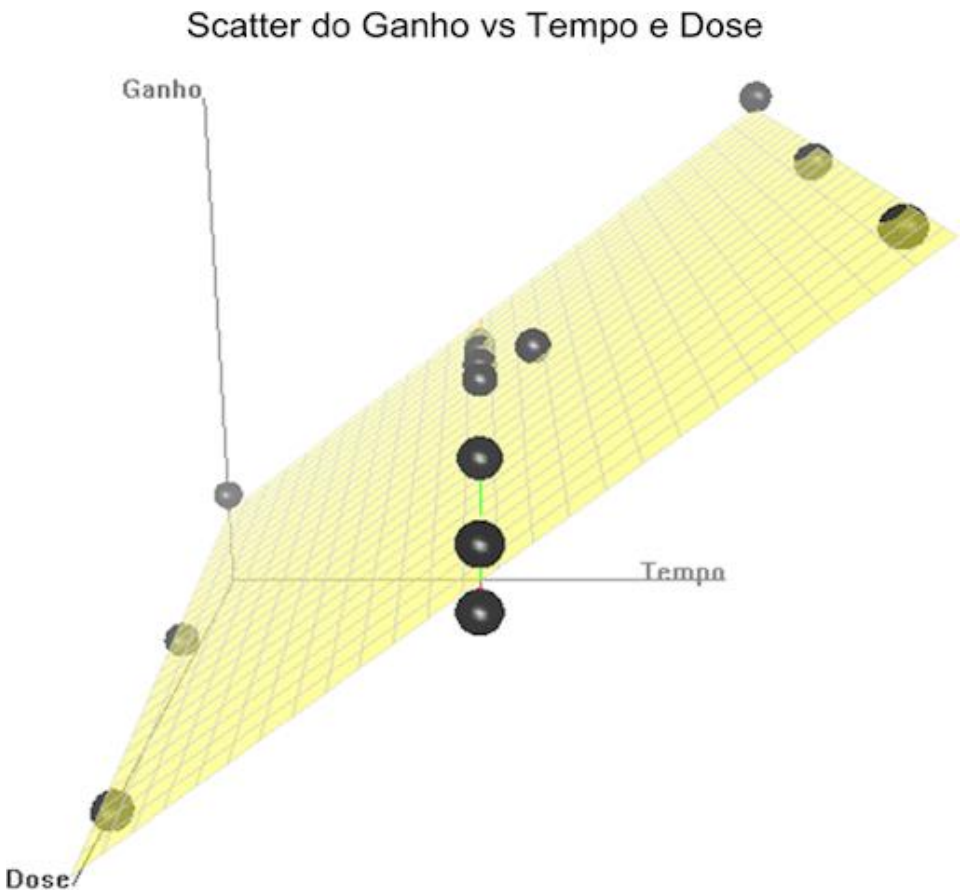
MODELO DE DE REGRESSÃO LINEAR MÚLTIPLA

Apostila Suely Ruiz Giolo



02/12 - Análise de Regressão Múltipla - ARM

09/12 - Análise de Regressão Múltipla - ARM

16/12 - Aula Prática (Exercício)

1. ARM (Interpretação, ANOVA, Coeficientes (Determinação e correlação), Matricial, MMQO)
2. IC
3. TH
4. Diagnósticos
5. Multicolinearidade
6. Diagnóstico de Influência
7. Métodos para tratar multicolinearidade
8. Seleção de variáveis e construção do modelo
9. Extrapolações
10. Validações MRLM
11. Regressão com parte categórica (Variáveis Dammy)
12. Regressão Polinomial
13. Exemplos

13/01 - ARM – Seleção de variáveis

20/01 - Variáveis Dummy

27/01 – 03/02 – 10/02 – 17/02

SELEÇÃO DO MELHOR MODELO

- ✓ R^2 ;
- ✓ R^2 ajustado;
- ✓ Qmresíduo;
- ✓ Sqres
- ✓ Cp de Mallwos

$$C_p = \frac{SQres(p)}{\sigma^2} - n + 2p$$

(Cp= p baixos)

Métodos passo a passo (Stepwise) - Seleção automática

- ✓ seleção passo à frente *(forward*)
- ✓ seleção passo atrás (*backward)*
- ✓ seleção passo a passo *(stepwise)*

- ✓ AIC e BIC (menores melhores)

Página 30 - Giollo

Seleção passo à frente *(Forward)*

Esse procedimento começa sem ***nenhuma regressora*** no modelo de regressão.

