PROCEDIMENTOS DO TIPO STEPWISE

MAE0350 – MODELOS DE REGRESSÃO I

Cristine Airi Kimura (13783096);

Gabriel Medeiros Cupini (13783116);

João Vitor Guandalini Vicente (15062582);

Olian Castanha Lima (13686261);

Vicenzo Carlim de Sousa (13686772).

STEPWISE

PASSO A PASSO

CONTEXTO: SELECIONAR A MELHOR EQUAÇÃO DE REGRESSÃO.

OBJETIVO: IDENTIFICAR O MENOR GRUPO DE VARIÁVEIS EXPLICATIVAS QUE SÃO "BOAS" DE ACORDO COM ALGUM **CRITÉRIO ESPECÍFICO.**


STEPWISE

X_1 X_2 X_3 X_4

Quantos modelos de regressão podem ser construídos a partir das quatro variáveis preditoras disponíveis?

- Primeiramente, existe **um** modelo sem variável explicativa alguma, isto é, $Y_i = \beta_0 + \epsilon_i$.
- Em seguida, temos **quatro** modelos com uma única variável explicativa (X_1, X_2, X_3, X_4).
- Em seguida, temos **seis** modelos com duas variáveis explicativas (X_1 e X_2; X_1 e X_3; X_1 e X_4; X_2 e X_3; X_2 e X_4; X_3 e X_4).
- Em seguida, temos **quatro** modelos com três variáveis explicativas (X_1, X_2 e X_3; X_1, X_2 e X_4; X_1, X_3 e X_4; X_2, X_3 e X_4).
- Por fim, temos **um** modelo com as quatro variáveis explicativas, isto é, $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$.

Resposta: *$2^4 = 16$ modelos.*

PROCEDIMENTO GERAL DA REGRESSÃO *STEPWISE*

"Escolhe-se uma equação contendo a melhor variável X e, depois, tentamos construir uma nova equação com adições subsequentes de X's, uma por vez, contanto que essas inserções valham a pena."

PROCEDIMENTO GERAL DA REGRESSÃO *STEPWISE*

*"Escolhe-se uma equação contendo a **melhor variável** X e, depois, tentamos construir uma nova equação com **adições subsequentes** de X's, uma por vez, contanto que essas inserções **valham a pena**."*

- A ordem da adição é determinada pelo teste *F*-parcial;
- O valor *F*-parcial mais alto é comparado com um valor "*F*-para-inserir" (selecionado ou padrão);
- Após a adição da variável, a equação é examinada.

PASSO A PASSO

1. Selecionar a variável independente Z_1 mais correlacionada com Y e encontrar a equação de regressão linear de primeira ordem $\hat{Y} = f(Z_1)$;
2. Checar se a variável Z_1 é significante;
   I. Se não for, abandonamos o modelo e adotamos $Y = \bar{Y}$;
   II. Caso contrário, procuramos pela segunda variável preditora para ser incluída na regressão.
3. Examinar os valores F-parciais de todas as variáveis preditoras que não estão na regressão. A variável Z_j com o maior valor (por exemplo, Z_2) é selecionada e uma segunda equação de regressão $\hat{Y} = f(Z_1, Z_2)$ é ajustada.
4. **Analisar**: verificar a significância global, observar a melhoria no R^2 e examinar os valores F-parciais para ambas as variáveis no modelo.
5. O menor desses dois F's parciais é comparado com um ponto percentual "*F-para-remover*" e a variável preditora correspondente é mantida ou não na equação dependendo se o teste é significativo ou não.

OBSERVAÇÕES SOBRE O PROCEDIMENTO

- O teste da variável preditora menos "útil" na equação atual é realizado em todas as etapas;
- A "melhor" variável explicativa candidata em uma fase anterior pode ser supérflua na fase posterior;
  - Para checar, utilizar o critério da *F*-parcial para cada variável na regressão;
  - Julgar a contribuição da variável "menos útil" naquele estágio.
    - Se a variável testada fornecer uma contribuição não significativa, ela é removida do modelo e é calculada a equação ajustada de regressão apropriada para as variáveis restantes no modelo.
- A "melhor" das variáveis (cujo valor *F*-parcial dado as preditoras na equação é o maior) é verificada no teste *F*-parcial de entrada;
  - Se for aprovado, ela será inserida e voltaremos a verificar todos os *F*-parciais em busca de variáveis;
  - Se falhar, uma nova remoção será tentada.
- Eventualmente, quando nenhuma variável na equação pode ser removida e a próxima melhor variável candidata não puder ser mantida, o processo termina.

