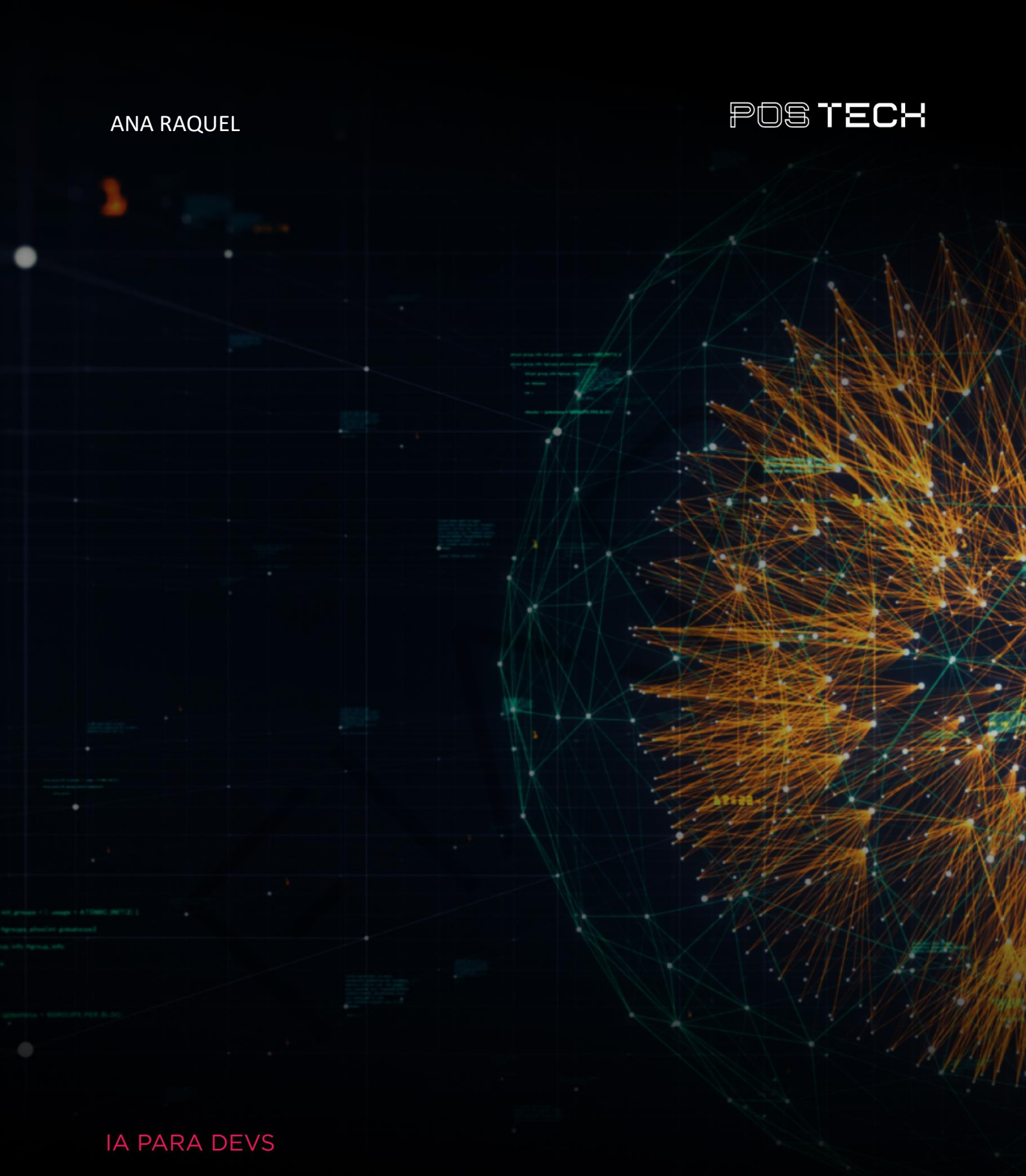
ANA RAQUEL

PDS TECH



IA PARA DEVS

MACHINE LEARNING

AULA 02

SUMÁRIO

O QUE VEM POR AÍ?

