Modelos de Regressão Múltipla - Parte IX

Erica Castilho Rodrigues

21 de Fevereiro de 2017

Métodos de Seleção de Variáveis

Métodos de Seleção de Variáveis

- Devemos incluir no modelo todas as variáveis explicativas disponíveis?
- Ou devemos escolher apenas um subconjunto delas?
- Podemos usar procedimentos automáticos para auxiliar na escolha desse subconjunto.
- Alguns métodos que podem ser usados:
    - Todas as Regressões Possíveis (*All Regressions*)
    - Inclusão Passo a Frente (*Forward*)
    - Eliminação Passo Atrás (*Backward*)
    - Seleção Passo-a-Passo (*Stepwise*)

- Procedimentos automáticos não devem substituir o bom senso e o conhecimento sobre o problema.
- Depois da escolha das variáveis explicativas devemos:
  - verificar a significância das variáveis;
  - verificar adequação do modelo;
  - fazer análise de resíduos.

Critérios de Seleção da melhor Regressão

- Modelo "cheio": modelo com todas variáveis explicativas disponíveis.
- Precisamos escolher uma variável a ser descartada.
- Podemos escolher com base em alguma das medidas:
  - menor valor de S^2 (estimativa de σ^2);
  - maior valor de $R^2_{ajustado}$

$$R^2_{ajustado} = 1 - (1 - R^2)\left(\frac{n-1}{n-p}\right);$$

  - Estatística C_p de Mallows

$$C_p = \frac{SQE}{S^2} - (n - 2p)$$

quanto mais próxima essa estatística estiver de p, melhor é o modelo.

Todas as Regressões Possíveis (*All Regressions*)

- Testa de maneira iterativa todos os subconjuntos possíveis de variáveis explicativas.
- Se temos k variáveis, qual o número total de subconjuntos possíveis?

Todas as Regressões Possíveis (*All Regressions*)

- Testa de maneira iterativa todos os subconjuntos possíveis de variáveis explicativas.
- Se temos k variáveis, qual o número total de subconjuntos possíveis?

Todas as Regressões Possíveis (*All Regressions*)

- Testa de maneira iterativa todos os subconjuntos possíveis de variáveis explicativas.
- Se temos k variáveis, qual o número total de subconjuntos possíveis?

  $2^k - 1$ pois não contamos o modelo se nenhuma variável.

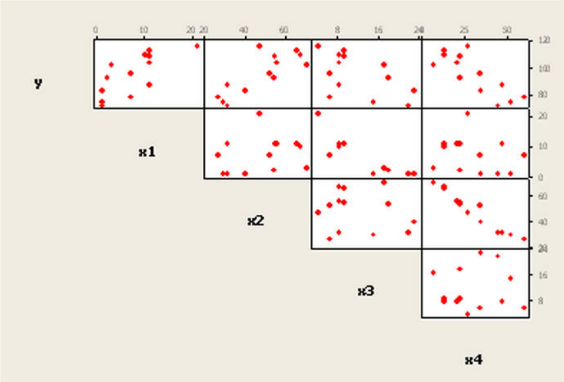
- Por exemplo, se $k = 10$, temos $2^{10} - 1 = 1023$ possibilidades.
- Escolhemos algum dos modelos com base nos critérios.
- Se o número de variáveis é grande, o método é computacionalmente inviável.

Exemplo:

- Estamos analisando a quantidade de calor envolvida no endurecimento do cimento (Y).
- As variáveis explicativas consideradas são as seguintes:
  - porcentagem de aluminato de tricálcio (X_1);
  - porcentagem de de silicato tricálcico (X_2);
  - porcentagem de tetracálcico alumino ferrita (X_3);
  - porcentagem de silicato dicálcico (X_4).

Exemplo: (continuação)

- A figura a seguir mostra os gráficos de dispersão entre as variáveis.



Exemplo: (continuação)

- Vamos usar primeiro o $R^2_{ajustado}$ como critério.
- Podemos verificar qual o ponto em que incluir mais variáveis não é vantajoso.
- Geralmente esse critério é combinado com outros critérios.
- A tabela a seguir mostra alguns dos modelos possíveis.

