Projeto 2

Gabriel Victor Soares dos Santos

19/11/2021

1. Introdução

Este trabalho tem como objetivo a aplicação de técnicas estatísticas fazendo análises descritivas, exploratórias e inferenciais sobre um modelo de regressão linear múltipla. Além de servir como método avaliativo para a matéria **ME613** - Análise de Regressão da Universidade Estadual de Campinas.

A gorjeta, pela definição do dicionário Michaelis, é uma pequena quantidade de dinheiro com que se gratifica um serviço além do preço estipulado, e é normalmente dada em bares, restaurantes, hotéis, entre outros estabelecimentos. Essa quantidade monetária varia dependendo da satisfação do cliente atendido, entretanto, por compor uma parte significativa dos salários desses trabalhadores, a porcentagem mínima de gorjeta é regulada de acordo com as leis trabalhistas de cada país. Nos Estados Unidos, por exemplo, é habitual calcular a parte e dar 15% do valor da conta, enquanto em outros países a gojeta é inclusa na conta na forma de uma *"taxa de serviços"*, como no Brasil onde a taxa é de 10%. Existem também países cujo o ato de dar gorjeta é considerado desrespeitoso e mal visto.

Dentre os bancos de dados disponíveis, foi escolhido "*tips*", um conjunto de dados coletados sobre as gorjetas recebidas, no começo dos anos 1990, por um garçom de um restaurante no Estados Unidos. Analisaremos as informações coletadas a respeito dos clientes que frequentaram esse restaurante e criaremos um modelo de regressão múltipla para a predição das gorjetas.

2. Metodologia

Este banco de dados apresenta 7 variáveis diferentes sendo elas a gorjeta recebida, o total da conta, o gênero do(a) cliente pagante, se é fumante, o dia frequentado entre quinta-feira e domingo, período do dia, nesse caso almoço ou janta, e tamanho do grupo por mesa atendida. Criou-se a variável razão que é a proporção da gorjeta em relação ao total da conta.

Para obter os resultados e respostas acerca da problematização apresentada neste trabalho, será feita a análise do conjunto de dados sumarizando informações relevantes em tabelas e gráficos, e por último a criação e refinamento do modelo de regressão múltipla adequado. O trabalho transcorrerá a partir o conteúdo ensinado em aula do quais, que dentre os métodos apresentados, utilizou-se o critério de análise de variância (ANOVA) e *Backward Elimination* para a seleção das variáveis preditoras, a um nível de significancia de 5%. Já para a testagem dos coeficientes optou-se por analisa-los a um nível de 1%.

Neste trabalho é utilizada a linguagem de programação R e o programa *RStudio* para os testes, cálculos e criação das tabelas e gráficos.

3. Descrição Dados

O conjunto é composto de 7 variáveis, com 244 observações ao todo sendo a maioria das variáveis classificadas como categóricas, com respostas de 2 a 4 níveis, ou seja, há 2 ou 4 tipos de resposta para cada variável. A

Tabela 1 apresenta as medidas sumárias das variáveis numéricas, a Tabela 2 é uma tabela de contingência sobre o gênero dos clientes e se são fumantes e a Tabela 3 mostra a quantidade de clientes ao longo do dia pelos dias da semana.

Table 1: Médias e Desvios Padrões das Variáveis Numéricas

	Média	Desvio Padrão	Valor Mínimo	Valor Máximo
Total da Conta	19.79	8.90	3.07	50.81
Gorjeta	3.00	1.38	1.00	10.00
Tamanho do Grupo	2.57	0.95	1.00	6.00

Podemos ver na Tabela 1 que os clientes que frequentaram o restaurante iam em pares, ou trios, e a média do total das contas é de 19.79 dólares, o que dada a taxa de 15% da seção 1, é proporcional a média de gorjetas de 3 dólares.

Table 2: Gênero do Pagante e se o Grupo é Fumante

Gênero/Fumante	Não	Sim	Total
Feminino	33	54	87
Masculino	60	97	157
Total	93	151	244

Sobre a Tabela 2, 64% (157) dos clientes que pagaram a conta foram os homens. Em relação ao fumo, 61% (151) dos clientes fumavam, desses 64% (97) eram homens e os outros 36% (54) mulheres.

