APLICAÇÃO DE REGRESSÃO LINEAR MÚLTIPLA PARA DETERMINAÇÃO DOS GASTOS DE CLIENTES PARA EMPRESA DO SETOR VAREJISTA

Kellen Dayelle Endler (Universidade Federal do Paraná - UFPR) kellen_dayelle@hotmail.com,
Cassius Tadeu Scarpin (Universidade Federal do Paraná - UFPR) cassiusts@gmail.com
Maria Teresinha Arns Steiner (Universidade Federal do Paraná - UFPR) maria.steiner@ufpr.br
Tamires de Almeida Sfeir (Universidade Federal do Paraná - UFPR) tas.sfeir@gmail.com

Resumo

Entender o comportamento dos consumidores pode fazer toda diferença no momento de estabelecer as estratégias de negócios de empresas. O presente trabalho consiste de uma pesquisa exploratória que visa estimar os gastos de clientes de um grupo de lojas varejistas situadas no município de Curitiba-PR utilizando-se a técnica de Regressão Linear Múltipla. A pesquisa realizada partiu do pressuposto de que é possível prever a quantidade de gastos gerados pelos clientes em determinada loja, analisando suas características e seu perfil de consumo. Considerou-se como variável resposta (dependente), o gasto realizado pelo consumidor e como variáveis explicativas, as características dos clientes, como idade, sexo, endereço, renda, distância do cliente até a loja. Como resultados, obteve-se o coeficiente correlação de 0,5433, que mostra uma correlação moderada, indicando que parte dos valores gastos pelos clientes são explicados por outras variáveis que não constam no modelo. Os valores de erros para os valores de teste são MAE=119,1538 e MSE =214,2868. É importante observar que outras variáveis adicionais poderiam também ser levadas em consideração para contribuir com um modelo mais eficiente, como número de filhos e grau de escolaridade. Porém, tais informações ainda não são levantadas pelo cadastro da empresa. Estas limitações encorajam a realização de trabalhos futuros e a aplicação de outras técnicas de processamento, como as Redes Neurais Artificiais e Árvores de Decisão, por exemplo.

Palavras-Chaves: Regressão Linear Múltipla; Gastos de Clientes; Varejo.

1. Introdução

Em um mundo global e dinâmico, as empresas precisam competir para construir um relacionamento lucrativo e duradouro com os clientes. Analisar as características do cliente e suas preferências, bem como suas inter-relações consistem em temas cruciais na inteligência de negócios. Afinal, a identificação dos consumidores de acordo com suas características (de demografia, comportamentos, valores etc.) e perfil de consumo pode contribuir com as estratégias de *marketing* das organizações, de modo a direcionar os estímulos mais adequados a cada perfil.

Neste contexto, tem-se em vista por exemplo, o crescente número de consumidores que estão utilizando a internet para realizar suas compras. Os consumidores então, passaram a ter acesso a um grande volume de propagandas de produtos. Assim, é fundamental compreender como esses clientes se relacionam com produtos, marcas, direcionando seus produtos ou serviços, suas promoções e sua comunicação para essa audiência de forma segmentada.

Esta pesquisa se insere nesse contexto e tem por objetivo propor o uso da técnica de Regressão Linear Múltipla para determinar a relação das características de clientes com seus gastos. O estudo utiliza dados reais de um grupo de lojas varejistas situadas em Curitiba-PR. Com base em informações armazenadas em sua base de dados obtidas pelo do cadastro de clientes, como a idade, sexo, endereço, itens consumidos em cada compra procurou-se avaliar a associação entre essas variáveis, ou uma provável relação de causa e efeito entre elas com o valor gasto pelo cliente.

Esse artigo encontra-se organizado como segue: na seção 2 tem-se a revisão da literatura, que descreve a técnica da Regressão Linear Múltipla bem como os pressupostos necessários para que ela possa ser utilizada; na seção 3, descrevem-se a metodologia utilizada; na seção 4 apresentam-se e discutem-se os resultados da aplicação da metodologia proposta; e na seção 5 são apresentadas as principais conclusões.

2. Regressão Linear Múltipla

A origem do termo "regressão" deu-se pelo estatístico inglês Francis Galton quando estudou o relacionamento das alturas de pais e filhos (GUJARATI, 2006). As aplicações desta técnica são numerosas e ocorrem em quase todos os campos científicos, sendo que a seguir, é apresentada de forma resumida esta técnica.

A regressão múltipla envolve três ou mais variáveis, portanto, estimadores. Ou seja, ainda uma única variável dependente, porém duas ou mais variáveis independentes (explanatórias) (COELHO; CUNHA, 2009). A finalidade da análise de regressão, segundo os mesmos autores é estimar os valores da variável dependente com base nos valores conhecidos ou fixados das variáveis independentes.

De acordo com Hair Júnior et al. (2009), a regressão múltipla é um modelo realista porque, no mundo em que vivemos, a previsão quase sempre depende de vários fatores. Mesmo quando estamos interessados no efeito de apenas uma das variáveis, é aconselhável incluir as outras capazes de afetar a variável dependente, efetuando uma análise de regressão múltipla, por 2 razões:

a) Para reduzir os resíduos estocásticos. Reduzindo-se a variância residual (erro padrão da estimativa), aumenta a força dos testes de significância;
b) Para eliminar a tendenciosidade que poderia resultar se simplesmente ignorássemos uma variável que afeta Y substancialmente.

