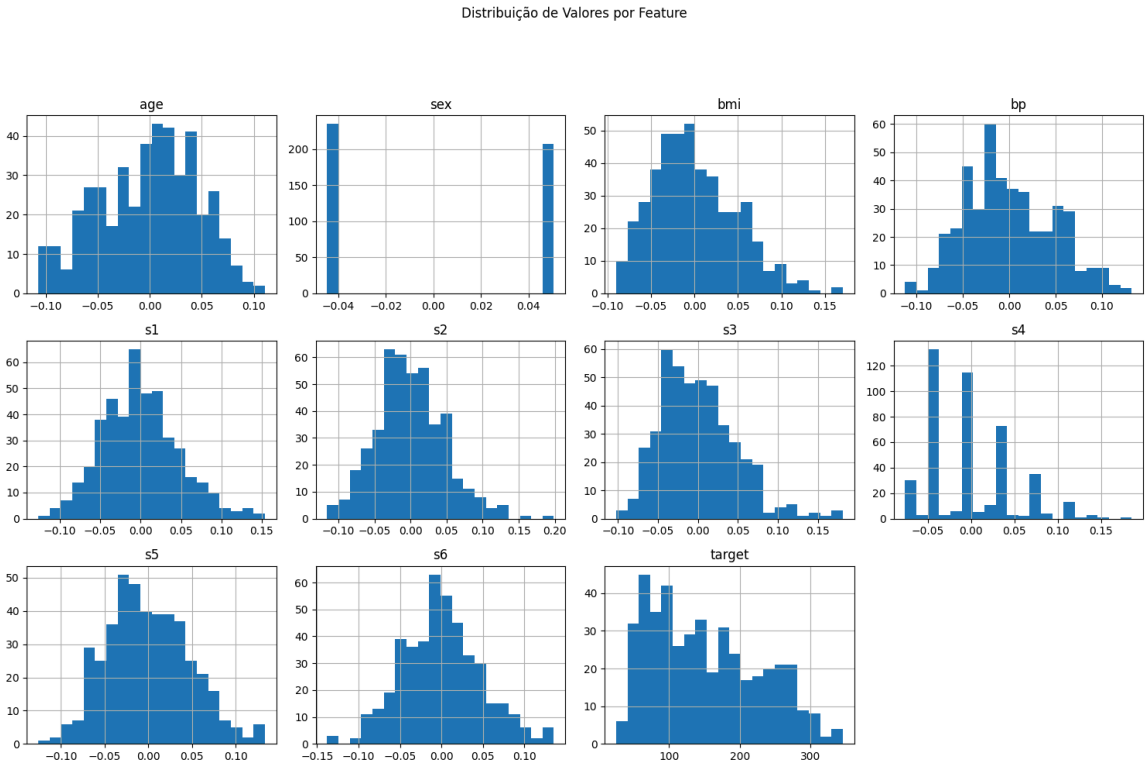
Teste de Regressão Linear — Relatório Final

