UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE ENGENHARIA DE COMPUTAÇÃO

KAZUKI MONTEIRO YOKOYAMA

Estudo Empírico Sobre a Lei de Metcalfe e o Efeito de Rede

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Engenharia da Computação

Orientador: Prof. Dr. Raul Fernando Weber
Co-orientador: Prof. Dr. Márcio Valk

Porto Alegre
2016

“They didn’t know it was impossible
So they did it.”
— Mark Twain

AGRADECIMENTOS

Agradeço primeiramente à minha mãe, Eliana, e ao meu irmão, Fábio, que sempre me apoiaram e incentivaram incondicionalmente. Sem vocês nada disso teria sido possível. Vocês são minha família.

Ao meu pai, Yoshitaka, que deixou muitas saudades. Muito obrigado pelo exemplo e pela formação como pessoa.

À minha namorada, Aline, por tudo. O seu suporte, compreensão, amor e carinho foram o meu combustível durante essa jornada.

Agradeço também às pessoas maravilhosas que tive o prazer de encontrar em Porto Alegre. Em especial, muito obrigado por toda força e incentivo, Srª Rainilde Back, Srº Afonso Sehnem e Srª Ivone. Muito obrigado aos meus amigos de turma pela parceria, confiança e bons e inesquecíveis momentos. Obrigado ao pessoal do Mconf pela paciência e compreensão.

Agradeço enormemente aos meus orientadores, Raul Fernando Weber e Márcio Valk, pela paciência, dedicação e atenção durante todo processo. Sem seus conselhos, este trabalho teria sido muito mais difícil.

Por fim, agradeço também a todos aqueles que, de uma forma ou de outra, fizeram parte da minha formação e, estando perto ou longe, sempre me deserajam o melhor e apoiaram em todos os momentos.

RESUMO

O efeito de rede, fenômeno que ocorre quando um produto ou serviço torna-se mais atraente quanto mais pessoas o utilizam, já foi bem discutido na literatura econômica. Suas implicações já mostraram ter grande influência no rumo da tecnologia e possivelmente continuarão tendo com o rápido avanço das telecomunicações digitais. Diversos modelos teóricos foram propostos para explicar o efeito de rede e tentar quantificá-lo. A lei de Metcalfe, dentre outras, fornece uma luz sobre a forma como o valor de uma rede varia de acordo com o seu tamanho. Apesar da sua popularidade, a lei de Metcalfe foi pouco testada com dados reais e sua validade ainda é incerta. Este trabalho propõe uma análise empírica de alguns modelos de valor de redes, incluindo a lei de Metcalfe, utilizando dados de uma rede social *online*. O foco da investigação são os procedimentos econométricos necessários para a correta análise dos dados. Para isso, é feito uso de técnicas estatísticas como regressão e séries temporais na avaliação de cada modelo.

Palavras-chave: Efeito de rede. lei de Metcalfe. redes sociais. econometria. séries temporais.

Empirical Study On Metcalfe's Law and the Network Effect

ABSTRACT

The network effect, phenomenon where a product or service becomes more attractive as the number of people using it increases, has been well discussed in economic literature. Its implications have been shown to have great influence in technological change and possibly will continue to have given the rapid progress of digital telecommunications. Several theoretical models have been proposed to explain the network effect and try to quantify it. Metcalfe's Law, among others, gives us an explanation on how the value of a network varies according to its size. Despite its popularity, Metcalfe's Law has been little tested against actual data and its validity remains unknown. This work proposes an empirical analysis of some models related to network value, including Metcalfe's Law itself, using data from an online social network. The focus of the research are the econometric procedures needed to correctly analyze the data. For this, statistical techniques such as regression and time series are employed to evaluate the models.

Keywords: network effect, Metcalfe's law, social networks, econometrics, time series.

LISTA DE ABREVIATURAS E SIGLAS

ADF	*Augmented Dickey-Fuller*
AIC	*Akaike Information Criterion*
FAC	Função de Autocorrelação
BLUE	*Best Linear Unbiased Estimators*
KPSS	Kwiatkowski–Phillips–Schmidt–Shin
MQO	Mínimos Quadrados Ordinários
MRLC	Modelo de Regressão Linear Clássico
MRLCN	Modelo de Regressão Linear Clássico Normal
UAM	Usuários Ativos Mensalmente

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Ao longo da história da informática, muitas "leis" anedóticas foram concebidas e algumas sobreviveram ao teste do tempo perpetuando-se na cultura da computação. Hoje, tem-se como exemplo a bem conhecida lei de Moore que indica que a quantidade de transistores em circuitos integrados densos dobra a cada dois anos aproximadamente. Ainda que Gordon Moore não tivesse usado a palavra "lei" uma vez sequer em seu artigo original, sua afirmação persiste como uma referência na evolução da tecnologia e iconiza bem o papel que essas relações possuem.

Em comum, essas proposições possuem um caráter empírico, observacional pouco científico, e não devem ser confundidas com, por exemplo, as leis físicas de validade universal. Elas baseiam-se, portanto, apenas em observações e muitas vezes são revistas e reformuladas para se adequarem às mudanças na realidade. Como tratam-se de relações empíricas, podem carecer de uma base ou fundamento teórico que suporte-as, abrindo caminho para investigações mais aprofundadas sobre natureza de tais relações. Curiosamente, muitas dessas relações são leis epônimas, ou seja, levam o nome de seus criadores.

Diversas proposições acerca do universo tecnológico já foram feitas. A performance de sistemas computacionais encontra lugar na lei de Amdahl, enquanto a velocidade das conexões para usuários finais é abordada na lei de Nielsen. Como quer que seja, cada um desses enunciados teve sua parcela de influência, grande ou pequena, no que é a tecnologia hoje e possivelmente no que será ainda dentro dos próximos anos.

Uma observação de grande influência no curso da tecnologia deve-se a Robert Metcalfe. Sua origem remonta ao surgimento das primeiras placas *Ethernet* há quatro décadas, mas que só tornou-se popular com uma obra de George Gilder (GILDER, 1993). A relação básica é que o valor de uma rede é proporcional ao número de nodos que a compõem. Muita discussão tem sido levantada ao longo dos anos sobre a validade dessa relação e outros modelos já foram propostos abordando a mesma questão (ODLYZKO; TILLY, 2005), (REED, 1999), (BECKSTROM, 2009), (STEIN, 2009).

Dois dos principais pontos em comum a todos os trabalhos sobre esse tema são o efeito de rede e a determinação do valor de uma rede. O primeiro explica a importância dos modelos ao mesmo tempo em que fornece fundamentação do ponto de vista da teoria econômica. A definição de valor no contexto dos modelos é importante para dar significado mais claro ao que está tentando-se modelar afinal de contas.

Há décadas estamos cercados por redes de comunicações por todos os lados. Desde o

surgimento da telefonia até o presente momento, as redes foram as responsáveis por conectar milhões de pessoas ao redor de todo mundo e, a cada dia que passa, mais e mais pessoas fazem parte de alguma rede. A Internet, a maior infraestrutura de informação global, serve como plataforma para milhares de redes de diferentes naturezas. Hoje é dificíl encontrar alguém que não esteja conectado à Web ou não faça parte de alguma comunidade digital como Facebook ou LinkedIn.

No entanto, nem sempre os usuários dessas redes questionam-se do porquê fazerem parte delas. Em grande parte, isso deve-se ao fato de seus famialiares, amigos e colegas de trabalho também serem usuários de tais redes. Dificilmente alguém continuaria a participar dos mesmos grupos se todos aqueles que ele julga importantes deixassem de participar. De forma geral, quanto mais conexões que julgamos importantes fizerem parte de uma rede, mais propensos estaremos de juntarmo-nos a ela. Essas conexões não precisam ser necessariamente diretas ou interpessoais. Por exemplo, sob o ponto de vista da Web, essas conexões podem ser as diversas páginas que exibem conteúdo gerado por outras pessoas. Quanto maior o número de páginas, logo de conteúdo, disponível na Web, mais importante será o acesso a ela. Essa dependência do valor de um produto ou serviço em relação ao número de usuários já possuindo ou utilizando tal serviço é chamado *efeito de rede*.

A literatura sobre o efeito de rede do ponto de vista da teoria econômica é vasta. (KATZ; SHAPIRO, 1985) analisaram os mercados com efeito de rede durante a grande evolução das indústrias de atelecomunicações e da popularização dos computadores pessoais e ainda hoje são referência no assunto. (ROHLFS, 1974), então pesquisador da Bell Labs, investigou o lado da demanda por serviços de telefonia fortemente influenciados pelo efeito de rede. (CHIU; NG, 2011) dão uma visão geral sobre a economia das redes e discutem o crescimento da Internet do ponto de vista econômico.

O termo “valor de uma rede” é extremamente vago por si só. Antes dos modelos serem apresentados, é preciso definir bem o que deseja-se dizer com “valor de uma rede”. A princípio, esse termo pode denotar: (1) o valor agregado de uma rede ou (2) o valor percebido por cada usuário nessa rede. Objetivamente, este trabalho preocupa-se com a primeira interpretação do termo. É esse tipo de valor que é equacionado em cada um dos modelos que serão apresentados e também é o mais simples de ser quantificado. A interpretação do valor individual pode ser traduzida como a utilidade percebida por cada usuário. Apesar de importante, não apenas por si só, mas também para as teorias sobre o valor total das redes, essa visão do valor não será tratada com maiores detalhes aqui. (SWANN, 2002) aborda a questão da forma funcional da utilidade em relação ao tamanho da rede. (FEIJÓO; GÓMEZ-BARROSO; VOIGT, 2014) discutem o

problema da quantificação do valor percebido pelos usuários a partir dos relatórios financeiros de empresas do setor tecnológico.

A proposta deste trabalho é analisar quantitativamente os modelos de valor agregado de redes utilizando as técnicas estatísticas apropriadas. Serão tratados quatro modelos diferentes, dentre eles o de Metcalfe, e tentará se verificar se é possível chegar à uma conclusão a partir dos dados disponíveis.

O restante do trabalho está organizado como segue. O Capítulo 2 descreve os quatro modelos sobre o valor das redes e a história e teoria por trás de suas equações que servem como fundamento teórico para a investigação. A teoria econométrica e as principais técnicas estatísticas utilizadas neste trabalho são brevemente apresentadas no Capítulo 3. O capítulo 4 discute os trabalhos anteriores principalmente os de cunho quantitativo e tenta identificar melhorias no procedimento de análise. O capítulo 5 exibe os dados obtidos e descreve os experimentos realizados. Por fim, o Capítulo 6 encerra o trabalho com as conclusões.

2 MODELOS

Diversos modelos acerca do valor agregado de redes foram propostos e a diferença entre um e outro é a taxa de crescimento do valor em relação ao seu número de nodos. Enquanto Metcalfe considera um crescimento quadrático do valor, outros como Sarnoff, sugerem um crescimento linear, mais modesto. Recentemente, foi proposto um modelo logarítmico que reside entre o linear e o quadrático (BRISCOE; ODLYZKO; TILLY, 2006). Levando em consideração a capacidade de formar subgrupos em uma rede, um modelo de variação exponencial surge. Esses modelos são conhecidos na literatura como lei de Metcalfe, lei de Sarnoff, lei de Odlyzko-Briscoe e lei de Reed. Uma maior atenção será dada à lei de Metcalfe a seguir por ser o modelo em maior evidência e o objetivo inicial deste trabalho.

2.1 A lei de Metcalfe

Em 1973, a Ethernet surgiu em um memorando escrito por Robert Metcalfe que circulou pelos laboratórios da Xerox Palo Alto Research (PARC) mostrando como uma rede LAN (*Local Area Network*) poderia funcionar (METCALFE, 2013). Anos depois, Bob Metcalfe - cofundador e então vice-presidente de vendas e marketing da 3Com - postulou que o valor de uma rede é proporcional ao quadrado do seu número de nodos. Além disso, observou que o custo dessa rede é linearmente proporcional ao seu número de nodos. Isso sendo verdade, deveria existir um número de nodos tal que o valor da rede igualaria-se ao seu custo e, desse momento em diante, o valor da rede seria superior ao seu custo. Essa quantidade de nodos iria compor a chamada massa crítica.

Inicialmente, suas afirmações serviram com o propósito comercial de convencer investidores a comprarem placas de comunicação Ethernet além da massa crítica, que acreditava-se ser em torno de 30 nodos, o que acabou funcionando. Munida de um *slide* de 35mm que representa graficamente o que viria a ser chamada lei de Metcalfe, a força tarefa da 3Com fez com que as vendas passassem de algumas centenas de placas vendidas ao mês para milhares (METCALFE, 2013).

Mesmo depois do sucesso da Ethernet, sua mensagem sobre o crescimento do valor de uma rede continuou causando influência no ramo. Reed Hundt, na época presidente da Comissão Federal de Comunicações dos Estados Unidos, afirmou que a lei de Metcalfe e a lei de Moore dão-nos a melhor base para entendermos a Internet (KAPROWSKI, 1996 apud ODLYZKO; TILLY, 2005). Na mesma linha de pensamento, Marc Andressen, coautor do

primeiro grande navegador da Web - Mosaic - e cofundador do Netscape, atribuiu o rápido desenvolvimento da Web, por exempo com o crescimento da base de assinantes da AOL, à lei de Metcalfe (BRISCOE; ODLYZKO; TILLY, 2006).

O princípio por trás do pensamento de Metcalfe é de que, numa rede conectada, o valor percebido por cada um dos n nodos é proporcional à quantidade de outros nodos, ou seja, $n-1$. Sendo assim, o valor total percebido na rede seria $n(n-1)$ ou, aproximadamente, n^2 (METCALFE, 2013). Visto de outro forma, em uma rede com n nodos, a quantidade total de conexões possíveis entre quaisquer dois nodos é $n(n-1)/2$ que tem ordem n^2, assintoticamente (BRISCOE; ODLYZKO; TILLY, 2006).

Uma hipótese implícita do modelo é que cada novo nodo introduz o mesmo valor. Essa hipótese é muito forte e já foi apontada como a principal falha do modelo antes (ODLYZKO; TILLY, 2005), (BRISCOE; ODLYZKO; TILLY, 2006). É plausível que um dado usuário da rede não valorize de forma igual todas suas conexões, tão menos todas as conexões possíveis na rede. Ainda é possível que novos usuários adicionados à rede diminuam seu valor total se esses usuários introduzirem ruído na forma de *spams*, virus e outras ações indesejadas.

Segundo (BRISCOE; ODLYZKO; TILLY, 2006), se a lei de Metcalfe fosse válida, as empresas teriam incentivos muito fortes para interconectarem-se independente do tamanho de suas redes. No entanto na prática, esse é um cenário incomum e, em geral, apenas redes com tamanhos similares tendem a fundirem-se. No modelo de Metcalfe, ambas as redes, independente de seus tamanhos, teriam o mesmo ganho com a interconexão, o que contraria o que é observado. Olhando com mais detalhes para o processo de interconexão, (HOVE, 2014) afirma que é possível que redes de tamanhos muito diferentes resistam a uma fusão mesmo se a lei de Metcalfe for válida. Isso é possível se for analisado não só o valor agregado das redes pós-fusão, mas também a utilidade derivada por cada usuário após o evento.

Outra suposição feita pelo modelo de Metcalfe é de que existem conexões entre os nodos da rede. Se não houver tais conexões, um usuário não pode perceber valor devido a presença de outro usuário já que ambos não podem comunicar-se.

$$\text{Lei de Metcalfe: } V \propto N^2,$$

onde V é o valor da rede e N é a sua quantidade de nodos. Até o restante do trabalho, essa notação será seguida.

Como exemplo, considere uma dada rede que possuindo 100 usuários é avaliada em \$1.000,00. Se seu valor tiver o crescimento como descrito pela lei de Metcalfe, ao alcançar 110 usuários, seu valor será de $\$1.000,00(110^2/100^2) = \$1.210,00$. Cabe obsevar aqui que essas

relações não fazem distinção da unidade monetária adotada e o símbolo $ será utilizado daqui em diante por mera convenção.

2.2 A lei de Sarnoff

A lei de Sarnoff apresenta a menor taxa de crescimento dentre as avaliadas neste trabalho. Ela sugere que o valor da rede é diretamente proporcional à sua quantidade de nodos. Isso é particularmente verdade para redes de *broadcast* onde há somente um nodo emissor e os emais são apenas receptores.

Nesse tipo de rede faz sentido pensar que para cada novo nodo receptor adicionado à rede, o valor total aumente na mesma proporção. Porém em redes onde a comunicação é bilateral, como aquelas que este trabalho tratará, é razoável imaginar que cada nodo perceba valor também de sua comunicação com os demais.

Esse modelo de crescimento foi primeiro proposto por David Sarnoff, considerado o Pai da Televisão Americana (METCALFE, 2013).

$$\text{Lei de Sarnoff: } V \propto N.$$

Novamente considerando a rede com 100 nodos e mensurada em $1.000,00, dessa vez, ao chegar aos 110 nodos, seu valor será de $\$1.000,00(110/100) = \$1.100,00$, abaixo daquele encontrado no caso da lei de Metcalfe como esperado.

2.3 A lei de Reed

A lei de Reed, como ficou conhecida, é fruto de algumas observações sobre a formação de grupos em redes por David Reed. Em seu artigo "*That Sneaky Exponential - Beyond Metcalfe's Law to the Power of Community Building*", Reed argumenta que muito valor pode ser obtida a partir da formação de grupos nas redes, se esta suportar tal formação. Exemplos de grupos são listas de emails, salas de conversa online e grupos de discussão (REED, 1999). As redes que suportam pgrupos dessa forma foram chamadas de GFN.[1]

Em termos de valor da rede, a lei afirma que GFNs podem ter crescimento exponencial com o número de nodos. A ideia por trás dessa afirmação é de que, numa rede com N nodos, a quantidade total de subconjuntos que pode-se formar é dada por:

[1] Do inglês, *Group Forming Networks*.

$$C_N^0 + C_N^1 + C_N^2 + C_N^3 + \cdots + C_N^N = 2^N.$$

Retirando-se os subconjuntos triviais, ou seja, aqueles sem participantes e aqueles com apenas um participante representados respectivamente por C_N^0 (= 1) e C_N^1 (= N), restam $2^N - N - 1$ subgrupos não triviais, que tem crescimento exponencial assintoticamente.

$$\text{Lei de Reed: } V \propto N^2.$$

Esse modelo parece superestimar a criação de valor mesmo em redes onde usuários são capazes de formar grupos. Com um crescimento exponencial (no sentido matemático), em algum momento a inclusão de apenas um novo nodo dobraria o valor agregado da rede e esta passaria a valer mais do que toda economia global junta (ODLYZKO; TILLY, 2005).

Como foi feito anteriormente para a lei de Metcalfe e de Sarnoff, considere a rede de 100 nodos, valor $1.000,00 e que ahora ela permite que seus participantes formem grupos. Seu valor, ao possuir 110 usuários, chegará a $\$1.000,00(2^{110}/2^{100}) = \$1.024.000,00$, o que é um aumento bastante expressivo no valor considerando uma adição de apenas 10 novos nodos.

2.4 A lei de Odlyzko-Tilly

O modelo apresentado a seguir será chamado aqui de lei de Odlyzko-Tilly - ainda que não haja consenso sobre o nome - por ser uma contribuição dos pesquisadores Andrew Odlyzko e Benjamin Tilly (ODLYZKO; TILLY, 2005). Essa lei difundiu-se publicamente através de um artigo publicado no periódico IEEE Spectrum que possui contribuição de Bob Briscoe, além dos dois autores anterior (BRISCOE; ODLYZKO; TILLY, 2006).

A lei baseia-se na ideia da "longa cauda"[2] descrita quantitativamente pela lei de Zipf. A lei de Zipf é outra observação empírica capaz de explicar uma vasta gama de fenômenos muito bem. Ela diz que elementos de uma dada coleção, uma vez ordenados por algum critério, apresentam valores sempre decrescentes proporcionais a $1/n$ sendo n a posição do elemento na ordenação (BRISCOE; ODLYZKO; TILLY, 2006).

Para cada um dos N diferentes nodos da rede, o valor de cada um dos $N - 1$ nodos restantes seguiria a lei de Zipf. Explicando melhor, considere um nodo dessa rede, n_1. Ordenando de forma decrescente os outros $N - 1$ nodos pelo valor de sua conectividade com n_1, teria-se que o primeiro nodo ordenado teria valor proporcional a 1, o segundo, $1/2$, o terceiro $1/3$ e

[2] *Long tail*, do inglês.

assim por diante até o $n-1$-ésimo nodo que contribuiria com valor proporcional a $1/(n-1)$. Dessa forma, o valor total percebido por n_1, V_{n_1}, seria

$$V_{n_1} = \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{N-1}.$$

Sabendo que

$$\lim_{N\to\infty}(\sum_{k=1}^{N} 1/k - ln(N)) = \gamma,$$

onde γ é a constante de Euler-Mascheroni (≈ 0.577215) e $ln(N)$ denota o logarítmo natural de N, pode-se considerar que $V_{n_1} \to ln(N)$ para N suficientemente grande. Como temos N nodos na rede, repetiria-se o mesmo processo para os demais $(n_2, n_3, \ldots, n_N)$ encontrando os valores percebidos por eles $(V_{n_2}, V_{n_3}, \ldots, V_{n_N})$ e portanto o valor total seria $Nln(N)$.

Lei de Odlyzko-Tilly: $V \propto Nln(N)$.

Esse modelo seria capaz de explicar a hesitação de redes maiores fundirem-se com redes menores, fenômeno observado na prática. Sendo o crescimento logarítmico, o ganho de uma rede maior devido a interconexão com uma rede menor seria menor que o ganho da rede menor. Isso justificaria a resistência de empresas maiores de unirem-se à empresas menores e, em alguns casos, aceitando a fusão mediante um pagamento. Além disso, o modelo prevê que, se ambas redes tiverem aproximadamente o mesmo tamanho, então o ganho para ambas é aproximadamente o mesmo. Isso traduz-se em um incentivo para tais redes interconectarem-se como observado no mundo real dos negócios (ODLYZKO; TILLY, 2005). Esse foi um dos principais argumentos contra a lei de Metcalfe e motivo para a proposta do modelo de Odlyzko-Tilly. Porém (HOVE, 2014) mostrou que esse argumento é inválido se forem levadas em consideração o ganho na utilidade percebida pelos usuários das redes e as mudanças nas posições estratégicas das empresas.

Voltando ao exemplo numérico da rede com 100 nodos e $\$1.000,00$, tem-se que o valor dessa rede, se a lei de Odlyzkko-Tilly valer, será de $\$1.000,00[(110ln(100))/(100ln(100))] = \$1.122,77$ quando for composta por 110 nodos.

3 MODELAGEM ECONOMÉTRICA

Frequentemente deparamos-nos com dados de diversas naturezas que precisam ser analisados de alguma forma. Isto é particularmente verdade na Computação onde o volume e formatos em que os dados mostram-se disponíveis é muito variada. No entanto, a análise exige grande cuidado já que quase sempre é possível pressionar os dados e fazê-los contar a história que queremos, não a história que deveriam.

Neste trabalho, são apresentados alguns modelos referentes ao valor de redes relativos aos seus tamanhos. Além da análise teórica subjacente a cada modelo, é interessante testá-los frente a dados reais e, se possível, eleger aquele modelo que melhor explica o que é observado. Apesar dos modelos serem de interesse para a Computação e, com alguma extensão, ao mundo da Administração e Economia, o conjunto de técnicas a ser empregada é de domínio da Estatística. Mais especificamente, trataremos das técnicas estatísticas utilizadas extensivamente na Econometria, área que explora os modelos econômicos através de dados do mundo real.

As aplicações das técnicas que serão mostradas estendem-se para muito além do domínio sob estudo. Elas podem ser úteis em uma gama de situações interessantes para a Computação. Por exemplo, (SINHA; RAZ; CHOUDHURI, 2006) modelaram o *round-trip time* - RTT para pacotes na rede através de variáveis como distância geográfica entre os nodos e número de roteadores utilizando técnicas de regressão. (POTOK; VOUK; RINDOS, 1999) tentaram predizer, através da análise de regressão, o impacto do uso de programação orientada a objetos (OOP) na produtividade de programadores. Assim, é importante que seja dedicado algum espaço à apresentação dessas técnicas sob o ponto de vista estatístico.

Faz-se a observação de que, no entanto, o tratamento dado a seguir não é rigoroso e serve apenas como uma introdução (ou recapitulação) ao assunto. Os resultados são apenas apresentados, mas não provados. As provas e maiores detalhes e formalismos a respeito do tema fogem ao escopo do trabalho, mas podem ser facilmente encontrados na literatura.

3.1 A Econometria

A Econometria é um campo de estudo que utiliza técnicas estatísticas no contexto da Economia. Mais especificamente, a Econometria baseia-se nos desenvolvimentos de métodos estatísticos para estimação das relações econômicas, teste de teorias econômicas, validação e implementação de políticas governamentais e de negócios (WOOLDRIDGE, 2015). Muitos problemas de interesse econômico necessitam dessas técnicas como ferramental na tentativa

de solucioná-los. As questões mais frequentes são aquelas que tentam extrair relações entre diferentes variáveis utilizando algumas dessas variáveis para explicar as demais.

É muito importante, no entanto, que tenha-se um modelo econômico teórico por trás da investigação. Os dados por si só podem revelar quase qualquer tipo de relação mesmo que tal relação não exista de fato ou ainda que não faça nenhum sentido econômico. Sob essa visão, a Econometria tem o papel de prover as técnicas e métodos para avaliar quantitativamente os modelos econômicos na busca de evidências que os sustentem, mas não de fornecer tal modelo teórico. Uma vez que se disponha de um ou mais modelos (por exemplo, modelos concorrentes para uma mesma situação), é possível dar início à análise numérica dos dados observados que encontram-se disponíveis. Importante notar que, em geral, encontram-se disponíveis dados observacionais e não experimentais.

O fenômeno econômico por trás da investigação pode ser de várias naturezas e de interesse para distintos grupos, às vezes não tão óbvios pela aparente distância do universo econômico como conhecido. Ao mesmo tempo, outras áreas do conhecimento que, à primeira vista não possuem relação com os problemas econômicos, podem beneficiar-se das técnicas de modelagem econômica e econométrica na resolução de suas próprias questões. É um pouco restrito imaginar que todos os métodos desenvolvidos pela Econometria sejam tão somente aplicáveis aos problemas de natureza puramente econômica.

3.1.1 Metodologia Econométrica

A metodologia de análise utilizada neste trabalho segue a abordagem econométrica tradicional/clássica. Ela consiste de algumas etapas que subdividem uma possível tarefa complexa em passos mais simples. Basicamente, esses passos são (GUJARATI, 2009):

- Enunciado da teoria ou hipóteses.
- Especificação matemática do modelo que representa a teoria.
- Especificação do modelo estatístico ou econométrico.
- Obtenção dos dados.
- Estimação dos parâmetros do modelo econométrico.
- Validação do modelo.
- Previsão e predição.
- Utilização do modelo para propósitos de políticas ou controle.

É importante entender como dá-se o processo de análise empírica para que os resultados e conclusões de toda investigação tenham significado válido. A seguir é detalhada um pouco mais cada uma das etapas descritas anteriormente.

A declaração da teoria ou das hipóteses sob estudo é uma das partes mais importantes de toda análise. Ela é basicamente a formulação do problema a ser estudado sendo portanto a base e, ao mesmo tempo, o objeto sob investigação. Durante essa etapa, a teoria investigada é discutida possivelmente em termos qualitativos e também são fornecidos os embasamentos teóricos necessários que possam sustentá-la. Como dito anteriormente, sem uma fundamentação teórica, corre-se o risco de se obter resultados que na realidade não fazem sentido ainda que os dados corroborem o contrário.

Como a formulação do problema ou da teoria pode ser de natureza basicamente qualitativa, é necessário traduzí-la para a linguagem matemática. Durante a modelagem matemática, a teoria é descrita em termos de uma equação ou um conjunto de equações. O resultado é um modelo que, embora simplifique de forma considerável a realidade, ainda possua a capacidade de explicar seus aspectos mais importantes. É importante que os objetos resultantes desta etapa (variáveis, equações entre outros) possuam um significado no mundo real.

Ainda que a formulação matemática do modelo seja um bom modo de descrever a teoria, ela ainda é uma especificação determística, exata do problema. Muitas vezes, porém, a realidade conta com altos graus de inexatidão e isso deve, em algum momento, ser considerado. A especificação econométrica ou estatística do problema baseia-se na sua formulação matemática e tenta ainda endereçar a questão da incerteza inerente aos processos reais. Isso geralmente dá-se através da introdução de termos aleatórios ou estocásticos no conjunto de equações da modelagem matemática. Esses termos aleatórios tentam agrupar os vários fatores que influenciam na teoria, mas que não são contabilizados explicitamente na forma de variáveis ou equações do modelo.

À partir da formulação do problema ou da teoria e da sua especificação matemática e estatística, é possível iniciar a investigação prática. Para isso, é necessário obter dados para a estimação dos parâmetros do modelo. Antes de mais nada, os dados necessários para se responder a questão proposta devem estar disponíveis e este nem sempre é o caso. Ademais, os dados corretos devem ser obtidos para se abordar o problema específico. Além disso, a natureza dos dados possui papel fundamental na decisão do tipo de análise que deve ser feita. Para diferentes tipos de dados existem diferentes tipos de técnicas e a omissão da técnica correta pode levar a resultados equivocados.

De posse dos dados e do modelo estatístico, chega-se a hora de estimar seus parâmetros.

O resultado desta etapa são valores numéricos para os parâmetros que descrevem o modelo. Para isso, são empregadas diversas ferramentas estatísticas de estimação sendo a principal delas a análise de regressão. Essa será a técnica principal deste trabalho.

À essa altura da investigação, o modelo com seus parâmetros estimados já representa uma aproximação da realidade. No entanto, ainda é necessário avaliar se essa representação condiz com as nossas expectativas. Os testes de hipótese tentam verificar se as estimativas dos parâmetros do modelo são razoáveis dada a teoria. Assim, eles possuem o poder de suportar uma teoria ou refutá-la. Os métodos empregados aqui são baseados no ramo da estatística chamado Inferência Estatística.

Se a teoria não for refutada, o modelo considerado torna-se uma boa base para a predição de algumas variáveis. Como dito, a realidade conta com uma inerente incerteza que nem sempre será capturada pelo modelo. Na verdade, durante a estimação de um modelo deseja-se utilizar dados amostrais para inferir, na média, algum parâmetro populacional desconhecido. Porém, isso não diminui de forma alguma a importância da predição. Em muitos casos, o objetivo final de toda análise é poder realizar predições de algumas variáveis de interesse. O resultado de uma predição é um intervalo de confiança onde acredita-se, com algum grau de certeza, onde estará o valor real. Importante lembrar que a qualidade de um modelo para predição não deve ser medida pela sua capacidade de acertar uma predição e sim pela sua capacidade de medir o quanto pode-se errar.

Finalmente, o resultado da modelagem, estimação e validação pode ser usado como base para o desenvolvimento de políticas públicas ou privadas. O modelo pode assim exercer influência em diversos seguimentos e os efeitos podem, com certo grau de certeza, ser previstos.

É importante enfatizar que nem sempre dispõe-se de uma teoria única à princípio, mas de uma coleção de teorias concorrentes que competem pela melhor representação da realidade. Esse é o caso neste trabalho. Aqui não há um modelo único que propõe-se explicar a relação entre o valor de uma rede e seu tamanho. Na verdade, são considerados quatro diferentes modelos, cada qual com sua fundamentação teórica e implicações. Esses modelos podem ser vistos como rivais e somente um, se algum, é o verdadeiro modelo que descreve a relação entre valor e tamanho de uma rede.

3.1.2 Tipos de Dados

Os dados disponíveis podem ser de diferentes naturezas. Como ressaltado anteriormente, tipos diferentes de dados podem exigir tratamentos diferentes na sua análise. Alguns

métodos econométricos são aplicáveis com nenhuma ou pouca modificação a diferentes estruturas (WOOLDRIDGE, 2015).

Muitos conjuntos de dados são do tipo corte-transversal. Todas observações de um conjunto dessa natureza são consideradas amostradas em um determinado instante de tempo. A noção de "mesmo instante de tempo", em alguns casos, é um pouco relaxada. Por exemplo, durante o Censo Demográfico, diversas famílias são visitadas num período de semanas ou até meses. No entanto, os dados coletados ainda são considerados como observações do mesmo instante de tempo.

Por outro lado, alguns dados podem ser observados em intervalos ao longo do tempo. Tais dados constituem as ditas séries temporais. Exemplos clássicos de séries temporais são o PIB, a inflação e taxa de desemprego de um país. A frequência com que os dados são coletados pode ser diária, semanal, mensal, anual e assim por diante.

Uma característica importante das séries temporais é a dependência temporal entre suas observações. Eventos passados possivelmente influenciarão os eventos futuros e isso deve ser contabilizado. Ao contrário dos dados de corte-transversal onde a ordem em que os dados apresentam-se não é relevante, nas séries temporais, a ordem cronológica de cada evento é vital. Outra característica notória das séries temporais é a presença de tendência e sazonalidade. Esses assuntos serão tratados mais a frente.

Devido a essa relação temporal entre as suas observações, técnicas específicas devem ser empregadas quando analisa-se dados dessa natureza. A análise de regressão, por exemplo, exige alguns cuidados extras quando conduzida em séries temporais quando comparada à análise em dados de corte-transversal. Esse será um dos principais temas deste trabalho.

