LCE 5861 - MODELOS LINEARES I

Prof. César Gonçalves de Lima (FZEA/USP)

cegdlima@usp.br

Baseado no livro:

Rencher, A. C.; Schaalje, G. B. (2008). *Linear Models in Statistics*, 2nd edition. New York: Wiley

"... *discute modelos lineares clássicos sob uma perspectiva de álgebra matricial, tornando o assunto acessível para leitores iniciantes neste assunto*".

Conteúdo do curso:

1. Introdução.
2. **Álgebra de matrizes** (Curso de verão)
3. Vetores e matrizes aleatórios.
4. Distribuição normal multivariada.
5. Distribuição de formas quadráticas.
6. Regressão linear simples.
7. Regressão linear múltipla: estimação.
8. Regressão múltipla: testes de hipóteses e intervalos de confiança.
9. Regressão múltipla: validação do modelo e diagnóstico.

10. Regressão múltipla: x's aleatórios.
11. Regressão múltipla: inferência bayesiana.
12. Modelos de análise de variância.
13. Análise de variância com um fator: caso balanceado.
14. Análise de variância com dois fatores: caso balanceado.
15. Análise de variância: dados desbalanceados.
16. Análise de covariância.
17. Introdução aos modelos lineares mistos.

BIBLIOGRAFIA COMPLEMENTAR:

Draper, N.R. and Smith, H. (1998). *Applied Regression Analysis*. 3ht Edition. Wiley, New York.

Graybill, F.A. (2000). *Theory and Application of the Linear Model*, Duxbury Press.

Faraway, J.J. (2004). *Linear Models with R (Texts in Statistical Science)*. London: Chapman & Hall/CRC.

Faraway, J.J. (2006).*Extending the Linear Model with R*: *generalized linear, mixed effects and nonparametric regression models*. London: Chapman & Hall/CRC.

Harville, D.A. (2000). *Matrix Algebra from a Statistician's Perspective*. 3ht Edition. New York: Springer-Verlag.

Hocking, R.R. (2003). *Methods and Applications of Linear Models*: *Regression and the Analysis of Variance*. 3h ed. New York. Wiley-Interscience.

Littell, R. C; Stroup, W. W; Freund, R. J. (2002) *SAS for Linear Models*. 4h ed. Wiley Series, SAS Institute.

Morrison, D.F. (1983). *Applied Linear Statistical Methods*. Englewood Cliffs, NJ: Addison-Wesley.

Neter, J.; Kutner, M.H.; Nachtsheim, C.; Wasserman,W. (2004). *Applied Linear Statistical Models*. 5ht edition. McGraw Hill/ Irwin.

Searle, S. R. (1997). *Linear Models.* New York : Wiley

Searle, S.R., (2006). *Linear models for unbalanced data*. New York : Wiley-Interscience.

"***Essentially, all models are wrong, but some are useful***"

Box, G. E. P.; Draper, N. R. (1987). Empirical Model-Building and Response Surfaces, p. 424, Wiley. ISBN 0471810339.

CAPÍTULO 1 – INTRODUÇÃO

Os métodos estatísticos são amplamente usados como parte do processo de aprendizagem do método científico.

Na Biologia, Física, Ciências Sociais, Ciências Agrárias, como também no agronegócio e engenharias, os **modelos lineares** são úteis nos estágios de planejamento da pesquisa e na análise dos dados resultantes dela.

Nas seções 1.1, 1.2 e 1.3 vamos apresentar uma breve introdução aos:

- **Modelos de regressão linear simples**
- **Modelos de regressão linear múltipla**
- **Modelos de análise de variância**.

1.1. MODELO DE REGRESSÃO LINEAR SIMPLES

Objetivo: Modelar a relação de dependência (causal) entre duas variáveis quantitativas por meio de uma reta.

Para esta relação linear, nós usamos um modelo da forma:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (1.1)$$

Onde:

- y é a variável dependente ou **variável resposta**.
- x é a variável independente ou **variável preditora**.
- ε é a variação aleatória ($erro$) que ocorre na variável resposta, y, e que não foi explicada pela sua relação com a variável preditora, x.

Erro não significa engano ou equívoco. É um termo estatístico que representa as flutuações aleatórias nas medidas feitas nos indivíduos, que evidenciam o efeito de fatores não controlados ou são erros de medidas.

