Seleção de modelos de regressão linear múltipla

NBR-14.653/2

Palestrante

Engenheiro Antonio Pelli Neto

Mestre em Inteligência Computacional

Pelli Sistemas Engenharia

Roteiro

Introdução à Regressão Linear – NBR 144.653-2

Problema da seleção de variáveis

Métodos: Forward, Backward, Stepwise

Critérios: R^2, R^2 ajustado, AIC, BIC, Cp

Teste F e ANOVA

Pressupostos e testes

Diagnóstico de resíduos e influência

Multicolinearidade (VIF)

Validação cruzada

Recomendações práticas

Introdução à Regressão Linear Múltipla

A regressão linear múltipla é uma técnica estatística que permite analisar a relação entre uma variável dependente (Y) e múltiplas variáveis independentes (X).

$$Y = X\beta + \varepsilon, \text{ com } E(\varepsilon)=0, Var(\varepsilon)=\sigma^2 I$$

Onde:

- Y é o vetor da variável resposta
- X é a matriz de variáveis preditoras
- β é o vetor de parâmetros a estimar
- ε são os erros aleatórios

Anexo A

A.3 Testes de Significância

A.3.1 O nível de significância máximo admitido nos demais testes estatísticos (aqueles não citados na Tabela 1) não deve ser superior a 10%.
A.3.2 A significância de subconjuntos de parâmetros, quando pertinente, pode ser testada pela análise da variância por partes.
A.3.3 Os níveis de significância utilizados nos testes citados em A.3 serão compatíveis com a especificação da avaliação.

A qualidade do modelo depende da seleção adequada de variáveis e da verificação dos pressupostos estatísticos.

Por que selecionar variáveis?

Desafios:

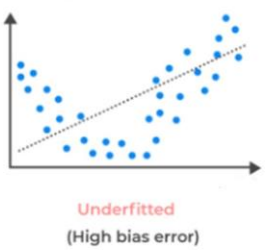
- Overfitting vs underfitting (equilíbrio viés-variância)
- Interpretabilidade e custo de medição
- Multicolinearidade e estabilidade dos coeficientes

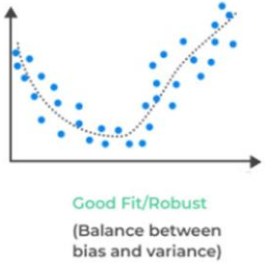
Objetivos conflitantes:

- Maximizar informação explicativa
- Minimizar número de variáveis (simplicidade e generalização)

Estratégias:

- Critérios de informação (AIC, BIC, R^2 ajustado)
- Validação cruzada para avaliação de desempenho
- Conhecimento substantivo da área de aplicação





Forward Selection (Seleção Progressiva)

Ideia principal: Começar com modelo vazio e adicionar variáveis uma a uma.

Algoritmo:

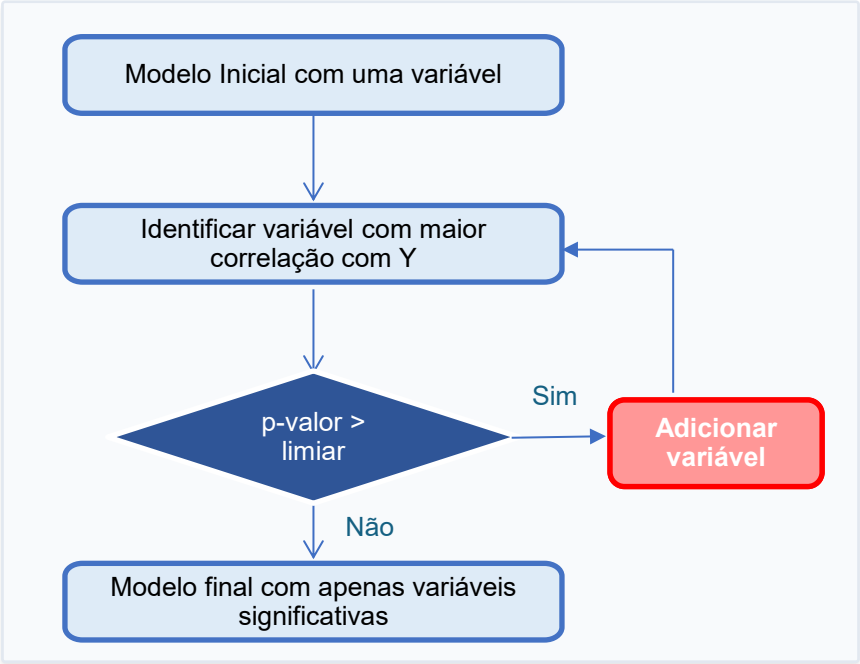
- Iniciar com modelo sem variáveis explicativas
- Adicionar a variável com maior correlação com Y
- Testar significância (teste F ou p-valor)
- Adicionar a próxima variável com maior correlação parcial
- Parar quando nenhuma nova variável for significativa

