Regressão Linear

PNV-3421 – Processos Estocásticos

Prof. Dr. João Ferreira Netto

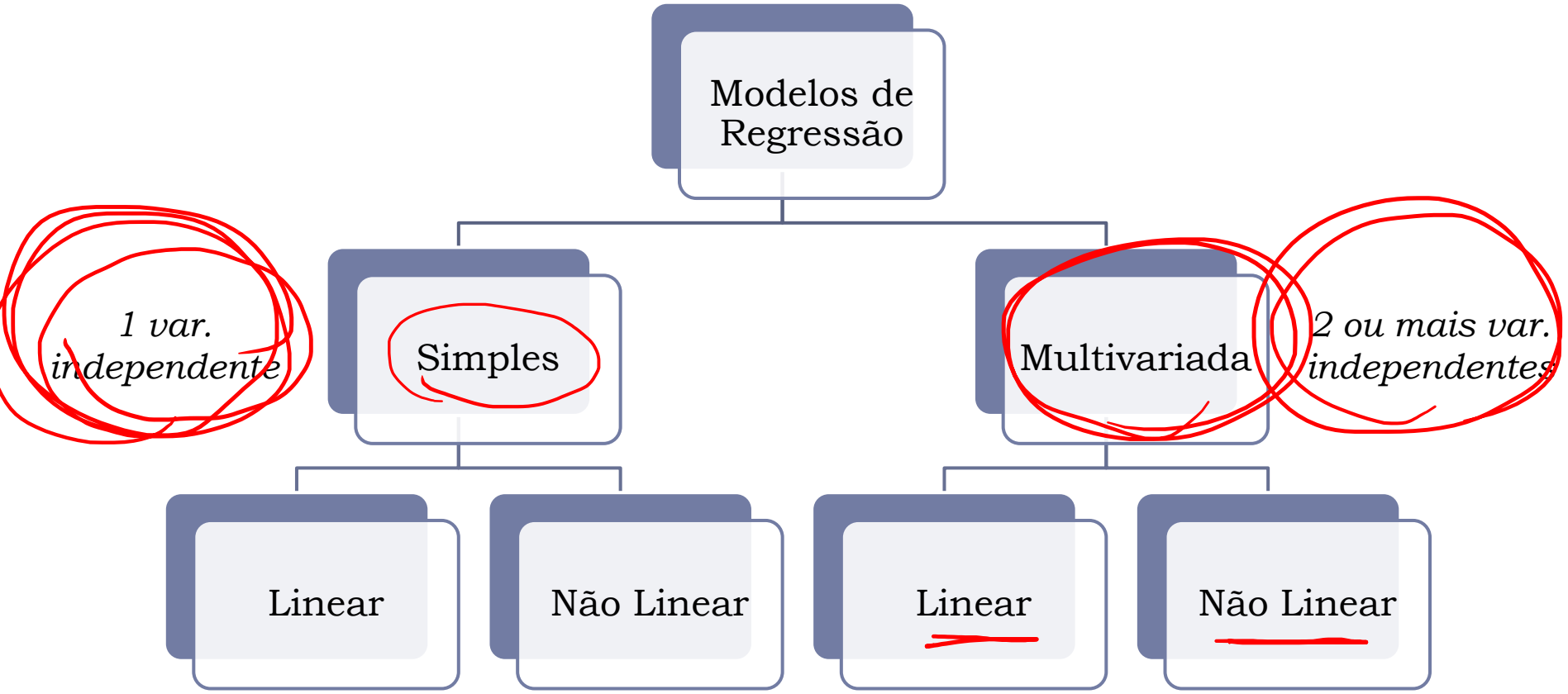
Bibliografia Principal

- Hanke, J.E & Reitsch A.G. (1998) Business Forecasting. 6th Edition, Prentice Hall, Upper Sadle River, NJ.

Modelos de Regressão

- Expressar a relação entre uma variável dependente e variáveis explanatórias (independentes), por meio de uma equação, com o objetivo de previsão.
- As variáveis podem ser numéricas ou indicativas de uma categoria.

Modelos de Regressão



Regressão Linear Multivariada

Regressão Linear – Exemplo Introdutório

- No exemplo introdutório usado em regressão linear simples, buscou-se uma expressão linear, baseada na variável independente 'preço de venda' (X), a fim de explicar o nível de vendas (Y).

#	X	Y
1	1,3	10
2	2,0	6
3	1,7	5
4	1,5	12
5	1,6	10
6	1,2	15
7	1,6	5
8	1,4	12
9	1,0	17
10	1,1	20

$R^2 = 75\%$

~75%

e os outros 25%?