Talvez um dos objetivos mais comum em estatística seja entender o quanto uma variável X (ou várias X1, X2 e assim por diante) pode ser associada a uma variável Y e o quanto essa variável X pode ser utilizada para prever a Y. Não existe nexo que seja mais forte entre a estatística e a ciência de dados quanto à **predição de uma variável target dado algumas variáveis características**. Nesta aula, você aprenderá como funciona um dos modelos clássicos de predição: a **regressão linear**. Então, vamos entender como funciona esse tipo de aprendizado de máquina?!

HANDS ON

Utilizando o Python como linguagem, você vai aprender na prática como construir um modelo de regressão linear múltipla para prever os valores médios das casas nos distritos da Califórnia, considerando uma série de características desses distritos.

SAIBA MAIS

Regressão Linear Simples

Podemos definir a regressão linear como um modelo estatístico que mede a relação entre uma variável característica com uma outra variável alvo. Por exemplo, vamos supor que uma empresa de sorvetes queria criar um modelo preditivo para prever as vendas de sorvete. Vamos supor que a nossa base de dados contenha uma variável X composta pela temperatura e uma outra variável Y composta pelas vendas. O objetivo desse modelo é entender como a temperatura influencia as vendas, ou seja, conforme a temperatura aumenta, as vendas também aumentam. A correlação é uma outra maneira de medir como duas variáveis estão correlacionadas, mas a diferença é que, enquanto a correlação mede a força de uma associação entre duas variáveis, a regressão quantifica a natureza do relacionamento.

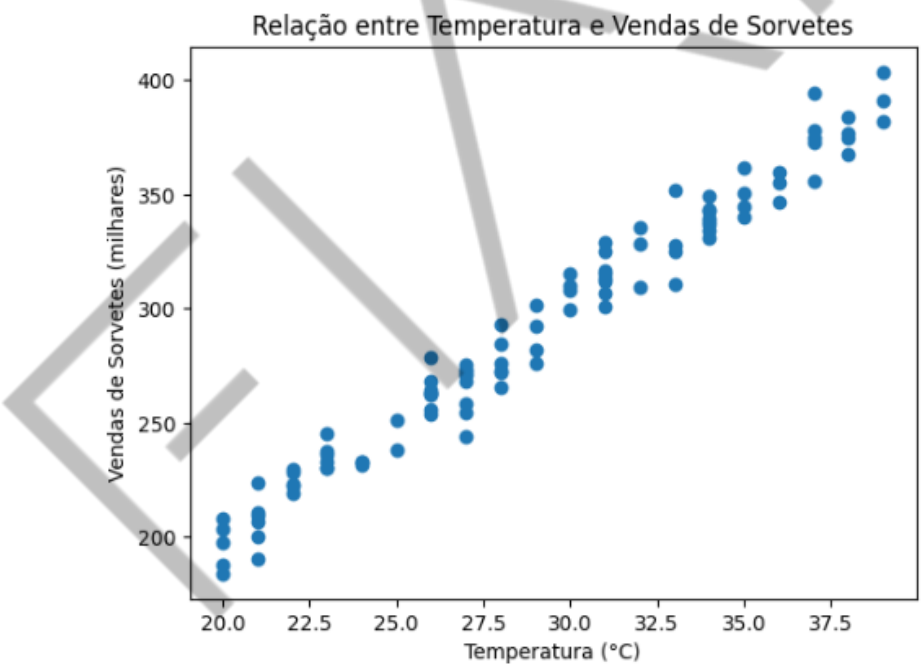


Figura 1 - Relação entre temperatura e vendas de sorvetes
Fonte: elaborado pela autora (2024)

Equação da Regressão Linear Simples

A regressão linear estima o quanto Y mudará quando X mudar em uma certa quantidade. Para descobrir o valor de Y, é utilizada a seguinte função linear (ou seja, uma linha):

