

**UNIVERSIDADE FEDERAL DO PARANÁ
CENTRO DE PESQUISA E PÓS-GRADUAÇÃO EM
ADMINISTRAÇÃO DA UFPR**

ADAM PAUL PATTERSON

**MODELAGEM DE REGRESSÃO DO MERCADO DE E-COMMERCE
NO BRASIL**

CURITIBA

2014

ADAM PAUL PATTERSON

**MODELAGEM DE REGRESSÃO DO MERCADO DE E-COMMERCE
NO BRASIL**

Artigo científico apresentado ao curso de MBA
em Finanças da Universidade Federal do Paraná,
como requisito à obtenção do título de especialista.
Orientador: Prof. José Roberto Frega

CURITIBA

2014

SUMÁRIO

1. INTRODUÇÃO	4
1.1 Background e objetivos	4
1.2 Justificativa do estudo	4
1.3 Referencial teórico – Regressão múltipla	5
2. INTRODUÇÃO AO MERCADO DE E-COMMERCE	6
2.1 Descrevendo e entendendo os dados	7
2.2 Teorema central do limite	9
2.3 Elaborando uma projeção “ <i>baseline</i> ”	10
3. REGRESSÃO MÚLTIPLA	15
3.1 Overview do método	15
3.2 Definindo as variáveis	16
3.3 Construindo o modelo	18
3.4 Testando o modelo	20
3.5 “Regressando” o futuro	22
4. CONCLUSÕES	24
4.1 Considerações finais	24
4.2 Próximos passos	24
REFERÊNCIAS BIBLIOGRÁFICAS	25

1.0 INTRODUÇÃO

1.1 Background e objetivos

“E-commerce”, que em português significa comércio eletrônico, é o comércio que realiza suas transações financeiras por meio de dispositivos e plataformas digitais, tal como computadores, celulares e tablets. Como uma indústria que surgiu relativamente recentemente nos últimos 15 anos, não existe um grande material de estudos estatísticos do crescimento histórico e previsões econômicas de sua potencial evolução futura. Assim, existe um “gap” para tal análise.

O estudo a seguir possui quatro objetivos macros:

1. Explorar o crescimento deste setor nos últimos anos e seu “drivers” chave;
2. Analisar seu crescimento e tendências históricas e mapear um “forecast baseline” para comparar com nosso modelo de regressão principal;
3. Elaborar um modelo econômico através de uma regressão múltipla para estimar a trajetória futura do mercado de E-Commerce brasileiro;
4. Testar a assertividade e confiabilidade estatística deste modelo

Um *guiding principle* deste estudo é realizar toda esta análise e modelagem econômica no Microsoft Excel© para demonstrar sua usabilidade em um ambiente profissional amplo onde este programa é muito conhecido e usado sem a necessidade de ter um pacote de software analítico mais complexo para o usuário geral.

A ideia é aplicar uma principia da ‘democratização’ na modelagem financeira de forma clara e simples para o profissional sem um grande conhecimento estatístico para extrapolar inteligência dos *growth paths* econômicos do seu próprio mercado.

1.2 Justificativa do estudo

- i. E-commerce é um mercado relativamente novo que se desenvolveu de acordo com a popularização da internet na década de 1990 e, como tal, há muito menos informação disponível sobre o mercado – especialmente para o mercado brasileiro – em comparação com indústrias mais maduras e desenvolvidas. Assim, existe um potencial “falta de informação” para este mercado.

- ii. Além disso, o autor acredita que o conceito de mapear previsões econômicas em Excel de forma simples também se acha um “gap” no atual mercado analítico profissional.

1.3 Referencial Teórico - Regressão múltipla

Segundo o Kasznar e Gonçalves (2012: p2):

“A Regressão Múltipla é um dos inúmeros modelos estatísticos explanatórios causais referentes ao tratamento de séries temporais de dados. Sua base estatística advém da Regressão Linear, que se restringe a duas variáveis e a apenas uma equação funcional do primeiro grau ($Y = a + bX$) de ajustamento”.

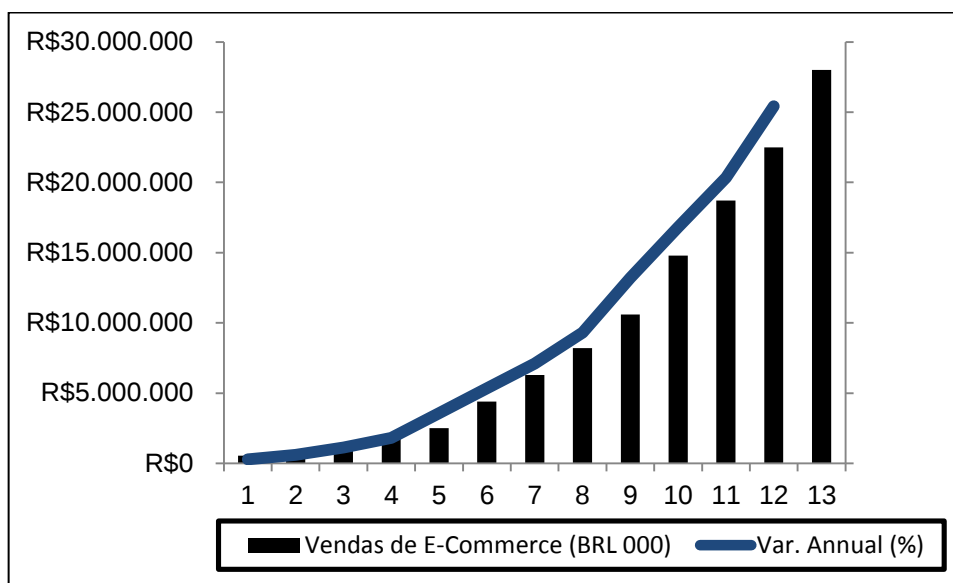
A análise de Regressão Múltipla é uma metodologia estatística de previsão de valores de uma ou mais variáveis de resposta (Dependentes) através de um conjunto de variáveis explicativas (Independentes).

Sua aplicação é especialmente importante, pois permite que se estime o valor de uma variável com base num conjunto de outras variáveis. Quanto mais significativo for o peso de uma variável isolada, ou de um conjunto de variáveis explicativas, mais se poderá afirmar que alguns fatores afetam o comportamento de uma variável de resposta especificamente procurada, do que outros.

2.0 INTRODUÇÃO AO MERCADO DE E-COMMERCE – SNAPSHOT SETORIAL

O mercado de e-commerce no Brasil vem crescendo em uma tendência bastante positiva. O volume do setor foi de R\$ 28,8 bilhões em 2013, crescendo, nominalmente, 28%, em comparação a 2012 quando atingiu um valor de R\$ 22,5 bilhões¹. O ano passado apresentou um resultado melhor que 2012 em virtude de maior crescimento econômico visto e da aceleração das vendas de dispositivos móveis como tablets e smartphones. O número de consumidores virtuais também saltou em 2013, superando 50 milhões de pessoas que já realizam uma compra online no Brasil.