Exemplos de uso da regressão linear simples:

Variável resposta (y)	Variável preditora (x)
Salário atual	Número de anos de educação
Ganho de peso	Consumo de alimento
Temperatura de ebulição da água	Altitude
Pressão arterial	Dose de uma droga
Produção de uma gramínea	Quantidade de adubo aplicado

Em todos os exemplos:

- A linearidade do modelo em (1.1) é uma suposição.
- Podemos adicionar **suposições** sobre a distribuição probabilística do erro (ε) e a independência dos valores observados de y. Geralmente assumimos que $\varepsilon \sim N(0, \sigma^2)$
- Usando valores observados de x e y, nós **estimamos** β_0 e β_1 e fazemos **inferências**, como: calcular intervalos de confiança e realizar testes de hipóteses sobre esses parâmetros (Capítulo 6).
- Também podemos usar o modelo estimado para **prever** o valor da resposta y para um particular valor de x.

Exemplo: Relação entre a quantidade de nitrito e absorbância em amostras de mortadela.

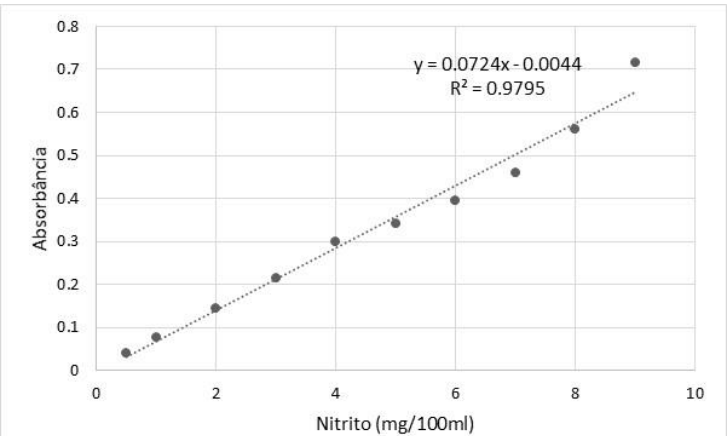

Gráfico de dispersão de absorbância e quantidade de nitrito

1.2. MODELO DE REGRESSÃO LINEAR MÚLTIPLA

Situação: A variável resposta y é influenciada por mais de uma variável preditora (x).

Exemplo: A produção de uma gramínea pode depender (somente) das quantidades oferecidas de nitrogênio, fósforo e potássio.

De um modo geral, essas quantidades de nutrientes são controladas, definidas ou escolhidas pelo pesquisador.

A produção também depende de outras variáveis que não podem ou não serão controladas pelo pesquisador, como aquelas associadas ao ambiente, por exemplo.

Um **modelo linear de regressão múltipla** relacionando y com diversas variáveis preditoras tem a forma geral:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \qquad (1.2)$$

em que os parâmetros β_0, β_1, β_2,..., β_k são chamados **coeficientes parciais de regressão**.

Note que o modelo (1.2) é linear nos parâmetros β's, mas existem modelos que **não são lineares** nos β's.

Exemplo: O modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_4 sen(x_2) + \varepsilon$$

é linear nos parâmetros, mas o modelo

$$y = \beta_0 + \beta_1 e^{\beta_2 (x - x_0)} + \varepsilon$$

é não linear nos parâmetros.

- Um modelo fornece uma **estrutura teórica** que possibilita um melhor entendimento do fenômeno de interesse.
- É uma construção matemática que pode representar bem o mecanismo que gerou as observações que temos em mãos, como a solução de uma equação diferencial, por exemplo.
- Em muitos casos o modelo é a simplificação idealizada de uma situação real e muito complexa.
- Os modelos empíricos fornecem **aproximações úteis** das relações complexas existentes entre as variáveis.
- Essas relações podem ser associativas ou causais, mas somente as relações causais serão estudadas nas próximas aulas!

Os modelos de regressão são usados para:

- **Predição:** Boas estimativas dos parâmetros individuais $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ são necessárias para fazer uma boa predição de y.

- **Seleção de variáveis:** A ênfase do estudo está na determinação da importância de cada variável preditora, x_i, em modelar a variação em y.

  Neste caso as variáveis preditoras que explicam uma quantidade importante de variação em y são mantidas no modelo e aquelas que contribuem pouco podem/devem ser descartadas.

Estimação e procedimentos inferenciais serão discutidos nos Capítulos 7 a 10.