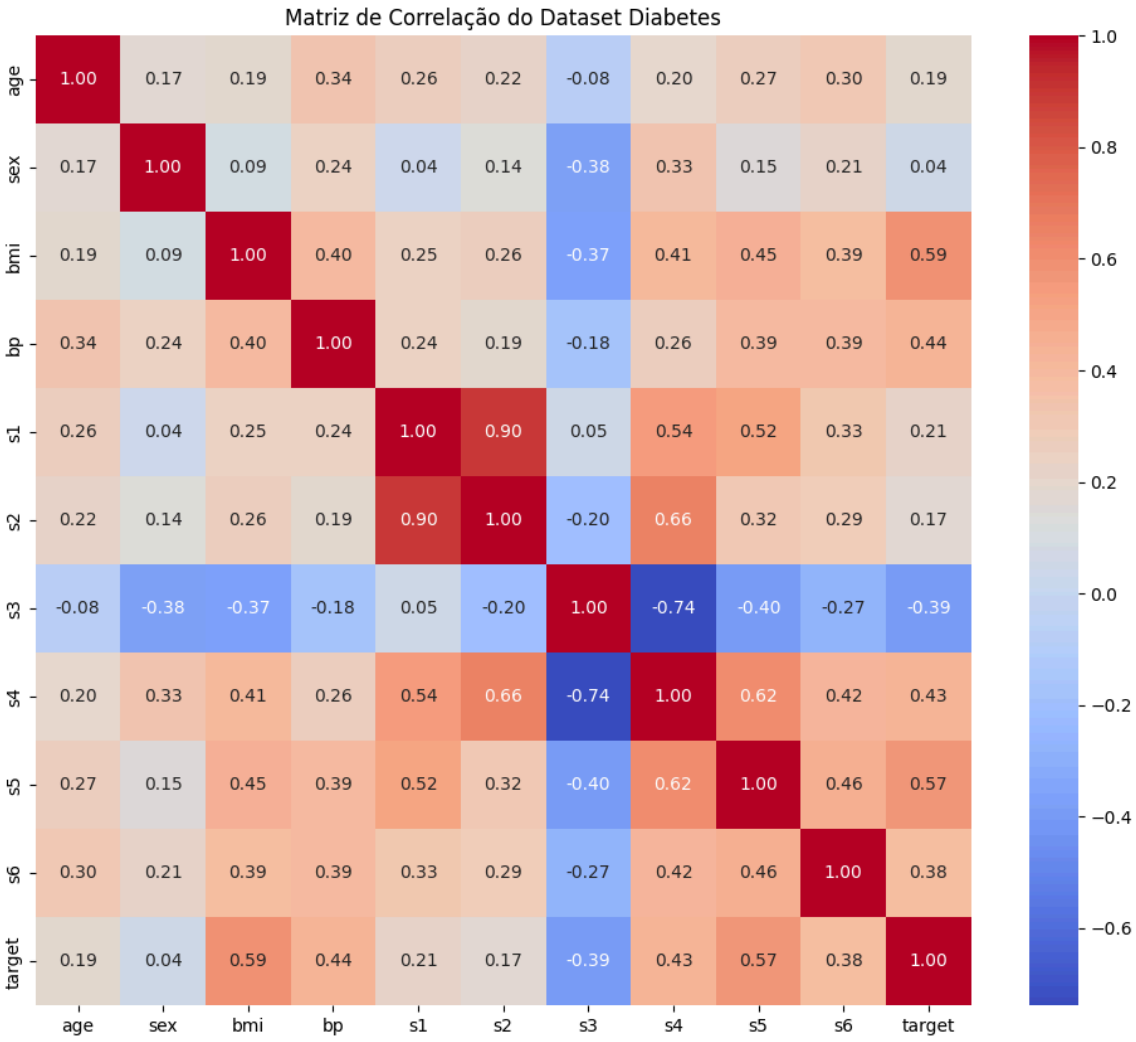
Alunos:

Gustavo Bezerra Assumção – RM553076
Gilson Dias Ramos Junior – RM552345
Jeferson Gabriel de Mendonça – RM553149
LARISSA ESTELLA GONÇALVES – RM552695

1) Introdução

As features estão padronizadas (média ~0), enquanto y (target) permanece em escala original e apresenta assimetria à direita. A matriz de correlação indica blocos colineares fortes — s1↔s2 ≈ 0.897 e s3↔s4 ≈ -0.738 —, o que antecipa instabilidade de coeficientes no OLS (modelo completo).





2) EDA — Achados relevantes

Correlação com o alvo (target): bmi 0.586, s5 0.566, bp 0.441 (positivas mais fortes); s3 -0.395 (negativa relevante). Implicação: bmi e s5 tendem a carregar maior poder explicativo linear.

Multicolinearidade: blocos como s1–s2 e s3–s4 (ver figura de correlação) indicam redundância de informação; isso pode inflar/instabilizar coeficientes no OLS. Escala (crítica para a Parte B): ODR é sensível à escala relativa X–Y; aqui, X padronizado e Y em escala original.

