

UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ENGENHARIA DE SÃO CARLOS
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO

**PREVISÃO DO MERCADO DE PAGAMENTOS
ELETRÔNICOS BRASILEIRO POR REGRESSÃO
MÚLTIPLA**

Trabalho de conclusão de curso II

Pedro Henrique Bianchi Samora

Orientador: Prof. Dr. Mateus Cecílio Gerolamo

São Carlos

Dezembro/2014

PEDRO HENRIQUE BIANCHI SAMORA

**PREVISÃO DO MERCADO DE PAGAMENTOS
ELETRÔNICOS BRASILEIRO POR REGRESSÃO
MÚLTIPLA**

Trabalho de Conclusão de Curso
apresentado à Escola de Engenharia de
São Carlos da Universidade de São Paulo,
para obtenção do título de Graduação em
Engenharia de Produção Mecânica

Orientador: Prof. Dr. Mateus Cecílio Gerolamo

São Carlos

Dezembro/2014

Dedicatória

À minha família, base de tudo.

Agradecimentos

A Deus o dom da vida.

Aos meus pais a educação que me possibilitou chegar até aqui, a humildade que me permitiu aprender com meus erros e o amor que me fez quem eu sou.

A Carolin, minha namorada, que me ajudou a superar os momentos difíceis durante a produção deste trabalho e contribuiu com ideias e sugestões.

Aos professores Mateus e Herick a inesgotável paciência e a enorme dedicação em difundir conhecimento.

A todos os amigos que estiveram comigo durante estes anos de universidade, de turma e república, a excelente experiência proporcionada pela nossa convivência.

A USP, que muito me ensinou, desenvolveu e transformou.

Aos amigos do trabalho na Cielo o ambiente de trabalho de descontração e repleto de boas experiências e conselhos.

Resumo

SAMORA, P. H. B. **Previsão do Mercado de Pagamentos Eletrônicos Brasileiro por Regressão Múltipla**. Trabalho de Conclusão de Curso – Escola de Engenharia de São Carlos – Universidade de São Paulo (USP), 2014.

A análise de mercado é ferramenta estratégica fundamental para minimização de riscos e auxílio de planejamento de uma empresa. A importância desta ferramenta é ainda maior em mercados de grande concorrência, como é o caso dos pagamentos eletrônicos no Brasil. Este trabalho tem por objetivo propor uma equação que seja capaz de prever o volume transacionado pelo mercado de pagamentos eletrônicos. Para isso, foi usada a regressão múltipla em integração com séries temporais na forma de dados públicos disponibilizados por órgãos regularizadores. A utilização de séries temporais em modelos de regressão requer cuidados especiais para evitar problemas que podem enviesar a amostra, como heteroscedasticidade e autocorrelação. O trabalho acompanha o processo desde a seleção da técnica, seleção das variáveis e obtenção da equação final. Todo o desenvolvimento prático da regressão é suportada pelo *software* STATA 12. A equação final, composta por cinco variáveis independentes, atinge resultados satisfatórios, que são observados ao compará-la aos resultados reais da série histórica.

Palavras-chave: Análise de mercado; previsão de mercado; séries temporais; regressão múltipla; pagamentos eletrônicos; análise multivariada.

Abstract

SAMORA, P. H. B. **Market forecast for the Brazilian electronic payment market via multiple regression**. Senior project – Engineering School of São Carlos – University of São Paulo (USP), 2014.

In a company, the market analysis is the fundamental strategic tool for minimizing risks as well as an aid in its planning process. The importance of this tool is even greater in highly competitive markets, such as the electronic payments market in Brazil. The main objective of this paper is the proposal of an equation that is able to predict the volume traded by the Brazilian electronic payments market. To achieve this objective, multiple regression was used in integration with time series in the form of public data made available by regulating organs. The use of time series regression models requires special attentiveness to avoid problems that may bias the sample such as heteroscedasticity and autocorrelation. The present paper records the process since the selection of the technique, selection of the variables through to obtaining the final equation. All the practical development of the regression is supported by the software STATA 12. The final equation, composed of five independent variables, achieves satisfactory results, which are observed when comparing it to the actual results of the historic time series.

Keywords: Market analysis; market forecast; time series; multiple regression; electronic payments; multivariate analysis.

Sumário

1. INTRODUÇÃO	12
1.1. CONTEXTUALIZAÇÃO E JUSTIFICATIVAS	12
1.2. HIPÓTESES E QUESTÕES DE PESQUISA	13
1.3. OBJETIVO DE PESQUISA	14
1.4. ESTRUTURA DO TEXTO DO TCC	14
2. REVISÃO DE LITERATURA	16
2.1. SETOR DE PAGAMENTOS ELETRÔNICOS NO BRASIL	16
FUNCIONAMENTO DOS PAGAMENTOS ELETRÔNICOS NO BRASIL	17
2.2. ANÁLISE MULTIVARIADA	19
CRITÉRIOS PARA ESTIMADORES	21
REGRESSÃO SIMPLES	22
REGRESSÃO MÚLTIPLA	23
MULTICOLINEARIDADE, HETEROSCEDASTICIDADE E AUTOCORRELAÇÃO	24
SÉRIES TEMPORAIS	29
GRAUS DE LIBERDADE	32
CRITÉRIOS DE INFORMAÇÃO E SELEÇÃO DO MODELO	33
2.3. CONSOLIDAÇÃO DOS CONCEITOS	34
3. MÉTODO DE PESQUISA	37
4. ANÁLISE DOS RESULTADOS	42
5. CONCLUSÕES E DISCUSSÕES	46
REFERÊNCIAS BIBLIOGRÁFICAS	48
ANEXOS	50
ANEXO A – ESTATÍSTICAS DE DURBIN-WATSON COM 1% DE SIGNIFICÂNCIA	50
ANEXO B – ESTATÍSTICAS DE DURBIN-WATSON COM 5% DE SIGNIFICÂNCIA	51
APÊNDICES	52
APÊNDICE A – CÁLCULO DO CRITÉRIO DE INFORMAÇÃO DA VARIÁVEL PIB NOMINAL	52
APÊNDICE B – CÁLCULO DO CRITÉRIO DE INFORMAÇÃO DA VARIÁVEL CONSUMO DAS FAMÍLIAS	52
APÊNDICE C – CÁLCULO DO CRITÉRIO DE INFORMAÇÃO DA VARIÁVEL VENDAS DO VAREJO	53

Sumário

APÊNDICE D – CÁLCULO DO CRITÉRIO DE INFORMAÇÃO DA VARIÁVEL PEA	53
APÊNDICE E – CÁLCULO DO CRITÉRIO DE INFORMAÇÃO DA VARIÁVEL POPULAÇÃO OCUPADA	54
APÊNDICE F – CÁLCULO DO CRITÉRIO DE INFORMAÇÃO DA VARIÁVEL TAXA DE DESEMPREGO	54
APÊNDICE G – CÁLCULO DO CRITÉRIO DE INFORMAÇÃO DA VARIÁVEL MASSA SALARIAL	55
APÊNDICE H – CÁLCULO DO CRITÉRIO DE INFORMAÇÃO DA VARIÁVEL ENDIVIDAMENTO DAS FAMÍLIAS	55
APÊNDICE I – CÁLCULO DO CRITÉRIO DE INFORMAÇÃO DA VARIÁVEL INFLAÇÃO – IPCA	56
APÊNDICE J – CÁLCULO DO CRITÉRIO DE INFORMAÇÃO DA VARIÁVEL INFLAÇÃO – IGP-M	56
APÊNDICE K – CÁLCULO DO CRITÉRIO DE INFORMAÇÃO DA VARIÁVEL SELIC	57

Lista de Figuras

Figura 1– Ciclo de uma compra com cartão	17
Figura 2– A colinearidade vista segundo o gráfico de Venn	25
Figura 3– Modelos de séries temporais	29
Figura 4– Transformação da variável ‘mercadomi’	37
Figura 5 – Exemplo de cálculo do Critério de Informação	38
Figura 6– Teste de cointegração Engle-Granger	40
Figura 7– Obtenção da regressão múltipla.....	41
Figura 8– Obtenção da regressão múltipla.....	43
Figura 9– Resultado do teste de Durbin-Watson.....	44
Figura 10 – Cálculo do critério de informação da variável PIB nominal	52
Figura 11 – Cálculo do critério de informação da variável consumo das famílias	52
Figura 12 – Cálculo do critério de informação da variável vendas do varejo	53
Figura 13 – Cálculo do critério de informação da variável PEA	53
Figura 14 – Cálculo do critério de informação da variável população ocupada	54
Figura 15 – Cálculo do critério de informação da variável taxa de desemprego	54
Figura 16 – Cálculo do critério de informação da variável massa salarial	55
Figura 17 – Cálculo do critério de informação da variável endividamento das famílias	55
Figura 18 – Cálculo do critério de informação da variável inflação – IPCA	56
Figura 19 – Cálculo do critério de informação da variável inflação – IGP-M.....	56
Figura 20 – Cálculo do critério de informação da variável Selic.....	57

Lista de Gráficos

Gráfico 1– Evolução do mercado de pagamentos eletrônicos	16
Gráfico 2– Evolução do mercado de pagamentos eletrônicos brasileiro	35
Gráfico 3– Erros residuais vs. valores ajustados	42
Gráfico 4– Comparação valor real e equação de regressão obtida.....	45

Lista de Quadros

Quadro 1– Atores do ciclo de pagamento eletrônico.....	18
Quadro 2– Data em que ocorre a remuneração dos envolvidos	19

Lista de Tabelas

Tabela 1 – Resultado das regressões lineares	39
Tabela 2 – Aplicação da seleção forward.....	40
Tabela 3 – Estatística de Durbin-Watson: pontos dU e dL com 1% de significância	50
Tabela 4 – Estatística de Durbin-Watson: pontos dU e dL com 5% de significância	51

Lista de abreviaturas

ABECS – Associação Brasileira de Empresas de Cartões de Crédito e Serviços

AIC – Critério de informação de Akaike (*Akaike information criterion*)

AR – Autorregressivo

BIC – Critério de informação Bayesiano

BLUE – Melhor estimador linear não-viesado (*Best linear unbiased estimator*)

IBGE – Instituto brasileiro de geografia e estatística

IGP-M – Índice geral de preços - Mercado

IPCA – Índice nacional de preços ao consumidor amplo

MDR – *Merchant discount rate*

MQO – Mínimos quadrados ordinários

PEA – População economicamente ativa

PIB – Produto Interno Bruto

Selic – Sistema especial de liquidação e de custódia

SQR – Soma dos quadrados dos erros residuais

SQE – Soma dos quadrados da regressão

SQT – Soma dos quadrados total

1. Introdução

1.1. Contextualização e Justificativas

O mercado de pagamentos eletrônicos é relativamente novo do Brasil e surgiu concomitantemente à expansão do uso de cartões de crédito e débito como forma de pagamento em estabelecimentos comerciais, que se expandiu mais significativamente no início do século XXI, segundo dados da Associação Brasileira de Empresas de Cartões de Crédito e Serviços (ABECS, 2014).

Por ser um mercado muito jovem, ainda possui taxas de crescimentos elevadas, de cerca de 17% ao ano (ABECS, 2014), cenário que deve se manter por alguns anos. No entanto, ao mesmo tempo em que abre oportunidades para as empresas consolidadas do segmento, também permite que novas empresas almejem parcelas consideráveis do mercado. Novas empresas entram no mercado constantemente, e apresentam riscos para as empresas já consolidadas, uma vez que brigam por fatias do mercado. Esse cenário exige que todos os *players* estejam atentos às estratégias de seus concorrentes, de forma a estudar as iniciativas deles, suportando a elaboração de planos estratégicos coerentes com a visão de mercado e apoiando a tomada de decisões.

Amorim (2008) destaca que, entre outros problemas do mundo atual, a previsão por séries temporais é usada na análise de mercado, com objetivo de minimizar riscos e auxiliar o planejamento estratégico e as tomadas de decisões. Por tais motivos, a análise de mercado é uma ferramenta muito bem vista por executivos.

Segundo Corrar et al (2007), as análises podem ser classificadas de acordo com a quantidade de variáveis que são consideradas simultaneamente: univariadas, quando uma variável é analisada, bivariadas, quando duas variáveis são analisadas, ou multivariadas, quando mais de duas variáveis são analisadas ao mesmo tempo.

As técnicas classificadas como univariadas e bivariadas estão, em geral, ligadas a comportamentos lineares. No entanto, quando séries temporais são analisadas, é comum verificar que os dados não apresentam um comportamento linear.

O mercado de pagamentos eletrônicos, por exemplo, está ligado ao consumo no varejo, que é influenciado por diversas variáveis, como o consumo das famílias, a inflação, as taxas de juros, produto interno bruto (PIB), entre outros.

1. Introdução

A grande quantidade de variáveis nos induz, portanto, ao uso das análises classificadas como multivariadas. Entre as técnicas mais comuns, segundo Corrar et al (2007), estão a regressão múltipla, regressão logística, análise fatorial, análise de conglomerados e redes neurais.

Estudos abordando os temas considerados foram realizados anteriormente. Conde (2012), por exemplo, abordou a ligação entre o processo de inovação no mercado de cartões com ferramentas de *business intelligence*. O autor ainda afirma a importância do uso de ferramentas como as análises multivariadas no processo de utilização de bases complexas. No entanto, a abordagem das ferramentas teve por objetivo o acompanhamento da performance dos indicadores de negócio.

Já Régis e Artes (2008) fizeram uso do modelo multiestado de Markov de forma a analisar o risco associado ao uso de cartões de crédito. Partindo da premissa de que os possuidores de cartão de crédito podem assumir vários estados (em dia, rotativo, em atraso, etc), o modelo multiestado de Markov defende que a probabilidade de transição entre estes estados depende do tempo entre as transições e covariáveis referentes aos clientes.

Por outro lado, Amadeu (2011) faz uso de modelos de análise para predição por séries temporais, que nada mais são do que séries de observações de uma determinada variável ao longo do tempo. Amadeu compara o resultado da predição a partir de modelos lineares e também a partir de redes neurais, que são modelos computacionais inspirados nos neurônios, capazes de aprender e reter conhecimento, usando históricos de variáveis influentes como entrada para as predições.

1.2. Hipóteses e Questões de Pesquisa

O mercado de pagamentos eletrônicos pode ser influenciado por diversas variáveis, algumas econômicas, como PIB, outras variáveis de renda familiar, desemprego, inflação, entre outras. Este trabalho se baseará na hipótese que uma seleção destas variáveis, com seus respectivos pesos, melhor se relaciona com o volume total monetário que é transacionado pelas máquinas de cartão de crédito.

Portanto, a questão que deve guiar este trabalho é a seguinte:

Questão de pesquisa: Qual é uma possível equação para prever o volume transacionado pelo mercado de pagamentos eletrônicos?

1.3. Objetivo de Pesquisa

O objetivo deste trabalho é propor uma equação para explicação do comportamento e previsão do volume transacionado pelo mercado de pagamentos eletrônicos brasileiro. Para isso, será necessário atender alguns objetivos intermediários que ajudam a responder a questão principal.

Os objetivos intermediários citados passam por pontos críticos na obtenção de um modelo de previsão do mercado. O primeiro ponto a ser discutido é qual a técnica de análise multivariada deve ser utilizada. Depois, devem ser definidas as variáveis independentes que serão usadas e como elas se relacionam com a variável dependente. Por fim, é necessário saber se o modelo proposto possui uma capacidade suficientemente boa de predição.