EXEMPLO – DADOS DE HALD (1952, p. 647)

- Dados originalmente provenientes de "*Efeito da composição do cimento de Portland no calor liberado durante o endurecimento*", por H. Woods, H. H. Steinour e H. R. Starke, *Industrial and Engineering Chemistry*, 24, 1932, 1207-1214.
- As variáveis são:

X_1 = quantidade de aluminato tricálcico, 3 CaO · Al_2O_3.

X_2 = quantidade de silicato tricálcico, 3 CaO · SiO_2.

X_3 = quantidade de ferroaluminato tetracálcico, 4 CaO · Al_2O_3 · Fe_2O_3.

X_4 = quantidade de silicato dicálcico, 2 CaO · SiO_2.

Y (resposta) = calor liberado em calorias por grama de cimento.

- X_1, X_2, X_3 e X_4 foram mensuradas como porcentagem do peso dos clínqueres com os quais o cimento foi feito.

X_1	X_2	X_3	X_4	Y	Soma dos X's
7	26	6	60	78,5	99
1	29	15	52	74,3	97
11	56	8	20	104,3	95
11	31	8	47	87,6	97
7	52	6	33	95,9	98
11	55	9	22	109,2	97
3	71	17	6	102,7	97
1	31	22	44	72,5	98
2	54	18	22	93,1	96
21	47	4	26	115,9	98
1	40	23	34	83,8	98
11	66	9	12	113,3	98
10	68	8	12	109,4	98

EXEMPLO – DADOS DE HALD (1952, p. 647)

- Matriz de correlação para os dados

Matrix M1

1.00000	0.22858	-0.82413	-0.24545	0.73072
0.22858	1.00000	-0.13924	-0.97295	0.81625
-0.82413	-0.13924	1.00000	0.02954	-0.53467
-0.24545	-0.97295	0.02954	1.00000	-0.82131
0.73072	0.81625	-0.53467	-0.82131	1.00000

EXEMPLO – DADOS DE HALD (1952, p. 647)

- Tabela dos testes *F*-parciais

TABLE 15.6. Partial *F*-Values for the Hald Data[a]

Subscripts of Variables in Regression	Partial *F*-Values of Variables Not in Regression			
	1	2	3	4
—	12.60	21.96	4.40	22.80
1	—	208.58	0.31	159.30
2	146.52	—	11.82	0.43
3	5.81	36.68	—	100.36
4	108.22	0.17	40.29	—
12	—	—	1.83	1.86
13	—	220.55	—	208.24
14	—	5.03	4.24	—
23	68.72	—	—	41.65
24	154.01	—	96.94	—
34	22.11	12.43	—	—
123	—	—	—	0.04
124	—	—	0.02	—
134	—	0.50	—	—
234	4.34	—	—	—

[a] For example, $F_{3|14} = 4.24$ is the partial *F*-value for b_3 given that (b_1 and b_4) are already in the regression equation. Remember that an intercept β_0 is fitted in all of the equations but is not specifically mentioned in the table.

EXEMPLO – DADOS DE HALD (1952, p. 647)

- Os testes de entrada e saída serão feitos considerando o nível $\alpha = 0{,}05$. Geralmente, é melhor escolher os mesmos níveis de significâncias para ambos os testes.

1) Escolher a primeira variável (mais correlacionada com a resposta e de maior valor *F*-parcial para apenas uma variável no modelo).

$r_{4Y} = -0{,}821$ é a maior correlação com Y e $F_{4|-} = 22{,}80$ é o maior valor *F*-parcial.

2) Teste para X_4. O valor F de 22,80 é maior que $F(1, 11; 0{,}05) = 4{,}84$. Manter X_4.

3) Procurar o próximo "melhor X", que é X_1, pois

$F_{1|4} = 108{,}22$ é maior que $F_{2|4} = 0{,}17$ e que $F_{3|4} = 40{,}29$.

Inserir X_1 e calcular a equação $\hat{Y} = f(X_1, X_4)$.

$$\hat{Y} = 103{,}10 + 1{,}440X_1 - 0{,}614X_4.$$

EXEMPLO – DADOS DE HALD (1952, p. 647)

Temos $R^2 = 0{,}972$ e um F global = 176,63, que é claramente significante. Isso completa o passo de entrada.