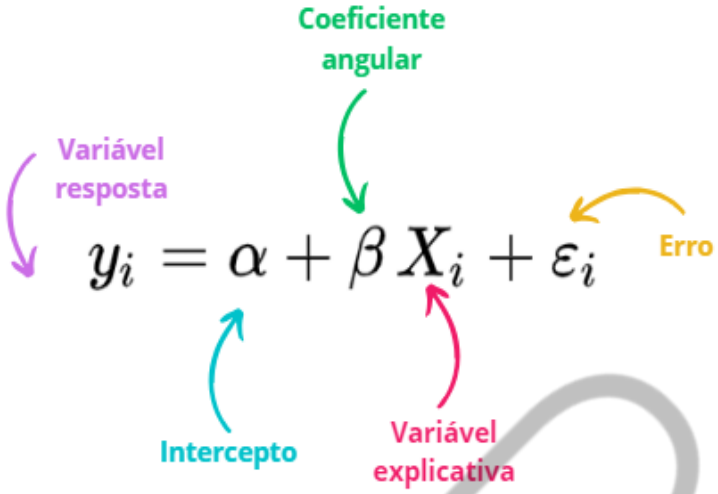


Figura 2 - Equação Regressão Linear Simples
Fonte: elaborado pela autora (2024)

Podemos ler essa equação dizendo que **Yi** é igual a **β** vezes **Xi**, mais uma constante **α**, mais um erro **ε**i. O **α** nesta equação é conhecido como o **interceptor da linha da regressão** (ou seja, é o valor previsto quando X=0). O **β** é conhecido como **declive da equação** para X. A variável **X** é conhecida como a **variável preditora** (ou também independente) e **Y** é a **variável resposta** (alvo da predição). O **ε** seriam os **resíduos da regressão**, ou seja, são a diferença entre os valores observados e ajustados (podemos definir também como o erro da equação).

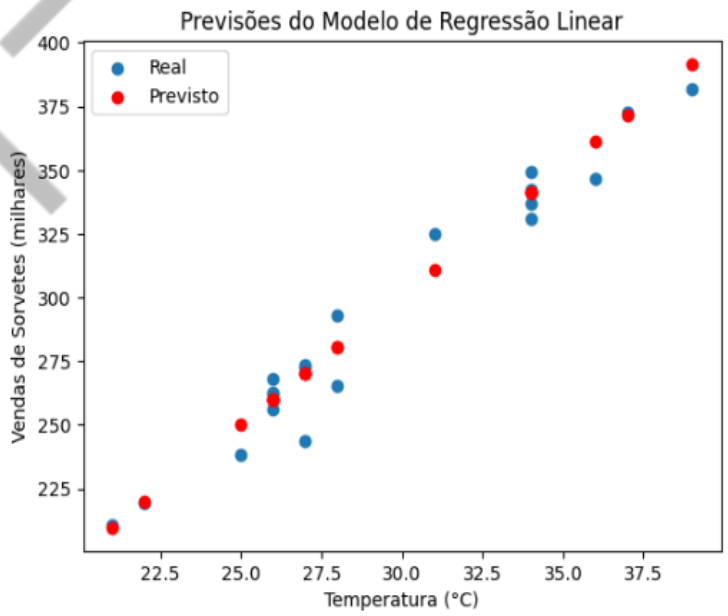


Figura 3 - Modelo de regressão linear - real x previsto

Fonte: elaborado pela autora (2024)

Se fizermos uma associação do exemplo das vendas de sorvetes com esta equação, teríamos o seguinte:

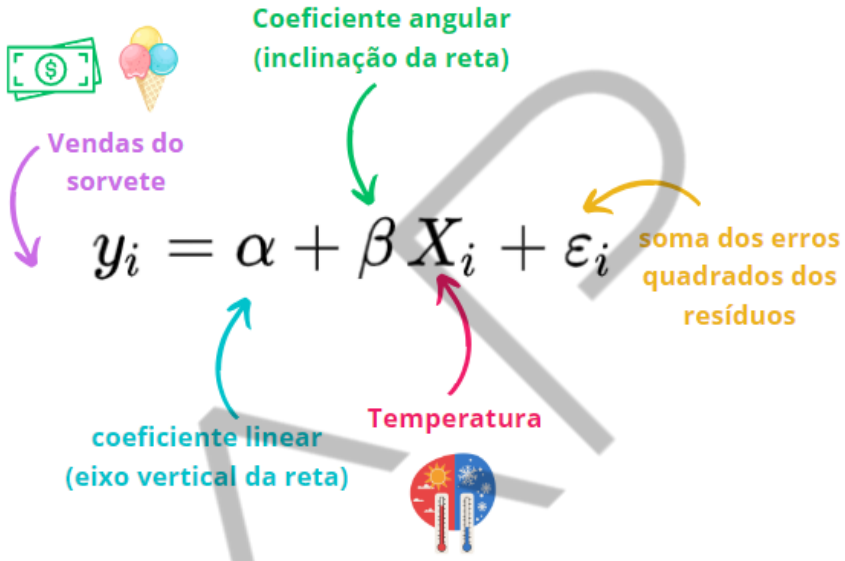


Figura 4 - Equação Regressão Linear Simples sobre preços de sorvetes explicado pela temperatura
Fonte: elaborado pela autora (2024)

O objetivo da regressão linear é encontrar os valores de **α** e **β** que **minimizam a soma dos erros quadrados dos resíduos ε.** Os resíduos são as diferenças entre os valores observados (vendas dos sorvetes) e os valores previstos de vendas na target do modelo (alvo). Queremos aqui na equação, que a reta esteja cada vez mais próxima dos valores reais.

Mínimos quadrados

Como o modelo é ajustado aos dados? Na prática, a linha de regressão é a estimativa que minimiza a soma dos valores quadrados dos resíduos, também conhecida como a “Soma dos Erros Quadrados” (ou “Residual Sum of Squares” ou “RSS”). O método de minimizar a soma dos resíduos quadrados é chamado de “Regressão de Mínimos Quadrados” ou “Regressão de Mínimos Quadrados Ordinários (ou ainda “Ordinary Least Squares” - OLS).

Regressão linear múltipla

Podemos colocar a regressão linear múltipla como uma extensão da regressão linear simples, onde simplesmente a equação se estende para acomodar múltiplas preditoras. Para esse tipo de regressão, existe mais de uma variável independente para prever a variável dependente. Ainda colocando o exemplo do sorvete, imagine nesse caso que, além da temperatura para prever as vendas, acrescentamos a variável de promoção de marketing. A equação dessa regressão múltipla seria desta forma:

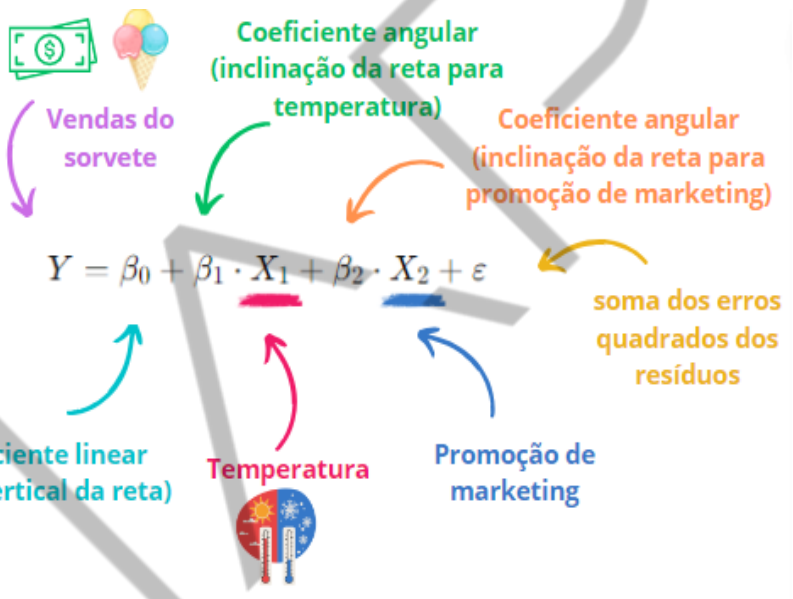


Figura 5 - Equação Regressão Linear Múltipla sobre preços de sorvetes explicado pela temperatura e promoção de marketing
Fonte: elaborado pela autora (2024)

Em vez de uma linha, agora temos um modelo linear - o relacionamento entre cada coeficiente e sua variável (característica) é linear. No exemplo do sorvete, utilizamos apenas duas variáveis independentes, mas o modelo de regressão linear múltipla pode ter muitas variáveis.