Outros tipos de dados como painel/longitudinal existem e são comuns em estudos econométricos, porém não são de interesse neste trabalho e portanto não serão tratados. O foco aqui será a abordagem econométrica sobre dados de corte-transversal e, principalmente, na modelagem e regressão de séries temporais.

3.1.3 Os Conceitos de Regressão

Como já dito, a análise de regressão é uma das principais ferramentas da Econometria. Informalmente, pode-se introduzir a regressão como a análise da dependência de uma variável (variável dependente) em relação a uma ou mais variáveis (variáveis explanatórias ou independentes) com o intuito de estimar e/ou predizer a média populacional da primeira em termos de valores fixos em amostras repetidas das últimas (GUJARATI, 2009). Cabe notar que

apesar de tentador, não se deve interpretar a regressão como uma espécie causalidade. O fato de uma variável poder ser explicada em termos de outras variáveis não implica uma relação causa-consequência entre elas.

A seguir será apresentada a regressão simples ou bivariada onde uma variável é explicada por apenas uma outra variável. Como o objetivo deste trabalho é verificar a relação entre duas variáveis (valor e tamanho de uma rede), o caso simples é suficiente. A extensão dos conceitos para o caso multivariado, ou seja, onde uma variável é explicada por duas ou mais variáveis, exige apenas a extensão dos conceitos para a álgebra matricial.

A introdução de alguma terminologia faz-se necessária neste ponto. É comum as variáveis envolvidas numa análise de regressão receberem nomes diferentes em algumas fontes da literatura econométrica e estatística. Alguns dos principais deles estão resumidos abaixo:

Tabela 3.1 – Variáveis de uma regressão

Variável dependente	**Variável independente**
Variável explicada	Variável explanatória
Variável endógena	Variável exógena
Regressando	Regressor
Variável controlada	Variável de controle

(Fonte: Próprio autor)

A variável dependente é aquela que é explicada pelo modelo. Variável independente é qualquer variável que explica o comportamento do modelo. Os modelos de equações simultâneas onde uma mesma variável ora toma o papel de regressor, ora o de regressando não será discutido aqui. No entanto, devido à natureza do problema em questão, um estudo que leve em conta tanto o efeito da variável *valor* na variável *tamanho* e vice-versa certamente tem muito a acrescentar.

Basicamente, a regressão de uma variável dependente Y em relação à variável independente X tenta estimar e fazer inferências sobre os parâmetros populacionais β_1 e β_2 do seguinte modelo:

$$Y_i = \beta_1 + \beta_2 X_i + u_i,$$

onde u_i denota o termo estocástico do i-ésimo elemento da população. Porém, como contamos apenas com amostras da população, não é possível determinar exatamente os parâmetros β_1 e β_2 e eles podem nunca ser conhecidos. Pode-se também descrever o modelo através do valor esperado condicional da variável dependente em relação à variável independente através da equação:

$$\mathbf{E}[Y|X_i] = \beta_1 + \beta_2 X_i.$$

Na verdade, tudo que dispomos são dados observados e portanto devemos lidar com valores estimados. A representação amostral do modelo é dada por:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i,$$

onde o "chapéu" nas variáveis representa que tais variáveis são estimadores das verdadeiras variáveis populacionais desconhecidas. De forma análoga, o modelo da média condicional é

$$\hat{Y}_i = \beta_1 + \beta_2 X_i,$$

onde $\hat{Y}_i$ denota o estimador de $\mathbf{E}[Y|X_i]$.

O termo u_i é chamado *termo estocástico de erro* e representa os diversos fatores que influenciam Y_i que não são explicados por X_i. Cabe notar que, apesar de considerar-se que u_i é um termo aleatório, ainda não se faz suposições sobre sua distribuição, apenas que $\mathbf{E}[u_i|X_i] = 0$. O seu estimador $\hat{u}_i$ é chamado *resíduo* e sua interpretação no contexto amostral é análoga ao do termo estocástico de erro no contexto populacional. Além disso, a variável dependente X é considerada sistemática, não estocástica. Em outras palavras, considera-se que os valores de X são sempre os mesmos para amostras repetidas.

Como neste trabalho serão utilizadas técnicas de regressão linear cabe esclarecer o significado de *linear*. A linearidade considerada aqui é a relação linear entre os **parâmetros** β e a esperança condicional de Y, $\mathbf{E}[Y|X_i]$. Não é assumida qualquer dependência linear entre Y e X e portanto modelos como

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3$$

são perfeitamente válidos, enquanto o modelo

$$Y_i = \beta_1 + \frac{1}{\beta_2} X_i$$

seria considerado não linear e, consequentemente, inválido.

3.2 Métodos de Regressão

Existem diferentes formas de obter-se a estimação dos parâmetros populacionais a partir dos dados observados. Talvez as duas formas mais comuns são os mínimos quadrados ordinários (MQO) e o de máxima verossimilhança (MV). Ambos serão abordados neste trabalho por razões diferentes. Apesar de não ser o mais adequado para o tipo de problema que é abordado neste trabalho, o MQO é apresentado pelas suposições que faz a respeito do modelo. O método MV é o mais indicado neste caso e será utilizado na análise de regressão.

3.2.1 O Método de Mínimos Quadrados Ordinários

O método de Mínimos Quadrados Ordinários deve-se ao matemático alemão Carl Friedrich Gauss e tem sua base bastante intuitiva. Por sua simplicidade teórica e propriedade estatísticas atrativas, o MQO tornou-se uma das técnicas mais estudadas e populares na análise de regressão.

Considerando a função de regressão amostral (FRA) bivariada

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i,$$

pode-se escrever o resíduo como

$$\hat{u_i} = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i.$$

O resíduo pode ser interpretado como a diferença entre os reais valores observados e os valores previstos pelo modelo. Intuitivamente, quanto menor tal diferença, melhor o modelo. A ideia do MQO portanto é escolher os parâmetros β de forma a minimizar o total dessa diferença. No entanto, é possível mostrar que minimizar $\sum \hat{u_i}^2 = \sum (Y_i - \hat{Y}_i)^2$ é um critério melhor do que minimizar simplesmente $\sum \hat{u_i} = \sum (Y_i - \hat{Y}_i)$.

A tarefa resume-se a achar os valores de β_1 e β_2 que minimizam

$$\sum \hat{u_i}^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

e pode ser resolvida com um simples ferramental de cálculo univariável. Os parâmetros assim obtidos são

$$\hat{\beta}_2 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

e

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X},$$

onde $\bar{X}$ e $\bar{Y}$ denotam a média aritmética de X e Y, respectivamente.

Os estimadores $\hat{\beta}_1$ e $\hat{\beta}_2$ obtidos assim são chamados de *estimadores de mínimos quadrados*. Além disso, por serem funções dos dados amostras, esses estimadores constituem variáveis aleatórias e portanto possuem seus valores variando de amostra para amostra. Assim, é interessante ter-se uma forma de avaliar a dispersão dessas variáveis. Suas equações são mostradas abaixo

$$\mathbf{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum(X_i - \bar{X})^2}$$

e

$$\mathbf{Var}(\hat{\beta}_1) = \frac{\sigma^2 \sum X_i^2}{n \sum(X_i - \bar{X})^2}.$$

A variância desconhecida do termo de erro, σ^2, também pode ser estimada. O estimador de MQO dessa variância é simplesmente a razão entre a soma dos quadrados dos resíduos (*residual sum of squares* - RSS) pelo grau de liberdade da regressão simples, $n - 2$, ou seja

$$\hat{\sigma}^2 = \frac{\sum \hat{u_i}^2}{n - 2}.$$

O erro-padrão da regressão (*standard error* - se) é computado como a raiz quadrada da variância estimada:

$$\mathbf{se} = \hat{\sigma} = \sqrt{\frac{\sum \hat{u_i}^2}{n - 2}}.$$

3.2.2 O Teorema de Gauss-Markov

Apesar de interessante e simples, o MQO exige que certas condições sejam satisfeitas para que os estimadores assim obtidos possam servir para inferência dos verdadeiros parâmetros populacionais. Verificar se as suposições feitas pelo método foram consideradas em um

trabalho empírico é uma tarefa necessária na sua revisão, pois muitas vezes a não manutenção das suposições pode conduzir a resultados equivocados (**fonte**).

O Teorema de Gauss-Markov enuncia que se o modelo de regressão atende às suposições do *Modelo de Regressão Linear Clássico* (MRLC), então os estimadores obtidos serão *BLUE* (*Best Linear Unbiased Estimators*) ou *melhores estimadores lineares não viesados*. Em outras palavras, os estimadores obtidos pelo MQO serão os estimadores de menor variância dentre toda classe de estimadores não viesados, ou seja, serão estimadores eficientes. Um estimador não viesado $\hat{\beta}$ do parâmetro populacional β é aquele que satisfaz a equação

$$\mathbf{E}[\hat{\beta}] = \beta.$$

A questão é quais condições devem ser atendidas para que isso ocorra. Brevemente, as suposições do *modelo de regressão linear clássico* são:

1. **Modelo de regressão linear**: Como observado anteriormente, o tipo de linearidade considerada aqui é em relação aos parâmetros do modelo de regressão. O modelo linear simples pode ser representado como a seguir

$$Y_i = \beta_1 + \beta_2 X_i + u_i.$$

2. X **é não estocástico**: Em amostras repetidas, os valores de X são sempre os mesmos. Essa suposição é importante já que considera-se no modelo populacional o valor esperado da variável dependente Y condicionado à variável independente X, ou seja, $\mathbf{E}[Y|X_i]$. Portanto, faz sentido que os valores de X não devam variar entre uma amostra e outra.

3. **Valor esperado de** u_i **é zero**: Para cada valor da variável explanatória X, a média condicional do erro é zero.

$$\mathbf{E}[u_i|X_i] = 0.$$

Isso implica que cada Y populacional correspondente a um X seja distribuído em torno de sua média. Esta também é uma suposição razoável dado que supõe-se que o termo de erro u_i não afeta sistematicamente a média de Y.

Por ora, não se faz nenhuma suposição acerca da distribuição de probabilidade do termo de erro além de que ela deve ser simétrica e centrada em zero. Mais a frente, estenderemos essa suposição para que o termo de erro tenha distribuição normal para fins de inferência.

4. **Termos de erro são homocedásticos**: Para cada valor de X, a variância de u_i é a mesma para todas observações. Simbolicamente:

$$\mathbf{Var}(u_i|X_i) = \sigma^2, \text{para todo } i.$$

Essa suposição refere-se à propriedade chamada *homocedasticidade* (*homo* - igual, *skedasis* - dispersão). Ela significa que a dispersão dos termos de erro u_i em torno da FRP mantém-se constante ao longo da variável X de forma que os valores de Y não se tornem nem mais, nem menos dispersos.

Em contrapartida, o fenômeno da *heterocedasticidade* (*hetero* - diferente, *skedasis* - dispersão) ocorre quando a variância dos termos de erro são condicionadas à variável independente. Quando isso acontece, a dispersão da variável dependente Y pode aumentar ou diminuir ao longo da variável independente X. Isso pode ser escrito como

$$\mathbf{Var}(u_i|X_i) = \sigma_i^2.$$

Note o subscrito i na variância σ^2 indicando que a variância não é mais constante e sim variável com cada X_i.

Essa é uma das suposições mais importantes do modelo. Frequentemente, ela não é atendida devido a própria natureza do problema em questão. Faz-se necessário então o desenvolvimento de métodos que permitam identificar a presença de heterocedasticidade e de medidas de correção, caso necessário. Alguns desses métodos serão tratados mais adiante.

5. **Termos de erro não autocorrelacionados**: A correlação entre quaisquer dois u_i e u_j, com $i \neq j$, é zero. Ou seja,

$$\mathbf{Cov}(u_i, u_j) = 0, \text{para quaisquer } i \text{ e } j,\ i \neq j.$$

A equação acima enuncia que quaisquer dois termos de erro são não correlacionados. Isso significa dizer que os fatores que influenciam o modelo, mas que não foram incorporados nele, não possuem correlação serial. Posto de outra forma, correlação serial no modelo $Y_t = \beta_1 + \beta_2 X_t + u_t$ significa que o termo de erro u_{t-1} possui algum grau de influência (positiva ou negativa) sobre u_t e portanto em Y_t.

Assim como a suposição 4, essa suposição tem enorme importância nos modelos aqui tratados. Isso porque os dados utilizados neste trabalho são originários de séries temporais e estas apresentam correlações seriais naturalmente.

6. **A covariância entre X_i e u_i é zero**: os termos de erro u e a variável explanatória X são não correlacionados.

$$\mathbf{Cov}(X_i, u_i) = 0.$$

A explicação para essa exigência é bastante intuitiva. O modelo considera que tanto X, quanto u possuem efeitos separados (aditivos) em Y. Porém, se X e u forem correlacionados, torna-se difícil avaliar o efeito individual de cada um em Y.

7. **O número de observações deve ser maior que o número de parâmetros a serem estimados**: Um modelo com k variáveis explanatórias exige um número n de observações, $n > k$, para ser estimado.

Aqui, não será necessário preocupar-se com essa suposição já que, na regressão simples, bastam duas observações para que seja possível estimar o modelo.

8. **Existe variabilidade nos valores de** X: Os valores de X numa dada amostra não podem ser todos iguais. Ou seja,

$$0 < \mathbf{Var}(X) < \infty.$$

Se todas observações são de um mesmo valor de X, então $X_i = \bar{X}$, para todo i. Isso torna impossível estimar β_2 e consequentemente β_1. De forma geral, ambas variável independente e dependente devem variar.

9. **O modelo de regressão está corretamente especificado**: Não existe viés ou erro de especificação no modelo utilizado na análise empírica.

Existem diversas formas de especificar um modelo incorretamente. É possível que inclua-se no modelo variáveis explanatórias demais ou de menos, que escolha-se a forma funcional errada ou que considere-se uma distribuição errada para as variáveis. De qualquer forma, cada um desses equívocos pode levar a resultados errôneos distintos.

10. **Não existe perfeita multicolinearidade**: Não há uma perfeita relação linear entre as variáveis explanatórias.

Essa suposição é particularmente importante no caso de regressão multivariada. Esse não é o caso aqui e portanto não será dada maior atenção.

3.2.3 Hipótese de Normalidade dos Termos de Erro

Anteriormente, foi mostrado como se obter estimadores ($\hat{\beta}_1$, $\hat{\beta}_2$ e $\hat{\sigma}^2$) para os verdadeiros parâmetros populacionais de interesse (β_1, β_2 e σ^2, respectivamente). No entanto, esses estimadores naturalmente variam de amostra para amostra e são, portanto, variáveis aleatórias.

De forma geral, os objetivos da análise de regressão vão além da estimação dos parâmetros. De posse dos estimadores, é interessante também ser capaz de realizar inferências sobre eles, ou seja, de quantificar o erro que está comentendo-se ao utilizar aquela estimativa como se fosse o valor populacional. Para que isso seja possível, é necessário primeiro descobrir as distribuições de probabilidades dos estimadores.

É possível mostrar que, no caso dos estimadores de MQO, as distribuições dos parâmetros $\hat{\beta}_1$ e $\hat{\beta}_2$ dependem tão somente da distribuição dos termos de erro, u_i. No entanto, o método de MQO não faz nenhuma suposição quanto essa distribuição. Torna-se bastante conveniente assumir que u_i segue uma distribuição normal estendendo assim o Modelo de Regressão Linear Clássico para o *Modelo de Regressão Linear Clássico Normal* (MRLCN).

A distribuição normal é escolhida por diversas razões, algumas listadas a seguir:

- A distribuição de probabilidade da soma das variáveis não introduzidas explicitamente no modelo, para um grande número delas, tende a seguir uma distribuição normal se elas forem independentes e identicamente distribuídas como pode ser mostrado pelo *Teorema do Limite Central*.
- A suposição de normalidade facilita a derivação da distribuição de probabilidade dos estimadores de MQO já que uma propriedade da distribuição normal é que qualquer função linear de variáveis normalmente distribuídas também é normalmente distribuída.
  A suposição de normalidade é afinal a seguinte:

$$\mathbf{E}[u_i] = 0;$$

$$\mathbf{E}[u_i - \mathbf{E}[u_i]]^2 = \mathbf{E}[u_i^2] = \sigma^2;$$

$$\mathbf{E}[u_i - \mathbf{E}[u_i]][u_j - \mathbf{E}[u_j]] = \mathbf{E}[u_i u_j] = 0, \, i \neq j,$$

  ou de forma mais compacta

$$u_i \sim NID(o, \sigma^2).$$

  Sob a hipótese de normalidade, os estimadores de MQO têm as seguintes propriedades desejáveis:

- O estimador $\hat{\beta}_1$ segue distribuição normal:

$$\mathbf{E}[\hat{\beta}_1] = \beta_1;$$

$$\mathbf{Var}[\hat{\beta}_1] = \sigma^2_{\hat{\beta}_1} = \frac{\sum X_i^2}{n\sum(X_i - \bar{X}_i)^2}\sigma^2,$$

ou seja,

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1}).$$

- Da mesma forma, o estimador $\hat{\beta}_2$ também segue uma distribuição normal:

$$\mathbf{E}[\hat{\beta}_2] = \beta_2;$$
$$\mathbf{Var}[\hat{\beta}_2] = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2}{\sum(X_i - \bar{X}_i)^2},$$

ou seja,

$$\hat{\beta}_2 \sim N(\beta_2, \sigma^2_{\hat{\beta}_2}).$$

Além disso, é possível mostrar que, seguindo a hipótese de normalidade dos termos de erro, os estimadores acima são os *melhores estimadores não viesados*.[1] A diferença para os estimadores *BLUE* é que não há mais a restrição à classe de estimadores lineares (RAO, 2009 apud GUJARATI, 2009).

3.2.4 Heterocedasticidade

No MRLC, foi feita uma suposição a respeito dos termos de erro de grande relevância. Considerou-se que os termos de erro, u_i, possuem variância constante, fenômeno chamado *homocedasticidade*. Isso significa que a variância do erro, $\mathbf{Var}[u_i]$, mantém-se constante e independe dos valores de X_i. A situação inversa (*heterocedasticidade*), quando a variância deixa de ser constante, representa uma violação da hipótese e tem importante implicação nas análises de regressão.

Partindo da suposição de que os termos de erro possuem média zero, é possível expressar a homocedasticidade como

$$\mathbf{E}[u_i^2] = \sigma^2, i = 1, 2, \ldots, n.$$

Note como o subscrito i presente no termo de erro não aparece na variância, explicitando que este é uma constante. Por outro lado, a heterocedasticidade refere-se à variâncias

[1] Do inglês, *BUE - Best Unbiased Estimators*.

condicionais não mais constantes

$$\mathbf{E}[u_i^2] = \sigma_i^2, i = 1, 2, \ldots, n,$$

onde o subscrito i na variância denota que ela agora varia.

Os motivos para que ocorra heterocedasticidade são vários. Esse é um fenômeno observado comumente no aprendizado humano, por exemplo. De acordo com que aprendemos, os nossos erros tendem a diminuir resultando em variância decrescente ao longo do tempo. O avanço das técnicas de coleta de dados também pode ser responsável pela existência de heterocedasticidade. Com o decorrer do tempo e melhoria das técnicas de coleta, as variacões tendem a reduzir-se, diminuindo assim a variância observada. Para finalizar, a presença de *outliers* também pode introduzir heterocedasticidade principalmente em pequenas amostras.

Porém, quais são as consequências da heterocedasticidade? O problema da variância não constante dos termos de erro tem implicação durante a inferência dos estimadores de MQO e não na sua estimação. Na presença de heterocedasticidade, as fórmulas para $\mathbf{Var}[\hat{\beta}_1]$ e $\mathbf{Var}[\hat{\beta}_2]$ não são mais válidas. Na verdade, os estimadores $\hat{\beta}_1$ e $\hat{\beta}_2$ deixam de ser *BLUE*. Apesar de permanecerem lineares, não viesados e consistentes, eles não são mais eficientes. Em outras palavras, mesmo de posse de algumas propriedades de interesse, os estimadores obtidos não são mais os melhores, ou seja, os de menor variância dentre os lineares e não viesados. Ainda é possível, através da técnica de *Mínimos Quadrados Generalizados* (MQG), obter estimadores *BLUE*.

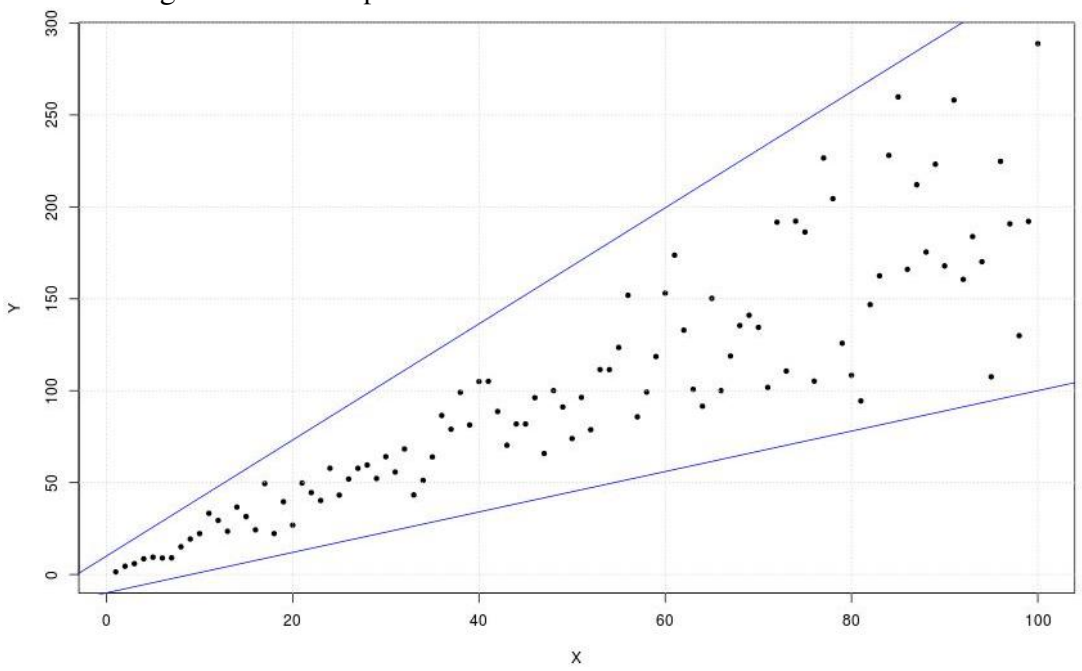
Na existência de heterocedasticidade, a variância do estimador de MQO $\hat{\beta}_2$ pode ser expressa como

$$\mathbf{Var}[\hat{\beta}_2] = \frac{\sum(X_i - \bar{X})^2\sigma_i^2}{[\sum(X_i - \bar{X})^2]^2}.$$

Note que, se $\sigma_i = \sigma$ para $i = 1, 2, \ldots, n$, a fórmula acima iguala-se à fórmula para variância de $\hat{\beta}_2$ já mostrada. Se a fórmula acima for utilizada na inferência considerando-se conhecidas as variâncias σ_i^2, as estatísticas produzidas (por exemplo, nos testes t e F) serão imprecisas ao resultarem em valores menores do que realmente são. Assim, alguns coeficientes que podem parecer estatisticamente não significantes podem ser, na verdade, significantes. Um exemplo de como a heterocedasticidade pode manifestar-se graficamente é dado na Figura 3.1.

Um caso ainda pior é quando utiliza-se os estimadores de MQO e ainda desconsidera-se a heterocedasticidade na variância dos estimadores dos coeficientes. Se isso acontecer, quaisquer conclusões e inferências serão não confiáveis. Em outras palavras, deve-se utilizar um

Figura 3.1 – Exemplo de heterocedasticidade: variância não constante



(Fonte: Próprio autor)

método mais correto na estimação dos coeficientes (por exemplo, MQG) e a fórmula correta para suas variâncias.

3.2.5 Autocorrelação

Uma das hipóteses do Modelo de Regressão Linear Clássico apresentado anteriormente era que os termos de erro eram não correlacionados (Hipótese 5). No entanto, quando lidando com dados de séries temporais, como neste trabalho, essa suposição não mais é válida. A autocorrelação[2] dos termos de erro é bastante comum em dados serialmente ordenados (principalmente quando o intervalo entre uma observação e outra é curto) e, portanto, deve ser levada em consideração.

A autocorrelação pode ser expressa como

$$\mathbf{E}[u_i u_j] \neq 0, i \neq j.$$

Como no caso da heterocedasticidade, os estimadores de MQO ainda são lineares, não viesados e consistentes, porém não eficientes. Isso significa que não são os estimadores com

[2] Alguns autores preferem explicitar a diferença dos termos *autocorrelação* e *correlação serial*. A primeira refere-se à correlação existente entre termos defasados de uma mesma série temporal, enquanto a segunda considera que a correlação ocorre entre duas séries distintas.

menor variância dentre a classe de estimadores não viesados. É importante notar que o Teorema de Gauss-Markov dá as condições suficientes para que os estimadores sejam *BLUE*, mas não necessárias. Isso significa que mesmo que algumas das condições do MRLC sejam violadas (como heterocedasticidade e autocorrelação), os estimadores obtidos ainda podem ser *BLUE* ainda que raramente. As condições necessárias e suficientes são dadas pelo Teorema de Krushkal.

Os métodos para tratamento de séries temporais levam em conta a existência de autocorrelação já que esta é uma das principais características desse tipo de dados. Voltaremos a este assunto quando discutirmos séries temporais.

3.3 Seleção de Modelo

No presente trabalho, tratamos da análise qualitativa e quantitativa de diversos modelos concorrentes acerca do valor das redes. É necessário portanto métodos para avaliar e selecionar o melhor modelo, se algum. A fim de ser considerado na análise, um modelo deve ser minimamente razoável. Uma lista do que um modelo deve primariamente tentar satisfazer e que servirá de base para nossa análise é dada por (HENDRY; RICHARD, 1983). O modelo

1. leva à predições que são logicamente possíveis;
2. deve ser consistente com a teoria que o suporta;
3. possui variáveis explanatórias que são não correlacionadas com os termos de erro;
4. exibe consistência nos seus parâmetros de forma que as predições possam ser confiáveis;
5. possui resíduos que são puramente aleatórios;
6. é capaz de explicar os demais modelos.

Na tentativa de selecionar e analisar modelos, é comum que se cometa erros de várias naturezas. Alguns deles são chamados *erros de especificação* já que são relacionados à má especificação das equações do modelo. Como exemplo, tem-se a inclusão de variáveis irrelavantes, bem como a omissão de variáveis relevantes. Além disso, pode ocorrer erro na especificação da forma funcional das equações, erros de medidas e outros.

Como todos modelos que serão tratados aqui são bastante simples em suas especificações, devemos prestar atenção especial ao problema da omissão de variáveis relevantes. Todos modelos, de Metcalfe a Odlyzko-Briscoe, consideram que a única variável explanatória para o valor da rede é o seu tamanho. Isso certamente não é verdade. O senso comum nos diz que vários outros fatores influenciam no valor da rede como, por exemplo, a sua conectividade e

estrutura, a velocidade com que a informação pode propagar-se por ela e muitos outros. No entanto, esses fatores não serão levados em consideração e as consequências disso devem ser explicitadas.

3.3.1 Omissão de variáveis relevantes

Suponha o seguinte modelo verdadeiro:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

e que escolhe-se, no entanto, avaliar o seguinte modelo:

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + \nu_i.$$

As consequências da omissão da variável X_{3i} do modelo são:

1. Se X_2, a variável inclusa, e X_3, a variável omitida, são correlacionadas, então tanto $\hat{\alpha_1}$ quanto $\hat{\alpha_2}$ serão viesados e não consistentes, mesmo assintoticamente. Ou seja,

$$\mathbf{E}[\hat{\alpha_1}] \neq \beta_1 \text{ e } \mathbf{E}[\hat{\alpha_2}] \neq \beta_2.$$

2. Se X_2 e X_3 são não correlacionados, então apenas $\hat{\alpha_1}$ é viesado.
3. De qualquer forma, a variância do termo de erro, σ^2, será estimada incorretamente.
4. A variância do estimador $\hat{\alpha_2}$ também será viesada.
5. Os intervalos de confiança e testes de hipótese sobre os estimadores poderão dar resultados errados.
6. As predições feitas pelo modelo serão imprecisas.

De posse dos dados, diversos testes podem ser feitos para verificar a adequação do modelo ao que era esperado e possivelmente detectar problemas como a omissão de variáveis que são relevantes no modelo. Não entraremos nos detalhes de cada método e apenas os citaremos brevemente.

Uma das formas mais imediatas de avaliar um modelo é através da análise dos resíduos da regressão. Com uma simples análise visual, pode ser possível detectar autocorrelação e heterocedasticidade, além de ajudar a identificar quando ocorreu omissão de variáveis. De forma mais geral, o método consiste em traçar o gráfico dos resíduos da regressão, $\hat{u_i}$, contra a

variável dependente, Y_i. Se houver erro de especificação, os resíduos exibirão padrões notáveis.

Outros dois métodos também estão disponíveis. O primeiro é utilizando a estatística d de Durbin-Watson e o segundo é o teste RESET.

3.3.2 Critérios de Seleção

Diante de modelos concorrentes, devemos ser capazes de escolher o melhor sistematicamente. Para isso, foram desenvolvidos alguns critérios que vão servir como métricas de desempenho dos modelos. É importante notar que existem dois cenários onde pode-se avaliar um modelo. O primeiro é utilizando a amostra usada para estimar o modelo. O segundo é verificando a adequação do modelo com observações futuras, não usadas na sua estimação. Aqui, será dada especial atenção a dois critérios bem simples: R^2 ajustado e o Critério de Informação de Akaike (*Akaike Information Criterion* - AIC).

O R^2 é uma medida de adequação bastante popular e é definido como:

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

sendo SQR a Soma dos Quadrados dos Resíduos e SQT a Soma dos Quadrados Totais.

Essa métrica, no entanto, apresenta três problemas. Primeiro, ele mede a qualidade da adequação do modelo aos dados da amostra utilizada na estimação do modelo provendo nenhuma garantia de que o modelo tenha desempenho similar quando exposto a novos dados nunca antes vistos. Segundo, as variáveis explanatórias devem ser as mesmas a fim de comparação utilizando o R^2. Por fim e mais importante, o R^2 sempre aumenta quando adiciona-se novos regressores. Isso significa que é possível inflar e projetar uma falsa qualidade de um modelo simplesmente aumentando-se o número de variáveis explanatórias.

Para resolver o último problema, recomenda-se a utilização do R^2 ajustado que penaliza a adição de novos regressores. A fórmula modificada segue:

$$\bar{R}^2 = 1 - (1 - R^2) - \frac{n-1}{n-k},$$

onde n é o número de observações e k o número de variáveis explanatórias.

O Critério de Informação de Akaike é uma medida que naturalmente já penaliza a adição de novos regressores ao modelo. Ele é definido como:

$$AIC = e^{2k/n}\frac{SQR}{n}$$

Às vezes, pode tornar-se conveniente apresentar o AIC em sua forma logarítmica:

$$lnAIC = \left(\frac{2k}{n}\right) + ln\left(\frac{RSS}{n}\right).$$

Quando comparando-se modelos, aquele que resultar no menor valor de AIC é considerado o melhor. Ao contrário do R^2 (normal ou ajustado), o AIC pode ser utilizado para avaliar a qualidade de predição de um modelo diante de novas amostras.

3.4 Regressão com Séries Temporais

Como discutido anteriormente, existe uma classe de dados que possui dependência explícita em relação à sua ordenação temporal. Quando a ordem em que observou-se a amostra torna-se relevante para sua análise, chamamos esses dados de séries temporais.

Tecnicamente, uma série temporal é uma realização possível de um processo estocástico. Logo, além de fazer sentido definir processo estocástico primeiro, isso ajuda a esclarecer a natureza das séries temporais. Um processo estocástico é uma sequência ordenada de variáveis aleatórias X_1, X_2, ..., X_t, $t \in Z$. De forma mais compactada, um processo estocástico é denotado por $\{X_t\}$.

O fato de uma série temporal ser uma sequência necessariamente ordenada tem um grande papel na análise desse tipo de dados e não pode ser facilmente desconsiderada. Considere como exemplo a série do Produto Interno Bruto (PIB) de um país divulgado trimestralmente. Nesse caso, existe grande diferença entre observar um PIB de R$ 50 bilhões no tempo t e R$ 100 bilhões em $t + 1$ ou o contrário. No primeiro caso, o PIB duplicou entre uma divulgação e outra, o que é algo bastante positivo para a economia do país, enquanto no segundo caso, ele reduziu-se à metade, um sinal nada bom. Como já dito, uma das suposições do MRLC é a ausência de correlação serial entre as observações. Certamente, esse não é um cenário que ocorre em dados de séries temporais já que cada observação é extremamente propensa a exercer influência nas observações seguintes e ser influenciada pelas observações passadas.

Séries temporais, apesar de bastante comuns em situações reais, impõem diversos novos desafios à análise. Para trabalhar com esse tipo de dados, é necessário antes estender o vocabulário através de novos conceitos. Em parte devido à importância prática dessa classe de dados e de seu extenso espectro de aplicações, que vai desde previsão meterológica até reconhecimento de voz, a literatura para sua análise é vasta e facilmente encontrada. Ver (GUJARATI, 2009), (WOOLDRIDGE, 2015), (SHUMWAY; STOFFER, 2010), (BUENO, 2008), (ENDERS, 2008)

e (HAMILTON, 1994).