Figura 1 – Evolução histórica do setor de E-commerce 2001-2013



Embora ainda em um estágio relativamente inicial de desenvolvimento de comércio eletrônico, esses números colocam o Brasil como o decimo maior mercado mundial em 2014 classificados por vendas de comércio eletrônico de varejo e o maior mercado na América de Sul. Ademais, há grade espaço para o crescimento futuro do setor puxado principalmente pelo aumento de usuários de internet a 110 milhões em 2015, o investimento publico na rede de internet no país e o crescimento da classe média nos últimos dez anos².

¹ <http://www.e-commerce.org.br/stats.php>

² <http://canaltech.com.br/noticia/mobile/Numero-de-usuarios-brasileiros-de-internet-movel-deve-aumentar-ate-2015/>

Todavia, existem desafios a este cenário positivo. Embora o mercado interno do Brasil esteja relativamente robusto, a cena internacional e indicadores macroeconômicos poderiam influenciar a trajetória no resto deste ano e para frente. A crise em curso na Europa (que não tem solução de curto prazo), a desaceleração esperada no crescimento chinês e o fato de que a economia dos EUA está realizando “escape velocity” e deve reduzir o afrouxamento monetário, poderiam gerar algum efeito negativo sobre o ritmo de crescimento econômico, afetando o setor de varejo do Brasil como um todo. Por outro lado, uma redução da taxa básica de juros e maior estabilidade depois antes da eleição presidencial em 2014 podem equilibrar esse impacto potencial.

Assim, apesar desse grande potencial de crescimento no mercado de e-commerce no Brasil e a importância de estratégias de comércio digital no setor de e-commerce não existe um amplo material de estudos econômicos focado no seu desempenho histórico e estimativas do seu comportamento futuro com base na análise estatística especialmente em um formato “user friendly”. Portanto, isso é o foco do estudo. Os próximos passos são mapear um fluxo de trabalho lógico e compreensivo para ser aplicado no processo de previsões econômicas.

2.1 Descrevendo e entendendo os dados

Um passo fundamental na modelagem financeira é analisar e entender sua base de dados. A análise dos dados é um processo de inspeção, limpeza, transformação, e modelagem de dados, com o objetivo de descobrir informações úteis, sugerindo conclusões e apoiar a tomada de decisão. A análise dos dados tem múltiplas facetas e abordagens, abrangendo diversas técnicas sob uma variedade de nomes, em diferentes negócios, ciência e domínios das ciências sociais.

Na modelagem econômica, alguns pesquisadores e analistas dividem análise de dados em estatística descritiva, análise exploratória de dados (*exploratory data analysis* ou EDA em inglês), e análise de dados de confirmação (*confirmatory data analysis* ou CDA em inglês). EDA centra-se na descoberta de novas características nos dados ("Knowledge Discovery") e análise preditiva centra-se na aplicação de modelos estatísticos e estruturais para a previsão ou classificação. CDA concentra-se em confirmar ou falsificar hipóteses existentes³.

Primeiramente, podemos analisar nossos dados com estatísticas descritivas. A estatística descritiva é a disciplina de descrever quantitativamente as principais características

³ <http://www.statistics.com/glossary/>

de um conjunto de informações e fornece resumos simples sobre a amostra e sobre as observações que foram feitas. Algumas medidas que são comumente usadas para descrever um conjunto de dados são medidas de tendência central e medidas de variabilidade ou dispersão⁴.

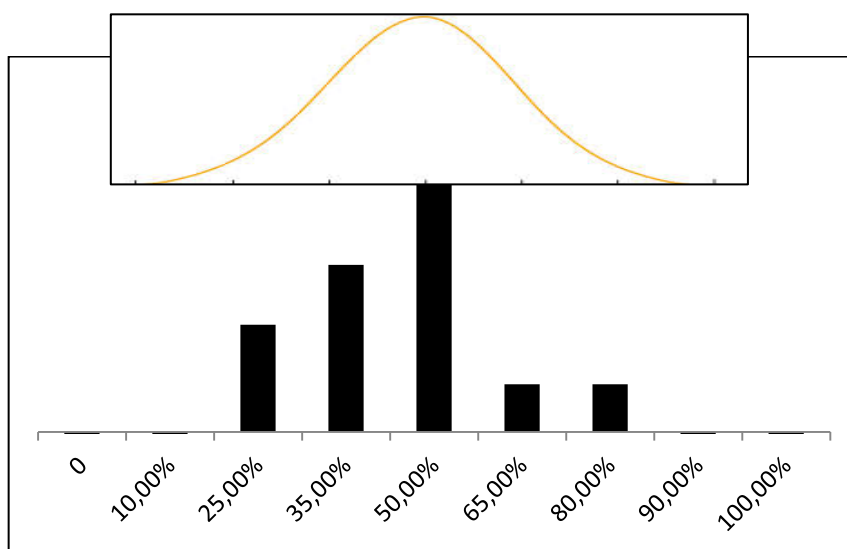
Observando nossos dados (veja a tabela 1) com o crescimento histórico podemos começar a descrever sua forma e comportamento. Podemos traçar que a evolução do setor é bastante positivo, com crescimento total 5085% entre 2001 e 2013, com CAGR (taxa de crescimento anual composta) de 38,95% e variância – ou dispersão estatística em relação à média – de 15,71%. O intervalo de crescimento histórico é de mínimo 20,32% ao máximo de 76,0%.

Tabela 1 – Crescimento histórico do setor de E-commerce 2001-2013 e estatísticas descritivas

Ano	Vendas de E-Commerce (BRL 000)	Estatísticas Descritivas	
2001	R\$ 540.000	Crescimento de 2001 a 2013	5085,19%
2002	R\$ 850.000	Crescimento anual médio	39,73%
2003	R\$ 1.180.000	CAGR	38,96%
2004	R\$ 1.750.000	Mediana	39,22%
2005	R\$ 2.500.000	Nível de confiança (95%)	9,98%
2006	R\$ 4.400.000	Mínimo	20,32%
2007	R\$ 6.300.000	Máximo	76,00%
2008	R\$ 8.200.000	Intervalo	55,68%
2009	R\$ 10.600.000	Desvio padrão	15,71%
2010	R\$ 14.800.000	Erro padrão	4,53%
2011	R\$ 18.700.000		
2012	R\$ 22.500.000		
2013	R\$ 28.000.000		

⁴ http://www.investopedia.com/terms/d/descriptive_statistics.asp#axzz2DxCoTnMM

Figura 2 – Histograma das percentuais de crescimento anual



2.2 Teorema central do limite

O teorema central de limite é uma teoria estatística que indica que, dado um tamanho suficientemente grande, a amostra de uma população com um nível limitado de variância, a média de todas as amostras a partir da mesma população será aproximadamente igual à média da população. Além disso, todas as amostras irão seguir um padrão de distribuição normal aproximada, com todas as variações, sendo aproximadamente igual a da variância da população dividida pelo tamanho de cada amostra (*Cowpertwait P.S.P., Metcalfe A.V, 2009*).