Best Subsets Regression: y versus x1, x2, x3, x4

Response is y

Vars	R-Sq	R-Sq(adj)	C-p	S	x1	x2	x3	x4
1	67.5	64.5	138.7	8.9639				X
1	66.6	63.6	142.5	9.0771		X		
2	97.9	97.4	2.7	2.4063	X	X		
2	97.2	96.7	5.5	2.7343	X			X
3	98.2	97.6	3.0	2.3087	X	X		X
3	98.2	97.6	3.0	2.3121	X	X	X	
4	98.2	97.4	5.0	2.4460	X	X	X	X

Exemplo: (continuação)

- A tabela mostra os dois melhores modelos para cada quantidade de variáveis.
- Com base no $R^2_{ajustado}$, qual modelo ou modelos parecem ser os melhores?

Exemplo: (continuação)

- A tabela mostra os dois melhores modelos para cada quantidade de variáveis.
- Com base no $R^2_{ajustado}$, qual modelo ou modelos parecem ser os melhores?
- Quando saímos do modelo com 1 variável para o modelo com 2:
  - o valor de $R^2_{ajustado}$ pula de 67,5 para 97,9.
- Com esse salto, vale a pena incluir uma variável no modelo?

Exemplo: (continuação)

- A tabela mostra os dois melhores modelos para cada quantidade de variáveis.
- Com base no $R^2_{ajustado}$, qual modelo ou modelos parecem ser os melhores?
- Quando saímos do modelo com 1 variável para o modelo com 2:
  - o valor de $R^2_{ajustado}$ pula de 67,5 para 97,9.
- Com esse salto, vale a pena incluir uma variável no modelo? Sim.

Exemplo: (continuação)

- Quando saímos do modelo com 2 variáveis para o modelo com 3:
  - o valor de $R^2_{ajustado}$ pula de 97,9 para 98,2.
- Com esse salto, vale a pena incluir uma variável no modelo?

Exemplo: (continuação)

- Quando saímos do modelo com 2 variáveis para o modelo com 3:
  - o valor de $R^2_{ajustado}$ pula de 97,9 para 98,2.
- Com esse salto, vale a pena incluir uma variável no modelo? Não.
- O modelo mais adequado é o modelo com as variáveis X_1 e X_2.
- Vejamos agora, se além do $R^2_{ajustado}$ usarmos S^2 como critério de escolha.
- O modelo com maior $R^2_{ajustado}$ e menor S^2 é modelo com as variávies X_1, X_2 e X_4.
- Um critério diferente, levou a escolha de um modelo diferente.

Exemplo: (continuação)

- Vamos usar agora como critério a estatística C_p de Mallows.
- Quanto mais próximo C_p estiver de p (número de variáveis), melhor é o modelo.
- Qual melhor modelo com base nesse critério?

Exemplo: (continuação)

- Vamos usar agora como critério a estatística C_p de Mallows.
- Quanto mais próximo C_p estiver de p (número de variáveis), melhor é o modelo.
- Qual melhor modelo com base nesse critério?
- O modelo com X_1, X_2 e X_3.
- Nesse caso temos que

$$C_3 = 3 .$$

- Critérios diferentes levaram a escolha de modelos distintos.

Exemplo:

- Sabe-se que se pode alterar a toxicidade de vários tipos de produtos químicos em mamíferos:
  - medicamentos, pesticidas e inceticidas.
- Isso é feito através da indução de atividade de enzimas do fígado.
- Variável resposta:desintoxicação do malatião (inceticida) (em percentagem, relativamente a um controle, não tratados, de frango).
- Temos cinco variáveis explicativas.
- Correspondentes a cinco tipos de enzimas induzidas.
- Registra-se a porcentagem de atividade enzimática em relação a um grupo controle.
- Não se sabe quais enzimas são mais imporantes para desentoxicação.
- Precisamos de um método para selecionar o melhor modelo.

Exemplo: (continuação)

- A figura a seguir mostra os resultados de ajuste com todas as regressões possíveis.