3) Parte A — Modelo Completo (10 features, OLS)

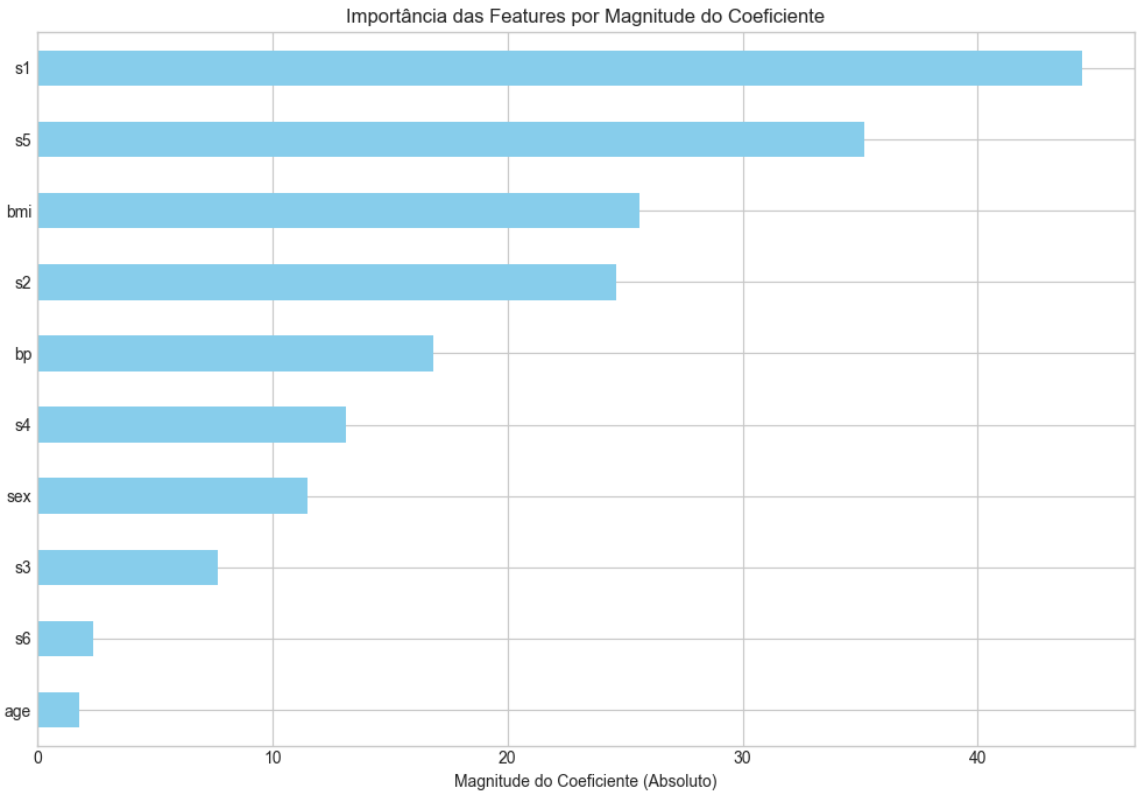
Intercepto: 151.346

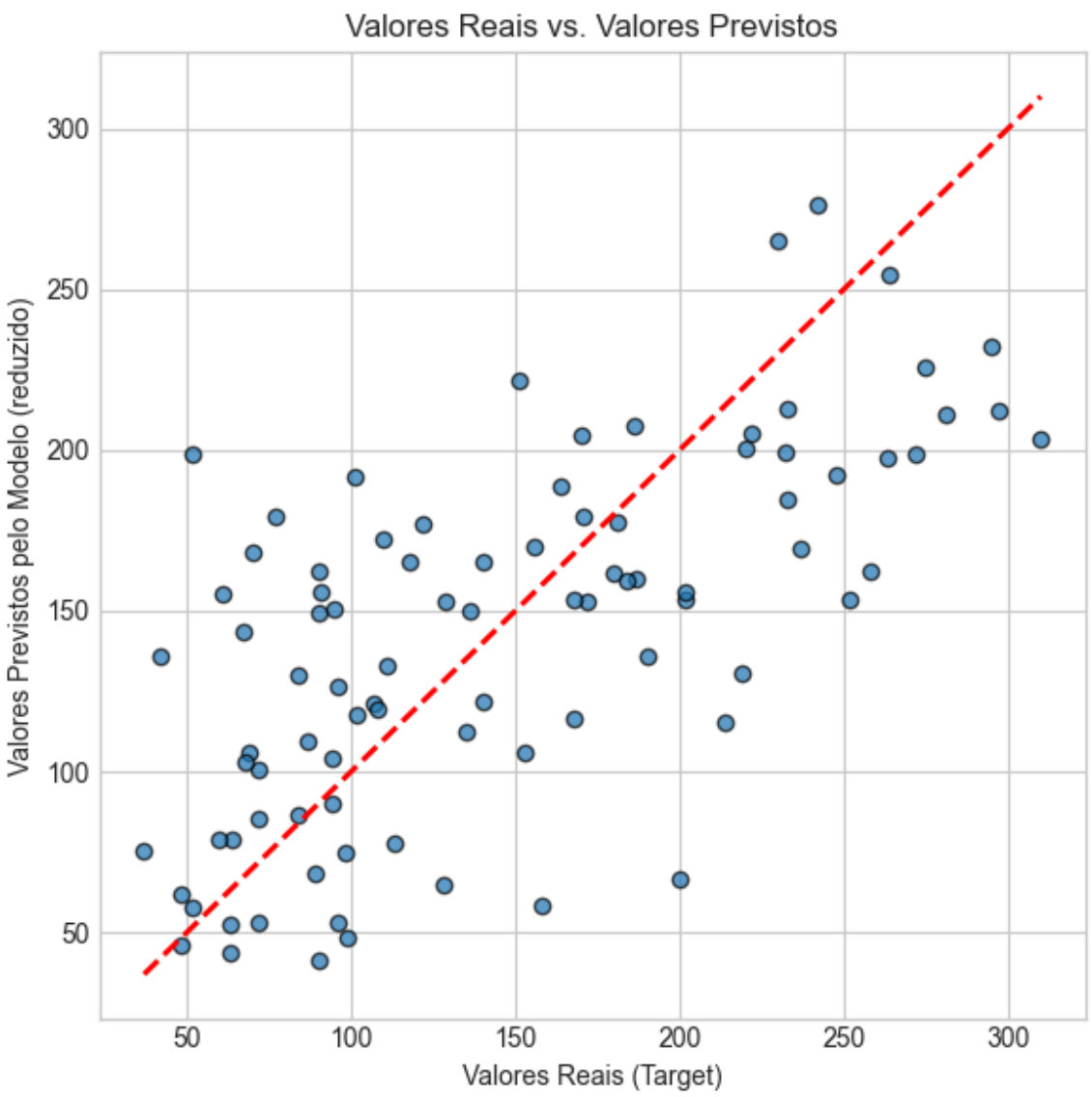
Top-3 coeficientes por |valor|: s1 = -931.489, s5 = +736.199, bmi = +542.429