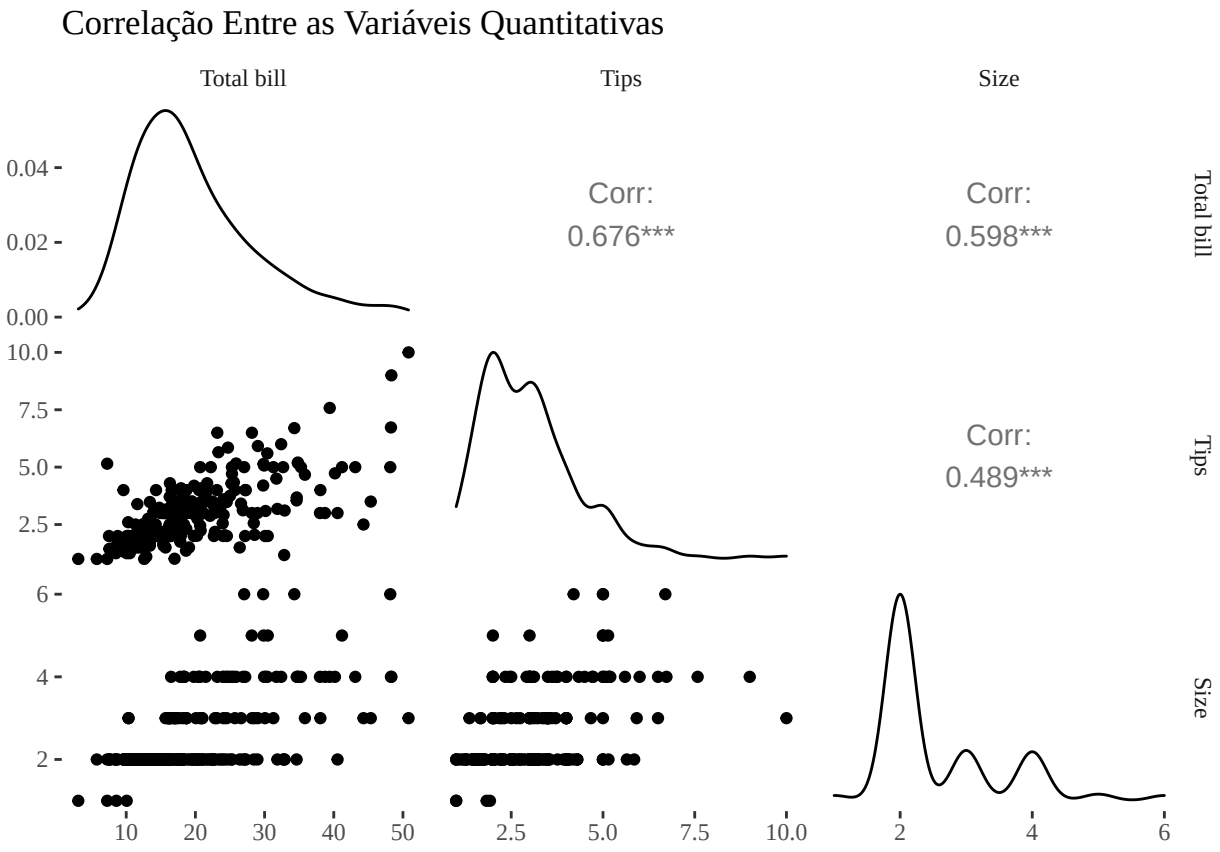
Table 3: Quantidade de mesas atendidas ao longo do dia pelos dias da semana

Período/Dia	Quinta	Sexta	Sábado	Domingo	Total
Almoço	61	7	0	0	68
Janta	1	12	87	76	176
Total	62	19	87	76	244

Na Tabela 3 temos que os almoços representam apenas 27% do total observado, indicando que os clientes preferem jantar no restaurante. Os dias preferidos são sábados e domingos com uma frequencia apresentada de 66%.

4. Análise exploratória

Usando a Correlação de Pearson é possível quantificar a correlação entre as variáveis, além de determinar se as variáveis são direta, ou inversamente, proporcionais.



