. Esse resultado quer mostrar que o IPCA pode ser utilizado como preditor do ICV. É dizer que o IPCA possui caráter preditivo sobre o ICV. É comum encontrar na literatura que "o IPCA causa, no sentido de Granger, o ICV". Assim, deve-se regredir o ICV sobre o IPCA para se encontrar o impacto de um sobre o outro. Só agora, as séries poderão ser avaliadas como em regressão linear usual.

Tab. 2 Resultados da regressão do ICV sobre o 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

A Tabela 2 pode ser interpretada como uma regressão simples usual. Aqui, a relação é dada por ICV = 0.95IPCA + 0.04, que indica que a cada aumento em uma unidade da variação do índice de preços ao consumidor amplo, há um aumento do índice do custo de vida na ordem de 0.95 com confiança de 95%. É dizer que uma variação positiva da inflação tem poder de aumentar a variação do índice do custo de vida.

Como o intercepto não se apresenta significativo, pode-se eliminá-lo da equação de regressão, reavaliando um modelo que só leve em consideração a inclinação da reta: $x_t = \beta_1 y_t + \varepsilon_t$; ou ainda, utilizar-se de centralização ($x_t = \beta_0 + \beta_1(y_t - \bar{y})$), a fim de tentar buscar um significado prático para o intercepto quando os valores da série do custo de vida se aproximam de seu valores médio.

Aqui, outras defasagens da variação do ICV e do IPCA podem ser acrescentadas ao modelo e os valores-p marginais dessas defasagens avaliados para saber quantas delas utilizar.

Além disso, testes usuais de normalidade (Shapiro & Wilk, 1965) e homoscedasticidade (Breusch & Pagan, 1979) precisam ser realizados para se verificar as suposições usuais de regressão. Além disso, o teste de Ljung e Box (1978) deve ser realizado para verificar se, após a modelagem, os resíduos são não autocorrelacionados. A ideia é que os resíduos se comportem da mesma maneira que um ruído branco.

Se alguma estrutura de autocorrelação for detectada nos resíduos, pode-se incorporar essa estrutura com uma gama de soluções que aparecem descritas em Bueno (2011) ou Morettin e Tolói (2006).

4. REFERÊNCIAS

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. Disponível em: https://www.dieese.org.br/analiseicv/icv.html. Acessado em 15 de julho de 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. Disponível em: https://ww2.ibge.gov.br/home/estatistica/indicadores/precos/inpc_ipca/defaultinpc.shtm. Acessado em 15 de julho de 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.

SOBRE OS AUTORES

- Cléber da Costa Figueiredo e Doutor em Estatística pela Universidade de São Paulo - USP, São Paulo, (Brasil). Professor da Escola Superior de Propaganda e Marketing – ESPM. E-mail: cfigueiredo@espm.br Orcid id: http://orcid.org/0000-0002-1632-6625

- Aldy Fernandes da Silva e Doutor em Engenharia pela Universidade de São Paulo - USP, São Paulo, (Brasil). Professor Pesquisador do Mestrado em Ciências Contábeis da Fundação Escola de Comércio Álvares Penteado – FECAP. E-mail: aldy@fecap.br Orcid id: http://orcid.org/0000-0003-3686-9288

Application of Regression Modeling to Data Observed Over Time

Cléber da Costa Figueiredo 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, some 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.

Para citar este artigo:

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

Para acessar este artigo: http://dx.doi.org/10.18568/1980-4865.13342-50