Passo 1: a *primeira* regressora a entrar no modelo é a que apresentar maior correlação simples com a resposta Y, isto é, maior r_{YXj} (j = 1, ..., *k*) ou, equivalentemente, o menor *p*-valor associado ao teste F tal que *p*-valor ≤ α,

Passo 2: a *segunda* regressora a entrar no modelo é, agora, a com maior correlação parcial com a resposta Y, isto é, maior $r_{Yj \bullet i}$ para j ≠ i e i o índice da regressora escolhida no primeiro passo.

Em outras palavras, a segunda regressora escolhida é aquela com maior estatística F parcial tal que *p*-valor ≤ α,

Passos subsequentes: análogo ao segundo passo, as demais regressoras são inseridas

Problema: Uma regressora escolhida em um determinado passo, permanece no modelo até o passo final, não havendo a possibilidade de exclusão da mesma em qualquer outro passo subsequente.

Quadro 2. Resultados da seleção *forward* (passo à frente) obtidos no *software* R.

```
Start:  AIC= 443.14
 Y ~ 1
       Df Sum of Sq     RSS       Cp  F value     Pr(F)
+ X4    1   1831.90  883.87 138.8038  22.7985 0.0005762 ***
+ X2    1   1809.43  906.34 142.5613  21.9606 0.0006648 ***
+ X1    1   1450.08 1265.69 202.6533  12.6025 0.0045520 **
+ X3    1    776.36 1939.40 315.3145   4.4034 0.0597623 .
<none>              2715.76 443.1410

Step:  AIC= 138.8
 Y ~ X4
       Df Sum of Sq    RSS       Cp  F value     Pr(F)
+ X1    1    809.10  74.76   5.5020 108.2239 1.105e-06 ***
+ X3    1    708.13 175.74  22.3876  40.2946 8.375e-05 ***
+ X2    1     14.99 868.88 138.2977   0.1725    0.6867
<none>              883.87 138.8038

Step:  AIC= 5.5
 Y ~ X4 + X1
       Df Sum of Sq    RSS     Cp F value   Pr(F)
+ X2    1    26.789 47.973 3.0222  5.0259 0.05169 .
+ X3    1    23.926 50.836 3.5010  4.2358 0.06969 .
<none>              74.762 5.5020

Step:  AIC= 3.02
 Y ~ X4 + X1 + X2
       Df Sum of Sq    RSS     Cp F value Pr(F)
<none>              47.973 3.0222
+ X3    1     0.109 47.864 5.0040  0.0182 0.896

Call:
lm(formula = Y ~ X4 + X1 + X2)

Coefficients:
(Intercept)           X4           X1           X2
    71.6483      -0.2365       1.4519       0.4161
```

Seleção passo atrás *(Backward)*

Esse procedimento começa com ***todas regressoras*** no modelo de regressão.

Passo 1: para cada regressora é calculado a estatística F parcial como se ela fosse a última regressora a entrar no modelo. A primeira regressora a ser eliminada do modelo é aquela que apresentar o menor valor para a estatística F parcial ou, o maior *p*-valor associado a essa estatística tal que *p*-valor > α,

Passo 2 e subsequentes: o modelo sem a regressora escolhida no primeiro passo é ajustado e novas estatísticas F parcias são calculadas para esse novo modelo.

Assim é repetido até que todas as regressoras que permanecerem no modelo apresentem *p*-valores associados aos testes F parciais menores ou iguais a α.

Problema: Uma regressora escolhida em um determinado passo para ser removida do modelo, não tem a possibilidade de vir a ser incluída em qualquer outro