Segundo o gráfico a menor correlação foi de 48%, entre a variável resposta e a variável Tamanho do Grupo, podendo classificar essa relação como sendo moderadamente positiva. A maior correlação foi de 67% entre a variável Gorjeta e Total da Conta, podendo considera-la uma correlação fortemente positiva.

Destaca-se a seguir alguns gráficos que podem nos indicar visualmente um pouco do que foi apresentado no gráfico de Correlação e nas tabelas da seção 2.

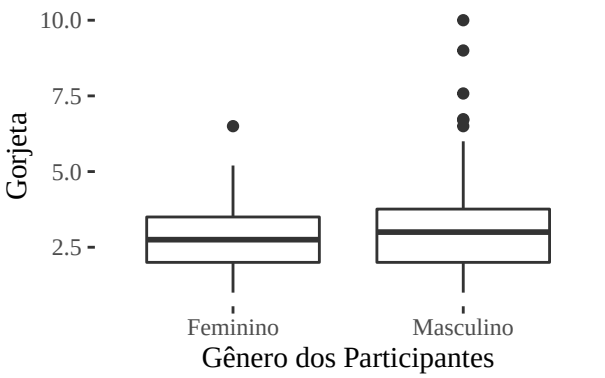


Gráfico 1

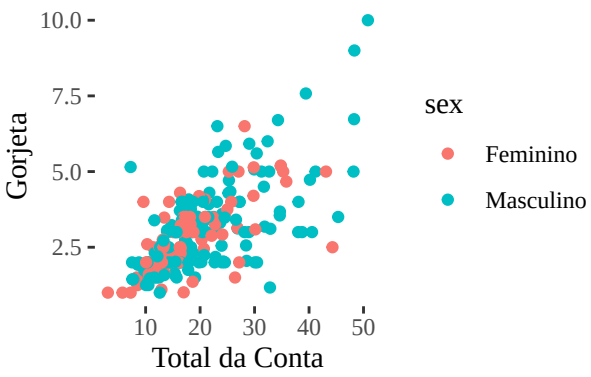


Gráfico 2

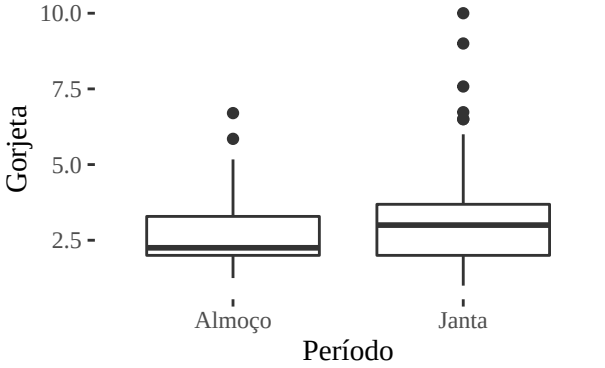


Gráfico 3

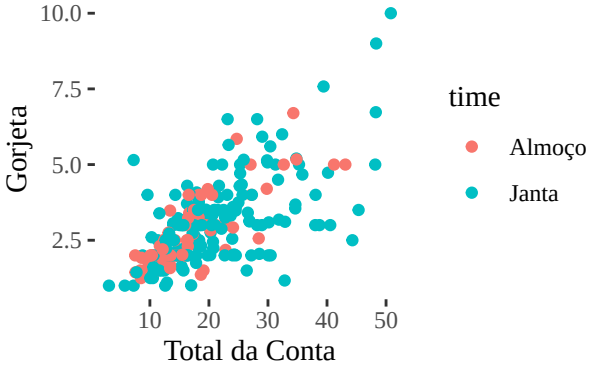


Gráfico 4

No gráfico de boxplot 1, a diferença da média de gorjeta é mínima. Seguindo a taxa de 15% de gorjeta, pode indicar uma qualidade estável no atendimento prestado pelo garçom. Entretanto no grafico de boxplot 2, a janta apresenta uma média muito maior de gorjetas dadas, já que, como visto na Tabela 3 da seção 2, os clientes frequentavam mais o restaurante no período da noite.