1.4. Estrutura do texto do TCC

O capítulo 1, foi a introdução do trabalho e apresentou os comentários iniciais, como motivação e justificativa do estudo, assim como o objetivo final do trabalho. Além disso, foi realizada uma breve contextualização, na qual se explica o funcionamento do mercado e quem são os interessados nos métodos de previsão de mercado.

No capítulo 2, ocorre uma revisão da literatura, de forma a entender melhor os conceitos de cada uma das análises que serão abordadas durante o trabalho, que está dividida em:

- Setor de pagamentos eletrônicos no Brasil;
 - Funcionamento dos pagamentos eletrônicos no Brasil;
- Análises multivariadas;
 - Critério para estimadores;
 - Regressão simples;
 - Regressão múltipla;
 - Multicolinearidade, heteroscedasticidade e autocorrelação;
 - Séries temporais;
 - Graus de liberdade;
 - Critérios de informação e seleção de modelos;

1. Introdução

Por fim, o capítulo 2 ainda apresentará uma consolidação dos conceitos, resumindo quais assuntos estudados serão usados nos capítulos seguintes.

O terceiro capítulo trará a aplicação dos conceitos estudados, justificando o uso de cada uma das ferramentas e variáveis. Também no capítulo 3 serão atingidos os objetivos intermediários.

O capítulo 4 apresentará os resultados obtidos, além do questionamento crítico da equação final e hipóteses de melhoria de trabalho.

Por fim, no quinto capítulo, haverá a conclusão e sumarização do trabalho.

2. Revisão de Literatura

2.1. Setor de pagamentos Eletrônicos no Brasil

O mercado de pagamentos eletrônicos é relativamente novo no Brasil e foi alavancado, em partes pela estabilização da moeda brasileira com o Plano Real, em 1994. A estabilidade permitiu uma grande revolução na forma como o brasileiro vê e usa os cartões de crédito e débito, que são os carros-chefe do mercado atualmente (FIGUEIREDO, 2006).

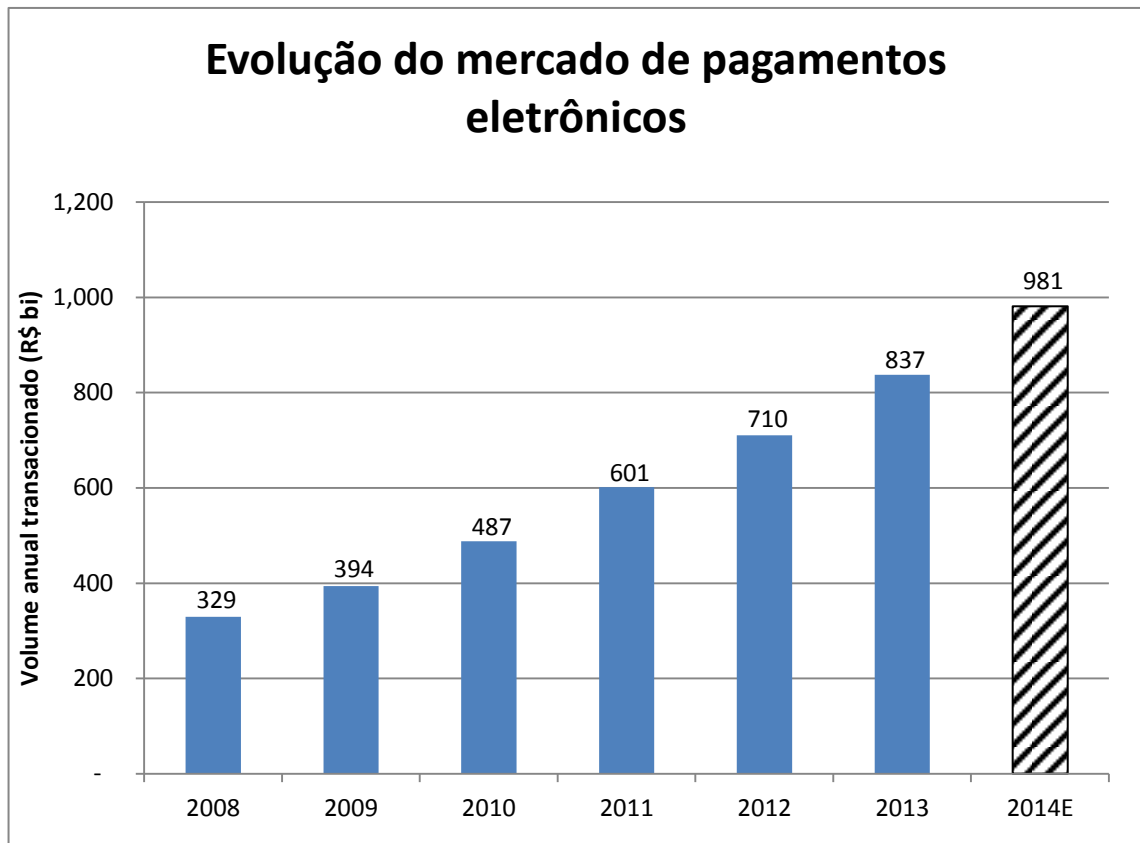


Gráfico 1– Evolução do mercado de pagamentos eletrônicos

Fonte: ABECS, 2014.

No entanto, ainda em 2014, o crescimento deve ser muito acelerado, cerca de 17%, alcançando a marca de R\$ 981 bilhões, segundo dados estimados pela associação brasileira das empresas de cartões de crédito e serviços (ABECS, 2014). Como mostra a Figura 1, as taxas de variação ano a ano do volume capturado pelas máquinas de cartão de crédito e débito são bem maiores do que as taxas de crescimento da economia. Esta discrepância nas taxas pode ser vista

quando comparados o mercado de pagamentos eletrônicos ao PIB que, em 2013, segundo o IBGE cresceu 2,3%.

Segundo Alves e Menezes (2007), outro fator preponderante para tal crescimento é a maior segurança e o conforto que as soluções eletrônicas proporcionam. O pagamento eletrônico, segundo os autores, surgiu da necessidade de saber com antecedência se a pessoa que vai realizar o pagamento está autorizada ou possui crédito para honrar a compra. Por tal motivo, há anos os cartões vêm substituindo os cheques como forma de pagamento.

Funcionamento dos pagamentos eletrônicos no Brasil

Os pagamentos eletrônicos podem ser divididos em três grandes frentes: débito, crédito à vista e crédito parcelado. O ciclo da compra é muito parecido nas três modalidades e diferem apenas quanto às datas de recebimento do valor da venda.

A Figura 1 ilustra o ciclo e cada um dos envolvidos quando é realizada uma compra com cartão de crédito ou débito.

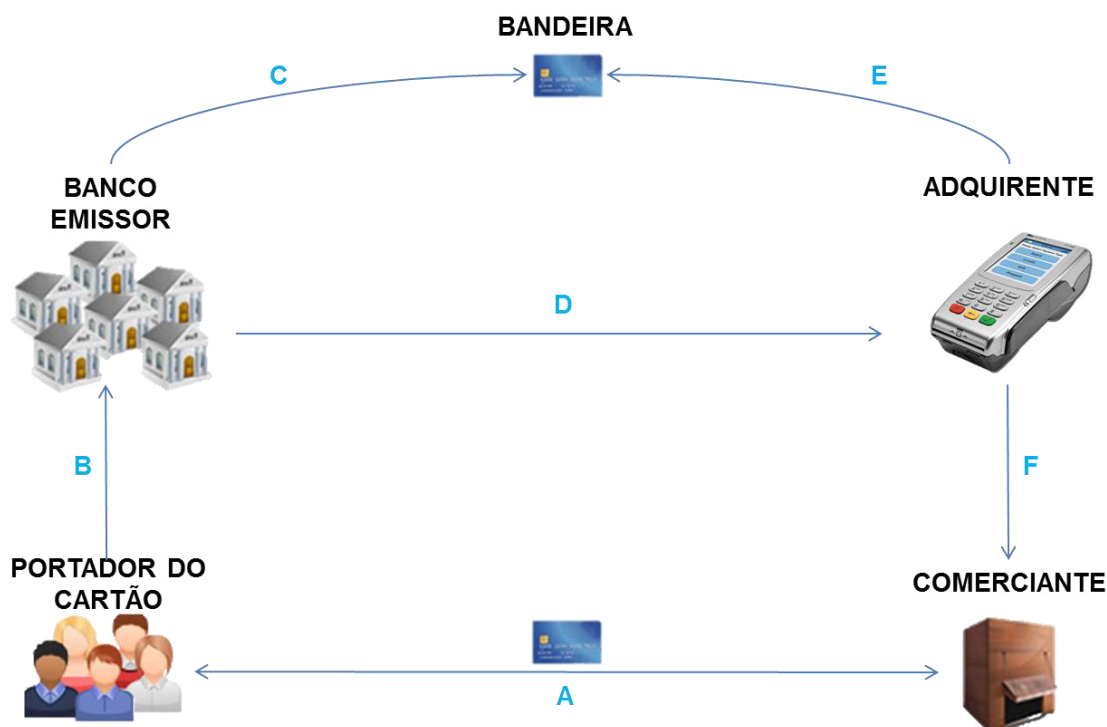


Figura 1– Ciclo de uma compra com cartão

Fonte: reprodução própria

O quadro 1 explica o que acontece em cada uma das ações, representadas pelas letras de A a F. Além disso, são citados os envolvidos em cada uma delas.

A diferença entre as três modalidades está nos prazos em que são movimentados os valores das transações realizadas. Essas diferenças são importantes para as empresas porque o MDR varia de acordo com a modalidade da compra e influencia diretamente nas receitas.

Em uma transação de débito, o portador do cartão faz a compra (A). O banco emissor recebe o dinheiro no mesmo dia (B). No dia seguinte, o banco emissor repassa o dinheiro para as adquirentes (D), que, também no dia seguinte, depositam o dinheiro para o comerciante (F).

Em uma transação de crédito, o portador do cartão faz a compra (A). O banco emissor recebe o dinheiro no mesmo dia (B). No dia seguinte, o banco emissor repassa o dinheiro para as adquirentes (D), que, também no dia seguinte, depositam o dinheiro para o comerciante (F).

Atores do ciclo de pagamento eletrônico	
A	O portador do cartão realiza a compra com o comerciante .
B	O portador do cartão repassa o valor da compra ao banco emissor . Além deste valor, ainda inferem sobre o portador as tarifas sobre o cartão. Exemplos de bancos emissores são: Bradesco, Banco do Brasil, Itaú, Caixa, etc.
C	O banco emissor repassa uma porcentagem da transação para a bandeira , conhecida como fee de bandeira. As bandeiras mais conhecidas são: VISA, MasterCard, Elo, AmericanExpress, etc.
D	O banco emissor recolhe uma outra porcentagem do valor recebido do portador , chamada de taxa de intercâmbio, e repassa o restante para a adquirente (também conhecida por credenciadora). As credenciadoras são também as donas das máquinas, responsáveis por afiliar e manter os comerciantes. Exemplos de credenciadoras são: Cielo, Rede, GetNet, etc.
E	As adquirentes repassam uma outra taxa à bandeira .
F	As adquirentes recolhem uma porcentagem, que somada às outras taxas compõe o MDR, e repassa o restante ao comerciante .

Quadro 1– Atores do ciclo de pagamento eletrônico

Fonte: reprodução própria

2. Revisão de Literatura

Quando a venda é realizada por crédito à vista, o banco emissor recebe o dinheiro 26 dias após a compra. 28 dias após a compra, o banco emissor repassa o dinheiro para as adquirentes, que, dois dias depois, depositam o dinheiro para o vendedor.

No crédito parcelado, o banco emissor recebe a primeira parcela em 26 dias e as próximas a cada 30 dias. O banco repassa o dinheiro para as adquirentes sempre dois dias depois (28, 58, 88, ...), que repassam para os comerciantes dois dias depois de receber (30, 60, 90, ...). Essa é, geralmente, a modalidade mais rentável para as empresas do mercado. O quadro 2 sintetiza as informações referentes aos prazos de pagamento.

Data em que ocorre a remuneração dos envolvidos			
	Débito	Crédito à vista	Crédito parcelado
A	D	D	D
B	D	D+26	D+26; D+56; ...
C	D+1	D+28	D+28; D+58; ...
D	D+1	D+28	D+28; D+58; ...
E	D+1	D+30	D+30; D+60; ...
F	D+1	D+30	D+30; D+60; ...

Quadro 2– Data em que ocorre a remuneração dos envolvidos

Fonte: reprodução própria

A taxa sobre a transação que é repassada pelo comerciante (MDR) é dividida entre as três responsáveis pela realização da transação: bancos emissores, aquirentes e bandeiras. Por ser uma taxa sobre o valor da transação, a receita das empresas neste mercado depende do valor total transacionado por cartões de crédito e débito.

2.2. Análise Multivariada

Na vida, sempre que se faz necessário tomar uma decisão, são levados em conta um grande número de fatores, ainda que imperceptivelmente. Estes fatores nem sempre pesam da mesma forma

no momento da tomada de decisão. Simples atividades cotidianas demonstram isto. Atravessar a rua, por exemplo, pode não parecer tarefa tão complexa, mas, para tal, nosso cérebro, naturalmente, considera dezenas de variáveis: quanto tempo tenho até o próximo carro?, Qual a velocidade que eu consigo alcançar?, A luz do semáforo está verde?, e muitas outras.

Estabelecer relações, encontrar ou propor leis explicativas é papel da ciência. Para isso, é necessário controlar, manipular e medir as variáveis que são consideradas relevantes para o entendimento do fenômeno analisado (VICINI, 2005).

No mundo corporativo, assim como quando atravessamos a rua, muitas são as informações disponíveis na tomada de decisão. Até recentemente, estas informações, no entanto, simplesmente desapareciam. Hoje, as empresas coletam e armazenam em bancos de dados, de forma a permitir a mineração para fins de melhoria na tomada de decisão (HAIR, 2009).

Segundo Hair et al. (2009), parte dessas informações pode ser analisada com estatística básica. No entanto, outra parte delas necessita técnicas de análise multivariada complexas, de modo a converter esta grande quantidade de informações em conclusões que possam ser úteis no momento de tomada de decisão.

Neto (2004) explica que a denominação “análise multivariada” corresponde a um grande número de métodos e técnicas que utilizam, simultaneamente, todas as variáveis na interpretação teórica do conjunto de dados obtido. Muitas destas análises são extensões da análise univariada e da análise bivariada. Um exemplo é a regressão simples, com uma variável preditora, que é estendida para incluir diversas variáveis preditoras (HAIR, 2009).

Frente à grande quantidade de métodos de análises multivariadas com finalidades bem distintas entre si, Neto (2009) defende que é essencial saber o que pretende se afirmar a partir dos dados que serão utilizados na análise. Ele cita que, quando o objetivo é, por exemplo, verificar o quanto as amostras se relacionam, métodos como análise por agrupamento hierárquico e análise por componentes principais podem ser utilizados. Por outro lado, quando a finalidade é fazer previsão, ou seja, quando existem muitas variáveis independentes e se quer encontrar uma dependente, a regressão linear múltipla e as redes neurais são algumas das técnicas indicadas.

Critérios para estimadores

A existência de diversas técnicas para se obter a equação desejada gera uma dúvida bastante comum para o pesquisador: Como medir qual é a melhor técnica?

Kennedy (2009) afirma que a resposta a esta pergunta varia de acordo com o econometrista responsável, mas que pode ser desde custo computacional e eficiência a minimização dos erros do modelo.

Tratando-se de estimadores que requerem cálculo sobre o resultado do modelo, os mais comuns são coeficiente de determinação, o método dos mínimos quadrados ordinários e a estimativa de máxima verossimilhança.

