UNIVERSIDADE ESTADUAL DA PARAÍBA - UEPB

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

CURSO DE BACHARELADO EM ESTATÍSTICA

ANÁLISE DE REGRESSÃO LINEAR PARA PREVER O PERFIL DO CLIENTE DE UMA LOJA DE ROUPAS

Lucas Manoel Batista de Albuquerque
Débora de Souza Cordeiro

Econometria

2023

1 INTRODUÇÃO

Este artigo mergulha nas complexidades da análise do perfil de clientes de uma loja de roupas, explorando como a regressão linear pode ser uma ferramenta poderosa na busca por informações. Examinaremos fatores como o tempo dedicado ao app e *website* da loja, o tempo que a pessoa é cliente da loja, o tempo médio de interação com os funcionários em relação ao valor total gasto no ano em reais como variável resposta, e como esses dados podem ser modelados para prever comportamentos futuros, personalizar estratégias de *marketing* e aprimorar a experiência do cliente.

2 OBJETIVOS

- Realizar uma regressão linear para prever se o valor total gasto no ano de cada cliente é influenciado pelo tempo que este cliente interage com os funcionários da loja, pelo tempo gasto no aplicativo e *website* da loja e pelo tempo que a pessoa é cliente da loja;
- Oferecer opiniões ou dicas para o mercado de lojas de roupa online em geral.

3 METODOLOGIA

3.1 ANÁLISE DE REGRESSÃO LINEAR MÚTIPLA

Percebe-se que o problema deste trabalho envolve mais de uma variável preditora. Sendo assim, iremos dar foco na regressão linear múltipla, pois é ele que nos dá a condição de trabalhar com mais de uma variável.

Para que o modelo de regressão linear múltiplo seja definido, supõe-se que se tenhamos $X_1, X_2, \dots, X_p$ variáveis preditoras em relação a uma variável y_i independente, sendo o modelo dado da seguinte forma:

$$y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_{i2} + X_{ip}\beta_{ip} + \epsilon_i, \qquad i = 1, 2, \dots, n. \tag{1}$$

Outra maneira de visualizar o modelo de regressão linear múltipla é em forma matricial, dada por:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{2}$$

em que $\boldsymbol{y}$ é a nossa variável a ser predita, $\boldsymbol{X}$ é a matriz de variáveis dependentes, $\boldsymbol{\beta}$ é a matriz dos parâmetros estimados e $\boldsymbol{\epsilon}$ é a matriz dos erros associados.

Algumas suposições e pressuposições do modelo de regressão linear múltiplo:

- A variável $\boldsymbol{y}$ tem de seguir uma distribuição normal;
- A variável $\boldsymbol{\epsilon}$ tem de seguir uma distribuição normal com médio zero e variância constante;
- Os erros são não correlacionados dois a dois e apresentam homogeneidade;

- As variáveis explicativas não podem ter uma correlação muito alta (acima de 0,9 ou abaixo de -0,9). Caso apresentem uma correlação dessa magnitude elas possuem multicolinearidade. Utilizaremos para esta análise o coeficiente de correlação de Pearson, pois é adequado para variáveis do tipo contínuas e que sigam normalidade. Calcula-se o coeficiente de correlação de Pearson seguindo a seguinte fórmula:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(z_i - \bar{z})^2}} = \frac{cov(X,Z)}{\sqrt{var(X).var(Z)}}$$

onde $x_1, x_2, \dots, x_n$ e $z_1, z_2, \dots, z_n$ são os valores medidos de ambas as variáveis. Temos também que $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ e $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} z_i$ são as médias aritméticas de ambas as variáveis.

- Deve haver uma relação linear entre a variável resposta $\boldsymbol{y}$ e as variáveis preditoras $\boldsymbol{X}$.

3.2 DETECTANDO MULTICOLINEARIDADE

De acordo com Gujarati e Poter (2011), o FIV mostra como a variância de um estimador é inflada pela presença de multicolinearidade. O FIV é definido como:

$$FIV = \frac{1}{1 - R_j^2} \qquad (3)$$

Onde FIV representa o Fator de Inflação da Variância, R_j^2 representa o coeficiente de determinação parcial de X_j em relação as demais variáveis explicativas e mostra como a variação de um estimador é inflada pela presença da multicolinearidade.