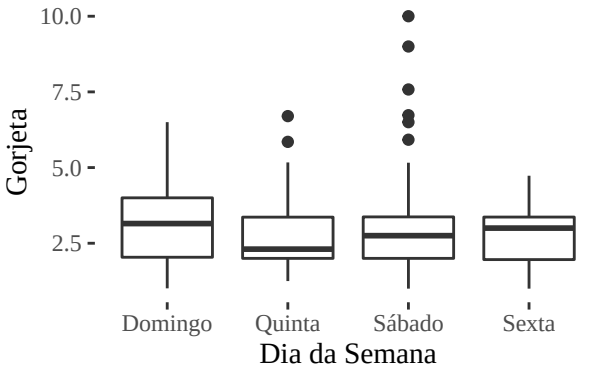


Gráfico 5

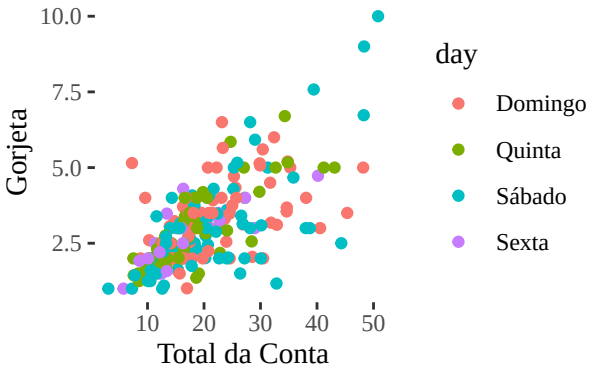


Gráfico 6

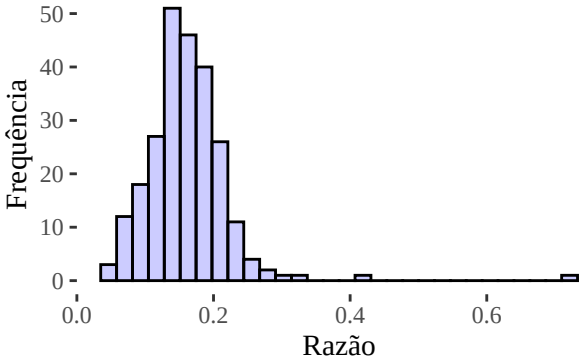


Gráfico 8

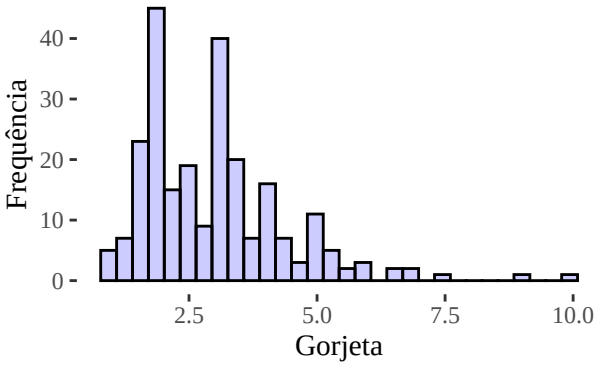


Gráfico 8

No gráfico de boxplot 3, sábado apresenta mais outliers que os outros dias, domingo a média de gorjetas foi maior, enquanto quinta-feira a média foi inferior a 2 dólares e meio. No gráfico 8, vemos que a proporção de gorjetas dadas tem média 16%, e que pelo gráfico 8 as gorjetas mais frequentes são de 2 e 3 dólares.

5. Analise Inferencial

O modelo de regressão utilizado nesse trabalho será:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i,p-1} + \epsilon_i \tag{1}$$

, onde:

- $\beta_0, \beta_1, ..., \beta_{p-1}$ são parâmetros.
- $X_{i1}, ..., X_{i,p-1}$ são constantes conhecidas.
- $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.
- $i = 1, 2, ..., n$.

5.1 Modelagem estatística

Serão construídos modelos de regressão multipla entre as variáveis resposta, neste caso Gorjeta, e as demais variáveis, como descrito na seção 2, avaliaremos primeiro a Análise de Variância (ANOVA) para encontrar se há diferença entre a distribuição das variáveis, a um nível de significancia de 5%.