Explanatory variables in model	R^2 (in %)	R_a^2 (in %)	C_p	Explanatory variables in model	R^2 (in %)	R_a^2 (in %)	C
x_1, x_2, x_3, x_4, x_5	79.3	53.3	6.0	x_1, x_2	71.9	63.8	1.4
x_1, x_2, x_3, x_4	78.4	61.0	4.2	x_1, x_3	61.2	50.2	3.5
x_1, x_2, x_3, x_5	78.8	61.8	4.1	x_1, x_4	61.6	50.6	3.4
x_1, x_2, x_4, x_5	78.0	60.4	4.2	x_1, x_5	64.7	54.6	2.8
x_1, x_3, x_4, x_5	79.0	62.2	4.1	x_2, x_3	47.8	32.9	6.1
x_2, x_3, x_4, x_5	65.2	37.4	6.7	x_2, x_4	42.1	25.6	7.2
x_1, x_2, x_3	78.3	67.5	2.2	x_2, x_5	59.3	47.7	3.8
x_1, x_2, x_4	76.5	64.8	2.5	x_3, x_4	39.6	22.4	7.6
x_1, x_2, x_5	72.0	58.0	3.4	x_3, x_5	38.0	20.3	8.0
x_1, x_3, x_4	63.0	44.5	5.1	x_4, x_5	34.7	16.1	8.6
x_1, x_3, x_5	77.6	66.3	2.3	x_1	44.8	38.0	4.6
x_1, x_4, x_5	77.9	66.8	2.3	x_2	21.9	12.2	9.0
x_2, x_3, x_4	48.1	22.1	8.0	x_3	37.9	30.1	6.0
x_2, x_3, x_5	64.2	46.3	4.9	x_4	34.7	26.6	6.6
x_2, x_4, x_5	60.3	40.4	5.7	x_5	0.3	0.0	13.2
x_3, x_4, x_5	39.7	9.5	9.6				

Exemplo: (continuação)

- De acordo com o $R^2_{ajustado}$ qual é o melhor modelo?

Exemplo: (continuação)

- De acordo com o $R^2_{ajustado}$ qual é o melhor modelo?
- O modelo com X_1, X_2 e X_3. ($R^2_{ajustado} = 67,5\%$).
- Usando a estatísitca C_p, qual melhor modelo?

Exemplo: (continuação)

- De acordo com o $R^2_{ajustado}$ qual é o melhor modelo?
- O modelo com X_1, X_2 e X_3. ($R^2_{ajustado} = 67,5\%$).
- Usando a estatísitca C_p, qual melhor modelo?
- Modelo com X_1 e X_2 ($C_p = 1,4$).
- Notamos ainda que o modelo com X_1, X_2 e X_3:
  - além de ter o maior $R^2_{ajustado}$ tem a segunda melhor estatística C_p (2,22).
- O modelo com as variávies X_1 e X_2 também tem um valor razoável para o $R^2_{ajustado}$ (63,8%).

Exemplo: (continuação)

- Portanto dois modelos que parecem adequados são os seguintes:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

- Os ajustes dos dois modelos fica

$$\hat{Y} = 30,6 + 0,159 X_1 + 0,107 X_2$$

$$\hat{Y} = -16,5 + 0,135 X_1 + 0,0899 X_2 + 0,441 X_3$$

Inclusão Passo a Frente (Forward)

- Ordenamos as preditoras em ordem decrescente de acordo com a:
  - “intensidade de sua relação” com Y (coeficiente de correlação linear, p.ex.).
- Ajuste um modelo com a primeira variável da lista (X_3, p.ex.).
- Teste sua significância usando teste F sequencial (ou outros critérios, como o teste t).
- Se for significante, entra no modelo e o algoritmo continua.
- Se não for significante, o procedimento pára.
- Suponha que X_3 tenha entrado no modelo.

Inclusão Passo a Frente (Forward)

- Calcule a estatística F parcial de cada uma das variáveis dado X_3.
- Selecione a que tiver o maior valor (X_2, p.ex.)
- Teste sua significância usando teste F sequencial.
- Se for significante, entra no modelo e o algoritmo continua.
- Se não for significante, o procedimento pára.
- Suponha que X_2 tenha entrado no modelo.

Inclusão Passo a Frente (Forward)

- Calcule a estatística F parcial de cada uma das variáveis dado X_2 e X_3.
- Selecione a que tiver o maior valor (X_4, p.ex.)
- Teste sua significância usando teste F sequencial.
- Se for significante, entra no modelo e o algoritmo continua.
- Se não for significante, o procedimento pára.
- E assim por diante.

Observações:

- Uma vez que a variável entra no modelo, ela não sai mais.
- Mesmo que sua contribuição deixe de ser significante quando uma outra variável entrar.

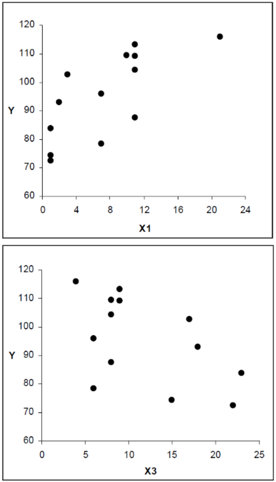
Exemplo:

- Vamos aplicar o método *Forward* para os dados abaixo.