Temos que, quando R_j^2 aproxima-se de 1, o FIV aproxima-se do infinito. Se não houver colinearidade entre as variáveis explicativas, o FIV será 1. Para apontar que existe colinearidade entre as variáveis usaremos um valor limite para o FIV igual a 4, ou seja, caso o FIV seja maior do que 4 para algumas das variáveis explicativas, então ela é uma variável que apresenta colinearidade e será removida do modelo.

3.3 METODO DE ESTIMAÇÃO

Para estimar os parâmetros de um modelo de regressão linear múltiplo, podemos recorrer ao método dos mínimos quadrados, que nos permite encontrar uma reta que minimize a distância entre os pontos observados e a reta estimada, fazendo, em média, a soma dos desvios quadráticos ser igual a zero. Da equação (2), temos:

$$\sum_{i=1}^{n} \boldsymbol{\epsilon_i^2} = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})$$
$$= (\boldsymbol{y}' - \boldsymbol{\beta' X'})(\boldsymbol{y} - \boldsymbol{X\beta})$$

$$= \boldsymbol{y'y - y'X\beta - \beta'X'y + \beta'X'X\beta}$$

$$\boldsymbol{Z = y'y - 2y'X\beta + \beta'X'X\beta}$$

A função $\boldsymbol{Z}$ deve ser derivada em relação a $\boldsymbol{\beta}$ e igualada a zero para o se obter o ponto de mínimo para os valores de $\boldsymbol{\beta}$, portanto:

$$\frac{\partial \boldsymbol{y'y}}{\partial \boldsymbol{\beta}} - \boldsymbol{2}\frac{\partial \boldsymbol{y'X\beta}}{\partial \boldsymbol{\beta}} + \frac{\partial \boldsymbol{\beta'X'X\beta}}{\partial \boldsymbol{\beta}} = \boldsymbol{-2(X'y) + 2X'X\beta}$$

Denominando por $\widehat{\boldsymbol{\beta}}$ o vetor que anula a derivada, podemos escrever:

$$\begin{aligned} -2\boldsymbol{X'y} + \boldsymbol{2X'X\widehat{\beta}} &= \boldsymbol{0} \\ \boldsymbol{X'X\widehat{\beta}} &= \boldsymbol{X'y} \\ \boldsymbol{\widehat{\beta}} &= \boldsymbol{(X'X)^{-1}X'y} \end{aligned}$$

Desta forma temos o estimador de $\boldsymbol{\beta}$ via métodos dos mínimos quadrados, desde que a inversa de $\boldsymbol{X'X}$ exista.

3.4 ANÁLISE DE RESÍDUOS

Para que seja confirmada a validade de um modelo de regressão linear é necessário que os resíduos sigam uma distribuição normal com média igual a zero e variância constante. A condição de normalidade dos resíduos será testada pelo teste de Shapiro-Wilk, que testa a hipótese nula de que uma amostra veio de uma população normalmente distribuída. Além disso, para verificar a normalidade dos resíduos, usamos o gráfico de probabilidade normal, o QQ-plot (Quantil de probabilidade esperado para a distribuição normal, em função dos resíduos). Após o esboço deste gráfico pode-se verificar que, se os erros possuírem distribuição normal, os pontos devem estar alinhados mais ou menos sobre uma reta, caso contrário, os dados não apresentam indícios de normalidade.

4 SOBRE OS DADOS

Este conjunto de dados contém dados de clientes que compram roupas online. A loja oferece sessões de aconselhamento de estilo e roupas na loja. Os clientes chegam à loja, têm sessões/reuniões com um *personal stylist*, depois podem ir para casa e encomendar através de um aplicativo de celular ou site as roupas que desejam. A empresa está tentando decidir se concentrará seus esforços na experiência do aplicativo móvel ou no site.

O conjunto de dados contém as seguintes variáveis:

- Tempo médio em minutos da sessão com o funcionário da loja;
- Tempo médio em minutos usando o aplicativo da loja;
- Tempo médio em minutos usando o *website* da loja;
- Tempo que a pessoa é cliente da loja em anos;
- Gasto médio anual do cliente em reais R$ (Variável resposta).