O método dos mínimos quadrados ordinários (MQO) mede a qualidade do modelo de acordo com o valor estimado. O método calcula a distância absoluta dos pontos estimados para o valor real do ponto. Depois, soma-se os quadrados de cada um dos valores. A aplicação do MQO tem por objetivo ajustar ao máximo a curva do modelo à curva real dos pontos, ou seja, a nova equação do modelo adere os erros residuais.

O MQO não deve ser confundido, entretanto, com o coeficiente de determinação (R^2). O R^2 é uma estatística bastante frequente e seu valor indica porcentagem da variação da variável dependente que pode ser explicada devido a variações nas variáveis independentes.

Hill et al. (2008) chega ao valor de R^2 por meio da seguinte proporção:

$$R^2 = \frac{SQE}{SQT} = 1 - \frac{SQR}{SQT}$$

Onde SQT é a soma dos quadrados total, SQE é a soma dos quadrados que pode ser explicada pela variação nas variáveis independentes e SQR é a soma dos quadrados dos erros residuais, que é igual ao calculado no MQO.

Portanto, o R^2 varia de 0 a 1 e, quanto mais próximo de 1, mais o modelo explica a variação na variável dependente por meio das variáveis independentes e, portanto, melhor a predição do modelo.

Uma relação entre o MQO e o R^2 é que, como o MQO tem por objetivo minimizar a soma dos quadrados dos erros, ao aplicá-lo, conseqüentemente se maximiza o R^2 .

O método dos mínimos quadrados ordinários é o mais simples a ser aplicado. Porém, sua aplicação está condicionada a hipóteses que serão abordadas posteriormente neste trabalho.

No entanto, o valores dos estimadores não devem ser os únicos a serem levados em consideração no estudo. Analisar estritamente estes números pode esconder outros problemas aos quais as técnicas de análise estão sujeitas, como será abordado mais adiante.

Regressão Simples

Quando o problema possui apenas uma variável independente, então a técnica é denominada como regressão simples, enquanto o uso de duas ou mais variáveis independentes a transforma em regressão múltipla.

Segundo Hill et al. (2008), o modelo de regressão linear clássica possuem quatro hipóteses que asseguram que o método dos mínimos quadrados é o melhor estimador linear não-viesado (ou Best Linear Unbiased Estimator – BLUE).

A primeira hipótese é que a média de Y , para cada valor de V_1 é dado pela equação de regressão linear a seguir:

$$\hat{Y} = b_0 + b_1 V_1 \quad (1)$$

Onde, b_0 é a constante de interceptação e b_1 é o coeficiente de regressão, que atua como peso para a variável V_1 . No caso deste trabalho de conclusão de curso, a variável dependente (mercado de pagamentos eletrônicos) poderia ser uma constante somada ao valor do PIB multiplicado pelo seu peso (também um valor constante).

A segunda hipótese é de que, para cada valor de V_1 , os valores de Y são distribuídos sobre o seu valor médio seguindo distribuições probabilísticas que possuem a mesma variância σ^2 .

A terceira hipótese das regressões lineares simples afirma que os valores amostrais de Y não são correlacionados, ou seja, não possuem associação linear entre eles.

Na quarta hipótese, Hill et al. (2008), dizem que a variável V_1 não deve ser aleatória e pode assumir, pelo menos, dois valores diferentes.

Os valores de b_0 e b_1 são encontrados pelo método dos mínimos quadrados e merecem interpretação aprofundada. b_0 , que é a constante de interceptação, tem sua explicação baseada nas características da variável independente. Se for factível a variável V_1 ter valor zero e a variável dependente ter valor diferente de zero, b_0 é considerado como um valor inicial para a variável dependente. No entanto, em situações onde a variável independente não pode ter valor igual à zero, a constante b_0 ajuda na melhoria da predição, mas não tem um valor explicativo.

Já o coeficiente de regressão b_1 é a variação estimada da variável dependente para cada unidade de variação da variável dependente.

Regressão Múltipla

No entanto, em muitos casos em que ocorrem o estudo de regressão, a variável dependente pode ser influenciada por diversas outras variáveis. Com isso, surge a necessidade de expandir a regressão simples adicionando novas variáveis independentes na análise.

Hair et al. (2009) definem regressão múltipla como uma técnica estatística que pode ser usada para analisar a relação entre uma única variável dependente e algumas variáveis independentes (ou preditoras). O objetivo desta técnica é usar o valor das variáveis independentes já conhecidos para prever o resultado da única variável dependente definida pelo responsável pela análise.

No resultado final de uma regressão múltipla, a cada variável independente é atribuído um peso, que representa a contribuição relativa daquela variável para o resultado da variável dependente.

Uma característica da regressão múltipla é que ela só deve ser usada quando todas as variáveis, dependentes ou independentes, são mensuráveis. A outra condição para que seja aplicada a análise é a definição de qual das variáveis será a variável dependente e quais das variáveis restantes serão consideradas independentes.

Além destas condições, a regressão múltipla, assim como a regressão linear simples, possui algumas hipóteses que, caso violadas, podem levar o pesquisador a conclusões incorretas.

A primeira hipótese é de que a média dos valores de Y segue a equação a seguir:

$$\hat{Y} = b_0 + b_1V_1 + b_2V_2 + b_3V_3 + \dots + e \quad (2)$$

Onde e é o erro residual, que nada mais é do que a diferença entre o resultado da equação e o valor real da variável dependente.

A segunda hipótese diz que o valor esperado de Y depende apenas das variáveis independentes e de seus parâmetros b_i , o que ocorre se, e somente se, o valor esperado para e seja zero.

Na terceira hipótese, Hill et al. (2008), afirmam que a variância dos valores de Y deve ser igual à variância dos valores de e , ou seja, $\text{var}(Y_i) = \text{var}(e_i) = \sigma^2$, para qualquer valor de i .

A quarta hipótese é de que, tanto para os valores de Y , quanto para os valores de e , duas observações da variável dependente não são correlacionadas, isto é, se uma observação de Y foi acima do valor esperado para Y , então uma observação subsequente não é mais ou menos provável de ser maior do que o valor esperado para Y .

Na quinta hipótese, diz-se que os valores para cada uma das variáveis V_i não devem ser aleatórios e também não podem ser função linear de outra variável independente.

Por fim, a sexta hipótese da regressão múltipla afirma que os valores de Y são distribuídos de acordo com uma distribuição normal tal que $Y_i \sim N[(b_0 + b_1V_1 + b_2V_2 + b_3V_3), \sigma^2]$, o que é o mesmo que dizer que o erro e também é distribuído normalmente tal que $e \sim N(0, \sigma^2)$.

Como dito anteriormente, a regressão múltipla estuda a relação de uma variável dependente para com outras variáveis independentes. No entanto, é necessária a devida atenção à escolha de variáveis para um modelo. O uso incorreto do conjunto de variáveis independentes pode causar a omissão de variáveis independentes relevantes para o modelo. Omissão de uma variável, de acordo com a equação, seria impor que seu estimador (b_i) seja igual a zero. Esta restrição, no entanto, pode enviesar o estimador das variáveis incluídas na equação.

As consequências de se omitir uma variável relevante pode gerar a estratégia de incluir todas as variáveis possíveis no modelo. Porém, isso somente causará uma complicação desnecessária para o modelo e inflação das variâncias devido às variáveis irrelevantes que serão adicionadas.

A existência de viés por causa das variáveis omitidas ou a inflação das variâncias por causa das variáveis irrelevantes tornam a especificação do conjunto de variáveis uma tarefa muito importante, assim como muito complicada.

Encontrada a equação de regressão, deve-se avaliar a sua qualidade. Para isso, mede-se a aderência do modelo à curva. A forma mais comum de medição é a partir dos coeficientes de determinação (R^2).

Multicolinearidade, heteroscedasticidade e autocorrelação

Alguns fenômenos são capazes de impactar negativamente o cálculo de uma equação de regressão múltipla. Estes fenômenos violam as hipóteses do modelo clássico de regressão múltipla e, portanto, os mínimos quadrados ordinários deixam de ser o melhor estimador linear não-viesado, o que pode levar o pesquisador a tirar conclusões incorretas da sua análise. Os três principais causadores

deste problema são multicolinearidade, heteroscedasticidade e autocorrelação, e serão abordados com mais detalhes a seguir.

A multicolinearidade é um fenômeno que viola a quinta hipótese da regressão múltipla, que diz que nenhuma das variáveis independentes é uma combinação linear das outras variáveis independentes. Deve-se frisar que o procedimento de estimação só é rompido em caso de multicolinearidade exata, que é o relacionamento linear exato entre alguns regressores, mas a existência de um relacionamento linear aproximado, chamado apenas de multicolinearidade, já é suficiente para ser considerado uma violação da hipótese do modelo de regressão múltipla. (GUJARATI, 2006)

No caso da multicolinearidade, a consequência para o modelo é uma inflação da variância dos estimadores. Isso ocorre porque não é possível calcular com confiança o impacto que cada um dos parâmetros têm sobre a variável dependente, uma vez que os impactos se confundem. Gujarati (2006) explica, na figura 2, o que é este efeito segundo o diagrama de Venn.

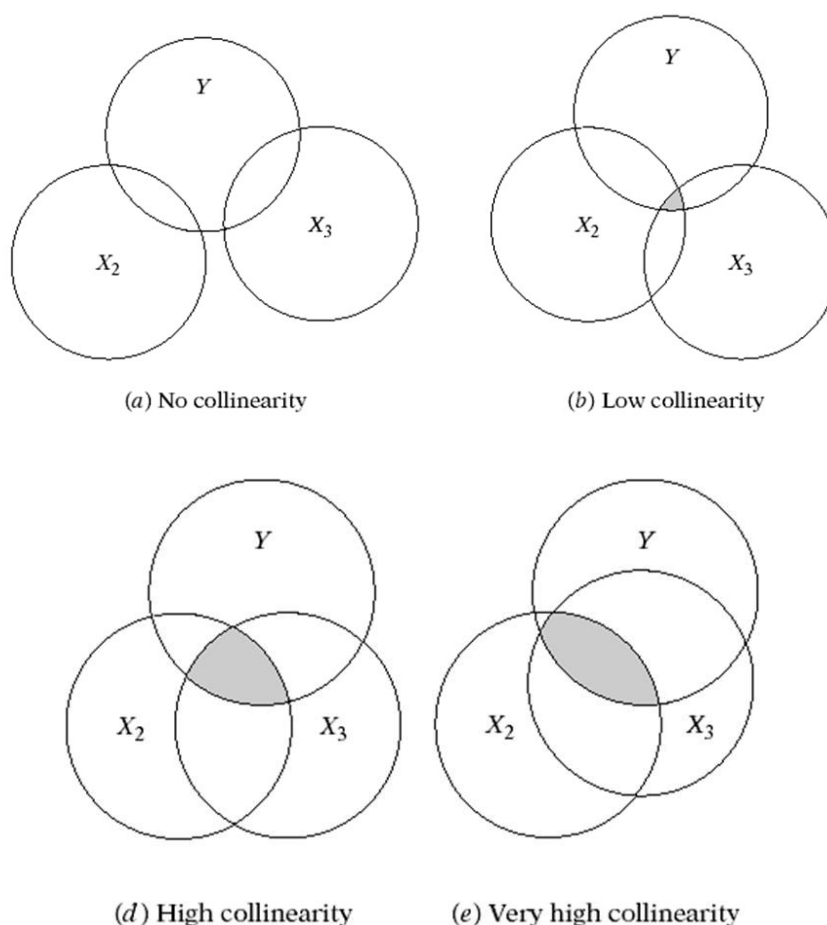


Figura 2– A colinearidade vista segundo o gráfico de Venn.

Fonte: GUJARATI, 2006, p. 277.

2. Revisão de Literatura

As principais causas da multicolinearidade são: amostra de tamanho limitado, adição de termos polinomiais quando o intervalo a que pertence a variável dependente é pequeno e também quando o modelo contém mais variáveis independentes do que observações na amostra.

Uma forma de identificar multicolinearidade é realizando uma regressão auxiliar para a variável independente a partir das outras variáveis independentes. Uma equação de regressão auxiliar geral para V_2 seria:

$$V_{i2} = b_1V_{i1} + b_3V_{i3} + \dots + b_KV_{iK} + erro \quad (3)$$

Caso o resultado de R^2 seja alto (acima de 0,8, por exemplo), é um indício de que uma variação na variável independente pode, em grande parte, ser explicada pela variação de outras variáveis.

Um outro fenômeno que prejudica o estudo de regressão é a heteroscedasticidade. Segundo Hill et al. (2008), heteroscedasticidade é uma violação da terceira hipótese da regressão múltipla, que diz que a variância dos valores da variável dependente Y é igual à variância do erro e , que, para qualquer observação deve ser igual ao quadrado do desvio padrão.

A heteroscedasticidade implica que a variância dos valores de Y seja uma função das variáveis independentes (V_1, V_2, \dots). Para exemplificar, Hill et al. (2008) consideram o caso de gastos com alimentação de acordo com a renda de uma família.

Os autores afirmam que, em uma família de alta renda pode ter a opção de comidas mais extravagantes (como caviar) assim como pode comer comidas mais simples (macarrão), enquanto famílias de baixa renda têm de optar por comidas mais simples. Desta forma, a variância dos valores de gasto com alimentação aumenta conforme aumenta a renda familiar, o que caracteriza uma violação de uma das hipóteses de regressão múltipla.

Existem alguns métodos utilizados para a detecção do problema de heteroscedasticidade. Entre os mais comuns, estão a plotagem dos erros residuais, o teste de Goldfeld-Quandt e o teste de Breusch-Pagan.

No caso da plotagem de erros residuais, a observação é muito complicada na existência de mais de uma variável independente, fazendo-se necessário o uso das técnicas mais avançadas.

O teste de Goldfeld-Quandt requer que a amostra seja dividida em duas subamostras, de forma a tentar identificar variâncias significativamente diferentes entre dois grupos. Este teste é mais facilmente aplicado em modelos com mais de um variável independente, mas é necessário identificar anteriormente a variável que pode ser responsável pela heteroscedasticidade, o que é uma limitação.

Além disso, o teste requer uma amostra suficientemente grande para ser dividida em outras duas subamostras.

Por fim, o teste de Breusch-Pagan assume que a variância depende de um conjunto de variáveis independentes. Esta dependência pode ser, por exemplo, linear ou exponencial. A técnica testa a hipótese de que todos os termos que multiplicam as variáveis explanatórias são zero, o que significa que a variância tomaria um valor constante e caracterizaria a não existência de heteroscedasticidade.

O próximo passo do teste de Breusch-Pagan é, então, identificar se as variáveis independentes ajudam a explicar as variações do erro residual, confirmando ou rejeitando a hipótese de existência de heteroscedasticidade.

Um vez identificada a ocorrência de heteroscedasticidade, surge a necessidade de correção. Para isso, é necessária a transformação do modelo, com o auxílio dos mínimos quadrados generalizados, de forma a atingir a homoscedasticidade.

A transformação que deve ser aplicada depende da relação entre a variância e as variáveis independentes e pode ser linear, como divisão pela raiz quadrada da variável independente, ou exponencial, como a aplicação do logaritmo neperiano.

O terceiro problema que pode impactar as análises de regressão múltipla é a autocorrelação. Este problema viola a quarta hipótese da regressão múltipla, que diz que um valor de Y não é correlacionado com outros valores de Y ou, em outras palavras, se uma observação de Y for acima do valor esperado de Y, então a observação subsequente não será mais ou menos provavelmente maior do que o valor esperado de Y.