Para este trabalho, será suficiente estudar as condições necessárias para a realização de regressão entre duas séries, algo que nem sempre é trivial de se observar. É importante notar que a análise de séries temporais é extremamente mais profunda e extensa do que aquilo que será superficialmente tratado aqui. Por exemplo, assuntos como predição de séries e análise espectral são alguns dos temas mais recorrentes na área, mas que não acrescentarão muito ao objetivo do trabalho e portanto não serão discutidos. Por outro lado, os conceitos de estacionariedade e cointegração desempenham papel importante quando trata-se de regressão e assim serão tratados com maior dedicação. Também chama-se a atenção para o fato de que, se os dados fossem de seção transversal (e não de séries temporais), não haveria necessidade de tratar esse assunto aqui e toda atenção poderia ser prestada tão somente à análise de regressão.

3.4.1 Estacionariedade

Uma das principais propriedades de um processo estocástico é a estacionariedade. Essa característica é essencial para que diferentes técnicas de análise e predição de séries temporais seja sucedida. Costuma-se falar de estacionariedade com o sentido de estacionariedade fraca. Informalmente, um processo estocástico possui estacionariedade fraca se sua média e variância são constantes ao longo do tempo e a covariância entre dois instantes depende apenas da distância temporal entre eles e não dos instantes de tempo específicos. Matematicamente, se Y_t denotar um processo estocástico com estacionariedade fraca, então

$$\mathbf{E}[Y_t] = \mu;$$

$$\mathbf{Var}[Y_t] = \sigma^2;$$

$$\gamma_k = \mathbf{E}[(Y_t - \mu)(Y_{t+k} - \mu)],$$

onde γ_k denota a covariância (ou autocovariância) com defasafem k, ou seja, a covariância entre Y_t e Y_{t+k}. Se k for zero, γ_0 é simplesmente a variância σ^2. Uma série que não apresenta essas propriedades é dita não estacionária.

Infelizmente, grande parte das séries reais são não estacionárias. Logo, deve ser dedicado um tempo a compreender essas séries e como manipulá-las. O exemplo clássico de séries não estacionárias é o modelo de passeio aleatório[3] que pode ser subdividido em dois tipos: sem

[3]Do inglês, *random walk model*.

drift e com *drift*.

O passeio aleatório é representado por

$$Y_t = Y_{t-1} + u_t,$$

onde u_t representa o distúrbio estocástico com distribuição normal no tempo t, ou seja, $u_t \sim N(0, \sigma^2)$. Nesse modelo, cada observação depende da observação diretamente anterior e de um termo estocástico. Como cada termo anterior também depende de um termo estocástico, é fácil notar que o impacto desses termos é propagado para o futuro. O passeio aleatório se lembrará do choque para sempre.

O passeio aleatório com *drift* possui um termo a mais, δ, chamado parâmetro do *drift*:

$$Y_t = \delta + Y_{t-1} + u_t.$$

Também é interessante introduzir os conceitos de tendência estocástica e determinística. As séries que serão apresentadas na parte prática do trabalho claramente apresentam sinais de tendência e por isso é justo apresentar esses conceitos aqui. Antes, são apresentados dois operadores importantes na análise de séries temporais.

O operador diferença (Δ) computa simplesmente a diferença entre seu argumento e seu antecessor. Por exemplo:

$$\Delta Y_t = Y_t - Y_{t-1} \text{ e}$$

$$\Delta^2 Y_t = \Delta \Delta Y_t = \Delta(Y_t - Y_{t-1}) = Y_t - 2Y_{t-1} + Y_{t+2}.$$

O operador defasagem (L), por sua vez, retorna o antecessor de seu argumento:

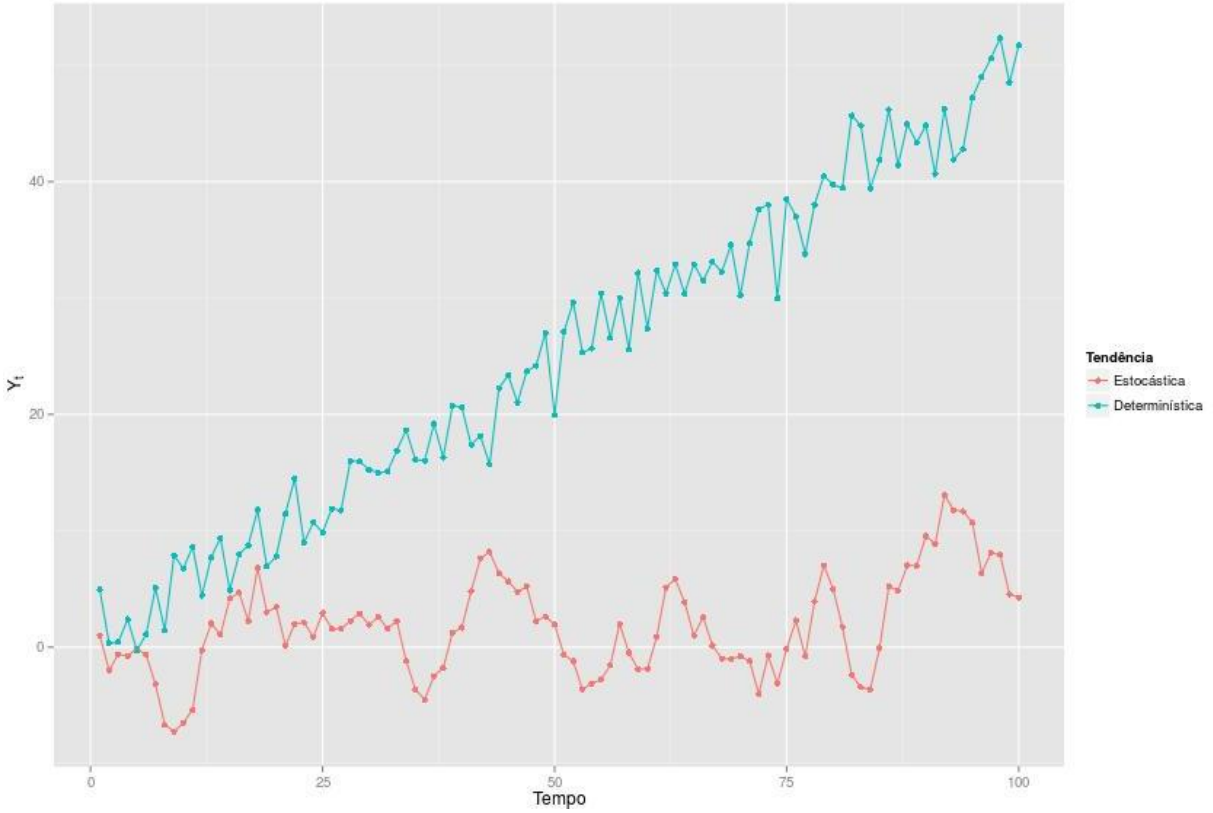
$$LY_t = Y_{t-1}.$$

Retornando, o passeio aleatório com *drift* é um exemplo de modelo que apresenta tendência estocástica, enquanto o modelo

$$Y_t = \beta_1 + \beta_2 t + u_t$$

apresenta tendência determinística. A diferença sutil nas duas formas é mais facilmente entendida através da Figura 3.2.

Figura 3.2 – Séries temporais com tendência estocástica e determinística



(Fonte: Próprio autor)

A série chamada com tendência estocástica foi gerada a partir do modelo de passeio aleatório $Y_t = Y_{t-1} + u_t$ com $u_t \sim N(0, 2.5)$. A série com tendência determinística ilustra a tendência em torno da reta seguindo o modelo $Y_t = 0.5t + u_t$, onde u_t segue a mesma distribuição anterior. Note como a tendência estocástica gera um efeito de memória: observações próximas tendem a seguir a mesma tendência (subida ou descida). Na tendência determinística não há esse efeito, ou seja, observações próximas não compartilham uma tendência necessariamente (uma subida pode ser seguida de uma descida e vice-versa).

O que tem de interessante nesses dois modelos é que ambos, apesar de originalmente não estacionários, podem ser transformados em séries estacionárias. A primeira é possível através da simples diferenciação da série original:

$$\Delta Y_t = \delta + u_t.$$

Como δ é uma constante e u_t tem distribuição com média zero, ΔY_t constitui uma série estacionária. Por outro lado, a tendência determinística pode ser removida através do processo de *detrending*. No caso mostrado acima, subtraindo-se de Y_t sua média, $\beta_1 + \beta_2 t$, chega-se a

$$Y_t - \mathbf{E}[Y_t] = u_t$$

que não passa de um ruído branco estacionário. Como nos casos reais as séries dificilmente serão originalmente estacionárias, ambas as técnicas são bastante úteis como uma primeira abordagem na tentativa de obter séries estacionárias. Em especial, a diferenciação, além de extremamente comum na prática, servirá para introduzir o conceito de ordem de integração. É possível que uma mesma série apresente os dois tipos de tendência. A tendência estocástica só pode ser removida através de diferenciação. Vale lembrar que a diferenciação também é um processo capaz de remover a tendência determinística ao custo de adicionar ruído.

3.4.2 Ordem de Integração e Regressão Espúria

Como foi apresentado, uma série temporal pode, a princípio, não ser estacionária, mas possivelmente pode tornar-se através de alguma das manipulações mostradas. Dizemos que uma série possui ordem de integração d quando necessita de d diferenciações para tornar-se estacionária. Por exemplo, o passeio aleatório com *drift* precisou de apenas uma diferenciação para alcançar estacionariedade e portanto podemos dizer que possui ordem de integração um.[4] Em geral, denota-se uma série temporal Y_t com ordem de integração d por $Y_t \sim I(d)$. Em particular, se essa série não necessita de diferenciação para tornar-se estacionária, ou seja, ela já é estacionária a princípio, então ela possui ordem de integração zero, ou seja, $Y_t \sim I(0)$.

Conhecer a ordem de integração de uma série é um dos primeiros passos que devem ser tomados para analisá-la corretamente. Um dos exemplos mais clássicos de não observância da ordem de integração é o caso da regressão espúria (YULE, 1926 apud GUJARATI, 2009). Duas séries não estacionárias podem apresentar aparente relação mesmo que os dois processos nada tenham em comum. O exemplo a seguir com uma amostra de 100 observações ilustra bem isso. Considere duas séries Y_t e X_t modeladas por passeio aleatório, ou seja,

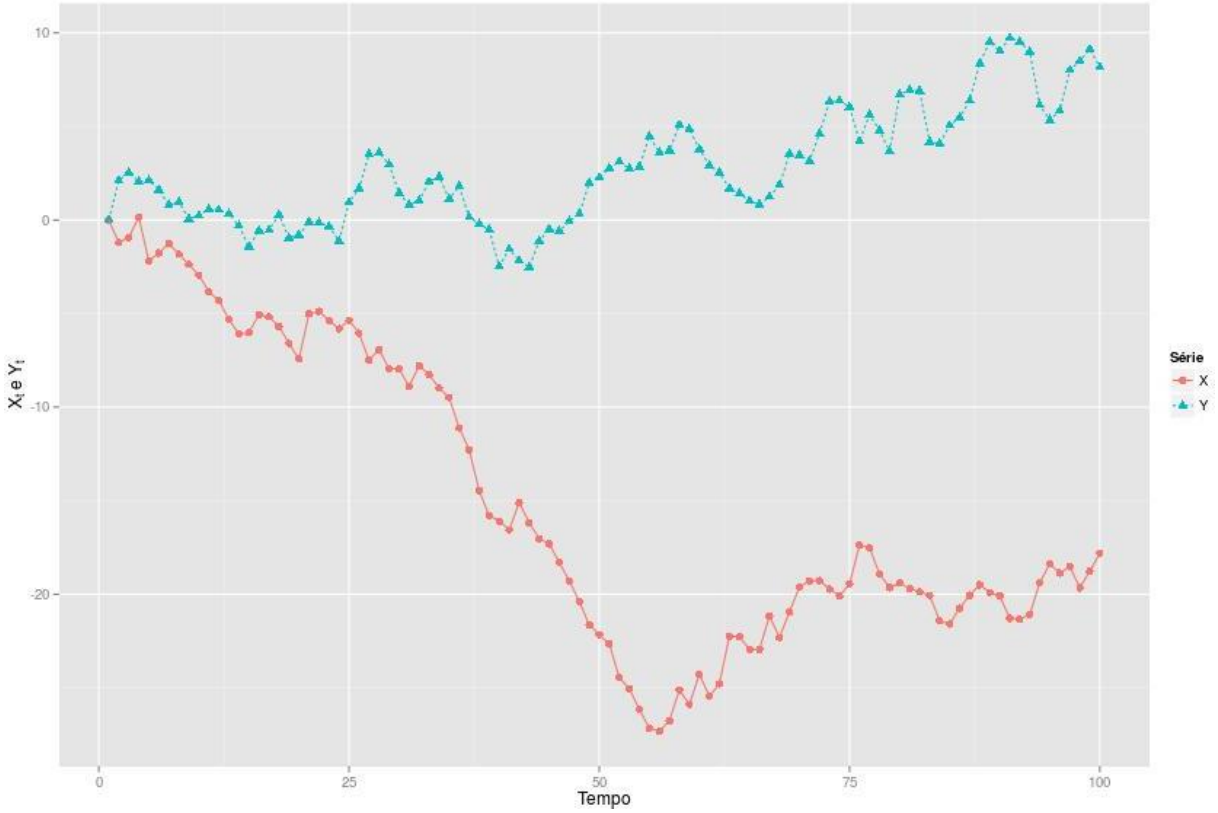
$$Y_t = Y_{t-1} + u_t;$$

$$X_t = X_{t-1} + v_t,$$

onde u_t e v_t são ambos termos estocásticos gerados a partir da distribuição normal, $u_t, v_t \sim N(0, 1)$.

[4] Outra forma de enunciar isso é dizendo que o passeio aleatório com *drift* é integrado de ordem um.

Figura 3.3 – Dois passeios aleatórios



(Fonte: Próprio autor)

Certamente, Y_t e X_t não possuem relação e isso deveria ficar aparente quando regredimos Y_t contra X_t:

$$Y_t = \beta_1 + \beta_2 X_t + w_{1t}.$$

Porém, como ambas as séries são não estacionárias integradas de primeira ordem, ou seja são $I(1)$, o resultado da regressão é um tanto inusitado

Tabela 3.2 – Regressão entre dois passeios aleatórios

Coeficiente	**Estimativa**	**Estatística** t	**Valor-p**
$\widehat{\beta_1}$	-0.03768	-0.065	0.948
$\widehat{\beta_2}$	-0.18876	-5.559	2.35e-07

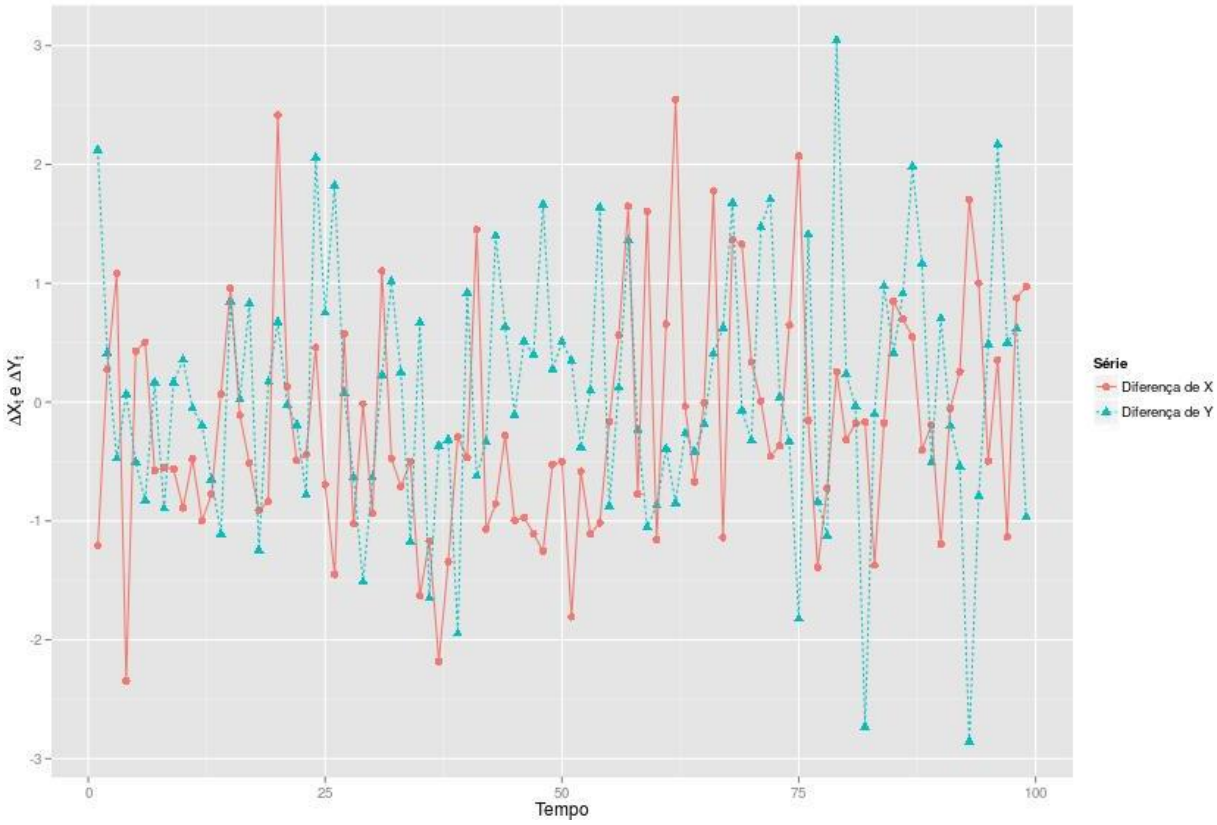
(Fonte: Próprio autor)

O valor de R^2 ajustado é significativamente diferente de zero (0.232) e a estimativa do parâmetro de X_t ($\widehat{\beta_2}$) também é estatisticamente significativa a qualquer nível de significância prático (repare como o valor-p é baixo, 2.34e-07). Sem conhecer o verdadeiro processo gerador dos dados, poderia-se acreditar que existe relação, ainda que fraca, entre as duas séries. A

regressão por si já não faria sentido, já que os dados foram gerados independentes propositalmente. Além disso, qualquer inferência deve ser feita com extremo cuidado.

Sabe-se que a não estacionariedade, responsável pela regressão espúria, pode ser removida de passeios aleatórios através de diferenciação. A Figura 3.4 mostra as duas séries após serem diferenciadas:

Figura 3.4 – Dois passeios aleatórios diferenciados



(Fonte: Próprio autor)

Como espera-se, nas séries diferenciadas, não se deve mais observar o fenômeno da regressão espúria já que agora serão analisadas duas séries estacionárias. A saída da regressão $\Delta Y_t = \gamma_1 + \gamma_2 \Delta X_t + w_{2t}$ evidencia isso:

Tabela 3.3 – Regressão entre dois passeios aleatórios diferenciados

Coeficiente	**Estimativa**	**Estatística** t	**Valor-p**
$\widehat{\gamma_1}$	0.06453	0.609	0.544
$\widehat{\gamma_2}$	-0.09959	-0.935	0.352

(Fonte: Próprio autor)

Quando a regressão ocorre entre duas séries estacionárias intencionalmente não relacionadas, o valor do R^2 ajustado, como esperado, é muito baixo (-0.001289).[5] Ademais, a estimativa do coeficiente de ΔX_t ($\widehat{\gamma_2}$) é estatisticamente não significativa (valor-p igual a 0.352) explicitando a inexistência de influência de X_t em Y_t. Esse exemplo, acima de tudo, deve deixar claro que deve-se sempre tomar muito cuidado ao realizar regressão com dados de séries temporais.

3.4.3 Testes de Raiz Unitária e Estacionariedade

Conhecendo o que é estacionariedade e como ela é realmente importante, resta saber: Como testamos se uma dada série é estacionária? Para responder essa questão, serão apresentados os fundamentos dos testes estatísticos mais utilizados. De qualquer forma, o ponto de partida é quase sempre o mesmo, o teste da raiz unitária.

Considere o modelo simples $Y_t = \rho Y_{t-1} + u_t$, com $-1 \leq \rho \leq 1$ sendo u_t ruído branco. Se $|\rho| = 1$, então tem-se o modelo do passeio aleatório que é sabidamente não estacionário, mais especificamente, é $I(1)$. Por outro lado, se $|\rho| < 1$, então tem-se uma série estacionária.[6] A ideia por trás do teste da raiz unitária é justamente descobrir se a estimativa do coeficiente de Y_{t-1}, ρ, é igual ou menor que 1.

A forma clássica de fazer a análise é subtraindo-se Y_{t-1} de ambos os lados da equação do modelo o que nos leva a

$$Y_t - Y_{t-1} = (\rho - 1)Y_{t-1} + u_t$$

$$\Delta Y_t = (\rho - 1)Y_{t-1} + u_t$$

$$\Delta Y_t = \delta Y_{t-1} + u_t$$

onde $\delta = \rho - 1$. Com essa nova equação, pode-se simplesmente fazer a regressão de Y_t contra Y_{t-1} e, mais importante, analisar a significância da estimativa do coeficiente de Y_{t-1}, $\hat{\delta}$. Se $\delta = 0$, então $\rho = 1$ e portanto tem-se uma série não estacionária. Se, por sua vez, $\rho \neq 1$, então o modelo é de uma série estacionária. No entanto, apesar de tentador, não se pode levar adiante um teste t já que sob a hipótese nula de que $\delta = 0$ a estatística t segue uma distribuição diferente. Graças a Dickey e Fuller, podemos continuar o nosso teste de hipóteses.

[5] A definição do R^2 ajustado permite resultados negativos. Em termos práticos, considera-se que seja igual a zero.

[6] O caso em que $|\rho| > 1$ leva a uma série explosiva que não será tratada aqui.

3.4.3.1 O Teste de Dickey-Fuller

O Teste de Dickey-Fuller (DICKEY; FULLER, 1979 apud GUJARATI, 2009) utiliza a estatística τ para descobrir sobre a existência de raiz unitária em uma série.[7] Essa estatística possui valores críticos diferentes dependendo do modelo em questão que pode ser

$$\Delta Y_t = \delta Y_{t-1} + u_t, \text{ quando } Y_t \text{ é um passeio aleatório.}$$

$$\Delta Y_t = \beta_1 + \delta Y_{t-1} + u_t, \text{ quando } Y_t \text{ é um passeio aleatório com drift.}$$

$\Delta Y_t = \beta_1 + \beta_2 t + \delta Y_{t-1} + u_t$, quando Y_t é um passeio aleatório com drift e tendência determinística.

Não importando qual seja o modelo em estudo, o teste segue sempre as mesmas hipóteses:

$$\begin{cases} H_0 : \delta = 0 \\ H_1 : \delta < 0 \end{cases}$$

A hipótese nula representa a hipótese da existência da raiz unitária, ou seja, não estacionariedade. A hipótese alternativa significa a estacionariedade da série. Note como trata-se de um teste unilateral. Sob a hipótese de não estacionariedade da série, temos a estatística τ de Dickey-Fuller no lugar da estatística t de Student clássica. No entanto, quando a rejeita-se a hipótese nula, a estatística t volta a tornar-se válida para nossas inferências.

Como no caso da computação da estatística t, deve-se achar o valor τ dividindo-se a estimativa do parâmetro de Y_{t-1}, $\hat{\delta}$, encontrado na regressão pela estimativa do seu erro-padrão. Esse valor deve então ser comparado com os valores críticos da estatística τ para determinado nível de significância e tamanho da amostra. Como já dito, os valores críticos de τ dependem da suposição do modelo.

O teste de Dickey-Fuller compara então o valor encontrado para τ de $\hat{\rho}$ com o valor crítico. De forma geral, se $|\tau| > |\tau_{critico}|$, então deve-se rejeitar a hipótese nula de não estacionariedade em favor da hipótese alternativa de estacionariedade. Em suma, esperamos valores bastante negativos para τ se queremos observar uma série estacionária. Quando realizando o teste de Dickey-Fuller, é razoável supor cada modelo apresentado anteriormente e analisar o resultado das regressões individuais. É possível, por exemplo, que algum modelo resulte em um valor de τ positivo o que implicaria que a série é explosiva ($\rho > 1$) e fazendo com que esse

[7] Mais formalmente deveria ser dito que trabalha-se com a estimativa de τ, ou seja, $\hat{\tau}$. De qualquer forma, τ continuará a ser usada.

modelo fosse descartado.

3.4.3.2 O Teste de Dickey-Fuller Aumentado

Como dito anteriormente, supõe-se que o termo de erro do modelo, u_t carregue a informação de todas variáveis explanatórias que não foram explicitamente postas na equação. Porém a especificação correta do modelo faz necessário que tais variáveis sejam explícitas para garantir a suposição de não correlação entre os termos de erro. O teste de Dickey-Fuller também baseia-se na suposição da não correlação entre os termos de erro que, no entanto, pode eventualmente existir. Para levar em conta isso, foi proposto o teste de Dickey-fuller Aumentado.

A mecânica do teste aumentado é repetir o teste de Dickey-Fuller já apresentado, mas agora adicionando-se também alguns termos defasados de ΔY_t, ou seja, ΔY_{t-1}, ΔY_{t-2}, ΔY_{t-3} e assim por diante. A ideia por trás disso é explicitar alguns termos defasados (removendo-os do termo de erro) já que eles também podem ser necessários para explicar a variável de saída no modelo. O parâmetro de interesse continua sendo δ. É possível decidir qual a extensão da defasagem a ser utilizada no teste através de alguns dos critérios de seleção de modelos já apresentados como, por exemplo, Akaike.

O modelo a ser estimado para passeio aleatório com *drift* e tendência determinística contando com m defasagens seria então:

$$\Delta Y_t = \beta_1 + \beta_2 t + \delta Y_{t-1} + \sum_{i=1}^{m} \alpha_i \Delta Y_{t-i} + \epsilon_t.$$

Note como foi utilizado ϵ_t no lugar de u_t para evidenciar que os termos de erro dos dois modelos diferem pelas variáveis defasadas. O teste segue o mesmo procedimento apresentado para o teste de Dickey-Fuller.

3.4.3.3 O Teste de Phillips-Perron

O teste de Phillips-Perron (PHILLIPS; PERRON, 1988 apud GUJARATI, 2009) também aborda o problema da autocorrelação e também lida com heterocedasticidade nos termos de erro. Diferentemente do teste de Dickey-Fuller Aumentado que adiciona termos defasados à equação, o teste de Phillips-Perron faz correções na estatística $\hat{\tau}$ de Dickey-Fuller através de métodos estatísticos não paramétricos.

A distribuição assintótica da estatística $\hat{\tau}$ de Phillips-Perron é a mesma que aquela sem tais correções não paramétricas de Dickey-Fuller. Outro ponto importante é que o teste de Phillips-Perron aparenta ter desempenho inferior ao teste de Dickey-Fuller Aumentado em

amostras pequenas (DAVIDSON; MACKINNON, 2004).

3.4.4 Função de Autocorrelação

Quando tratamos de dados de séries temporais, algumas medidas adicionais aparecem e mostram-se bastante úteis na identificação de propriedades como estacionariedade e ordem de modelos autorregressivos e de médias móveis. Esses últimos são o papel central dos modelos ARIMA e da modelagem Box-Jenkins que não trataremos a fundo aqui. Nosso foco será no uso dessas medidas na identificação da estacionariedade.

A função de autocorrelação - FAC é simplesmente a razão da covariância com defasagem k e a variância da série:

$$\rho_k = \frac{\gamma_k}{\gamma_0},$$

onde γ_k foi definido quando apresentamos estacionariedade fraca. Note que $\rho_0 = 1$. Uma das ferramentas para se visualizar e trabalhar com a FAC é o correlograma. O correlograma é o gráfico de ρ_k contra a defasagem k.

No entanto, na prática dispomos apenas de uma das realizações possíveis para o processo estocástico. Em outras palavras, quando analisamos uma série, dispomos apenas de uma amostra. A função de autocorrelação resultante é devidamente chamada de função de autocorrelação amostral, $\hat{\rho_k}$. Como trataremos apenas de séries temporais observadas, ou seja, amostras, daqui em diante deixaremos o formalismo e chamaremos a função de autocorrelação amostral apenas de função de autocorrelação - FAC.

$$\hat{\rho_k} = \frac{\hat{\gamma_k}}{\hat{\gamma_0}},$$

onde

$$\hat{\gamma_k} = \frac{\sum(Y_t - \bar{Y})(Y_{t+k} - \bar{Y})}{n-k} \text{ e}$$

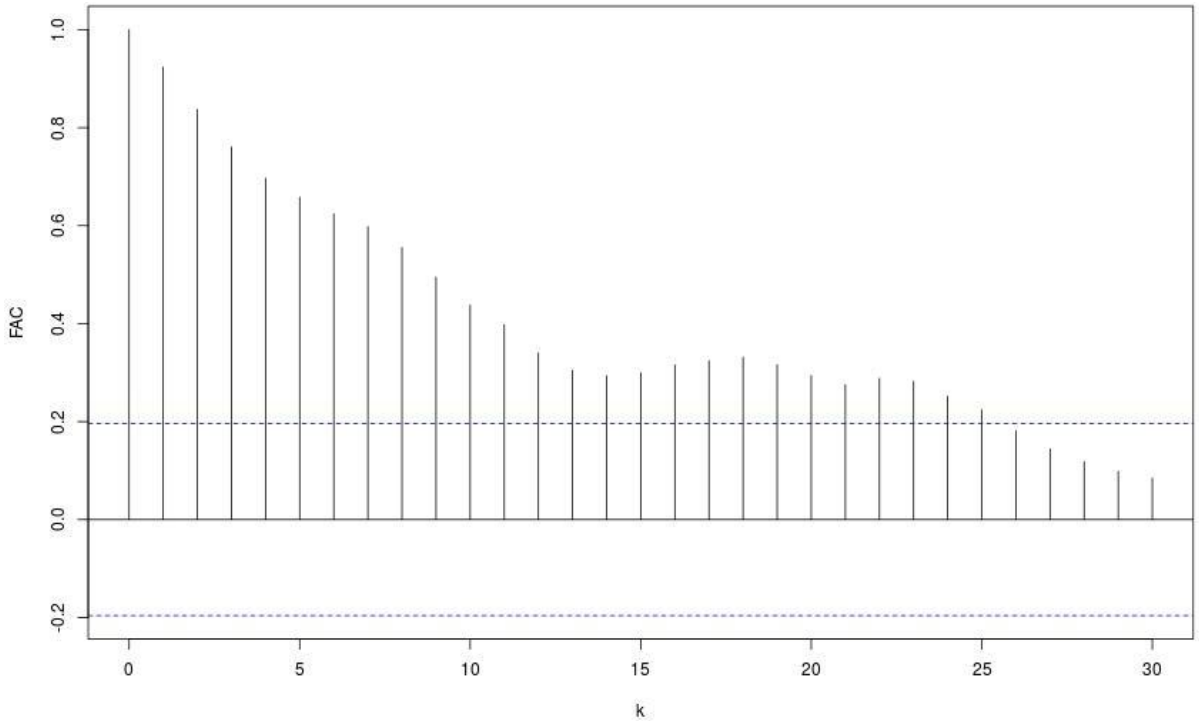
$$\hat{\gamma_0} = \frac{\sum(Y_t - \bar{Y})^2}{n-1}.$$

Novamente, o correlograma que obteremos na análise será o correlograma amostral, o gráfico de $\hat{\rho_k}$ contra k. Assim como faremos para a função de autocorrelação amostral, chamaremos o correlograma amostral apenas de correlograma.

Para exemplificar como o correlograma pode nos ajudar a identificar se uma série é ou não estacionária, usaremos as mesmas séries do exemplo de regressão espúria mostradas anteriormente. Naquele exemplo, a duas séries, Y_t e X_t, eram originalmente passeios aleatórios, ou seja, não estacionárias. No entanto, suas primeiras diferenças, ΔY_t e ΔX_t, eram estacionárias. Em outras palavras, por construção, Y_t, $X_t \sim I(1)$.