Vantagens:

- Funciona quando número de variáveis > amostra
- Computacionalmente eficiente

Limitações:

- Não considera efeitos conjuntos entre variáveis
- Pode ignorar colinearidade entre preditores

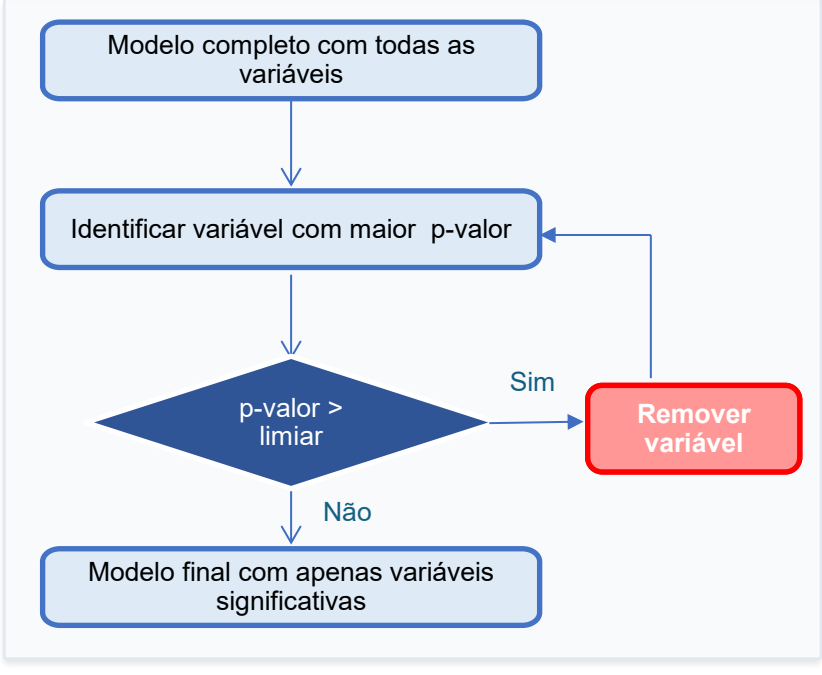


Backward Elimination (Eliminação Regressiva)

Ideia principal: Começar com modelo completo e remover variáveis uma a uma.

Algoritmo:

- Iniciar com modelo contendo todas as variáveis
- Identificar a variável menos significativa (maior p-valor)
- Remover se p-valor > limiar (Graus de Fundamentação / Precisão)
- Reajustar modelo e repetir o processo
- Parar quando todas as variáveis forem significativas



Vantagens:

- Considera efeitos conjuntos entre variáveis
- Melhor para lidar com multicolinearidade
- Indicado quando há teoria prévia para inclusão de variáveis

Limitações:

- Requer que número de observações > variáveis
- Computacionalmente mais intensivo para grandes conjuntos

Stepwise Regression (Método Misto)

Ideia principal: Combina adições (forward) e remoções (backward) em um processo iterativo.

Algoritmo:

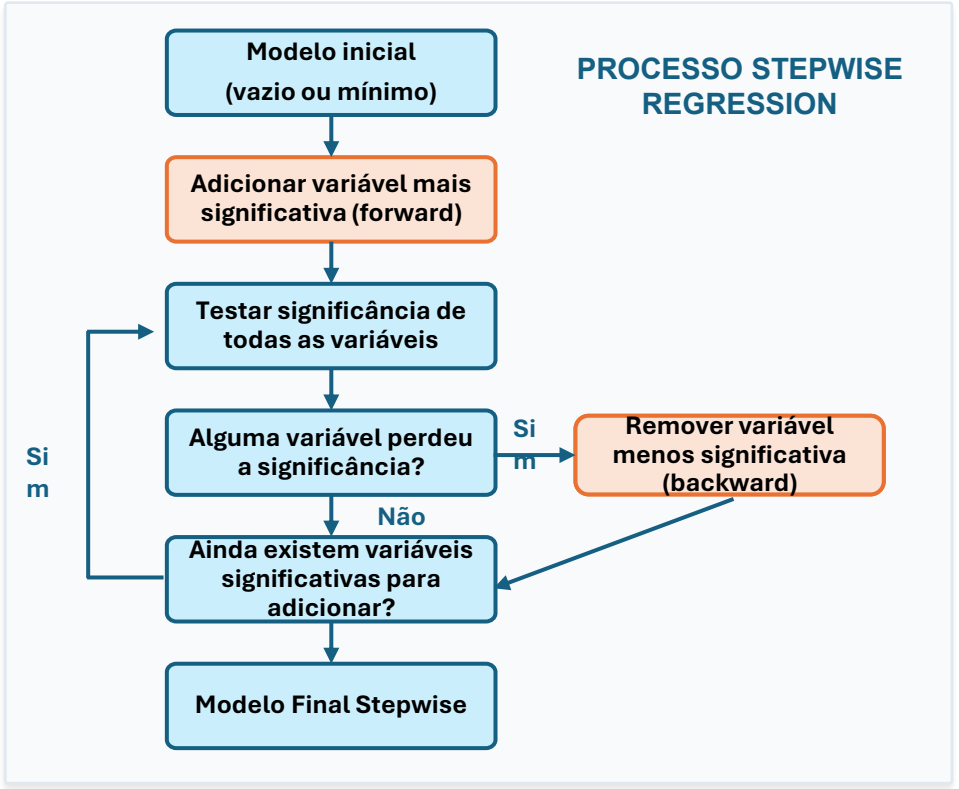
- Inicia com modelo vazio ou com variáveis de controle
- Adiciona variável mais significativa (Forward)
- Reavalia significância de todas as variáveis já no modelo
- Remove variáveis que perderam significância (Backward)
- Repete ciclo até estabilidade (nenhuma adição/remoção)

Vantagens:

- Avaliação dinâmica das variáveis já incluídas
- Mais eficiente que "best subset" para muitas variáveis

Cuidados:

- Instabilidade na seleção de variáveis
- Coeficientes e p-valores enviesados
- Não substitui conhecimento teórico do domínio



R^2 e R^2 Ajustado

Coeficiente de Determinação (R^2)

$$R^2 = 1 - SQRes / SQT$$

SQRes = Soma dos Quadrados dos Resíduos

SQT = Soma Total dos Quadrados

Coeficiente de Determinação Ajustado

$$R^2Ajustado = 1 - [(n-1)/(n-p)] \times (1-R^2)$$

n = número de observações

p = número de parâmetros (incluindo intercepto)

> Importante: Enquanto R^2 sempre aumenta (ou permanece igual) ao adicionar variáveis, o R^2Ajustado penaliza a inclusão de variáveis não significativas.

Interpretação:

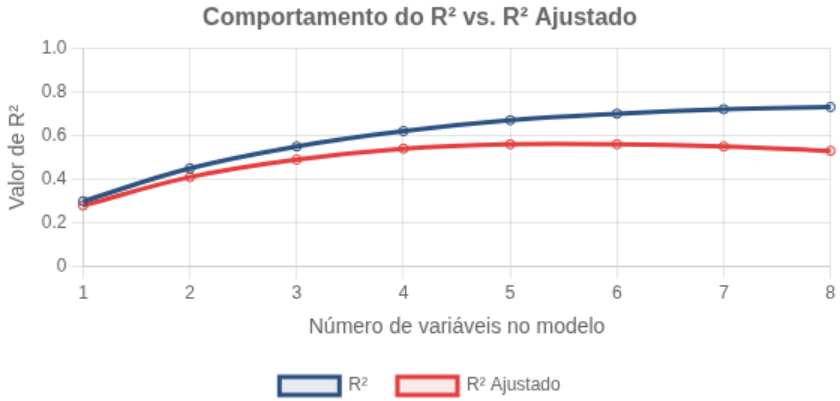
- R^2 mede a proporção da variância explicada pelo modelo
- R^2 varia de 0 (péssimo ajuste) a 1 (ajuste perfeito)

Uso na seleção de modelos:

- Compare modelos com as mesmas variáveis dependentes
- Quando comparar modelos com diferentes números de preditores, prefira R^2Ajustado
- R^2Ajustado pode diminuir se variáveis irrelevantes forem adicionadas

Limitações:

- Não mede validade preditiva fora da amostra
- Sensível a outliers e observações influentes
- Afetado por problemas de multicolinearidade



AIC, BIC e Cp de Mallows

Critérios de informação: Balanceiam a qualidade do ajuste com o número de parâmetros.