Quadro 3. Resultados da seleção *backward* (passo atrás) obtidos no *software* R.

```
Start:  AIC= 5
 Y ~ X1 + X2 + X3 + X4
       Df Sum of Sq    RSS     Cp F value   Pr(F)
- X3    1     0.109 47.973 3.0222  0.0182 0.89592
- X4    1     0.247 48.111 3.0453  0.0413 0.84407
- X2    1     2.972 50.836 3.5010  0.4968 0.50090
<none>              47.864 5.0040
- X1    1    25.951 73.815 7.3436  4.3375 0.07082 .

Step:  AIC= 3.02
 Y ~ X1 + X2 + X4
       Df Sum of Sq    RSS       Cp  F value     Pr(F)
- X4    1      9.93  57.90   2.6830   1.8633   0.20540
<none>               47.97   3.0222
- X2    1     26.79  74.76   5.5020   5.0259   0.05169 .
- X1    1    820.91 868.88 138.2977 154.0076 5.781e-07 ***

Step:  AIC= 2.68
 Y ~ X1 + X2
       Df Sum of Sq     RSS      Cp F value     Pr(F)
<none>               57.90   2.683
- X1    1    848.43  906.34 142.561 146.523 2.692e-07 ***
- X2    1   1207.78 1265.69 202.653 208.582 5.029e-08 ***

Call:
lm(formula = Y ~ X1 + X2, data = exe5)
Coefficients:
(Intercept)           X1           X2
    52.5773       1.4683       0.6623
```

Seleção passo atrás *(Stepwise)*

Passo 1: a *primeira* regressora a entrar no modelo é a que apresentar maior correlação simples com a resposta Y, isto é, maior r_{YXj} (j = 1, ..., *k*) ou, o menor *p*-valor associado ao teste F tal que *p*-valor ≤ α1,

Passo 2: a *segunda* regressora a entrar no modelo é a que apresentar maior correlação parcial com a resposta Y, isto é, maior $r_{Yj \bullet i}$ para j ≠ i e i o índice da regressora escolhida no primeiro passo. Em outras palavras, a segunda regressora escolhida é aquela com maior estatística F parcial tal que *p*-valor ≤ α1,

Passo 3 e Passos subsequentes: todas as regressoras são reacessadas por meio de suas respectivas estatísticas F parciais. E havendo regressoras no modelo, segundo α2 estabelecido, que se mostrem redundantes na presença das demais, remove-se a que apresentar menor significância para o modelo.

Quadro 4. Resultados da seleção *stepwise* (passo a passo) obtidos no *software* R.

```
Start:  AIC= 443.14
 Y ~ 1
       Df Sum of Sq     RSS       Cp  F value     Pr(F)
+ X4    1   1831.90  883.87 138.8038  22.7985 0.0005762 ***
+ X2    1   1809.43  906.34 142.5613  21.9606 0.0006648 ***
+ X1    1   1450.08 1265.69 202.6533  12.6025 0.0045520 **
+ X3    1    776.36 1939.40 315.3145   4.4034 0.0597623 .
<none>               2715.76 443.1410

Step:  AIC= 138.8
 Y ~ X4
       Df Sum of Sq     RSS       Cp  F value     Pr(F)
+ X1    1    809.10   74.76   5.5020 108.2239 1.105e-06 ***
+ X3    1    708.13  175.74  22.3876  40.2946 8.375e-05 ***
+ X2    1     14.99  868.88 138.2977   0.1725 0.6866842
<none>               883.87 138.8038
- X4    1   1831.90 2715.76 443.1410  22.7985 0.0005762 ***

Step:  AIC= 5.5
 Y ~ X4 + X1
       Df Sum of Sq     RSS       Cp  F value     Pr(F)
+ X2    1     26.79   47.97   3.0222   5.0259   0.05169 .
+ X3    1     23.93   50.84   3.5010   4.2358   0.06969 .
<none>               74.76   5.5020
- X1    1    809.10  883.87 138.8038 108.2239 1.105e-06 ***
- X4    1   1190.92 1265.69 202.6533 159.2952 1.815e-07 ***

Step:  AIC= 3.02
 Y ~ X4 + X1 + X2
       Df Sum of Sq     RSS       Cp  F value     Pr(F)
- X4    1      9.93   57.90   2.6830   1.8633   0.20540
<none>               47.97   3.0222
+ X3    1      0.11   47.86   5.0040   0.0182   0.89592
- X2    1     26.79   74.76   5.5020   5.0259   0.05169 .
- X1    1    820.91 868.88 138.2977 154.0076 5.781e-07 ***

Step:  AIC= 2.68
 Y ~ X1 + X2
       Df Sum of Sq     RSS       Cp  F value     Pr(F)
<none>               57.90   2.6830
+ X4    1      9.93   47.97   3.0222   1.8633    0.2054
+ X3    1      9.79   48.11   3.0453   1.8321    0.2089
- X1    1    848.43  906.34 142.5613 146.5227 2.692e-07 ***
- X2    1   1207.78 1265.69 202.6533 208.5818 5.029e-08 ***
Call:
lm(formula = Y ~ X1 + X2)
Coefficients:
(Intercept)           X1           X2
    52.5773       1.4683       0.6623
```