2.1 Modelo Linear Geral de Regressão

Um modelo geral de regressão linear pode ser descrito segundo Fávero et al. (2009) conforme a equação (1).

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 +_{\dots} + \beta_n X_n + \delta \quad (1)$$

Onde:

Y: variável dependente métrica;

α: intercepto do eixo y;

β_k: coeficientes angulares de cada variável ($k = 1,2, \dots, n$);

X_k: são as variáveis explicativas (métricas ou *dummies*);

δ: termo de erro.

Cabem aqui mencionar alguns pressupostos básicos a serem considerados segundo Cunha e Coelho (2009):

- A variável Y é aleatória;
- A esperança matemática dos resíduos é nula, ou seja, a média dos resíduos é nula;
- A variância de δ (termos de erro) é constante e igual a σ^2 (condição de homoscedasticidade dos resíduos). A homoscedasticidade (variância constante dos resíduos) é uma propriedade essencial, que deve ser garantida, sob pena de invalidar toda a análise estatística. Havendo heterocedasticidade, podem ser feitas transformações nas variáveis;
- Os resíduos são independentes entre si. A verificação da autocorrelação pode ser feita pela análise do gráfico dos resíduos cotejados com os valores preditos, onde este deve apresentar pontos dispersos aleatoriamente, sem nenhum padrão definido ou pelo teste de Durbin-Watson.
- Os resíduos têm distribuição normal (distribuição de Gauss). A verificação da normalidade pode ser feita pelos testes de aderência não paramétricos, como por exemplo, o de Kmogorov-Sminorv.

Assim como na regressão simples, o parâmetro α é calculado por meio do seu estimador a, conforme (2).

$$a = \overline{Y_l} - b_1 \overline{X}_{1i} -_{\dots} b_n \overline{X}_{ni} \quad (2)$$

Onde:

b_k:representa os estimadores dos betas para as n variáveis X utilizadas no modelo.

Já a estimativa dos betas segue-se o mesmo procedimento que para a regressão linear simples, mas é feito por etapa e por variáveis. A equação (3) apresenta a estimativa de β_1.

$$\beta_1 = \frac{cov(X_1, Y).Var(X_2) - cov(X_2, Y).cov(X_1, X_2)}{Var(X_1).Var(X_2) - [cov(X_1, X_2)]^2} \quad (3)$$

Enquanto uma regressão simples de duas variáveis resulta na equação de uma reta, um problema de três variáveis implica num plano, e um problema de k variáveis implica em um hiperplano.

2.2 Poder Explicativo do Modelo

Stock e Watson (2004) definem R^2 como "a fração da variância da amostra de Y_i explicada (ou prevista) pelos regressores". Fávero et al. (2009) descreve que a capacidade explicativa do modelo é analisada pelo R² da regressão, conhecido por coeficiente de ajuste ou de explicação. Ou seja, para se medir o quanto a variabilidade total dos dados é explicada pelo modelo de regressão, compara-se a Soma de Quadrados da Regressão com a Soma de Quadrados Total e tem-se o coeficiente de determinação ou de correlação múltipla ao quadrado R^2 conforme apresenta-se em (3) segundo Fávero et al. (2009).

$$R^2 = \left(\frac{\sum_{i=1}^{n}(X_i - \bar{X}).(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 . \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \right)^2 , 0 < R < 1 \quad (4)$$

O R² pode variar entre 0 e 1 (0 a 100%); porém, é praticamente impossível a obtenção de um R²=1. Quando o ajuste é bom o modelo explica boa parte da variação total e consequentemente o valor de R^2 é próximo de 1. O coeficiente de determinação é uma medida da qualidade do ajuste.

Quando há o intuito de comparar o coeficiente de ajuste (R^2) entre dois modelos ou entre um mesmo modelo com tamanhos de amostra diferentes, faz-se necessário o R^2 ajustado, cuja expressão está descrita em (5).

$$R^2 = 1 - \frac{n-1}{n-k}(1 - R^2) \tag{5}$$

Onde:

n: é o tamanho da amostra;

k: número de parâmetros do modelo de regressão (número de variáveis explicativas mais o intercepto);

2.3 Significância do modelo e das variáveis explicativas

Algumas inferências adicionais podem ser feitas a partir do modelo de regressão múltipla. Uma das mais comuns, para Hill et al. (2003), é o teste F, que auxilia na avaliação da significância do modelo. Em (6) estão as suas hipóteses nula e verdadeira.

$$\begin{gathered} H_0: \beta_1 = \beta_2 =_{...} = \beta_k = 0 \\ H_1: existe\ pelo\ menos\ um\ \beta_i \neq 0 \end{gathered} \tag{6}$$

Este teste auxilia na verificação inicial sobre a existência do modelo que está sendo usado, uma vez que se os betas forem estatisticamente iguais a zero, o comportamento de alteração de cada uma das variáveis explicativas não influenciará em absolutamente nada o comportamento da variável dependente. Este teste baseia-se na comparação da soma dos quadrados dos erros do modelo de regressão múltipla original, com a soma dos quadrados dos erros de um modelo de regressão em que se supõe verdadeira a hipótese nula (FÁVERO et al., 2009). A estatística F pode ser escrita utilizando o R^2 conforme em (7).

$$F = \frac{\frac{R^2}{k-1}}{\frac{1-R^2}{(n-k)}} \tag{7}$$

O teste F, porém não define qual, ou quais, das variáveis explicativas consideradas são estatisticamente significantes ou não para influenciar o comportamento da variável y.