AIC (Akaike Information Criterion):

$$AIC = -2\log(L) + 2p$$
$$AIC = n\log(SQRes/n) + 2p$$

BIC (Bayesian Information Criterion):

$$BIC = -2\log(L) + p\cdot\log(n)$$
$$BIC = n\log(SQRes/n) + p\cdot\log(n)$$

Cp de Mallows:

$$Cp = SQRes/\hat{\sigma}^2 - (n - 2p)$$

Objetivo:

- Minimizar AIC e BIC
- Cp próximo a p indica modelo não enviesado

Comparação entre critérios:

Critério	Penalização	Preferência	Amostra
AIC	2p	Moderada	Qualquer tamanho
BIC	p·log(n)	Alta	Grandes (n>100)
Cp	2p	Moderada	Qualquer tamanho
R^2 Ajustado	(n-1)/(n-p)	Baixa	Qualquer tamanho

Principais diferenças:

- BIC penaliza mais modelos complexos
- BIC favorece modelos mais simples que AIC
- Cp considera explicitamente o viés do modelo
- Consistência assintótica: BIC > AIC

AIC modificado com Matriz Jacobiana

Problema: AIC padrão não permite comparar modelos com diferentes transformações na variável resposta.

Por quê? Transformações alteram a escala da verossimilhança, tornando AICs incomparáveis.

Solução: Ajuste Jacobiano

$$AIC_{ajustado} = AIC_{original} + 2 \times \sum_{i=1}^{n} \log|J_i|$$

Onde J_i é a derivada da transformação em relação a y_i

Princípio: Retorna à escala original dos dados, permitindo comparação justa entre:

- Modelos com variável resposta sem transformação
- Modelos com diferentes transformações (log, razão, Box-Cox)
- Modelos com transformações dependentes de variáveis independentes

Quando usar:

- Seleção de modelo + seleção de transformação simultâneas
- Comparação entre modelos com diferentes escalas

Exemplos de ajustes Jacobiano:

Transformação	Jacobiano	Ajuste ao AIC
$\log(Y)$	$1/Y_i$	$+2 \times \sum \log(Y_i)$
$Y/X1$	$1/X1,i$	$+2 \times \sum \log(X1,i)$
$\sqrt{Y}$	$1/(2\sqrt{Y_i})$	$+2 \times \sum \log(2\sqrt{Y_i})$
Box-Cox (λ)	$(Y_i+c)^{\lambda-1}$	$+2 \times \sum (\lambda-1)\log(Y_i+c)$

Exemplo aplicado:

Modelo 1: $Y \sim X1 + X2$ (AIC = 157,3)

Modelo 2: $\log(Y) \sim X1 + X2$ (AIC = 42,8)

AIC ajustado do Modelo 2: $42,8 + 2 \times \sum \log(Y_i) = 148,5$

Conclusão: Modelo 2 é melhor (148,5 < 157,3)

Importante: Sem ajuste Jacobiano, estaríamos erroneamente comparando valores em escalas diferentes, levando a conclusões equivocadas sobre qual modelo é superior.

Teste F e ANOVA do Modelo

Hipóteses: H0: $\beta 1 = \beta 2 = \ldots = \beta p\text{-}1 = 0$ vs H1: pelo menos um $\beta j \neq 0$

O teste F avalia a significância conjunta de todos os parâmetros, determinando se o modelo como um todo explica a variabilidade da resposta.

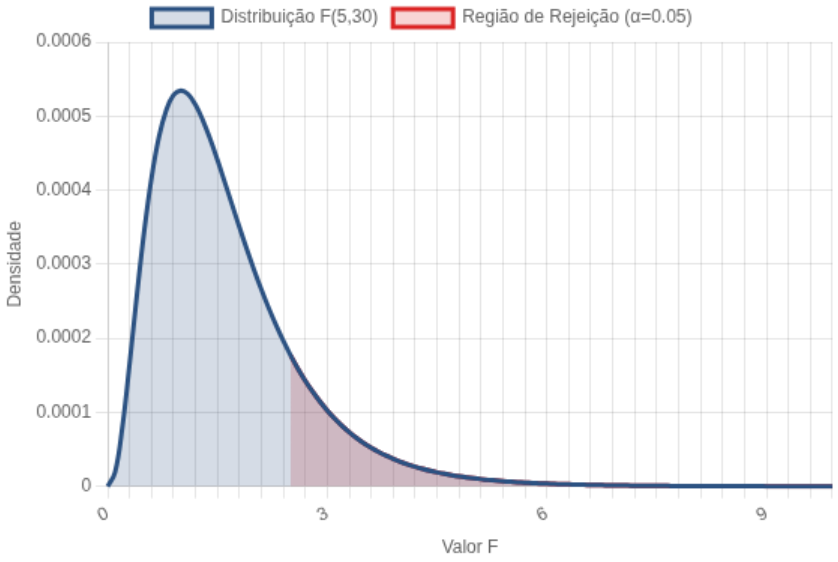
Tabela ANOVA

Fonte de Variação	GL	SQ	QM	F	p-valor
Regressão	k	SQReg	SQReg/k	QMReg/QMRes	P(F > Fobs)
Resíduo	n-k-1	SQRes	SQRes/(n-k-1)	-	-
Total	n-1	SQTotal	-	-	-

Decisão: Rejeitar H0 se Fobs > Fcrítico ou p-valor < α

Interpretação: Um F significativo indica que pelo menos uma variável independente contribui significativamente para o modelo.