O correlograma de uma série não estacionária (aqui Y_t) tem uma forma como a mostrada a seguir

Figura 3.5 – Função de autocorrelação amostral de uma série não estacionária

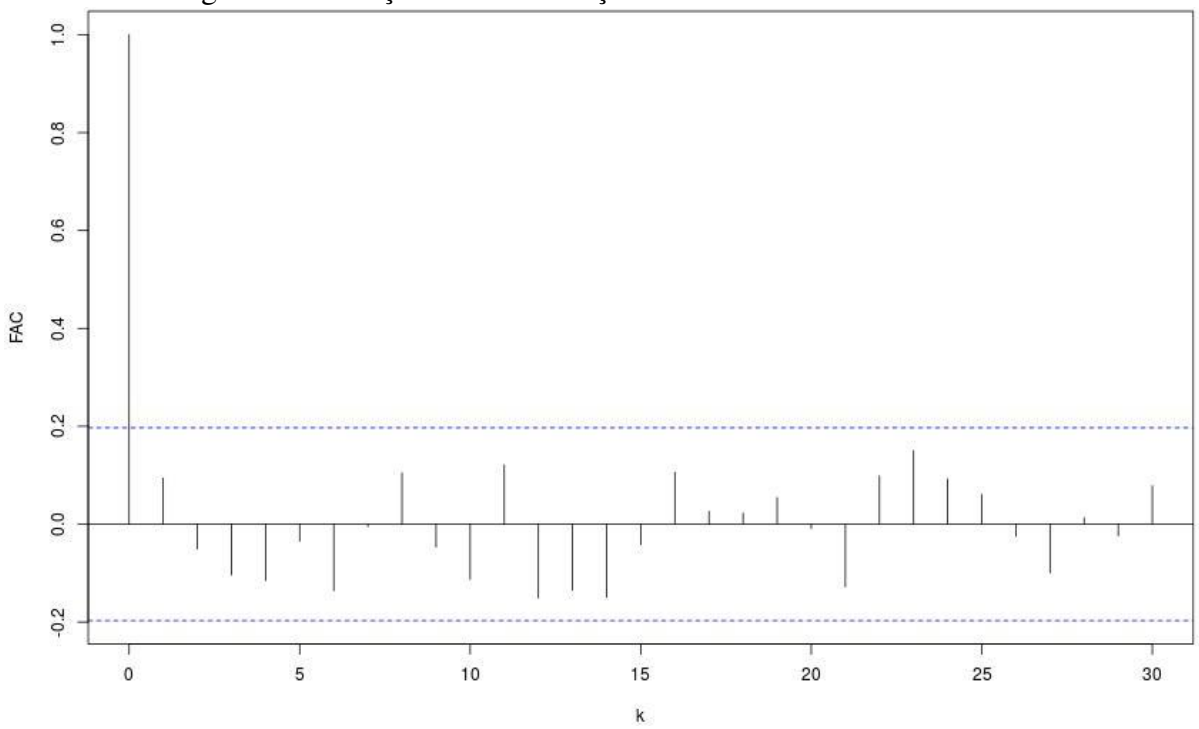


(Fonte: Próprio autor)

Note como o decaimento da função de autocorrelação até a defasagem $k = 30$ é extremamente lenta. Isso significa que observações dessa série, ainda que muito afastadas temporalmente, ainda mantêm correlação. Dois pontos merecem observação. O primeiro é como a FAC com $k = 0$ realmente tem o valor 1 como antes enunciado. O segundo são as linhas horizontais em $FAC = 0.2$ e $FAC = -0.2$. Elas denotam o intervalo em que considera-se $\hat{\rho}_k = 0$. Uma aproximação empírica para a definição desse intervalo comumente utilizada é $\pm 1.96/\sqrt{n} \approx \pm 2/\sqrt{n}$.

Por outro lado, a primeira diferença é sabidamente estacionária. Seu correlagrama (3.6) mostra um cenário bem diferente do anterior. Agora, a função de autocorrelação decai muito mais depressa comparado ao exemplo acima.

Figura 3.6 – Função de autocorrelação amostral de uma série estacionária



(Fonte: Próprio autor)

Observando o correlograma é possível distinguir, em muitos casos, uma série não estacionária de uma série estacionária. O correlograma também é uma ferramente poderosa na identificação da ordem de modelos autorregressivos e de médias móveis usados na clássica modelagem Box-Jenkins.

Acima foi dito que a FAC decai com a distância temporal das observações, mais lentamente em séries não estacionárias e mais rapidamente em séries estacionárias. Agora é preciso descobrir como decidir se um dado coeficiente $\hat{\rho}_k$ para um determinado k é estatisticamente diferente de zero ou não.

Alguns testes foram desenvolvidos e dois deles tornaram-se bastante populares. Em ordem cronológica, o primeiro deles é o Teste de Box-Pierce (BOX; PIERCE, 1970 apud GUJARATI, 2009). O segundo, o Teste de Ljung-Box (LJUNG; BOX, 1978 apud GUJARATI, 2009), é na verdade um aprimoramento do primeiro e é a ferramenta mais utilizada nessa tarefa.

O teste de Box-Pierce testa se todos coeficientes de autocorrelação são simultaneamente iguais a zero e tem a seguinte configuração (BUENO, 2008)

$$\begin{cases} H_0 : \rho_0 + \rho_1 + \cdots + \rho_m = 0 \\ H_1 : \rho_0 + \rho_1 + \cdots + \rho_m \neq 0 \end{cases}$$

onde m é a defasagem máxima que estamos testando. Sob a hipótese nula, pode-se

utilizar a estatística Q de Box-Pierce que é definida como

$$Q = n \sum_{k=1}^{m} \hat{\rho_k}^2,$$

onde n é o tamanho da amostra. Em grandes amostras, a estatística Q computada acima segue aproximadamente a distribuição χ^2 com m graus de liberdade. Matematicamente, $Q \sim \chi^2_m$, para n grande. Se Q ultrapassar o valor crítico da distribuição χ^2 no nível de significância especificado, então pode-se rejeitar a hipótese nula de que todos coeficientes de autocorrelação da série são ao mesmo tempo iguais a zero. Do contrário, tem-se que ao menos um dos coeficientes deve ser diferente de zero.

O teste de Ljung-Box é uma variante do teste de Box-Pierce seguindo as mesmas hipóteses anteriores. No entanto, agora computamos a estatística LB da seguinte forma

$$LB = n(n+2) \sum_{k=1}^{m} \left(\frac{\hat{\rho_k}^2}{n-k} \right),$$

que também segue uma distribuição χ^2_m para n grande. O teste de Ljung-Box possui desempenho melhor em pequenas amostras do que o teste de Box-Pierce. No presente trabalho, utilizaremos ambos quando o objetivo for analisar a significância dos coeficientes de autocorrelação.

3.4.5 Remoção de não estacionariedade

Os perigos de trabalhar com séries não estacionárias, principalmente quando tratamos de regressão que pode levar ao fenômeno da regressão espúria, já foram apresentados e resta saber o que fazer quando dispomos desse tipo de série. É interessante que disponha-se de métodos adequados para a transformação de uma série não estacionária em uma série estacionária. Tais métodos dependem se a série em questão pode ser tornada estacionária por diferenças ou se possui tendência estacionária.

Quando uma série possui raízes unitárias, ela pode tornar-se estacionária tomando-se suas primeiras diferenças. Foi exatamente esse o processo mostrado na seção que tratamos de regressão espúria. A série original era não estacionária com apenas uma raiz unitária. Sua primeira diferença mostrou-se no entanto uma série estacionária.

Em geral, se uma série Y_t possui d raízes unitárias, basta diferenciá-la d vezes para torná-la estacionária. Tal série é dita integrada de ordem d, ou seja, $Y_t \sim I(d)$. A importância

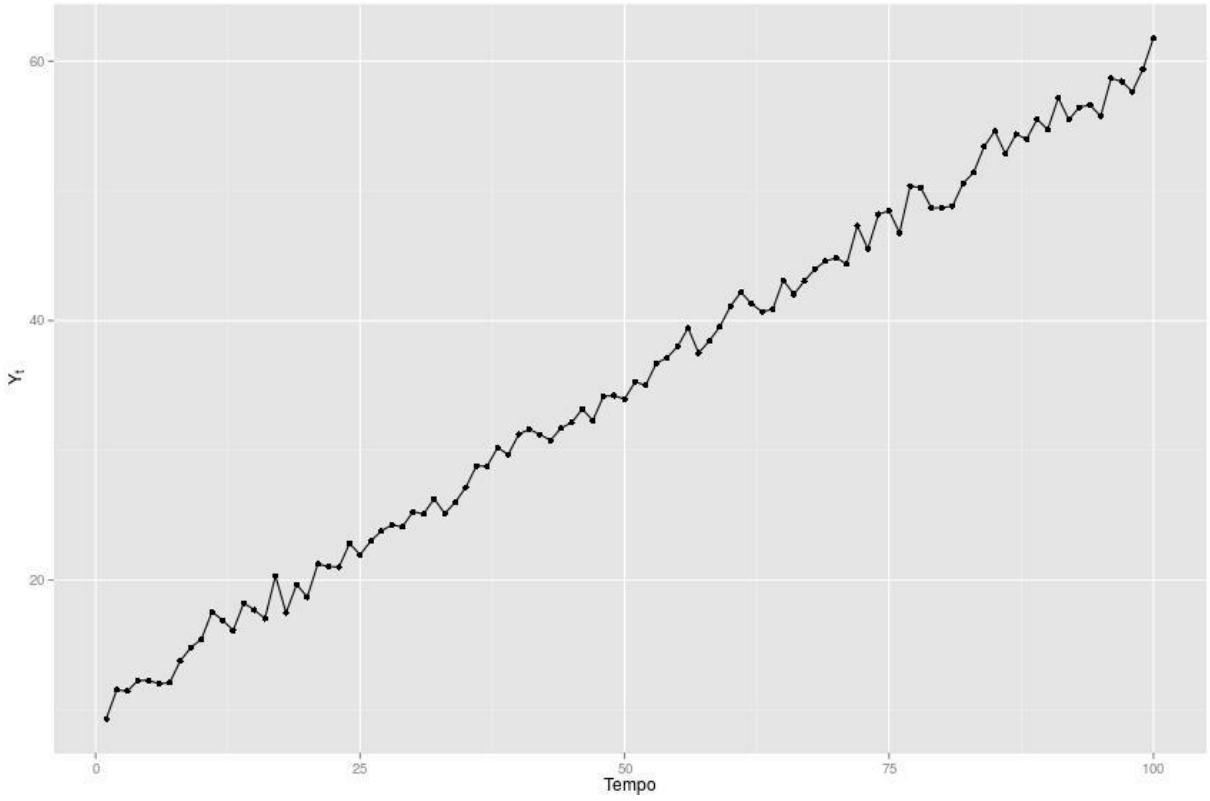
dessa classificação tornará-se mais evidente quando falarmos de cointegração.

Por outro lado, se uma série possui estacionariedade em torno de uma tendência determinística, então os resíduos da regressão da série pela tendência serão estacionários. Para ilustrar isso, considere a série Y_t com a seguinte forma

$$Y_t = \beta_1 + \beta_2 t + u_t.$$

Nota-se a presença de um termo t relativo ao tempo das observações. Novamente, recorrendo aos recursos computacionais, podemos criar uma série com o modelo acima com $\beta_1 = 10$, $\beta_2 = 0.5$ e $u_t \sim N(0, 1)$ para 100 observações. O gráfico é mostrado a seguir

Figura 3.7 – Exemplo de série com tendência determinística



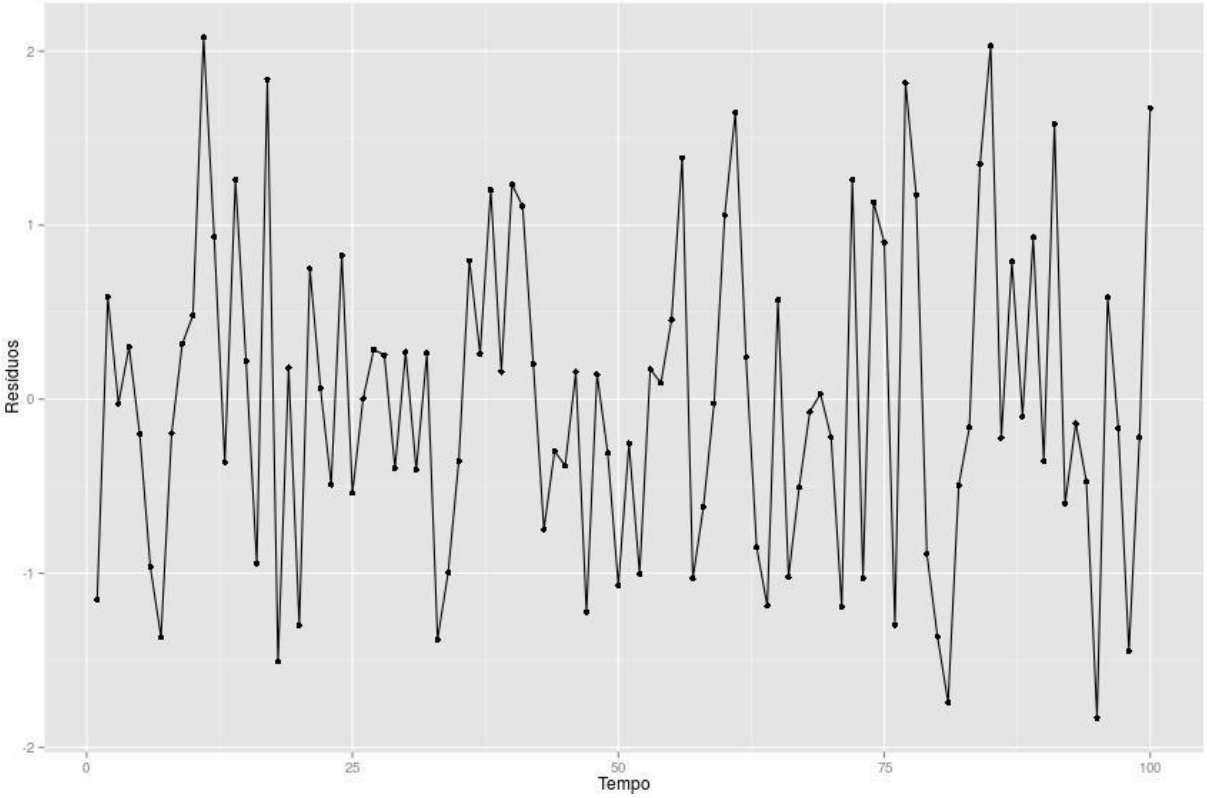
(Fonte: Próprio autor)

Existe uma clara tendência de subida pela análise do gráfico. A série estacionária pode ser obtida através da regressão de Y_t pela tendência e ficando com os resíduos

$$u_t = Y_t - \beta_2 - \beta_1 t.$$

Um rápido teste de Dickey-Fuller sugere fortemente que os resíduos são estacionários (valor-p inferior a 0.01). O gráfico dos resíduos é mostrado abaixo

Figura 3.8 – Resíduos da regressão da série com tendência determinística

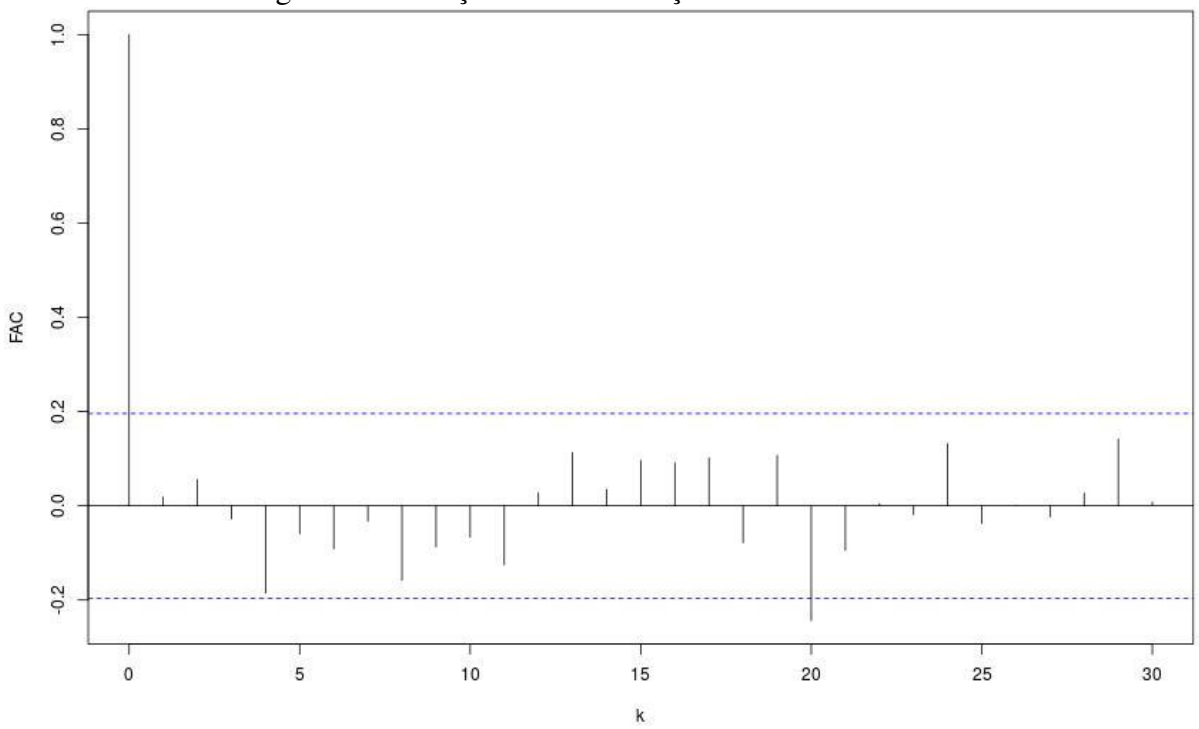


(Fonte: Próprio autor)

A função de autocorrelação amostral da 3.9 sugere que os resíduos podem ser estacionários. Assim, u_t é chamada série temporal retificada. É possível ainda que Y_t tenha tendência de formas mais complicadas, por exemplo, $Y_t = \beta_1 + \beta_2 t + \beta_3 t^2 + u_t$, nas quais procede-se da mesma maneira.

É importante notar que o tipo de eliminação da não estacionariedade da série a ser realizada depende de sua natureza. Aplicar um método quando a origem da não estacionariedade é outra pode levar a sérios problemas de erro de especificação. Como já dito, a diferenciação é capaz de remover a tendência determinística, mas o custo disso é a introdução de ruído na forma de uma variância bem maior nos resíduos.

Figura 3.9 – Função de autocorrelação amostral dos resíduos



(Fonte: Próprio autor)

3.4.6 Cointegração

É possível que duas séries, mesmo que ambas não estacionárias, compartilhem de uma tendência em comum de forma que a regressão de um pela outra não resulte em uma regressão espúria. No entanto, elas devem satisfazer algumas propriedades para que isso aconteça.

Basicamente, (1) ambas séries devem possuir a mesma ordem de integração e (2) os termos de erro da regressão de uma pela outra, u_t devem ser estacionários, ou seja, $u_t \sim I(0)$. A intuição por trás disto é que, apesar das séries não serem estacionárias por si só, a combinação linear delas, na forma dos resíduos da regressão, é estacionária. Quando tal situação ocorre, dizemos que as séries são cointegradas. Assim, a metodologia tradicional de análise de regressão que foi mostrada até então pode continuar valendo ainda que as séries sejam não estacionárias desde que os resíduos sejam $I(0)$.

É preciso então testar a estacionariedade dos resíduos da regressão. Pode-se imaginar que os testes de Dickey-Fuller ou Dicley-Fuller Aumentado podem resolver a questão. No entanto, como os resíduos da regressão são baseados nas estimativas dos parâmetros do modelo, os valores críticos dos testes de Dickey-Fuller e Dickey-Fuller Aumentado são ligeiramente diferentes. Engle e Granger (ENGLE; GRANGER, 1987 apud GUJARATI, 2009) recalcularam esses valores gerando o que conhece-se por teste de Engler-Granger e Engler-Granger Aumen-

tado. Esses testes já estão disponíveis na maior parte dos softwares estatísticos, em especial, no *R*.

Pondo de forma mais clara, quando trabalhando com séries univariadas, podemos testar por estacionariedade utilizando os testes de Dickey-Fuller. No entanto, quando tratamos da relação entre várias séries, onde cada um possui raízes unitárias, a análise de estacionariedade toma a forma de teste de cointegração através de Engle-Granger e Engle-Granger Aumentado. Por fim, é importante notar que as séries podem possuir dois tipos de tendência, estocástica e determinística, e isso deve ser levado em consideração no modelo quando analisando os resíduos.

4 TRABALHOS ANTERIORES

Dois trabalhos recentes trataram da avaliação da lei de Metcalfe e demais modelos utilizando dados reais. O próprio Robert Metcalfe tentou validar seu modelo através de dados do Facebook na ocasião do 40º aniversário da Ethernet (METCALFE, 2013). Mais recentemente, (ZHANG; LIU; XU, 2015) analisaram empiricamente os modelos de Metcalfe, Sarnoff, Reed e Odlyzko com base no Facebook e na Tencent. Ambos trabalhos concluíram que a lei de Metcalfe é um bom modelo para quantificar o efeito de rede. Porém, alguns aspectos, principalmente de caráter estatístico, poderiam ser melhorados.

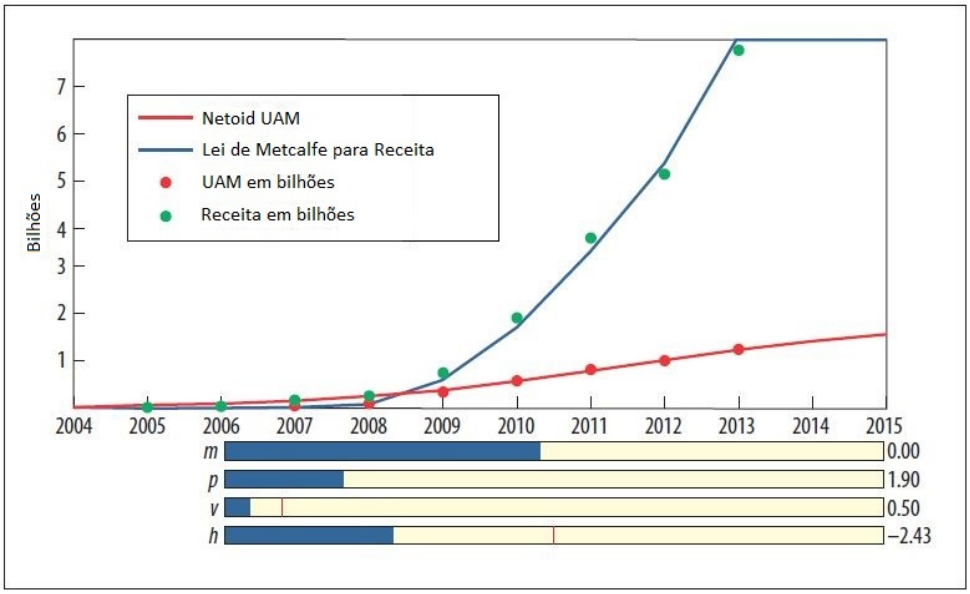
Em uma edição da IEEE Spectrum, Robert Metcalfe fez uma retrospectiva da evolução da Ethernet desde a sua criação em um memorando seu que circulou pelo Xerox PARC até o ano de 2013. Em seu artigo, Metcalfe não se limita à descrever a jornada da Ethernet até os dias, mas também discute o impacto que sua invenção causou na tecnologia e o papel do seu modelo sobre o valor da rede na sua adoção. Sem dúvida, a Ethernet foi um passo de grande importância nas telecomunicações, crescendo além de todas expectativas e tornando-se um dos padrões mais conhecidos das comunicações digitais.

Metcalfe utiliza dados de dez anos de Facebook para verificar a adequação dos modelos para o valor da rede e crescimento do número de usuários. O primeiro modelo é a clássica lei de Metcalfe, $V \sim N^2$. O segundo modelo é sobre o crescimento do número de usuários seguindo uma função sigmoide. A receita é utilizada como *proxy* para o valor agregado da rede, enquanto o número médio de usuários ativos por mês (UAM) é utilizado como *proxy* para o tamanho da rede. A amostra soma 10 observações compreendidas entre 2004 e 2013, inclusive. Certamente, essa amostra é extremamente pequena para tentar fazer qualquer tipo de inferência.

A forma como os coeficientes dos modelos são estimados não é a mais precisa. Após os dados serem plotados, os parâmetros das curvas da lei de Metcalfe e Netoid são ajustados até obter-se boa aproximação visual. Não são dados muitos detalhes a respeito de como a análise é feita além de que são utilizados *sliders* da linguagem Python para ajustar os parâmetros dos modelos. Obviamente, "boa aproximação visual" é um termo subjetivo e pouco reproduzível. O gráfico resultante encontra-se na Figura 4.1.

As aproximações de Metcalfe são visualmente atrativas e parecem sugerir que seus modelos realmente encaixam-se nos dados. No entanto é importante notar que estimativas baseadas tão somente na sua qualidade visual não são a melhor abordagem para o problema. O que Metcalfe considerou como os melhores parâmetros podem não ser os melhores na ótica de outra pessoa e provavelmente também não serão os melhores segundo algum critério estatístico (por

Figura 4.1 – Avaliação empírica de Metcalfe



(Fonte: (METCALFE, 2013) modificado.)

exemplo, estimadores de MQO). Além disso, seguindo uma metodologia de análise mais rigorosa, deveria ser considerado que os dados constituem séries temporais dando-se o tratamento devido.

A conclusão de Metcalfe é que seus modelos são capazes de adequar-se bem aos dados observados. Apesar de carecer de algum cuidado estatístico, o artigo de Metcalfe serviu bem como uma primeira abordagem empírica do problema e, de certa forma, instigou mais investigações na mesma linha, incluindo o presente trabalho.

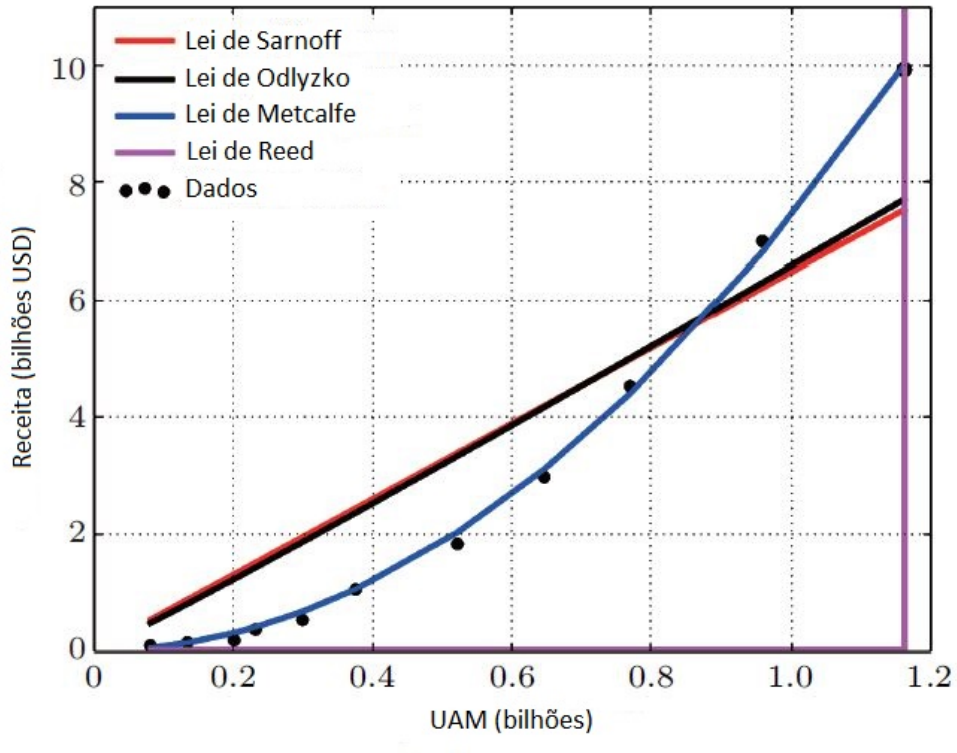
Em 2015, (ZHANG; LIU; XU, 2015) publicaram um artigo que diz validar a Lei de Metcalfe utilizando dados do Facebook e da Tencent. Os pesquisadores comparam os modelos de Metcalfe, Sarnoff, Odlyzko e Reed para descobrir qual deles melhor explica o crescimento do valor dessas redes em função do seu número de usuários e analisam o modelo de crescimento Netoid de usuários de Metcalfe. Novamente, a receita é utilizada como variável *proxy* para o valor da rede e o número médio de usuários ativos mensalmente serve como *proxy* para o tamanho das redes. Dessa vez, o custo da rede também é quantificado e modelado. A diferença entre a receita e o lucro líquido é utilizada como variável *proxy* para o custo. Aqui, concentraremo-nos na análise das leis de crescimento da rede.

Os dados são obtidos através dos relatórios financeiros das empresas. Foram usadas 11 observações para cada uma das empresas. Para o Facebook, os dados de UAM (Usuários Ativo Mensalmente), receita e custo abrangem o período de 2004 a 2014, enquanto para a Tencent o intervalo é de 2003 a 2013.

O método dos Mínimos Quadrados é empregado na regressão dos dados através da fun-

ção *leastsq* disponível no pacote *SciPy* da linguagem Python. Após estimar cada um dos quatro modelos por MQO, é utilizada a raiz quadrada do erro quadrático médio (*Root-Mean-Square Deviations* - RMSD) para escolher aquele que melhor explica os dados. Os gráficos dos dados reais e das regressões do valor contra o tamanho das redes para o Facebook e para a Tencent são mostrados Figuras 4.2 e 4.3, respectivamente.

Figura 4.2 – Avaliação empírica de Zhang et al. - Facebook

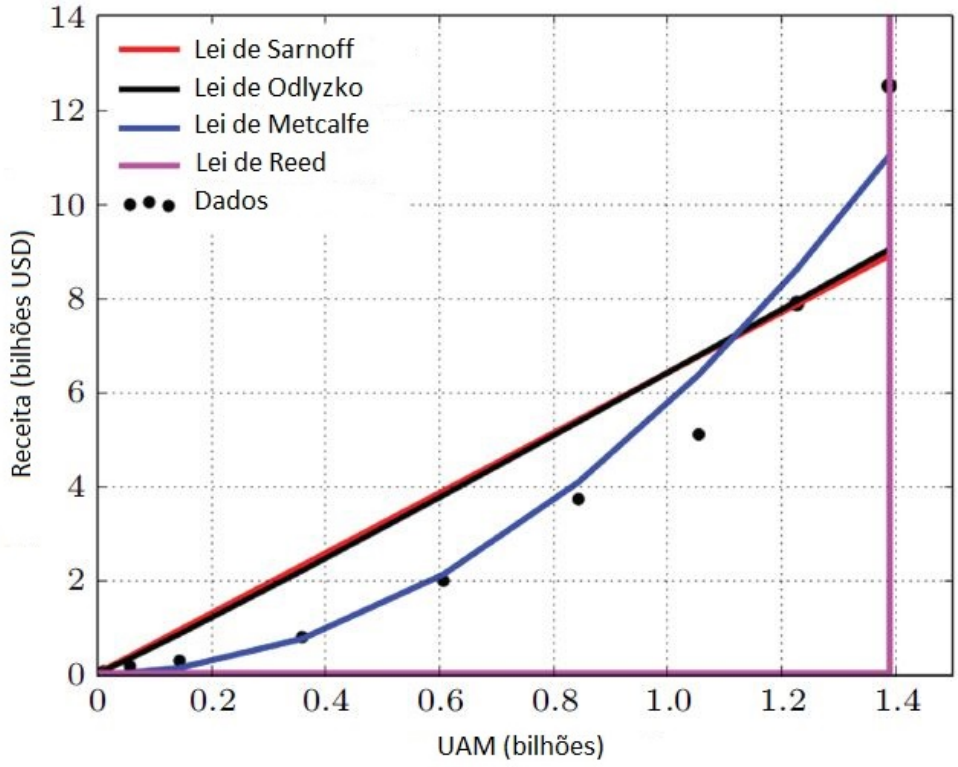


(Fonte: (ZHANG; LIU; XU, 2015) modificado.)

Visualmente, a lei de Metcalfe aparenta ser o melhor modelo, principalmente para a série do Facebook. As leis de Sarnoff e Odlyzko têm comportamentos bem semelhantes para os dados considerados. Por fim, a lei de Reed passa longe dos dados reais. A tabela resumindo as estimativas de MQO para cada modelo e seus respectivos RMSD é apresentada na Figura 4.4.

Como pode-se observar, tanto para os dados da Tencent quanto para os dados do Facebook, a lei de Metcalfe é a que apresenta o menor valor de RMSD (0.12 e 0.64, respectivamente). Com isso, (ZHANG; LIU; XU, 2015) afirmam que a lei de Metcalfe é o melhor dentre os modelos concorrentes apresentados para explicar o crescimento do valor da rede em relação ao seu número de usuários.

Figura 4.3 – Avaliação empírica de Zhang et al. - Tencent



(Fonte: (ZHANG; LIU; XU, 2015) modificado.)

Figura 4.4 – Resumo das regressões de Zhang et al.

	Dados da Tencent		Dados do Facebook	
	Funções do Valor	RMSDs	Funções do Valor	RMSDs
Lei de Sarnoff	$V_{\text{Tencent}} = 6.46n$	1.27	$V_{\text{Facebook}} = 6.39n$	1.51
Lei de Odlyzko	$V_{\text{Tencent}} = 0.22 \times n\log(n)$	1.19	$V_{\text{Facebook}} = 0.21 \times n\log(n)$	1.45
Lei de Metcalfe	$V_{\text{Tencent}} = 7.39 \times 10^{-9} \times n^2$	0.12	$V_{\text{Facebook}} = 5.70 \times 10^{-9} \times n^2$	0.64
Lei de Reed	$V_{\text{Tencent}} = 2^{-1.16\times10^{9}} \times (2^n - 1)$	4.06	$V_{\text{Facebook}} = 2^{-1.39\times10^{9}} \times (2^n - 1)$	4.88

(Fonte: (ZHANG; LIU; XU, 2015) modificado.)

Apesar do apelo de ser uma forma simples de avaliar quantitativamente as leis, a análise de regressão desenvolvida precisa de alguns cuidados. Primeiramente, pouquíssimos dados são utilizados na regressão, o que implica que a inferência é menos confiável. Por se tratar de dados de séries temporais, seria necessário avaliar a estacionariedade das séries antes de efetuar a regressão. Como foi apresentado, regressões com séries não estacionárias podem levar a resultados espúrios. Por fim, nenhuma investigação pós-regressão, como teste de homocedasticidade, foi feita.

