Modelo de Regressão Linear Múltipla para Previsão de Safra Agrícola

Seminário - Data Science

Guilherme Zanin, João Lucas Criveli, Docente: Clayton Reginaldo Pereira

25 de junho de 2025

Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP)
Faculdade de Ciências (FC) / Departamento de Computação (DCo)
Bauru, SP - Brasil

Sumário da Apresentação

Introdução

Introdução

- A soja é um dos principais produtos agrícolas do Brasil, com grande impacto econômico.
- Prever a produção é crucial para planejamento e políticas agrícolas.
- **Objetivo**: Desenvolver um modelo de regressão linear múltipla para prever a quantidade produzida e a área colhida, e testar hipóteses sobre as relações entre variáveis.

Descrição dos Dados

- **Fonte**: Tabela 1612 do IBGE (Produção Agrícola Municipal).
- **Variáveis**:
  - Área plantada (hectares)
  - Área colhida (hectares)
  - Quantidade produzida (toneladas)
  - Valor da produção (mil reais)
  - Estado (unidade da federação)
  - Ano (2010–2023)
- **Pré-processamento**: Tratamento de valores ausentes, conversão de tipos de dados, pivotamento.

Fundamentação Teórica

Regressão Linear Múltipla

A regressão linear múltipla é uma técnica estatística utilizada para modelar a relação entre uma variável dependente contínua e duas ou mais variáveis independentes. Sua forma geral pode ser representada por:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Onde:

- Y é a variável dependente (ex: área colhida ou valor da produção);
- $X_1, X_2, \ldots, X_p$ são as variáveis independentes (ex: área plantada, estado, ano);
- β_0 é o intercepto;
- $\beta_1, \ldots, \beta_p$ são os coeficientes associados a cada variável;
- ϵ é o termo de erro.

Métricas para avaliação

- **Coeficiente de determinação (R^2):** mede o quanto da variabilidade da variável dependente é explicada pelo modelo.
- **Erro quadrático médio (RMSE):** ndica o desvio médio das previsões em relação aos valores reais.

Desenvolvimento

Limpeza e transformação dos dados

```
df['Valor'] = df['Valor'].replace('-', np.nan)
df['Valor'] = pd.to_numeric(df['Valor'], errors='coerce')

df_pivot = df.pivot_table(index=['Unidade da Federação', 'Ano'],
           columns='Variável', values='Valor').reset_index()
```

Seleção de Variáveis

- **Variáveis independentes**: Área plantada, estado (codificado posteriormente), ano;
- **Variáveis dependentes**: Área colhida e valor da produção.

A codificação da variável `Estado` foi feita por meio de *one-hot encoding*. Essa técnica transforma uma variável categórica em múltiplas variáveis binárias (0 ou 1). Para cada categoria distinta (no caso, cada estado brasileiro), é criada uma nova coluna no conjunto de dados. Um valor 1 é atribuído à coluna correspondente ao estado de origem da observação, e 0 às demais.

Treinamento dos modelos

- **Divisão dos dados:** 80/20.
- **2 modelos treinados:** Previsão da quantidade produzida e previsão da área colhida.

Treinamento dos modelos

```
X = df_pivot_soja[['Estado', 'Area_Plantada', 'Ano']]
y_quantidade = df_pivot_soja['Quantidade_Produzida']
y_area_colhida = df_pivot_soja['Area_Colhida']

modelo_quantidade = LinearRegression()
modelo_quantidade.fit(X_train, y_quantidade_train)

modelo_area_colhida = LinearRegression()
modelo_area_colhida.fit(X_train, y_area_colhida_train)
```

Exemplo de aplicação do modelo

```
[21]: # Testa previsões para diferentes estados com mesmos valores
estados_teste = ['São Paulo', 'Mato Grosso', 'Paraná']
novos_dados = pd.DataFrame({
    'Estado': estados_teste,
    'Ano': [2025] * len(estados_teste),
    'Area_Plantada': [3694468.0] * len(estados_teste),
})

print(novos_dados.head(3))
novos_dados_encoded = preprocessor.transform(novos_dados)
quantidade_prevista = modelo_quantidade.predict(novos_dados_encoded)
area_colhida_prevista = modelo_area_colhida.predict(novos_dados_encoded)

# Imprime previsões para diferentes estados
print("\nPrevisões para diferentes estados (Ano=2023, Área Plantada=100000, Área Colhida=100000):")
for estado, qtd, area in zip(estados_teste, quantidade_prevista, area_colhida_prevista):
    print(f"{estado}: Quantidade = {qtd:.2f} toneladas, Área Colhida = {area:.2f} hectares")
```

```
        Estado   Ano  Area_Plantada
0    São Paulo  2025      3694468.0
1  Mato Grosso  2025      3694468.0
2       Paraná  2025      3694468.0

Previsões para diferentes estados (Ano=2023, Área Plantada=100000, Área Colhida=100000):
São Paulo: Quantidade = 14392360.36 toneladas, Área Colhida = 3696166.86 hectares
Mato Grosso: Quantidade = 7798158.80 toneladas, Área Colhida = 3677696.02 hectares
Paraná: Quantidade = 11019632.94 toneladas, Área Colhida = 3691825.15 hectares
```