Distribuição F



Pressupostos do Modelo e Consequências

1. Linearidade

Relação linear entre preditores e variável resposta

Consequência} Previsões enviesadas[subestimação ou superestimação

Correção: Transformações (log, Box-Cox).

2. Normalidade dos Resíduos

Erros aleatórios com distribuição normal

Consequência: Testes t e F inválidos; ICs imprecisos

Correção: Transformação de variáveis; bootstrap

3. Homocedasticidade

Variância constante dos erros

Consequência} Estimadores ineÃcientes-testes estatísticos inexatos

Correção: erro-padrão robusto (Park, White)

4. Independência

Ausência de autocorrelação nos resíduos

Consequência: Subestimação dos erros-padrão; inflação do R^2

Correção: erro-padrão robusto

5. Ausência de Multicolinearidade

Preditores não altamente correlacionados

Consequência} CoeÃcientes instáveis-erros?padrão inÀacionados

Correção: remover/combinar variáveis

6. Ausência de Pontos Influentes

Observações não devem ter influência desproporcional

Consequência: Distorção nos coeficientes; resultados não generalizáveis

Correção: Regressão robusta; análise de influência; validação cruzada

Teste RESET de Ramsey (Especificação do Modelo)

Objetivo: Detectar erro de especificação funcional

Como funciona: Verifica se combinações não lineares das variáveis explicativas ajudam a explicar a variável resposta.

Procedimento:

Ajuste o modelo original: $Y = X\beta + \varepsilon$

Obtenha os valores ajustados ($\hat{y}$)

Inclua potências de $\hat{y}$ no modelo: $\hat{y}^2$, $\hat{y}^3$, etc.

Teste se os coeficientes dessas potências são zero

Modelo auxiliar:

$$Y = X\beta + \gamma_1\hat{y}^2 + \gamma_2\hat{y}^3 + \ldots + \gamma_{k-1}\hat{y}^k + \varepsilon$$

Hipóteses:

H_0: $\gamma_1 = \gamma_2 = \ldots = \gamma_{k-1} = 0$ (bem especificado)

H_1: pelo menos um $\gamma \neq 0$ (má especificação)

Nota: Rejeitar H_0 pode indicar variáveis omitidas, transformações necessárias (log, raiz) ou termos de interação faltantes.

Teste estatístico:

$$F = [(R^2\text{novo} - R^2\text{original})/q]/[(1 - R^2\text{novo})/(n - p - q)]$$

q = número de termos adicionados (k-1)

n = tamanho da amostra

p = número de parâmetros no modelo original

Interpretação:

p-valor < 0,05 → rejeita H_0 → má especificação

p-valor ≥ 0,05 → não rejeita H_0 → bem especificado

Quando usar:

Verificar linearidade da forma funcional

Após ajuste inicial do modelo

Antes de fazer inferências dos parâmetros

Limitações:

Indica problema, mas não sua solução específica

Sensível à presença de outliers

Transformação Box-Cox

Objetivo: Transformar variável resposta não-normal para:

- Aproximar da distribuição normal
- Estabilizar a variância (homocedasticidade)
- Melhorar linearidade nas relações

Definição matemática:

$T_\lambda(Y) = (Y^\lambda - 1)/\lambda$, se $\lambda \neq 0$
$T_\lambda(Y) = \log(Y)$, se $\lambda = 0$

Quando usar:

- Resíduos não-normais no diagnóstico
- Variância não constante (heterocedasticidade)
- Teste RESET indicando não-linearidade
- Falta de ajuste mesmo após incluir termos polinomiais

Importante:
Para comparar modelos com diferentes transformações, aplique ajuste Jacobiano ao AIC (slide anterior)

Transformações comuns:

Valor de λ	Transformação	Aplicação típica
-1	1/Y	Assimetria positiva forte
-0,5	$1/\sqrt{Y}$	Valores muito dispersos
0	log(Y)	Crescimento exponencial
0,5	$\sqrt{Y}$	Contagens, dados Poisson
1	Y	Já normal (sem transformação)
2	Y^2	Assimetria negativa moderada

Vantagens:

- Família paramétrica com interpretação clara
- Seleção objetiva por máxima verossimilhança
- Inclui transformações tradicionais como casos especiais

Cuidados:

- Requer Y > 0 (aplicar deslocamento se Y ≤ 0)
- Verificar normalidade após transformação
- Considerar interpretabilidade prática
- Para valores de λ muito distantes de 1, considerar outros modelos

Regressão Linear

Tabela 1 – Fundamentação

Item	Descrição	Grau		
		III	II	I
1	Caracterização do imóvel avaliando	Completa quanto a todas as variáveis analisadas	Completa quanto às variáveis utilizadas no modelo	Adoção de situação paradigma
...	...	...	...	...
5	Nível de significância (somatório do valor das duas caudas) máximo para a rejeição da hipótese nula de cada regressor (teste bicaudal)	10%	20%	30%
6	Nível de significância máximo admitido para a rejeição da hipótese nula do modelo através do teste F de Snedecor	1%	2%	5%

A.3 Testes de Significância

A.3.1 O nível de significância máximo admitido nos demais testes estatísticos (aqueles não citados na Tabela 1) não deve ser superior a 10%.

A.3.3 Os níveis de significância utilizados nos testes ciados em A.3 serão compatíveis com a especificação da avaliação.

Limiar – Testes de Significância

- Define-se matematicamente as hipóteses H_0 e H_1 mutuamente exclusivas e exaustivas.
- Constrói-se uma região de rejeição fixando a probabilidade do Erro Tipo I. A partir de região de rejeição será tomada uma decisão sobre H_0 e H_1
- Se a estatística observada pertencer a região de rejeição, rejeita-se a hipótese H_0 para o nível de significância prefixado.
- Se a estatística observada não pertencer a região de rejeição, aceita-se a hipótese H_0 para o nível de sindicância prefixado. Neste ultimo caso, por influencia dos argumentos de Fisher, diz-se
- atualmente que:

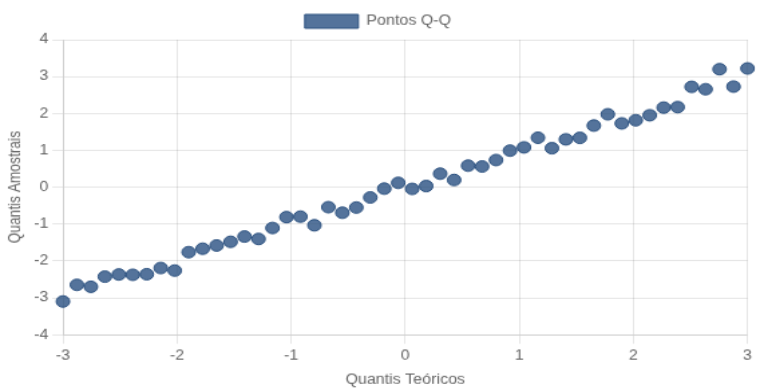
"não foi possível encontrar evidencias para rejeitar a hipótese nula".

Testes de Normalidade dos Resíduos

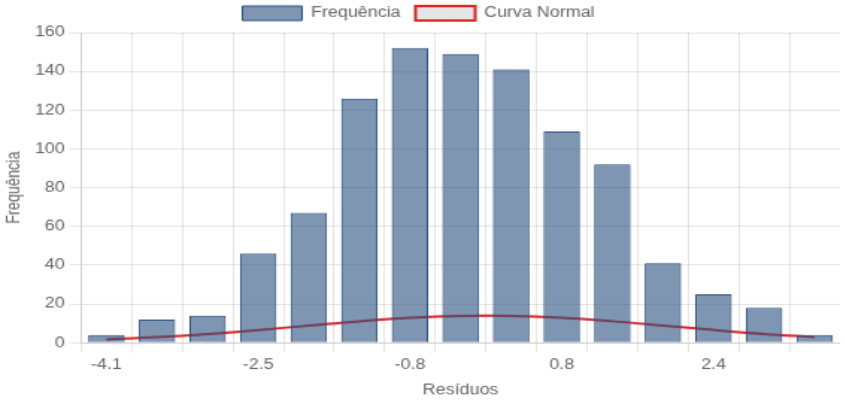
A normalidade dos resíduos é um pressuposto importante para inferência em modelos de regressão, afetando diretamente a validade dos testes t e F e dos intervalos de confiança.

- Shapiro-Wilk
- Kolmogorov-Smirnov
- Anderson-Darling
- Jarque-Bera

Gráfico Q-Q Plot



Histograma de Resíduos



Quando a normalidade é violada: Considerar transformações (Box-Cox), bootstrapping, ou novas abordagens

Interpretação: Pontos alinhados à reta indicam normalidade. Desvios sistemáticos sugerem assimetria ou caudas pesadas.