Portanto, é preciso que se avalie se cada um dos parâmetros relacionados ao modelo é estatisticamente significante. A estatística t tem as hipóteses apresentadas em (8).

$$\begin{gathered} H_0: \beta_i = 0 \\ H_1: \beta_i \neq 0 \end{gathered} \tag{8}$$

Neste caso, a estatística do teste é descrita por (9).

$$T = \frac{\hat{\beta}_i - \beta_i}{S(\hat{\beta}_i)} \quad (9)$$

Onde

S: significa o erro padrão de cada parâmetro em análise.

Na resposta dos testes de hipóteses, um valor é comparado com o nível de significância previamente escolhido, sendo chamado de *p-valor*, isto é, valor do poder do teste.

A regra de decisão é se $t_{calculado} \geq t_{tabelado}$, rejeita-se H_0, e conclui-se ao nível α de significância que a variável não pode ser eliminada do modelo, pois explica bem a regressão linear. se $t_{calculado} < t_{tabelado}$, aceita-se a hipótese H_0, e ao nível de significância, conclui-se que a variável pode ser eliminada do modelo sem tanto dano para a explicação da regressão.

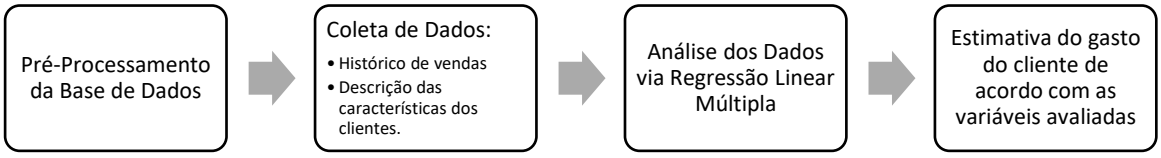
2.4 Procedimento *Stepwise*

Gujarati (2006) afirma que através da estatística t, os pesquisadores podem decidir pelo "melhor" conjunto de variáveis explicativas para um modelo de regressão a partir do método *Stepwise*. Conforme Hair Júnior et al. (2009) afirmam, a regressão *Stepwise* (passo a passo) é uma abordagem sequencial em que a equação é estimada com um conjunto de variáveis independentes, que são seletivamente adicionadas ou modeladas. Todas as variáveis independentes inicialmente são examinadas para inclusão na análise de regressão, e então, as variáveis independentes são acrescentadas uma de cada vez. Esse procedimento pode ser realizado através da seleção forward, que incorpora uma variável por vez usando a variável de maior coeficiente de correlação com a variável dependente; ou pela eliminação backward, que incorpora inicialmente todas as variáveis e vai eliminando as variáveis menos significativas para cada etapa.

3. Metodologia

Este estudo foi realizado utilizando-se dados reais em um grupo de lojas de uma empresa do setor varejista. Um esquema da metodologia adotada para este estudo está apresentado na Figura 1 que ilustra os procedimentos adotados para realização deste estudo.

Figura 1 - Representação esquemática da pesquisa



Fonte: Os autores (2020)

Inicialmente, o procedimento limpeza dos dados foi realizado na base disponibilizada pela empresa. No processo de limpeza, foram eliminados eventos de compras com dados faltantes ou informações incorretas. As informações coletadas na empresa em estudo são fornecidas pelos próprios clientes, que precisam registrar-se e fornecer dados pessoais antes de comprar. Nesta empresa, o fornecimento de dados pessoais está associado à algumas vantagens relacionadas à descontos que beneficiam o cliente.

Além das informações disponibilizadas pela empresa (idade, gênero, endereço e itens preferidos) foram inferidas informações sobre a renda das pessoas com base no seu endereço. Com base nos dados do CENSO (2010), foram registrados os valores de renda de acordo com cada um dos 2.395 setores censitários existentes no município de Curitiba. Posteriormente, os dados foram integrados em uma base de dados única que é constituída por variáveis categóricas, contínuas e binárias. Para a pesquisa, considerou-se como variável resposta (dependente) o gasto (EXPENDITURE) realizado pelo consumidor e como variáveis explicativas as variáveis descritas na Tabela 1. Os itens preferidos não foram descritos em razão da confidencialidade.