Para exemplificar este problema, Hill et al. (2008) usam séries temporais, como taxas de inflação ou taxas de desemprego. Segundo os autores, um período de altas taxas de desemprego provavelmente será sucedido por outro período com altas taxas de desemprego. Isto ocorre porque uma ação que acontece em determinado período, ainda vai ter efeito nos períodos seguintes. Desta forma, o erro residual de uma observação é função do erro residual da observação anterior, o que vai contra a hipótese da regressão múltipla.

Um modelo de regressão que captura os efeitos de autocorrelação seria descrito por:

$$Y_t = \hat{Y} + \rho Y_{t-1} + e_t \quad (4)$$

Onde ρ é o parâmetro que descreve a dependência do erro em relação à observação anterior e e_t é um termo de erro aleatório, independente, com média zero e variância constante σ_e^2 . A equação

(4) define um modelo conhecido por modelo regressivo de ordem 1, ou, de forma mais simples, modelo AR(1).

É importante frisar que, no caso da heteroscedasticidade e da autocorrelação, os mínimos quadrados continua sendo um estimador linear e não-viesado, mas não pode mais ser considerado o melhor, ou seja, é possível encontrar um outro estimador com menor variância.

Como forma de identificar a existência de autocorrelação, um dos testes mais comuns é o teste de Durbin-Watson. Ele testa a hipótese nula de que não existe autocorrelação entre os erros de uma regressão. A estatística de Durbin-Watson varia entre 0 e 4 e valores próximos a 2 apontam para a não existência de autocorrelação, enquanto valores próximos a 0 indicam autocorrelação positiva e valores próximos a 4 indicam autocorrelação negativa.

Como estas definições de autocorrelação podem ser arbitrárias, Durbin-Watson disponibilizaram duas tabelas – com diferentes significâncias estatísticas – com os valores críticos, que devem ser consultados de acordo com a quantidade de variáveis e o tamanho da amostra. Estas tabelas se encontram nos anexos A e B.

As regras das tabelas são as seguintes:

- Se $d_{\text{calc}} < dL$ – há autocorrelação positiva
- Se $dL < d_{\text{calc}} < dU$ – zona de indecisão
- Se $dU < d_{\text{calc}} < 4 - dU$ – não há autocorrelação
- Se $4 - dU < d_{\text{calc}} < 4 - dL$ – zona de indecisão
- Se $4 - dL < d_{\text{calc}}$ – há autocorrelação negativa

A existência de zonas de indecisão é, no entanto, uma grande restrição inerente ao método de Durbin-Watson e, apesar da simplicidade, torna-o pouco defendido na literatura. Outros métodos surgem como opção caso seja obtido um resultado que indique indecisão, como é o caso do teste de Breusch-Godfrey, que é um teste baseado no multiplicador de Lagrange.

Séries temporais

Uma série temporal tipicamente observa e descreve o comportamento de uma determinada variável dependente ao longo do tempo. Quando este tipo de dados é analisado, algumas especificidades devem ser consideradas.

Greene (2008) afirma que as séries temporais podem ser divididas em dois principais tipos: as séries temporais estacionárias e as não-estacionárias. A noção de estacionariedade está diretamente ligada ao comportamento da série temporal.

A principal característica de uma série temporal estacionária é que a média, a variância e a estrutura de autocorrelação independem do tempo. A figura 3 exemplifica a diferença entre os dois tipos de comportamento. Os gráficos de vendas mensais de casas e de oscilações diárias no índice Dow Jones são séries temporais estacionárias, enquanto os outros dois gráficos representam séries temporais não estacionárias, o de contratos do tesouro americano tem tendência de queda e o de produção mensal de eletricidade na Austrália tem tendência ascendente.

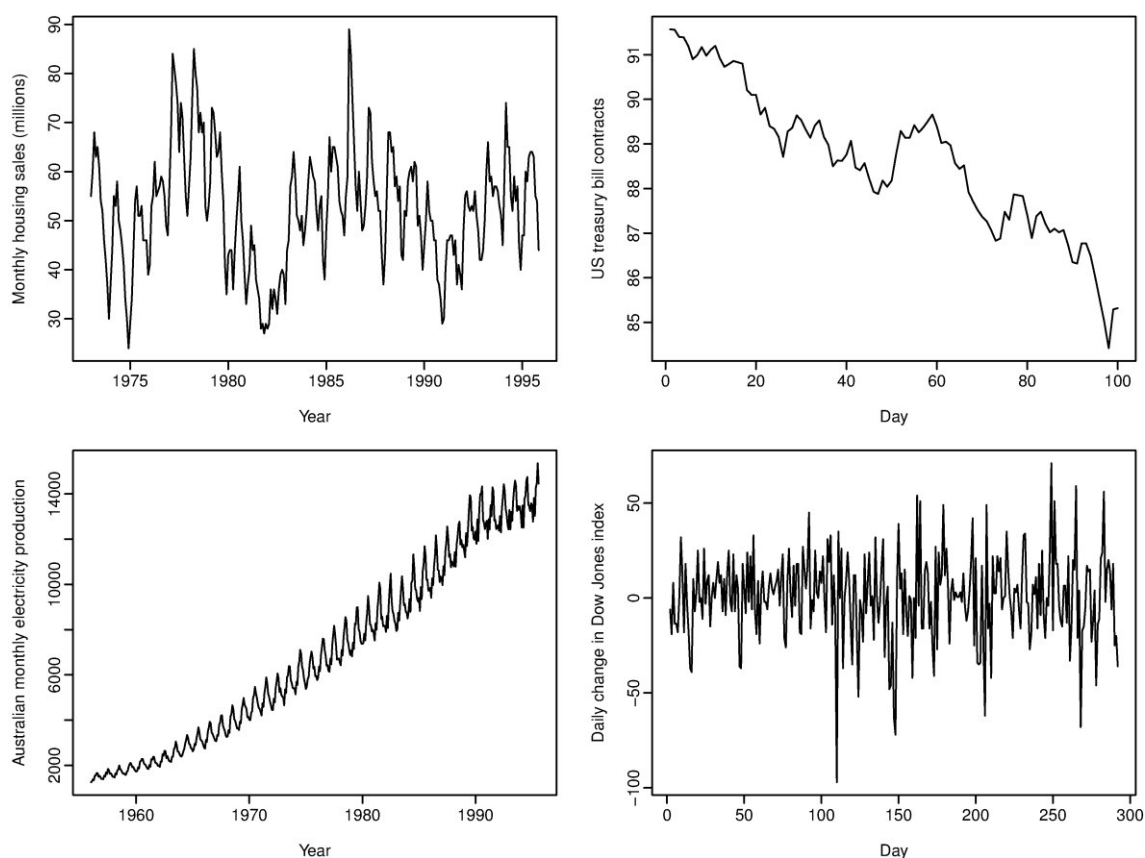


Figura 3– Modelos de séries temporais

Fonte: OTEXTS, 2014.

As séries temporais precisam de uma revisão nas hipóteses que foram descritas para os outros modelos de análise de dados abordados principalmente porque elas possuem uma particularidade: apresentam autocorrelação entre suas observações, isto é, a observação em determinado espaço de tempo t depende do valor da observação no espaço de tempo $t-1$ e, além disso, influenciará a observação no espaço de tempo $t+1$.

Os modelos autorregressivos são particularmente adequados para situações em que os erros residuais vão se diluindo ao longo do tempo, como nas séries temporais estacionárias (LOPES, 2001). Os modelos regressivos de primeira ordem, AR(1), foram ligeiramente abordados anteriormente. No entanto, é muito comum que efeitos sejam percebidos após mais de um período. Por isso, um modelo auto regressivo mais geral de p -ésima ordem, AR(p), poderia ser escrito por:

$$Y_t = \hat{Y} + \rho^1 e_{t-1} + \rho^2 e_{t-2} + \dots + \rho^p Y_{t-p} + e_t \quad (5)$$

Considerada a restrição $|\rho| < 1$, a equação acima se caracteriza como uma série temporal estacionária, pois o efeito de períodos antigos tendem a ser zero a cada novo período. Outra forma de ver o comportamento estacionário é observar a própria equação. Como o valor esperado para e_s , para qualquer que seja o valor de n , é sempre zero, os valor esperado de Y_t será sempre igual à média \hat{Y} , independente do período, o que caracteriza a estacionariedade.

Como visto anteriormente, o uso dos mínimos quadrados ordinários ainda é possível na existência de autocorrelação. No entanto, ele deixa de ser o melhor estimador. Por isso, é necessária a devida atenção durante a realização da análise, já que ela pode levar a conclusões incorretas.

Tratando-se de séries temporais não-estacionárias, a inclinação, seja ela ascendente ou descendente, pode ser obtida por meio da inserção de um termo que seja multiplicado por uma variável do período. É possível entender esse comportamento seguindo estes passos:

$$\begin{aligned} Y_1 &= \alpha + Y_0 + e_1 \\ Y_2 &= \alpha + Y_1 + e_2 = Y_1 + \alpha + e_2 = \alpha + (\alpha + Y_0 + e_1) + e_2 = 2\alpha + Y_0 + \sum_{s=1}^2 e_s \\ &\vdots \\ Y_t &= \alpha + Y_{t-1} + e_t = t\alpha + Y_0 + \sum_{s=1}^t e_s \end{aligned} \quad (6)$$

Assumindo que o valor esperado de e_s , para qualquer que seja o valor de s , é zero, então $E(Y_t) = t\alpha + Y_0$, ou seja, a média é dependente do período t , o que viola a condição de estacionariedade.

Não necessariamente é possível observar o comportamento de estacionariedade ou não-estacionariedade de uma série temporal e, por isso, alguns testes são usados para determinar o comportamento de determinado conjunto de dados. O mais comum deles é o teste de Dickey-Fuller.

Este teste é baseado na restrição do modelo AR(p) onde foi assumido $|\rho| < 1$ para modelos estacionários. Nas equações acima, é possível notar a ausência da constante ρ , mas, na verdade, foi imposta a restrição $\rho = 1$, transformando o modelo da equação em não-estacionário.

Os modelos testam a estacionariedade da série temporal através do valor de ρ . Testes nesta natureza são conhecidos por testes de raiz unitária. O teste de Dickey-Fuller identifica o valor de ρ a partir da variação da variável dependente, ou seja, $\Delta Y_t = Y_t - Y_{t-1}$.

As equações que trataram de séries temporais até o momento tratavam o valor da variável dependente somente em relação à observação anterior por simplificação, isto é, é possível incluir na equação 6, por exemplo, variáveis independentes.

A adição de variáveis independentes em séries temporais não-estacionárias, no entanto, pode ocasionar um novo problema, as chamadas regressões espúrias. A existência de regressões espúrias é perigosa pois pode levar o pesquisador a tomar conclusões incorretas.

Um exemplo das consequências de uma regressão pode ser explicado se forem analisados por exemplo, a população brasileira ao longo do tempo e o PIB da China, que são duas séries temporais com tendência ascendente. É sabido que estas duas variáveis não possuem a menor influência uma na outra. No entanto, uma regressão com as duas séries retornará um R^2 bastante alto, simplesmente porque ambas apresentam comportamento não-estacionário ascendente. Esta conclusão, no entanto, não faz sentido e pode enviesar a análise.

A regra geral é não usar séries temporais não-estacionárias em modelos de regressão. Apesar disso, muitas variáveis importantes poderiam ser omitidas, o que também enviesaria a análise. Uma forma de testar a espuriedade de uma regressão é a partir dos testes de cointegração.

Um teste bastante comum de cointegração é o teste de Engle e Granger. Este teste aborda sempre duas séries temporais não-estacionárias, Y_t e X_t . É esperado que a diferença ou uma combinação linear entre elas, como $e_t = Y_t - \beta_1 - \beta_2 X_t$, seja também não-estacionária. Porém, existe um caso especial onde $e_t = Y_t - \beta_1 - \beta_2 X_t$ é estacionário e, neste caso, Y_t e X_t são ditas cointegradas. Mas, se os erros residuais forem não-estacionários, então elas não são cointegradas e qualquer aparente relação entre elas é espúria. Para isso, faz-se o uso do Z-score, com o objetivo de aceitar ou rejeitar a hipótese nula de que as curvas não são cointegradas.

De maneira mais objetiva, os testes de cointegração são responsáveis por comparar as curvas das variáveis independentes e da variável dependente a fim de detectar uma relação na segunda derivada, como por exemplo, em curvas sazonais.

Graus de liberdade

Além das variáveis dependentes e independentes, outro elemento deve ser considerado no momento de aplicar uma regressão múltipla. O tamanho da amostra pode ser mais influente do que as próprias variáveis. Os efeitos do tamanho da amostra são vistos diretamente no poder estatístico do modelo e na generalização do resultado.

A regra geral de generalização do resultado é que a razão entre a quantidade de observações e a quantidade de variáveis independentes não deve ser menor do que 5:1. Caso a proporção fique abaixo, corre-se o risco de superajustar a variável à amostra, tornando os resultados muito específicos. O conceito de superajuste está ligado à noção de graus de liberdade. (CORRAR et al., 2007)

Como cada observação representa uma unidade separada e independente de informação, se cada uma das variáveis fosse dedicada a prever uma observação diferente, então a previsão perfeita ocorrerá mesmo que todos os valores de variáveis sejam aleatórios. Em tal cenário, os parâmetros estimados não teriam generalização, uma vez que se relacionam apenas com a amostra.

Quando, por outro lado, se aumenta o tamanho da amostra, seria possível prever uma observação com sua respectiva variável, mas não as outras observações. Como o objetivo é conseguir o melhor modelo de regressão, é de interesse do responsável pelo modelo que este consiga a melhor precisão preditiva para a maior (ou mais generalizável) amostra. O grau de generalização é representado pelo grau de liberdade, que, como definem Corrar et al. (2007), pode ser calculado da seguinte forma:

$$\text{Graus de liberdade}(df) = \text{Tamanho da amostra} - \text{Número de parâmetros}$$

Quanto maior o grau de liberdade, mais generalizados são os resultados. Portanto, a meta é conseguir a melhor precisão preditiva com o máximo de graus de liberdade.

Critérios de informação e seleção do modelo

A enorme quantidade de variáveis possíveis aliada às restrições no quantidade máxima de variáveis dentro de um modelo geram um conflito e grande discussão. A seleção das melhores variáveis e, conseqüentemente, o melhor modelo são obtidos pelos métodos baseados na teoria da informação.

Dal Bello (2010) diz que a seleção de componentes é um processo complexo e que o processo pode ser fortemente impactado pelo efeito de colinearidade entre variáveis, que pode enviesar o modelo, tornando variáveis significativas perante alguns termos, mas insignificantes quando na presença de outros termos.

Alguns dos modelos de seleção podem, portanto, se tornar arbitrários. É o caso da seleção *stepwise* e da seleção *forward*. O primeiro modelo consiste em remover as variáveis com menor índice de significância para o modelo, até que um modelo com resultado suficientemente bom seja alcançado. Já a seleção *forward* inclui, uma a uma, novas variáveis, aceitando-as caso haja melhora no índice de significância do modelo. Como esses métodos ignoram a possível existência de colinearidade, eles obtêm a otimização heurística do modelo, e não o melhor modelo possível.

Na verdade, obter o melhor modelo é uma tarefa bastante complexa. Uma alternativa, segundo Dal Bello (2010), é considerar todas as combinações possíveis de termos e número de parâmetros e utilizar os critérios de informação para selecionar o melhor modelo. Para isso, destaca-se, como mais comum, o critério de informação de Akaike, AIC, que se origina da minimização da distância de Kullback-Leibler, que é uma medida de distância entre duas distribuições de probabilidade, no caso, o modelo verdadeiro e um modelo candidato.