Interpretação: Distribuição simétrica em forma de sino sugere normalidade. Assimetrias ou múltiplos picos indicam violações.

Testes de Homocedasticidade

Homocedasticidade: Variância constante dos resíduos, essencial para inferência válida em regressão linear.

Principais Testes

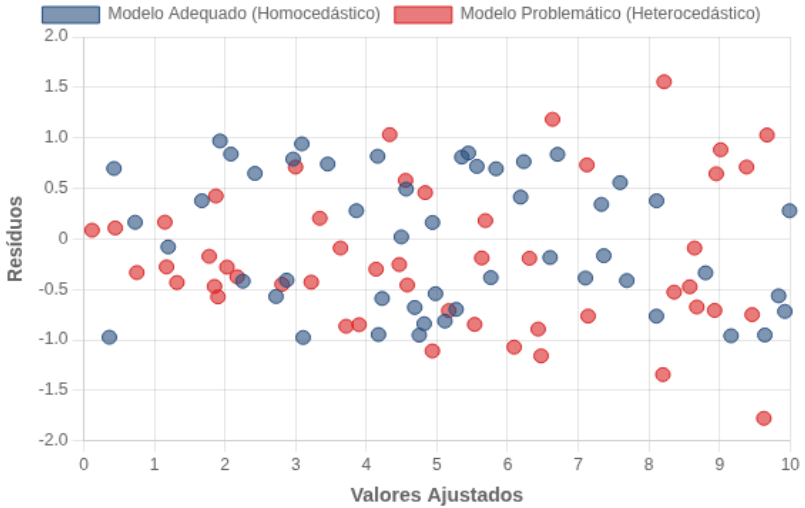
Teste	Descrição	Hipótese Nula
Breusch-Pagan	Verifica se a variância residual depende das variáveis independentes	Variância constante
Koenker	Versão mais robusta do BP, menos sensível a desvios da normalidade	Variância constante
Goldfeld-Quandt	Divide os dados em dois grupos e compara variâncias	Razão de variâncias = 1

Correções

Se detectada heterocedasticidade:

- Transformação da variável resposta (log, raiz quadrada)
- Mínimos quadrados ponderados (WLS)
- Erros-padrão robustos (Park ou White)

Diagnóstico Visual



Interpretação: Em um modelo adequado, os resíduos devem estar distribuídos aleatoriamente em torno de zero, sem padrões como "funil", "megafone" ou tendências sistemáticas.

Multicolinearidade e VIF

Multicolinearidade: Situação em que variáveis independentes são altamente correlacionadas entre si, gerando instabilidade nos coeficientes estimados.

Sintomas: Coeficientes com sinais inesperados; magnitudes irrealistas; erros-padrão inflacionados; testes t não significativos, mas F significativo.

Matriz de Correlação

Id	Variável	Transf.	VIF	Alias	x1	x2	x3	x4	x5	x6	x7	y
	Área Privativa	1/x	6,3084	x1	1	-0,4024	0,7763	-0,5260	-0,6713	-0,1217	-0,2480	-0,4927
	Local (Indice Fi...	ln(x)	3,1609	x2	-0,4024	1	0,0313	0,5073	0,7473	0,0516	0,1060	0,8760
	Quartos	1/x	4,5110	x3	0,7763	0,0313	1	-0,3554	-0,3277	-0,2523	-0,1821	-0,1713
	Vagas	x	1,8814	x4	-0,5260	0,5073	-0,3554	1	0,6572	0,2425	0,1243	0,6777
	Padrão	x	4,8078	x5	-0,6713	0,7473	-0,3277	0,6572	1	0,2864	0,0803	0,9062
	Idade	1/x	1,3597	x6	-0,1217	0,0516	-0,2523	0,2425	0,2864	1	-0,0786	0,2994
	Evento (Oferta...	x	1,1058	x7	-0,2480	0,1060	-0,1821	0,1243	0,0803	-0,0786	1	0,1231
	Valor/m²	y		y	-0,4927	0,8760	-0,1713	0,6777	0,9062	0,2994	0,1231	1

Fator de Inflação da Variância (VIF)

Variável	VIF	Interpretação
X_1 (Renda)	1.8	Sem problema
X_2 (Idade)	2.3	Sem problema
X_3 (Área)	4.7	Moderado
X_4 (Padrão)	9.2	Problemático
X_5 (Conservação)	12.5	Crítico

VIF = $1/(1-R^2_j)$, onde R^2_j é o R^2 da regressão da variável j sobre todas as outras.