4) Verificar uma possível saída. Os dois F's parciais são $F_{1|4} = 108{,}22$ e $F_{4|1} = 159{,}30$. A mais recente variável adicionada X_1 é a mais fraca, mas é significante pois $108{,}22 > F(1, 10; 0{,}05) = 4{,}96$. Vamos manter X_4 e X_1.

5) Selecionamos, como a próxima variável para entrar, a mais parcialmente correlacionada com a resposta (dadas as variáveis X_4 e X_1 já na regressão). Parece ser a variável X_2, já que $F_{2|14} = 5{,}03$ é maior que $F_{3|14} = 4{,}24$. A equação é:

$$\hat{Y} = 71{,}65 + 1{,}452X_1 + 0{,}416X_2 - 0{,}237X_4,$$

com $R^2 = 0{,}98234$ e um F global = 166,83 significante.

EXEMPLO – DADOS DE HALD (1952, p. 647)

6) Verificar uma possível saída. Os três *F's* parciais são:

$$F_{1|24} = 154{,}01;\ F_{2|14} = 5{,}03 \text{ e } F_{4|12} = 1{,}86$$

Claramente, $F_{4|12} = 1{,}86 < F(1, 9; 0{,}05) = 5{,}12$. Então, devemos remover X_4. Embora $F_{2|14} = 5{,}03$ também seja menor que 5,12, não agimos em relação a X_2, pois primeiramente recalculamos a equação $\hat{Y} = f(X_1, X_2)$. Removemos apenas um X por vez.

7) A equação ajustada com X_1 e X_2 é:

$$\hat{Y} = 52{,}58 + 1{,}468X_1 + 0{,}662X_2.$$

Os *F's* parciais são $F_{1|2} = 146{,}52$ e $F_{2|1} = 208{,}58$. Então, ambas as variáveis são mantidas. $R^2 = 0{,}979$ e F global = 229,50.

8) Agora, olhamos para uma nova variável candidata e verificamos que $F_{3|12} = 1{,}83$ e $F_{4|12} = 1{,}86$. Embora X_4 seja sutilmente melhor, é inútil inseri-la. Portanto, o procedimento termina com o resultado dado no passo 7).

SELEÇÃO FORWARD

- A seleção *forward* é um procedimento mais simples (porém, menos efetivo), em que apenas a variável mais recentemente adicionada é testada.
- Ela garante que as variáveis inseridas não são subsequentemente removidas, o que pode ser desejável para aplicações específicas.
- Os autores **não** recomendam o uso da seleção **forward**, exceto quando é especificamente desejado não remover nunca as variáveis que foram retidas em estágios anteriores.
- Para o exemplo, no passo 6), apenas a variável X_2 teria sido testada (e rejeitada) e não teríamos visto que X_4 era a variável "mais fraca".

ELIMINAÇÃO BACKWARD

- Ao contrário do método *forward*, o *backward stepwise* começa com todas as variáveis incluídas. Em cada iteração, o algoritmo avalia o impacto da remoção de cada variável e exclui aquela que menos contribui para o modelo, até que nenhuma variável atenda ao critério de exclusão.
- Os passos básicos desse procedimento são:

1. Uma equação de regressão contendo todas as variáveis é calculada;
2. O valor do teste F-parcial é calculado para cada variável preditora tratada como se fosse a última variável a ser inserida no modelo de regressão;
3. O menor valor do teste F-parcial (F_L) é comparado com a nível de significância pré-selecionado ou padrão (F_0).
   I. Se $F_L < F_0$, remova a variável X_L e recalcule a equação de regressão para as variáveis remanescentes. Repita o passo 2.;
   II. Se $F_L > F_0$, adote a equação de regressão calculada.

ELIMINAÇÃO BACKWARD

Considerações sobre a eliminação ***backward***:

- Computacionalmente, é um método de seleção muito econômico;
- É possível ter uma dimensão de todas as variáveis na equação, isto é, produz um senso ao estatístico de que nenhuma variável está faltando;
- Pode não fazer sentido se a matriz $X'X$ for mal condicionada, devido a erros de aproximação.

