Common Pitfalls in Statistical Analysis: Linear Regression Analysis — Relatório e Síntese

Lucas Oliveira Antonetti[1] and Bruno de Oliveira Monchelato[1]

[1]Disciplina: Inteligência Artificial, Prof. Clayton Pereira

Resumo

Este trabalho analisa os conceitos de regressão linear e as armadilhas estatísticas comuns descritas por Aggarwal e Ranganathan (2017). Através de uma simulação computacional em Python, reproduzimos a relação entre Circunferência Média do Braço (CMB) e Índice de massa corporal (IMC) para discutir a validade das premissas do modelo. O estudo foca não apenas no ajuste da reta, mas na interpretação crítica do intercepto, na restrição da extrapolação de dados e na necessidade vital da análise de resíduos para validar o modelo.
Palavras Chave: Regressão linear, Armadilhas estatísticas, Análise de resíduos, Interpretação de coeficientes.

Introdução

A regressão linear é uma ótima ferramenta para prever valores de uma variável dependente a partir de uma independente. No entanto, sua aplicação mecânica sem verificação de premissas pode levar a conclusões errôneas, conforme alertam Aggarwal e Ranganathan (2017).

O objetivo deste trabalho é simular o cenário clínico citado pelos autores, a relação entre CMB e IMC, para evidenciar na prática as limitações do modelo linear. O foco será especificamente em três pontos críticos: a interpretação do intercepto, a homocedasticidade dos resíduos e os perigos da extrapolação.

Metodologia

Para demonstrar os conceitos, foram gerados dados sintéticos em Python seguindo os parâmetros estatísticos do estudo de referência:

- **Variável Independente (X):** CMB com distribuição normal ($\mu = 27, \sigma = 4$), simulando uma população de pacientes hospitalizados.
- **Variável Dependente (Y):** IMC calculado pela equação $Y = -0.042 + 0.972X + \epsilon$, onde ϵ é o erro aleatório.

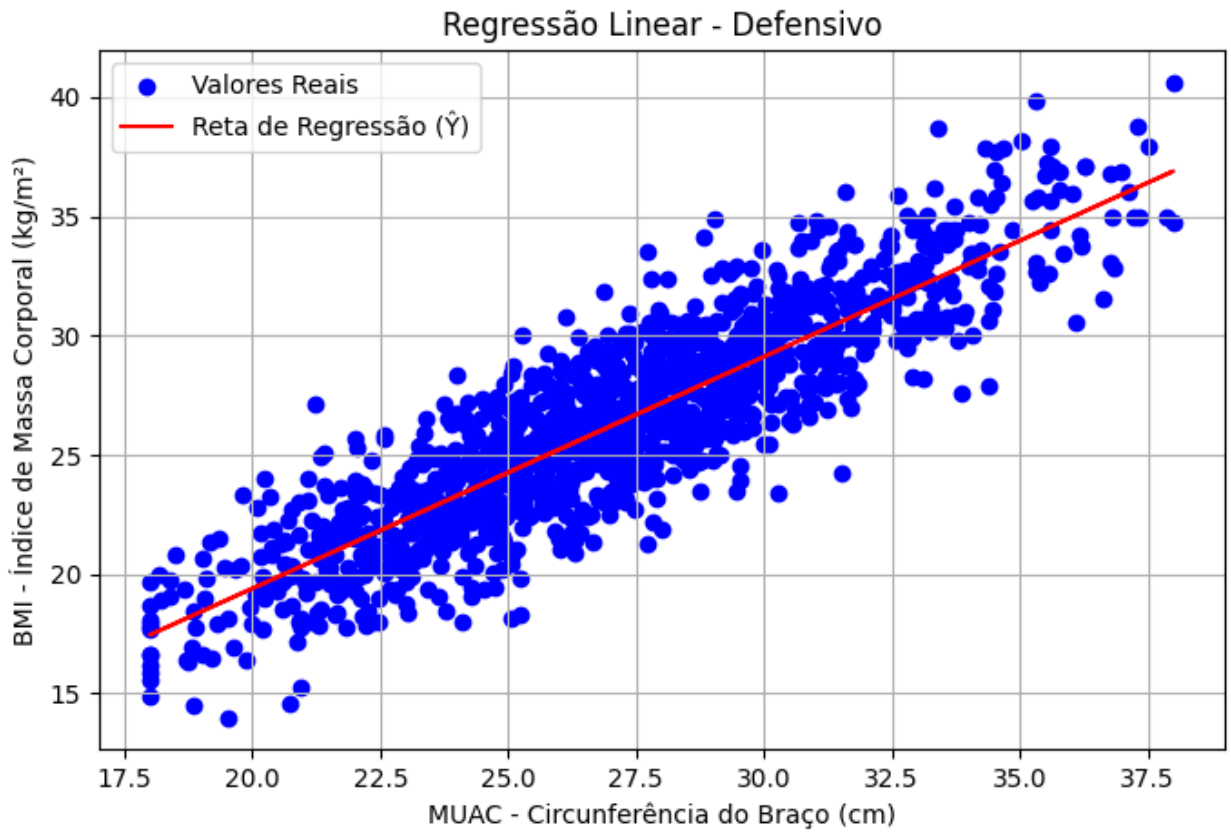

Figure 1: Regressão linear entre o Índice de Massa Corporal e a Circunferência de Braço dos pacientes

A análise consistiu em ajustar o modelo de mínimos quadrados e avaliar os diagnósticos do modelo sob a ótica das falhas comuns em pesquisa médica.