A equação do AIC pode ser expressa por

$$AIC = \frac{(-2l)}{n} + \frac{(-2k)}{n}$$

onde l é a função de verossimilhança, n é o tamanho da amostra e k é o número de restrições.

Quando calculado o AIC, o valor, de fato, não possui nenhum significado. Há a necessidade de comparar o valor de AIC de algum modelo com outros modelos diferentes e, então, o melhor modelo será aquele com o valor de AIC menor.

Uma outra opção em relação ao AIC é o BIC, critério de informação bayesiano. Em relação ao AIC, o BIC tem por característica punir com maior intensidade os modelos com maiores restrições,

ou seja, o BIC pune os graus de liberdade. No entanto, a diferença existente entre os dois modelos não é tão grande.

2.3. Consolidação dos Conceitos

Os conceitos abordados acima dão uma dimensão da complexidade envolvida no processo de modelagem de uma análise multivariada. É muitas vezes necessário fazer escolhas quanto às variáveis, tomar cuidado com os problemas que podem induzir a resultados ilusórios, observar o comportamento da sua variável dependente. Portanto, uma consolidação do que foi discutido por todo este capítulo e posterior abordagem do problema será descrita a seguir.

A variável dependente a ser estudada é o volume financeiro trimestral transacionado pelas máquinas de cartão de crédito. Os dados serão observados a partir do primeiro trimestre de 2006, o que corresponde a uma amostra de 34 observações, até o segundo trimestre de 2014. Se observado o comportamento da variável no gráfico abaixo, é possível tirar algumas informações.

O gráfico 2 plota o volume transacionado pelas máquinas de cartão de crédito no Brasil trimestralmente.

Primeiramente, observa-se que se trata de uma série temporal não-estacionária, o que requer alguma atenção especial. Além disso, existe sazonalidade da curva, o que torna necessário o teste de cointegração para eliminação de regressões espúrias.

A grande quantidade de variáveis que podem impactar o resultado final do mercado de pagamentos eletrônicos induz a utilização da regressão múltipla como técnica de modelagem. Já como estimador, será usado o método dos mínimos quadrados ordinários. Vale ressaltar que, como as observações da variável tem comportamento de série temporal não-estacionária, o método dos mínimos quadrados não é mais o melhor estimador linear não-viesado. Outros estimadores como método da máxima verossimilhança, poderiam ser usados, mas a complexidade das aplicações destes foge ao escopo deste trabalho de conclusão de curso.

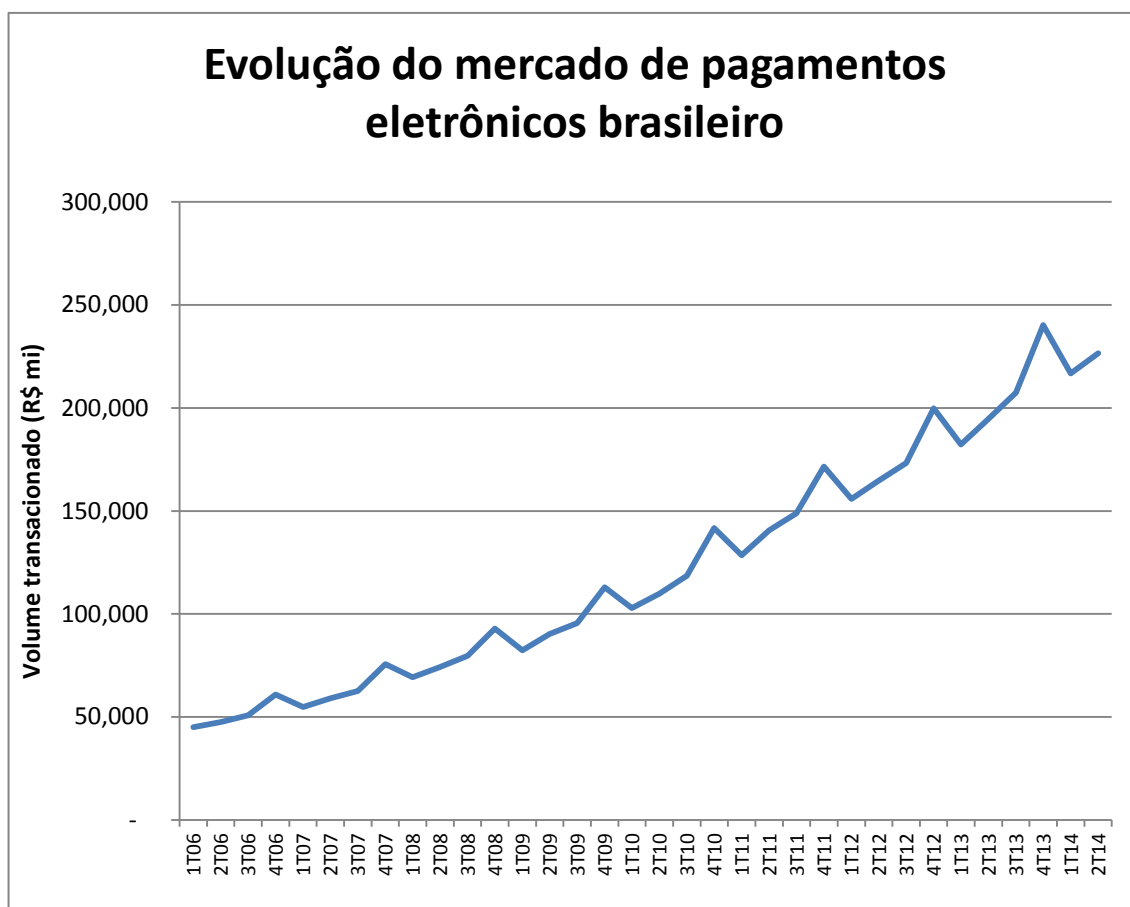


Gráfico 2– Evolução do mercado de pagamentos eletrônicos brasileiro

Fonte: ABECS, 2014.

Uma vez definido o método, é necessária a escolha das variáveis. Os métodos de seleção *stepwise* e *forward* podem ser bastante aleatórios, escondendo multicolinearidade ou outros problemas, o que acarreta em grandes chances de enviesamento da análise. Por outro lado, considerar todas as combinações possíveis de variáveis e identificar a melhor por meio de critérios de informação é muito mais complexo do que o desejado. Portanto, o método de seleção tem um perfil entre estes dois extremos.

Primeiro, será calculada uma regressão simples para cada uma das variáveis independentes em relação à variável dependente. Depois, a seleção das melhores variáveis explanatórias se dará por meio do critério de informação. As variáveis selecionadas no segundo passo comporão a equação de regressão final.

O passo seguinte será o teste de cointegração de Engle-Granger com a regressão e todas as suas variáveis. O *software* que será usado não possui esta função. No entanto, Schaffer (2010) desenvolveu um código que realiza a tarefa, e será utilizado também neste trabalho.

2. Revisão de Literatura

Deve-se, por fim, mitigar os efeitos de homoscedasticidade e multicolinearidade. No caso do primeiro, a todos os termos da regressão será incluído o logaritmo neperiano. Esta ação deve ser suficiente para eliminar os efeitos de homoscedasticidade. Já a existência de multicolinearidade poderá ser estudada quando a equação estiver pronta. Neste caso, o *software* a ser utilizado, que será comentado adiante, apontará a existência deste efeito.

Como variáveis explanatórias, existem muitas opções. As principais variáveis independentes são as variáveis macroeconômicas e, a partir de dados na literatura, é possível citar algumas. É importante que existam dados passados dessas variáveis nos mesmos moldes da variável dependente. A seguir, uma lista com as possíveis variáveis que podem ser usadas:

- | | | |
|------------------------------|--------------------------------|--------------------------|
| - PIB Nominal | - Massa salarial | - Quantidade de cartões |
| - Consumo das Famílias | - Carteira de crédito | - Quantidade de máquinas |
| - Volume de vendas do varejo | - Carteira de consumo | - Taxa Selic |
| - PEA | - Carteira de consumo imediato | - Inflação |
| - População ocupada | - Endividamento das famílias | |

A maioria das variáveis citadas são dados públicos do governo brasileiro e estão disponíveis no site do Instituto Brasileiro de Geografia e Estatística, IBGE. Portanto, não seria problema a utilização destas. Outras variáveis, no entanto, possuem restrição de dados, como é o caso da quantidade de cartões e quantidade de máquinas. Estas variáveis, portanto, não serão utilizadas durante a análise de regressão múltipla deste trabalho.

Alguns dos passos que serão seguidos pelo método, como discutido durante este capítulo, exigem muitos cálculos ou cálculos complexos. Por isso, um *software* auxiliará esta tarefa. O *software* em questão é o STATA 12.0, que é um programa de estatística utilizado para análises econométricas. As funções do STATA variam desde as mais simples, como calcular média e desvio-padrão, até as mais complexas, como regressão linear múltipla e outros métodos de regressão.

3. Método de Pesquisa

Os primeiro passo para obtenção da equação de regressão múltipla é a aplicação do logaritmo neperiano em todas as variáveis¹, dependentes ou independentes. Como visto, esta transformação tem como objetivo mitigar os efeitos de heteroscedasticidade no modelo. Uma outra vantagem da utilização da função logarítmica no modelo é a obtenção do efeito de elasticidade, isto é, pode-se, assim, calcular o efeito da variação de 1% de alguma variável independente na variável dependente. Em outras palavras, se o PIB aumentar em 1%, em qual proporção aumentaria o mercado de pagamentos eletrônicos.

Na Figura 4, por exemplo, faz-se a transformação da variável dependente 'mercadormi', que é a variável alvo do modelo, com o respectivo resultado no banco de dados.

	mercadormi	lnmercado
1	44949.9	10.7133
2	47497.3	10.76843
3	50756.1	10.83479
4	60939.5	11.01764
5	54833.6	10.91206
6	58974.6	10.98486
7	62621.3	11.04486
8	75619.5	11.23347
9	69292.1	11.14609
10	74305.7	11.21594
11	79735.7	11.28647
12	92938.9	11.4397
13	82366.3	11.31893
14	90183.9	11.40961
15	95573	11.46765
16	112921	11.63444
17	102905	11.54156
18	109863	11.60699
19	118559	11.68316
20	141627	11.86095
21	128520	11.76384
22	140534	11.85321
23	148833	11.91058
24	171456	12.05208
25	155818	11.95644
26	164631	12.01146
27	173267	12.06259
28	199854	12.20534

Figura 4– Transformação da variável 'mercadormi'

Fonte: Geração própria através do *software* STATA.

¹ No STATA, a transformação é realizada com a função $\text{gen lnvar1} = \ln(\text{var1})$, onde *var1* é a variável a ser transformada

3. Método de Pesquisa

O passo de transformação é aplicado a todas as variáveis e integram o banco de dados.

O segundo passo é definir a amostra como série temporal.² Para isso, o *software* possui uma função especial, que pode incluir um termo temporal diário, mensal, trimestral ou anual.³

A seleção das variáveis para o modelo passa pela regressão linear de cada uma das possíveis variáveis independentes e posterior cálculo do critério de informação de Akaike. A figura 5 mostra como são realizados, no STATA 12, estes dois passos.

```
. reg lnmercado lnplibnominal
```

Source	SS	df	MS	Number of obs = 34		
Model	8.20682886	1	8.20682886	F(1, 32)	= 2999.16	
Residual	.087564058	32	.002736377	Prob > F	= 0.0000	
				R-squared	= 0.9894	
				Adj R-squared	= 0.9891	
Total	8.29439291	33	.25134524	Root MSE	= .05231	

lnmercado	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lnplibnominal	1.965518	.0358903	54.76	0.000	1.892412	2.038624
_cons	-1.728955	.2435933	-7.10	0.000	-2.225139	-1.232772


```
. estimate stats
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	34	-24.26064	53.10576	2	-102.2115	-99.1588

Figura 5 – Exemplo de cálculo do Critério de Informação

Fonte: Geração própria através do *software* STATA.

No primeiro comando, é possível analisar outras informações importantes sobre a variável. O R^2 e o R^2 ajustado já estão presentes para cada uma das variáveis explanatórias. Além disso, o comando resulta em uma equação de regressão linear simples para a variável dependente. A equação neste caso seria representada por:

$$\ln(\text{Mercado}) = -1,728955 + 1,965518 * \ln(\text{PIB})$$

² Os comandos do STATA que definem a amostra como série temporal são: *gen tempo = q(2006q1)+_n+1, tset tempo, quarterly*

³ Neste caso, a variável tempo é definida trimestralmente, com seu primeiro valor no primeiro trimestre de 2006.

3. Método de Pesquisa

O segundo comando resulta no critério de informação, que será utilizado para selecionar as variáveis para o modelo. O procedimento é repetido para todas as outras variáveis independentes e está disponível nos apêndices deste trabalho. O resumo de todos os resultados encontra-se na tabela 1.

Para definição da quantidade de variáveis independentes e quais serão estas, será utilizado o método *forward*, que consiste em utilizar, inicialmente, todas as variáveis possíveis. A partir daí, é excluída uma variável por vez, até que se obtenha o melhor resultado.

Resultados das regressões lineares		
Variável	R ² Ajustado	AIC
PIB Nominal	0,9891	-102,2
Consumo das famílias	0,9890	-101,8
Varejo	0,9556	-54,5
PEA	0,9598	-57,8
População ocupada	0,9764	-76,0
Taxa de desemprego	0,9604	-58,3
Massa salarial	0,9865	-95,0
Endividamento das famílias	0,9612	-59,0
Inflação (IPCA)	0,4599	30,5
Inflação (IGPM)	0,1538	45,2
Selic	0,4457	31,4

Tabela 1 – Resultado das regressões lineares

Fonte: Geração própria através do *software* STATA.

Neste caso, as variáveis serão excluídas de acordo com seus valores de critério de informação na regressão simples com a variável dependente. Quando o valor do critério de informação do modelo aumentar após a exclusão de uma variável, então tem-se o modelo final.

A tabela 2 ilustra a variação dos critérios de informação e Akaike e Bayesiano de acordo com a quantidade de variáveis.

Apesar de o método BIC punir com mais intensidade os graus de liberdade, o resultado é o mesmo para ambos os casos (AIC e BIC). O modelo que atinge o menor critério de informação entre os

3. Método de Pesquisa

testados é aquele que possui oito variáveis independentes. São elas: PIB nominal, consumo das famílias, massa salarial, população ocupada, endividamento das famílias, taxa de desemprego, população economicamente ativa e volume de vendas do varejo.

Aplicação da seleção <i>forward</i>		
Quantidade de variáveis	AIC	BIC
11	-133,46	-115,87
10	-137,11	-120,31
9	-139,09	-123,76
8	-141,00	-127,26
7	-120,38	-108,17
6	-122,24	-111,55
5	-117,80	-108,65
4	-117,50	-109,86
3	-117,32	-111,22

Tabela 2 – Aplicação da seleção *forward*

Fonte: Geração própria através do *software* STATA.

Com as variáveis escolhidas, o teste de cointegração é realizado a partir do comando abaixo. O resultado também está descrito na figura 6.

```
. egranger lnmercado lnpibnominal lnconsumo lnvarejoamp lnpea lnpopocup lntaxadesemp lnmassasalarial lnendividamento
Replacing variable _egresid...