Regressão Linear – Exemplo Introdutório

- Para este exemplo, o coeficiente de determinação (r^2), que indica o percentual da variação em Y que pode ser explicada pela variação em X, era de $(-0{,}8635)^2 = 0{,}75$.
- Assim, existe 25% de variação em Y que não é consequência da (não é explicada pela) variação em X.

Regressão Linear – Exemplo Introdutório

- Uma nova variável independente será introduzida na análise, e consiste na verba gasta em publicidade (milhares de reais).

#	Preço	Publicidade	Vendas
1	1,3	9,0	10
2	2,0	7,0	6
3	1,7	5,0	5
4	1,5	14,0	12
5	1,6	15,0	10
6	1,2	12,0	15
7	1,6	6,0	5
8	1,4	10,0	12
9	1,0	15,0	17
10	1,1	21,0	20

X_1 X_2 Y

Regressão Linear Multivariada

- Reta (curva) de Regressão: $\hat{Y} = b_0 + b_2 X_2 + b_3 X_3$
- A melhor regressão é aquela que minimiza a soma das diferenças quadráticas (distância) entre os pontos e a reta – *método dos mínimos quadrados*, e deriva da resolução do sistema abaixo:

$$\begin{cases} \sum Y = n\, b_0 + b_2 \sum X_2 + b_3 \sum X_3 \\ \sum X_2 Y = b_0 \sum X_2 + b_2 \sum {X_2}^2 + b_3 \sum X_2 X_3 \\ \sum X_3 Y = b_0 \sum X_3 + b_2 \sum X_2 X_3 + b_3 \sum {X_3}^2 \end{cases}$$

Regressão Linear Multivariada
Coeficiente de Correlação

#	Y	X_2	X_3	X_2Y	X_3Y	X_2X_3	Y^2	X_2^2	X_3^2
1	10,0	1,3	9,0	13,0	90,0	11,7	100,0	1,7	81,0
2	6,0	2,0	7,0	12,0	42,0	14,0	36,0	4,0	49,0
3	5,0	1,7	5,0	8,5	25,0	8,5	25,0	2,9	25,0
4	12,0	1,5	14,0	18,0	168,0	21,0	144,0	2,3	196,0
5	10,0	1,6	15,0	16,0	150,0	24,0	100,0	2,6	225,0
6	15,0	1,2	12,0	18,0	180,0	14,4	225,0	1,4	144,0
7	5,0	1,6	6,0	8,0	30,0	9,6	25,0	2,6	36,0
8	12,0	1,4	10,0	16,8	120,0	14,0	144,0	2,0	100,0
9	17,0	1,0	15,0	17,0	255,0	15,0	289,0	1,0	225,0
10	20,0	1,1	21,0	22,0	420,0	23,1	400,0	1,2	441,0
Soma	**112,00**	**14,40**	**114,00**	**149,30**	**1.480,00**	**155,30**	**1.488,00**	**21,56**	**1.522,00**
Média	**11,2**	**1,44**	**11,4**						

Regressão Linear Multivariada

- Sistema:

$$112 = 10b_0 + 14.4b_2 + 114b_3$$
$$149.3 = 14.4b_0 + 21.56b_2 + 155.3b_3$$
$$1480 = 114b_0 + 155.3b_2 + 1522b_3$$

- O sistema tem solução, cujos valores são: $b_0 = 16.4064; b_2 = -8.2476; b_3 = 0.5851$.

$$\hat{Y} = 16{,}4 - 8{,}25 * X_2 + 0{,}59 * X_3$$

Regressão Linear Multivariada

- Em R (“linear model”):

```
regressão <- lm(Vendas ~ Preço + Publicidade, data = ex0)
```

1. Soma dos Erros Quadráticos

- Soma dos erros quadráticos (SSE - *sum of squares for errors*), é a diferença entre os pontos e a curva de regressão.
- Permite aferir quanto que a curva se adere aos dados.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

1. Soma dos Erros Quadráticos

Y	X_2	X_3	$\hat{Y}$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
10,0	1,3	9,0	10,95	-0,95	0,90
6,0	2,0	7,0	4,01	1,99	3,97
5,0	1,7	5,0	5,31	-0,31	0,10
12,0	1,5	14,0	12,23	-0,23	0,05
10,0	1,6	15,0	11,99	-1,99	3,95
15,0	1,2	12,0	13,53	1,47	2,16
5,0	1,6	6,0	6,72	-1,72	2,96
12,0	1,4	10,0	10,71	1,29	1,66
17,0	1,0	15,0	16,94	0,06	0,00
20,0	1,1	21,0	19,62	0,38	0,14
112,00					**15,90**
11,2					**SSE**

2. Erro Padrão da Estimativa

- O erro padrão da estimativa consiste no valor padrão pelo qual o valor real difere do valor estimado pela regressão.
- O erro médio é zero.
- Se o erro padrão σ_ε for baixo, os erros tenderão a ficar próximos de zero, indicando que o modelo de regressão se adere aos dados. Também indicará que o uso de um modelo linear é válido.
- Um estimador de σ_ε pode ser dado por s_ε.