Com o aumento do tamanho da amostra, a distribuição amostral da média, $\bar{X}$, pode ser aproximada por uma distribuição normal com média μ e desvio padrão $\sigma / \sqrt{n}$ onde:

μ é a média da população

σ é o desvio padrão da população

n é o tamanho da amostra

Assim, se tomarmos repetidamente amostras aleatórias independentes de tamanho n de qualquer população, em seguida, quando n é grande, a distribuição dos meios de amostragem irão aproximar-se de uma distribuição normal.

Quão grande é grande o suficiente? De um modo geral, um tamanho de amostra de 30 ou mais é considerado grande o suficiente para o teorema do limite central para fazer efeito. Quanto mais perto a distribuição da população é a uma distribuição normal, menos amostras necessárias para demonstrar o teorema. Populações que são fortemente enviesadas ou têm vários modos podem exigir amostras maiores.

Isso é importante porque o campo das estatísticas é baseado no fato de que raramente é viável ou prática de coletar todos os dados a partir de uma população inteira. Em vez disso, podemos reunir um subconjunto de dados de uma população, e, em seguida, usar as estatísticas para essa amostra tirar conclusões sobre a população.

Neste estudo, como nosso objetivo fundamental é mostrar a metodologia atrás de modelagem de regressão de forma compreensiva, o “como fazer” é mais importante e assim a ideia é utilizar uma margem de tolerância em termos das necessidades estatísticas gerais. Por isso, nossa amostra somente possui 13 anos de dados – de 2001 a 2013. Também, é importante ressaltar que como o mercado em questão é o e-commerce do Brasil, o que é relativamente novo, não existem dados históricos muito além do nosso período de análise. De mesmo modo que em termos gerais elaboraríamos uma regressão com mais variáveis independentes aperfeiçoar a realidade e assertividade do modelo.

2.3 Elaborando uma projeção “baseline”

Um *baseline* é um valor inicial, que pode ser usado para comparar passado, presente e previsão de valores futuros. A linha de base pode ser considerada como o ponto de partida em relação a quais os valores futuros são medidos. Na modelagem, a linha de base é o ponto de referência para avaliar as variações entre tendências e previsões. As duas principais formas de elaborar tal baseline são através de media moveis e linhas de tendências exponenciais. Nesta seção mapeamos uma breve introdução a media moveis e previsões exponenciais.

As médias móveis vêm em diversas formas, mas o seu propósito subjacente é auxiliar os analistas a acompanhar as tendências de mercados e ativos financeiros por suavizar as flutuações do dia-a-dia, ou ruídos, representados pelas oscilações mais fortes. Com isso torna-se mais simples entender o comportamento de crescimento e ajudar na visualização de tendências. Aqui é apresentado uma média móvel de três meses, a equação é dada por:

$$\text{Onde: } MMS_{t+1} = \frac{\sum_{i=t+1-n}^t A_i}{n}$$

MMS_{t+1} = Média móvel simples no final do período t (pode ser usado como uma previsão para o período $t + 1$)

A_i = Procura atual no período i

n = Número de períodos incluídos em cada média

Também podemos calcular uma média móvel ponderado que é qualquer média que tem fatores de multiplicação para dar pesos diferentes para os dados em diferentes posições na janela de amostra. Matematicamente, a média móvel é a convolação dos pontos de referência, com uma função de ponderação fixo (*Hamilton, James, Time Series Analysis, 1994*). Aqui utilizamos os seguintes pesos:

período t-3 = 10%

período t-2 = 20%

Período t-1 = 70%

Onde:

$$MMP = \frac{np_M + (n-1)p_{M-1} + \dots + 2p_{(M-n+2)} + p_{(M-n+1)}}{n + (n-1) + \dots + 2 + 1}$$

O media móvel exponencial é uma extensão da média móvel simples, utilizando a suavização ponderada de observações passadas.

O ajuste exponencial é outra forma de obter a tendência de uma série temporal.

Apresenta algumas vantagens em relação às médias móveis:

- Permite realizar previsões de curto prazo (para o período seguinte da série), o que não é possível por médias móveis.
- Leva em conta todos os valores previamente observados ao período sob análise, e não somente o “mais próximo” dele como ocorre nas médias móveis⁵.

Os pesos designados para os valores observados decrescem ao longo do tempo, ou seja, o valor observado mais recentemente recebe o maior peso, o valor anterior o segundo maior e o valor observado inicialmente recebe o menor peso: isso é bom senso, imagina-se que os dados mais recentes devem ter mais influência nas previsões do que os mais antigos. A equação é dada por:

$$MME_x = \alpha \cdot x_t + (1 - \alpha) \cdot s_{t-1}$$

Onde:

a = alfa (“smoothing factor”).