Métricas de avaliação de modelos de regressão

A métrica de desempenho mais importante é a “Raiz Quadrada do Erro Quadrático” (ou “Root Mean Squared Error” - RMSE), que é a raiz quadrada do erro quadrado médio nos valores previstos em Y.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - p_i)^2}$$

Figura 6 - Equação do RMSE
Fonte: Mario Filho (2023)

O RMSE é a raiz quadrada do MSE (“Erro Quadrático Médio” - Mean Squared Error). O MSE é a média dos quadrados das diferenças entre os valores reais e os valores previstos.

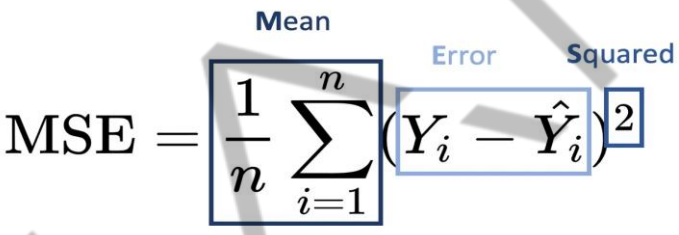


Figura 7 - Equação do MSE
Fonte: SUBOPTIMaL (s.d)

Na prática, para regressão linear, a diferença entre o RMSE e o MSE é muito pequena. Para trabalhar com a regressão linear no Python, as bibliotecas **statsmodels** e **sklearn** são muito boas. Que tal dar uma espiada nessas bibliotecas oficiais?

Em específico, a statsmodel possui algo muito próximo do summary() que mostra um resumo dos resultados da regressão (quem é “amante” da linguagem R, vai gostar de utilizar na biblioteca statsmodels).

Para instalar a biblioteca do **statsmodels**, você pode realizar o seguinte comando:

! pip install statsmodels

Caso também não tenha instalado a biblioteca do Pandas, utilize o seguinte código para realizar este procedimento:

! pip install pandas

Depois de todo o ambiente instalado, vamos analisar um exemplo de regressão linear múltipla dos resultados no cenário de vendas de sorvete:

```
import statsmodels.api as sm
import pandas as pd

# Criar um DataFrame com dados fictícios
data = {'Vendas_Sorvetes': [200, 300, 400, 350, 500],
        'Temperatura': [28, 30, 32, 29, 33],
        'Promocao_Marketing': [1000, 1200, 800, 900, 1100]}

df = pd.DataFrame(data)

# Adicionar uma constante para o termo de intercepto
df['Intercepto'] = 1

# Definir as variáveis independentes (X)
X = df[['Intercepto', 'Temperatura', 'Promocao_Marketing']]

# Definir a variável dependente (Y)
Y = df['Vendas_Sorvetes']

# Criar e ajustar o modelo de regressão linear múltipla
modelo = sm.OLS(Y, X).fit()

# Imprimir os resultados do modelo
print(modelo.summary())
```

Vamos observar alguns dos principais resultados em summary()?

OLS Regression Results

Dep. Variable:	Vendas_Sorvetes	R-squared:	0.845
Model:	OLS	Adj. R-squared:	0.690
Method:	Least Squares	F-statistic:	5.456
Date:	Mon, 20 Nov 2023	Prob (F-statistic):	0.155
Time:	21:25:14	Log-Likelihood:	-25.458
No. Observations:	5	AIC:	56.92
Df Residuals:	2	BIC:	55.74
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercepto	-1102.3256	497.538	-2.216	0.157	-3243.058	1038.407
Temperatura	49.4186	15.004	3.294	0.081	-15.139	113.976
Promocao_Marketing	-0.0500	0.197	-0.254	0.823	-0.897	0.797

Omnibus:	nan	Durbin-Watson:	1.622
Prob(Omnibus):	nan	Jarque-Bera (JB):	0.632
Skew:	0.604	Prob(JB):	0.729
Kurtosis:	1.745	Cond. No.	1.81e+04

Figura 8 - Summary statsmodels
Fonte: elaborado pela autora (2024)

Sob a coluna **"coef"**, você encontrará os valores dos coeficientes estimados para cada variável. No exemplo dado, você terá coeficientes para a constante (intercepto), temperatura e promoção de marketing.