2. Erro Padrão da Estimativa

$$s_{y.x} = s_\varepsilon = \sqrt{\frac{\sum\left(Y - \hat{Y}\right)^2}{n-k}} = \sqrt{\frac{SSE}{n-k}}$$

- k – número de parâmetros sendo estimados pelo modelo
- $s_\varepsilon = \sqrt{\frac{15{,}90}{7}} = 1{,}51$

b_0
b_1
b_2

- Em R: `summary(regressão)`

```
Residual standard error: 1.507 on 7 degrees of freedom
```

3. Teste de Hipótese Coeficiente Angular

- Quando não houver relação linear entre duas variáveis, os coeficientes angulares das variáveis independentes será igual a 0.
- Inferência sobre β_2 e β_3 por meio de um teste de hipótese, para cada um dos coeficientes (b_2 e b_3):
- $H_0: \beta_2 = 0$
- $H_1: \beta_2 \neq 0$
- A estatística de teste é: $t = \frac{b_2 - \beta_2}{s_{b_2}}$

4. Análise de Variação dos Resíduos

- É importante conhecer qual é a variação na variável dependente Y que está associada à variação na variável dependente X.
- Vamos considerar a variação dos valores de Y em torno de sua média $\bar{Y}$ (chamaremos de SSTO - *total sum of squares*).

$$SSTO = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

4. Análise de Variação dos Resíduos

- Vamos separar o valor de SSTO em duas componentes: uma será a variação dos valores previstos pelo modelo de regressão $\hat{Y}$ em relação à média $\bar{Y}$ (SSR – *sum of squares due regression*); a outra medirá a variação dos valores em relação aos valores previstos (SSE – *sum of squares due to error*), que é uma variação não explicada pelo modelo.

4. Análise de Variação dos Resíduos

$$\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Variação total = variação explicada pelo modelo de previsão + variação não explicada pelo modelo.

4. Análise de Variação dos Resíduos

Y	X_2	X_3	$Y - \bar{Y}$	$(Y - \bar{Y})^2$	$\hat{Y}$	$(\hat{Y} - \bar{Y})$	$(\hat{Y} - \bar{Y})^2$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
10,0	1,3	9,0	-1,20	1,44	10,95	-0,25	0,06	-0,95	0,90
6,0	2,0	7,0	-5,20	27,04	4,01	-7,19	51,74	1,99	3,97
5,0	1,7	5,0	-6,20	38,44	5,31	-5,89	34,68	-0,31	0,10
12,0	1,5	14,0	0,80	0,64	12,23	1,03	1,05	-0,23	0,05
10,0	1,6	15,0	-1,20	1,44	11,99	0,79	0,62	-1,99	3,95
15,0	1,2	12,0	3,80	14,44	13,53	2,33	5,43	1,47	2,16
5,0	1,6	6,0	-6,20	38,44	6,72	-4,48	20,06	-1,72	2,96
12,0	1,4	10,0	0,80	0,64	10,71	-0,49	0,24	1,29	1,66
17,0	1,0	15,0	5,80	33,64	16,94	5,74	32,89	0,06	0,00
20,0	1,1	21,0	8,80	77,44	19,62	8,42	70,92	0,38	0,14
112,00				**233,60**			**217,70**		**15,90**
11,2				**SSTO**			**SSR**		**SSE**

4. Análise de Variação dos Resíduos

- **Coeficiente de determinação**

$$r^2 = 1 - \frac{SSE}{SSTO} = \frac{SSR}{SSTO} = \frac{217{,}7}{233{,}6} = 0{,}932$$

- r^2 indica a proporção da variação em Y que pode ser explicada pela variação em X_2 e X_3.
- Aproximadamente 93,2% das vendas podem ser explicadas pela variação de preço e do valor gasto em publicidade, enquanto que 6,8% das vendas são atribuídas a outros fatores.