Limites práticos: VIF > 5 (preocupante); VIF > 10 (problemático)

Soluções: Excluir/combinar variáveis; Padronização; PCA.

Diagnóstico de Resíduos e Pontos Influentes

Identificar observações que exercem influência desproporcional sobre as estimativas do modelo

Métricas de Diagnóstico

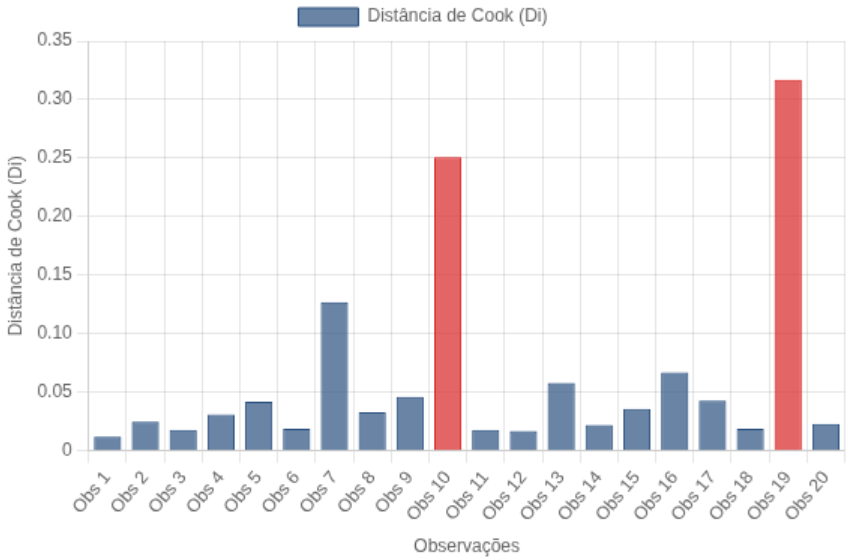
Resíduos studentizados: Padronizam os resíduos pela sua variância estimada. Valores |rse(i)| > 2 indicam potenciais outliers.

Leverage (hii): Mede a influência potencial de cada observação nos valores ajustados. Pontos com hii > 2p/n são considerados de alta alavancagem.

Distância de Cook (Di): Medida da influência global da observação i nos coeficientes estimados. Valores Di > 4/(n-p) merecem investigação.

Ações recomendadas: Não excluir observações automaticamente; verificar coleta de dados e erro de medição; considerar transformações ou modelos alternativos; reportar análises com e sem pontos influentes.

Gráfico de Distância de Cook



Validação Cruzada

Objetivo: estimar desempenho fora da amostra, evitar overfitting e comparar modelos de forma confiável.

Hold-out

Divisão simples dos dados:

- Treino: 70-80% dos dados
- Teste: 20-30% restantes
- Limitação: dependência da divisão escolhida

k-fold

Divisão em k subconjuntos:

- Divide dados em k partes (geralmente k=5 ou 10)
- Treina em k-1 partes e testa na restante
- Repete k vezes e calcula média das métricas

Variantes: Leave-One-Out (LOOCV), estratificado, bootstrap.

Boas práticas:

Pipeline completo para evitar vazamento de informação

Usar métricas adequadas (RMSE/MAE para regressão)

Estratificar os dados quando necessário

Escolher modelo por CV, mas reportar coeficientes do modelo final treinado em todos os dados

Recomendações Práticas

Fluxo de trabalho sugerido para análise de regressão:

1. Exploração e pré-processamento: análise descritiva, correlações, transformações e tratamento de dados
2. Especificação inicial: identificar possíveis termos e interações com base em teoria e visualização
3. Ajuste e diagnóstico completo: verificar linearidade, normalidade, homocedasticidade, independência
4. Tratamento de violações: transformações, erros-padrão robustos, modelos alternativos (GLM, etc.)
5. Seleção de variáveis: critérios objetivos (AIC/BIC/penalização) + validação cruzada
6. Verificação de estabilidade: bootstrap para coeficientes, análise de multicolinearidade
7. Relatório final: coeficientes, intervalos de confiança, métricas de desempenho e limitações

Observação: manter variáveis de controle teoricamente importantes mesmo quando não significativas estatisticamente. O conhecimento deve complementar a análise quantitativa.

50 ANOS DE HISTÓRIA E EXCELÊNCIA

XXIII COBREAP

CONGRESSO BRASILEIRO DE ENGENHARIA DE AVALIAÇÕES E PERÍCIAS

Obrigado pela atenção!

Eng. Antonio Pelli Neto

(31) 3466-1557

pelli@pellisistemas.com.br