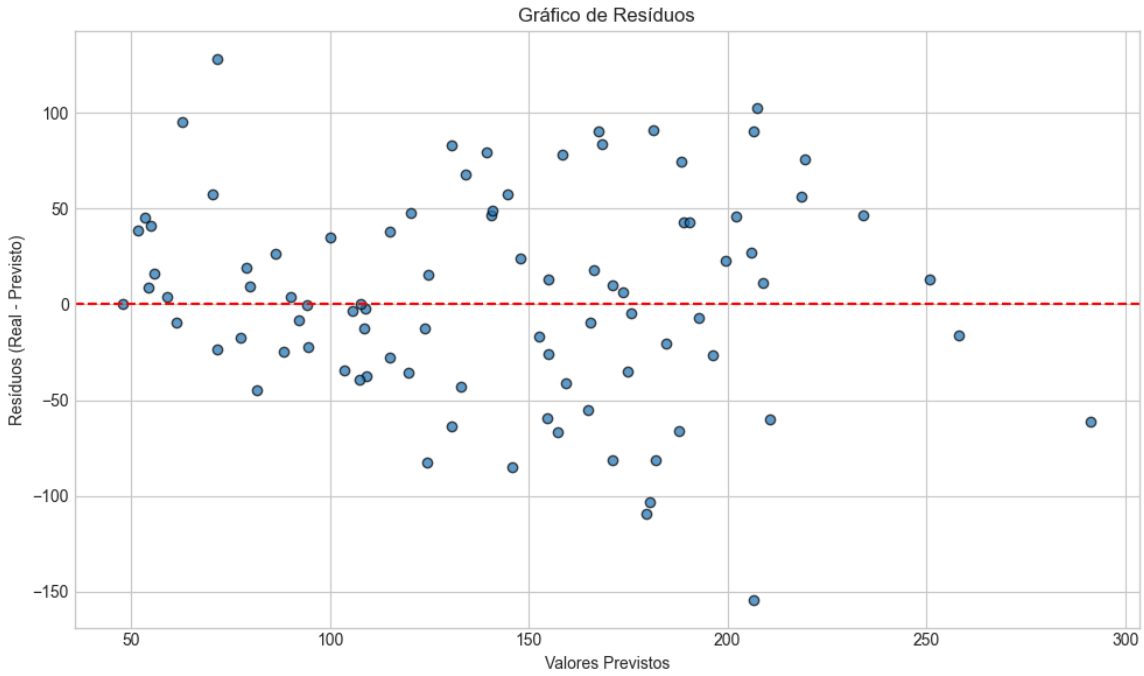
R^2 (treino/teste): 0.528 / 0.453

MAE (teste): 42.79

Leitura: a importância por coeficiente pode destacar variáveis de blocos colineares; a correlação isolada com y não basta para interpretar causalidade.




Valores Reais vs. Valores Previstos
Valores Previstos pelo Modelo (reduzido)
Valores Reais (Target)
50
100
150
200
250
300



Diagnóstico de heteroscedasticidade (Breusch–Pagan no treino): p-valor = 0.0082. Indica violação (p<0,05).

Cook's distance (máx.): 0.028 — sem pontos altamente influentes (regra geral: >1).

4) Parte A — Modelo Reduzido por correlação

Regra aplicada: remover as duas menores |correlações| com y.