Tabela 1 - Relação dos atributos relacionados aos eventos de compras

Atributos	**Categorias**
Idade (AGE)	Quantidade
Gênero (SEX)	0 = Feminino 1 = Masculino
Renda (INCOME)	1 = Sem rendimento nominal mensal 2 = Rendimento nominal mensal de até 1/2 salário mínimo 3 = Rendimento nominal mensal de mais de 1/2 a 1 salário mínimo 4 = Rendimento nominal mensal de mais de 1 a 2 salários mínimos 5 = Rendimento nominal mensal de mais de 2 a 3 salários mínimos 6 = Rendimento nominal mensal de mais de 3 a 5 salários mínimos 7 = Rendimento nominal mensal de mais de 5 a 10 salários mínimos 8 = Rendimento nominal mensal de mais de 10 a 15 salários mínimos 9 = Rendimento nominal mensal de mais de 15 a 20 salários mínimos 10 = Rendimento nominal mensal de mais de 20 salários mínimos
Endereço (ADDRESS)	[1,2395]
Distância do Cliente até a loja (DISTANCE)	Quantidade
Artigo Preferido (F)	[1,11]

Fonte: Os autores (2020)

O estudo da Regressão Linear Múltipla foi realizado em um *notebook* com sistema operacional de 64 *bits*, processador Intel(R) Core (TM) i7-6500U CPU de 2,60 GHz e 8,00 GB de memória RAM. Por sua vez, o *software* R (versão 1.2.1335) foi utilizado para a realização das análises dos dados e testes estatísticos. Este *software* tem sido muito utilizado por empresas por ser um programa de fonte aberta. Foram utilizados os pacotes *nortest* para o teste de Anderson-Darling (AD) e *forecast* para a transformação BoxCox das variáveis, conforme descreve-se na seção 4.

4. Resultados

Quanto à determinação do tamanho de amostra, Hair Júnior et al. (2009) afirma que a proporção mínima de observações por variáveis é 5:1, mas a proporção preferida é de 15:1 ou 20:1. O tamanho da amostra deve ser adequado para garantir poder estatístico e generalização, e quanto mais graus de liberdade (tamanho da amostra subtraído do número de parâmetros estimados) melhor é a generalização. Visto que são consideradas 16 variáveis dependentes, considerando-se a proporção 20:1, a proporção mínima de observações necessárias seria de 360. O conjunto de dados, possui 1000 eventos de vendas, e a fim de se avaliar os erros decorridos, a amostra foi dividida aleatoriamente em 800 dados de treinamento e 200 dados de teste.

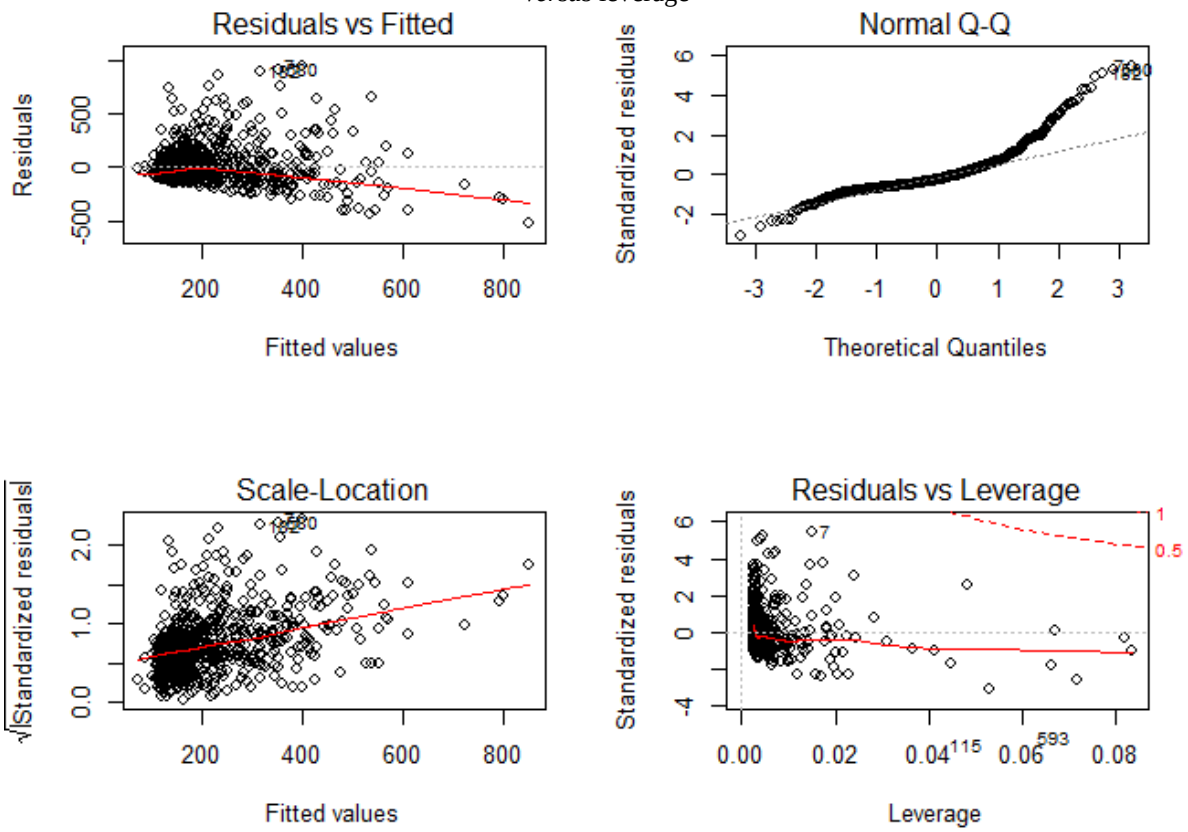
O procedimento utilizado aqui é o *Stepwise Backward* onde todas as variáveis são inicialmente incluídas no modelo e retiradas passo a passo em função da análise de significância. Avaliou-se a estatística *t* (*p*-valor) com 95% de confiança. Assim, para Sig. *t* menor do que 0,05,

confirma-se que existe a correlação. Ou seja, rejeita-se a hipótese nula (H_0) de que os parâmetros do modelo tenham sua relação com a variável dependente iguais a zero. Retirando-as do modelo, caso contrário.