5. Usando o Modelo para Previsão

- Para o preço $X_2 = 1{,}50$ e para o orçamento de publicidade (em milhares de R$) $X_3 = 10$ prever a quantidade a ser vendida.
- $\hat{Y} = 16.4064 - 8.2476X_2 + 0.5851X_3 = 16.4064 - 8.2476(1{,}50) + 0.5851(10) = 14{,}06.$

Exemplo de Colinearidade

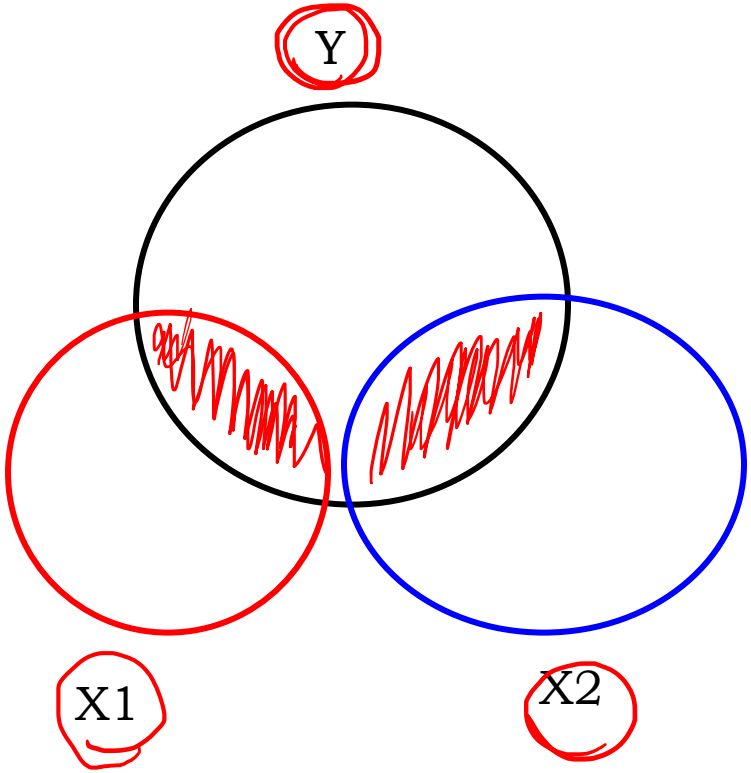
Colinearidade

- A colinearidade ocorre quando uma ou mais variáveis independentes são altamente relacionadas.

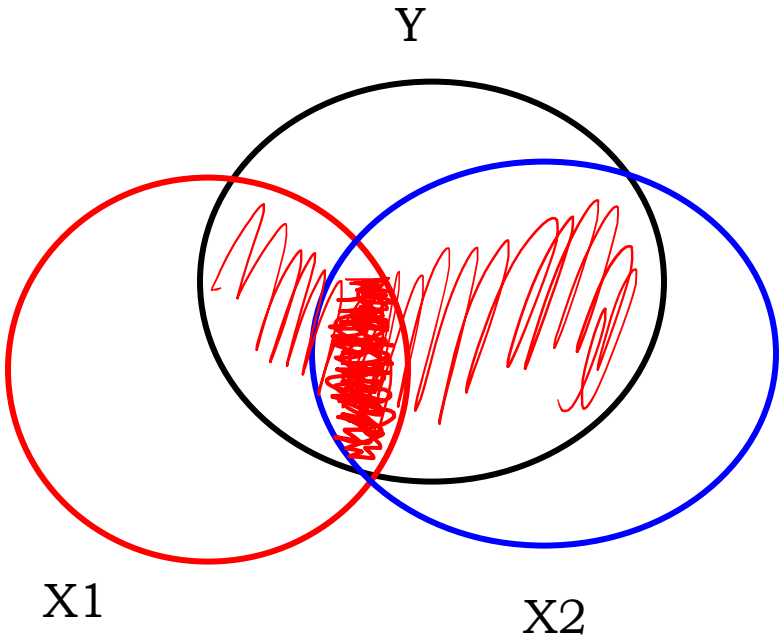
$$r_{x_j,x_k} = \frac{\sum_{i=1}^{n}(x_{ji} - \bar{x}_j)(x_{ki} - \bar{x}_k)}{(n-1)s_{x_j}s_{x_k}}.$$

s_{x_j} e s_{x_k} são os desvios padrão de X_j e X_k, respectivamente.