Se o modelo for adequado, verificaremos então a normalidade pelo Teste de Shapiro-Wilks, heterocedasticidade pelo Teste de Breusch-Pagan, a um nível de significancia de 1%.

Table 4: Coeficientes do Modelo

	x
Intercepto	0.8038173
Total da Conta	0.0944870
Gênero	-0.0324409
Fumante	-0.0864083
Sábado	-0.1214584
Domingo	-0.0254807
Quinta	-0.1622592
Período	0.0681286
Tamanho do Grupo	0.1759920

Teste ANOVA	p-valor	Resultado Significativo
Total Conta	<2e-16	Sucesso
Gênero	0.8190	Fracasso
Fumante	0.5561	Fracasso
Sábado	0.6953	Fracasso
Domingo	0.9369	Fracasso
Quinta	0.6804	Fracasso
Período Almoço	0.8783	Fracasso
Tamanho do Grupo	0.0500	Sucesso

Com base na análise de variância e usando o método de *Backward Elimination*, dentre o modelo atual as variáveis significativas para a regressão são o Total da Conta e o Tamanho do Grupo.

Portanto, a equação fica da seguinte maneira:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i \tag{2}$$

as variáveis do modelo são:

- X_1 = Total da Conta
- X_2 = Tamanho do Grupo

Dos quais apresentam os seguintes coeficientes:

$$Y_i = 0.6689 + 0.0927 X_1 + 0.1925 X_2 + \epsilon_i \tag{3}$$

Tendo um modelo aceitável por ANOVA, procediremos para os testes de normalidade e heterocedasticidade.

5.2 Teste de Shapiro-Wilks e Teste de Breusch-Pagan

O Teste de Shapiro-Wilks, que busca analisar a normalidade dos resíduos, tem como critério de decisão o p-valor for inferior a 0.01, o modelo será considerado um sucesso.

Teste de Shapiro-Wilks	p-valor	Resultado
Modelo	1.78e-05	Fracasso

Enquanto o Teste de Breusch-Pagan busca analisar a dispersão dos dados, também tendo como critério de decisão o p-valor for inferior a 0.01 para ser considerado um sucesso.

Teste de Breusch-Pagan	p-valor	Resultado
Modelo	$<$ 2.22e-16	Fracasso

Esse modelo não é adequado para descrever a relação entre a variável resposta e as possíveis variáveis preditoras.

5.3 Remodelando

Usando o procedimento de *Box-Cox*, determinamos o poder de tranformação apropriado para a variável Y, obtendo o valor de $\lambda = 0.1237$. Então criamos um novo modelo após a transformação e testamos a análise de variância.

Teste ANOVA	p-valor	Resultado Significativo
Total Conta	<2e-16	Sucesso
Gênero	0.6912	Fracasso
Fumante	0.6903	Fracasso
Sábado	0.5020	Fracasso
Domingo	0.9467	Fracasso
Quinta	0.6784	Fracasso
Período Almoço	0.9318	Fracasso
Tamanho do Grupo	0.0426	Sucesso

Usando outra vez ANOVA e o método de *Backward Elimination*, temos que do modelo atual as variáveis significativas para a regressão novamente são o Total da Conta e o Tamanho do Grupo. Tendo um modelo aceitável por ANOVA, procediremos para os testes de normalidade e heterocedasticidade.