5 EXPERIMENTOS

5.1 Dados

Os dados que serão utilizados na análise dos modelos foram obtidos dos relatórios aos investidores da Tencent[1], uma empresa de grande porte do setor de mídias digitais da China. Como a empresa é negociada em bolsa de valores (SEHK), algumas informações, principalmente as de cunho contábil, devem ser mantidas abertas para conhecimento dos investidores. Nesses relatórios, é possível conseguir informações como o número de usuários únicos ativos mensalmente[2], receita e custos para um período de três meses. A princípio, considerou-se utilizar também os dados do Facebook[3], mas como o número de observações, até o presente momento, é pequeno, optou-se por não analisá-los.

Como já discutido, é muito difícil mensurar valor, tanto agregado quanto individual, de uma rede. Na impossibilidade ou nos casos de grande dificuldade de observar diretamente dados para uma determinada variável do modelo, recorremos às chamadas variáveis *proxy*. Variáveis *proxy* são aquelas que transmitem significado muito próximo ao da variável do modelo e portanto servem como substitutas razoáveis durante a análise empírica. Neste estudo, a receita total será utilizada como *proxy* para o valor das redes, V.

Será utilizada a média do número de usuários ativos mensalmente observados ao longo de três meses do serviço de mensagem instantânea (IM) como variável *proxy* para o tamanho da rede, N. É importante notar que o número total de usuários ativos da Tencent é maior do que o número utilizado aqui haja visto que a empresa fornece diversos outros serviços.

Os dados para a Tencent iniciam-se no segundo trimestre de 2004 e, até o momento (segundo trimestre de 2016), soma 48 observações. É importante notar que esse número ainda é pequeno para uma análise apropriada de séries temporais, mas ainda assim é uma das melhores fontes de dados disponíveis. Muitos dos testes estatísticos que serão feitos possuem propriedades assintóticas, ou seja, propriedades que valem quando o número de observações na amostra é considerado grande. Esse não é o caso aqui, então é importante ressaltar que os resultados dos testes podem não ser conclusivos e devem ser interpretados com cautela.

Para fazer a análise, será utilizado o *software* estatístico *R* (R Core Team, 2015). O *R* é um ambiente de *software* aberto para estatística computacional. A variedade e qualidade de funcionalidades disponibilizadas pelo *R* na forma de pacotes foram uns dos principais fatores

[1] http://www.tencent.com/en-us/index.shtml.
[2] Do inglês, *Monthly Active Users* - MAUs.
[3] https://www.facebook.com.

na sua escolha. Outros bons motivos para utilizá-lo são código aberto, extensa literatura sobre o uso e comunidade ativa. Alguns pacotes auxiliarão a análise e a visualização dos dados: *tseries* (TRAPLETTI; HORNIK, 2016), *urca* (PFAFF, 2008), *ggplot2* (WICKHAM, 2009) e *reshape2* (WICKHAM, 2007).

Algumas estatísticas sobre a amostra são mostradas abaixo

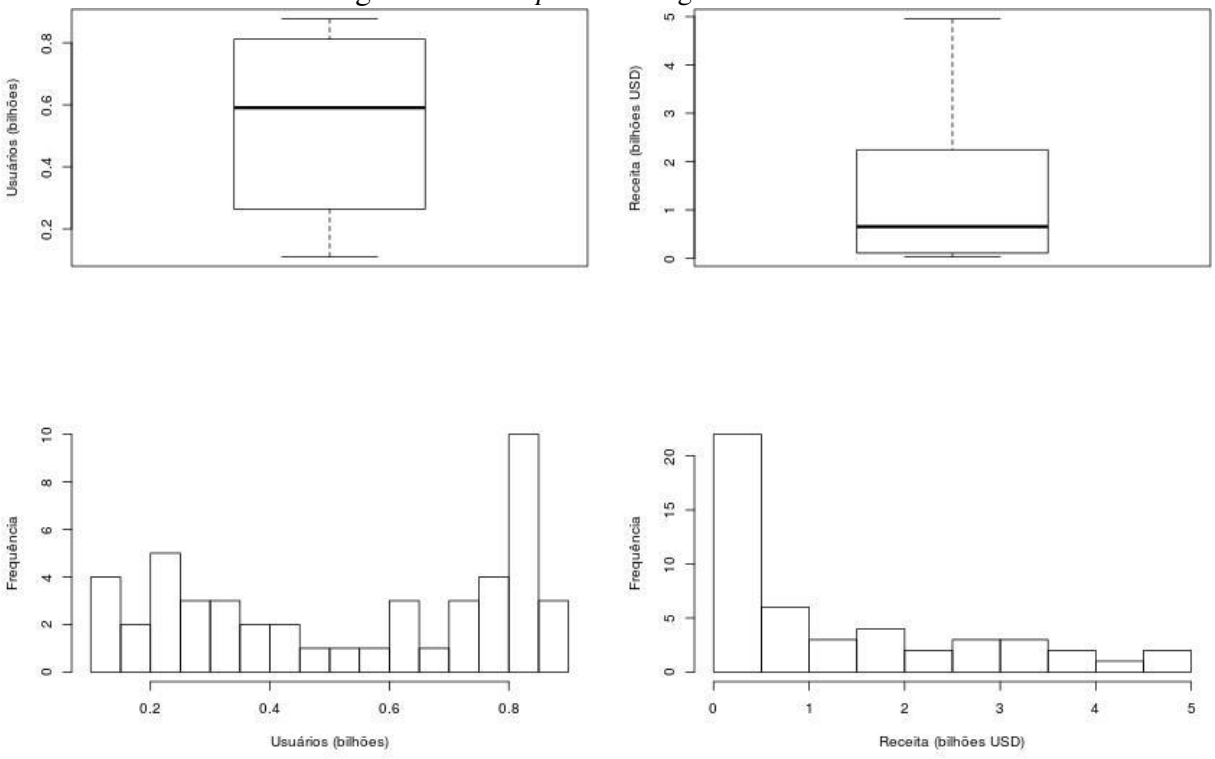
Tabela 5.1 – Estatísticas sobre os dados

Série	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Usuários	0.1101	0.2683	0.5906	0.5330	0.8097	0.8770
Receita	0.0327	0.1105	0.6533	1.2974	2.2029	4.9520

(Fonte: Próprio autor)

Note que *usuários* está em bilhões e *receita* está em bilhões de dólares americanos (USD). Observa-se o grande crescimento tanto na base de usuários da rede quanto no valor de sua receita ao longo dos doze anos de observação. As mesmas informações a respeito dos dados são apresentadas visualmente abaixo na forma de *boxplots*. Os histogramas mostrados na parte inferior da figura abaixo revelam um grande assimetria em ambos conjuntos de dados. Os dados de usuários parecem ter assimetria à esquerda, enquanto a receita exibe assimetria à direita.

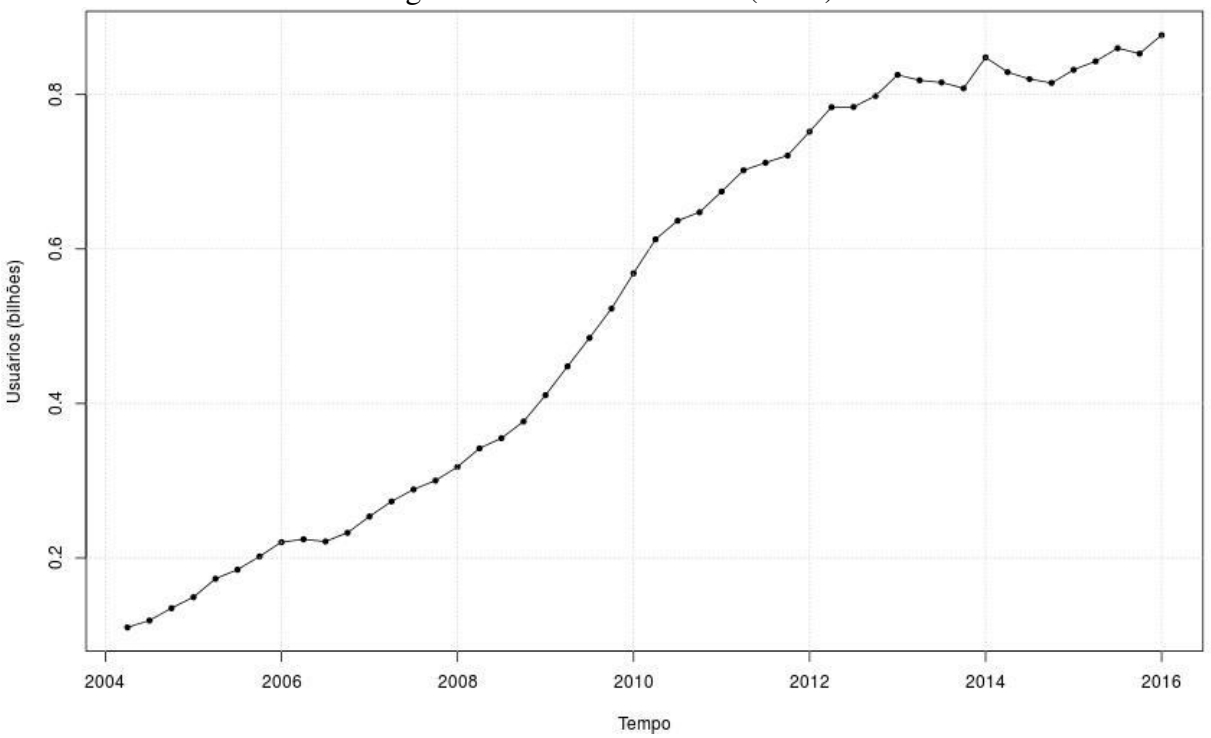
Figura 5.1 – *Boxplots* e histogramas dos dados



(Fonte: Próprio autor)

No entanto, a melhor forma de visualizar esses dados é através de um *scatterplot* como mostrado abaixo. A Figura 5.2 mostra a evolução do número de usuários ativos mensalmente ao longo de doze anos. Nota-se um crescimento acentuado até 2012 e um crescimento mais modesto a partir de então até 2016.

Figura 5.2 – Série de usuários (UAM)

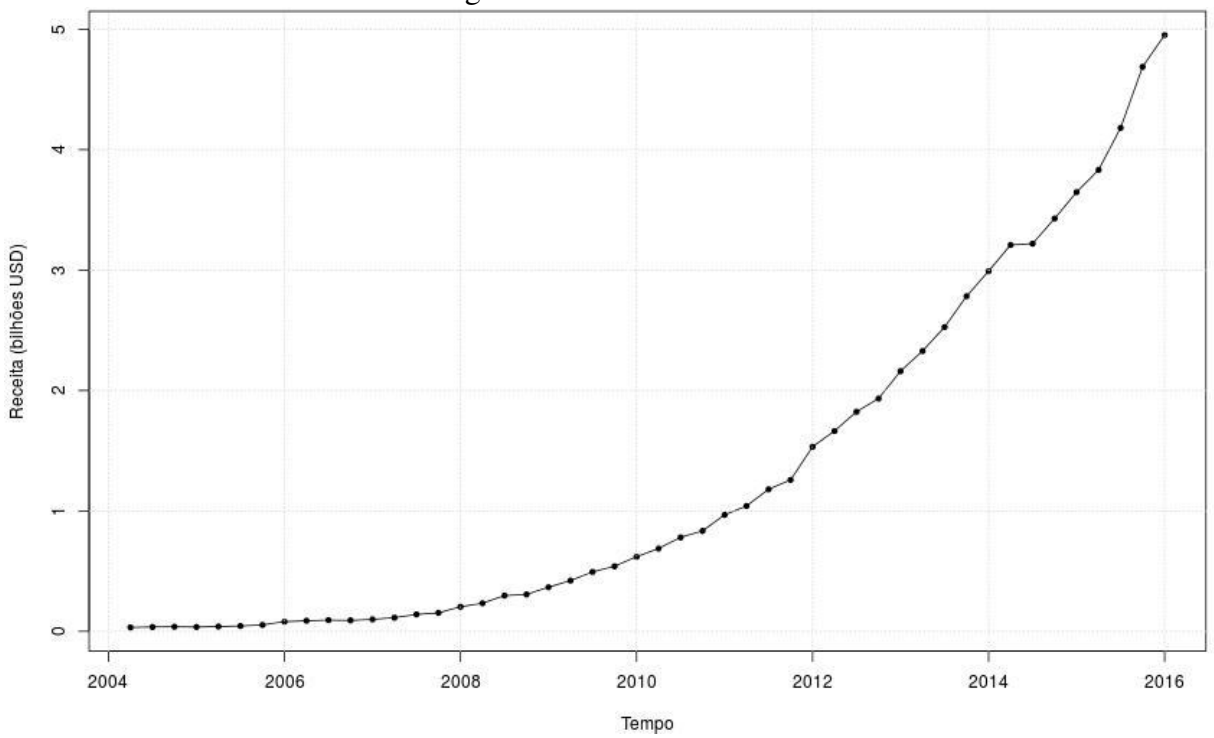


(Fonte: Próprio autor)

A série de receita é mostrada na Figura 5.3. O crescimento torna-se cada vez mais acentuado com o tempo e a evolução, entre uma observação e outra, é sempre positiva.

É fácil notar que ambas as séries não possuem uma média constante ao longo do tempo. As duas séries exibem claro crescimento com médias dependentes da variável temporal. Em outras palavras, a inspeção visual indica que ambos conjuntos de dados constituem séries temporais não estacionárias.

Figura 5.3 – Série de receita



(Fonte: Próprio autor)

5.2 Experimentos

O objetivo deste trabalho é analisar as quatro leis propostas para o valor das redes e número de usuários. Cada uma delas é representada por um modelo com sua própria variável independente (uma transformação da série do número de usuários) e um variável dependente (a receita total que é comum a todos modelos). Cada modelo é composto por apenas uma equação, apesar de um modelo de equações simultâneas ser intuitivamente o mais adequado na abordagem do problema.

Todas as leis concordam que um crescimento no número de usuários implica em um crescimento no valor agregado da rede, diferenciando-se pela forma funcional e taxa na qual isso ocorre. No entanto, não é difícil perceber que, ao mesmo tempo, um aumento no valor da rede atrairá novos usuários. Em outras palavras, o efeito de rede poderia ser melhor representado por um modelo de duas vias: o crescimento da rede implica em aumento no seu valor, enquanto a valorização da rede torna-a atrativa para que outros usuários participem dela. Essa abordagem fica como sugestão para trabalhos futuros.

As equações a serem estimadas de cada lei são mostradas abaixo:

Lei de Sarnoff: $V_t = \beta_{1S} + \beta_{2S} M_{tS} + u_{tS}$ onde $M_{tS} = N_t$;

$$\text{Lei de Metcalfe: } V_t = \beta_{1M} + \beta_{2M} M_{tM} + u_{tM} \text{ onde } M_{tM} = N_t^2;$$

$$\text{Lei de Odlyzko-Tilly: } V_t = \beta_{1OT} + \beta_{2OT} M_{tOT} + u_{tOT} \text{ onde } M_{tOT} = N_t \log N_t;$$

$$\text{Lei de Reed: } V_t = \beta_{1R} + \beta_{2R} M_{tR} + u_{tR} \text{ onde } M_{tR} = 2^{N_t},$$

onde V é a série da receita total e N a série de UAM e o subscrito t nos lembra que os dados são ordenados cronologicamente. Para cada um dos modelos, a tarefa consiste em estimar os parâmetros β_1 e β_2 se for razoável fazê-lo. Espera-se que os parâmetros β_1 sejam estatisticamente zero já que redes sem usuários intuitivamente não possuem valor agregado algum. Os coeficientes β_2 são esperados positivos dado que um aumento no número de usuários representa um aumento no valor da rede.

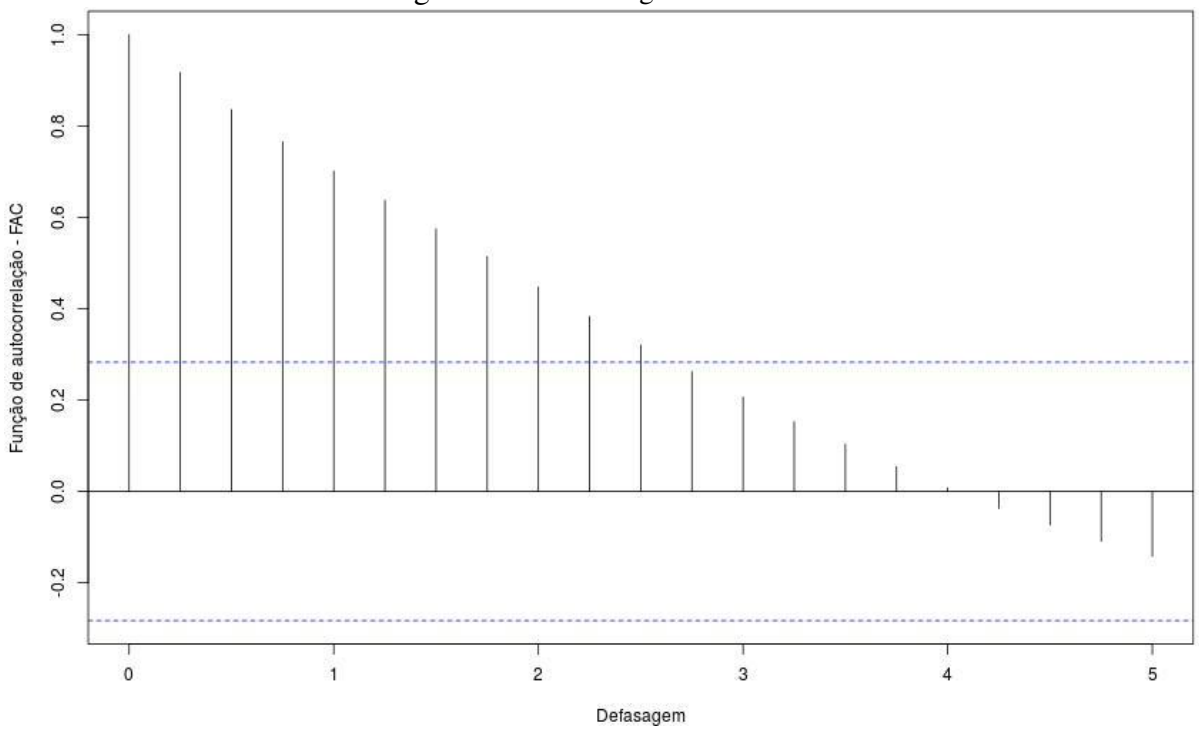
5.2.1 Série da Receita

Como foi apresentado na seção teórica, pode-se testar a existência de raizes unitárias e estacionariedade de séries temporais através dos testes de Dickey-Fuller, Phillips-Perron e KPSS. Como o conjunto de dados é pequeno, um só teste não é conclusivo para uma tomada de decisão. Assim, alguns testes com os mesmos objetivos e outros métodos de análise serão utilizados em conjunto para as tomadas de decisão sobre estacionariedade, homocedasticidade, normalidade dos resíduos e outras hipóteses.

O primeiro teste de Dickey-Fuller Aumentado (pacote *urca*) surpreendentemente apresenta estatística-teste 5.1291, indicando estacionariedade (os valores críticos a 1%, 5% e 10% são -2.62, -1.95 e -1.61, respectivamente). O segundo teste ADF (pacote *tseries*) indica fortemente não estacionariedade com um valor-p superior a 0.99 com ordem de defasagem 3. O mesmo acontece com o teste de Phillips-Perron (pacote *tseries*) com valor-p novamente superior a 0.99 e mesma ordem de defasagem. Por fim, o teste KPSS rejeita a hipótese nula de estacionariedade em tendência com valor-p inferior a 0.01. Assim, três dos quatro testes indicam não estacionariedade. O correlograma mostrado na Figura 5.4 corrobora com essa hipótese

O decaimento da FAC é bastante lento indicando não estacionariedade da série. Seguindo os intervalos de significância mostrados, a correlação só torna-se estatisticamente nula a partir da 11ª defasagem. Assim, consideraremos que a série de receita não é estacionária como se havia suspeitado pela inspeção visual das série.

Figura 5.4 – Correlograma da receita



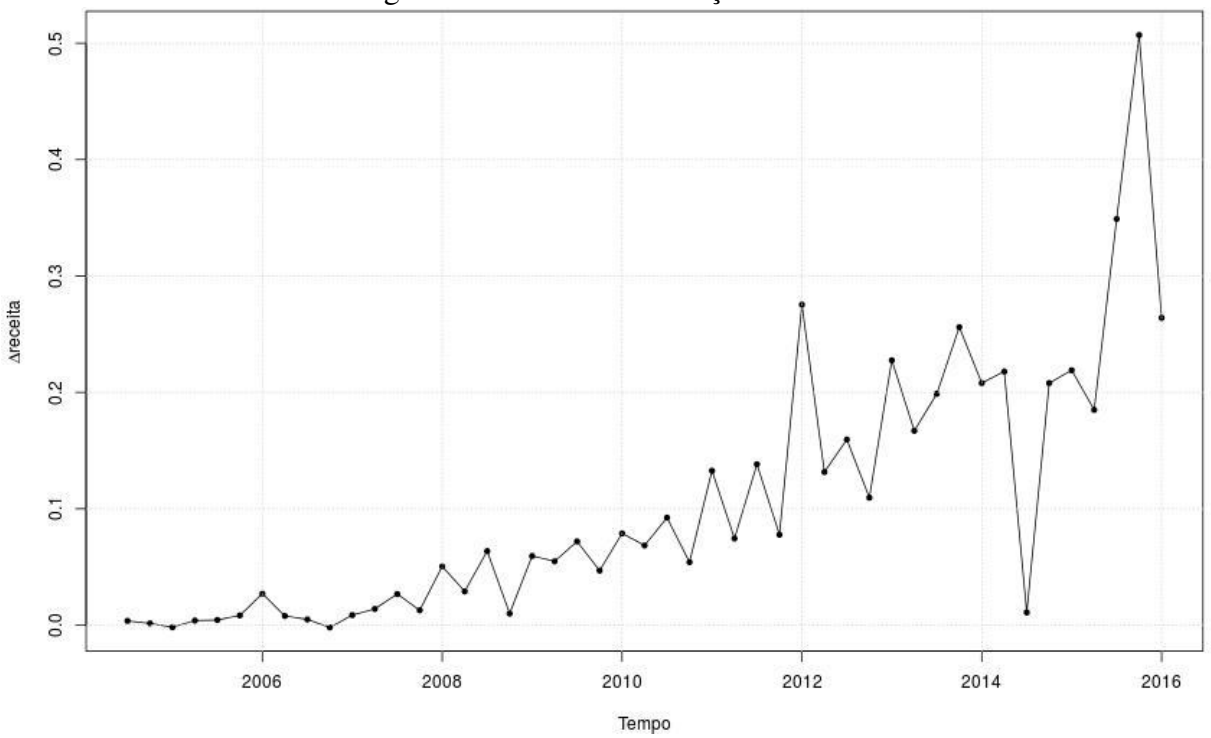
(Fonte: Próprio autor)

A diferenciação será utilizada para a removação da não estacionariedade. É possível que esse processo introduza ruído na forma de um aumento na variância da série como já discutido. A seguir são analisadas as duas primeiras diferenças da série.

A primeira diferença de receita, $\Delta receita$, mostrada na Figura 5.5 não se parece em nada com uma série estacionária. Os testes de Dickey-Fuller Aumentado e KPSS confirmam isso. O teste de Dickey-Fuller Aumentado (pacote *urca*) apresenta estatística-teste de 1.7587 indicando fortemente que a série é não estacionária. O teste de Dickey-Fuller Aumentado com tendência linear (do pacote *tseries*) tem valor-p de 0.4704 e não rejeita a hipótese nula de não estacionariedade. O teste de Phillips-Perron (pacote *tseries*), por outro lado, indica estacionariedade com valor-p inferior à 0.01. O teste KPSS (pacote *tseries*) para estacionariedade em tendência rejeita a hipótese nula com valor-p menor de 0.0425. O correlograma mostrado na Figura 5.6 indica novamente a não estacionariedade:

O decaimento das correlações ainda é lento e podemos manter a hipótese da não estacioariedade da primeira diferença da série de receita. Na Figura 5.7 é mostrado o gráfico da segunda diferença.

Figura 5.5 – Primeira diferença da receita



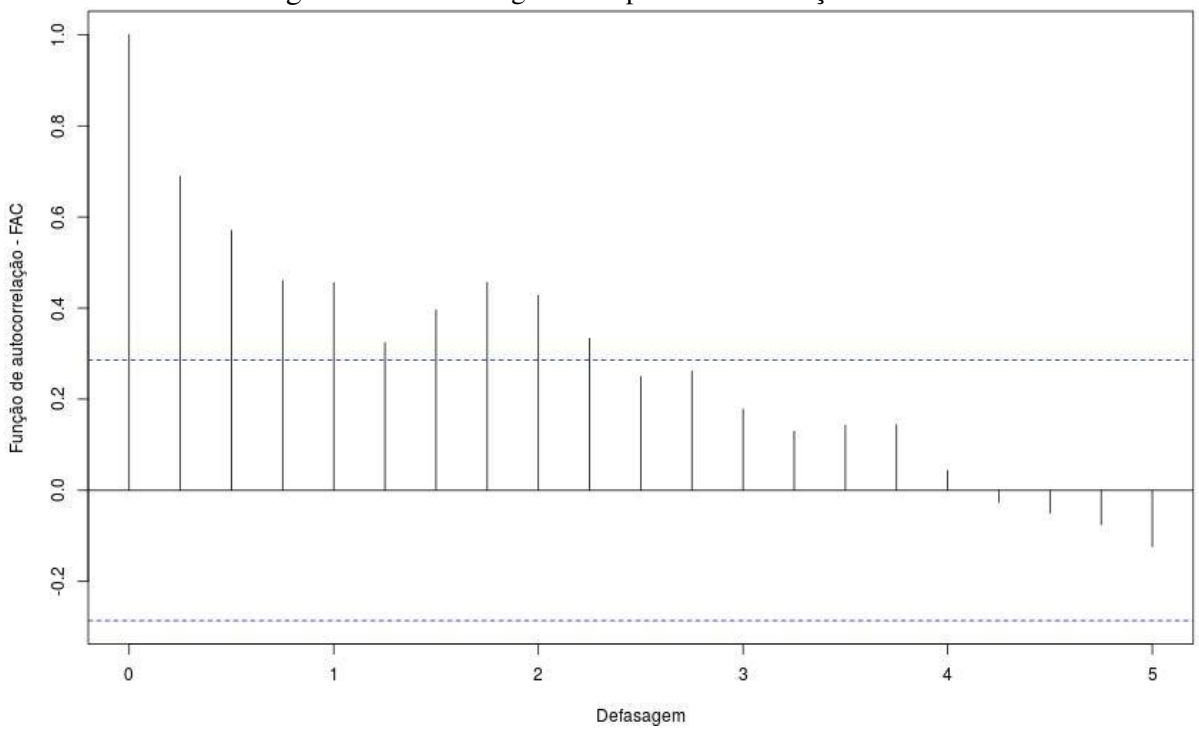
(Fonte: Próprio autor)

A variância da série diferenciada aparenta aumentar com o tempo. Isso caracteriza, por si só, não estacionariedade. Lembrando que, no contexto da estacionariedade fraca, ambas média e variância devem ser independentes do tempo. De qualquer forma, é válido continuar com os testes estatísticos e com as análises dos correlogramas.

Os testes para a segunda diferença da série de receita também sugerem mais fortemente estacionariedade. O teste de Dickey-Fuller Aumentado (pacote *urca*) sem tendência linear e com ordem de defasagem 5 encontrou estatística-teste igual a -4.21, rejeitando portanto a hipótese de não estacionariedade. O teste de Dickey-Fuller Aumentado (pacote *tseries*) com ordem de defasagem 3 tem valor-p de 0.1051, mantendo a hipótese nula. Os testes de Phillips-Perron e KPSS indicam ausência de raiz unitária e estacionariedade, respectivamente. Para embasar melhor a hipótese de que a receita é $I(2)$, observa-se os correlogramas da Figura 5.8.

Apesar dos correlogramas assemelharem-se mais daquilo que é esperado para uma série estacionária, ainda é possível notar que algumas correlações parecem significativamente diferentes de zero em ambos correlogramas (em $k = 1$ e 4 e $k = 1$ e 5, respectivamente). Neste momento, é interessante seguir com um teste para significância das correlações amostrais como o teste de Ljung-Box.

Figura 5.6 – Correlograma da primeira diferença da receita



(Fonte: Próprio autor)

O teste de Ljung-Box produz fortes indícios de que nem todas correlações são significativamente iguais a zero. A estatística LB computada foi 11.319 e o valor-p foi igual a 0.00077, rejeitando a hipótese nula. Apesar disso, será feita a consideração de que a segunda diferença é estacionária baseado principalmente nos testes de raiz unitária e estacionariedade feitos anteriormente. Certamente, este não é o cenário ideal, mas trata-se de um melhor esforço frente aos poucos e complicados dados disponíveis.

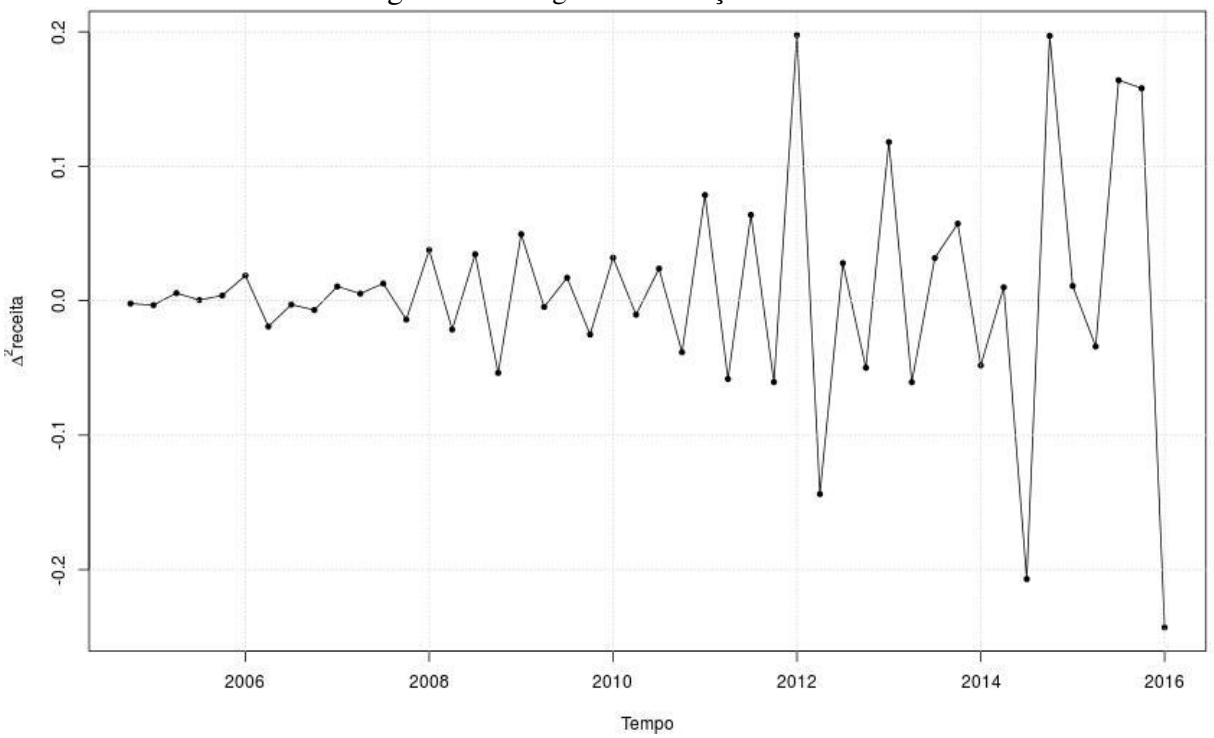
5.2.2 Lei de Sarnoff

O primeiro modelo a ser considerado é o de Sarnoff. Ele é o mais simples de todos, não necessitando de transformação na variável explanatória. Antes de proceder com qualquer regressão, deve-se verificar se as duas séries, receita e usuários, cointegram.

Já foi mostrado que a série de receita é $I(2)$. Para que haja cointegração, a série de usuários também deve ser $I(2)$. Caso não o seja, não faz muito sentido continuar com a regressão. A verificação da ordem de integração da série de usuários segue o mesmo esquema da série de receita.

O gráfico da Figura 5.9 mostra a segunda diferença da série de usuários. A segunda diferença parece bem mais com uma série estacionária.

Figura 5.7 – Segunda diferença da receita



(Fonte: Próprio autor)

Todos os testes feitos concordam com a estacionariedade e que, portanto, a série de usuários é $I(2)$. O correlograma na Figura 5.10 mostra que a hipótese é razoável.