Row	X1	X2	X3	X4	Y
1	7	26	6	60	78,5
2	1	29	15	52	74,3
3	11	56	8	20	104,3
4	11	31	8	47	87,6
5	7	52	6	33	95,9
6	11	55	9	22	109,2
7	3	71	17	6	102,7
8	1	31	22	44	72,5
9	2	54	18	22	93,1
10	21	47	4	26	115,9
11	1	40	23	34	83,8
12	11	66	9	12	113,3
13	10	68	8	12	109,4

Exemplo: (continuação)

- A figura a seguir mostra os gráficos de dispersão das variáveis.



Exemplo: (continuação)

- A figura a seguir mostra os resultados do método *Forward*.

Stepwise Regression: Y versus X1; X2; X3; X4
Forward selection.
Alpha-to-Enter: 0,25

Response is Y on 4 predictors, with N = 13

Step	**1**	**2**	**3**
Constant	117,57	103,10	71,65
X4	-0,738	-0,614	-0,237
T-Value	-4,77	-12,62	-1,37
P-Value	**0,001**	0,000	0,205
X1		1,44	1,45
T-Value		10,40	12,41
P-Value		**0,000**	0,000
X2			0,42
T-Value			2,24
P-Value			**0,052**
S	8,96	2,73	2,31
R-Sq	67,45	97,25	98,23
R-Sq(adj)	64,50	96,70	97,64
Mallows C-p	138,7	5,5	3,0

Exemplo: (continuação)

- Os passos do algoritmo são os seguintes:
  - Passo 1:
$$\text{entra } X_4 \Rightarrow \text{modelo}(Y, X_4)$$
  - Passo 2:
$$\text{entra } X_1 \Rightarrow \text{modelo}(Y, X_1, X_4)$$
  - Passo 3:
$$\text{entra } X_2 \Rightarrow \text{modelo}(Y, X_1, X_2, X_4)$$
  - Passo 4:
$$\text{não entra } X_3 \Rightarrow \text{modelo}(Y, X_1, X_2, X_4)$$
- Modelo selecionado:
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \epsilon \,.$$

Eliminação Passo Atrás (*Backward*)

- Ajuste um modelo com todas as preditoras (X_1, X_2, X_3, X_4, p.ex.).
- Para cada preditora, calcule:
  - a estatística F parcial como se ela tivesse sido a última a entrar no modelo

$$F(X_1|X_2, X_3, X_4) \quad F(X_2|X_1, X_3, X_4) \quad F(X_3|X_1, X_3, X_4)$$
$$F(X_4|X_1, X_2, X_3)\ .$$

- Escolha aquela com o menor valor (X_2 com $F(X_2|X_1, X_3, X_4)$, p.ex.)
- Teste sua significância usando teste F parcial (ou outro critério, como teste t por ex.).
- Se for significante, o procedimento pára.
- Se não for significante, sai do modelo e o algoritmo continua.
- Suponha que X_2 tenha saído do modelo.

Eliminação Passo Atrás (*Backward*)

- Ajuste um modelo com as preditoras X_1, X_3 e X_4.
- Para cada preditora, calcule:
  - a estatística F parcial como se ela tivesse sido a última a entrar no modelo

$$F(X_1|X_3, X_4) \quad F(X_3|X_1, X_4) \quad F(X_4|X_1, X_3)\ .$$

- Escolha aquela com o menor valor (X_4 com $F(X_4|X_1, X_3)$, p.ex.)
- Teste sua significância usando teste F parcial.
- Se for significante, o procedimento pára.
- Se não for significante, sai do modelo e o algoritmo continua.
- Repita o procedimento até que nenhuma variável seja eliminada.

Observações

- Uma vez que a variável sai no modelo, ela não entra mais.
- Mesmo que sua contribuição passe a ser significante quando uma outra variável entrar.