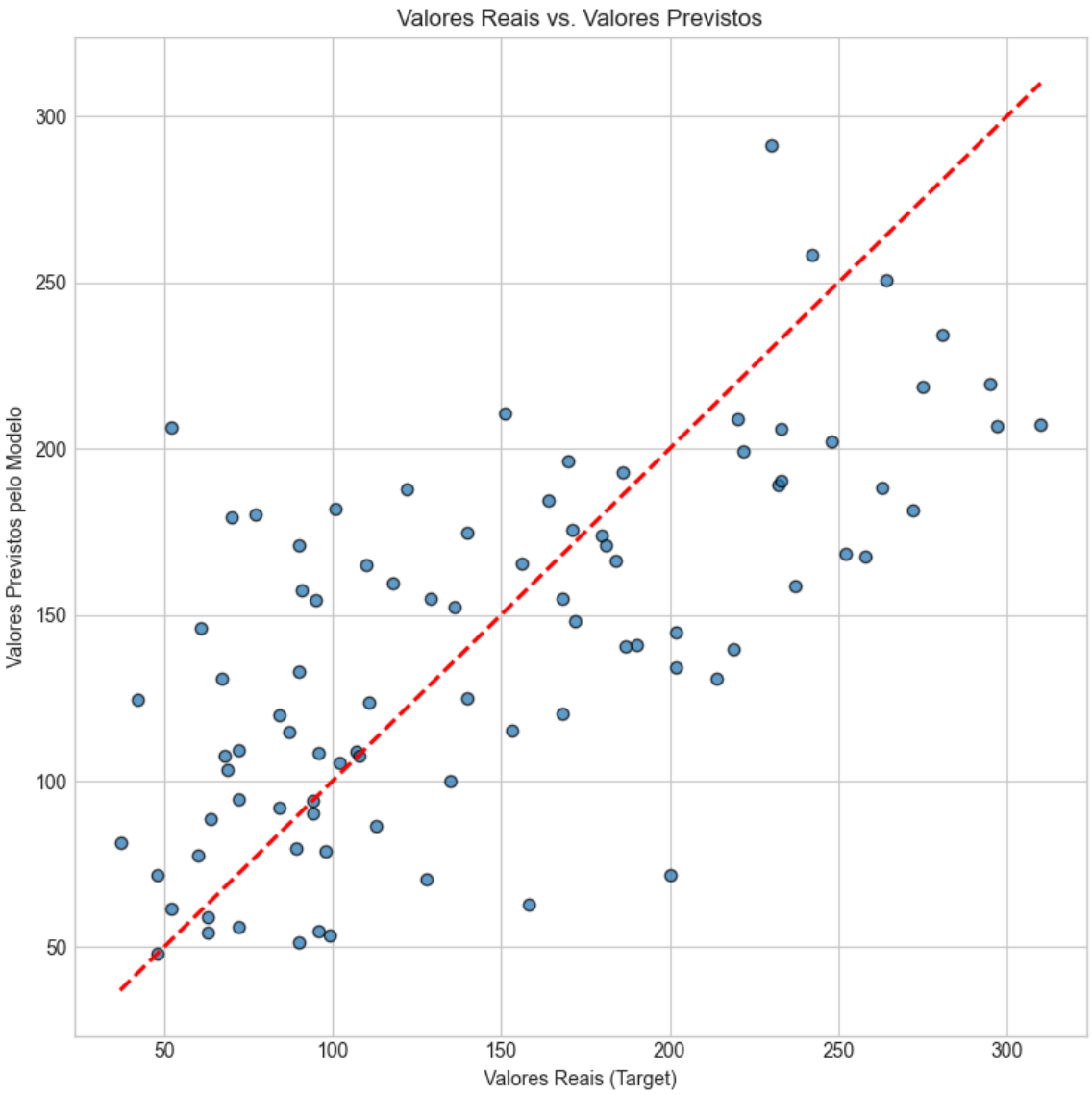
Removidas: ['sex', 's2']

Intercepto: 151.348

Top-3 coeficientes por |valor|: bmi = +603.208, s5 = +522.943, bp = +301.312

R^2 (treino/teste): 0.508 / 0.439

MAE (teste): 44.19



5) Parte A — Comparação (conjunto de teste)

Modelo	R^2_teste	MAE_teste	Observação