A coluna **"std err"** mostra o erro padrão dos coeficientes. Quanto menor, melhor, pois indica uma estimativa mais precisa.

O **valor-p (P>|t|)** é usado para testar a hipótese nula de que o coeficiente associado a cada variável é igual a zero. Valores pequenos (geralmente < 0,05) indicam que você pode rejeitar a hipótese nula e considerar o coeficiente significativo.

O **"R-squared (R2)"** fornece informações sobre o ajuste geral do modelo. O valor do R2 pode variar entre 0 e 1, quanto mais próximo de 1, melhor, pois indica que o modelo explica uma maior proporção da variabilidade nos dados. O R2 é também uma das principais métricas de avaliação do modelo de regressão. Neste caso, podemos concluir que o modelo explica 85% da variabilidade nas vendas de sorvetes.

A **estatística t** - e sua imagem espelhada, o valor p - mede até onde o coeficiente é “estatisticamente significativo”, ou seja, fora da faixa do que um arranjo aleatório casual de variáveis preditoras e alvo poderiam causar. Quanto maior a estatística t e menor o valor p, mais significativo é o preditor.

Os valores de RMSE e MSE não vem por padrão no summary do statsmodels, mas você pode analisá-los importando do sklearn conforme código a seguir:

```
! pip install scikit-learn
from sklearn.metrics import mean_absolute_error, mean_squared_error
import numpy as np


# realizar as previsões
y_pred = modelo.predict(X)


# Calcular MAE, MSE e RMSE
mae = mean_absolute_error(Y, y_pred)
mse = mean_squared_error(Y, y_pred)
rmse = np.sqrt(mse)


# Imprimir as métricas
print(f'MAE: {mae}')
print(f'MSE: {mse}')
print(f'RMSE: {rmse}')
```

Além das métricas RMSE e MSE também podemos analisar a **MAE** (erro médio absoluto). Para analisar essa métrica é muito simples, um valor pequeno para MAE significa que suas previsões estão próximas das reais. Vamos analisar os resultados dessas três métricas a seguir, aplicadas sobre a regressão linear múltipla de vendas de sorvetes.

```
MAE: 36.27906976744235
MSE: 1548.8372093023268
RMSE: 39.35526914280128
```

Figura 9 - Métricas de validação de regressão
Fonte: elaborado pela autora (2024)

Um MAE 37 (arredondando) significa que nossa previsão de vendas de sorvete, em média, está se desviando em 37 unidades das vendas reais. Um valor de MSE baixo significa que o modelo está errando menos, sendo mais preciso. Neste caso, o valor está em média. O quadrado do desvio entre previsões e valores reais é de 1548 (arredondando). O RMSE fornece uma interpretação mais intuitiva, pois está na mesma escala da variável dependente. Analisando este caso das vendas de sorvetes, as vendas desviam-se em 39 (arredondando) unidades das vendas reais.

O QUE VOCÊ VIU NESTA AULA?

Nesta aula, você aprendeu sobre um dos modelos de aprendizado de máquina clássico: a regressão linear. O processo de estabelecer um relacionamento entre uma ou múltiplas variáveis preditoras e uma variável resultante é um dos trabalhos mais requisitados no mundo da estatística e da ciência de dados.

REFERÊNCIAS

FILHO, M. **RMSE (Raiz Do Erro Quadrático Médio) Em Machine Learning**. 2023. Disponível em: <https://mariofilho.com/rmse-raiz-do-erro-quadratico-medio-em-machine-learning/> Acesso em: 06 fev. 2024.

GÉRON, A. **Estatística prática para cientistas de dados: 50 conceitos essenciais**. O'Reilly Media, Inc., 2019

HALTAKOV, V. **Mean Squared Error (MSE)**. [s.d]. Disponível em: <https://suboptimal.wiki/explanation/mse/> Acesso em: 06 fev. 2024.

PALAVRAS-CHAVE

Regressão linear, Regressão linear múltipla, R2, RMSE.