COMENTÁRIOS

- OS procedimentos de seleção *forward*, eliminação *backward* e seleção *stepwise,* não necessariamente levam a escolha do mesmo modelo final;

- Recomenda-se que todos os procedimentos sejam aplicados na esperança de que haja alguma concordância entre eles, ou mesmo para aprender algo mais sobre a estrutura dos dados;

- O modelo final obtido por qualquer um dos procedimentos deve ser analisado quanto ao seu sentido prático. Analistas inexperientes podem concluir por um modelo não realístico.

COMENTÁRIOS

Para o modelo escolhido sugere-se, portanto, que o analista responda as questões a seguir:

- O modelo obtido é razoável? Isto é, as regressoras no modelo fazem sentido à luz do problema real?

- O modelo é útil para o propósito pretendido? (custos para a coleta dos dados, regressoras observáveis na prática etc.).

- Os coeficientes de regressão são razoáveis? Isto é, os sinais e magnitude dos coeficientes são realísticos e seus erros-padrão relativamente pequenos?

- O modelo apresenta bom ajuste aos dados? (análise de resíduos, diagnóstico de influência etc.).

EXEMPLO

Tabela 4: Dados observados em um estudo envolvendo quatro regressoras.

Observação i	Y_i	X_{i1}	X_{i2}	X_{i3}	X_{i4}
1	78,5	7	26	6	60
2	74,3	1	29	15	52
3	104,3	11	56	8	20
4	87,6	11	31	8	47
5	95,9	7	52	6	33
6	109,2	11	55	9	22
7	102,7	3	71	17	6
8	72,5	1	31	22	44
9	93,1	2	54	18	22
10	115,9	21	47	4	26
11	83,8	1	40	23	34
12	113,3	11	66	9	12
13	109,4	10	68	8	12

Fonte: Montgomery e Peck, 1992.

Tabela 5. Resumo de todas as regressões possíveis com o intercepto e sem interações.

no. de regressoras	no. de parâmetros	regressoras no modelo	SQres	R^2_p	R^2_a	QMres	C_p
Nenhuma	1	Nenhuma	2715,76	0	0	226,31	443,14
1	2	X_1	1265,68	0,5339	0,4915	115,06	202,55
1	2	X_2	906,33	0,6662	0,6359	82,39	142,49
1	2	X_3	1939,40	0,2858	0,2209	176,31	315,16
1	2	X_4	883,86	0,6745	0,6449	80,35	138,73
2	**3**	**X_1 X_2**	**57,90**	**0,9786**	**0,9744**	**5,79**	**2,68**
2	3	X_1 X_3	1227,07	0,5481	0,4578	122,70	198,10
2	3	X_1 X_4	74,76	0,9724	0,9669	7,47	5,50
2	3	X_2 X_3	415,44	0,8470	0,8164	41,54	62,44
2	3	X_2 X_4	868,88	0,6800	0,6160	86,88	138,23
2	3	X_3 X_4	175,73	0,9352	0,9223	17,57	22,37
3	4	X_1 X_2 X_3	48,11	0,9822	0,9763	5,34	3,04
3	4	X_1 X_2 X_4	47,97	0,9823	0,9764	5,33	3,02
3	4	X_1 X_3 X_4	50,83	0,9812	0,9750	5,64	3,50
3	4	X_2 X_3 X_4	73,81	0,9728	0,9637	8,20	7,34
4	5	X_1 X_2 X_3 X_4	47,86	0,9823	0,9735	5,98	5,00

Tabela 6. Matriz de correlações simples.

	X_1	X_2	X_3	X_4
X_1	1,0			
X_2	0,229	1,0		
X_3	**-0,824**	-0,139	1,0	
X_4	-0,245	**-0,973**	0,030	1,0
Y	0,731	0,816	-0,535	-0,821

Tabela 7. Estimativas por MQO para as 16 regressões consideradas.

