Regressão Linear

Estatística Inferencial

MBA CDIA

ENAP - Escola Nacional de Administração Pública

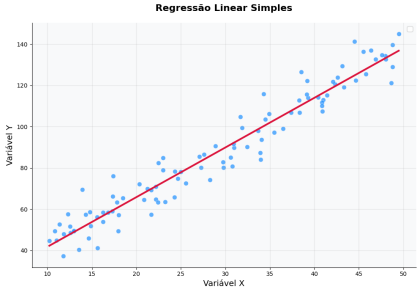
2025

Agenda

- Introdução à Regressão Linear
- Conceitos Fundamentais
- Regressão Linear Simples
- Análise de Diagnóstico
- Regressão Linear Múltipla
- Interpretação de Resultados
- Aplicações Práticas

O que é Regressão Linear?

- Técnica estatística que modela a relação entre uma variável dependente (Y) e uma ou mais variáveis independentes (X)
- Objetivo: encontrar a melhor linha reta que se ajuste aos dados
- Minimiza a soma dos quadrados dos resíduos (Método dos Mínimos Quadrados Ordinários - MQO)
- Equação geral: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$



Pressupostos da Regressão Linear

- **Linearidade:** Relação linear entre X e Y
- **Independência:** Observações independentes entre si
- **Homocedasticidade:** Variância constante dos resíduos
- **Normalidade:** Resíduos seguem distribuição normal
- **Ausência de multicolinearidade:** Variáveis explicativas não perfeitamente correlacionadas

Notação:

- $Y \sim N(\mu, \sigma^2)$
- $\epsilon \sim N(0, \sigma^2)$

Regressão Linear Simples

Modelo com uma variável explicativa:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- β_0 = intercepto (valor de Y quando X = 0)
- β_1 = coeficiente angular (mudança em Y para uma unidade de X)
- ϵ = termo de erro aleatório

Objetivo: Estimar $\hat{\beta}_0$ e $\hat{\beta}_1$ que minimizam:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

Exemplo: IDEB vs. Taxa de Abandono Escolar

Dataset: Educação Municipal (IDEB vs. Abandono Escolar)

[PLACEHOLDER: Scatter plot Taxa Abandono vs IDEB com linha de regressão]

Modelo: $IDEB = \hat{\beta}_0 + \hat{\beta}_1 \times Taxa\_Abandono$
Interpretação:

- Para cada 1% de aumento na taxa de abandono, esperamos uma redução de $|\beta_1|$ pontos no IDEB
- β_0 representa o IDEB quando não há abandono escolar

Output Típico da Regressão

Variable	Coef	Std Error	t-value	P¿—t—
Intercept	6.8420	0.1124	60.87	0.000
Taxa_Abandono	-0.0892	0.0087	-10.25	0.000

Métricas de Qualidade:

- R-squared: 0.4821 (48,21% da variação do IDEB explicada)
- Residual Std Error: 0.634
- F-statistic: 105.1 ($p < 0.001$)

Análise de Diagnóstico

Métricas de Avaliação:

- **RSE (Residual Standard Error):** Desvio padrão dos resíduos
- **R^2 (Coeficiente de Determinação):** Proporção da variância explicada
- **RMSE (Root Mean Squared Error):** $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

Gráficos de Diagnóstico:

- Resíduos vs. Valores Ajustados
- Q-Q Plot (Normalidade dos Resíduos)
- Scale-Location (Homocedasticidade)
- Resíduos vs. Leverage (Outliers)

[MOSTRAR]

Regressão Linear Múltipla

Modelo com múltiplas variáveis:

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

Exemplo - Fatores que Impactam o IDEB:

$$IDEB = \beta_0 + \beta_1 \times Taxa\_Abandono + \beta_2 \times Distorcao\_Idade + \beta_3 \times Horas\_Aula + \epsilon$$

Vantagens:

- Controla o efeito de múltiplas variáveis simultaneamente
- Maior poder explicativo (R^2 mais alto)
- Reduz viés de variáveis omitidas

Desafios:

- Multicolinearidade entre indicadores educacionais
- Interpretação mais complexa dos efeitos
- Overfitting com muitas variáveis explicativas

Interpretação dos Coeficientes

Modelo múltiplo:

$$IDEB = 2.1 - 0.082 \times Taxa\_Abandono - 0.034 \times Distorcao\_Idade + 0.0012 \times Horas\_Aula$$

Interpretação:

- **Taxa_Abandono (-0.082):** Mantendo outros fatores constantes, cada 1% de aumento no abandono reduz o IDEB em 0.082 pontos
- **Distorcao_Idade (-0.034):** Mantendo outros fatores constantes, cada 1% de aumento na distorção idade-série reduz o IDEB em 0.034 pontos
- **Horas_Aula (0.0012):** Mantendo outros fatores constantes, cada hora-aula adicional no ano aumenta o IDEB em 0.0012 pontos

Cuidados:

- Interpretação é *ceteris paribus* (tudo mais constante)
- Correlação Causalidade
- Validade limitada ao intervalo dos dados observados

Seleção de Variáveis

Métodos de Seleção:

- **Forward Selection:** Adiciona variáveis progressivamente
- **Backward Elimination:** Remove variáveis não significativas
- **Stepwise:** Combinação de ambos os métodos

Critérios de Avaliação:

- **AIC (Akaike Information Criterion):** Penaliza complexidade
- **BIC (Bayesian Information Criterion):** Penalização mais forte
- **Adjusted R^2:** R^2 ajustado pelo número de variáveis
- **Cross-validation:** Validação cruzada