NÍVEIS DE SIGNIFICÂNCIA PARA PROCEDIMENTOS DE SELEÇÃO

Considerações extras sobre o procedimento *stepwise*:

- É típico que o aumento do R^2 seja registrado para cada variável adicionada;
- Os valores *F*-para-entrar e *F*-para-remover não precisam necessariamente ser iguais, mas algumas considerações devem ser feitas:
  - Com o *F*-para-remover sendo maior do que o *F*-para-entrar, é possível que variáveis sejam removidas do modelo logo após serem inseridas. Essa situação pode causar um efeito cíclico de inserção e remoção de variáveis em programas que permitem reinserção.

NÍVEIS DE SIGNIFICÂNCIA PARA PROCEDIMENTOS DE SELEÇÃO

- Os valores *F* escolhidos para os testes possuem, cada um, equivalentes valores-p; entretanto, as estatísticas *F* usadas para a realização dos testes de entrada não têm a típica distribuição *F* (ao contrário do que é presumido pelo uso do teste *F*), mas sim uma distribuição cuja função de densidade de probabilidade é de difícil obtenção.
  - Os verdadeiros valores-p desses testes são maiores do que os obtidos pelos programas.

Procedimento alternativo (criado para melhorar a seleção de variáveis)**:**

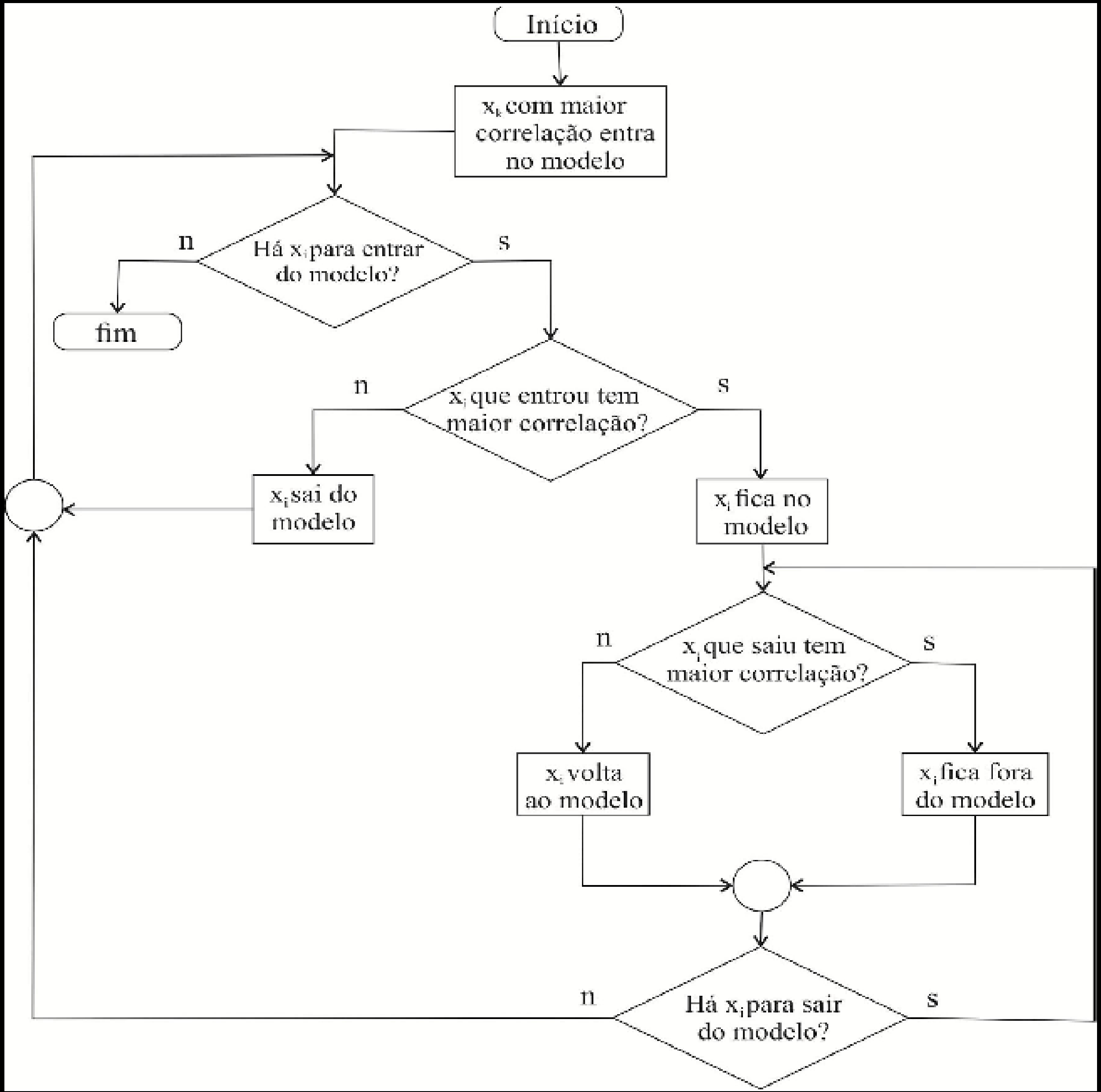
- *Rodar a regressão* stepwise, *registrar a quantidade* q *de variáveis explicativas usadas no modelo final e, então, realizar regressões para todos os conjuntos de* q *variáveis do conjunto total e escolher o melhor modelo.*

