Revisão e Notas sobre Regressão Linear

Pedro Rupf Pereira Viana

2 de fevereiro de 2026

1 Introdução e Objetivos

A regressão linear simples representa um dos pilares fundamentais da análise estatística, servindo como ferramenta essencial para investigar e modelar relações entre ariáveis em diversos campos do conhecimento humano. Surgida no contexto das ciências naturais e sociais no século XIX, com contribuições pioneiras de figuras como Francis Galton e Karl Pearson, essa técnica permite que pesquisadores e analistas explorem como uma variável pode influenciar outra de maneira previsível. Em essência, a regressão linear simples busca descrever a associação linear entre uma variável independente (também chamada de preditora ou explicativa) e uma variável dependente (ou resposta), assumindo que mudanças na primeira possam explicar variações na segunda.

Diferentemente de métodos mais avançados, como a regressão múltipla ou não linear, a versão simples foca em apenas duas variáveis, o que a torna acessível para iniciantes e útil para análises preliminares. Seu apelo reside na simplicidade: ao invés de demandar computações intensivas, ela oferece insights rápidos sobre padrões de dados, facilitando decisões informadas em áreas como economia, biologia, engenharia e ciências sociais. Por exemplo, em estudos econômicos, pode-se usar a regressão linear simples para examinar como o nível de educação de uma população afeta sua renda média, revelando tendências que orientam políticas públicas.

2 Definição

No cerne da regressão linear simples está a ideia de uma linha reta que melhor representa a relação entre as duas variáveis. Imagine um gráfico de dispersão, onde pontos de dados são plotados com base em observações reais: a variável independente no eixo horizontal e a dependente no vertical. A regressão busca traçar uma linha que passe o mais próximo possível desses pontos, minimizando as discrepâncias entre os valores observados e os previstos pela linha.

O propósito principal é duplo: descritivo e preditivo. Descritivamente, ela quantifica a força e a direção da relação – se positiva (ambas variáveis aumentam juntas) ou negativa (uma aumenta enquanto a outra diminui). Preditoriamente, permite estimar valores futuros da variável dependente com base em valores conhecidos da independente. Essa capacidade preditiva é particularmente valiosa em cenários práticos, como na previsão de vendas de um produto com base no investimento em publicidade, onde a regressão oferece uma base empírica para projeções.

É importante destacar que a regressão linear simples não implica causalidade. Mesmo que haja uma associação forte, fatores externos não considerados podem influenciar o resultado. Por isso, analistas devem interpretar os achados com cautela, integrando-os a teorias subjacentes e evidências adicionais.

Do ponto de vista matemático, a regressão linear simples é um modelo estatístico que descreve a relação entre uma variável resposta (dependente) Y e uma única variável explicativa (independente) X, assumindo que essa relação pode ser aproximada por uma reta. O modelo populacional é escrito como:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad , \quad i = 1, 2, \ldots, n \tag{2.1}$$

em que:

- β_0 = intercepto (valor esperado de Y quando $X = 0$)

- β_1 = coeficiente angular (inclinação) $\rightarrow$ mudança esperada em Y para cada aumento de 1 unidade em X
- ε_i = termo de erro aleatório, com $\mathrm{E}(\varepsilon_i) = 0$ e $\mathrm{Var}(\varepsilon_i) = \sigma^2$

3 Componentes e Suposições Básicas

A estrutura da regressão linear simples envolve elementos chave: a interceptação (o ponto onde a linha cruza o eixo vertical, representando o valor esperado da variável dependente quando a independente é zero) e a inclinação (que indica quanto a variável dependente muda para cada unidade de alteração na independente). Esses componentes formam a equação da linha, que serve como modelo para os dados.

Para que o modelo seja válido, certas suposições devem ser atendidas. Primeiramente, assume-se linearidade: a relação entre as variáveis deve ser aproximadamente reta, sem curvas ou padrões não lineares evidentes. Em segundo lugar, a independência das observações é crucial, significando que os dados não devem ser influenciados uns pelos outros, como em séries temporais onde eventos passados afetam os futuros. Terceiro, a homocedasticidade implica que as variações nos erros (diferenças entre valores observados e previstos) sejam constantes ao longo da linha, evitando que os erros aumentem ou diminuam sistematicamente. Por fim, os erros devem seguir uma distribuição normal, o que facilita inferências estatísticas sobre a confiabilidade do modelo.

Quando essas suposições são violadas, o modelo pode levar a conclusões enviesadas. Ferramentas diagnósticas, como gráficos de resíduos, ajudam a verificar essas condições, permitindo ajustes ou a adoção de métodos alternativos.