Colinearidade & Diagramas de Venn



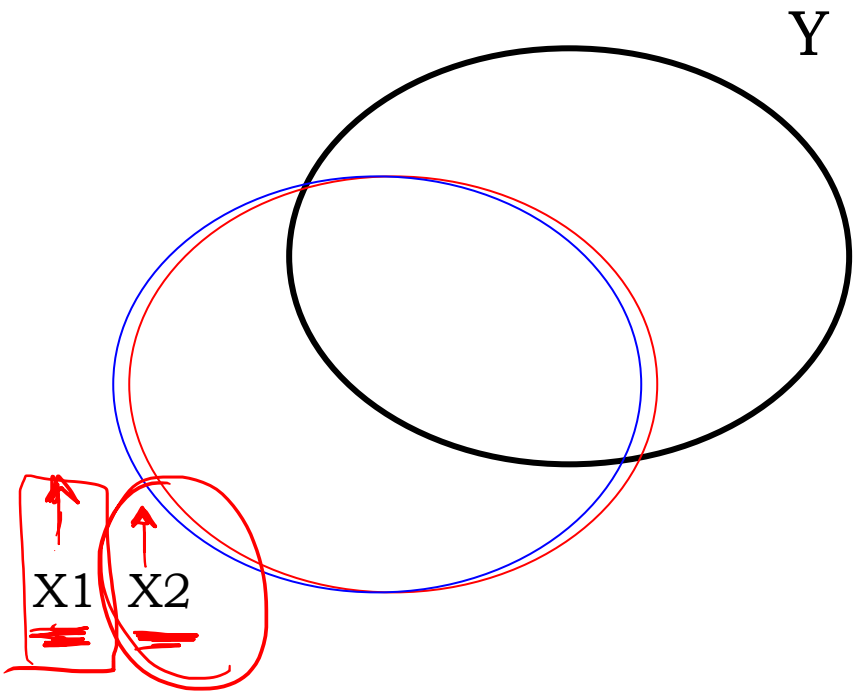
Não existe correlação entre X1 e X2. Cada variável explica parte da variação de Y.



Existe correlação entre X1 e X2. Parte da variação de Y pode ser atribuída a cada variável.

Fonte: utminers.utep.edu/crboehmer/Multivariate%20Regression.ppt

Colinearidade Perfeita & Diagrama de Venn



X2 é função de X1. Portanto, incluir X2 no modelo de previsão não irá explicar nenhuma variação de Y que é atribuída a X1.

Fonte: utminers.utep.edu/crboehmer/Multivariate%20Regression.ppt

Colinearidade

- A colinearidade tem como consequência:
  1. À medida que novas variáveis são acrescentadas ao modelo, o coeficiente de uma variável pode mudar de sinal, adquirindo sinal contrário ao de sua correlação.
  2. As estimativas dos coeficientes dos modelos de regressão variam grandemente de amostra para amostra.
  3. Torna-se difícil o entendimento e a interpretação da contribuição de cada variável junto à variável dependente.

Colinearidade - Exemplo

- Considere o seguinte exemplo que procura explicar o gasto anual de uma família com alimentação (Y), em centenas de reais, por meio das variáveis independentes de renda anual em milhares de reais (X_2) e tamanho da família (X_3).

Colinearidade - Exemplo

Família	Gasto Anual (Y)	Renda Anual (X_2)	Tamanho da Família (X_3)
A	24	11	6
B	8	3	2
C	16	4	1
D	18	7	3
E	24	9	5
F	23	8	4
G	11	5	2
H	15	7	2
I	21	8	3
J	20	7	2

Colinearidade - Exemplo

- `cor(ex1)`

```
          Y        X2        X3
Y  1.0000000 0.8838733 0.7368500
X2 0.8838733 1.0000000 0.8666181
X3 0.7368500 0.8666181 1.0000000
```

matriz de colinearidade

Colinearidade - Exemplo

- Apesar de ambas variáveis serem correlacionadas de forma positiva com Y, o coeficiente de X_3 é negativo, sugerindo que com o aumento de uma pessoa na família, o gasto com alimentação cai independente da renda familiar, o que é inconsistente.

Procedimento para Seleção de Variáveis Independentes

Seleção de Variáveis – Exemplo

- Um dos canais de distribuição de uma empresa é a venda direta, por meio de uma equipe de vendedores que visitam os clientes. A empresa está interessada em prever se um funcionário será um bom vendedor, e para isso pretende confrontar o nível de vendas do primeiro mês de trabalho (Y) com as seguintes variáveis: nota no teste de aptidão para vendas (X_2), idade do vendedor (X_3), nota em uma prova de ansiedade (X_4), anos de experiência anterior (X_5), e média ponderada do ensino médio em escala de 0 a 5 (X_6).