Exemplo 4. Num estudo com framboesas realizado na Seção de Horticultura do ISA foram analisados frutos de 14 plantas com o intuito de estudar a relação entre y: teor de sólidos solúveis brix (em graus Brix) e

x_1: diâmetro do fruto (em cm); x_2: altura do fruto (em cm);

x_3: peso do fruto (em g) e x_4: pH do fruto

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	7.11286	0.56301	12.63	<.0001
x1	1	1.73566	0.57318	3.03	0.0143
x2	1	-1.11733	0.44476	-2.51	0.0332
x3	1	-0.24050	0.17679	-1.36	0.2068
x4	1	0.26444	0.24008	1.10	0.2993

Usando o método *Stepwise* de seleção de variáveis:

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	F Value	Pr > F
1	x1		1	0.4314	0.4314	9.10	0.0107
2	x2		2	0.3186	0.7499	14.01	0.0032

Parameter Estimates (MODELO FINAL)

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	6.62693	0.43702	6.46684	229.95	<.0001
x1	2.18960	0.39323	0.87198	31.01	0.0002
x2	-1.39849	0.37358	0.39411	14.01	0.0032

O modelo final ajustado fica:

$$\hat{y} = 6{,}6269 + 2{,}1896x_1 - 1{,}3985x_2$$

Com este modelo podemos estimar o teor de sólidos solúveis brix (y) das framboesas, medindo somente o seu diâmetro (x_1) e a sua altura (x_2).

O coeficiente de determinação ajustado, $R^2_{aj} = 0{,}70$, indica que o diâmetro e a altura explicam 70% da variabilidade do teor de sólidos solúveis brix das framboesas.

O modelo completo, com as 4 variáveis regressoras, apresenta um $R^2_{aj} = 0{,}71$, indicando que as variáveis x_3 (peso) e x_4 (pH) são pouco importantes para estimar o teor de sólidos solúveis brix (y) das framboesas

Outras estatísticas usadas na seleção de modelos de regressão múltipla e que serão estudadas em um próximo seminário:

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	BIC	Variables in Model
4	**0.7093**	**0.7988**	5.0000	-46.4140	-39.4757	x1 x2 x3 x4
2	0.7045	0.7499	**3.1838**	**-47.3726**	**-43.9885**	**x1 x2**
3	0.7031	0.7716	4.2132	-46.6436	-41.9523	x1 x2 x3
3	0.6846	0.7574	4.8505	-45.7961	-41.6426	x1 x2 x4
3	0.5550	0.6577	9.3112	-40.9750	-39.6747	x1 x3 x4
2	0.5494	0.6187	9.0528	-41.4668	-40.6050	x1 x3
2	0.4876	0.5665	11.3901	-39.6685	-39.4909	x3 x4
3	0.4719	0.5938	12.1696	-38.5786	-38.5198	x2 x3 x4
1	0.3840	0.4314	15.4316	-37.8712	-38.0734	x1
2	0.3714	0.4681	15.7894	-36.8058	-37.6135	x2 x4
2	0.3284	0.4317	17.4156	-35.8801	-36.9785	x1 x4
1	0.2752	0.3310	19.9235	-35.5941	-36.2890	x4
1	-.0345	0.0451	32.7080	-30.6137	-32.2062	x2
1	-.0767	0.0061	34.4517	-30.0535	-31.7323	x3
2	-.1090	0.0616	33.9707	-28.8575	-31.7444	x2 x3

1.3. MODELOS DE ANÁLISE DE VARIÂNCIA (ANOVA)

Situação: Estamos interessados em comparar diversas populações (grupos) ou comparar diversas condições em um experimento.

Exemplo 1: Suponha que um pesquisador deseje comparar o rendimento (y) de quatro catalisadores em um processo industrial. Se n observações são feitas para cada catalisador, um modelo para as $4n$ observações pode ser expresso como:

$$y_{ij} = \mu_i + \varepsilon_{ij} \qquad (1.3)$$

$$i = 1, 2, 3, 4, \qquad j = 1, 2, \ldots, n$$

em que μ_i é o rendimento médio associado ao i-ésimo catalisador.

Dados de rendimento do processo industrial, por catalisador.

Catalisador			
1	2	3	4
y_{11}	y_{21}	y_{31}	y_{41}
y_{12}	y_{22}	y_{32}	y_{42}
…	…	…	…
y_{1n}	y_{2n}	y_{3n}	y_{4n}
$\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$	$\bar{y}_4$

Uma hipótese de interesse a ser testada é que os rendimentos médios (populacionais) dos 4 catalisadores são iguais entre si, ou seja,

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

Para julgar esta hipótese usaremos as médias amostrais $\bar{y}_1$, …, $\bar{y}_4$.

O modelo em (1.3) também pode ser expresso de uma forma alternativa (mais comum!), chamada **superparametrizada**:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad (1.4)$$

$$i = 1, 2, 3, 4, \qquad j = 1, 2, \ldots, n$$

Nesta forma, α_i é o **efeito** do i-ésimo catalisador na produção e a hipótese de interesse pode ser expressa como:

$$H_0: \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$$

Veremos detalhes dessas duas parametrizações em modelos com um e dois fatores de tratamento, nos Capítulos 12 a 15.