Engle-Granger test for cointegration          N (1st step) =      34
                                                N (test)      =      33
```

	Test Statistic	1% Critical Value	5% Critical Value	10% Critical Value
Z(t)	-6.540	-7.212	-6.281	-5.833

Critical values from MacKinnon (1990, 2010)

Figura 6– Teste de cointegração Engle-Granger

Fonte: Geração própria no *software* STATA.

O teste de cointegração de Engle-Granger mostra que $Z(t) = -6,540$ é maior do que o valor crítico com confiança de 1% que é $V_{C1\%} = -7,212$. Portanto, com 1% de confiança, não se pode rejeitar a hipótese nula de que não há cointegração entre as curvas. No entanto, o valor de $Z(t)$ é menor do

3. Método de Pesquisa

que o valor crítico com confiança de 5%, ou seja, $Vc_{5\%} = -6,281$, o que significa que, com 5% de confiança, é possível rejeitar a hipótese nula e aceitar a hipótese alternativa de que há cointegração entre as curvas.

Após selecionar as variáveis com os melhores critérios de informação e testar a cointegração de Engle-Granger, que confirmou que, com 5% de confiança, a regressão não é espúria, isto é, existe cointegração entre as curvas, o próximo passo é rodar a regressão com as variáveis selecionadas. No *software* STATA, o comando usado e os resultados obtidos constam na figura 7.

```
. reg lnmercado lnpiynominal lnconsumo lnvarejoamp lnpea lnpopocup lntaxadesemp lnmassasalarial lnendividamento
```

Source	SS	df	MS	Number of obs = 34		
Model	8.27593234	8	1.03449154	F(8, 25) = 1395.09		
Residual	.018538115	25	.000741525	Prob > F = 0.0000		
				R-squared = 0.9978		
				Adj R-squared = 0.9970		
Total	8.29447045	33	.251347589	Root MSE = .02723		

lnmercado	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lnpiynominal	.420869	.2340321	1.80	0.084	-.061129	.9028671
lnconsumo	-.1975839	.4534477	-0.44	0.667	-1.131477	.7363091
lnvarejoamp	.8905307	.183235	4.86	0.000	.5131512	1.26791
lnpea	-4.796759	4.393867	-1.09	0.285	-13.8461	4.25258
lnpopocup	2.230006	4.271476	0.52	0.606	-6.567262	11.02728
lntaxadesemp	.1444761	.3315571	0.44	0.667	-.5383785	.8273306
lnmassasalarial	1.507612	.4010913	3.76	0.001	.6815488	2.333675
lnendividamento	-.2528309	.2318355	-1.09	0.286	-.7303049	.2246432
_cons	24.30503	10.10591	2.41	0.024	3.49152	45.11853


```
. estimate stats
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	34	-24.2608	79.49897	9	-140.9979	-127.2607

Figura 7– Obtenção da regressão múltipla

Fonte: Geração própria através do *software* STATA.

O resultado obtido é a equação de regressão

$$\begin{aligned} \ln(\text{Mercado}) = & -24,305 + 0,420 * \ln(\text{PIB}) - 0,198 * \ln(\text{Consumo das famílias}) \\ & + 2,230 * \ln(\text{População ocupada}) + 1,507 * \ln(\text{Massa salarial}) + 0,144 \\ & * \ln(\text{Taxa de desemprego}) - 0,253 * \ln(\text{Endividamento das famílias}) \\ & - 4,797 * \ln(\text{População economicamente ativa}) \end{aligned}$$

e a análise deste resultado será feita no capítulo 4.

4. Análise dos Resultados

A primeira observação que deve ser feita quanto ao resultado obtido é que a equação de regressão resultante do comando utilizado na figura 7 é aquela que minimiza o valor da soma dos quadrados do erro residual, neste caso, 0,00074. No entanto, como visto na revisão bibliográfica, séries temporais possuem autocorrelação nos erros residuais e a existência deste fenômeno não garante o método dos mínimos quadrados ordinários como o melhor estimador linear não-viesado. Portanto, é impossível garantir que a equação obtida é a melhor equação de regressão possível.

Outra observação que pode ser feita é em relação ao alto valor do R^2 ajustado = 0,9970. A revisão bibliográfica explica que muitos efeitos podem enviesar este número e que não necessariamente a regressão com este alto valor seja uma regressão tão boa quanto parece. Um destes efeitos pode ser a heterocedasticidade. No entanto, a aplicação do termo logarítmico em todos os termos da equação tem por objetivo a mitigação da heterocedasticidade dos erros.

No STATA, podemos plotar os resíduos em um gráfico para analisar a existência de heteroscedasticidade dos erros. O gráfico 3 é o resultado desta plotagem de dados.

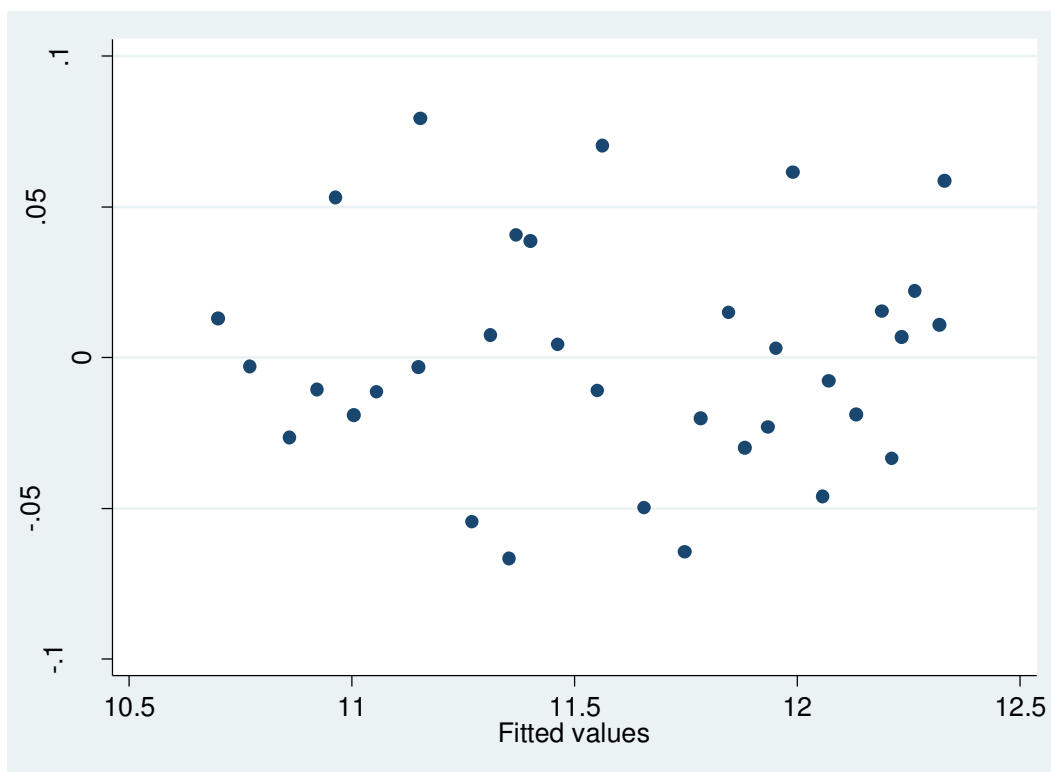


Gráfico 3– Erros residuais vs. valores ajustados

Fonte: Geração própria através do *software* STATA.

4. Análise dos Resultados

Visualmente, é possível perceber que não há aumento ou diminuição significativa do desvio padrão dos erros de acordo com o valor da variável dependente. Entende-se, portanto, que não existe heteroscedasticidade dos erros no modelo.

Um ponto de atenção são os valores negativos dos coeficientes das variáveis consumo das famílias e população economicamente ativa. Enquanto espera-se que o aumento destas variáveis implicaria em um aumento nos gastos com cartão de crédito, o resultado observado é exatamente o contrário. Da mesma forma, aumentar a taxa de desemprego causa, de acordo com o modelo, um aumento nos gastos com cartão de crédito, o que também é contraintuitivo.

O motivo destes paradoxos é, provavelmente, causa da multicolinearidade, ou seja, a variação destas variáveis pode ser parcialmente – ou totalmente – explicada pela variação em outras variáveis. Neste caso, o consumo das famílias é um número que está dentro do PIB, por exemplo, enquanto a taxa de desemprego, a população economicamente ativa e a população ocupada podem ser variáveis que explicam umas as outras.

```
. reg lnmercado lnpibnominal lnvarejoamp lnpopocup lnmassasalarial lnendividamento
```

Source	SS	df	MS	Number of obs =	34
Model	8.27189999	5	1.65438	F(5, 28) =	2052.36
Residual	.022570459	28	.000806088	Prob > F =	0.0000
				R-squared =	0.9973
				Adj R-squared =	0.9968
Total	8.29447045	33	.251347589	Root MSE =	.02839

lnmercado	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lnpibnominal	.3492239	.2086063	1.67	0.105	-.0780867 .7765346
lnvarejoamp	.9192946	.1793078	5.13	0.000	.5519992 1.28659
lnpopocup	-1.505657	.9684138	-1.55	0.131	-3.489363 .4780483
lnmassasalarial	1.30935	.156994	8.34	0.000	.9877627 1.630938
lnendividamento	-.2000645	.145615	-1.37	0.180	-.4983434 .0982144
_cons	13.00178	8.895444	1.46	0.155	-5.219707 31.22327

```
. estimate stats
```

Model	Obs	ll (null)	ll (model)	df	AIC	BIC
.	34	-24.2608	76.15315	6	-140.3063	-131.1481

Figura 8– Obtenção da regressão múltipla

Fonte: Geração própria através do *software* STATA

4. Análise dos Resultados

Para corrigir este problema, existem duas maneiras. A primeira, mais complexa, exige a correção dos dados e estudo mais profundo das variáveis. A segunda solução é tirar uma das variáveis causadoras do problema. Opta-se, portanto, pela exclusão das três variáveis citadas anteriormente e, assim, o modelo passa a ter cinco variáveis dependentes. O resultado aparece na figura 8.

Como já era esperado, o valor de R^2 tem leve redução. Além disso, os valores de t-student para todas as cinco variáveis são agora significativos. Além disso, o critério de informação do modelo com estas cinco variáveis restantes é bastante próximo do modelo com oito variáveis anterior.

Outro efeito que pode causar enviesamento da análise é a autocorrelação dos erros, que implica que o valor acima ou abaixo da média de uma observação não torna a observação seguinte mais ou menos provável de ser acima da média. Para estudar a existência deste efeito, o STATA disponibiliza o teste de Durbin-Watson, que trabalha com a hipótese nula H_0 de que a análise está livre de autocorrelação. O resultado obtido a partir do comando aparece na figura 9.

```
. dwstat  
  
Durbin-Watson d-statistic( 6, 34) = 2.450266
```

Figura 9– Resultado do teste de Durbin-Watson

Fonte: Geração própria através do *software* STATA

Portanto, $d_{calc} = 2,45$, com $k = 6$ e $n = 34$. De acordo com a tabela 3 no anexo A, para um índice de significância de 1%, os valores de dU e dL são, respectivamente, 1,677 e 0,896. Com isso, temos que $4 - dU < d_{calc} < 4 - dL$, o que significa que o teste aponta para uma zona de indecisão.

Para o índice de significância de 5%, de acordo com a tabela 4 no anexo B, os valores de dU e dL são 1,891 e 1,079, respectivamente. O mesmo resultado, de zona de indecisão, é obtido. Assim, com o teste de Durbin-Watson, não é possível aceitar nem rejeitar a hipótese nula de que não existe autocorrelação dos erros da regressão.

No gráfico 4 são traçados os valores para o mercado a partir da equação de regressão obtida em comparação com os valores reais divulgados pela ABECS. É possível notar grande similaridade entre as duas curvas.

Além disso, fica bastante claro que as inflexões no gráfico traçado com os dados reais são acompanhadas pelo modelo proposto, o que indica a existência de cointegração do modelo à curva, como já havia sido testado anteriormente.

4. Análise dos Resultados



Gráfico 4– Comparação valor real e equação de regressão obtida

Fonte: Geração própria.

5. Conclusões e Discussões

Através de um estudo do mercado de pagamentos eletrônicos e de uma revisão da bibliografia de análises multivariada, chegou-se a uma equação de regressão múltipla para o volume transacionado pelo mercado de pagamentos eletrônicos no Brasil. Durante o desenvolvimento deste assunto, percebeu-se que a obtenção de uma equação não é tarefa trivial e que os resultados, por muitas vezes, podem levar o pesquisador a tomar conclusões incorretas.

Exemplos disso, são os efeitos que inflam o valor do R^2 , como multicolinearidade e autocorrelação, mas que, apesar disso, não melhoram o poder de predição do modelo de regressão. Omissão de variáveis importantes ou inclusão de variáveis inúteis também podem enviesar o resultado final, assim como a utilização de regressões espúrias. Portanto, cada passo de obtenção de um modelo de regressão requer um questionamento crítico dos resultados.

Neste modelo de previsão do mercado de pagamentos eletrônicos, entendeu-se que as variáveis que melhor explicam as variações no mercado são o PIB (nominal), o volume de vendas do varejo, a população ocupada, a taxa de endividamento das famílias e a massa salarial. Os testes de efeitos que podem enviesar o modelo, como heteroscedasticidade e autocorrelação foram realizados. Enquanto a existência do primeiro foi descartada, os resultados dos testes para autocorrelação foram inconclusivos.

A zona de indefinição na qual se encontrou o teste de autocorrelação exige uma investigação mais específica sobre o assunto e, caso comprovada a existência, ela deve ser corrigida. No entanto, a tarefa de investigação pode ser muito complexa, assim como o passo de correção. Deve-se ter em mente que a existência de autocorrelação pode ser um dos efeitos do alto valor de R^2 , mas que sua investigação e correção fogem ao escopo deste trabalho devido à complexidade inerente.

O resultado final obtido foi a equação de regressão

$$\ln(\text{Mercado}) = 13,002 + 0,349 * \ln(\text{PIB}) + 0,919 * \ln(\text{Volume de vendas do Varejo}) \\ - 1,506 * \ln(\text{População ocupada}) + 1,309 * \ln(\text{Massa salarial}) - 0,200 \\ * \ln(\text{Taxa de endividamento das famílias})$$

que, como visto, não necessariamente representa a melhor equação de regressão. Isto ocorre porque a possível existência de autocorrelação não garante o método dos mínimos quadrados ordinários como o melhor estimador linear não-viesado (BLUE).

5. Conclusões e Discussões

Outro ponto que mereceu atenção foram os coeficientes de algumas variáveis terem valor contraintuitivo. A multicolinearidade pode ter sido a causa deste problema e esta foi reduzida com a exclusão de três variáveis independentes.

Como mostrou o gráfico 3, equação encontrada apresenta resultados bem similares aos reais. No entanto, existe a possibilidade de melhorias, como é o caso da investigação de autocorrelação. Esta investigação podem ser consideradas desafios futuros para o estudo de previsão do mercado de pagamentos eletrônicos.

Referências Bibliográficas

ALVES, A. A. e MENEZES, O.M.S. *Cartão de Crédito Private Label*. Novatec, 2007.

AMADEU, J. R. *Desenvolvimento de Modelos para a Previsão de Spread do Preço de Energia Elétrica no Mercado de Curto Prazo*. Monografia (Graduação) – Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2011.

AMORIM, M. C. *Previsão de séries temporais usando séries exógenas e combinação de redes neurais aplicada ao mercado financeiro*. Dissertação de Mestrado. Centro de Informática, Universidade Federal de Pernambuco, 2008.

BRETERNITZ, V. J. e SILVA, L. A. *Big Data: Trazendo Novas Oportunidades e Desafios*. 10th International Conference on Information Systems and Technology Management – CONTECSI, 2013.

CONDE, D. M. R. *A Relação Entre Business Intelligence e o Processo de Inovação no Mercado de Cartões de Crédito Brasileiro: Um Estudo de Caso*. Dissertação de Mestrado. FGV São Paulo, 2012.