5.4 Teste de Shapiro-Wilks e Teste de Breusch-Pagan

Teste de Breusch-Pagan	p-valor	Resultado
Modelo	0.0337246	Sucesso

Teste de Shapiro-Wilks	p-valor	Resultado
Modelo	0.7030176	Sucesso

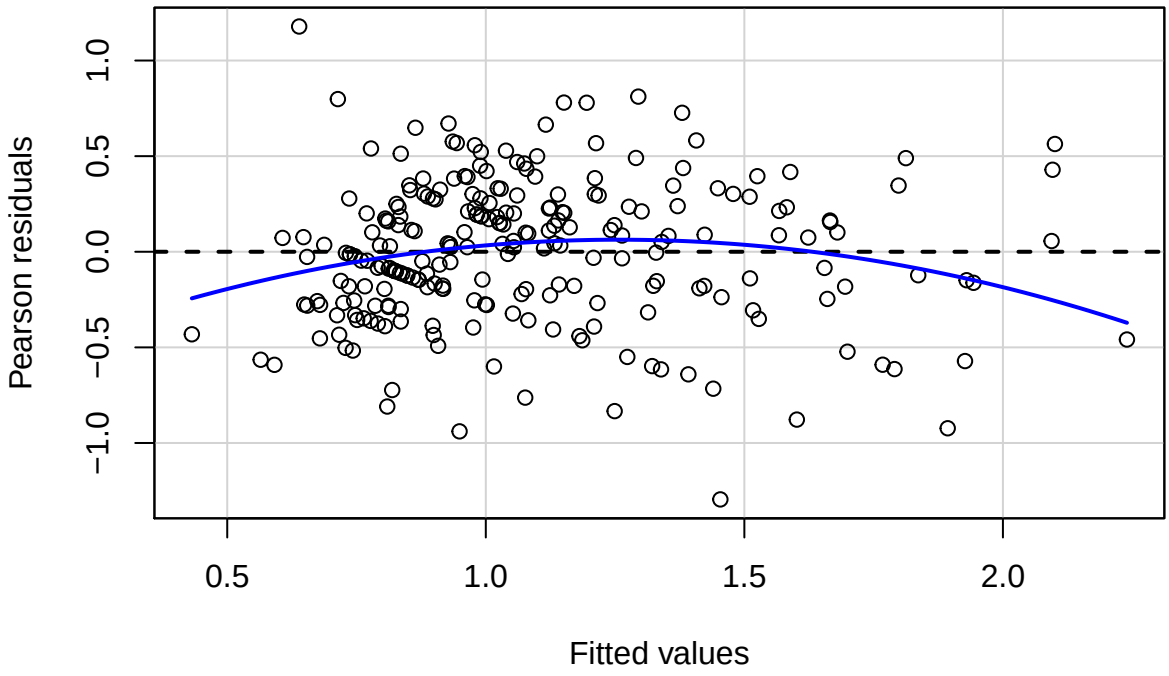
o p-valor do teste desse modelo é inferior a 0.01. Portanto não há diferença entre a distribuição dessas variáveis. Como o modelo atendeu a todas as exigências propostas o usaremos como modelo válido para a nossa análise, que apresenta R^2 em um valor de 0.4512, R^2 ajustado no valor de 0,4326, uma estatística F de 24,16 com 8 e 235 graus de liberdade, além de um p-valor menor que $2.2e^{-16}$, obtendo assim a seguinte equação:

$$Y_i = 0.2593 + 0.0318X_1 + 0.0744X_2 + \epsilon_i \tag{4}$$

onde as variáveis do modelo são:

- X_1 = Total da Conta
- X_2 = Tamanho do Grupo

Uma vez analisados os elementos que compóem a equação, temos a seguir um gráfico de resíduos do modelo selecionado:



6. Conclusão

Ao interpretar o modelo, percebemos que menos variáveis realmente influenciam o valor da gorjeta. Por exemplo, pelo que vimos nos gráficos da seção 4, nem o dia da semana e nem o período do dia foram variáveis preditoras no modelo final, mesmo se considerarmos que as pessoas frequentam mais restaurantes nas noites de sexta, ou sábado. Através da analise dos dados aqui apresentados fica evidente que a gorjeta recebida tem apenas relação com total da conta e o tamanho do grupo.

Lembrando que essas estatísticas são baseadas nos dados coletados, talvez com uma amostra maior mais variáveis poderiam ser mais relevantes e seria necessário um novo modelo de regressão.

7. Bibliografia

RISCO. In: MICHAELIS: moderno dicionário da língua portuguesa. São Paulo: Companhia Melhoramentos, 1998-(Dicionários Michaelis). Disponível em: https://michaelis.uol.com.br/moderno-portugues/busca/portugues-brasileiro/gorjeta/. Acesso em: 18, nov. 2021.

KUTNER, M. H., NACHTSHEIM, C., NETER, J., & LI, W. (2005). Applied Linear Statistical Models.