Exemplo 2. Um experimento de competição de variedades de cana-de-açúcar foi instalado em um delineamento inteiramente casualizado, com 5 repetições por tratamento. Foram obtidas as seguintes produções (ton/ha):

	CB5034	CB6245	IAC 6258	IAC 6529	IAC 6814	IAC 6538
	112,3	125,3	118,4	127,9	130,1	115,2
	121,0	119,7	120,5	128,3	122,4	123,2
	114,3	120,8	119,7	129,5	126,7	117,8
	115,8	120,5	118,3	126,5	127,3	120,8
	117,2	122,3	117,8	127,3	128,9	116,4
$y_{i\bullet}$	580,6	608,6	594,7	639,5	635,4	593,4
$\bar{y}_{i\bullet}$	116,12	121,72	118,94	127,90	127,08	118,68

Quadro de Análise de Variância (ANOVA)

Causa de Variação	g.l.	SQ	QM	F
Variedade	5	576,2480	115,2496	18,42 **
Resíduo	24	150,1440	6,2560	
Total	29	726,3920		

$s^2 = 6{,}2560$ $\bar{y}_{\bullet\bullet} = 121{,}74$ ton/ha $CV = 2{,}05\%$

Testar: $\begin{cases} H_0: \mu_1 = \mu_2 = \ \ldots \ = \mu_6 \\ H_1: \text{pelo menos duas médias diferem entre si.} \end{cases}$

Como $F_{calc} > F_{tab} = 2{,}62 \Rightarrow$rejeita-se H_0 ao nível $\alpha = 5\%$ e conclui-se que existem pelo menos duas variedades de cana-de-açúcar com produções médias diferentes entre si.

Complicando um pouco: Suponha que no Exemplo 1 o pesquisador também deseje comparar o efeito de três níveis de temperatura e n observações são tomadas em cada uma das $4*3 = 12$ combinações catalisador-temperatura. O que fazer?

Neste caso o modelo (de médias de casela) envolvendo os dois fatores pode ser expresso como:

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \tag{1.5}$$

Em que: y_{ijk} é o rendimento da k-ésima repetição submetida ao i-ésimo catalizador e j-ésima temperatura; μ_{ij} é a média da (ij)-ésima combinação catalisador-temperatura e ε_{ijk} é o erro experimental associado à observação y_{ijk}.

Na forma superparametrizada o modelo fica:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

$$i = 1, 2, 3, 4 \quad j = 1, 2, 3 \quad k = 1, 2, \ldots, n$$

em que

- α_i é o efeito do i-ésimo catalisador
- β_j é o efeito do j-ésimo nível de temperatura
- γ_{ij} é o efeito da interação do i-ésimo catalisador e j-ésimo nível de temperatura.

Note que em (1.5) o pesquisador **escolhe** os tipos de catalisador e os níveis de temperatura a serem comparados e aplica diferentes tratamentos aos objetos ou unidades experimentais sob estudo.

Rendimento do processo industrial, por catalisador e temperatura

Catalisador	Temperatura	Dados				Média
1	1	y_{111}	y_{112}	...	y_{11n}	$\bar{y}_{11}$
	2	y_{121}	y_{122}	...	y_{12n}	$\bar{y}_{12}$
	3	y_{131}	y_{132}	...	y_{13n}	$\bar{y}_{13}$
2	1	y_{211}	y_{212}	...	y_{21n}	$\bar{y}_{21}$
	2	y_{221}	y_{222}	...	y_{22n}	$\bar{y}_{22}$
	3	y_{231}	y_{232}	...	y_{23n}	$\bar{y}_{23}$
3	1	y_{311}	y_{312}	...	y_{31n}	$\bar{y}_{31}$
	2	y_{321}	y_{322}	...	y_{32n}	$\bar{y}_{32}$
	3	y_{331}	y_{332}	...	y_{33n}	$\bar{y}_{33}$
4	1	y_{411}	y_{412}	...	y_{41n}	$\bar{y}_{41}$
	2	y_{421}	y_{422}	...	y_{42n}	$\bar{y}_{42}$
	3	y_{431}	y_{432}	...	y_{43n}	$\bar{y}_{43}$

Comentários gerais:

Também podemos comparar as médias de variáveis medidas em grupos naturais de unidades, como em um estudo para avaliar os ganhos de peso dos machos e das fêmeas de certa espécie animal.

Modelos de análise de variância (ANOVA) serão tratados com detalhes nos Capítulos 12-15.

Tópicos adicionais relacionados aos modelos lineares, como análise de covariância e modelos mistos, serão cobertos nos Capítulos 16 e 17.