3.1 Principais Suposições do Modelo Clássico

Para que as propriedades desejáveis dos estimadores sejam válidas, assume-se:

1. **Linearidade nos parâmetros:** a média condicional é linear em X

$$\mathrm{E}(Y_i|X_i) = \beta_0 + \beta_1 X_i$$

2. **Independência dos erros:** ε_i e ε_j são independentes para $i \neq j$
3. **Homoscedasticidade:** variância constante

$$\mathrm{Var}(\varepsilon_i|X_i) = \sigma^2 \quad \text{(constante para todo } X_i\text{)}$$

4. **Normalidade (necessária para testes exatos em amostras pequenas):**

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

5. **Não-colinearidade (trivial no caso simples):** X não é constante

3.2 Método de Mínimos Quadrados Ordinários (MQO)

O método de mínimos quadrados busca os valores $\hat{\beta}_0$ e $\hat{\beta}_1$ que minimizam a soma dos quadrados dos resíduos:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \tag{3.1}$$

As soluções analíticas (fórmulas fechadas) são:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \tag{3.2}$$

onde $\bar{X}$ e $\bar{Y}$ são as médias amostrais.

A reta ajustada é, então:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{3.3}$$

4 Ajuste do Modelo: Uma Visão Conceitual

O ajuste da regressão linear simples geralmente emprega o método dos mínimos quadrados, uma abordagem intuitiva que visa reduzir ao máximo as distâncias verticais entre os pontos de dados e a linha proposta. Em termos simples, calcula-se a soma das diferenças ao quadrado (para enfatizar erros maiores) e seleciona-se a linha que minimize esse total. Esse processo resulta em estimativas para a interceptação e a inclinação, que podem ser interpretadas diretamente.

Por exemplo, em um estudo sobre o impacto do tempo de estudo na nota de um exame, a inclinação poderia indicar que cada hora adicional de estudo aumenta a nota em uma certa quantidade de pontos. A interceptação representaria a nota esperada sem estudo algum, embora nem sempre tenha interpretação prática. Softwares como *R*, *Python* (com bibliotecas como **statsmodels**) ou até planilhas como o Excel facilitam esse cálculo, democratizando o acesso à técnica.

Uma vez ajustado, o modelo deve ser avaliado quanto à sua qualidade. Medidas como o coeficiente de determinação (R^2) indicam a proporção da variabilidade na variável dependente explicada pela independente – um valor próximo de 1 sugere um ajuste forte, enquanto próximo de 0 indica pouca relação. No entanto, R^2 não é infalível; em amostras pequenas, pode superestimar a força da associação.

A significância estatística dos coeficientes também é testada, verificando se a relação observada é provável de ocorrer por acaso. Intervalos de confiança e valores-p ajudam nessa avaliação, fornecendo uma medida de incerteza. Em contextos acadêmicos, esses elementos são essenciais para discutir a robustez dos achados, evitando generalizações precipitadas.

Do ponto de vista matemático, R^2 representa a proporção da variabilidade total de Y explicada pelo modelo:

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} \tag{4.1}$$

No caso da regressão simples, $R^2 = r_{XY}^2$, onde r_{XY} é o coeficiente de correlação linear de Pearson.

5 Aplicações Práticas

A versatilidade da regressão linear simples a torna ubíqua. Na biologia, pode modelar o crescimento de plantas em função da quantidade de luz solar recebida, auxiliando experimentos agrícolas. Em finanças, analisa a relação entre o retorno de um investimento e o risco associado, orientando portfólios. Nas ciências sociais, explora como o tempo gasto em redes sociais afeta o bem-estar psicológico, contribuindo para debates sobre saúde mental.

Um exemplo clássico é o estudo de Galton sobre a altura de pais e filhos, que demonstrou a "regressão à média"– filhos de pais muito altos tendem a ser mais próximos da média populacional. Essa aplicação inicial pavimentou o caminho para avanços em genética e hereditariedade. Em engenharia, a técnica é usada para prever o desgaste de materiais com base no tempo de uso, otimizando manutenção preventiva.

6 Aplicações em Análise e Ciência de Dados

Em Análise de Dados, a RLS é usada para modelar relações simples entre variáveis, como prever *churn* de clientes com base em tempo de uso de um app, ou estimar o impacto de variáveis ambientais em áreas agrícolas. Em Ciência de Dados, serve como baseline para modelos mais complexos (ex.: machine learning), ajudando a identificar features iniciais em pipelines de previsão.

6.1 Estimação por Mínimos Quadrados Ordinários (MQO)

Usado em Análise de Dados para otimizar modelos preditivos, como estimar ROI de campanhas de marketing. Em Ciência de Dados, é a base para feature engineering em pipelines de Machine Learning:

```
from sklearn.linear_model import LinearRegression

# Dados ficticios: publicidade (X) vs. vendas (Y)
X = np.array([[100], [200], [300], [400], [500]])
Y = np.array([10, 20, 25, 35, 40])

model_sk = LinearRegression().fit(X, Y)
print(f'Intercepto: {model_sk.intercept_}, Coeficiente:
    {model_sk.coef_[0]}, R2: {model_sk.score(X, Y)}$')
```

Listing 1: Métricas em um modelo de regressão linear simples

Saída esperada: Intercepto = 3.000, Coeficiente = 0.075 e $R^2 \approx 0.9583333333333334$.