VARIAÇÕES DOS MÉTODOS

- Os procedimentos não necessariamente selecionam absolutamente o melhor modelo, mas sim um aceitável para aquele conjunto de dados;
- Proposta de procedimento alternativo:
  - Rodar o procedimento de regressão *stepwise* com níveis de aceitação e rejeição;
  - Quando o procedimento parar, determinar o número de variáveis do modelo final;
  - Use as ‘j’ em ‘k’ (total) variáveis selecionadas para montar o melhor modelo;
- Quando o procedimento seleciona 2 candidatos muito próximos para o modelo, dizemos que há informações insuficientes nos dados para fazer uma decidir um único modelo;
- O procedimento também falha quando uma grande quantidade de variáveis seriam mais apropriadas para descrever o modelo, mas nunca são analisadas pelo algoritmo de *stepwise;*
- Pode-se forçar o programa a selecionar determinada variável na equação de regressão, quando o estatístico possui a informação de que essa variável tem efeito conhecido.

CONCLUSÃO E COMENTÁRIOS DOS AUTORES

- Na maioria dos casos, as equações serão as mesmas do método dos mínimos quadrados, mas isso não é garantido;
- É válido destacar também que estatísticos diferentes podem ter interpretações distintas sobre os métodos, isto é, é comum haver escolhas inconsistentes que resultarão em equações totalmente distintas entre os vários métodos que podem ser utilizados para formar a equação de regressão;
- Para os autores (*Draper e Smith*), o método prático de regressão *stepwise* é o favorito, principalmente devido ao avanço tecnológico e a capacidade dos computadores para rodar o algoritmo. Vale lembrar que nenhuma técnica é a melhor sempre, permitindo a utilização de diversos métodos distintos como ferramentas úteis para a descrição dos dados.



Resumo:

- A regressão *stepwise* envolve a adição procedural das variáveis explicativas ao modelo, até não haver mais nenhuma significante para a explicação de *Y*, junto da remoção, após cada adição, das variáveis no modelo que não são mais significantes para a explicação de *Y*.
  - É feita a regressão de *Y* pela variável X_1 mais correlacionada com ela;
  - Realiza-se o teste *F* dessa variável. Caso não rejeitemos que $\beta_1 = 0$, o procedimento termina, sendo adotado o modelo $Y = \bar{Y}$; caso contrário, o procedimento continua;
  - São realizados testes *F*-parciais para cada uma das variáveis ainda fora do modelo e a com o maior valor-*F* é adicionada ao modelo, caso esse valor-*F* seja maior do que o valor escolhido como aceitável (*F*-para-entrar);
  - São realizados testes *F*-parciais para cada uma das variáveis que estão no modelo e são removidas do modelo as variáveis com valores-*F* abaixo do valor escolhido como aceitável (*F*-para-remover);
  - São repetidos os últimos passos, até que nenhuma variável fora do modelo passe no teste para ser adicionada e nenhuma variável dentro do modelo falhe no teste para ser mantida.

Fonte: Seleção de variáveis *stepwise* aplicadas em redes neurais artificiais para previsão de demanda de cargas elétricas. Anna Diva P. Lotufo.

REFERÊNCIAS BIBLIOGRÁFICAS

- DRAPER, N. R.; SMITH, H. Applied Regression Analysis – Third Edition (1998). *Wiley series in probability and statistics,* p. 335-348.