Seleção de Variáveis – Exemplo

- Identifique quais as variáveis mais relevantes para explicar o desempenho da equipe de vendas.

Seleção de Variáveis (Dados 1/2)

Unidades Vendidas em Mês de Experiência (Y)	Teste de Aptidão (X_2)	Idade (X_3)	Teste de Ansiedade (X4)	Experiência (anos) (X_5)	Média Ponderada Ensino Médio (X_6)
44	10	21,1	4,9	0	2,4
47	19	22,5	3,0	1	2,6
60	27	23,1	1,5	0	2,8
71	31	24,0	0,6	3	2,7
61	64	22,6	1,8	2	2,0
60	81	21,7	3,3	1	2,5
58	42	23,8	3,2	0	2,5
56	67	22,0	2,1	0	2,3
66	48	22,4	6,0	1	2,8
61	64	22,6	1,8	1	3,4
51	57	21,1	3,8	0	3,0
47	10	22,5	4,5	1	2,7
53	48	22,2	4,5	0	2,8
74	96	24,8	0,1	3	3,8
65	75	22,6	0,9	0	3,7

Seleção de Variáveis (Dados 2/2)

Unidades Vendidas em Mês de Experiência (Y)	Teste de Aptidão (X_2)	Idade (X_3)	Teste de Ansiedade (X4)	Experiência (anos) (X_5)	Média Ponderada Ensino Médio (X_6)
33	12	20,5	4,8	0	2,1
54	47	21,9	2,3	1	1,8
39	20	20,5	3,0	2	1,5
52	73	20,8	0,3	2	1,9
30	4	20,0	2,7	0	2,2
58	9	23,3	4,4	1	2,8
59	98	21,3	3,9	1	2,9
52	27	22,9	1,4	2	3,2
56	59	22,3	2,7	1	2,7
49	23	22,6	2,7	1	2,4
63	90	22,4	2,2	2	2,6
61	34	23,8	0,7	1	3,4
39	16	20,6	3,1	1	2,3
62	32	24,4	0,6	3	4,0
78	94	25,0	4,6	5	3,6

Seleção de Variáveis

- `cor(ex2)`

	Y	X2	X3	X4	X5	X6
Y	1.0000000	0.6761204	0.8119558	-0.2958598	0.5498340	0.6217841
X2	0.6761204	1.0000000	0.2566494	-0.2219880	0.3496392	0.3177716
X3	0.8119558	0.2566494	1.0000000	-0.3192239	0.5575954	0.6969216
X4	-0.2958598	-0.2219880	-0.3192239	1.0000000	-0.2786892	-0.2443816
X5	0.5498340	0.3496392	0.5575954	-0.2786892	1.0000000	0.3121288
X6	0.6217841	0.3177716	0.6969216	-0.2443816	0.3121288	1.0000000

Seleção de Variáveis – Método Exaustivo

- O método exaustivo consiste em gerar todas as possíveis combinações de variáveis, e avaliar o coeficiente de determinação resultante.
- Para n variáveis independentes, haverá 2^n possibilidades.
- Será solução a combinação que resultar no maior coeficiente de determinação.
- Característica desejável: menor número possível de variáveis, visando garantir a maior independência entre as variáveis.

Seleção de Variáveis (1/3)

Variáveis Independentes					# Parâmetros	Graus de Liberdade	r^2
X_2					2	28	0,457
X_3					2	28	0,637
X_4					2	28	0,088
X_5					2	28	0,302
X_6					2	28	0,387
X_2	X_3				3	27	0,8948
X_2	X_4				3	27	0,479
X_2	X_5				3	27	0,569
X_2	X_6				3	27	0,641
X_3	X_4				3	27	0,642

Seleção de Variáveis (2/3)

Variáveis Independentes					# Parâmetros	Graus de Liberdade	r^2
X_3	X_5				3	27	0,657
X_3	X_6				3	27	0,646
X_4	X_5				3	27	0,324
X_4	X_6				3	27	0,409
X_5	X_6				3	27	0,527
X_2	X_3	X_4			4	26	0,8951
X_2	X_3	X_5			4	26	0,8948
X_2	X_3	X_6			4	26	0,8953
X_2	X_4	X_5			4	26	0,575
X_2	X_4	X_6			4	26	0,646

Seleção de Variáveis (3/3)

Variáveis Independentes					# Parâmetros	Graus de Liberdade	r^2
X_2	X_5	X_6			4	26	0,701
X_3	X_4	X_5			4	26	0,659
X_3	X_4	X_6			4	26	0,65
X_3	X_5	X_6			4	26	0,669
X_4	X_5	X_6			4	26	0,531
X_2	X_3	X_4	X_5		5	25	0,8951
X_2	X_3	X_4	X_6		5	25	0,8955
X_2	X_3	X_5	X_6		5	25	0,8953
X_2	X_4	X_5	X_6		5	25	0,701
X_3	X_4	X_5	X_6		5	25	0,671
X_2	X_3	X_4	X_5	X_6	6	24	0,8955

Seleção de Variáveis – Método Exaustivo

- Tabela resumo, contendo a melhor seleção de variáveis (de 1 a 5).