Figura 1: Exemplo de Previsão Agrícola

Avaliação dos modelos

Avaliação dos modelos

- RMSE (Quantidade produzida): 2.461.662.352.880,37
- R^2 (Quantidade produzida): 0,9424
- RMSE (Área colhida): 444.540.571,60
- R^2 (Área colhida): 0,9999

Gráficos de dispersão

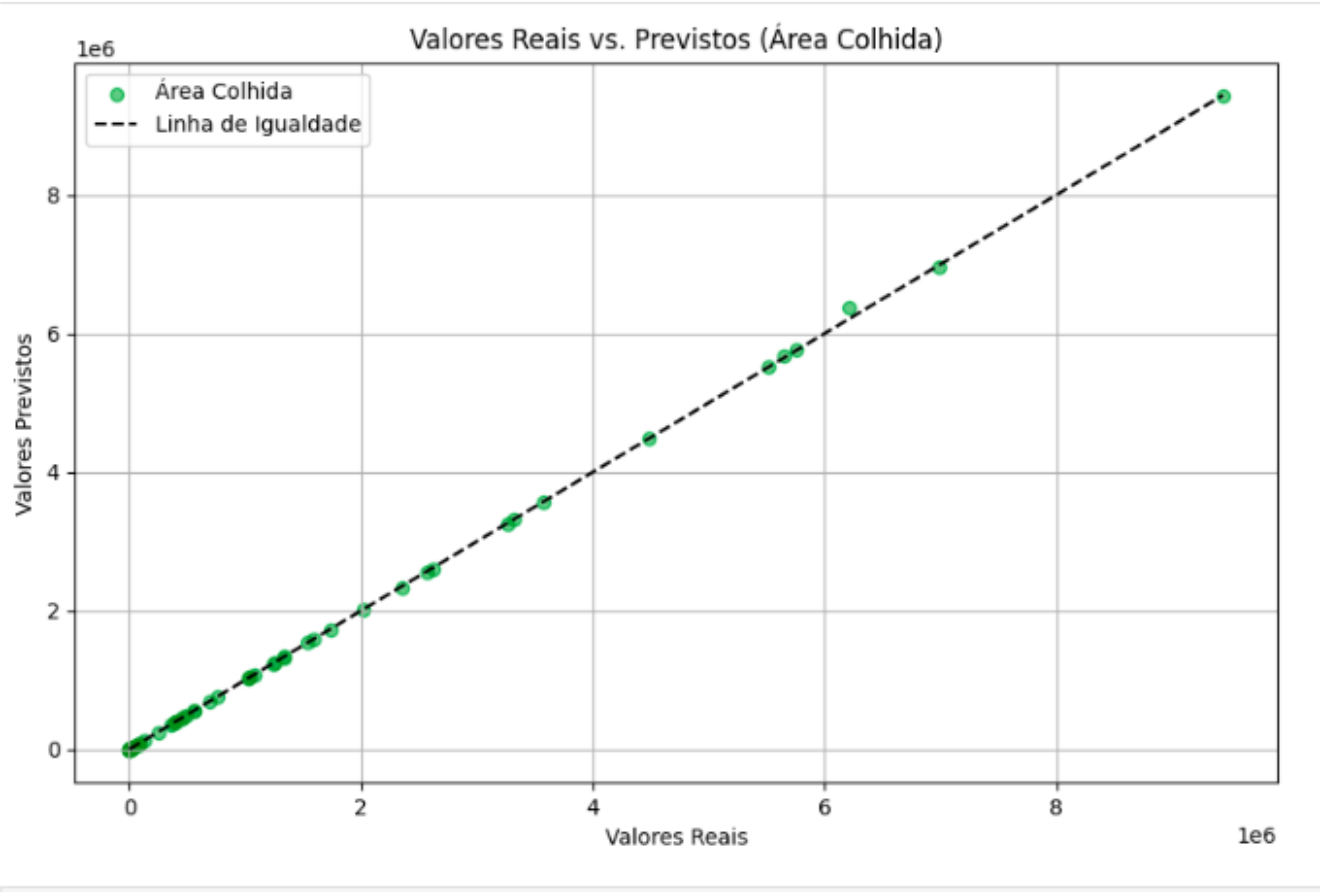

Figura 2: Valores Reais x Previstos - Área Colhida

Gráficos de dispersão

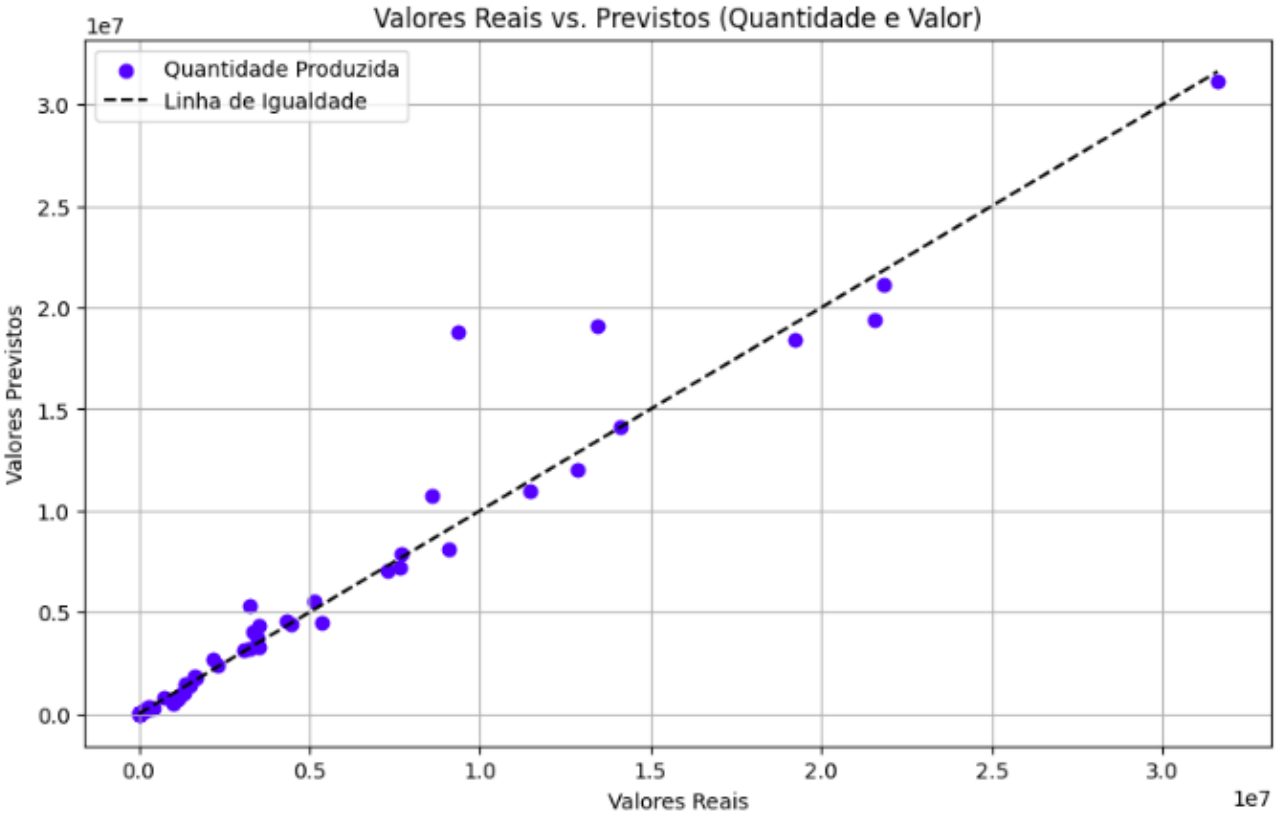

Figura 3: Valores Reais x Previstos - Quantidade e Valor

Gráficos de resíduos

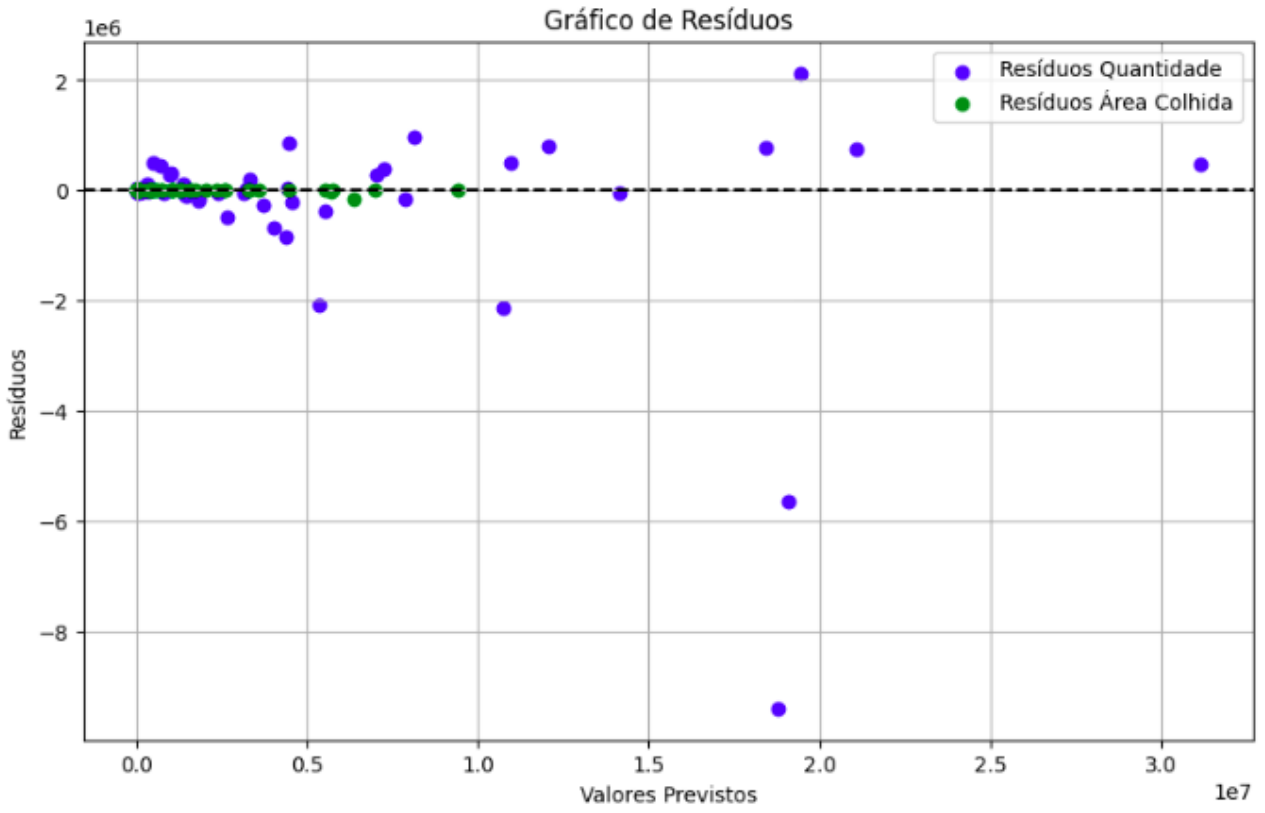

Figura 4: Resíduos

Testes de hipótese

Testes de Hipótese

- **Teste 1**: Significância de Area_Plantada na Quantidade_Produzida
  - H_0: $\beta_{\text{Area\_Plantada}} = 0$