xt = período

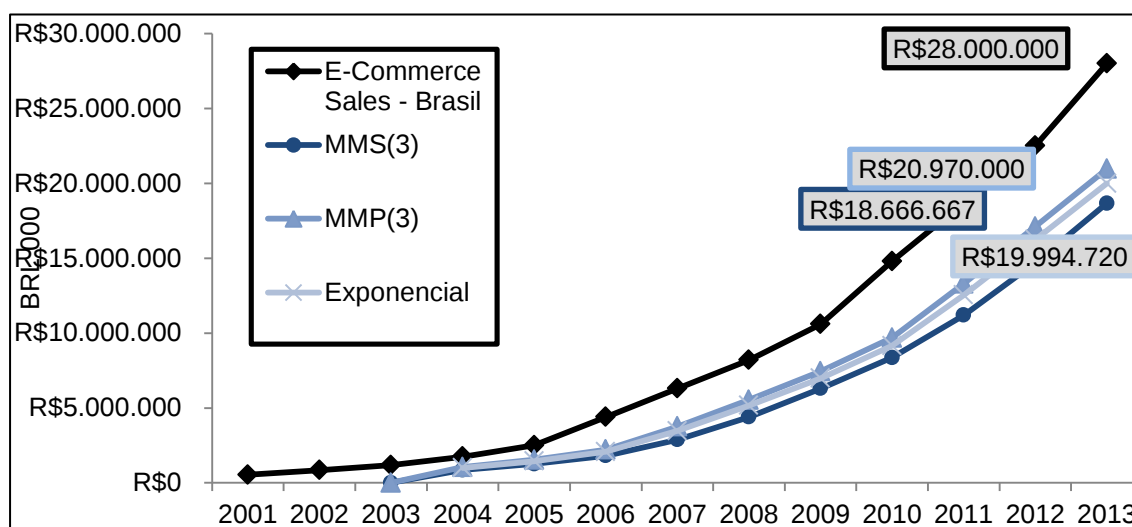
st-1 = período -1

⁵ <http://www.economicsnetwork.ac.uk>

Tabela 2 – Media moveis

Ano	Vendas de e-Commerce	MMS(3)	MMP(3)	Exponencial	Error
2001	R\$ 540.000	-	-	R\$ 540.000	-
2002	R\$ 850.000	-	-	R\$ 540.000	R\$ 310.000
2003	R\$ 1.180.000	-	-	R\$ 726.000	R\$ 454.000
2004	R\$ 1.750.000	✓ R\$ 856.667	R\$ 1.050.000	R\$ 998.400	R\$ 751.600
2005	R\$ 2.500.000	✓ R\$ 1.260.000	R\$ 1.546.000	R\$ 1.449.360	R\$ 1.050.640
2006	R\$ 4.400.000	✓ R\$ 1.810.000	R\$ 2.218.000	R\$ 2.079.744	R\$ 2.320.256
2007	R\$ 6.300.000	✓ R\$ 2.883.333	R\$ 3.755.000	R\$ 3.471.898	R\$ 2.828.102
2008	R\$ 8.200.000	✓ R\$ 4.400.000	R\$ 5.540.000	R\$ 5.168.759	R\$ 3.031.241
2009	R\$ 10.600.000	✓ R\$ 6.300.000	R\$ 7.440.000	R\$ 6.987.504	R\$ 3.612.496
2010	R\$ 14.800.000	✓ R\$ 8.366.667	R\$ 9.690.000	R\$ 9.155.001	R\$ 5.644.999
2011	R\$ 18.700.000	✓ R\$ 11.200.000	R\$ 13.300.000	R\$ 12.542.001	R\$ 6.157.999
2012	R\$ 22.500.000	✓ R\$ 14.700.000	R\$ 17.110.000	R\$ 16.236.800	R\$ 6.263.200
2013	R\$ 28.000.000	✓ R\$ 18.666.667	R\$ 20.970.000	R\$ 19.994.720	R\$ 8.005.280

Figura 3 – Mapeando um baseline



As medias moveis ajudam o analista a melhor entender a evolução dos dados, suavizar as flutuações de curto prazo e destacar tendências de longo prazo ou ciclos. Com esta perspectiva podemos traçar nossa previsão baseline através de regularização exponencial dupla atribui pesos exponenciais diminuindo à medida que a observação envelhece. Em outras palavras, as recentes observações são dadas relativamente mais peso na previsão do que as observações mais velhas. Dupla suavização exponencial é melhor na análise de tendências.

Dupla suavização exponencial - também conhecido como a suavização exponencial do Holt - é um refinamento do modelo de suavização exponencial simples, mas acrescenta outro componente que leva em conta qualquer tendência nos dados. Simples modelos de suavização

exponencial trabalhar melhor com os dados onde não há tendência ou sazonalidade componente aos dados (Shumway R. H., *Stoffer Time Series Analysis and its Applications*, 2011, 2;14). Quando os dados apresentam uma tendência crescente ou decrescente ao longo do tempo, previsões de suavização exponencial simples tendem a ficar para trás das observações. Dupla suavização exponencial é projetada para resolver este tipo de séries de dados, tendo em conta qualquer tendência nos dados.

. Nota-se que a suavização exponencial dupla ainda não aborda a sazonalidade. Para as previsões melhor exponencialmente usando dados onde não se espera ou que se sabe ser variação sazonal nos dados, use suavização exponencial tripla.

Tal como acontece com suavização exponencial simples, em modelos de suavização exponencial dupla observações passadas são dados pesos exponencialmente menores conforme as observações envelhecem. Em outras palavras, as recentes observações são dadas relativamente mais peso na previsão do que as observações mais velhas. A equação é dada por:

$$f_t = a.Y_t + (1-a) (m-1 + bt-1)$$

onde:

Y_t é o valor observado no tempo t .

f_t é a previsão no tempo t .

bt a inclinação é calculado no tempo t .

a - representando alfa - é a primeira constante de amortecimento, utilizada para suavizar as observações. Em nossa análise usamos uma alfa de 0,6 e beta de 0,4.

Para prever além de x_t a equação é dada por:

$$F_{t+m} = s_t + mb_t$$

Onde:

S_t = o dado do base do ultimo período

m = o numero de períodos para prever (por exemplo, dois para dois períodos além do ultimo período)

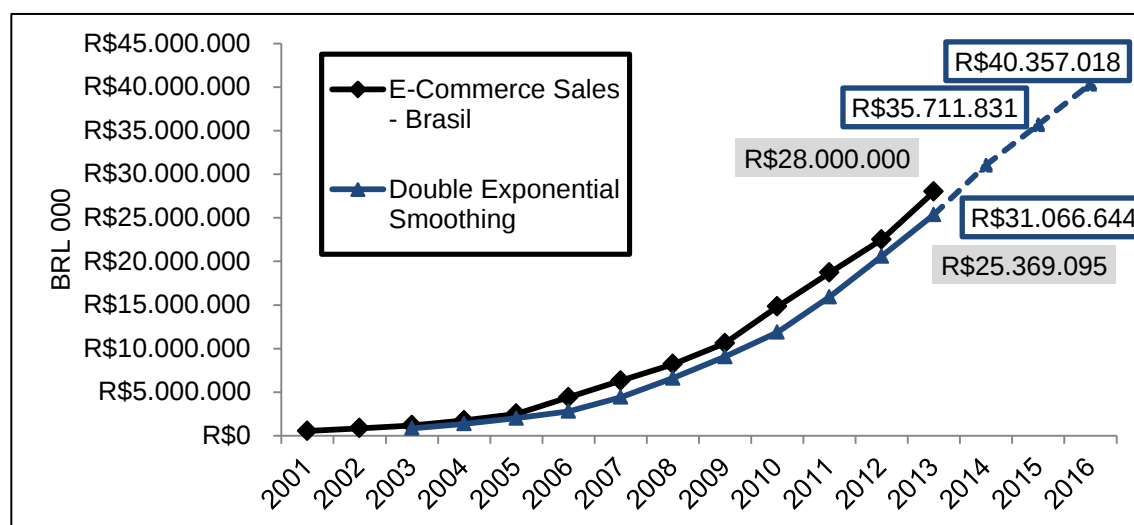
bt = o ultimo dado de tendência

Utilizando este método chegamos com nossas previsões baseline até 2016. Para 2016 a projeção é de R\$ 40.357B que seria um aumento de 44,0% sobre as vendas do setor em 2013.