O primeiro modelo de regressão múltipla gerado foi ajustado e obteve *p*-valor de $2.2e^{-16}$ para o teste F, que denota que as variáveis do modelo foram significativas conjuntamente. No entanto, o teste de Shapiro-Wilk para os resíduos rejeitou a hipótese de normalidade dos resíduos (W = 0,84005, p-valor < 2.2e-16). O teste de Anderson-Darling apresentou valores A = 36.257, p-valor < 2.2e-16, reforçando a hipótese de que os dados não seguem distribuição normal. O teste de Durbin-Watson, por sua vez, foi de DW = 2,0097, e p-valor =0,5554, constatando-se a não rejeição da hipótese nula, indicando a correlação serial ou dependência dos resíduos.

Hair Júnior et al (2009) ressalta que os testes de significância como o de Shapiro-Wilks ou Kolmogorov-Smirnov são muito sensíveis em amostras grandes (que excedem 1000 observações). Assim, uma alternativa é usar testes gráficos para avaliar o grau real de desvio da normalidade. Na Figura 2, percebe-se que os valores dos resíduos não estão perfeitamente em uma reta (quantil-quantil), indicando a falta de normalidade.

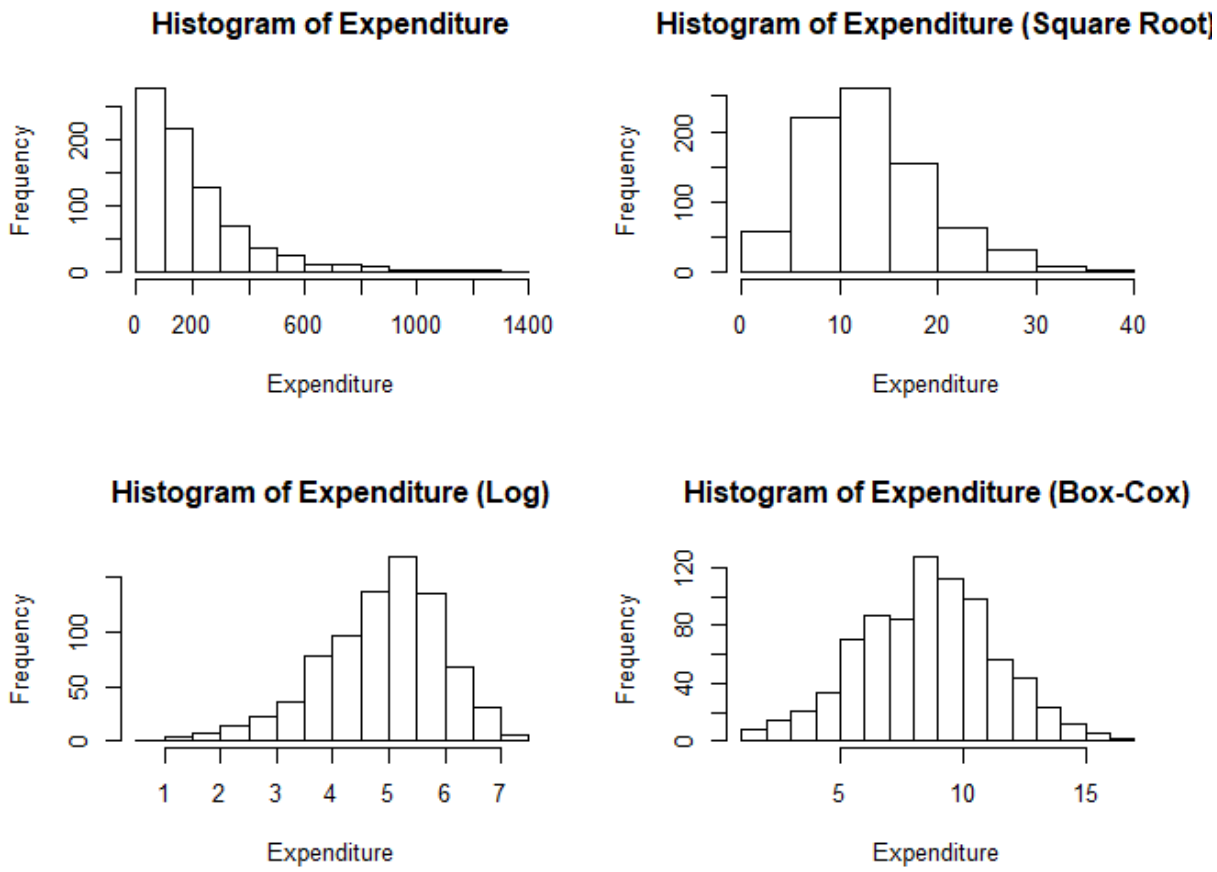
Figura 2 - Representação gráfica dos resíduos versus valores ajustados, Quantil-Quantil da normal, resíduos *versus* leverage