Princípio da Parcimônia: Prefira modelos mais simples com desempenho similar

Limitações e Cuidados

Quando NÃO usar Regressão Linear:

- Relações não-lineares complexas
- Variável dependente categórica (usar regressão logística)
- Violação severa dos pressupostos
- Presença de muitos outliers

Cuidados Importantes:

- **Extrapolação:** Não fazer previsões fora do intervalo dos dados observados
- **Causalidade:** Regressão mostra associação, não causalidade
- **Multicolinearidade:** Verificar correlação entre variáveis explicativas
- **Outliers:** Podem distorcer significativamente os resultados
- **Contexto das políticas:** Considerar fatores externos não modelados

Exercício Prático 1: Regressão Linear Simples

Dataset: Dados Educacionais (Taxa de Abandono vs IDEB)

```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score
import matplotlib.pyplot as plt

from google.colab import files

# Fazer upload do arquivo do PC
uploaded = files.upload()
# Pegar o nome do arquivo enviado
filename = list(uploaded.keys())[0]

# Ler o arquivo CSV
df = pd.read_csv(filename)

# Preparar dados
X = df[['taxa_abandono']].values
y = df['ideb'].values

# Dividir em treino e teste
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Ajustar modelo
modelo = LinearRegression()
modelo.fit(X_train, y_train)
```

Exercício Prático 2: Regressão Linear Múltipla

Dataset: Dados Educacionais (Múltiplos Fatores que Impactam o IDEB)

```
# Modelo múltiplo
X_multi = df[['taxa_abandono', 'distorcao_idade', 'horas_aula']].values
X_train_multi, X_test_multi, y_train_multi, y_test_multi = train_test_split(
    X_multi, df.ideb, test_size=0.3, random_state=42)

# Ajustar modelo múltiplo
modelo_multi = LinearRegression()
modelo_multi.fit(X_train_multi, y_train_multi)

# Previsões e avaliação
y_pred_multi = modelo_multi.predict(X_test_multi)
rmse_multi = np.sqrt(mean_squared_error(y_test_multi, y_pred_multi))
r2_multi = r2_score(y_test_multi, y_pred_multi)

print("Modelo Múltiplo - Fatores Realistas do IDEB:")
print(f'Intercepto: {modelo_multi.intercept_:.4f}')
variaveis = ['taxa_abandono', 'distorcao_idade', 'horas_aula']
for i, var in enumerate(variaveis):
    print(f'Coef {var}: {modelo_multi.coef_[i]:.6f}')
print(f'RMSE: {rmse_multi:.4f}')
print(f'R²: {r2_multi:.4f}')
```

Exercício Prático 3: Análise de Diagnóstico

```
import seaborn as sns
from scipy import stats

# Gráficos de diagnóstico para modelo IDEB realista
fig, axes = plt.subplots(2, 2, figsize=(12, 10))
fig.suptitle('Diagnóstico: Fatores Realistas do IDEB', fontsize=16)

# 1. Resíduos vs Valores Ajustados
residuos = y_test_multi - y_pred_multi
axes[0,0].scatter(y_pred_multi, residuos, alpha=0.6, color='steelblue')
axes[0,0].axhline(y=0, color='red', linestyle='--')
axes[0,0].set_xlabel('IDEB Predito')
axes[0,0].set_ylabel('Resíduos')
axes[0,0].set_title('Resíduos vs IDEB Predito')

# 2. Q-Q Plot
stats.probplot(residuos, dist="norm", plot=axes[0,1])
axes[0,1].set_title('Q-Q Plot - Normalidade dos Resíduos')

# 3. Histograma dos resíduos
axes[1,0].hist(residuos, bins=20, alpha=0.7, density=True, color='lightgreen')
axes[1,0].set_xlabel('Resíduos')
axes[1,0].set_ylabel('Densidade')
axes[1,0].set_title('Distribuição dos Resíduos')

# 4. IDEB Real vs Predito
axes[1,1].scatter(y_test_multi, y_pred_multi, alpha=0.6, color='orange')
axes[1,1].plot([y_test_multi.min(), y_test_multi.max()],
               [y_test_multi.min(), y_test_multi.max()], 'r--', lw=2)
axes[1,1].set_xlabel('IDEB Real')
```

Aplicações em Políticas Públicas

Exemplos de Uso:

- **Educação:** Relação entre abandono escolar e desempenho no IDEB
- **Saúde:** Impacto de variáveis socioeconômicas na expectativa de vida
- **Segurança:** Efeito de políticas de segurança na criminalidade
- **Economia:** Análise do impacto de políticas fiscais no crescimento
- **Meio Ambiente:** Relação entre regulamentação e qualidade do ar

Vantagens para o Setor Público:

- Simplicidade na interpretação
- Base para tomada de decisões baseada em evidências
- Identificação de fatores-chave
- Previsão de cenários

Conclusões

Pontos-chave da Regressão Linear:

- Ferramenta fundamental para análise de relações entre variáveis
- Importante verificar pressupostos antes da aplicação
- Interpretação cuidadosa dos coeficientes
- Útil para previsão dentro do intervalo dos dados

Próximos Passos:

- Regularização (Ridge, Lasso, Elastic Net)
- Regressão Logística para variáveis categóricas
- Modelos não-lineares
- Validação cruzada e técnicas de seleção de modelos

Dúvidas?