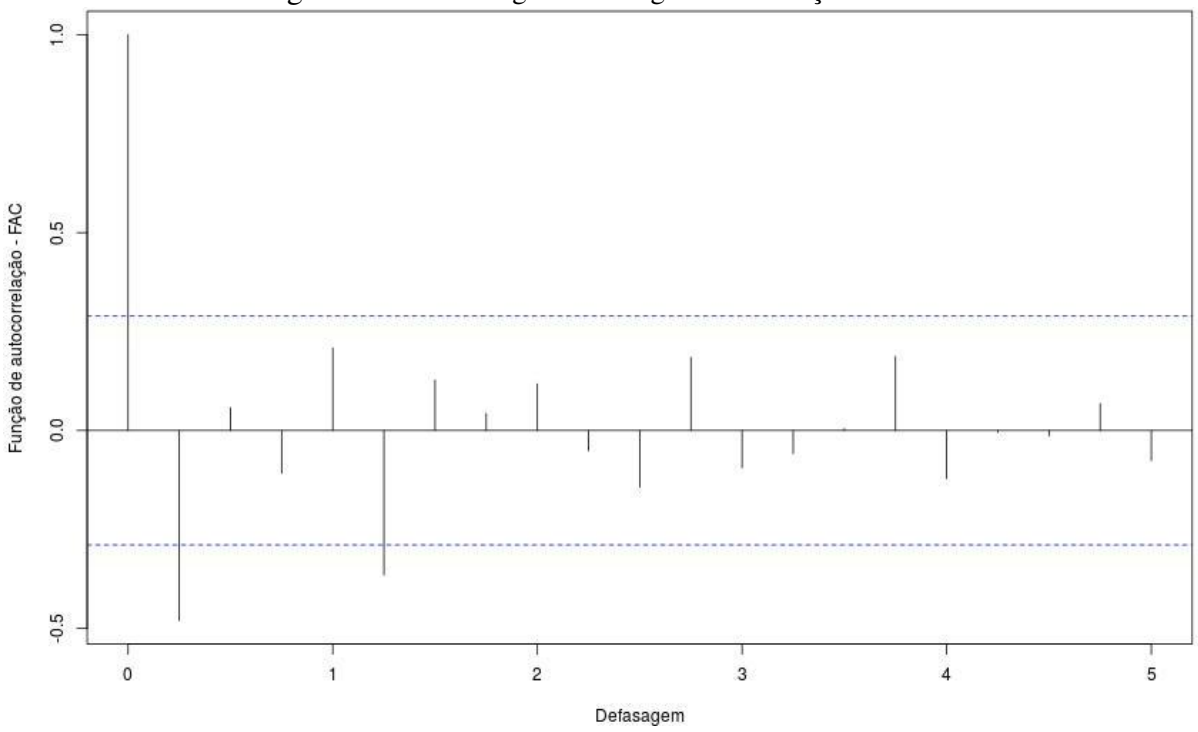
Para que haja cointegração, ainda é necessário que os resíduos da regressão sejam estacionários. Para realizar a regressão, será utilizada a clássica função *lm* do pacote *stats* do *R* para a regressão por Mínimos Quadrados Ordinários (MQO). Os resíduos são mostrados na Figura 5.11.

O gráfico deixa claro que os resíduos não são estacionários. Logo, não se pode considerar que as séries cointegram. A análise do modelo continuará, apesar das séries serem não estacionárias e não cointegrarem. A justificativa para isso é que existem poucos dados para se estabelecer confiança nos testes e ainda existe a possibilidade das séries cointegrarem. Além disso, a regressão possui embasamento teórico suficiente para fazer sentido. Alerta-se para o possível e provável caso de omissão de variável relevante. Isso pode explicar o comportamento dos resíduos.

Segue-se com a regressão

$$V_t = \widehat{\beta_{1S}} + \widehat{\beta_{2S}} N_t + \widehat{u_{tS}},$$

Figura 5.8 – Correlograma da segunda diferença da receita



(Fonte: Próprio autor)

onde o chapéu denota um estimador para o parâmetro populacional. Os valores mais relevantes da regressão são mostrados abaixo

Tabela 5.2 – Resultado da regressão do modelo de Sarnoff

Coeficiente	**Estimativa**	**Estatística** t	**Valor-p**
$\widehat{\beta_{1S}}$	-1.1366	-4.631	3.0e-05
$\widehat{\beta_{2S}}$	4.5669	11.083	1.4e-14

(Fonte: Próprio autor)

O valor de R^2 ajustado é 0.7216 e o Critério de Informação de Akaike para esse modelo é 113.7397. A seguir é mostrado os resultados dos testes de homocedasticidade do modelo e normalidade dos resíduos:

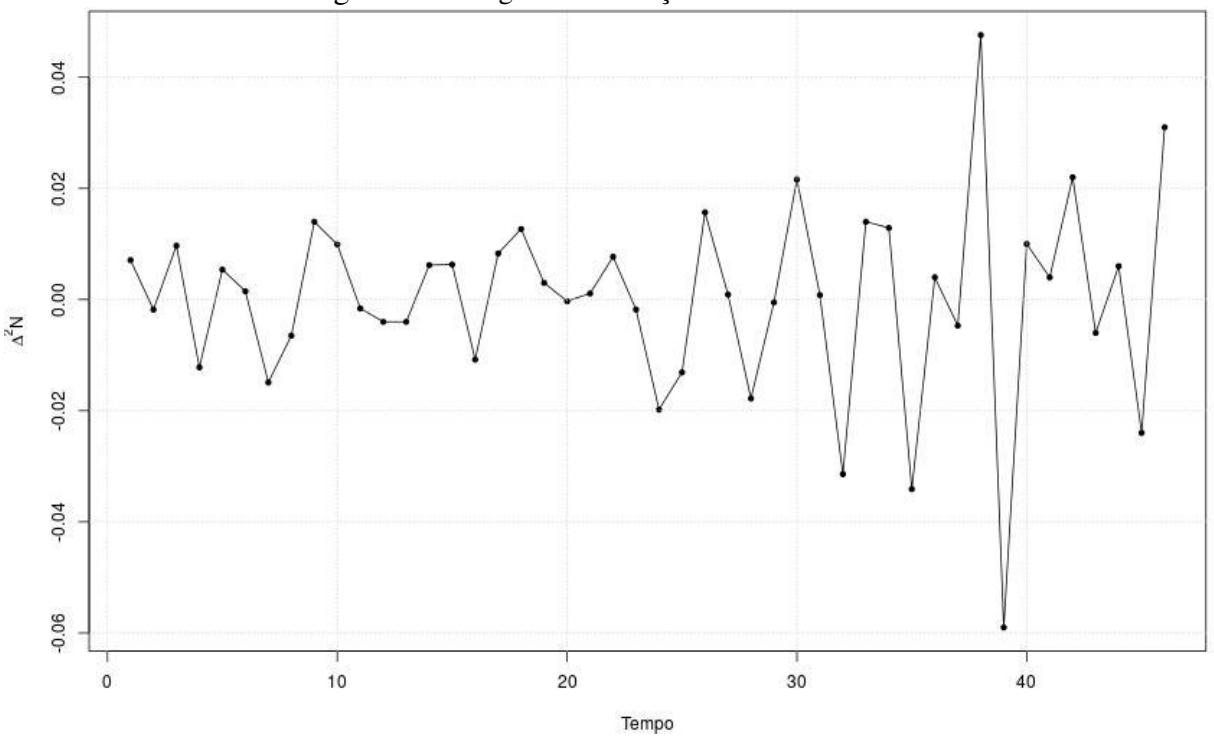
Tabela 5.3 – Homocedasticidade e normalidade dos resíduos no modelo de Sarnoff

Teste	**Hipótese nula**	**Estatística**	**Valor-p**
Breusch-Pagan	Homocedasticidade	10.835	<0.001
Shapiro-Wilk	Normalidade	0.9441	0.0234

(Fonte: Próprio autor)

Nota-se a presença de heterocedasticidade e que os resíduos não são normalmente distribuídos. As implicações da heterocedasticidade na inferência já foram discutidas.

Figura 5.9 – Segunda diferença de usuários - $\Delta^2 N$



(Fonte: Próprio autor)

5.2.3 Lei de Metcalfe

O segundo modelo é o de Metcalfe. Ele é o principal modelo deste trabalho e necessita de uma simples transformação na variável explanatória. Novamente, verifica-se a possibilidade de cointegração das séries.

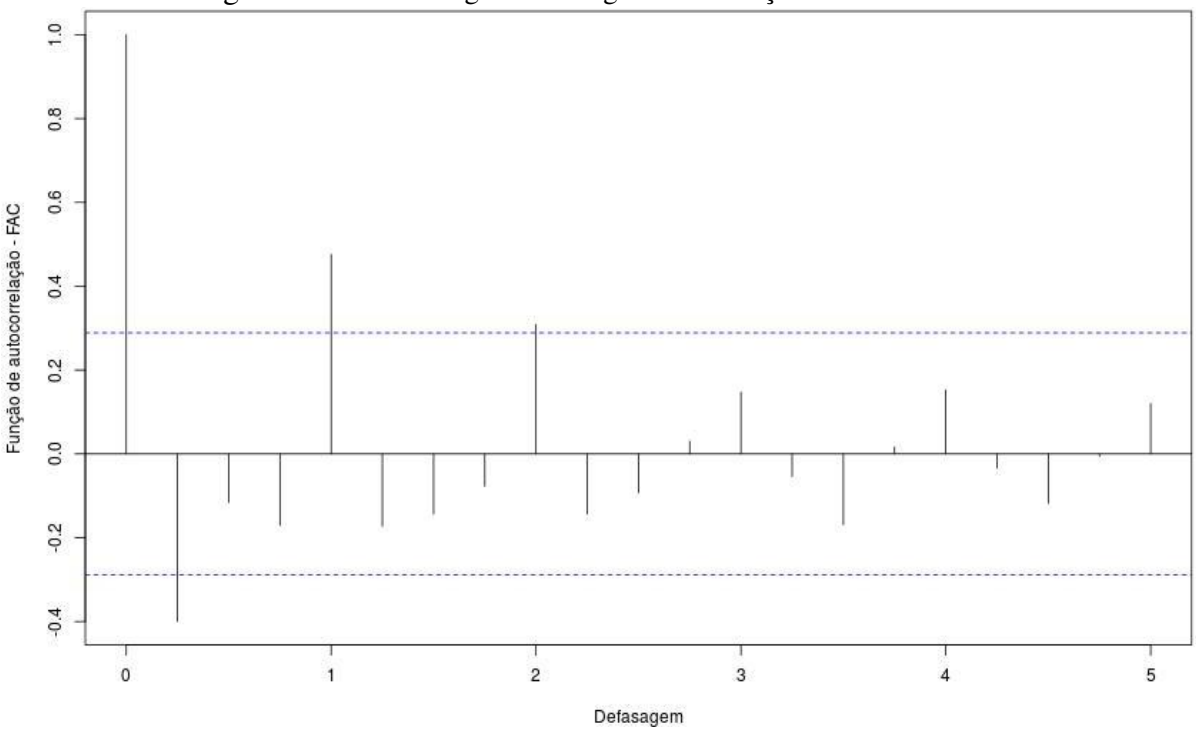
Dessa vez, para que haja cointegração, a série de usuários transformada também deve ser $I(2)$.

Assim como aconteceu com a segunda diferença no modelo de Sarnoff, a série diferenciada de Metcalfe apresenta significativo aumento na variância com o tempo. Isso pode ser fruto do processo de diferenciação empregado quando a série apresenta tendência determinística.

Os resultados mostram que a série de usuários ao quadrado também é $I(2)$. Novamente, todos os testes (Dickey-Fuller, Phillips-Perron e KPSS) concordam com a estacionariedade. O correlograma da Figura 5.13 demonstra forte correlação para $k = 1$ e $k = 4$, mas mantém-se a hipótese de estacionariedade devido aos testes anteriores.

Infelizmente, a regressão revela que os resíduos são claramente não estacionários e portanto não é possível afirmar que as séries cointegram. Da mesma forma como feito na análise

Figura 5.10 – Correlograma da segunda diferença de usuários - $\Delta^2 N$



(Fonte: Próprio autor)

da lei de Sarnoff acima, aqui se prosseguirá com a análise apesar da não cointegração das séries. A equação a ser estimada é

$$V_t = \widehat{\beta_{1M}} + \widehat{\beta_{2M}} N_t^2 + \widehat{u_{tM}}.$$

Os valores mais relevantes da regressão são mostrados abaixo

Tabela 5.4 – Resultado da regressão do modelo de Metcalfe

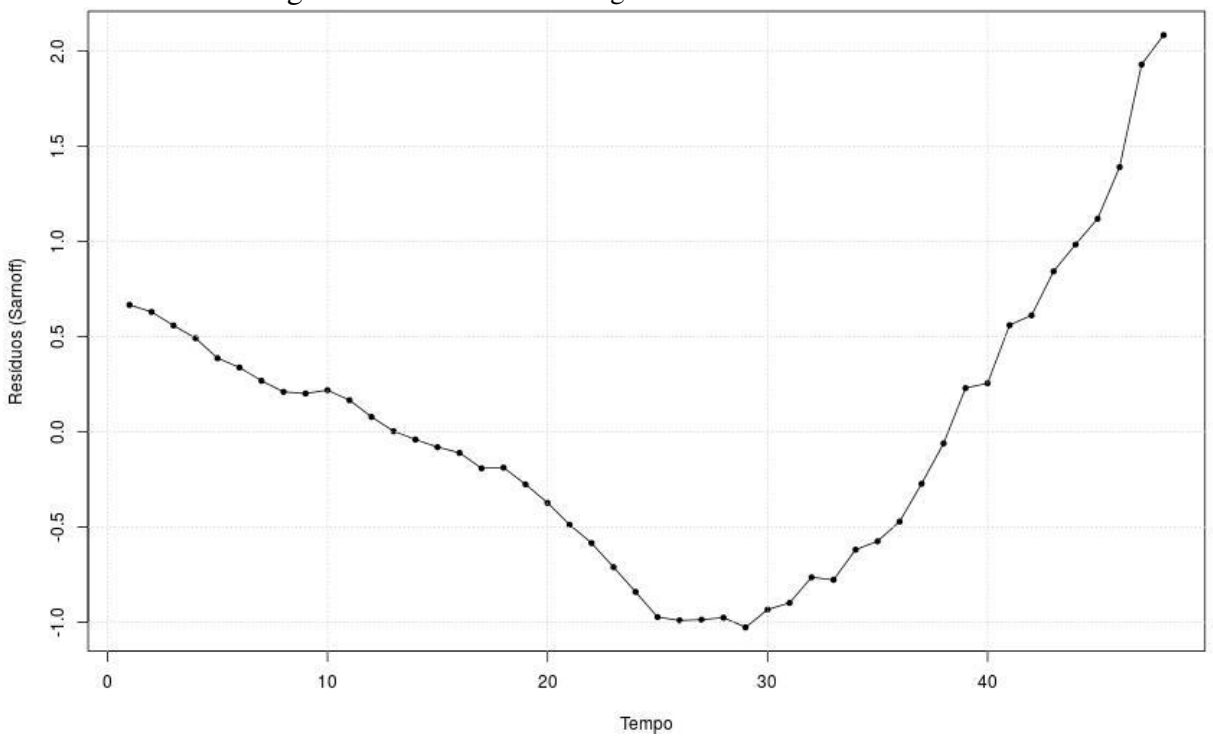
Coeficiente	**Estimativa**	**Estatística** t	**Valor-p**
$\widehat{\beta_{1M}}$	-0.3677	-2.514	0.0155
$\widehat{\beta_{2M}}$	4.6934	14.390	<2e-16

(Fonte: Próprio autor)

O valor de R^2 ajustado é 0.8143 e o Critério de Informação de Akaike para esse modelo é 94.3076. Na Tabela 5.5 são mostrados os resultados dos testes de homocedasticidade do modelo e normalidade dos resíduos.

Como no modelo de Sarnoff, existe heterocedasticidade e os resíduos não são normalmente distribuídos. Esse resultado já era esperado dada a similaridade dos resíduos com a regressão de Sarnoff.

Figura 5.11 – Resíduos da regressão do modelo de Sarnoff



(Fonte: Próprio autor)

Tabela 5.5 – Homocedasticidade e normalidade dos resíduos no modelo de Metcalfe

Teste	**Hipótese nula**	**Estatística**	**Valor-p**
Breusch-Pagan	Homocedasticidade	15.041	<0.001
Shapiro-Wilk	Normalidade	0.92377	0.004048

(Fonte: Próprio autor)

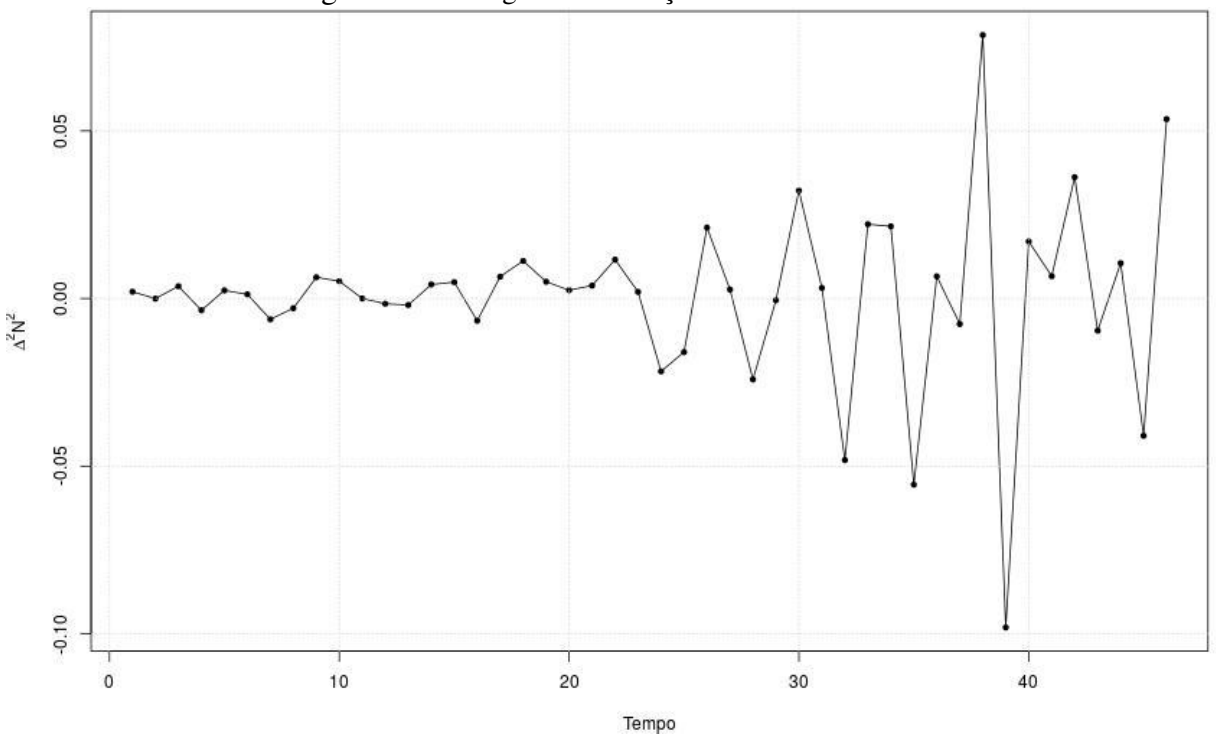
5.2.4 Lei de Odlyzko-Tilly

O modelo de Odlyzko-Tilly é o terceiro a ser avaliado. Novamente, verifica-se a possibilidade de cointegração das séries.

Como em todos os casos anteriores, a série diferenciada de Odlyzko-Tilly apresenta significativo aumento na variância com o tempo e a explicação é a mesma. Para testar a estacionariedade, procede-se com os usuais testes de Dickey-Fuller, Phillips-Perron e KPSS.

Os testes apontam que a série de usuários transformada também é $I(2)$. O correlograma que encontra-se na Figura 5.16 de novo demonstra forte correlação para $k = 1$ e $k = 4$ com uma forte diminuição nas correlaçõe seguintes, mas mantém-se a hipótese de estacionariedade devido aos testes anteriores.

Figura 5.12 – Segunda diferença de usuários - $\Delta^2 N^2$



(Fonte: Próprio autor)

Apesar da mudança na forma dos resíduos como demonstrado na Figura 5.17, ainda não é possível dizer que eles constituem uma série estacionária. A cointegração deve ser novamente rejeitada.

$$V_t = \widehat{\beta_{1OT}} + \widehat{\beta_{2OT}} N_t \log(N_t) + \widehat{u_{tOT}}.$$

Os valores mais relevantes da regressão são mostrados abaixo

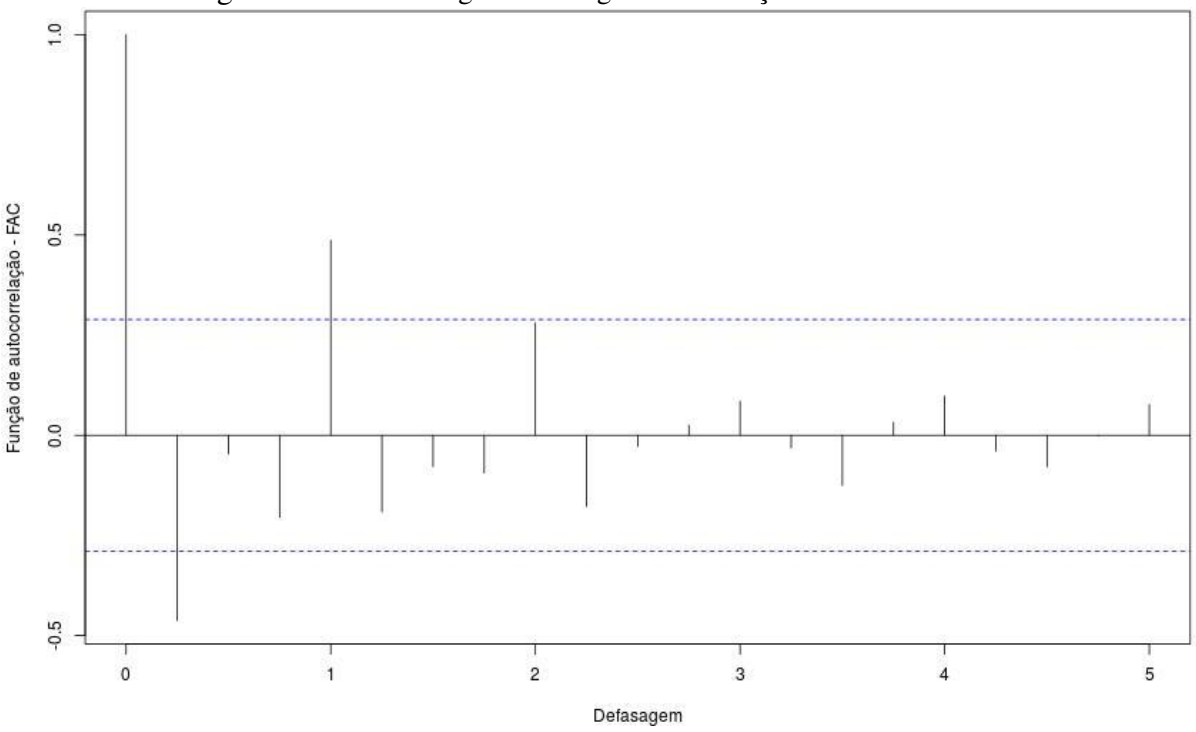
Tabela 5.6 – Resultado da regressão do modelo de Odlyzko-Tilly

Coeficiente	**Estimativa**	**Estatística t**	**Valor-p**
$\widehat{\beta_{1OT}}$	5.2868	16.98	<2e-16
$\widehat{\beta_{2OT}}$	15.2505	13.44	<2e-16

(Fonte: Próprio autor)

O valor de R^2 ajustado é 0.7927 e o Critério de Informação de Akaike para esse modelo é 99.58123. Chama-se a atenção para o a estimativa do intercepto, $\widehat{\beta_{1OT}}$, que aparenta ser estatisticamente significativa. Como dito anteriormente, espera-se que o intercepto do modelo seja nulo já que o valor da rede deve ser zero quando a rede não tiver usuários. Esse resultado contraria as expectativas da investigação. Na Tabela 5.7 são mostrados os resultados dos testes de homocedasticidade do modelo e normalidade dos resíduos.

Figura 5.13 – Correlograma da segunda diferença de usuários - $\Delta^2 N^2$



(Fonte: Próprio autor)

Tabela 5.7 – Homocedasticidade e normalidade dos resíduos no modelo de Odlyzko-Tilly

Teste	**Hipótese nula**	**Estatística**	**Valor-p**
Breusch-Pagan	Homocedasticidade	5.0476	0.02466
Shapiro-Wilk	Normalidade	0.9837	0.7369

(Fonte: Próprio autor)

Ao contrário dos testes anteriores, dessa vez não se rejeita a hipótese de que os dados são provenientes de uma população normal. No entanto, a heterocedasticidade ainda está presente.

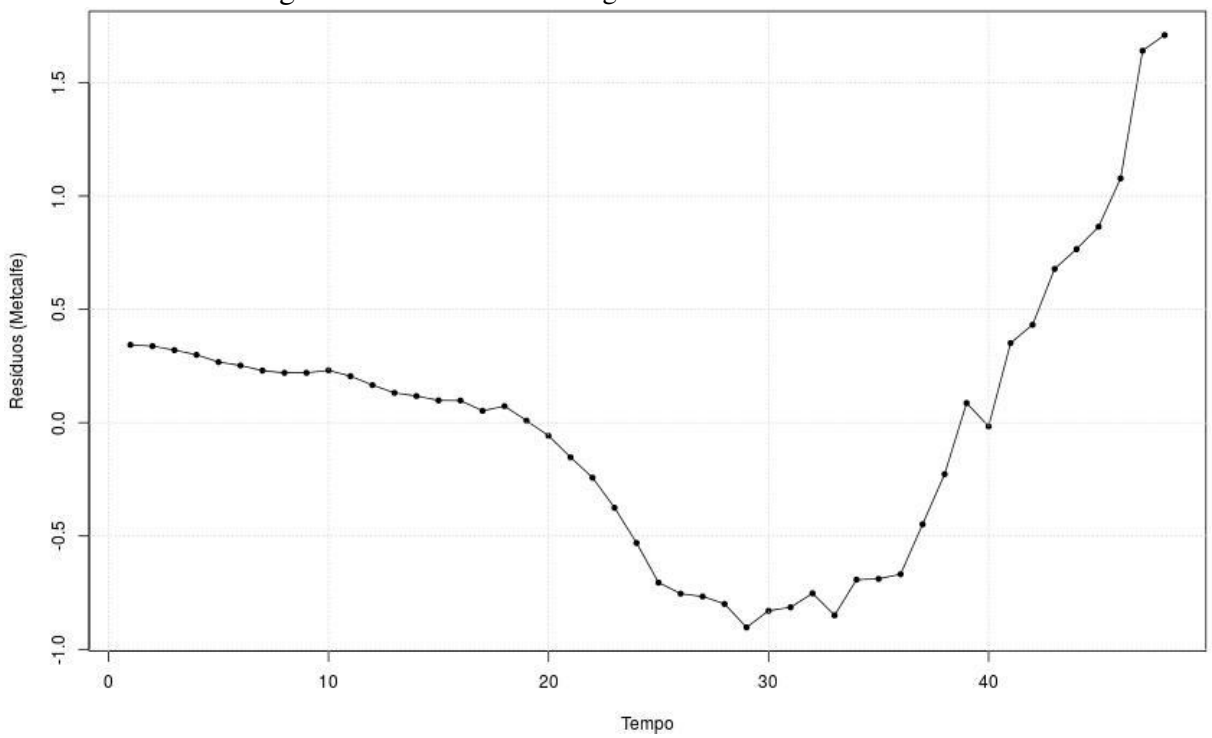
5.2.5 Lei de Reed

O modelo de Reed é o último a ser estudado. A Figura 5.18 mostra a segunda diferença da série de usuários transformada para este caso.

A série diferenciada de Reed exibe o mesmo aumento significativo na variância com o tempo. A ordem de integração da série pode ser confirmada através dos testes estatísticos.

Os resultados são praticamente os mesmos para os modelos anteriores. Mantém-se a hipótese de estacionariedade da segunda diferença, ou seja, a série com a transformação de Reed é $I(2)$. Na Figura 5.19 é mostrado seu correlograma.

Figura 5.14 – Resíduos da regressão do modelo de Metcalfe



(Fonte: Próprio autor)

De qualquer forma, ainda não é possível dizer que os resíduos são $I(0)$. De novo, não há cointegração.

$$V_t = \widehat{\beta_{1R}} + \widehat{\beta_{2R}} 2^{N_t} + \widehat{u_{tR}}.$$

Os valores mais relevantes da regressão são mostrados abaixo

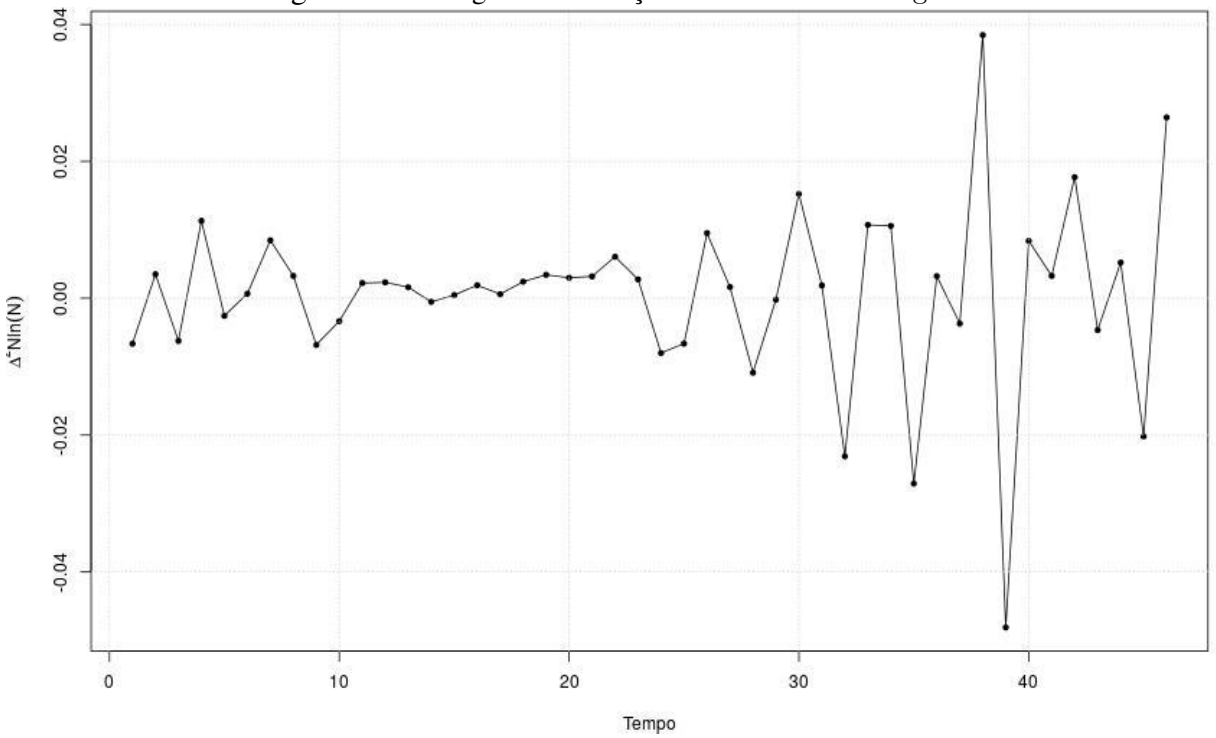
Tabela 5.8 – Resultado da regressão do modelo de Reed

Coeficiente	**Estimativa**	**Estatística** t	**Valor-p**
$\widehat{\beta_{1R}}$	-5.6178	-9.768	8.55e-13
$\widehat{\beta_{2R}}$	4.7000	12.217	4.83e-16

(Fonte: Próprio autor)

O valor de R^2 ajustado é 0.7593 e o Critério de Informação de Akaike para esse modelo é 106.7595. Novamente a estimativa do intercepto, $\widehat{\beta_{1OT}}$, estatisticamente significativa. Na Tabela 5.9 são mostrados os resultados dos testes de homocedasticidade do modelo e normalidade dos resíduos.