CAPÍTULO 2. ÁLGEBRA DE MATRIZES

Nas diversas fases de análise de dados realizadas usando **modelos lineares**, iremos precisar de um bom conhecimento de Álgebra de Matrizes e de Estatística Matemática

Sugestão: Reveja/estude os conceitos apresentados nas aulas do curso de verão sobre Álgebra de Matrizes.

CAPÍTULO 3. VETORES E MATRIZES ALEATÓRIOS

3.1. INTRODUÇÃO

A ***ANOVA*** (***AN****alysis* ***O****f* ***VA****riance*) é um processo aritmético de decomposição da variação total dos valores observados, que é expressa por meio de somas de quadrados.

Os parâmetros que se busca estimar e/ou comparar são **formas lineares** ou **formas quadráticas** das observações.

O termo **modelo linear** aparece em situações nas quais a média de uma variável aleatória y (variável resposta) pode ser expressa como uma **função linear** de $(p + 1)$ parâmetros desconhecidos:

$$E(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

As variáveis x_i são denominadas **variáveis explicativas**, **preditoras** ou **covariáveis** e os seus valores podem ser:

- **Dicotômicos** (0 ou 1): x é uma **variável indicadora** da presença ($x = 1$) ou ausência ($x = 0$) de uma determinada característica para aquela observação.

  Exemplo: Colunas da matriz de um delineamento experimental.

- **Estabelecidos** ou **fixados** pelo pesquisador que planeja a pesquisa e observa as respostas, y.

  Exemplo: Doses de um nutriente químico ou biológico.

- **Observados simultaneamente** com a resposta y. Neste caso, a variável x também é uma variável aleatória e é chamada de covariável.

  Exemplo: Quando se observa a produção (y) de certa cultura, também se observa a temperatura ambiente, o número de perfilhos, a pluviosidade, o nível de infestação por plantas daninhas *etc*., que são candidatas naturais a covariáveis, pois podem interferir na produção.

Vetor aleatório é um vetor cujos elementos são observações de uma variável aleatória.

Variável aleatória é aquela cujos valores resultam de um experimento aleatório.

Exemplos: Face obtida no lançamento de um dado, número sorteado na Megasena, peso de um frango de corte aos 42 dias de idade escolhido ao acaso de um grande galpão de produção, produção de uma parcela experimental de milho que recebeu certo adubo químico *etc*.

Formalmente, "**variável aleatória** é uma função que associa a cada elemento do espaço amostral um número em $\mathbb{R}$".

Podemos distinguir duas estruturas diferentes de vetores aleatórios:

1) **Um vetor $n\times1$ contendo os valores de uma única variável medida em cada um dos n diferentes indivíduos ou unidades experimentais.**

   Neste caso, admite-se que os n valores, $y_1, y_2, \ldots, y_n$, da variável aleatória y são não correlacionadas e têm a mesma variância.

   Exemplo: Consideremos o modelo de regressão múltipla em que a produção de certa forrageira é função das quantidades de nitrogênio (x_1), fósforo (x_2) e potássio (x_3) presentes no solo:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i,\ i = 1, \ldots, n$$

Tratando os x's como constantes (valores escolhidos pelo pesquisador), neste modelo de regressão linear múltipla nós temos somente dois vetores aleatórios:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \text{e} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{3.1}$$

onde os y_i's são as produções observadas, mas os erros ε_i's não são observados.

2) **Um vetor $p \times 1$ consistindo da medida de p diferentes variáveis feitas em um único indivíduo ou unidade experimental.**

Neste caso, admite-se que as p **variáveis aleatórias** podem ser correlacionadas e ter variâncias diferentes.

Exemplo: Em um estudo sobre desempenho de alunos na disciplina de Modelos Lineares foram avaliadas diversas características dos alunos: nota final na disciplina, idade, número de horas de estudo por semana, nota em Cálculo, em Matrizes e em Probabilidade *etc*.

De um modo geral, um vetor desse tipo é escrito como:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix}$$

3.2. MÉDIA, VARIÂNCIA, COVARIÂNCIA E CORRELAÇÃO.

Definição: Se $f(y)$ é a função densidade de probabilidade ($f.d.p.$) da variável aleatória contínua, y, a **média** (populacional) ou o **valor esperado** de y é definido como:

$$\mu = E(y) = \int_{-\infty}^{\infty} y\, f(y)\, dy \qquad (3.2)$$

O **valor esperado** (média) de uma **função da variável aleatória** y é definido como:

$$E[u(y)] = \int_{-\infty}^{\infty} u(y)\, f(y)\, dy \qquad (3.3)$$

e pode ser obtido diretamente, sem a necessidade de se obter a densidade de probabilidade de $u(y)$.