COOPER, D.R.; SCHINDLER, P.S. *Business research methods*. McGraw-Hill/Irwin, 2003.

CORRAR, L.J.; PAULO, E. ; DIAS FILHO, J.M *Análise Multivariada para os Cursos de Administração, Ciências Contábeis e Economia*. Atlas, 2007.

DAL BELLO, L. H. A. *Modelagem em experimentos mistura-processo para otimização de processos industriais*. Orientador: Antonio Fernando de Castro Vieira – Rio de Janeiro: Pontifícia Universidade Católica do Rio de Janeiro, 2010.

FEIJÓ, B. V. *A Revolução dos Dados*. Revista Exame PME – Pequenas e Médias Empresas, São Paulo, p. 30-43, set. 2013.

FIGUEIREDO, R. P. *A Evolução do Sistema de Pagamentos Brasileiro e o Desaparecimento do Cheque: Realidade ou Exagero?*. Dissertação de Mestrado. Ibmec São Paulo, 2006.

GREENE, W. H. *Econometrics Analysis*. 6 ed. Pearson, 2008.

GUJARATI, D. N. *Econometria Básica*. Rio de Janeiro: Elsevier, 2006.

Referências Bibliográficas

HAIR, J. F.; ANDERSON, R.E.; TATHAM, R.L.; BLACK, W.C. 2009. *Análise multivariada de dados*. 6. ed. Porto Alegre: Bookman.

HILL, R. C.; GRIFFITHS, W. E.; LIM, G. C. *Principles of Econometrics*. 3 ed. Wiley, 2008.
KENNEDY, P. *Manual de Econometria*. 6 ed. Rio de Janeiro: Elsevier, 2009.

LOPES, A. C. B. S. *Econometria II: Modelização Univariada de Séries Temporais: uma Introdução*. Universidade Técnica de Lisboa, 2001.

MACKINNON, J. G. *Critical Values for Cointegration Tests*. Working Papers 1227, Queen's University, Department of Economics, 1990.

MACKINNON, J. G.; ERICSSON, N. R. *Distributions of Error Correction Tests for Cointegration*. Econometric Society World Congress 2000 Contributed Papers 0561, Econometric Society, 2000.

NETO, M. M. J. *Estatística multivariada*. Revista de Filosofia e Ensino. 9 maio 2004. Disponível em: http://www.criticanarede.com/cien_estatistica.html. Acesso em: 22 setembro 2014.

OTEXTS, *Time series components*. Disponível em: <https://www.otexts.org/fpp/6/1>. Acesso em 13/11/2014.

RÉGIS, D. E. e ARTES, R. *Modelo Multiestado de Markov em Cartões de Crédito*. Insper Working Paper, 2008.

SAVIN, N.E.; WHITE, K.J. *The Durbin-Watson Test for Serial Correlation with Extreme Sample Sizes or Many Regressors*. *Econometrica* 45, 1977.

SCHAFFER, M.E. 2010. *egranger: Engle-Granger (EG) and Augmented. Engle-Granger (AEG) cointegration tests and 2-step ECM estimation*. Disponível em <http://ideas.repec.org/c/boc/bocode/s457210.html>. Acesso em 15/11/2014.

VICINI, L. *Análise multivariada da teoria à prática*. Orientador Adriano Mendonça Souza. - Santa Maria : UFSM, CCNE, 2005.

Anexos

ANEXO A – Estatísticas de Durbin-Watson com 1% de significância

Tabela 2 – Estatística de Durbin-Watson: pontos dU e dL com 1% de significância																				
n	k = 1		k = 2		k = 3		k = 4		k = 5		k = 6		k = 7		k = 8		k = 9		k = 10	
	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU
6	0.390	1.142	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
7	0.435	1.036	0.294	1.676	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8	0.497	1.003	0.345	1.489	0.229	2.102	-	-	-	-	-	-	-	-	-	-	-	-	-	-
9	0.554	0.998	0.408	1.389	0.279	1.875	0.183	2.433	-	-	-	-	-	-	-	-	-	-	-	-
10	0.604	1.001	0.466	1.333	0.340	1.733	0.230	2.193	0.150	2.690	-	-	-	-	-	-	-	-	-	-
11	0.653	1.010	0.519	1.297	0.396	1.640	0.286	2.030	0.193	2.453	0.124	2.892	-	-	-	-	-	-	-	-
12	0.697	1.023	0.569	1.274	0.449	1.575	0.339	1.913	0.244	2.280	0.164	2.665	0.105	3.053	-	-	-	-	-	-
13	0.738	1.038	0.616	1.261	0.499	1.526	0.391	1.826	0.294	2.150	0.211	2.490	0.140	2.838	0.090	3.182	-	-	-	-
14	0.776	1.054	0.660	1.254	0.547	1.490	0.441	1.757	0.343	2.049	0.257	2.354	0.183	2.667	0.122	2.981	0.078	3.287	-	-
15	0.811	1.070	0.700	1.252	0.591	1.465	0.487	1.705	0.390	1.967	0.303	2.244	0.226	2.530	0.161	2.817	0.107	3.101	0.068	3.374
16	0.844	1.086	0.738	1.253	0.633	1.447	0.532	1.664	0.437	1.901	0.349	2.153	0.269	2.416	0.200	2.681	0.142	2.944	0.094	3.201
17	0.873	1.102	0.773	1.255	0.672	1.432	0.574	1.631	0.481	1.847	0.393	2.078	0.313	2.319	0.241	2.566	0.179	2.811	0.127	3.053
18	0.902	1.118	0.805	1.259	0.708	1.422	0.614	1.604	0.522	1.803	0.435	2.015	0.355	2.238	0.282	2.467	0.216	2.697	0.160	2.925
19	0.928	1.133	0.835	1.264	0.742	1.416	0.650	1.583	0.561	1.767	0.476	1.963	0.396	2.169	0.322	2.381	0.255	2.597	0.196	2.813
20	0.952	1.147	0.862	1.270	0.774	1.410	0.684	1.567	0.598	1.736	0.515	1.918	0.436	2.110	0.362	2.308	0.294	2.510	0.232	2.174
21	0.975	1.161	0.889	1.276	0.803	1.408	0.718	1.554	0.634	1.712	0.552	1.881	0.474	2.059	0.400	2.244	0.331	2.434	0.268	2.625
22	0.997	1.174	0.915	1.284	0.832	1.407	0.748	1.543	0.666	1.691	0.587	1.849	0.510	2.015	0.437	2.188	0.368	2.367	0.304	2.548
23	1.017	1.186	0.938	1.290	0.858	1.407	0.777	1.535	0.699	1.674	0.620	1.821	0.545	1.977	0.473	2.140	0.404	2.308	0.340	2.479
24	1.037	1.199	0.959	1.298	0.881	1.407	0.805	1.527	0.728	1.659	0.652	1.797	0.578	1.944	0.507	2.097	0.439	2.255	0.375	2.417
25	1.055	1.210	0.981	1.305	0.906	1.408	0.832	1.521	0.756	1.645	0.682	1.776	0.610	1.915	0.540	2.059	0.473	2.209	0.409	2.362
26	1.072	1.222	1.000	1.311	0.928	1.410	0.855	1.517	0.782	1.635	0.711	1.759	0.640	1.889	0.572	2.026	0.505	2.168	0.441	2.313
27	1.088	1.232	1.019	1.318	0.948	1.413	0.878	1.514	0.808	1.625	0.738	1.743	0.669	1.867	0.602	1.997	0.536	2.131	0.473	2.269
28	1.104	1.244	1.036	1.325	0.969	1.414	0.901	1.512	0.832	1.618	0.764	1.729	0.696	1.847	0.630	1.970	0.566	2.098	0.504	2.229
29	1.119	1.254	1.053	1.332	0.988	1.418	0.921	1.511	0.855	1.611	0.788	1.718	0.723	1.830	0.658	1.947	0.595	2.068	0.533	2.193
30	1.134	1.264	1.070	1.339	1.006	1.421	0.941	1.510	0.877	1.606	0.812	1.707	0.748	1.814	0.684	1.925	0.622	2.041	0.562	2.160
31	1.147	1.274	1.085	1.345	1.022	1.425	0.960	1.509	0.897	1.601	0.834	1.698	0.772	1.800	0.710	1.906	0.649	2.017	0.589	2.131
32	1.160	1.283	1.100	1.351	1.039	1.428	0.978	1.509	0.917	1.597	0.856	1.690	0.794	1.788	0.734	1.889	0.674	1.995	0.615	2.104
33	1.171	1.291	1.114	1.358	1.055	1.432	0.995	1.510	0.935	1.594	0.876	1.683	0.816	1.776	0.757	1.874	0.698	1.975	0.641	2.080
34	1.184	1.298	1.128	1.364	1.070	1.436	1.012	1.511	0.954	1.591	0.896	1.677	0.837	1.766	0.779	1.860	0.722	1.957	0.665	2.057
35	1.195	1.307	1.141	1.370	1.085	1.439	1.028	1.512	0.971	1.589	0.914	1.671	0.857	1.757	0.800	1.847	0.744	1.940	0.689	2.037
36	1.205	1.315	1.153	1.376	1.098	1.442	1.043	1.513	0.987	1.587	0.932	1.666	0.877	1.749	0.821	1.836	0.766	1.925	0.711	2.018
37	1.217	1.322	1.164	1.383	1.112	1.446	1.058	1.514	1.004	1.585	0.950	1.662	0.895	1.742	0.841	1.825	0.787	1.911	0.733	2.001
38	1.227	1.330	1.176	1.388	1.124	1.449	1.072	1.515	1.019	1.584	0.966	1.658	0.913	1.735	0.860	1.816	0.807	1.899	0.754	1.985
39	1.237	1.337	1.187	1.392	1.137	1.452	1.085	1.517	1.033	1.583	0.982	1.655	0.930	1.729	0.878	1.807	0.826	1.887	0.774	1.970
40	1.246	1.344	1.197	1.398	1.149	1.456	1.098	1.518	1.047	1.583	0.997	1.652	0.946	1.724	0.895	1.799	0.844	1.876	0.749	1.956
45	1.288	1.376	1.245	1.424	1.201	1.474	1.156	1.528	1.111	1.583	1.065	1.643	1.019	1.704	0.974	1.768	0.927	1.834	0.881	1.902
50	1.324	1.403	1.285	1.445	1.245	1.491	1.206	1.537	1.164	1.587	1.123	1.639	1.081	1.692	1.039	1.748	0.997	1.805	0.955	1.864
55	1.356	1.428	1.320	1.466	1.284	1.505	1.246	1.548	1.209	1.592	1.172	1.638	1.134	1.685	1.095	1.734	1.057	1.785	1.018	1.837
60	1.382	1.449	1.351	1.484	1.317	1.520	1.283	1.559	1.248	1.598	1.214	1.639	1.179	1.682	1.144	1.726	1.108	1.771	1.072	1.817
65	1.407	1.467	1.377	1.500	1.346	1.534	1.314	1.568	1.283	1.604	1.251	1.642	1.218	1.680	1.186	1.720	1.153	1.761	1.120	1.802
70	1.429	1.485	1.400	1.514	1.372	1.546	1.343	1.577	1.313	1.611	1.283	1.645	1.253	1.680	1.223	1.716	1.192	1.754	1.162	1.792
75	1.448	1.501	1.422	1.529	1.395	1.557	1.368	1.586	1.340	1.617	1.313	1.649	1.284	1.682	1.256	1.714	1.227	1.748	1.199	1.783
80	1.465	1.514	1.440	1.541	1.416	1.568	1.390	1.595	1.364	1.624	1.338	1.653	1.312	1.683	1.285	1.714	1.259	1.745	1.232	1.777
85	1.481	1.529	1.458	1.553	1.434	1.577	1.411	1.603	1.386	1.630	1.362	1.657	1.337	1.685	1.312	1.714	1.287	1.743	1.262	1.773
90	1.496	1.541	1.474	1.563	1.452	1.587	1.429	1.611	1.406	1.636	1.383	1.661	1.360	1.687	1.336	1.714	1.312	1.741	1.288	1.769
95	1.510	1.552	1.489	1.573	1.468	1.596	1.446	1.618	1.425	1.641	1.403	1.666	1.381	1.690	1.358	1.715	1.336	1.741	1.313	1.767
100	1.522	1.562	1.502	1.582	1.482	1.604	1.461	1.625	1.441	1.647	1.421	1.670	1.400	1.693	1.378	1.717	1.357	1.741	1.335	1.765
150	1.611	1.637	1.598	1.651	1.584	1.665	1.571	1.679	1.557	1.693	1.543	1.708	1.530	1.722	1.515	1.737	1.501	1.752	1.486	1.767
200	1.664	1.684	1.653	1.693	1.643	1.704	1.633	1.715	1.623	1.725	1.613	1.735	1.603	1.746	1.592	1.757	1.582	1.768	1.571	1.779

Tabela 3 – Estatística de Durbin-Watson: pontos dU e dL com 1% de significância

Fonte: Savin e White, 1977.