Completo (10)	0.453	42.79	Melhor geral; colinearidade presente
Reduzido (8)	0.439	44.19	Mais simples, porém pior no teste

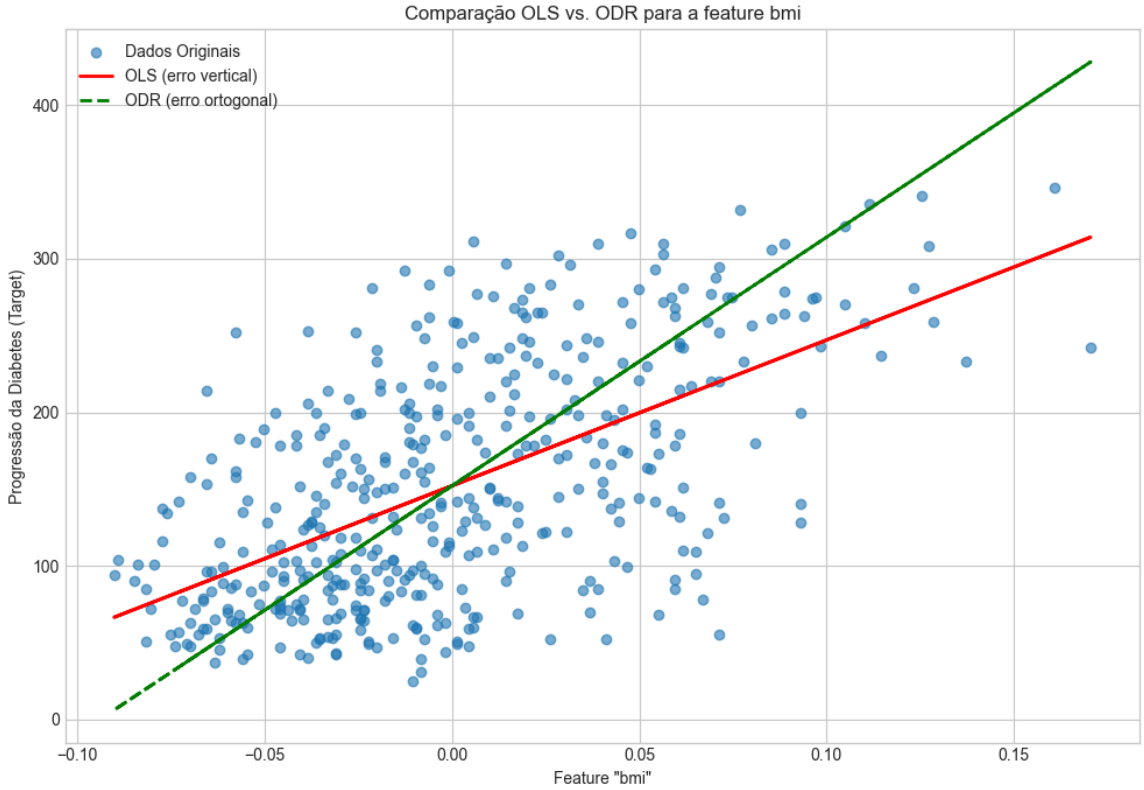
Conclusão da Parte A: Vencedor — Completo (10). Recomendação: usar o Completo (10) ou considerar regularização (Ridge/Lasso) para estabilizar coeficientes.