Tabela 3 – Exponencial regularização dupla e previsões (“baseline”)

Ano	Vendas de e-Commerce	Exp. Dup. (Base)	Exp. Dup. (Tendência)	D.Ex (Reg.)	Error
2001	R\$ 540.000	-	-	-	-
2002	R\$ 850.000	R\$ 540.000	R\$ 310.000	-	-
2003	R\$ 1.180.000	R\$ 982.000	R\$ 389.200	R\$ 850.000	R\$ 330.000
2004	R\$ 1.750.000	R\$ 1.522.720	R\$ 480.112	R\$ 1.371.200	R\$ 378.800
2005	R\$ 2.500.000	R\$ 2.201.699	R\$ 599.432	R\$ 2.002.832	R\$ 497.168
2006	R\$ 4.400.000	R\$ 3.440.679	R\$ 983.161	R\$ 2.801.132	R\$ 1.598.868
2007	R\$ 6.300.000	R\$ 5.174.304	R\$ 1.433.439	R\$ 4.423.840	R\$ 1.876.160
2008	R\$ 8.200.000	R\$ 7.244.646	R\$ 1.815.581	R\$ 6.607.743	R\$ 1.592.257
2009	R\$ 10.600.000	R\$ 9.676.136	R\$ 2.185.126	R\$ 9.060.227	R\$ 1.539.773
2010	R\$ 14.800.000	R\$ 13.036.758	R\$ 2.890.423	R\$ 11.861.263	R\$ 2.938.737
2011	R\$ 18.700.000	R\$ 17.036.309	R\$ 3.555.900	R\$ 15.927.181	R\$ 2.772.819
2012	R\$ 22.500.000	R\$ 21.355.325	R\$ 4.013.770	R\$ 20.592.209	R\$ 1.907.791
2013	R\$ 28.000.000	R\$ 26.421.457	R\$ 4.645.187	R\$ 25.369.095	R\$ 2.630.905

2014 (e)	R\$ 31.066.644
2015 (e)	R\$ 35.711.831
2016 (e)	R\$ 40.357.018



Com esta inteligência de tendência podemos avançar para o foco principal do estudo que é análise de regressão múltipla e previsões de crescimento do setor de e-commerce.

3.0 REGRESSÃO MULTIPLA

3.1 Overview do método

A análise de regressão inclui muitas técnicas para modelagem e análise de diversas variáveis, quando o foco é sobre a relação entre uma variável dependente e uma ou mais variáveis independentes. A análise de regressão é usada para entender que entre as variáveis independentes estão relacionadas com a variável dependente, e explorar as formas destas relações. A análise de regressão pode ser usada para inferir relações causais entre as variáveis independentes e dependentes (*Shumway R. H., Stoffer (Time Series Analysis and its Applications, 2011, 7;1)*)

As variáveis que são usadas para explicar outras variáveis são chamadas de motivos (ou independente) variáveis. A variável dependente é o que você mede na previsão. A variável dependente responde a variável independente. Ela é chamada de dependente, pois "depende" da variável independente. Os modelos de regressão são utilizados para prever uma variável a partir de uma ou mais outras variáveis.

Os modelos de regressão fornecem ao cientista uma ferramenta poderosa, permitindo previsões sobre tendências econômicas, realizadas através de informações históricas.

Assim, nosso principal objetivo é determinar o melhor conjunto de parâmetros bi, de tal forma que o modelo prevê valores experimentais da variável dependente com a maior precisão possível. É preciso também avaliar se o nosso próprio modelo é adequado para encaixar os dados experimentais observados (ou seja, se nós escolhemos a forma matemática correta do mesmo) e se todos os termos do nosso modelo são significativos.

O formato geral da equação de Regressão Linear Múltipla é dado por equação:

$$Y = a + b_1X_1 + b_2X_2 + \dots + b_k X_k$$

Onde:

Y é a Variável Dependente;

a corresponde a interseção aonde y começa

b_k corresponde aos coeficientes das Variáveis Independentes; e

X_k são as variáveis independentes.

3.2 Definindo as variáveis

Neste estudo estimamos que podemos prever a atividade econômica no setor de e-commerce com duas variáveis independentes, no caso; o consumo das famílias e o número de usuários de internet no Brasil.

O consumo das famílias é a atividade “que consiste na fruição de bens e serviços pelas individuais, famílias e pelas empresas pertencentes à iniciativa privada, que como agente econômico, utilizam o rendimento que obtêm na atividade produtiva para comprar bens e serviços necessários à satisfação de suas necessidades”⁶, e no total compõe em torno de 60% da economia brasileira⁷. Assim, estamos usando este indicador como um “proxy” direto de atividade econômica com seu foco em medir a demanda específica de bens e serviços ou seja os produtos principais fornecidos pelo e-commerce. Neste caso consideramos o consumo das famílias sendo mais aplicável do que produto interno bruto que além de consumo das famílias também é uma função de investimento, consumo do governo e comércio líquido.

Nossa segunda variável independente é o número de usuários de internet no Brasil. Por definição, um comprador de e-commerce também tem que ser um usuário de internet. Assim, estamos usando este indicador como um “proxy” pelo tamanho total do mercado, a premissa sendo que quanto mais pessoas usando a internet, maior o público potencial para atividade de e-commerce.

O coeficiente de correlação, r , dá-nos uma indicação sobre a força da relação linear entre x e y . Ele pode variar de 0 (que indica ausência de correlação) a 1 (que indica correlação perfeita).

Realizamos um teste de hipótese da “significância do coeficiente de correlação” para decidir se a relação linear dos dados da amostra é forte o suficiente para usar para modelar a relação entre a população.

Com dois graus de liberdade o teste de significância para r de Pearson é calculado da seguinte forma:

$$t = \frac{r \sqrt{N - 2}}{\sqrt{1 - r^2}}$$

Tabela 4 – As variáveis dependentes e ii

⁶ <http://pt.wikipedia.org/wiki/Consumo>

⁷ <http://data.worldbank.org/indicator/NE.CON.PETC.ZS>

Correlações (Pearson)

	<i>Vendas de E-Commerce</i>	<i>Consumo das Famílias</i>	<i>Usuários de Internet no Brasil</i>
Vendas de E-Commerce	1		
Consumo das Famílias	0,9734	1	
Usuários de Internet no Brasil	0,9420	0,9833	1

Significâncias das Correlações

	<i>Vendas de E-Commerce</i>	<i>Consumo das Famílias</i>	<i>Usuários de Internet no Brasil</i>
Vendas de E-Commerce	1		
Consumo das Famílias	14,0809	1	
Usuários de Internet no Brasil	9,3095	17,8934	1

Valores P

	<i>Vendas de E-Commerce</i>	<i>Consumo das Famílias</i>	<i>Usuários de Internet no Brasil</i>
Vendas de E-Commerce	1		
Consumo das Famílias	0,0000%	1	
Usuários de Internet no Brasil	0,0002%	0,0000%	1

Podemos dizer que nossas variáveis possuem fortes correlações positivas entre si com todos os valores de r acima de 0,90%, e assim são adequados para fins de previsões de regressão. Ademais, com 95% de certeza que tais correlações são altamente significados em media dos valores p sendo bastante abaixo da alfa de 5%.