Fonte: Os autores (2020)

Devido a isto, procedeu-se uma transformação de dados visando a modificação de alguma variável de modo a corrigir as violações das suposições estatísticas e/ou melhorar as relações entre as variáveis. Fez-se tentativas de transformações tanto na variável dependente quanto nas variáveis independentes monitorando melhorias obtidas. Na Figura 3 apresenta-se o histograma da variável gasto original, seguido de transformações de aplicação de raiz-quadrada, logarítmica e por fim a Box-Cox. Essa ilustração é importante, pois a definição da melhor forma funcional é, por vezes, uma questão empírica a ser decidida a favor do melhor ajuste de variáveis.

Figura 3 - Histograma da variável dependente em relação às funções de transformação



Fonte: Os autores (2020)

Para cada uma das transformações, foram avaliados os testes de normalidade, cujos valores estão descritos na Tabela 2. Pode-se observar que para o teste de Shapiro-Wilk para a variável dependente, apenas para a transformação BoxCox a hipótese de normalidade dos resíduos não é rejeitada (0,3208> α = 0,05). Da mesma forma, para o teste de Anderson-Darling, o p-valor foi maior do que o nível de significância (0,5016> α = 0,05), o que permite concluir que os dados seguem a distribuição normal.

Tabela 2 - Testes de normalidade para a variável dependente

Teste	**Original, [p-valor]**	**Raiz Quadrada, [p-valor]**	**Logaritmo, [p-valor]**	**BoxCox, [p-valor]**
Shapiro-Wilk	W = 0,81223, [2.2e-16]	W = 0,97208, [3.086e-11]	W = 0,96385, [3.649e-13]	W = 0,9987, [0,8388]
Anderson-Darling	33.302, [2.2e-16]	A = 3.7711, [2.08e-09]	A= 6.8375, [2.2e-16]	A = 0,8388, [0,6927]

Fonte: Os autores (2020)

Assim a transformação de Box-Cox foi adotada para a variável dependente, com o valor ótimo para transformação sendo $\lambda = 0{,}20$, Após a transformação ótima de Box-Cox, realizou-se uma nova análise com a variável transformada. Na Tabela 3, tem-se as estimativas dos parâmetros do modelo ajustado com a variável transformada.

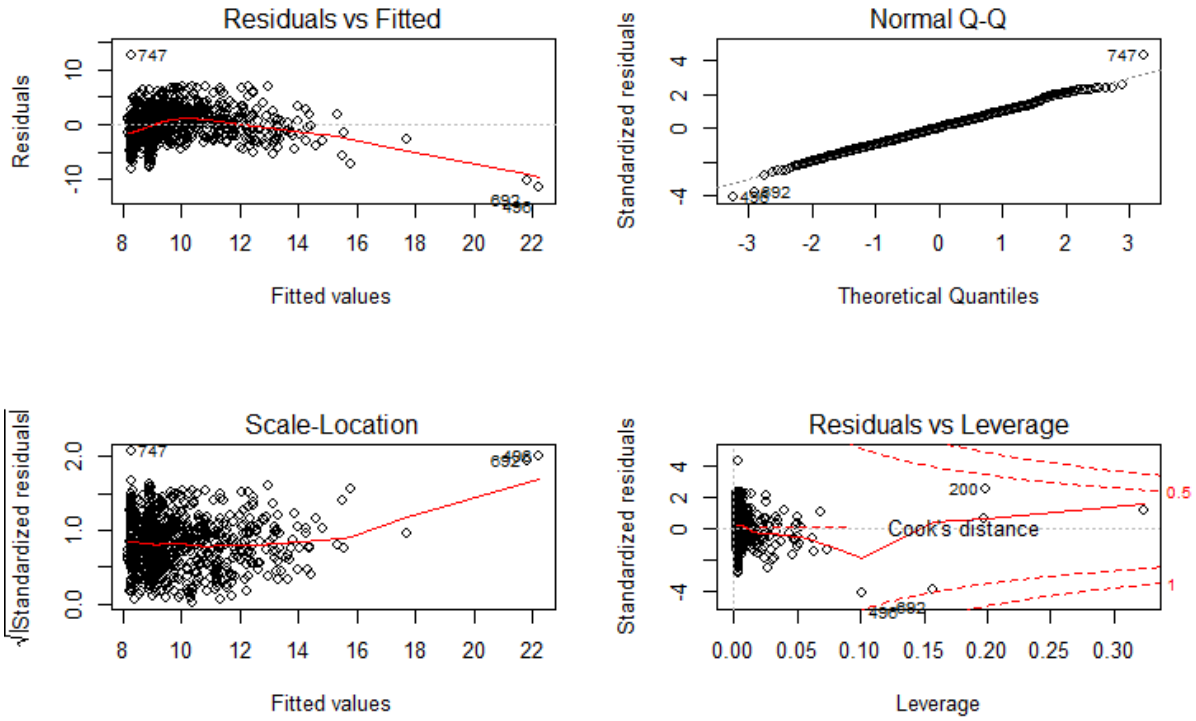
Tabela 3 - Ajuste do modelo de regressão linear múltipla utilizando *Stepwise* (dados treinamento)