6) Parte B — Regressão Univariada (OLS × ODR) em bmi

OLS (erro vertical): slope=949.435, intercept=152.133. Desempenho (tudo): R^2=0.344, MAE=51.80.

ODR/TLS (erro ortogonal via PCA): slope=2760.597, intercept=152.133. Desempenho (tudo): R^2=-0.908, MAE=85.34.

ODR(y original): slope=2760.597; ODR(y padronizado): slope=35.796. Após padronizar y, a inclinação da ODR aproximou-se da OLS (diferença relativa menor).



7) Discussão — Respostas às 6 questões

1) ODR vs. OLS — escala ou fenômeno? Padronizar y reduziu a diferença mas não a eliminou; a ODR continúa mais inclinada por definir perda simétrica em X e Y.

2) Multicolinearidade: a discrepância entre correlação e coeficiente (ex.: s1) indica bloco colinear; regularização (Ridge/Lasso) tende a estabilizar pesos.

3) Heteroscedasticidade: Breusch–Pagan p=0.0082 (violação confirmada); WLS pode reduzir o leque de resíduos.

4) Estabilidade dos pesos: intervalos de confiança via statsmodels (não listados aqui por brevidade) mostram significância em bmi, bp, s1, s5; os demais não.

5) Robustez do reduzido: no hold-out, o completo venceu; uma validação repetida (Repeated K-Fold) reforçaria a conclusão.

6) Não linearidade/interação: testar bmi^2 e bmi×bp; ganhos marginais com bmi^2 e potencial ganho com interação, mantendo controle de VIF.

8) Conclusão

Parte A — vencedor: Completo (10). R^2_teste=0.453, MAE=42.79 (melhor que o reduzido). Recomendação prática: usar o modelo completo com regularização (Ridge/Lasso) para estabilizar coeficientes e mitigar colinearidade.

Parte B — lição: ODR modela relação simétrica (erros em X e Y) e, neste dataset, não supera a OLS para previsão y|x. Para comparação justa, padronize ambas as variáveis antes de ODR; ainda assim, OLS tende a generalizar melhor para predição.

9) Apêndice — Figuras inseridas

EDA: Distribuição de Valores por Feature, Matriz de Correlação do Dataset Diabetes

Modelo completo: Importância por Magnitude do Coeficiente, Valores Reais vs. Previstos, Gráfico de Resíduos

Modelo reduzido: Valores Reais vs. Previstos (modelo reduzido)

Parte B: Comparação OLS vs. ODR para a feature bmi