Variáveis Independentes					# Parâmetros	Graus de Liberdade	r^2
X_3					2	28	0,637
X_2	X_3				3	27	0,8948
X_2	X_3	X_6			4	26	0,8953
X_2	X_3	X_4	X_6		5	25	0,8955
X_2	X_3	X_4	X_5	X_6	6	24	0,8955

Seleção de Variáveis – Método Aditivo

- O método aditivo consiste em selecionar, em cada iteração, uma variável independente, que resulte no maior valor da estatística t (em módulo) para o coeficiente angular.
- Na primeira iteração, a escolha coincide com a variável com o maior coeficiente de correlação. A variável escolhida será designada de X_2, e o modelo de regressão será $\hat{Y} = b_{0(1)} + b_{2(1)}X_2$.

Seleção de Variáveis – Método Aditivo

- Na iteração subsequente, a variável X_2 é mantida, e será escolhida a variável X_3 que resulte no maior valor da estatística t (em módulo) para o coeficiente angular b_3. Se o teste indicar que o coeficiente é diferente de zero, para o nível de significância considerado, então a expressão para a ser expressa por $\hat{Y} = b_{0(2)} + b_{2(2)}X_2 + b_{3(2)}X_3$.
- OBS: Os coeficientes dos termos mantidos de uma iteração para a outra provavelmente irão variar. Assim $b_{0(1)} \neq b_{0(2)}$, e assim por diante...

Seleção de Variáveis – Método Aditivo

- Continuar o procedimento até que não haja uma variável a ser adicionada, que possua coeficiente angular significativamente diferente de zero.

Variáveis “Dummy”

Variáveis "Dummy" para Categorias

- Quando os dados representarem uma determinada categoria ao invés de uma variável quantitativa, estes poderão ser substituídos por valores numéricos.
- Exemplo: para diferenciar o desempenho entre dois terminais A e B. Fazer com que uma determinada variável X seja 1 quando os dados se referirem ao terminal A, e 0 em caso contrário (terminal B).

	T_1	T_2	T_3
trim 1	1	0	0
trim 2	0	1	0
trim 4	0	0	0

Exercícios

Exercício 1

- Uma empresa de energia necessita prever o consumo de energia para o 3º e o 4º trimestre com base no consumo registrado nos últimos 16 anos.
- O modelo de regressão é baseado em categorias indicativas do trimestre em questão, estruturadas da seguinte maneira: S_2, S_3, S_4 - serão 1, se o trimestre t for, respectivamente, o 1º, 2º e 3º trimestre de um ano, e 0 em caso contrário.
- Curva de regressão: $\hat{Y} = b_0 + b_2 S_2 + b_3 S_3 + b_4 S_4$.