Abaixo na Tabela 1 estão expostas as correlações de Pearson para cada uma das variáveis:

Tabela 1 – Correlações de Pearson para as variáveis

---	Média sessão	Tempo no App	Tempo no website	Tempo que é cliente	Gasto Anual
Média sessão	1	-0.028	-0.035	0.060	0.355
Tempo no App	-0.028	1	0.082	0.029	0.499
Tempo no website	-0.035	0.082	1	-0.047	-0.002
Tempo que é cliente	0.060	0.029	-0.047	1	0.809
Gasto Anual R$	0.355	0.499	-0.002	0.809	1

Fonte: Autoria própria.

As correlações nos indicam que não há multicolinearidade nos dados pois as correlações entre as variáveis independentes são baixas. Em relação as correlações das variáveis independentes e a variável dependente vemos que apenas a variável independente relacionada ao tempo no website tem uma correlação quase nula, nos indicando que ela possa não ter influência no gasto anual dos clientes.

Na Tabela 2 temos as estatísticas descritivas para as variáveis que serão usadas para estimar o modelo de regressão linear múltiplo:

Tabela 2 – Estatísticas descritivas para as variáveis disponíveis no banco de dados

Variável	Mínimo	Mediana	Média	Máximo
Média sessão	29,53	33,08	33,05	36,14
Tempo no App	8,5	11,98	12,05	15,12
Tempo no website	33,91	37,07	37,06	40,01
Tempo que é cliente	0,27	3,53	3,53	6,92
Gasto anual R$	256,7	498,9	499,3	765,5

Fonte: Autoria própria.

5 RESULTADOS

Primeiramente ajustaremos um modelo de regressão linear múltiplo utilizando todas as variáveis disponíveis. Na Tabela 2 temos as estimativas dos parâmetros para o primeiro modelo estimado:

Tabela 2 – Estimativas dos parâmetros para o modelo usando todas as variáveis explicativas:

Variáveis	Estimativas	T calculado	p-valor
Intercepto	-1051,59	-45,736	<0,001
Média sessão	25,73	57,06	<0,001
Tempo no App	38,7	85,8	<0,001
Tempo no website	0,43	0,98	0,326
Tempo que é cliente	61,5	137,34	<0,001

Fonte: Autoria própria

Observamos que apenas a variável que representa o tempo médio que os clientes usam o website da loja não apresenta significância estatística a um nível de 5% suficiente para permanecer no modelo final. Refazendo mais uma vez as estimativas dos parâmetros sem esta variável, temos na Tabela 3 as seguintes estimativas:

Tabela 3 – Estimativas dos parâmetros para o modelo sem usar a variável tempo médio que o cliente passou usando o website

Variáveis	Estimativas	T calculado	p-valor
Intercepto	-1035,34	-64,78	<0,001
Média sessão	25,72	57,05	<0,001
Tempo no App	38,74	86,21	<0,001
Tempo que é cliente	61,56	137,46	<0,001

Fonte: Autoria própria.

Vemos que todas as variáveis foram significativas a um nível de 5% de significância, considerando este o modelo final a ser analisado e interpretado. O coeficiente de determinação para este modelo foi de 0,9843, ou seja, o modelo explica aproximadamente 98% da variação dos dados.

Aplicando a técnica do FIV ao modelo presente na Tabela 3, vemos que não há multicolinearidade, pois os valores do FIV estão todos abaixo de 4 como ilustra a Tabela 4:

Tabela 4: Valores FIV para as variáveis explicativas do segundo modelo

Variável	FIV
Média sessão	1,004
Tempo no App	1,001
Tempo que é cliente	1,004

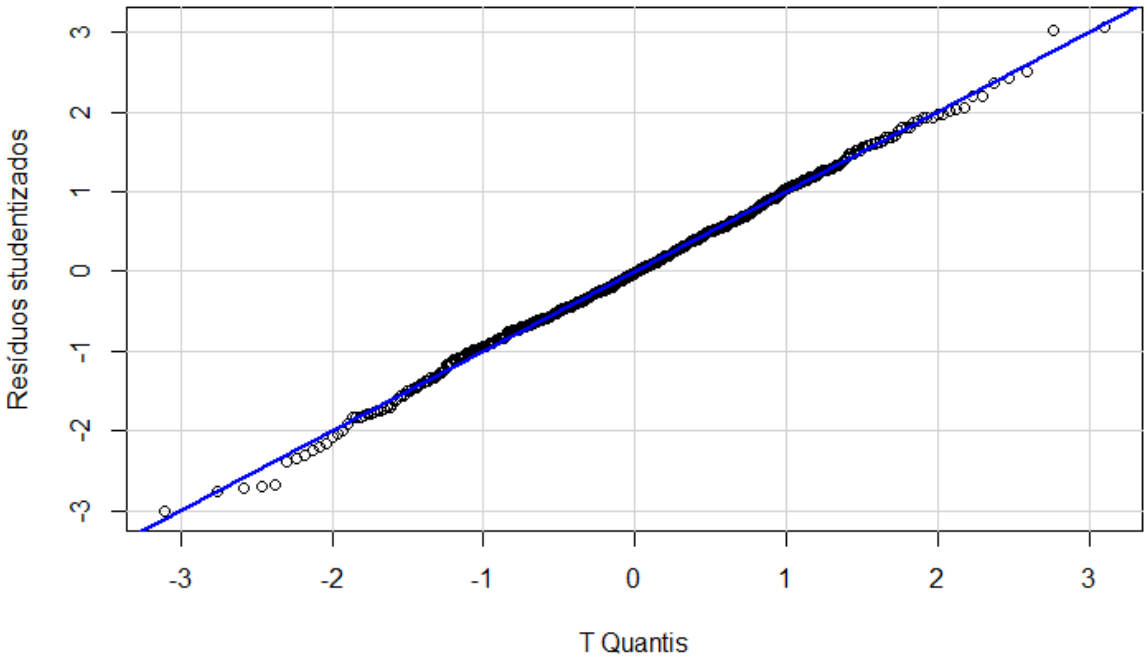
Fonte: Autoria própria.

O teste de Shapiro-Wilk nos indica que os resíduos para este modelo provem de uma distribuição normal. O p-valor para este teste é de 0,87 e sua estatística de

teste é 0,99, ou seja, não rejeitamos a hipótese nula de que os resíduos sejam provenientes de uma distribuição normal a um nível de 10% de significância.

Além disso, o gráfico QQ-Plot, que pode ser visto na Figura 1, corrobora os resultados deste teste, já que a maioria dos resíduos estão dispostos próximos da linha da distribuição normal:

Figura 1 - Gráfico QQ-Plot para os resíduos do modelo final ajustado



Fonte: Autoria própria.

6 CONCLUSÃO

Foi possível demonstrar por meio da análise de regressão múltipla que o gasto médio anual dos clientes da loja de roupas pode ser explicado pelo tempo médio em minutos da sessão com o funcionário da loja, tempo médio em minutos usando o aplicativo da loja e o tempo que a pessoa é cliente da loja em anos. A variável relacionada ao Tempo médio em minutos que usa o Web Site da loja não influencia no gasto dos clientes.

Para possíveis novas informações, poderíamos ter na base de dados variáveis como sexo do cliente, idade e/ou classe social, pois assim um modelo mais robusto e com interpretações mais diversas seria apresentado.

Pode-se dizer então que um investimento mais adequado para com o App da loja seja necessário, já que é um fator que influencia diretamente e de forma positiva nos lucros da empresa. Analogamente, é interessante observar qual o problema com o Web Site da loja e os motivos do tempo em que os clientes dedicam a ele não influenciam nos lucros da empresa.

7 REFERÊNCIAS

GUJARATI, D. N.; PORTER, D. C. **Econometria Básica** - 5.Ed. [s.l.] McGraw Hill Brasil, 2011.

CHARNET, R. et al. **Análise de Modelos de Regressão Linear com Aplicações**. 2a edição ed. [s.l.] Editora da Unicamp, 2008.

Linear Regression E-commerce Dataset. Disponível em: https://www.kaggle.com/datasets/kolawale/focusing-on-mobile-app-or-website/data.