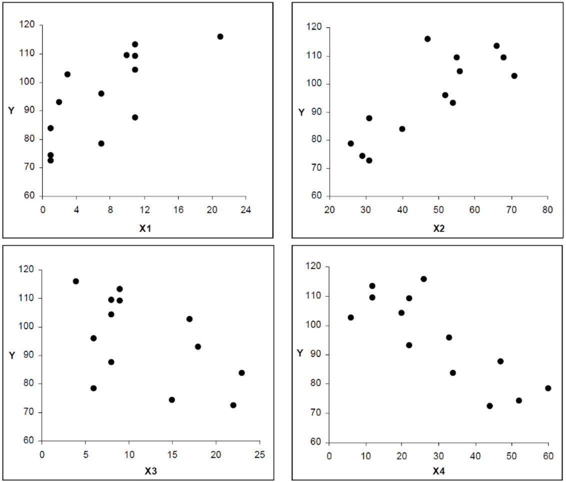
Exemplo:

- Vamos aplicar o método *Backward* para os dados abaixo.

Row	X1	X2	X3	X4	Y
1	7	26	6	60	78,5
2	1	29	15	52	74,3
3	11	56	8	20	104,3
4	11	31	8	47	87,6
5	7	52	6	33	95,9
6	11	55	9	22	109,2
7	3	71	17	6	102,7
8	1	31	22	44	72,5
9	2	54	18	22	93,1
10	21	47	4	26	115,9
11	1	40	23	34	83,8
12	11	66	9	12	113,3
13	10	68	8	12	109,4

Exemplo: (continuação)

- A figura mostra os gráficos de dispersão das variáveis.



Exemplo: (continuação)

- A figura a seguir mostra os resultados do método *Backward*.

```
Stepwise Regression: Y versus X1; X2; X3; X4
Backward elimination.
Alpha-to-Remove: 0,10

Response is Y on 4 predictors, with N = 13

Step               1       2       3
Constant       62,41   71,65   52,58

X1              1,55    1,45    1,47
T-Value         2,08   12,41   12,10
P-Value        0,071   0,000   0,000

X2             0,510   0,416   0,662
T-Value         0,70    2,24   14,44
P-Value        0,501   0,052   0,000

X3              0,10
T-Value         0,14
P-Value        0,896

X4             -0,14   -0,24
T-Value        -0,20   -1,37
P-Value        0,844   0,205

S               2,45    2,31    2,41
R-Sq           98,24   98,23   97,87
```

Exemplo: (continuação)

- Os passos do algoritmo são os seguintes:
  - Passo 1:

$$\text{saiu } X_3 \Rightarrow \text{modelo}(Y, X_1, X_2, X_4)$$

  - Passo 2:

$$\text{saiu } X_4 \Rightarrow \text{modelo}(Y, X_1, X_2)$$

  - Passo 3:

não sai nenhuma

- Modelo selecionado:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon\,.$$

Seleção Passo a Passo (*Stepwise*)

- É uma mistura de *Forward* e *Backward*.
- A variável que entra em um passo pode sair nos próximos.
- A variável que sai em um passo pode entrar nos próximos.

Seleção Passo a Passo (*Stepwise*)

- Começa como no *Forward*:
  - Ordene as preditoras em ordem decrescente da "intensidade de sua relação" com Y.
  - Ajuste um modelo com a primeira variável da lista (X_3, p.ex.).
  - Teste sua significância.
  - Se for significante, continua o algoritmo.
  - Se não for significante, o procedimento pára.
  - Calcule o F parcial de cada uma das variáveis dado X_3.
  - Selecione a que tiver o maior valor (X_2, p.ex.)
  - Teste sua significância.
  - Se for significante, continua o algoritmo.
  - Se não for significante, o procedimento pára.
- Aqui, verifique a possível saída de variáveis.

Seleção Passo a Passo (*Stepwise*)

- Passamos para o *Backward*:
  - Calcule as estatísticas
    $$F(X_2|X_3) \text{ e } F(X_3|X_2)$$
    e teste a contribuição de uma variável dado a outra.
  - Se ambas são significantes, as duas ficam no modelo.
  - Caso contrário, aquela que não for sai do modelo.
  - Vamos supor que nenhuma saiu.
- Prossiga com o *Forward*

Seleção Passo a Passo (*Stepwise*)

- Prosseguindo com *Forward*:
  - Calcule o F parcial das variáveis fora do modelo dado X_3 e X_2.
  - Selecione a que tiver o maior valor (X_1, p.ex.)
  - Teste sua contribuição usando Teste F seqüencial.
  - Se for significante, continua o algoritmo.
  - Se não for significante, o procedimento pára.
- Suponha que X_1 entrou no modelo.
- Novamente cheque uma possível saída de variáveis