Exercício 1

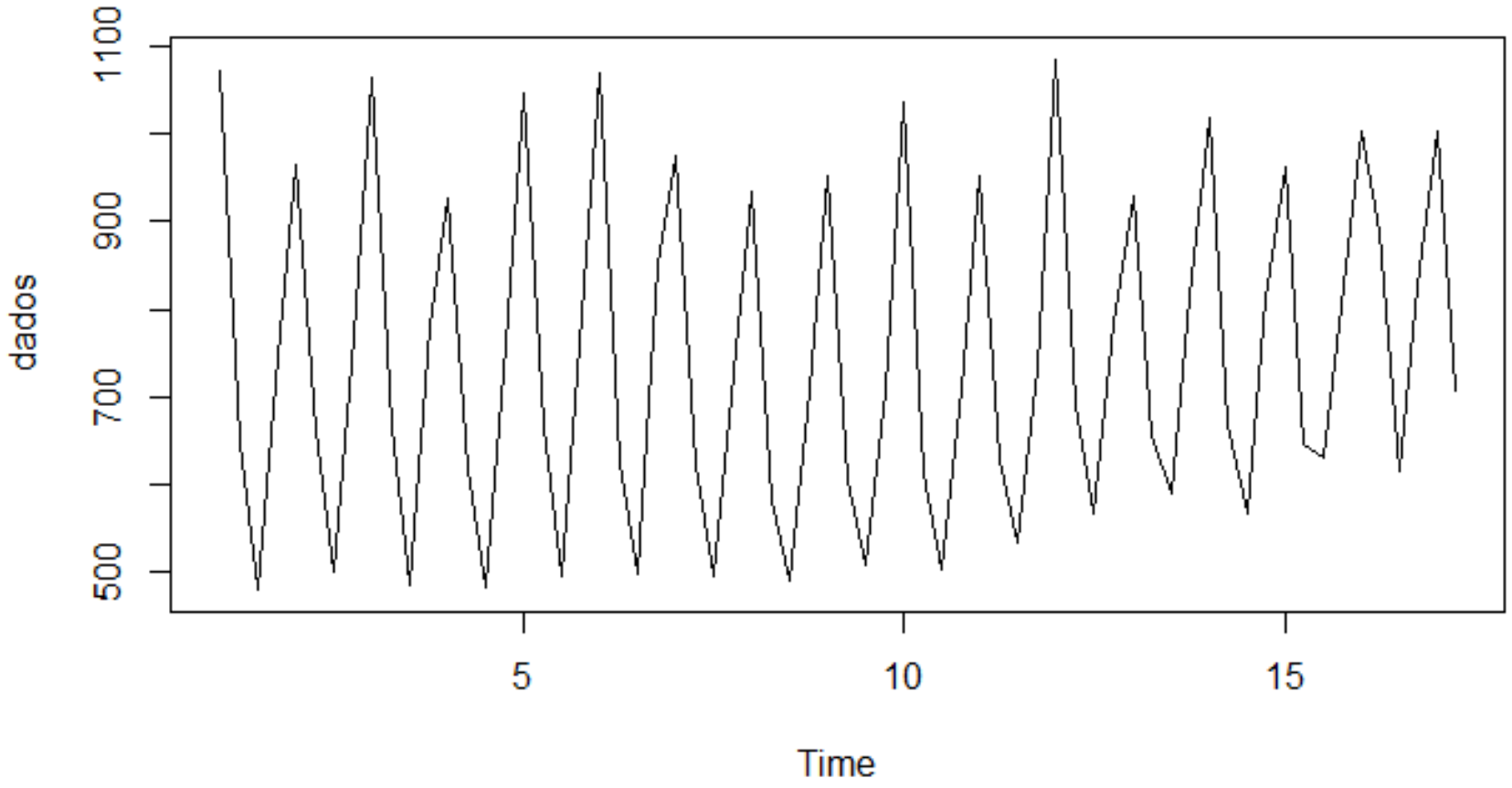
Ano	Trim.	kWh
2001	1	1071
	2	648
	3	480
	4	746
2002	1	965
	2	661
	3	501
	4	768
2003	1	1065
	2	667
	3	486
	4	780
2004	1	926
	2	618
	3	483
	4	757

Ano	Trim.	kWh
2005	1	1047
	2	667
	3	495
	4	794
2006	1	1068
	2	625
	3	499
	4	850
2007	1	975
	2	623
	3	496
	4	728
2008	1	933
	2	582
	3	490
	4	708

Ano	Trim.	kWh
2009	1	953
	2	604
	3	508
	4	708
2010	1	1036
	2	612
	3	503
	4	710
2011	1	952
	2	628
	3	534
	4	733
2012	1	1085
	2	692
	3	568
	4	783

Ano	Trim.	kWh
2013	1	928
	2	655
	3	590
	4	814
2014	1	1018
	2	670
	3	566
	4	811
2015	1	962
	2	647
	3	630
	4	803
2016	1	1002
	2	887
	3	615
	4	828
2017	1	1003
	2	706

Exercício 1



Exercício 1

- Estruturação dos dados, fazendo uso das variáveis $S_2, S_3, S_4 \ldots$

Ano	Trim.	kWh (Y)	S_2	S_3	S_4
2001	1	1071	1	0	0
	2	648	0	1	0
	3	480	0	0	1
	4	746	0	0	0
2002	1	965	1	0	0
	2	661	0	1	0
	3	501	0	0	1
	4	768	0	0	0
2003	1	1065	1	0	0
	2	667	0	1	0
	3	486	0	0	1
	4	780	0	0	0
2004	1	926	1	0	0
	2	618	0	1	0
	3	483	0	0	1
	4	757	0	0	0

Exercício 2

- Para o exemplo 2 relativo à seleção de vendedores, realize o processo aditivo de seleção de variáveis independentes.