Parâmetros	**Coeficientes**	**Erro padrão**	**Estatística *t***	**Valor-*p***
(Intercepto)	7,988	0,1765	45.263	$<2e^{-16}$
SEX	-3,761	0,1645	-2.286	0,02249
DISTANCE	-0,00008006	0,00002587	-3.095	0,00204
F_1	0,05317	0,004027	13.203	$<2e^{-16}$
F_2	-0,04232	0,01447	-2.924	0,00356
F_4	1,347	0,6353	2.120	0,03435

Fonte: Os autores (2020)

O teste de Shapiro-Wilks para os resíduos não rejeitou a hipótese de normalidade dos resíduos (W = 0,99773, p-valor = 0,3498). Da mesma forma, o teste de normalidade de Anderson-Darling, apresentou resultados de A = 0,24627, e p-valor = 0,756, que significa não rejeitar a hipótese nula de que os dados seguem uma distribuição normal. O teste de Durbin-Watson também não rejeitou a hipótese nula, (DW = 2.0189, p-valor = 0,6065), indicando que a transformação de Box-Cox corrigiu as suposições do modelo de regressão. Assim, na Figura 4 apresenta-se uma análise gráfica dos resíduos e por meio dela é possível verificar que houve melhora no ajuste da distribuição aos dados e que existe alguns pontos identificados que podem ser importantes e interferir na análise.

Figura 4 - Representação gráfica dos resíduos após transformação de Box-Cox



Fonte: Os autores (2020)

Os gráficos de diagnóstico apresentam os resíduos de quatro maneiras diferentes. O primeiro gráfico é um gráfico de dispersão dos valores de resíduos versus valores ajustados. Usado para verificar as suposições de relacionamento linear. Idealmente, o gráfico residual não mostrará nenhum padrão ajustado. Pode-se observar em evidência os pontos 743 e 541, que acabam por deslocar a linha vermelha da horizontal em zero. Esses pontos foram analisados e não há motivos evidentes para descartá-los da amostra.

O segundo gráfico (QQ-normal) mostra como a distribuição de resíduos se encaixa na distribuição normal. Isso representa os resíduos padronizados em relação aos quantis normais teóricos. O que se pode observar é que todos os pontos caem aproximadamente nessa linha de referência, o que reafirma a suposição de normalidade.

O terceiro apresenta (*Scale-Location*), um Gráfico de Escala-Localização, utilizado para verificar a homogeneidade da variância dos resíduos (homoscedasticidade). É essencialmente um gráfico de dispersão de resíduos normalizados com raiz quadrada absoluta e valores ajustados, com uma linha de regressão baixa. A linha horizontal com pontos igualmente dispersos é uma boa indicação de homoscedasticidade. Pode-se observar que a variabilidade

(variâncias) dos pontos residuais aumenta com o valor da variável de resultado ajustado, sugerindo variações não constantes nos erros residuais (ou heterocedasticidade).

O quarto gráfico mostra quais pontos têm maior influência na regressão (pontos de alavancagem). Qualquer coisa fora do grupo e fora das linhas "Distância de Cook" pode ter um efeito influente no ajuste do modelo. Esses pontos também foram verificados e não apresentaram motivos evidentes para descartá-los da amostra.

As estatísticas utilizadas para diagnóstico das suposições exigidas para a análise de regressão foram avaliadas para cada classe estão descritas na Tabela 4. Os resultados mostram que após a transformação de Box-Cox o modelo atende aos pressupostos de normalidade com significância de 5% para os testes de Shapiro-Wilk e de Anderson-Darling. Da mesma forma, todos os modelos apresentaram a confirmação para todas as classes da independência dos resíduos pelo teste de Durbin-Watson.

Tabela 4 - Resultado das estatísticas analisadas dos resíduos

Estatística	**Valores**
Transformação de Box-Cox (λ)	0,25
Shapiro-Wilk (SW)	0,9966
p-valor	0,08468
Anderson-Darling (AD)	0,26786
p-valor	0,6842
Durbin-Watson (DW)	2,0477
p-valor	0,7489

Fonte: Os autores (2020)

Por fim, a Tabela 5 apresenta os coeficientes de correlação e de determinação (R^2) do modelo proposto. Os coeficientes indicam correlações moderadas. O valor de R arredondado em torno de 0,5433 mostra uma correlação moderada. O teste F (66,5) apresentou p-valor<$2{,}2e^{-16}$ que indica que o modelo é significativo como um todo.