Além disso, podemos representar graficamente estas relações fortes positivos – e lineares – entre as variáveis independentes e a dependente:

Figura 4 – Gráfica de dispersão de vendas de e-commerce x consumo das famílias

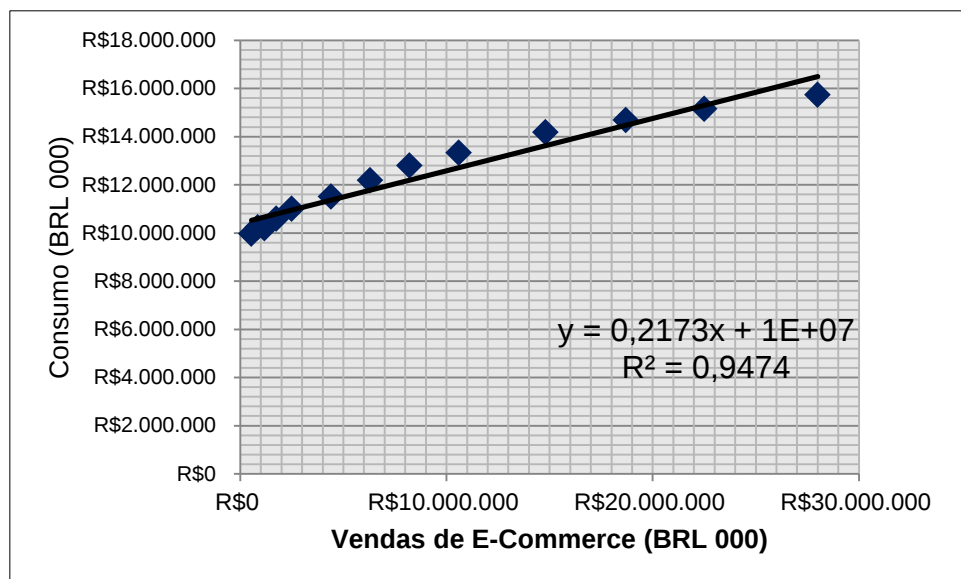
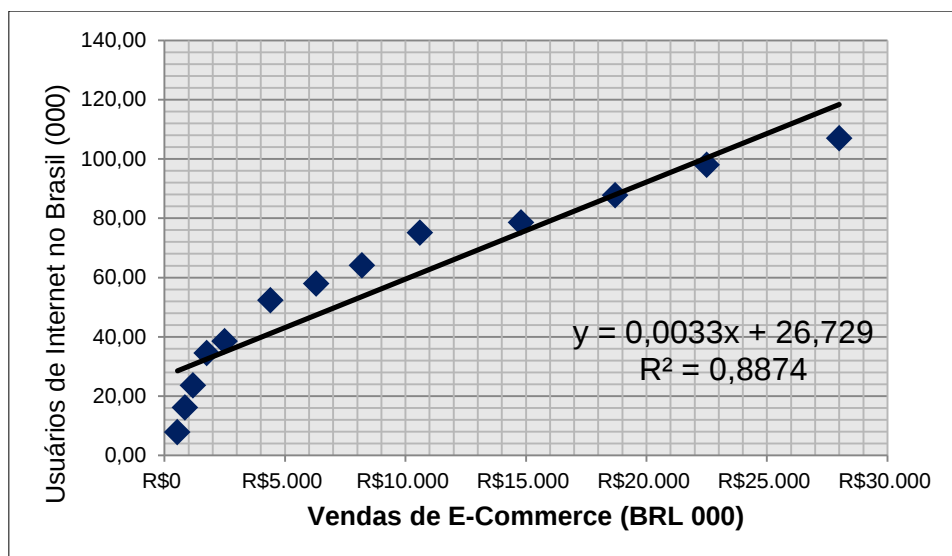


Figura 5 – Gráfica de dispersão de vendas de e-commerce x usuários de internet



3.3 Construindo o modelo

Estatisticamente, o relacionamento de nossas variáveis pode ser descrito por:

$$Y = a + b_1X_1 + b_2X_2$$

Ou simplesmente: vendas no mercado de e-commerce = f (consumo das famílias, usuários de internet).

Tabela 5 – As variáveis dependentes e independentes

Year	Vendas de E-Commerce	Consumo das Famílias	Usuários de Internet no Brasil
2001	R\$ 540.000.000	R\$ 996.411.792.347	7.870.886,81
2002	R\$ 850.000.000	R\$ 1.022.142.024.491	16.130.794,99
2003	R\$ 1.180.000.000	R\$ 1.018.907.881.381	23.607.425,89
2004	R\$ 1.750.000.000	R\$ 1.058.502.692.810	34.543.488,81
2005	R\$ 2.500.000.000	R\$ 1.100.250.536.181	38.552.190,00
2006	R\$ 4.400.000.000	R\$ 1.150.542.051.819	52.288.989,13
2007	R\$ 6.300.000.000	R\$ 1.217.872.483.715	57.943.761,28
2008	R\$ 8.200.000.000	R\$ 1.279.857.956.965	64.146.014,98
2009	R\$ 10.600.000.000	R\$ 1.332.810.319.244	75.098.703,09
2010	R\$ 14.800.000.000	R\$ 1.417.228.358.998	78.557.183,53
2011	R\$ 18.700.000.000	R\$ 1.468.466.287.781	87.719.708,25
2012	R\$ 22.500.000.000	R\$ 1.513.880.027.807	97.964.425,29
2013	R\$ 28.000.000.000	R\$ 1.572.921.348.892	106.943.392,80

Com bastante diferença de escala inerente nas variáveis, por exemplo, milhões de usuários de internet, bilhões de reais de vendas de e-commerce e trilhões de reais de consumo das famílias precisamos introduzir uniformidade à série de variáveis e homogeneizá-las

através do cálculo de suas variações percentuais, o que altera a função de relacionamento das variáveis para:

$\Delta\%$ Vendas de no mercado de e-commerce = f ($\Delta\%$ consumo das famílias, $\Delta\%$ usuários de internet).

Tabela 6 – As variáveis dependentes e independentes (variação percentual).

Year	Vendas de E-Commerce	Consumo das Famílias	Usuários de Internet no Brasil
2001	0,00%	0,00%	0,00%
2002	57,41%	104,94%	2,58%
2003	38,82%	46,35%	-0,32%
2004	48,31%	46,32%	3,89%
2005	42,86%	11,60%	3,94%
2006	76,00%	35,63%	4,57%
2007	43,18%	10,81%	5,85%
2008	30,16%	10,70%	5,09%
2009	29,27%	17,07%	4,14%
2010	39,62%	4,61%	6,33%
2011	26,35%	11,66%	3,62%
2012	20,32%	11,68%	3,09%
2013	20,32%	11,68%	3,90%

Desse modo, usando estes dados e rodando a fermento de regressão do Excel, os resultados obtidos são:

Tabela 7 – Resumo dos resultados de regressão

<i>Estatística de regressão</i>	
R múltiplo	0,79059
R-Quadrado	0,62504
R-quadrado ajustado	0,55005
Erro padrão	0,12680
Observações	13

ANOVA					
	<i>gl</i>	<i>SQ</i>	<i>MQ</i>	<i>F</i>	<i>F de significação</i>
Regressão	2	0,2680	0,1340	8,3347	0,0074
Resíduo	10	0,1608	0,0161		
Total	12	0,4288			

	<i>Coefficientes</i>	<i>Erro padrão</i>	<i>Stat t</i>	<i>valor-P</i>	<i>95% inferiores</i>	<i>95% superiores</i>
Interseção	0,0521	0,0917	0,5679	0,5827	-0,1522	0,2563
Usuários de Internet n	0,4875	0,1338	3,6424	0,0045	0,1893	0,7857
Consumo das Famílias	5,3003	1,9396	2,7327	0,0211	0,9787	9,6219

Observando no resultado da regressão, a leitura macro é que vendas de e-commerce são previstos crescer em 0,49% para cada aumento percentual de usuários de internet e 5,3% para cada aumento percentual de consumo de famílias.