Propriedades importantes: Para uma constante $a \in \mathcal{R}$ e funções $u(y)$ e $v(y)$ segue de (3.3) que:

$$E(ay) = aE(y) \tag{3.4}$$

$$E[u(y) + v(y)] = E[u(y)] + E[v(y)] \tag{3.5}$$

A **variância** (populacional) de uma variável aleatória y é definida como

$$\sigma^2 = var(y) = E(y - \mu)^2 = \int_{-\infty}^{\infty} (y - \mu)^2 f(y)\, dy \tag{3.6}$$

A raiz quadrada da variância de y é conhecida como **desvio padrão** (populacional):

$$\sigma = \sqrt{var(y)} = \sqrt{E(y - \mu)^2} \tag{3.7}$$

Usando (3.4) e (3.5), a variância de y pode ser expressa na forma:

$$\sigma^2 = var(y) = E(y)^2 - [E(y)]^2 \qquad (3.8)$$

Propriedade: Se a é uma constante, podemos usar (3.4) e (3.6) para mostrar que:

$$var(ay) = a^2 var(y) = a^2\sigma^2 \qquad (3.9)$$

Para avaliar o grau de relacionamento entre duas variáveis y_i e y_j de um vetor aleatório $[y_1, \ldots, y_p]'$ definimos a **covariância** (populacional) entre y_i e y_j como:

$$\sigma_{ij} = cov(y_i, y_j) = E[(y_i - \mu_i)(y_j - \mu_j)] \qquad (3.10)$$

em que $\mu_i = E(y_i)$ e $\mu_j = E(y_j)$ são as médias das variáveis y_i e y_j, respectivamente.

Usando (3.4) e (3.5), $cov(y_i, y_j)$ pode ser expressa na forma:

$$\sigma_{ij} = cov(y_i, y_j) = E(y_i\, y_j) - \mu_i \mu_j \qquad (3.11)$$

Nota: Vale notar que σ_{ij} assume valores em $\mathcal{R}$ e se $\sigma_{ij} = 0$ dizemos que as variáveis y_i e y_j não são correlacionadas.

Definição: Duas variáveis aleatórias y_i e y_j são ditas **independentes** se a sua densidade conjunta puder ser **fatorada** no produto de suas densidades marginais:

$$f(y_i, y_j) = f_i(y_i) f_j(y_j) \qquad (3.12)$$

Em que a **densidade marginal** de y_i, $f_i(y_i)$, é definida como

$$f_i(y_i) = \int_{-\infty}^{\infty} f(y_i, y_j)\, dy_j$$

Propriedades:

1) **Se** y_i e y_j são independentes **então** $E(y_i y_j) = E(y_i)E(y_j)$ (3.13)

2) **Se** y_i e y_j são independentes **então** $\sigma_{ij} = cov(y_i, y_j) = 0$ (3.14)

No primeiro tipo de vetor aleatório definido na Seção 3.1, as variáveis y_1, y_2,…, y_n são tipicamente **independentes** se foram obtidas de uma **amostra aleatória** de uma população de indivíduos.

Exemplo: Vetor com as alturas de $n = 6$ alunos da disciplina.

Neste caso, a condição de independência implica não haver correlação entre as medidas, ou seja, implica admitir que $\sigma_{ij} = 0$ para todo $i \neq j$.

Geralmente, para as variáveis $y_1, y_2, \ldots, y_p$ do vetor aleatório do segundo tipo tem-se $\sigma_{ij} \neq 0$ para alguns valores de i e j.

- É importante reforçar que $cov(y_i, y_j) = 0$ não implica em independência entre y_i e y_j. [ver Exemplo 3.2, pág. 64-67]

A **esperança condicional** da variável aleatória y para um dado valor de x é definida como:

$$E(y \mid x) = \int y\, f(y|x)\, dy,$$

em que $f(y|x) = \dfrac{f(x,y)}{f_1(x)}$ é a densidade condicional de y dado x.

Se a esperança condicional **não depender** dos valores de x, podemos concluir que as variáveis aleatórias y e x são **independentes**.

Problema: Como a covariância σ_{ij} depende da escala de medida das duas variáveis y_i e y_j, é difícil saber se existe um alto grau de dependência entre as duas variáveis.

Para **padronizar** a covariância, σ_{ij}, nós dividimos o seu valor pelo produto dos desvios padrões de y_i e y_j, obtendo assim a **correlação linear (populacional)** entre y_i e y_j:

$$\rho_{ij} = corr(y_i, y_j) = \frac{\sigma_{ij}}{(\sigma_i)(\sigma_j)} \qquad (3.17)$$

em que $\sigma_i = dp(y_i)$ e $\sigma_j = dp(y_j)$. Vale notar que $-1 \leq \rho_{ij} \leq 1$

Importante: Valores ρ_{ij} próximos a 1 ou a -1 indicam forte relação de dependência linear.