ANEXO B – Estatísticas de Durbin-Watson com 5% de significância

Tabela 3 – Estatística de Durbin-Watson: pontos dU e dL com 5% de significância																				
n	k = 1		k = 2		k = 3		k = 4		k = 5		k = 6		k = 7		k = 8		k = 9		k = 10	
	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU	dL	dU
6	0.610	1.400	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
7	0.700	1.356	0.467	1.896	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
8	0.763	1.332	0.559	1.777	0.367	2.287	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
9	0.824	1.320	0.629	1.699	0.455	2.128	0.296	2.588	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	2.414	0.243	2.822	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----
11	0.927	1.324	0.758	1.604	0.595	1.928	0.444	2.283	0.315	2.645	0.203	3.004	-----	-----	-----	-----	-----	-----	-----	-----
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.380	2.506	0.268	2.832	0.171	3.149	-----	-----	-----	-----	-----	-----
13	1.010	1.340	0.861	1.562	0.715	1.816	0.574	2.094	0.444	2.390	0.328	2.692	0.230	2.985	0.147	3.266	-----	-----	-----	-----
14	1.045	1.350	0.905	1.551	0.767	1.779	0.632	2.030	0.505	2.296	0.389	2.572	0.286	2.848	0.200	3.111	0.127	3.360	-----	-----
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220	0.447	2.471	0.343	2.727	0.251	2.979	0.175	3.216	0.111	3.438
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157	0.502	2.388	0.398	2.624	0.304	2.860	0.222	3.090	0.155	3.304
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104	0.554	2.318	0.451	2.537	0.356	2.757	0.272	2.975	0.198	3.184
18	1.158	1.391	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060	0.603	2.258	0.502	2.461	0.407	2.668	0.321	2.873	0.244	3.073
19	1.180	1.401	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023	0.649	2.206	0.549	2.396	0.456	2.589	0.369	2.783	0.290	2.974
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991	0.691	2.162	0.595	2.339	0.502	2.521	0.416	2.704	0.336	2.885
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.829	1.964	0.731	2.124	0.637	2.290	0.546	2.461	0.461	2.633	0.380	2.806
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940	0.769	2.090	0.677	2.246	0.588	2.407	0.504	2.571	0.424	2.735
23	1.257	1.437	1.168	1.543	1.078	1.660	0.986	1.785	0.895	1.920	0.804	2.061	0.715	2.208	0.628	2.360	0.545	2.514	0.465	2.670
24	1.273	1.446	1.188	1.546	1.101	1.656	1.013	1.775	0.925	1.902	0.837	2.035	0.750	2.174	0.666	2.318	0.584	2.464	0.506	2.613
25	1.288	1.454	1.206	1.550	1.123	1.654	1.038	1.767	0.953	1.886	0.868	2.013	0.784	2.144	0.702	2.280	0.621	2.419	0.544	2.560
26	1.302	1.461	1.224	1.553	1.143	1.652	1.062	1.759	0.979	1.873	0.897	1.992	0.816	2.117	0.735	2.246	0.657	2.379	0.581	2.513
27	1.316	1.469	1.240	1.556	1.162	1.651	1.084	1.753	1.004	1.861	0.925	1.974	0.845	2.093	0.767	2.216	0.691	2.342	0.616	2.470
28	1.328	1.476	1.255	1.560	1.181	1.650	1.104	1.747	1.028	1.850	0.951	1.959	0.874	2.071	0.798	2.188	0.723	2.309	0.649	2.431
29	1.341	1.483	1.270	1.563	1.198	1.650	1.124	1.743	1.050	1.841	0.975	1.944	0.900	2.052	0.826	2.164	0.753	2.278	0.681	2.396
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.071	1.833	0.998	1.931	0.926	2.034	0.854	2.141	0.782	2.251	0.712	2.363
31	1.363	1.496	1.297	1.570	1.229	1.650	1.160	1.735	1.090	1.825	1.020	1.920	0.950	2.018	0.879	2.120	0.810	2.226	0.741	2.333
32	1.373	1.502	1.309	1.574	1.244	1.650	1.177	1.732	1.109	1.819	1.041	1.909	0.972	2.004	0.904	2.102	0.836	2.203	0.769	2.306
33	1.383	1.508	1.321	1.577	1.258	1.651	1.193	1.730	1.127	1.813	1.061	1.900	0.994	1.991	0.927	2.085	0.861	2.181	0.796	2.281
34	1.393	1.514	1.333	1.580	1.271	1.652	1.208	1.728	1.144	1.808	1.079	1.891	1.015	1.978	0.950	2.069	0.885	2.162	0.821	2.257
35	1.402	1.519	1.343	1.584	1.283	1.653	1.222	1.726	1.160	1.803	1.097	1.884	1.034	1.967	0.971	2.054	0.908	2.144	0.845	2.236
36	1.411	1.525	1.354	1.587	1.295	1.654	1.236	1.724	1.175	1.799	1.114	1.876	1.053	1.957	0.991	2.041	0.930	2.127	0.868	2.216
37	1.419	1.530	1.364	1.590	1.307	1.655	1.249	1.723	1.190	1.795	1.131	1.870	1.071	1.948	1.011	2.029	0.951	2.112	0.891	2.197
38	1.427	1.535	1.373	1.594	1.318	1.656	1.261	1.722	1.204	1.792	1.146	1.864	1.088	1.939	1.029	2.017	0.970	2.098	0.912	2.180
39	1.435	1.540	1.382	1.597	1.328	1.658	1.273	1.722	1.218	1.789	1.161	1.859	1.104	1.932	1.047	2.007	0.990	2.085	0.932	2.164
40	1.442	1.544	1.391	1.600	1.338	1.659	1.285	1.721	1.230	1.786	1.175	1.854	1.120	1.924	1.064	1.997	1.008	2.072	0.952	2.149
45	1.475	1.566	1.430	1.615	1.383	1.666	1.336	1.720	1.287	1.776	1.238	1.835	1.189	1.895	1.139	1.958	1.089	2.022	1.038	2.088
50	1.503	1.585	1.462	1.628	1.421	1.674	1.378	1.721	1.335	1.771	1.291	1.822	1.246	1.875	1.201	1.930	1.156	1.986	1.110	2.044
55	1.528	1.601	1.490	1.641	1.452	1.681	1.414	1.724	1.374	1.768	1.334	1.814	1.294	1.861	1.253	1.909	1.212	1.959	1.170	2.010
60	1.549	1.616	1.514	1.652	1.480	1.689	1.444	1.727	1.408	1.767	1.372	1.808	1.335	1.850	1.298	1.894	1.260	1.939	1.222	1.984
65	1.567	1.629	1.536	1.662	1.503	1.696	1.471	1.731	1.438	1.767	1.404	1.805	1.370	1.843	1.336	1.882	1.301	1.923	1.266	1.964
70	1.583	1.641	1.554	1.672	1.525	1.703	1.494	1.735	1.464	1.768	1.433	1.802	1.401	1.838	1.369	1.874	1.337	1.910	1.305	1.948
75	1.598	1.652	1.571	1.680	1.543	1.709	1.515	1.739	1.487	1.770	1.458	1.801	1.428	1.834	1.399	1.867	1.369	1.901	1.339	1.935
80	1.611	1.662	1.586	1.688	1.560	1.715	1.534	1.743	1.507	1.772	1.480	1.801	1.453	1.831	1.425	1.861	1.397	1.893	1.369	1.925
85	1.624	1.671	1.600	1.696	1.575	1.721	1.550	1.747	1.525	1.774	1.500	1.801	1.474	1.829	1.448	1.857	1.422	1.886	1.396	1.916
90	1.635	1.679	1.612	1.703	1.589	1.726	1.566	1.751	1.542	1.776	1.518	1.801	1.494	1.827	1.469	1.854	1.445	1.881	1.420	1.909
95	1.645	1.687	1.623	1.709	1.602	1.732	1.579	1.755	1.557	1.778	1.535	1.802	1.512	1.827	1.489	1.852	1.465	1.877	1.442	1.903
100	1.654	1.694	1.634	1.715	1.613	1.736	1.592	1.758	1.571	1.780	1.550	1.803	1.528	1.826	1.506	1.850	1.484	1.874	1.462	1.898
150	1.720	1.747	1.706	1.760	1.693	1.774	1.679	1.788	1.665	1.802	1.651	1.817	1.637	1.832	1.622	1.846	1.608	1.862	1.593	1.877
200	1.758	1.779	1.748	1.789	1.738	1.799	1.728	1.809	1.718	1.820	1.707	1.831	1.697	1.841	1.686	1.852	1.675	1.863	1.665	1.874

Tabela 4 – Estatística de Durbin-Watson: pontos dU e dL com 5% de significância
 Fonte: Savin e White, 1977

Apêndices

APÊNDICE A – Cálculo do critério de informação da variável PIB nominal

```
. reg lnmercado lnpibnominal
```

Source	SS	df	MS	Number of obs = 34		
Model	8.20690727	1	8.20690727	F(1, 32) = 2999.22		
Residual	.087563181	32	.002736349	Prob > F = 0.0000		
Total	8.29447045	33	.251347589	R-squared = 0.9894		
				Adj R-squared = 0.9891		
				Root MSE = .05231		

lnmercado	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lnpibnominal	1.965528	.0358901	54.77	0.000	1.892422	2.038633
_cons	-1.729025	.2435921	-7.10	0.000	-2.225205	-1.232844


```
. estimate stats
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	34	-24.2608	53.10593	2	-102.2119	-99.15914

Figura 10 – Cálculo do critério de informação da variável PIB nominal
 Fonte: Geração própria através do *software* STATA

APÊNDICE B – Cálculo do critério de informação da variável consumo das famílias

```
. reg lnmercado lnconsumo
```

Source	SS	df	MS	Number of obs = 34		
Model	8.20593284	1	8.20593284	F(1, 32) = 2965.86		
Residual	.088537614	32	.0027668	Prob > F = 0.0000		
Total	8.29447045	33	.251347589	R-squared = 0.9893		
				Adj R-squared = 0.9890		
				Root MSE = .0526		

lnmercado	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lnconsumo	1.867125	.0342845	54.46	0.000	1.79729	1.936961
_cons	-.1363411	.2157356	-0.63	0.532	-.5757802	.303098


```
. estimate stats
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	34	-24.2608	52.91779	2	-101.8356	-98.78286

Figura 11 – Cálculo do critério de informação da variável consumo das famílias
 Fonte: Geração própria através do *software* STATA

APÊNDICE C – Cálculo do critério de informação da variável vendas do varejo

```
. reg lnmercado ln varejoamp
```

Source	SS	df	MS			
Model	7.93770153	1	7.93770153	Number of obs =	34	
Residual	.356768925	32	.011149029	F(1, 32) =	711.96	
Total	8.29447045	33	.251347589	Prob > F =	0.0000	
				R-squared =	0.9570	
				Adj R-squared =	0.9556	
				Root MSE =	.10559	

lnmercado	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ln varejoamp	2.355628	.0882831	26.68	0.000	2.175801	2.535455
_cons	-1.530032	.4924997	-3.11	0.004	-2.533222	-.5268435


```
. estimate stats
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	34	-24.2608	29.22556	2	-54.45111	-51.39839

Figura 12 – Cálculo do critério de informação da variável vendas do varejo
 Fonte: Geração própria através do *software* STATA

APÊNDICE D – Cálculo do critério de informação da variável PEA

```
. reg lnmercado lnpea
```

Source	SS	df	MS			
Model	7.97144845	1	7.97144845	Number of obs =	34	
Residual	.323021998	32	.010094437	F(1, 32) =	789.69	
Total	8.29447045	33	.251347589	Prob > F =	0.0000	
				R-squared =	0.9611	
				Adj R-squared =	0.9598	
				Root MSE =	.10047	

lnmercado	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lnpea	14.19486	.5051303	28.10	0.000	13.16594	15.22377
_cons	-131.2098	5.08206	-25.82	0.000	-141.5616	-120.8579


```
. estimate stats
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	34	-24.2608	30.91481	2	-57.82962	-54.7769

Figura 13 – Cálculo do critério de informação da variável PEA
 Fonte: Geração própria através do *software* STATA

APÊNDICE E – Cálculo do critério de informação da variável população ocupada

```
. reg lnmercado lnpopocup
```

Source	SS	df	MS			
Model	8.10492795	1	8.10492795	Number of obs =	34	
Residual	.189542503	32	.005923203	F(1, 32) =	1368.34	
Total	8.29447045	33	.251347589	Prob > F =	0.0000	
				R-squared =	0.9771	
				Adj R-squared =	0.9764	
				Root MSE =	.07696	

lnmercado	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lnpopocup	9.27446	.2507219	36.99	0.000	8.763756	9.785163
_cons	-81.00957	2.503666	-32.36	0.000	-86.10937	-75.90977


```
. estimate stats
```

Model	Obs	ll (null)	ll (model)	df	AIC	BIC
.	34	-24.2608	39.97763	2	-75.95527	-72.90254

Figura 14 – Cálculo do critério de informação da variável população ocupada
 Fonte: Geração própria através do *software* STATA

APÊNDICE F – Cálculo do critério de informação da variável taxa de desemprego

```
. reg lnmercado lntaxadesemp
```

Source	SS	df	MS			
Model	7.97633664	1	7.97633664	Number of obs =	34	
Residual	.31813381	32	.009941682	F(1, 32) =	802.31	
Total	8.29447045	33	.251347589	Prob > F =	0.0000	
				R-squared =	0.9616	
				Adj R-squared =	0.9604	
				Root MSE =	.09971	

lnmercado	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lntaxadesemp	-1.983784	.0700362	-28.33	0.000	-2.126443	-1.841125
_cons	6.330165	.1869123	33.87	0.000	5.949437	6.710893


```
. estimate stats
```

Model	Obs	ll (null)	ll (model)	df	AIC	BIC
.	34	-24.2608	31.17403	2	-58.34807	-55.29535

Figura 15 – Cálculo do critério de informação da variável taxa de desemprego
 Fonte: Geração própria através do *software* STATA

APÊNDICE G – Cálculo do critério de informação da variável massa salarial

```
. reg lnmercado lnmassasalarial
```

Source	SS	df	MS	
Model	8.18605732	1	8.18605732	Number of obs = 34
Residual	.108413131	32	.00338791	F(1, 32) = 2416.26
Total	8.29447045	33	.251347589	Prob > F = 0.0000
				R-squared = 0.9869
				Adj R-squared = 0.9865
				Root MSE = .05821

lnmercado	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lnmassasalarial	1.861152	.0378626	49.16	0.000	1.784028 1.938276
_cons	3.16984	.1718366	18.45	0.000	2.819821 3.51986

```
. estimate stats
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	34	-24.2608	49.47492	2	-94.94984	-91.89712

Figura 16 – Cálculo do critério de informação da variável massa salarial

Fonte: Geração própria através do *software* STATA

APÊNDICE H – Cálculo do critério de informação da variável endividamento das famílias

```
. reg lnmercado lnendividamento
```

Source	SS	df	MS	
Model	7.98241359	1	7.98241359	Number of obs = 34
Residual	.312056859	32	.009751777	F(1, 32) = 818.56
Total	8.29447045	33	.251347589	Prob > F = 0.0000
				R-squared = 0.9624
				Adj R-squared = 0.9612
				Root MSE = .09875

lnmercado	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lnendividamento	2.196047	.0767567	28.61	0.000	2.039699 2.352396
_cons	13.90284	.0821739	169.19	0.000	13.73545 14.07022

```
. estimate stats
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	34	-24.2608	31.50191	2	-59.00381	-55.95109

Figura 17 – Cálculo do critério de informação da variável endividamento das famílias

Fonte: Geração própria através do *software* STATA

APÊNDICE I – Cálculo do critério de informação da variável inflação – IPCA

```
. reg lnmercado lnipca
```

Source	SS	df	MS	Number of obs = 34		
Model	3.95067703	1	3.95067703	F(1, 32) =	29.10	
Residual	4.34379343	32	.135743545	Prob > F =	0.0000	
Total	8.29447045	33	.251347589	R-squared =	0.4763	
				Adj R-squared =	0.4599	
				Root MSE =	.36843	

lnmercado	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lnipca	1.545905	.2865541	5.39	0.000	.9622134	2.129597
_cons	16.18145	.8511598	19.01	0.000	14.4477	17.91521


```
. estimate stats
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	34	-24.2608	-13.2645	2	30.529	33.58172

Figura 18 – Cálculo do critério de informação da variável inflação – IPCA
 Fonte: Geração própria através do *software* STATA

APÊNDICE J – Cálculo do critério de informação da variável inflação – IGP-M

```
. reg lnmercado lnigpm
```

Source	SS	df	MS	Number of obs = 32		
Model	1.49892186	1	1.49892186	F(1, 30) =	6.64	
Residual	6.7760582	30	.225868607	Prob > F =	0.0152	
Total	8.27498006	31	.266934841	R-squared =	0.1811	
				Adj R-squared =	0.1538	
				Root MSE =	.47526	

lnmercado	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lnigpm	.2842494	.1103413	2.58	0.015	.0589025	.5095964
_cons	12.44667	.3371736	36.91	0.000	11.75807	13.13527


```
. estimate stats
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	32	-23.76604	-20.56859	2	45.13717	48.06865

Figura 19 – Cálculo do critério de informação da variável inflação – IGP-M
 Fonte: Geração própria através do *software* STATA

APÊNDICE K – Cálculo do critério de informação da variável Selic

```
. reg lnmercado lnselic
```

Source	SS	df	MS	Number of obs =	34
Model	3.83649216	1	3.83649216	F(1, 32) =	27.54
Residual	4.45797829	32	.139311822	Prob > F =	0.0000
Total	8.29447045	33	.251347589	R-squared =	0.4625
				Adj R-squared =	0.4457
				Root MSE =	.37324

lnmercado	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lnselic	-1.634997	.3115615	-5.25	0.000	-2.269627 -1.000367
_cons	7.950451	.6988216	11.38	0.000	6.526998 9.373904

```
. estimate stats
```

Model	Obs	ll (null)	ll (model)	df	AIC	BIC
.	34	-24.2608	-13.7056	2	31.4112	34.46393

Figura 20 – Cálculo do critério de informação da variável Selic
 Fonte: Geração própria através do *software* STATA