Regressoras no modelo	$\hat{\beta}_o$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
X_1	81,479	1,869			
X_2	57,424		**0,789**		
X_3	110,203			-1,256	
X_4	117,568				-0,738
X_1 X_2	52,577	1,468	0,662		
X_1 X_3	72,349	2,312		0,494	
X_1 X_4	103,097	1,440			-0,614
X_2 X_3	72,075		0,731	-1,008	
X_2 X_4	94,160		**0,331**		-0,457
X_3 X_4	131,282			-1,200	-0,724
X_1 X_2 X_3	48,194	1,696	0,657	0,250	
X_1 X_2 X_4	71,648	1,452	0,416		-0,237
X_1 X_3 X_4	111,684	1,052		-0,410	-0,643
X_2 X_3 X_4	203,642		**-0,923**	-1,448	-1,557
X_1 X_2 X_3 X_4	62,405	1,551		0,102	-0,144

i) modelo resultante da seleção *forward*: Y em X_4, X_1 e X_2
ii) modelo resultante da seleção *backward*: Y em X_1 e X_2
iii) modelo resultante da seleção *stepwise*: Y em X_1 e X_2.

Extrapolações

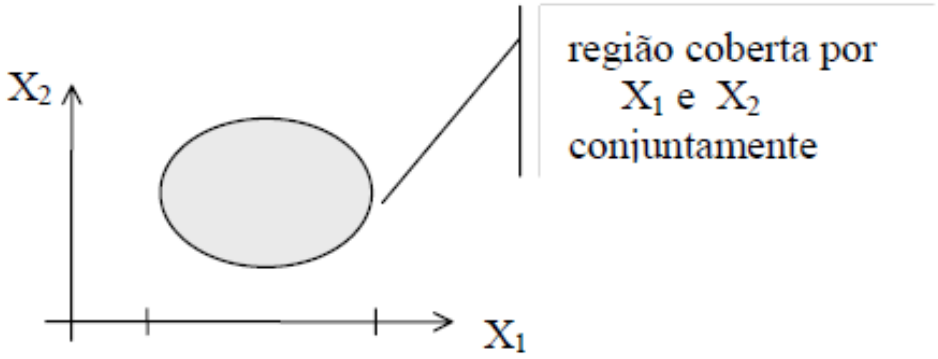


Figura 6 – Região conjunta de X1 e X2.

Para detectar se um ponto $\mathbf{x} = (1, x_{i1}, x_{i2}, \ldots, x_{i,k})$ pertence à região conjunta, usa-se o seguinte procedimento baseado nos elementos h_{ii} da diagonal da matriz **H**. Considere $h_{máx}$ = maior valor de h_{ii} . O conjunto de pontos **x** que satisfizerem:

$$\mathbf{x'(X'X)^{-1}x} \leq h_{máx}$$

estarão inclusos no elipsóide ou região conjunta definida pelas regressoras. Logo, se o interesse for o de predizer Y em $\mathbf{x}_o = [1 \quad x_{o1} \quad x_{o2} \quad \ldots. \quad x_{op}]$, a localização desse ponto relativa ao elipsóide será obtida por:

$$h_{oo} = \mathbf{x'_o (X'X)^{-1} x_o}.$$

Se $h_{oo} > h_{máx}$, $\mathbf{x}_o$ estará fora do elipsóide. Caso contrário, $\mathbf{x}_o$ estará dentro ou nos limites do elipsóide.

Validação do modelo

Será que o modelo escolhido funciona?

- Análise dos coeficientes do modelo e dos valores preditos por meio de:

- comparações com experimentos anteriores, quando existirem;
- resultados de simulação.

- Coleta de novos dados para verificar o desempenho preditivo do modelo.

- Partição (*split*) dos dados, que consiste em deixar parte dos dados originais fora da análise para investigar o desempenho preditivo do modelo com a parte não utilizada no ajuste.

PASSOS DA MRLM

1- Coleta dos dados

2- Define um modelo

3- Análise de resíduos

4- Multicolinearidade (VIF, k, determinante e autovalores de matriz r_{xx})

5- Pontos influentes (hii, distância de Cook, DIFFITS, DFBetas,Covratio)

6- Modelo final

8- Extrapolações