3.4 Testando o modelo

Um teste de hipótese estatística é um método de inferência estatística usando dados de um estudo científico. Em estatística, o resultado é chamado de estatisticamente significativa, se tiver sido prevista como improvável que tenha ocorrido por acaso, de acordo com uma probabilidade limite pré-determinado, o nível de significância (*Gershenfeld, Neil, The Nature of Mathematical Modeling, 2000*).

Estes testes são usados para determinar o que os resultados de um estudo levariam a uma rejeição da hipótese nula para um nível pré-determinado de significância. A significância estatística dos resultados obtidos na Análise de Regressão deve ser colocada antes do uso de tais resultados numa previsão. Neste presente estudo nossa hipótese nula é que não há nenhuma relação entre as nossas variáveis independentes e a variável dependente.

O coeficiente de determinação (r^2) é uma medida para avaliar quão bem o modelo de regressão múltipla explica e prevê resultados futuros. É expressa como um valor entre 0 e 1 onde o valor de um indica um encaixamento perfeito, e por conseguinte, um modelo de regressão linear múltipla muito confiável para previsões futuras. Um valor igual a zero, no entanto, indica que o modelo de regressão múltipla falha de prever com precisão o conjunto de dados.

Com r descrito por:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

R^2 é simplesmente a correlação quadrada, ou seja elevado a poder de dois. Como orientação ampla, valores entre 0 e 0,3 indicam uma relação linear positiva fraca; valores entre 0,3 e 0,7 indicam uma relação linear positiva moderada; valores entre 0,7 e 1 indicam uma relação linear positiva forte.

Em um modelo de regressão linear múltipla, o coeficiente de determinação ajustado (ar^2) mede a proporção da variação na variável dependente é responsável pelas variáveis explicativas. Coeficiente de determinação ajustado é geralmente considerada como uma

medida mais com mais precisão do que o coeficiente de determinação. O ar^2 ajustado será sempre menor ou igual do que ao coeficiente de determinação (ar^2). No output da regressão, podemos observar que o r^2 é 0,550 que indica que 55% das variações de y podem ser explicados através das variáveis independentes em nosso modelo, que com somente 13 observações é razoavelmente adequado. Além disso, apesar de que neste estudo nosso objetivo é demonstrar um modelo linear simples, em outros estudos, o r^2 pode ajudar a entender qual modelo, por exemplo, polinomial ou logarítmica, retorna o melhor encaixamento.

O teste F ou estatística F mostra se a regressão é significativa e a relação entre vendas de e-commerce e as variáveis independentes de consumo das famílias e usuários de internet são salientes. O valor do teste F é determinado pelo quociente entre a variância explicada e a variância que pode ser demonstrado estatisticamente na seguinte forma:

$$F = \frac{\sum (\vec{y}_i - \bar{Y})^2 / (k - 1)}{\sum (y_i - \bar{Y})^2 / (n - k)}$$

No Excel podemos comparar o valor F do resultado de regressão – 8,3347 – com o valor crítico de F através da função INVF. Com as graus de liberdade de 10 e 2 respectivamente, isto retorna um valor de 4,2565 e como isto é abaixo do valor F podemos afirmar que é improvável que um valor de F essa alta ocorreu por acaso e, portanto, que a equação de regressão é útil na previsão de vendas de e-commerce.

Do mesmo modo, o significado de F indica a probabilidade de que o resultado de regressão poderia ter sido obtido por acaso. Um baixo significado de F confirma a validade do resultado de regressão. Neste caso, a leitura é que há apenas 0,74% de chance de que o output de regressão era apenas uma ocorrência de acaso.

Os testes F determinam a adequação do modelo geral, entretanto, precisamos também, analisar a significância dos coeficientes da equação de regressão individualmente. O objetivo é provar que o valor aferido a cada coeficiente é significativamente diferente de 0 ou se tal valor ocorreu simplesmente ao acaso. Assim, os testes T focam nas variâncias dos coeficientes da regressão e, via sua raiz quadrada, determinar o erro padrão, o que determina se os valores dos coeficientes são significativamente diferente de 0 (*Bowerman, B.L., and R.T.O'Connel, Linear Statistical Models : An Applied Aproach (Second Edition, 1990).*

Tabela 8 – Testes t e análise de significância

	<i>Coefficientes</i>	<i>Erro padrão</i>	<i>Stat t</i>	<i>Valor t</i>	<i>Significante?</i>	<i>valor-P</i>
Internet Users	0,48747	0,13383	3,64243	2,22814	sim	0,4519%
Household Consumo	5,30031	1,93957	2,73272	2,22814	sim	2,1095%

Se o valor absoluto de t é suficientemente elevado, pode concluir-se que o declive de coeficiente é útil na previsão de vendas e-commerce. Podemos usar a função INVT para encontrar o valor crítico de t. Com um nível de 5% de confiança e 10 graus de liberdade o valor dado é 2,2814. Ambos os valores de t das variáveis independentes sejam superiores a este valor crítico t e, como tal, podemos afirmar que eles são úteis para predizer a nossa variável y.

Além disso, os valores-p de cada coeficiente fornece a probabilidade de que eles são resultados relevantes e não ocorreu por acaso. Quanto menor o P-valor, maior a probabilidade de que esse coeficiente ou Y-Intercept é válido. No nosso caso, com o valor de p inferior a alfa, a leitura é de que existe apenas uma possibilidade menor do que 2,2% que o resultado aconteceu ao acaso.

3.5 “Regressando” o futuro

Definamos os valores futuros das variáveis x na seguinte forma:

- Nos últimos três anos, o crescimento médio dos usuários de internet foi de 12%. O assumption para ser utilizado é que esta taxa de crescimento diminuir ligeiramente a 9% nos próximos três anos em função da grande adesão de usuários nos últimos anos. Assim, a tendência seria para maturação e estabilidade em vez do crescimento exponencial visto nos últimos anos.
- Nos últimos três anos, a taxa média de crescimento de consumo das famílias foi de 3,5%. Considerando os desafios econômicos atuais no Brasil, nossa expectativa é que isso também diminuir ligeiramente a 3% nos próximos três anos.

Dado este cenário, podemos inserir estes inputs na equação de regressão:

$$\begin{aligned}
 \hat{y} &= a + b_1X_1 + b_2X_2 \\
 \hat{y} &= \text{Intercepção de } y + \Delta\% \text{ usuários de internet} * \beta_1 + \Delta\% \text{ consumo das famílias} * \beta_2 \\
 \hat{y} &= 0,0520635761033303 + (9\% * 0,487468350763315) + (3\% * 5,30030931875065) \\
 \hat{y} &= 25,49\%
 \end{aligned}$$

O resultado de regressão é que considerando tal crescimento nas variáveis independentes como traçado acima o crescimento nas vendas de e-commerce é estimado a ser 25,49% ao ano em 2014 a 2016.