Resultados

O modelo ajustado resultou na equação $\hat{Y} \approx -0.042 + 0.972X$. O gráfico de dispersão (Fig. 1) mostra uma correlação positiva, e o R^2 obtido foi de 0.7752.

Métricas de ajuste

- MAE (erro absoluto médio): 1.6584;
- MSE (erro quadrático médio): 4.2912;
- RMSE (raiz do MSE): 2.0715;
- R^2: 0.7752;

- SSE (soma dos erros ao quadrado): 5891.7678;
- SST (soma total dos quadrados): 26208.1397;
- Média de Y ($\bar{Y}$): 26.4292.

Discussão

A Falácia do Intercepto

O modelo gerou um intercepto próximo de zero ou negativo. O artigo de referência destaca este como um erro de interpretação clássico: matematicamente, o intercepto é necessário para a reta, mas biologicamente, uma circunferência de braço (CMB) de 0 cm é impossível. Logo, interpretar esse coeficiente como o IMC base do paciente é um erro conceitual que a simulação confirma.

Extrapolação de Dados

A regressão é válida apenas dentro do intervalo dos dados observados, no caso, de 18 a 38 cm na simulação. Tentar prever o IMC de uma criança com CMB de 10 cm usando esta mesma equação seria uma armadilha, pois a relação linear pode não se sustentar fora da amostra original.

Análise de Resíduos

Para validar se o modelo linear é adequado, analisamos os resíduos (Fig. 2). A distribuição aleatória dos pontos em torno do zero confirma a homocedasticidade. Se houvesse um padrão, como por exemplo, um formato de cone, seria possível identificar a armadilha de usar um modelo linear para dados heterocedásticos, invalidando os testes de significância.

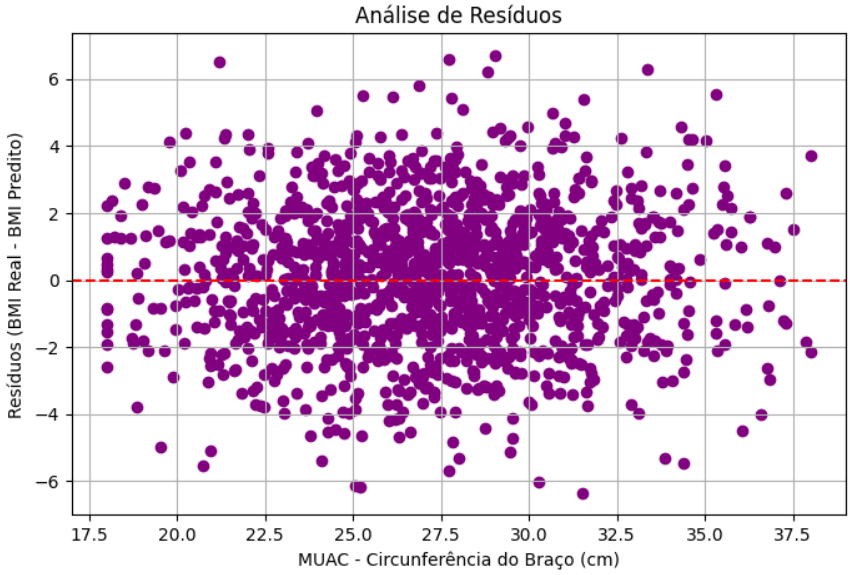

Figure 2: Análise dos Resíduos

Conclusão

Ao replicar o estudo de Aggarwal e Ranganathan, é possível confirmar que a regressão linear vai além de somente encontrar uma reta de ajuste. O sucesso da análise depende da capacidade do pesquisador de evitar armadilhas, como não interpretar o intercepto literalmente quando $X = 0$ não existe na realidade, e jamais confiar no R^2 sem inspecionar os resíduos. A simulação em Python provou-se eficaz para visualizar esses conceitos teóricos.

O modelo atual pode ser útil como ilustração e como ajuste inicial, mas não é suficientemente robusto para conclusões definitivas, procedimentos de diagnóstico e correções sugeridas devem ser aplicados e relatados para garantir transparência e reprodutibilidade nas pesquisas.

References

Aggarwal, Rakesh, and Priya Ranganathan. 2017. “Common pitfalls in statistical analysis: Linear regression analysis.” *Perspectives in clinical research* 8 (2): 100–102.