Tabela 5 - Resultado das estatísticas

Coeficiente de Correlação	R^2 (coeficiente Correlação ao quadrado)	R^2 ajustado	Estatística F	Graus de Liberdade Resíduos	Soma dos Quadrados dos Resíduos	**Dados Treinamento**		**Dados TESTE**	
						MAE	RMSE	MAE	RMSE
0,5433	0,2952	0,2907	66,5; $<2e^{-16}$	794	2,319	120,8408	196.4526	119.1538	214.2868

Fonte: Os autores (2020)

Avaliados os pressupostos necessários à análise, obtém-se a equação de Regressão Linear Múltipla que descreve a relação entre o valor gasto pelo cliente em relação às variáveis independentes escrita por (10).

$$\frac{Expenditure^{0,2} - 1}{0,2} = 7,988 - 3,761 * SEX - 0,00008006 * DISTANCE + 0,05317 * F_1 - 0,04232 * F_2 + 1,347 * F_4 \quad (10)$$

A variável “SEX”, que refere-se ao gênero do cliente, apresentou significância relacionando-a ao valor gasto. Visto que ela tem ponderação negativa, conclui-se as mulheres tendem a gastar mais do que os homens nestas lojas. Além disso a variável “DISTANCE” sendo também negativa denota que pessoas que moram mais longe da loja tendem a fazer compras com maiores gastos do que pessoas que moram mais próximas à loja em que a compra foi efetuada. Além disso, dentre as 11 classificações de produtos possíveis, os clientes com preferências pelos produtos do tipo 1, 2 e 4 representam aqueles com maiores gasto.

5. Conclusão

O conhecimento armazenado em bases de dados tem se mostrado ser um elemento fundamental à compreensão dos fenômenos que envolvem qualquer ambiente organizacional. A pesquisa realizada partiu do pressuposto de que é possível prever a quantidade de gasto gerada pelos clientes em determinada loja, considerando-se suas características e seu perfil de consumo. Considerou-se como variável resposta (dependente) o gasto realizado pelo consumidor e como

variáveis explicativas as características dos clientes, como idade, sexo, endereço, renda, distância do cliente até a loja.

Como resultados, obteve-se que o coeficiente de determinação múltipla (R^2) foi igual 0,2952 e o R^2 ajustado foi 0,2907. Este valor representa a proporção dos dados que são explicados através do conjunto de variáveis explanatórias selecionadas. O valor de *R* arredondado em torno de 0,5433 mostra uma correlação moderada, o que pode indicar que parte dos valores gastos pelos clientes, são explicados por outras variáveis que não constam no modelo. Os valores de erros para os valores de teste são MAE=119.1538 e MSE =214,2868.

Quanto às conclusões do modelo, o sexo do cliente é um fator significativo quanto ao valor gasto, sendo que mulheres tendem a gastar mais do que os homens nestas lojas. Além disso, clientes que moram mais longe da loja tendem a fazer compras com maiores gastos do que pessoas que moram mais próximas à loja em que a compra foi efetuada. Além disso, dentre as 11 classificações de produtos possíveis, os clientes com preferências pelos produtos do tipo 1, 2 e 4 representam aqueles com maiores gastos.

Dessa forma, pode-se observar que por mais instigantes que sejam os resultados, encontram-se limitados pela impossibilidade de se compreender, em toda a sua dimensão, as razões que levam alguém a gastar. Além disso, é importante observar que outras variáveis adicionais como, número de filhos, escolaridade, além de variáveis psicográficas como interesses, opiniões, necessidades e valores, também são importantes na determinação do gasto dos clientes. Atualmente, essas variáveis não são levantadas pelo cadastro da empresa, mas poderiam também ser levadas em consideração no modelo para melhorar a sua previsão.

Tais limitações, ao mesmo tempo, recomendam cautela como os resultados do presente trabalho, e encorajam o aprofundamento da investigação e o desenvolvimento de futuras pesquisas. Sugere-se portanto, a avaliação de outras técnicas de processamento, como forma de corroborar ou negar aquilo que acaba de ser relatado, como Redes Neurais Artificiais e Árvores de Decisão, por exemplo.

REFERÊNCIAS

CENSO, I. B. G. E. Disponível em:< http://www. censo2010. ibge. gov. br/>. Acesso em, v.23, 2010.

COELHO, A. C.; CUNHA, J. V.A. **Regressão linear múltipla.** Análise multivariada: para os cursos de administração, ciências contábeis e economia. São Paulo: Atlas, p.131-231, 2007.

CUNHA, J. V.; COELHO, A. C**. Análise multivariada**: para os cursos de administração, ciências contábeis e economia. FIPECAFI–Fundação Instituto de Pesquisas Contábeis, Atuariais e Financeiras. Regressão linear múltipla. São Paulo: Atlas, p.131-231, 2009.

FÁVERO, L. P.; BELFIORE, P. P.; SILVA, F. L., CHAN, B. L. **Análise de dados: modelagem multivariada para tomada de decisões.** Editora Campus: Rio de Janeiro, 2009.

GUJARATI, D. N. **Econometria Básica,** 3 ed. São Paulo/SP, Markron Books, 2006.

HAIR JR., J.F.; WILLIAM, B.; BABIN, B.; ANDERSON, R.E. Análise multivariada de dados. 6.ed. Porto Alegre: Bookman, 2009.

HILL, R. C.; GRIFFITHS, W. E. e JUDGE, G. G., **Econometria.** Saraiva, São Paulo. KIM, S.-J. e WU, E. (2008), "Sovereign credit ratings, capital flows and financial sector development in emerging markets". Emerging Markets Review, v.9(1), pp.17-39, 1999.

STOCK, J. H.; WATSON, M. W. Combination forecasts of output growth in a seven-country data set. **Journal of forecasting,** v.23, n. 6, p.405-430, 2004.