Figura 5.15 – Segunda diferença de usuários - $\Delta^2 N \log N$



(Fonte: Próprio autor)

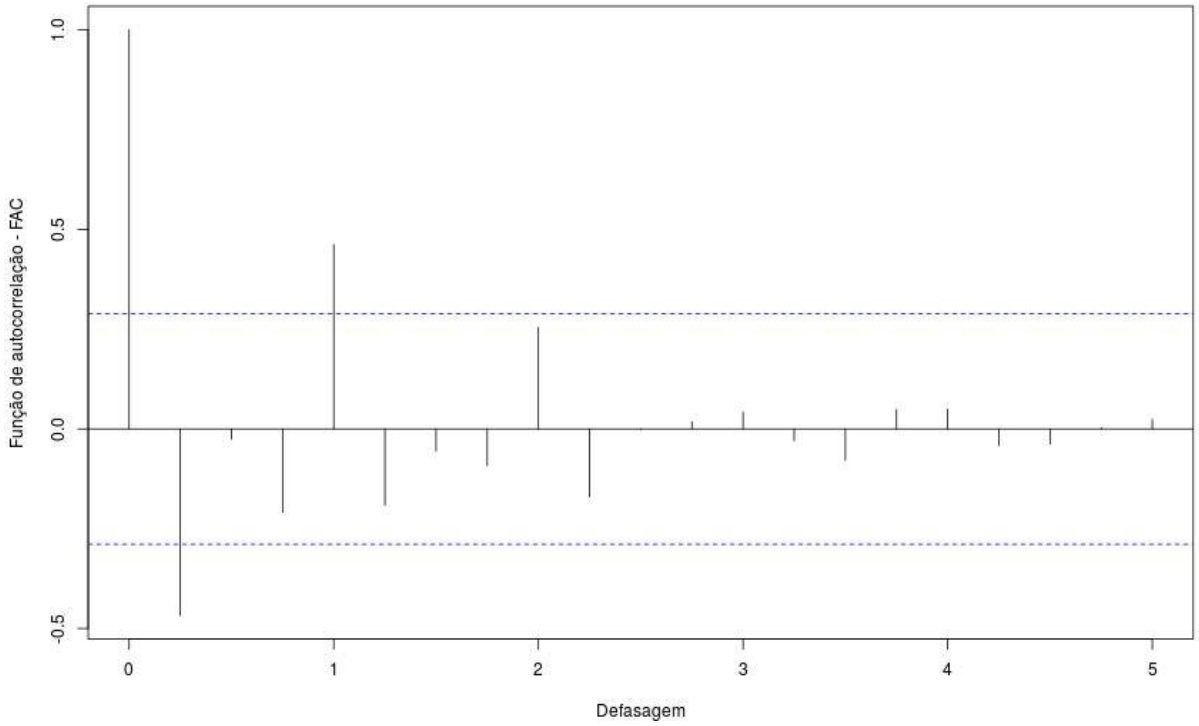
Tabela 5.9 – Homocedasticidade e normalidade dos resíduos no modelo de Reed

Teste	**Hipótese nula**	**Estatística**	**Valor-p**
Breusch-Pagan	Homocedasticidade	12.104	<0.001
Shapiro-Wilk	Normalidade	0.94244	0.02016

(Fonte: Próprio autor)

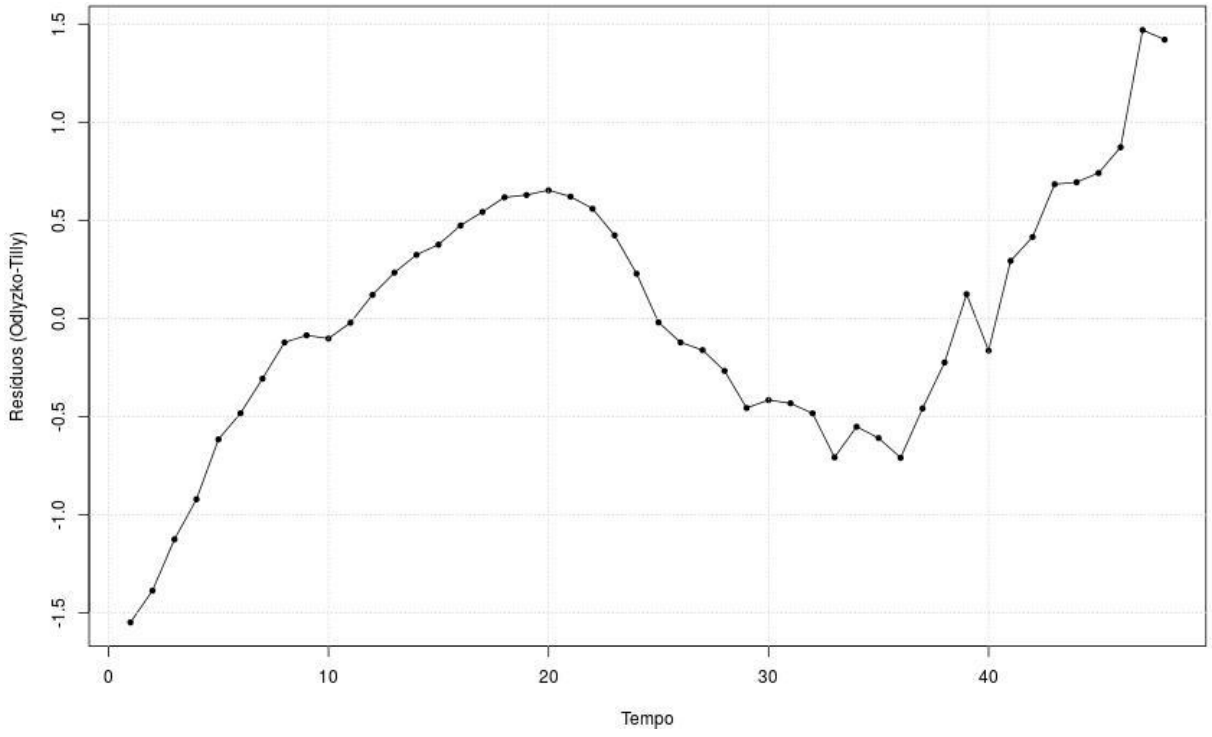
Como nos dois primeiros modelos, rejeita-se a homocedasticidade e os resíduos não parecem vir de uma distribuição normal. O tratamento do problema da heterocedasticidade fica como trabalho futuro.

Figura 5.16 – Correlograma da segunda diferença de usuários - $\Delta^2 N \log N$



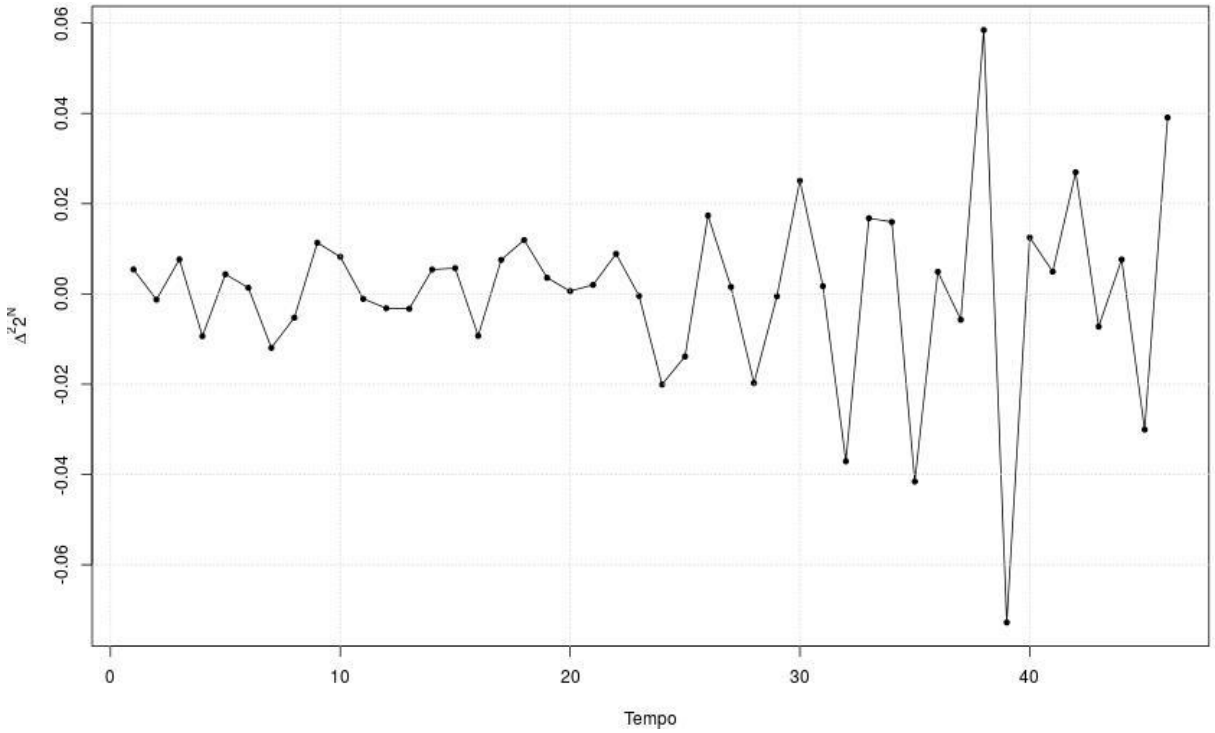
(Fonte: Próprio autor)

Figura 5.17 – Resíduos da regressão do modelo de Odlyzko-Tilly



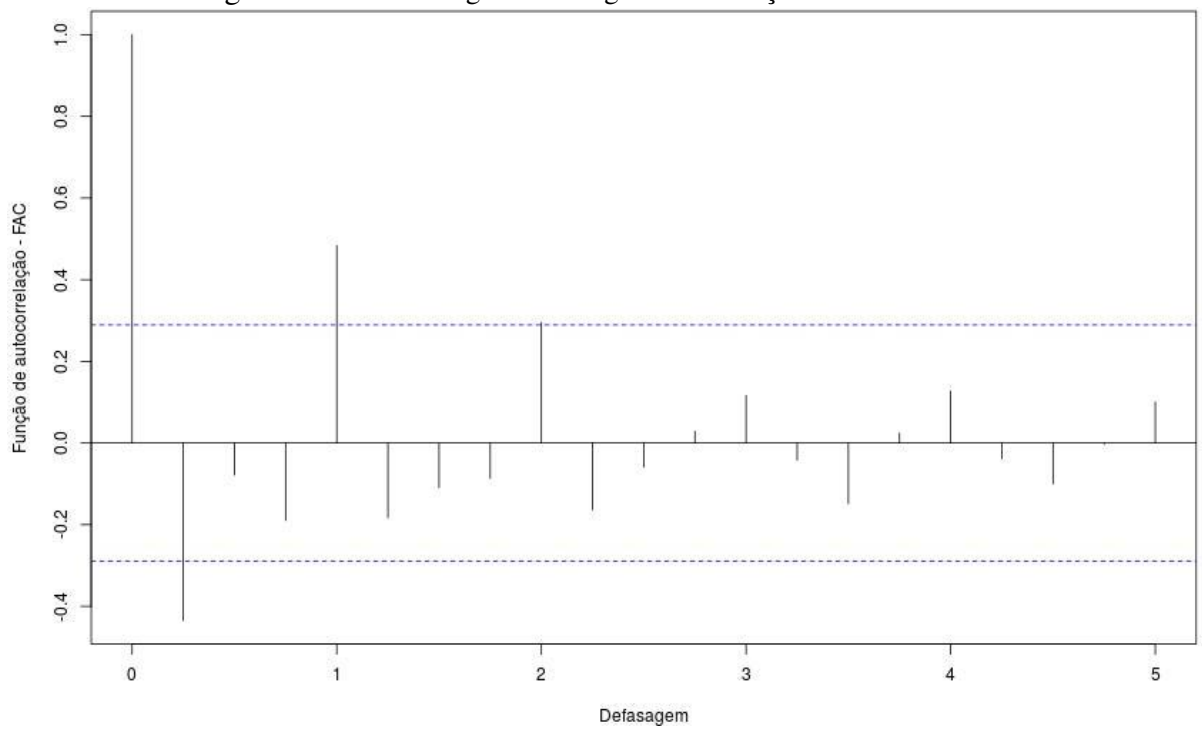
(Fonte: Próprio autor)

Figura 5.18 – Segunda diferença de usuários - $\Delta^2 2^N$



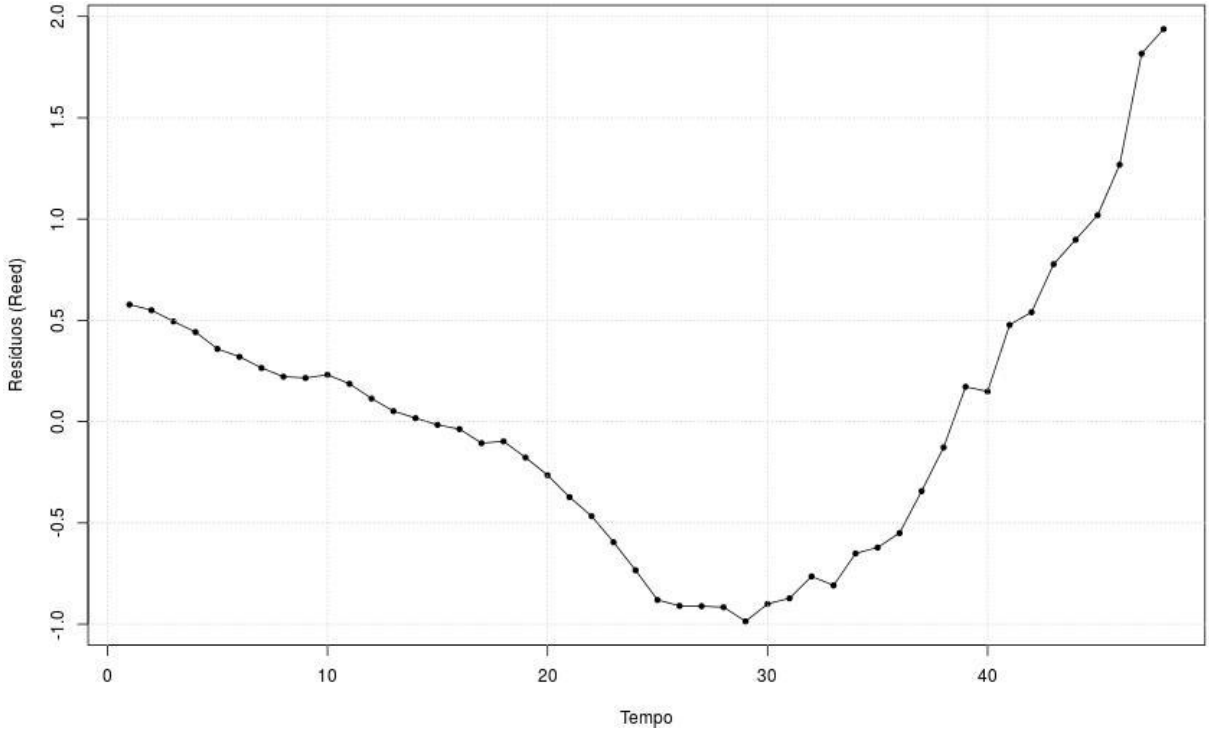
(Fonte: Próprio autor)

Figura 5.19 – Correlograma da segunda diferença de usuários - $\Delta^2 2^N$



(Fonte: Próprio autor)

Figura 5.20 – Resíduos da regressão do modelo de Reed



(Fonte: Próprio autor)

5.3 Resultados

Os dados não apresentaram as condições necessárias para que a análise de regressão fosse realizada da forma apropriada. As séries temporais originais mostraram-se claramente não estacionárias. Tanto a série que representa o valor quanto a série que representa o tamanho aparentam possuir ambos tipos de tendência: determinística e estocástica. O processo de remoção de tendência utilizado foi apenas a diferenciação.

Após as séries serem diferenciadas e feitas estacionárias, os resíduos das regressões foram analisados para cada um dos modelos. Como os resíduos foram não estacionários, considerou-se que as séries não cointegram. Para que elas sejam ditas cointegradas, elas devem possuir a mesma ordem de integração e os resíduos da equação cointegrante devem ser estacionários. Apesar de plausível a hipótese de ambas séries serem integradas de mesma ordem ($I(2)$), os resíduos da regressão são nitidamente não estacionários. Assim, não é possível avaliar com razoável confiança os modelos propostos com os dados utilizados.

No entanto, manteve-se a análise para os modelos. Um resumo das regressões é mostrado abaixo

Tabela 5.10 – Resumo dos resultados das regressões

Modelo	$\widehat{\beta_1}$	$\widehat{\beta_2}$	R^2 ajustado	**AIC**
Sarnoff	-1.1366	4.5669	0.7216	113.73
Metcalfe	-0.3677	4.6934	0.8143	94.3076
Odlyzko-Tilly	5.2868	15.2505	0.7927	99.58123
Reed	-5.6178	4.7000	0.7593	106.7595

(Fonte: Próprio autor)

Seguindo os critérios do maior R^2 ajustado e menor AIC, a lei de Metcalfe parece fornecer o melhor modelo para a quantificação do efeito de rede. A expectativa de um valor positivo para os coeficientes foi confirmada pelos dados. No entanto, cabe notar que todos os modelos mostraram-se inconsistentes com a teoria de que o valor do intercepto é nulo.

Dada relação de Metcalfe:[4]

$$V_t = \beta_0 + \beta_1 N_t^2,$$

Um aumento em 1 unidade no número de usuários (1 bilhão de usuários) leva a

[4]Idealmente, o intercepto β_0 é nulo.

$$V_{t+1} = \beta_0 + \beta_1(N_t + 1)^2$$

$$V_{t+1} = \beta_0 + \beta_1 N_t^2 + 2\beta_1 N_t + \beta_1,$$

$$V_{t+1} = V_t + (2\beta_1 N_t + \beta_1)$$

A diferença no valor é então

$$\Delta V = 2\beta_1 N_t + \beta_1.$$

Sendo a estimativa $\widehat{\beta_1}$, de β_1, 4.6934, um aumento em um usuário leva a um aumento de

$$\Delta V = 9.3868 N_t + 4.6934,$$

no valor (em bilhões USD), onde N_t é o número de usuários no instante t, antes do crescimento.

5.4 Trabalhos Futuros

Um dos maiores problemas deste trabalho foi a falta de dados. Para a análise de séries temporais, geralmente faz-se uso de uma grande quantidade de observações já que muitos testes e propriedades para esse tipo de dados são assintóticos, ou seja, valem principalmente quando o tamanho da amostra é grande. Como esse não foi o caso aqui, os testes não podem ser considerados definitivos ou conclusivos. Deve-se o fato de haver poucas observações disponíveis ao grande intervalo de tempo entre as divulgações dos relatórios financeiros para investidores de empresas de interesse, como Facebook, LinkedIn e Tencent, em conjunto com a recente abertura dessas empresas nas bolsas de valores. Futuramente, estarão disponíveis mais dados e esta investigação pode ser revisitada.

Como já foi dito, o efeito de rede pode ser melhor explicado por um modelo de equações simultâneas. Nesse modelo, não só o valor da rede é influenciado pelo tamanho dessa rede, como o tamanho é influenciado pelo valor. Esse modelo exigiria a introdução de novas variáveis endógenas e um tratamento mais sofisticado do ponto de vista econométrico, mas retrataria a teoria do efeito de rede de forma mais realística.

Uma outra possível melhoria na análise é a introdução de modelos dinâmicos. Nesse tipo de modelo, uma série não só é explicada em termos de outras séries, mas também através

de defasagens dela própria. É razoável imaginar que o valor de uma rede dependa não somente de sua extensão, mas também de seus valores passados. Além de utilizar defasagens da própria série na sua modelagem, é interessante observar e adicionar ao modelo outras variáveis que podem ajudar a explicar os dados, evitando possíveis problemas relacionados à omissão de variáveis relevantes.

6 CONCLUSÕES

No trabalho foi vista a importância das leis de crescimento do valor da rede e seu tamanho. Tais leis se propunham a quantificar o chamado efeito de rede, que foi tão importante durante o "*boom*" da era *dot-com*. Ao longo dos anos, diversos modelos sobre o assunto foram propostos diferenciando-se um do outro pela taxa de crescimento. Enquanto alguns modelos eram mais modesto, sugerindo crescimento linear, outros mostravam-se mais otimistas e advogavam a favor do crescimento exponencial. Como quer que seja, as teorias pouco haviam sido testadas com dados reais.

A investigação empírica desses modelos encontra alguns desafios naturais. Primeiro, as variáveis em questão são difíceis de serem observadas ou até mesmo quantificadas. O termo "valor de uma rede", por exemplo, é um tanto vago e escolher uma variável que represente-o e seja observável não é uma tarefa trivial. O tamanho da rede, apesar de parecer intuitivo, também é capaz de levantar questões. Segundo, mesmo que as variáveis do modelo estejam bem definidas, encontrar dados para realizar as análises pode ser um problema maior ainda. A pequena amostra disponível foi um dos maiores problemas enfrentados durante este trabalho. Muitas das técnicas estatísticas empregadas supõem um grande número de observações e quando este não é o caso, os resultados podem ser inconclusivos e inconsistentes. Foi preciso uma dose de intuição e bom-senso para concluir a investigação.

Os dados utilizados neste trabalho são séries temporais, o que exigiu uma análise anterior às regressões para validação dos modelos. As duas séries observadas - valor e tamanho - são não estacionárias. Isso levou à investigação da cointegração. No entanto, a cointegração também não teve suas condições atendidas. Apesar da das séries serem integradas de mesma ordem, os resíduos das equações cointegrantes foram claramente não estacionários. Os motivos para isso incluem amostra pequena e omissão de variáveis relevantes nos modelos. Ciente desses problemas, decidiu-se continuar com a análise de regressão.

Os resultados das regressões utilizando Mínimos Quadrados Ordinários mostraram que a lei de Metcalfe é a que melhor explica os dados observados segundo os critérios de maior R^2 ajustado e menor Critério de Informação de Akaike. Todos os modelos apresentaram heteroscedasticidade e resíduos não normais.

Como trabalho futuro foi sugerido o uso de modelos de equações simultâneas a fim de refletir de forma mais realística o efeito de rede e a adição de novas variáveis nos modelos. Modelos dinâmicos, que incluem defasagens da própria variável dependente, parecem uma alternativa intuitiva na melhoria dos modelos. A análise realizada neste trabalho pode ser revisitada

quando amostras maiores estiverem disponíveis.

Apesar dos resultados não terem sido muito favoráveis, o trabalho serviu como roteiro na análise exploratória de dados, principalmente daqueles de séries temporais. Além disso, foi mostrado um caminho para uma análise mais formal do assunto do ponto de vista estatístico.

REFERÊNCIAS

BECKSTROM, R. A new model for network valuation. **National Cyber Security Center Research Paper**, 2009.

BOX, G. E.; PIERCE, D. A. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. **Journal of the American statistical Association**, Taylor & Francis Group, v. 65, n. 332, p. 1509–1526, 1970.

BRISCOE, B.; ODLYZKO, A.; TILLY, B. Metcalfe's law is wrong-communications networks increase in value as they add members-but by how much? **Spectrum, IEEE**, IEEE, v. 43, n. 7, p. 34–39, 2006.

BUENO, R. D. L. da S. **Econometria de séries temporais**. [S.l.]: Cengage Learning, 2008.

CHIU, D. M.; NG, W. Y. Exploring network economics. **arXiv preprint arXiv:1106.1282**, 2011.

DAVIDSON, R.; MACKINNON, J. G. **Econometric theory and methods**. [S.l.]: Oxford University Press New York, 2004.

DICKEY, D. A.; FULLER, W. A. Distribution of the estimators for autoregressive time series with a unit root. **Journal of the American statistical association**, Taylor & Francis, v. 74, n. 366a, p. 427–431, 1979.

ENDERS, W. **Applied econometric time series**. [S.l.]: John Wiley & Sons, 2008.

ENGLE, R. F.; GRANGER, C. W. Co-integration and error correction: representation, estimation, and testing. **Econometrica: journal of the Econometric Society**, JSTOR, p. 251–276, 1987.

FEIJÓO, C.; GÓMEZ-BARROSO, J. L.; VOIGT, P. Exploring the economic value of personal information from firms' financial statements. **International Journal of Information Management**, Elsevier, v. 34, n. 2, p. 248–256, 2014.

GILDER, G. Metcalfe's law and legacy. **Forbes ASAP**, 1993.

GUJARATI, D. N. **Basic econometrics**. [S.l.]: McGraw-Hill, 2009.

HAMILTON, J. D. **Time series analysis**. [S.l.]: Princeton university press Princeton, 1994.

HENDRY, D.; RICHARD, R. The econometric analysis of economic time series. **International Statistical Review**, v. 51, p. 3–33, 1983.

HOVE, L. V. Metcalfe's law: not so wrong after all. **NETNOMICS: Economic Research and Electronic Networking**, Springer, v. 15, n. 1, p. 1–8, 2014.

KAPROWSKI, G. Fcc chair wants universal net access - and he's serious. **Wired News**, 1996.

KATZ, M. L.; SHAPIRO, C. Network externalities, competition, and compatibility. **The American economic review**, JSTOR, v. 75, n. 3, p. 424–440, 1985.

LJUNG, G. M.; BOX, G. E. On a measure of lack of fit in time series models. **Biometrika**, Biometrika Trust, v. 65, n. 2, p. 297–303, 1978.

METCALFE, R. Metcalfe's law after 40 years of ethernet. **Computer**, IEEE, v. 46, n. 12, p. 26–31, 2013.

ODLYZKO, A.; TILLY, B. **A refutation of Metcalfe's Law and a better estimate for the value of networks and network interconnections**. 2005. Acesso em: 16 jun. 2016. Available from Internet: <http://www.dtc.umn.edu/~odlyzko/doc/metcalfe.pdf>.

PFAFF, B. **Analysis of Integrated and Cointegrated Time Series with R**. Second. New York: Springer, 2008. ISBN 0-387-27960-1. Available from Internet: <http://www.pfaffikus.de>.

PHILLIPS, P. C.; PERRON, P. Testing for a unit root in time series regression. **Biometrika**, Biometrika Trust, v. 75, n. 2, p. 335–346, 1988.

POTOK, T. E.; VOUK, M.; RINDOS, A. Productivity analysis of object-oriented software developed in a commercial environment. **Software-Practice and Experience**, London, New York, Wiley Interscience [etc.], v. 29, n. 10, p. 833–848, 1999.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2015. Available from Internet: <https://www.R-project.org/>.

RAO, C. R. **Linear statistical inference and its applications**. [S.l.]: John Wiley & Sons, 2009.

REED, D. **That Sneaky Exponential - Beyond Metcalfe's Law to the Power of Community Building**. 1999. Acesso em: 16 jun. 2016. Available from Internet: <http://www.reed.com/dpr/locus/gfn/reedslaw.html>.

ROHLFS, J. A theory of interdependent demand for a communications service. **The Bell Journal of Economics and Management Science**, JSTOR, p. 16–37, 1974.

SHUMWAY, R. H.; STOFFER, D. S. **Time series analysis and its applications: with R examples**. [S.l.]: Springer Science & Business Media, 2010.

SINHA, P.; RAZ, D.; CHOUDHURI, N. Estimation of network distances using off-line measurements. **Computer communications**, Elsevier, v. 29, n. 16, p. 3295–3305, 2006.

STEIN, Y. **The Value of Being Linked In**. 2009. Acesso em: 16 jun. 2016. Available from Internet: <https://www.researchgate.net/profile/Yaakov_Stein/publication/235625563_The_Value_of_Being_Linked_In/links/0912f511ff0975431a000000.pdf>.

SWANN, G. P. The functional form of network effects. **Information Economics and Policy**, Elsevier, v. 14, n. 3, p. 417–429, 2002.

TRAPLETTI, A.; HORNIK, K. **tseries: Time Series Analysis and Computational Finance**. [S.l.], 2016. R package version 0.10-35. Available from Internet: <http://CRAN.R-project.org/package=tseries>.

WICKHAM, H. Reshaping data with the reshape package. **Journal of Statistical Software**, v. 21, n. 12, p. 1–20, 2007. Available from Internet: <http://www.jstatsoft.org/v21/i12/>.

WICKHAM, H. **ggplot2: Elegant Graphics for Data Analysis**. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. Available from Internet: <http://ggplot2.org>.

WOOLDRIDGE, J. **Introductory econometrics: A modern approach**. [S.l.]: Nelson Education, 2015.

YULE, G. U. Why do we sometimes get nonsense-correlations between time-series?–a study in sampling and the nature of time-series. **Journal of the royal statistical society**, JSTOR, v. 89, n. 1, p. 1–63, 1926.

ZHANG, X.-Z.; LIU, J.-J.; XU, Z.-W. Tencent and facebook data validate metcalfe's law. **Journal of Computer Science and Technology**, Springer, v. 30, n. 2, p. 246–251, 2015.

ANEXO A - DADOS DA TENCENT

Tabela 1: Dados de usuários e receita

Data[1]	UAM (milhões)[2]	Receita (milhões USD)[3]
2004-06-30	110.1	32.7
2004-09-30	119	36.4
2004-12-31	135	38
2005-03-31	149.2	36.3
2005-06-30	173.1	40.3
2005-09-30	184.8	44.8
2005-12-31	201.9	53.2
2006-03-31	220.5	80.3
2006-06-30	224.2	88.2
2006-09-30	221.4	93.2
2006-12-31	232.6	91.3
2007-03-31	253.7	100
2007-06-30	273.2	114
2007-09-30	288.7	140.8
2007-12-31	300.2	153.6
2008-03-31	317.9	204.1
2008-06-30	341.9	233.2
2008-09-30	355.1	296.9
2008-12-31	376.6	306.9
2009-03-31	410.8	366.4
2009-06-30	448	421.3
2009-09-30	484.9	493.3
2009-12-31	522.9	540.2

(Fonte: http://www.tencent.com/en-us/ir/news/2016.shtml)

[1]Data de fechamento do relatório.
[2]Usuários Ativos Mensalmente no serviço de Mensagens Instantâneas.
[3]Receita total.

Tabela 2: Dados de usuários e receita - continuação

Data	UAM (milhões)	Receita (milhões USD)
2010-03-31	568.6	619.1
2010-06-30	612.5	687.6
2010-09-30	636.6	780
2010-12-31	647.6	834.1
2011-03-31	674.3	966.8
2011-06-30	701.9	1041.3
2011-09-30	711.7	1179.6
2011-12-31	721	1257.4
2012-03-31	751.9	1532.8
2012-06-30	783.6	1664.4
2012-09-30	783.9	1823.9
2012-12-31	798.2	1933.5
2013-03-31	825.4	2161.1
2013-06-30	818.5	2328.1
2013-09-30	815.6	2526.9
2013-12-31	808	2783
2014-03-31	848	2991
2014-06-30	829	3209
2014-09-30	820	3220
2014-12-31	815	3428
2015-03-31	832	3647
2015-06-30	843	3832
2015-09-30	860	4181
2015-12-31	853	4688
2016-03-31	877	4952

(Fonte: http://www.tencent.com/en-us/ir/news/2016.shtml)

ANEXO B - TRABALHO DE CONCLUSÃO I
Lei de Metcalfe e o Valor das Redes

Kazuki Yokoyama[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

kmyokoyama@inf.ufrgs.br

Abstract. *Metcalfe's Law, which states that the value of a network is proportional to the square of the number of its nodes, served as basis to quantify the growth of communication networks, in special, the boom of the Internet in the 1990s and the dot-com bubble in the early 2000s. Today, the future of the Web is discussed in the light of new models such as Web 2.0 and Semantic Web and many wonder whether we are not facing a new era of speculations. Besides Metcalfe's growth model, many others theoretical models have been proposed and critized, but little empirical work has been done so far. The proposal of this work is to evaluate such models in a social network scenario by utilizing actual data from Facebook, currently the world's largest social network, and Tencent, China's largest social network.*

Resumo. *A lei de Metcalfe, que diz que o valor de uma rede é proporcional ao quadrado do seu número de nodos, serviu como base para quantificar o crescimento das redes de comunicação, em especial o grande boom da Internet na década de 1990 e a bolha das empresas dot-com até início dos anos 2000. No momento, se discute o futuro da Web à vista de novos modelos como a Web 2.0 e Semantic Web e muitos se perguntam se não estamos diante de uma nova era de especulações. Além do modelo de crescimento de Metcalfe, outros modelos teóricos foram propostos e criticados, porém pouco trabalho empírico foi realizado até o momento. A proposta deste trabalho é avaliar tais modelos no cenário de redes sociais utilizando dados reais do Facebook, atualmente a maior rede social do mundo, e da Tencent, maior rede social da China.*

1. Introdução

No universo da computação, "leis" que tentam prever o futuro são encontradas com certa facilidade. Na verdade, essas "leis" são mais profecias baseadas na observação do que uma verdade universal, algo que poderia realmente se chamar de lei. No entanto, elas continuam vivas, algumas sobrevivendo há décadas, se tornando parte da cultura corrente.

Algumas dessas leis são bastante conhecidas como, por exemplo, a lei de Moore. Ela recebe o nome de Gordon Moore, cofundador da Intel, que publicou um artigo entitulado "*Craming More Components onto Integrated Circuits*" em 1965 (Moore 1965). Ainda que Moore não tenha utilizado a palavra "lei" uma vez sequer em seu artigo, a sua observação ganhou notoriedade e ainda hoje persiste como referência para a evolução da tecnologia.

Outras leis são menos famosas como a lei de Wirth e a lei de Nielsen. A primeira foi popularizada por Niklaus Wirth, inventor da linguagem Pascal, em 1995 e declara que a execução do *software* está ficando mais lenta do que a aceleração do *hardware*. A segunda, formulada em 1998 por Jakob Nielsen, afirma que a velocidade de conexão dos usuários finais com a Internet cresce em torno de 50% ao ano (Ross 2003).

Várias outras leis epônimas foram concebidas ao longo da breve história da tecnologia (Amdahl, Rock, Grosch, para citar algumas) e, mesmo que entre erros e acertos, levaram a reflexões dos "comos" e "porquês" dessa rápida evolução.

Uma das leis mais controversas e discutidas é a lei de Metcalfe. Desde sua popularização em 1993 por George Gilder até os dias de hoje, a lei é alvo de críticas e estudos e um consenso não parece estar próximo. Ela afirma que o valor de uma rede varia com o quadrado do seu número de nodos (Gilder 1993). Ela também serve como referência quantitativa para o chamado *efeito de rede* que é uma externalidade positiva da formação de redes (GANGMIN 2008), (HENDLER et al. 2008).

O efeito de rede ou externalidade de rede se caracteriza pelo valor de um produto depender do número de usuários existentes (VARIAN et al. 1999) e foi a ideia propulsora por trás dos massivos investimentos durante o boom da Internet na década de 1990 (BRISCOE et al. 2006), (LASETER et al. 2011). Esse é um dos conceitos fundamentais da chamada *nova economia* (JANSEN 2006), (CUELLAR 2002).