  - H_1: $\beta_{\text{Area\_Plantada}} \neq 0$
  - Teste t
- **Teste 2**: Significância geral do modelo para Area_Colhida
  - H_0: Todos os coeficientes $= 0$
  - H_1: Pelo menos um coeficiente $\neq 0$
  - Teste F

Resultados - Teste 1

- **P-valor** para Area_Plantada: $7,76 \times 10^{-207}$
- **Conclusão**: Rejeita H_0; Area_Plantada tem efeito significativo na Quantidade_Produzida.
- **Interpretação**: Aumentar a área plantada está associado a maior produção de soja.

Resultados - Teste 2

- **P-valor** do teste F: $0,0$
- **Conclusão**: Rejeita H_0; o modelo para Area_Colhida é estatisticamente significativo.
- **Interpretação**: As variáveis (Area_Plantada, Estado, Ano) explicam a variância em Area_Colhida.
- Por meio de outros testes, constatou-se que a Área Plantada é o fator que mais influencia no valor da Área Colhida, e a variável "Ano" NÃO possui influência no resultado da previsão.

Conclusão

Discussão

- A significância de Area_Plantada destaca a importância de políticas de expansão agrícola.
- O modelo para Area_Colhida é válido, sugerindo influência de estado e ano.
- **Implicações**: Resultados orientam alocação de recursos e estratégias de plantio.

Conclusão

- **Resumo**: Area_Plantada é um preditor chave; modelo para Area_Colhida é significativo.
- **Recomendações**: Considerar Area_Plantada no planejamento agrícola.
- **Próximos passos**: Explorar modelos não lineares e variáveis adicionais (e.g., clima).

Perguntas?

Obrigado pela atenção!