3.3 VETOR DE MÉDIAS E MATRIZ DE COVARIÂNCIA PARA VETORES ALEATÓRIOS

3.3.1 Vetor de médias

O **valor esperado** de um vetor aleatório $\boldsymbol{y}$ $(p\times1)$ é definido como o vetor de valores esperados das p variáveis aleatórias y_1, y_2, …, y_p de $\boldsymbol{y}$:

$$E(\boldsymbol{y}) = E\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} E(y_1) \\ E(y_2) \\ \vdots \\ E(y_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} \tag{3.18}$$

em que $E(y_i) = \mu_i = \int_{-\infty}^{\infty} y_i\, f(y_i)\, dy_i$, usando a notação $f(y_i)$ para a densidade marginal da variável aleatória y_i.

Propriedade: Se $\boldsymbol{x}$ e $\boldsymbol{y}$ são dois vetores aleatórios de dimensões ($p \times 1$), segue de (3.18) que o valor esperado da soma desses vetores aleatórios é a soma de seus valores esperados:

$$E(\boldsymbol{x} + \boldsymbol{y}) = E(\boldsymbol{x}) + E(\boldsymbol{y}) \tag{3.19}$$

3.3.2 Matriz de covariâncias

As variâncias σ_1^2, σ_2^2, …, σ_p^2 de y_1, y_2, …, y_p e as covariâncias σ_{ij}, para todo $i \neq j$, podem ser convenientemente arranjadas em uma **matriz de covariâncias**, denotada por $\boldsymbol{\Sigma}$, $p \times p$, que tem a seguinte forma:

$$\boldsymbol{\Sigma} = cov(\boldsymbol{y}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} \quad (3.20)$$

- Quando as variáveis y's são variáveis aleatórias contínuas e não existe qualquer dependência linear entre essas variáveis, pode-se garantir que a matriz $\boldsymbol{\Sigma}$ é **simétrica** e **positiva definida**

- Se existir alguma relação linear entre as variáveis y's nós assumiremos que a matriz $\boldsymbol{\Sigma}$ é **positiva semidefinida**.

Por analogia com (3.18), nós definimos o valor esperado de uma **matriz aleatória $\boldsymbol{Z}$**, formada pelas medidas de p variáveis aleatórias (colunas) feitas em n indivíduos (linhas), como a matriz de valores esperados:

$$E(\boldsymbol{Z}) = E\begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix} = \begin{bmatrix} E(z_{11})\, E(z_{12}) & \cdots & E(z_{1p}) \\ E(z_{21})\, E(z_{22}) & \cdots & E(z_{2p}) \\ \vdots & \ddots & \vdots \\ E(z_{n1})\, E(z_{n2}) & \cdots & E(z_{np}) \end{bmatrix} \quad 3.21)$$

A matriz de covariâncias $\boldsymbol{\Sigma} = cov(\boldsymbol{y})$ em (3.20) pode ser expressa como o valor esperado de uma matriz aleatória.

- O (ij)-ésimo elemento da matriz $(\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})'$ é o produto da i-ésima linha $(\boldsymbol{y} - \boldsymbol{\mu})$ pela j-ésima coluna de $(\boldsymbol{y} - \boldsymbol{\mu})$, ou seja: $(y_i - \mu_i)(y_j - \mu_j)$.

  De (3.10) e (3.21) o (ij)-ésimo elemento de $E[(\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})']$ é $E(y_i - \mu_i)(y_j - \mu_j) = \sigma_{ij}$. Então:

$$\boldsymbol{\Sigma} = E[(\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})'] = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} \qquad (3.22)$$

  ou

$$\boldsymbol{\Sigma} = E(\boldsymbol{yy}') - \boldsymbol{\mu\mu}' \qquad (3.23)$$

  que é uma forma análoga a (3.8) e (3.11).