Hoje, se observa o crescimento acelerado de empresas do setor tecnológico, sobretudo daquelas ligadas à *Web* e mídias digitais. Um claro exemplo disso foi a discussão em torno do valor de IPO do Facebook na NASDAQ em 2012 que, por muitas vezes, foi superestimado (REUTERS 2012).

Frente ao papel implícito, porém importante que a lei de Metcalfe desempenha nesse novo cenário, é natural que se levante a questão se esse modelo é capaz de explicar adequadamente o crescimento de redes de mídias sociais. No entanto, essa lei não é a única que afirma quantificar o crescimento de redes. Diferentes pesquisadores também propuseram outros modelos como o de Sarnoff, Reed e, como será chamado aqui, Odlyzko-Tilly.

Este trabalho objetiva avaliar teoricamente cada um dos modelos já propostos e suas implicações, porém não se limitando a isso. Os dados reais trimestrais de receita e quantidade de usuários ativos do Facebook e da Tencent, duas redes sociais de grande escala, serão explorados a fim de se avaliar o modelo de dependência entre eles. Esses dados podem ser obtidos de forma simples através dos relatórios financeiros trimestrais dessas empresas já que são cotadas em bolsa de valores e portanto mantêm portais com o investidor.[1]

Tanto a receita quanto a quantidade de usuários são exemplos de séries temporais e o devido tratamento deve ser dado. Analisá-los sob a ótica de dados de corte transversal seria incorreto, pois ignoraria as relações intertemporais existentes nos dados e o resultado não refletiria a verdadeira relação entre as duas séries. O método de séries temporais, ainda que necessário, certamente adiciona complexidade ao tratamento dos dados, algo

[1]Facebook: http://investor.fb.com/releases.cfm
Tencent: http://www.tencent.com/en-us/ir/news/2015.shtml

que foi deixado de lado em outros trabalhos como (METCALFE 2013) e (ZHANG et al. 2015) e que não se pretende repetir aqui.

Esta primeira parte do trabalho consiste na apresentação dos modelos, de alguns trabalhos anteriores e na proposta de uma avaliação dos dados disponíveis. Na segunda parte do trabalho, serão apresentadas as técnicas de séries temporais, as técnicas de regressão a fim de verificar a relação entre as variáveis juntamente com todas as hipóteses envolvidas, a análise real dos dados e, por fim, conclusões.

Este trabalho é organizado como segue. A seção 2 introduz de forma mais detalhada a lei de Metcalfe, parte de sua história e os demais modelos existentes. A seção 3 discute sobre alguns trabalhos publicados no assunto analisando as técnicas empregadas e suas implicações. A seção 4 explica de forma sucinta a metodologia e os objetivos da segunda parte do trabalho, bem como o cronograma de atividades. A seção 5 faz as considerações finais. A bibliografia consultada se encontra na seção 6.

2. A lei de Metcalfe e outros modelos

Metcalfe não foi o primeiro a considerar o efeito de rede. Já em 1974, Jeffrey Rohlfs fez um trabalho sob a ótica da microeconomia onde discute as externalidades da rede, a utilidade derivada pelo usuário dessa rede e suas implicações, por exemplo, a precificação de serviços de comunicações (ROHLFS 1974).

Diversos outros modelos foram propostos e a diferença entre um e outro é a taxa de crescimento do valor em relação ao seu número de nodos. Enquanto Metcalfe considera um crescimento quadrático do valor, outros, como Sarnoff, sugerem um crescimento linear, mais modesto. Recentemente, foi proposto um modelo logarítmico que reside entre o linear e o quadrático (BRISCOE et al. 2006). Levando em consideração a capacidade de se formar subgrupos em uma rede, um modelo de variação exponencial surge. Esses modelos são conhecidos na literatura como lei de Metcalfe, lei de Sarnoff, lei de Odlyzko-Tilly e lei de Reed.

Uma maior atenção será dada a lei de Metcalfe a seguir por ser o modelo em maior evidência e o objetivo inicial deste trabalho.

2.1. A lei de Metcalfe

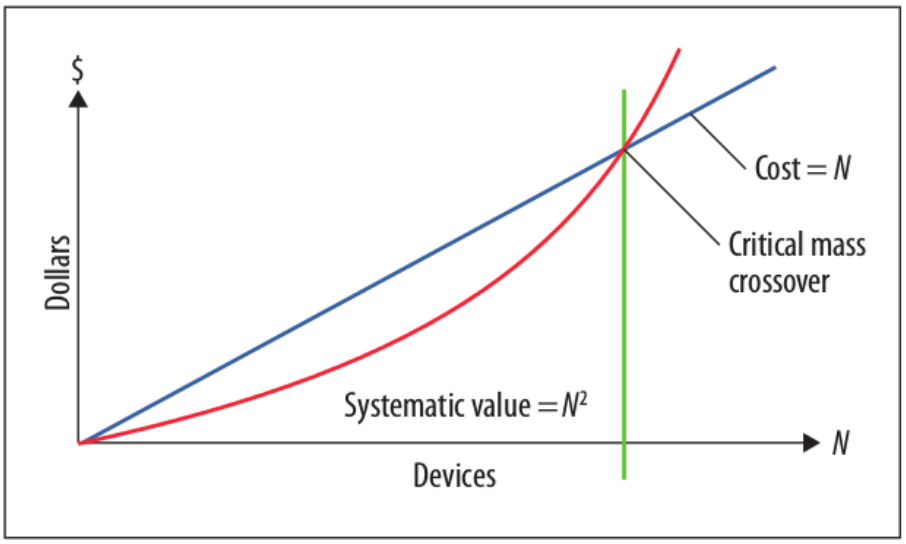
Em 1973, a Ethernet surgiu em um memorando escrito por Robert Metcalfe que circulou pelos laboratórios da Xerox Palo Alto Research Center (PARC) mostrando como uma rede LAN (*Local Area Network*) poderia funcionar (METCALFE 2013).

Alguns anos depois, Bob Metcalfe - cofundador e então vice presidente de vendas e marketing da 3Com - postulou que o valor de uma rede é proporcional ao quadrado do seu número de nodos. Além disso, observou que o custo dessa rede é linearmente proporcional ao seu número de nodos. Isso sendo verdade, deveria existir um número de nodos tal que o valor da rede se igualiaria ao seu custo e desse momento em diante, o valor da rede seria superior ao seu custo. Essa quantidade de nodos iria compor a massa crítica.

Inicialmente, suas afirmações serviram com o propósito comercial de convencer investidores a comprarem placas de comunicação Ethernet além da massa crítica que se acreditava ser em torno de 30 nodos, o que acabou funcionando. Munida de um *slide*

de 35mm que representa graficamente o que viria a ser chamada lei de Metcalfe, a força tarefa da 3Com fez com que as vendas passassem de algumas centenas de placas vendidas por mês para milhares (METCALFE 2013).

Figura 1: Relações valor e gasto por nodos na rede



Fonte: (METCALFE 2013).

Mesmo depois do sucesso da Ethernet, sua mensagem sobre o crescimento do valor de uma rede continuou causando influência no ramo. Reed Hundt, na época presidente da Comissão Federal de Comunicações dos Estados Unidos, afirmou que a lei de Metcalfe e a lei de Moore nos dão a melhor base para entendermos a Internet (KAPROWSKI 1996 apud ODLYZKO et al. 2005). Na mesma linha de pensamento, Marc Andressen, coautor do primeiro grande navegador da Web Mosaic e cofundador do Netscape, atribuiu o rápido desenvolvimento da Web, por exemplo com o crescimento da base de assinantes da AOL, à lei de Metcalfe (BRISCOE et al. 2006).

Em julho de 2005, a Rupert Murdoch's New Corp adquiriu a Intermix Media, dona do MySpace.com, por US$580 milhões e, em outubro do ano seguinte, o Google comprou o site de vídeos YouTube.com por US$1.65 bilhões. Apenas dois meses depois da venda do YouTube.com, a AT&T se fundiu com a BellSouth e se tornou controladora da Cingular, de provisionamento de redes cabeadas e sem fio, por US$86 bilhões (WEINMAN 2007). Esses são alguns exemplos das dimensões que empresas de tecnologia, principalmente da área de comunicações e mídias digitais, tomaram nos anos recentes.

O princípio por trás do pensamento de Metcalfe é de que, numa rede conectada, o valor percebido por cada um dos n nodos é proporcional a quantidade de outros nodos, ou seja, $n - 1$. Sendo assim, o valor total percebido na rede seria $n(n - 1)$ ou, aproximadamente, n^2 (METCALFE 2013). Visto de outra forma, em uma rede com n nodos, a quantidade total de conexões possíveis entre quaisquer dois nodos é $n(n - 1)/2$ que tem ordem n^2 assintoticamente (BRISCOE et al. 2006). Uma hipótese implícita do modelo é que cada novo nodo introduz o mesmo valor.

Essa hipótese é muito forte e já foi apontada como a principal falha do modelo

antes (ODLYZKO et al. 2005), (BRISCOE et al. 2006). Uma análise econômica mais detalhada pode ser encontrada em (ROHLFS 2001).

$$\text{Lei de Metcalfe: } V \propto N^2$$

onde V é o valor da rede e N é a sua quantidade de nodos. Até o restante do trabalho, essa notação será seguida.

Como exemplo, considere uma dada rede que possuindo 100 usuários é avaliada em $\$1.000,00$. Se seu valor tiver o crescimento como descrito pela lei de Metcalfe, ao alcançar 110 usuários, seu valor será de $\$1.000,00(110^2/100^2) = \$1.210,00$. Cabe observar aqui que essas relações não fazem distinção da unidade monetária adotada e o símbolo $\$$ será utilizado daqui em diante por mera convenção.

A lei de Metcalfe será apresentada com mais detalhes na segunda parte do trabalho.

2.2. A lei de Sarnoff

A lei de Sarnoff apresenta a menor taxa de crescimeto dentre as avaliadas neste trabalho. Ela sugere que o valor da rede é diretamente proporcional à sua quantidade de nodos. Isso é particularmente verdade para redes de *broadcast* onde há somente um nodo emissor e os demais são apenas receptores.

Nesse tipo de rede, faz sentido pensar que, para cada novo nodo receptor adicionado à rede, o valor total aumente na mesma proporção. Porém, em redes onde a comunicação é bilateral, como aquelas que este trabalho tratará, é razoável imaginar que cada nodo perceba valor também de sua comunicação com os demais nodos.

Esse modelo de crescimento foi primeiro proposto por David Sarnoff, considerado o Pai da Televisão Americana (METCALFE 2013).

$$\text{Lei de Sarnoff: } V \propto N$$

Novamente considerando a rede com 100 nodos e mensurada em $\$1.000,00$, dessa vez, ao chegar aos 110 nodos, seu valor será de $\$1.000,00(110/100) = \$1.100,00$, abaixo daquele encontrado no caso da lei de Metcalfe como esperado.

2.3. A lei de Reed

A lei de Reed, como ficou conhecida, é fruto de algumas observações sobre a formação de grupos em redes por David Reed. Em seu artigo "*That Sneaky Exponential - Beyond Metcalfe's Law to the Power of Community Building*", Reed argumenta que muito valor pode ser obtido a partir da formação de grupos nas redes, se esta suportar tal formação. Exemplos de grupos são listas de emails, salas de conversa online e grupos de discussão (REED). As redes que suportam grupos dessa forma foram chamadas de GFN[2].

Em termos de valor da rede, a lei afirma que GFN's podem ter crescimento exponencial com o número de nodos. A ideia por trás dessa afirmação é de que, numa rede com N nodos, a quantidade total de subconjuntos que se pode formar é dada por:

[2]do inglês, *Group Forming Networks*.

$$C_N^0 + C_N^1 + C_N^2 + C_N^3 + ... + C_N^N = 2^N$$

Retirando-se os subconjuntos triviais, ou seja, aqueles sem participantes e aqueles com apenas um participante representados respectivamente por C_N^0 $(= 1)$ e C_N^1 $(= N)$, restam $2^N - N - 1$ subgrupos não-triviais, que tem crescimento exponencial assintoticamente.

Lei de Reed: $V \propto 2^N$

Como foi feito anteriormente para a lei de Metcalfe e de Sarnoff, considere a rede de 100 nodos, valor $\$1.000,00$ e que agora ela permite que seus nodos formem grupos. Seu valor, ao possuir 110 nodos, chegará a $\$1.000,00(2^{110}/2^{100}) = \$1.024.000,00$, o que é um aumento bastante expressivo no valor considerando uma adição de apenas 10 novos nodos.

2.4. Lei de Odlyzko-Tilly

O modelo apresentado a seguir será chamado aqui de lei de Odlyzko-Tilly - ainda que não haja um consenso sobre o nome - por ser uma contribuição dos pesquisadores Andrew Odlyzko e Benjamin Tilly (ODLYZKO et al. 2005). Essa lei se difundiu publicamente através de um artigo publicado na IEEE Spectrum que possui contribuição de Bob Briscoe, além dos dois autores anteriores (BRISCOE et al. 2006).

A lei se baseia na ideia da "longa cauda"[3] descrita quantitativamente pela lei de Zipf. A lei de Zipf é outra observação empírica capaz de explicar uma vasta gama de fenômenos muito bem. Ela diz que elementos de uma dada coleção, uma vez ordenados por algum critério, apresentam valores sempre decrescentes proporcionais a $1/n$ sendo n a posição do elemento na ordenação (BRISCOE et al. 2006).

Para cada um dos N diferentes nodos da rede, o valor de cada um dos $N-1$ nodos restantes seguiria a lei de Zipf. Explicando melhor, considere um nodo dessa rede, n_1. Ordenando de forma decrescente os outros $N - 1$ nodos pelo valor de sua conectividade com n_1, teríamos que o primeiro nodo ordenado teria valor proporcional a 1, o segundo $1/2$, o terceiro $1/3$ e assim por diante até o n-1-ésimo nodo que contribuiria com valor proporcional a $1/(n-1)$. Dessa forma, o valor total percebido por n_1, V_{n_1} seria:

$$V_{n_1} = \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{N-1}$$

Sabendo que

$$\lim_{N\to\infty} (\sum_{k=1}^{N} 1/k - ln(N)) = \gamma$$

onde γ é a constante de Euler-Mascheroni (≈ 0.577215) e $ln(N)$ denota o logarítmo natural de N, pode-se considerar que $V_{n_1} \to ln(N)$ para N suficientemente grande. Como temos N nodos na rede, se repetiria o processo para os demais nodos (n_2, n_3,..., n_N) encontrando os valores percebidos por eles (V_{n_1}, V_{n_2},..., V_{n_N}) e portanto o valor total seria $Nln(N)$.

[3] *the long tail*, do inglês.

Lei de Odlyzko-Tilly: $V \propto Nln(N)$

Voltando ao exemplo numérico da rede com 100 nodos e $\$1.000,00$, tem-se que o valor dessa rede, se a lei de Odlyzko-Tilly valer, será de $\$1.000,00[(110ln(110))/(100ln(100))] = \$1.122,77$ quando for composta por 110 nodos.

Os quatro modelos estão resumidos na *Tabela 1*. Note que foi utilizado o sinal de proporcionalidade ($\propto$) para relacionar o valor e a quantidade de nodos em cada modelo. Para que a proporção se torne uma igualdade, é preciso que se introduza em cada modelo uma constante de proporcionalidade como será feito mais a frente.

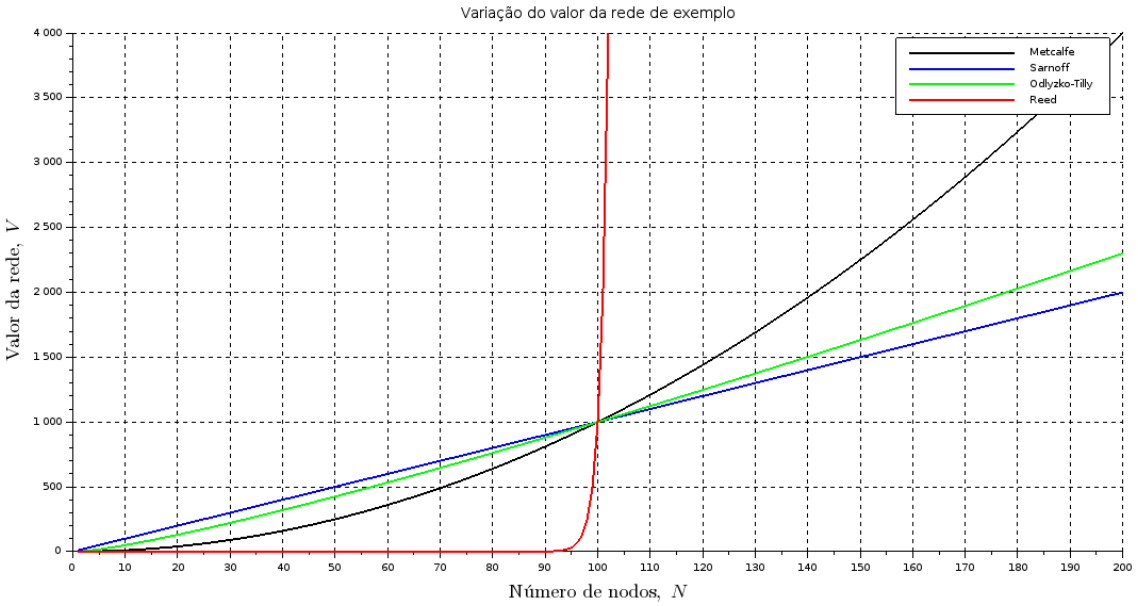
Tabela 1: Diferentes leis e suas relações de proporção

Modelo	**Valor, V, em função da quantidade de nodos, N**
Lei de Metcalfe	$V \propto N^2$
Lei de Sarnoff	$V \propto N$
Lei de Reed	$V \propto 2^N$
Lei de Odlyzko-Tilly	$V \propto Nln(N)$

Fonte: Próprio autor.

Para dar uma visão geral da diferença entre os quatro modelos, a rede dos exemplos anteriores (100 nodos e avaliada em $\$1.000,00$) teve sua curva Valor x Nodos traçada seguindo cada um dos modelos e é mostrada abaixo:

Figura 2: As quatro leis aplicadas à rede dos exemplos anteriores



Fonte: Próprio autor.

Repare como todas as curvas passam pelo ponto de 100 nodos e $\$1.000,00$, como

nos exemplos dados.[4] Como esperado, a lei de Odlyzko-Tilly possui um crescimento nem tão rápido quanto a lei de Metcalfe, nem tão lento quanto a lei de Sarnoff como se pode observar para valores acima de 100 nodos.

3. Trabalhos anteriores

Brevemente, serão abordados dois trabalhos que se propuseram a avaliar com dados reais a validade de uma ou mais leis dentre as discutidas. O primeiro artigo é do próprio Robert Metcalfe publicado na IEEE Spectrum na ocasião do 40º aniversário da Ethernet. O segundo se deve a três pesquisadores de universidades chinesas que verificaram a adequação dos modelos aos dados reais do Facebook e da Tencent (ZHANG et al. 2015), as mesmas empresas que serão analisadas aqui.

Em (METCALFE 2013), Bob Metcalfe ressalta que, apesar de sua lei estar entre nós há anos, nenhum trabalho havia proposto uma análise com dados reais seja para refutar seu modelo ou não. Com efeito, entre histórias pessoais e curiosidades da evolução da Ethernet, ele mesmo esboçou uma investigação através de dados de uma década de Facebook (de 2004 a 2013). Ambos utilizaram as respectivas receitas como variáveis *proxies*[5] para o valor da rede.

De posse da quantidade de usuários da rede social e de sua receita em dez anos, Metcalfe modelou o crescimento do número de usuários através do que ele chamou de função *netoid* e a evolução da receita como uma função quadrática do número de usuários tendo como parâmetro variável a constante de proporcionalidade. Na verdade, o que ele chamou de função *netoid* é a função de crescimento logístico bastante empregada na modelagem de crescimentos populacionais cuja forma geral é $Y_t = \frac{\beta_1}{1+e^{-(\beta_2+\beta_3 t)}} + u_t$, onde β_1, β_2 e β_3 são os parâmetros do modelo, t o tempo e u_t o termo de erro estocástico (GUJARATI 2004).

No entanto, a forma como conduziu sua análise prática careceu de alguns cuidados técnicos. Para a regressão entre a receita e a quantidade de usuários, apenas variou a constante de proporcionalidade, k_M, do modelo mostrado abaixo para que se ajustasse visualmente bem aos dados da amostra:

$$V = k_M N^2$$

O mesmo método "visual" foi utilizado para achar os parâmetros β_1, β_2 e β_3, os quais ele chamou de p, h e v respectivamente como mostrado no modelo abaixo:

$$Netoid = \frac{p}{1 + e^{-(vt+h)}}$$

[4]Os valores muito baixos para a lei de Reed (até $N \approx 90$) se devem ao fato do coeficiente de proporcionalidade ($1.000/2^{100} \approx 0$) ser muito baixo para esse exemplo.

[5]Nem sempre podemos observar determinadas variáveis diretamente de forma que devemos buscar outras variáveis que desempenhem o papel daquelas no modelo. A essas dá-se o nome de variável *proxy*.

Claramente, um exame visual dos dados pode revelar bastante informação importante e fornecer indicativos de como proceder na análise. Não obstante, variar parâmetros de um modelo até que se ajustem visualmente bem a uma amostra não é o procedimento mais correto para verificar a adequação de tal modelo à realidade. Quaisquer estimativas resultantes de tal procedimento são duvidosas e não oferecem bases sólidas para se fazer inferências ou previsões. Existem técnicas estatísticas capazes de fornecer bons estimadores[6] para esses modelos e elas devem ser usadas sempre que necessário.

Uma nova investigação foi proposta em (ZHANG et al 2015). Nesse artigo, os modelos de Metcalfe, Reed, Sarnoff e, como eles chamaram, Odlyzko são avaliados. Para isso, coletaram dados de número de usuários ativos mensais (MAUs[7]), receita e gastos de onze anos das redes sociais Facebook e Tencent (de 2003 a 2013). Como *proxy* para os gastos, foi utilizada a diferença entre a receita e o lucro líquido (*revenue - net profit*).

O procedimento utilizado para analisar a relação receita vs número de usuários foi simples. Primeiro, fizeram regressão simples através do método de Mínimos Quadrados Ordinários (MQO) de receita contra o número de usuários obtendo o valor dos parâmetros dos modelos. Segundo, escolheram o "melhor" modelo como aquele possuindo o menor valor RMSD (*root mean square deviation*)[8]. Concluíram assim que o modelo que melhor explica a evolução da receita pelo número de usuários é o de Metcalfe (ZHANG et al 2015).

A tabela abaixo mostra os modelos analisados. Em comparação com a *Tabela 1*, transformou-se cada relação de proporcionalidade em uma igualdade através da introdução de uma constante, o parâmetro a ser estimado do modelo.

Tabela 2: Diferentes leis e suas formas funcionais

Modelo	**Valor, V, em função da quantidade de nodos, N**
Lei de Metcalfe	$V = k_M N^2$
Lei de Sarnoff	$V = k_S N$
Lei de Reed	$V = k_R 2^N$
Lei de Odlyzko-Tilly	$V = k_{OT} N ln(N)$

Fonte: Próprio autor.

Entretanto, existem algumas hipóteses que devem ser atendidas para que as propriedades de MQO desejadas na regressão tenham validade (GUJARATI 2004).

Apesar de utilizar MQO na estimação dos parâmetros, (ZHANG et al. 2015) não atenta a tais suposições e não se sabe até que ponto elas são atendidas. Como consequência, não se pode ter certeza da validade dos resultantes.

Cabe ressaltar que a utilização do valor de RMSD assume que os erros sejam não viesados e tenham distribuição normal (CHAI et al 2014), no entanto estas hipóteses também não foram verificadas.

[6] O conceito de "bom" será tratado com mais detalhes na segunda parte do trabalho.
[7] do inglês, *Monthly Active Users*.
[8] Também chamado de *root mean square error* - RMSE.

Uma última observação: ambos trabalhos ignoram o fato de os dados de receita e número de usuários observados serem dados de séries temporais e não de corte transversal. Isso significa que uma análise prévia importante dos dados não foi realizada nos dois trabalhos citados. Por exemplo, para que a relação entre as duas variáveis possa ser estimada por MQO, é necessário que se verifique que ambas séries são estacionárias. Caso não o sejam, a regressão pode produzir resultados espúrios (HILL et al 2010). No entanto, em alguns casos, ainda é possível realizar regressão através de transformações adequadas às séries ou observando que elas são cointegradas.

4. Metodologia e organização

Na sequência deste trabalho, os modelos apresentados anteriormente serão avaliados com as amostras de dados obtidas das empresas Facebook e Tencent. Até o momento, estão disponíveis 12 amostras da primeira[9] e 44 da segunda[10] com periodicidade trimestral em ambas. É possível que esses números sejam maiores na ocasião da segunda parte do trabalho.

Tentar-se-á verificar se algum dos modelos descreve de forma razoável a realidade do valor das redes à luz das ferramentas estatísticas disponíveis e a abordagem de séries temporais será utilizada na modelagem dos dados.

Para isso, primeiro se analisará a natureza dos dados e se tentará modelar as séries de receita e número de usuários através da metodologia Box-Jenkins. Durante esse procedimento, questões relativas à estacionariedade, existência de tendência e ordem de integração, por exemplo, serão levantadas e discutidas. É possível que as poucas observações em cada amostra configurem um problema, de forma que os testes estatísticos feitos possam ser tornar pouco confiáveis. Na impossibilidade de se obter um teste definitivo, em muitos casos se recorrerá a mais de uma ferramenta estatística de mesmo propósito comparando os múltiplos resultados achados a fim de se extrair o melhor indicativo possível.

Depois, as regressões para os modelos apresentados serão feitas, se possível, e seus resultados avaliados. As hipóteses para que as regressões tenham validade serão também discutidas. Dada a natureza do problema e sua determinação no chamado *efeito de rede*, espera-se que haja problema de equações simultâneas, ou seja, o número de usuários influencia o valor da rede ao mesmo tempo em que o valor da rede influencia a decisão de novos usuários de se juntarem a ela.

Na segunda parte do trabalho, as seguintes tarefas serão realizadas:

1. **Estudo**: para avaliar de forma correta as relações entre o valor de uma rede e sua quantidade de nodos (usuários no contexto de redes sociais), diversas técnicas específicas deverão ser usadas e seus resultados julgados. Para tanto, nesta etapa serão estudados o embasamento teórico de tais técnicas, suas aplicações e limitações. Alguns assuntos importantes nesse desenvolvimento são o modelo clássico de regressão linear (englobando todas as hipóteses, o problema da inferência e casos especiais) e a modelagem de séries temporais para o tratamento adequado dos dados.

[9] Do segundo trimestre de 2012 ao primeiro trimestre de 2015.
[10] Do segundo trimestre de 2004 ao primeiro trimestre de 2015.

2. **Análise**: nesta etapa, as técnicas estudadas anteriormente serão utilizadas na extração de informações das amostras. Será empregada a linguagem *R* na manipulação e análise dos dados. Essa escolha poupará tempo de implementação de um *software* próprio que certamente poderia introduzir erros. Além disso, a linguagem *R* tem sido largamente empregada pela comunidade científica em análises estatísticas e gráficas.
3. **Avaliação**: uma vez obtidos os resultados, resta avaliá-los e explicá-los. Nesta etapa, cabe interpretar, no contexto do problema exposto, o que os dados revelaram e o que isso pode implicar.

Para a realização das tarefas, é proposto o seguinte cronograma:

1. Estudo do modelo de regressão linear clássico.
2. Estudo da modelagem de séries temporais e de regressão nesses casos.
3. Análise exploratória dos conjuntos reais de dados.
4. Avaliação e explicação dos resultados.
5. Redação da segunda parte deste trabalho.
6. Apresentação do Trabalho de Graduação 2.

Tabela 3: Cronograma para a segunda parte do Trabalho de Graduação

Tarefa	2015						
	Junho	Julho	Agosto	Setembro	Outubro	Novembro	Dezembro
1	X						
2		X	X				
3			X	X			
4				X	X		
5					X	X	
6							X

Fonte: Próprio autor.

5. Considerações finais

Este artigo teve o objetivo de introduzir o tema da relação entre valor observado pela formação de uma rede e o número de elementos que a compõem. Apesar da origem da discussão remontar ao ínicio da era das telecomunicações, o assunto ainda é relevante pois, como nunca antes, as redes de comunicações se tornaram necessárias e ubíquas. Assim, entender o que leva essas formações a assumirem papéis cada vez mais importantes e como esse processo ocorre é de grande interesse tanto para a criação de novas redes, como para a manutenção daquelas já existentes e também na explicação do fracasso de muitas outras. Trata-se portanto de um trabalho de aspecto econômico de relevância para a computação.

Ainda que o assunto seja de interesse prático, pouco trabalho com dados do mundo real foi realizado, em oposição ao bastante que se fez em seus aspectos mais teóricos. Isso deixou uma lacuna entre teoria e observação que ainda não foi explorada em toda sua

capacidade. A intenção da segunda parte do trabalho é avaliar como a teoria proposta até então se encaixa à realidade observada.

Para isso, é necessário que se explore os dados da forma correta, utilizando os métodos de análise adequados. Isso significa que uma fração maior da teoria estatística, em comparação com aquela empregada nos trabalhos anteriores, deve ser aplicada, o que torna o trabalho certamente mais complexo mas, ao mesmo tempo, é capaz de entregar resultados mais confiáveis.

6. Bibliografia

ANDERSON, C. The Long Tail, **Wired**, Issue 12.10, out. 2004. Disponível em: < http://archive.wired.com/wired/archive/12.10/tail.html >. Acesso em: 28 mai. 2015.

BRISCOE, B.; ODLYZKO, A; TILLY, B. Metcalfe's law is wrong, **IEEE Spectrum**, [s.l.], v. 43, n. 7, p. 34-39, jul. 2006.

CHAI T.; DRAXLER R. R. Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? - Arguments Against Avoiding RMSE in the Literature, **Geoscientific Model Development**, vol. 7, n. 3, p. 1247-1250, fev. 2014.

CUELLAR, S. S. The New Economy, Network Effects and Market Structure, **National Business and Economic Society Conference**, Maui, Havaí, mar. 2002.

GANGMIN, L. Economic Sense of Metcalfe's Law, **Proceedings of 17th International World Wide Web Conference**, Pequim, China, abr. 2008.

GILDER, G. **Metcalfe's Law and Legacy**, Forbes ASAP, set. 1993.

HILL, R. C; GRIFFITHS W. E.; JUDGE G. G. **Econometria**, 3ª edição, Saraiva, 2010.

GUJARATI, D. N. **Basic Econometrics**, 4ª edição, McGraw Hill, 2004.

HENDLER, J.; GOLDBECK, J. Metcalfe's Law, Web 2.0, and the Semantic Web. **Journal of Web Semantics**, [s.l], vol. 6, n. 1, p. 14-20, 2008.

JANSEN, D. W. **The New Economy and Beyond, Past, Present and Future**, Texas, EUA, Bush Series in the Economics of Public Policy, Edward Elgar Publishing, p. 98, 2006.

KAPROWSKI, G. FCC Chair wants universal Net access - and he's serious, **Wired News**, nov. 1996.

LASETER, T. M; RABINOVICH, E. **Internet Retail Operations: Integrating Theory and Practice for Managers**, [s.l.], CRC Press, 2011.

METCALFE B. Metcalfe's law after 40 year of Ethernet, **IEEE Computer**, [s.l.], v. 46, n. 12, p. 26-31, dez. 2013.

MOORE G. Cramming More Components onto Integrated Circuits, **Electronics**, [s.l.], v. 38, n. 8, p. 114-117, abr. 1965.

ODLYZKO, A.; TILLY, B. **A refutation of Metcalfe's Law and a better estimate for the value of networks and network interconnections**, Minneapolis, EUA, mar. 2005.

ORAN, O.; Barr, A. **Facebook prices at top of range in landmark IPO**. Nova Iorque, São Francisco, EUA. 17 mai. 2012. Disponível em <

http://www.reuters.com/article/2012/05/17/us-facebook-idUSBRE84G14Q20120517 >. Acesso em: 28 mai. 2015.

POWERS, D. M. W. Applications and Explanations of Zipf's Law, **Proceedings of the Joint Conference on New Methods in Language Processing and Computational Natural Language Learning, Association for Computation Linguistics**, p. 151-160, Somerset, Nova Jersey, 1998.

REED, DAVID P. **That Sneaky Exponential - Beyond Metcalfe's Law to the Power of Community Building**. Disponível em < http://www.reed.com/dpr/locus/gfn/reedslaw.html >. Acesso em: 28 mai. 2015.

ROHLFS, J. H. **A Theory of Interdependent Demand for a Communications Service**, Bell Laboratories, New Jersey, EUA, 1974.

ROHLFS, J. H. **Bandwagon Effects and the Internet**, Strategic Policy Research Inc., Maryland, EUA, 2001.

VARIAN, H. R.; SHAPIRO, C. **Information Rules: A Strategic Guide to the Network Economy**, [s.l], Harvard Business Press, 1999.

WEINMAN J. Is Metcalfe's Law Way Too Optimisc? **Business Communications Review**, [s.l.], p. 18-27, ago. 2007.

ZHANG, X. Z.; LIU, J. J.; XU, Z. W. Tencent and Facebook Data Validate Metcalfe's Law, **Journal of Computer Science and Technology**, vol. 30, n. 2, p. 246-251, mar. 2015.

Metcalfe's Law: Right? Wrong? **IEEE Spectrum**, [s.l.], v.43, n. 11, p. 10, nov. 2006.