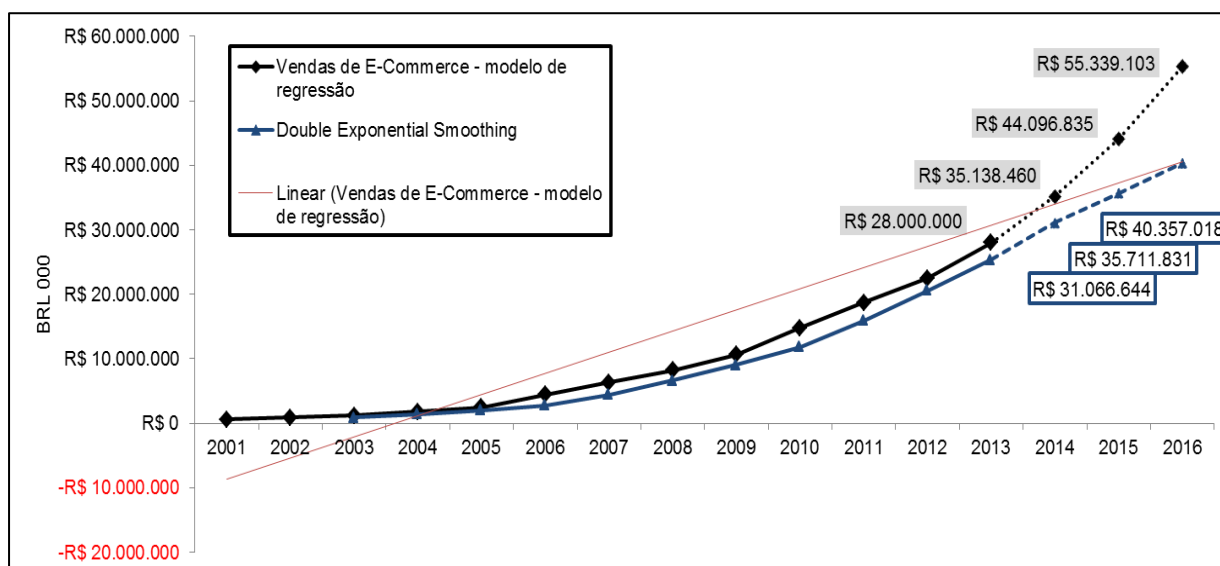
Utilizando esta taxa de crescimento estimada chegamos a um valor de R\$ 55.339B em 2016; um aumento total de 97,64% de 2013 que continua a tendência positiva desde o início das observações em 2001.

A projeção exponencial dupla é mais conservadora do que o modelo de regressão, ou seja, há um aumento mais lento com a linha de regressão do que com o dobro de alisamento. Neste estudo o nosso modelo preferido é o modelo de regressão que o autor espera que melhor represente as amplas perspectivas de crescimento do mercado de e-commerce no Brasil nos próximos anos.

Tabela 9 – Estimativas para crescimento futuro do mercado de e-commerce

Year		Vendas de E-Commerce (BRL 000)	Vendas de E-Commerce ($\Delta\%$)	Consumo das Famílias	Usuários de Internet no Brasil
2014 (e)	R\$	35.138.460.202,57	25,49%	9,00%	3,00%
2015 (e)	R\$	44.096.835.193,12	25,49%	9,00%	3,00%
2016 (e)	R\$	55.339.103.160,45	25,49%	9,00%	3,00%

Figura 6 – Gráfica de previsão do modelo de regressão x “base” exponencial



4 CONCLUSÕES

4.1 Considerações finais

Este estudo teve como objetivo macro apresentar uma metodologia de modelagem de series temporal e regressão para prever o crescimento no mercado de e-commerce no Brasil de maneira compreensiva, eficaz e no Excel para atender as necessidades de inteligência de mercado para um Profissional sem acesso ao software estatístico avançado. Tal um “workflow” foi mapeado começando um snapshot atual do mercado, análise de tendências históricas, elaboração de um “baseline” através de regularização exponencial e finalmente uma previsão de regressão do crescimento do mercado até 2016. Além disso, uma validação estatística básica do modelo foi mostrada. Na introdução, o autor chamou esta abordagem a “democratização” de modelagem financeira e isso foi o “guiding principle” utilizado ao longo do estudo.

4.2 Próximos passos

Os Pressupostos da modelagem de regressão são:

- Estacionariedade das variáveis independentes
- Linearidade
- Independência dos resíduos
- Homocedasticidade
- Normalidade dos resíduos

Como o foco principal do estudo foi para mostrar de forma *straightforward* a elaboração de um modelo de regressão para facilitar a modelagem econômica, os testes para obter tais certificações estatísticas não entraram em nosso escopo. Todavia, um estudo futuro, ou um segundo estudo, poderia incluir tais análises como o próximo passo.

REFERÊNCIAS BIBLIOGRÁFICAS

- I. Bowerman, B.L., and R.T.O'Connel, Linear Statistical Models : An Applied Aproach (Second Edition), Boston : PWS-Kent, 1990.
- II. Box, George; Jenkins, Gwilym (1976), Time Series Analysis: forecasting and control, rev. ed., Oakland, California: Holden-Day
- III. Cowpertwait P.S.P., Metcalfe A.V. (2009), Introductory Time Series with R, Springer.
- IV. Draper, N.R., & H.Smith, Applied Regression Analysis, (Second Edition), New York : John Wiley & Sons, 1981.
- V. Durbin J., Koopman S.J. (2001), Time Series Analysis by State Space Methods, Oxford University Press.
- VI. Gershenfeld, Neil (2000), The Nature of Mathematical Modeling, Cambridge University Press, ISBN 978-0-521-57095-4, OCLC 174825352
- VII. Goldberger, A.S., Econometric Theory, New York : John Wiley & Sons, 1964.
- VIII. Hamilton, James (1994), Time Series Analysis, Princeton University Press, ISBN 0-691-04289-6
- IX. Makidrakis, S. and Wheelright S.C., Forecasting Methods and Applications, New York : John Wiley & Sons, 1978.
- X. Priestley, M. B. (1981), Spectral Analysis and Time Series, Academic Press. ISBN 978-0-12-564901-8
- XI. Shasha, D. (2004), High Performance Discovery in Time Series, Springer, ISBN 0-387-00857-8
- XII. Shumway R. H., Stoffer (2011), Time Series Analysis and its Applications, Springer.
- XIII. Wiener, N. (1949), Extrapolation, Interpolation, and Smoothing of Stationary Time Series, MIT Press.
- XIV. Woodward, W. A., Gray, H. L. & Elliott, A. C. (2012), Applied Time Series Analysis, CRC Press.