Seleção Passo a Passo (*Stepwise*)

- Passamos para o *Backward*:
  - Calcule as estatísticas

$$F(X_1|X_2, X_3) \quad F(X_2|X_1, X_3) \quad \text{e } F(X_3|X_1, X_2)$$

    e teste a contribuição de uma variável dado a outra.
  - Selecione a de menor valor (X_2, p.ex.)
  - Teste sua contribuição usando Teste F seqüencial.
  - Se for significante, continua o algoritmo.
  - Se não for significante, o procedimento pára.
- Suponha que X_2 tenha saído.
- Prossiga com o *Forward*

Seleção Passo a Passo (*Stepwise*)

- Prosseguindo com *Forward*:
  - Calcule o F parcial das variáveis fora do modelo dado X_1 e X_2.
  - Selecione a que tiver o maior valor (X_4, p.ex.)
  - Teste sua contribuição usando Teste F seqüencial.
  - Se for significante, continua o algoritmo.
  - Se não for significante, o procedimento pára.
- Suponha que X_4 entrou no modelo.
- Novamente cheque uma possível saída de variáveis

Seleção Passo a Passo (*Stepwise*)

- Passamos para o *Backward*:
  - Calcule as estatísticas

$$F(X_1|X_3,X_4) \quad F(X_3|X_1,X_4) \quad \text{e } F(X_4|X_1,X_3)$$

    e teste a contribuição de uma variável dado a outra.
  - Selecione a de menor valor.
  - Teste sua contribuição usando Teste F seqüencial.
  - Se for significante, continua o algoritmo.
  - Se não for significante, o procedimento pára.
- O procedimento continua até que se chegue a um modelo já foi tentado antes.
- O modelo escolhido é o do passo anterior.

Exemplo:

- Vamos aplicar o método *Stepwise* para os dados abaixo.

Row	X1	X2	X3	X4	Y
1	7	26	6	60	78,5
2	1	29	15	52	74,3
3	11	56	8	20	104,3
4	11	31	8	47	87,6
5	7	52	6	33	95,9
6	11	55	9	22	109,2
7	3	71	17	6	102,7
8	1	31	22	44	72,5
9	2	54	18	22	93,1
10	21	47	4	26	115,9
11	1	40	23	34	83,8
12	11	66	9	12	113,3
13	10	68	8	12	109,4

Exemplo: (continuação)

- A figura a seguir mostra os resultados do método *Backward*.

Stepwise Regression: Y versus X1; X2; X3; X4

Alpha-to-Enter: 0,15   Alpha-to-Remove: 0,15

Response is Y on 4 predictors, with N = 13

Step	**1**	**2**	**3**	**4**
Constant	117,57	103,10	71,65	52,58
X4	-0,738	-0,614	-0,237	
T-Value	-4,77	-12,62	-1,37	
P-Value	**0,001**	**0,000**	**0,205**	
X1		1,44	1,45	1,47
T-Value		10,40	12,41	12,10
P-Value		**0,000**	0,000	**0,000**
X2			0,416	0,662
T-Value			2,24	14,44
P-Value			**0,052**	**0,000**
S	8,96	2,73	2,31	2,41
R-Sq	67,45	97,25	98,23	97,87
R-Sq(adj)	64,50	96,70	97,64	97,44
Mallows C-p	138,7	5,5	3,0	2,7

Exemplo: (continuação)

- Os passos do algoritmo são os seguintes:
  - Passo 1:
    $$\text{entra } X_4 \Rightarrow \text{modelo}(Y, X_4)$$
  - Passo 2:
    $$\text{entra } X_1 \Rightarrow \text{modelo}(Y, X_1, X_4)$$
  - Passo 3:
    $$\text{fica } X_1 \Rightarrow \text{modelo}(Y, X_1, X_4)$$
  - Passo 4:
    $$\text{entra } X_2 \Rightarrow \text{modelo}(Y, X_1, X_2.X_4)$$
  - Passo 5:
    $$\text{sai } X_4 \Rightarrow \text{modelo}(Y, X_1, X_2)$$
  - Passo 6:
    $$\text{não entra } X_4 \Rightarrow \text{modelo}(Y, X_1, X_2)$$

Exemplo: (continuação)

- Modelo selecionado:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \, .$$

- Mesmo modelo selecionado no *Backward*.
- Mas o *Stepwise* é um algoritmo mais completo.