Vamos ilustrar (3.22) para $p = 2$:

$$\boldsymbol{\Sigma} = E[(\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})'] = E\left\{\begin{bmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \end{bmatrix}[y_1 - \mu_1, y_2 - \mu_2]\right\}$$

$$= E\begin{bmatrix} (y_1 - \mu_1)^2 & (y_1 - \mu_1)(y_2 - \mu_2) \\ (y_2 - \mu_2)(y_1 - \mu_1) & (y_2 - \mu_2)^2 \end{bmatrix}$$

$$= \begin{bmatrix} E(y_1 - \mu_1)^2 & E(y_1 - \mu_1)(y_2 - \mu_2) \\ E(y_2 - \mu_2)(y_1 - \mu_1) & E(y_2 - \mu_2)^2 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} var(y_1) & cov(y_1, y_2) \\ cov(y_2, y_1) & var(y_2) \end{bmatrix}$$

3.3.3. Variância generalizada

Uma **medida de variabilidade geral** na população dos $\boldsymbol{y}$'s pode ser definida como o determinante da matriz de variâncias e covariâncias do vetor aleatório $\boldsymbol{y}$:

$$\textbf{Variância generalizada} = det(\boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}| \qquad (3.24)$$

Ideia: Se $|\boldsymbol{\Sigma}|$ é um valor pequeno então os $\boldsymbol{y}$'s estão concentrados mais próximos do vetor de médias $\boldsymbol{\mu}$ do que quando $|\boldsymbol{\Sigma}|$ é um valor grande.

Problema: Um pequeno valor de $|\boldsymbol{\Sigma}|$ também pode indicar que as variáveis $y_1, y_2, \ldots, y_p$ são fortemente correlacionadas!

Neste caso, os $\boldsymbol{y}$'s tendem a ocupar um **subespaço** do espaço p-dimensional e $\boldsymbol{\Sigma}$ deixa de ser positiva definida.

3.3.4. Distância padronizada

Para obter uma medida útil da "distância entre um vetor aleatório $\boldsymbol{y}$ e o seu vetor de médias $E(\boldsymbol{y}) = \boldsymbol{\mu}$", sugere-se levar em conta as variâncias e as covariâncias dos y_i's do vetor aleatório $\boldsymbol{y}$.

Por analogia ao caso univariado, em que $z = (y - \mu)/\sigma$ tem média 0 e variância 1, a **distância padronizada de um vetor aleatório, $\boldsymbol{y}$, ao vetor de médias, $\boldsymbol{\mu}$**, é definida como:

$$\textbf{Distância padronizada} = (\boldsymbol{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) \qquad (3.25)$$

que é muitas vezes chamada **Distância de Mahalanobis**.

O uso de $\boldsymbol{\Sigma}^{-1}$ em (3.25) padroniza as variáveis y_i, que passam a ter média igual a zero, variância igual a 1 e todas elas passam a ser não correlacionadas entre si.

3.4. MATRIZ DE CORRELAÇÕES

A partir da matriz de covariâncias, $\mathbf{\Sigma}$, podemos obter **a matriz de correlações**, $\mathbf{P}_\rho$, que é definida como:

$$\mathbf{P}_\rho = (\rho_{ij}) = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix} \tag{3.26}$$

em que $\rho_{ij} = \dfrac{\sigma_{ij}}{\sqrt{\sigma_{ii}\ \sigma_{jj}}}$ é a correlação entre as variáveis y_i e y_j, como definido em (3.17).

Por exemplo: a segunda linha de $\mathbf{P}_\rho$, contém a correlação entre y_2 e cada uma das outras variáveis y_i's.

Definindo:

$$\mathbf{D}_\sigma = [diag(\mathbf{\Sigma})]^{1/2} = diag(\sigma_1, \sigma_2, \cdots, \sigma_p) \tag{3.27}$$

Por (2.31), podemos obter $\mathbf{P}_\rho$ a partir de $\mathbf{\Sigma}$ e vice-versa:

$$\mathbf{P}_\rho = \mathbf{D}_\sigma^{-1}\, \mathbf{\Sigma}\, \mathbf{D}_\sigma^{-1} \tag{3.28}$$

$$\mathbf{\Sigma} = \mathbf{D}_\sigma\, \mathbf{P}_\rho\, \mathbf{D}_\sigma \tag{3.29}$$

Exercício: Verifique as igualdades (3.28) e (3.29) para o caso de um vetor aleatório $\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$ com matriz de variâncias e covariâncias $\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$

3.5. VETOR DE MÉDIAS E MATRIZ DE COVARIÂNCIAS DE VETORES ALEATÓRIOS PARTICIONADOS

Suponha que o vetor aleatório $\boldsymbol{v}$ seja particionado em dois subconjuntos de variáveis, denotados por $\boldsymbol{y}$ $(p \times 1)$ e $\boldsymbol{x}$ $(q \times 1)$:

$$\boldsymbol{v} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_p \\ x_1 \\ \vdots \\ x_q \end{bmatrix}$$

Exemplo: O vetor $\boldsymbol{y}$ é formado por p variáveis zootécnicas (peso, consumo, conversão alimentar etc) e o vetor $\boldsymbol{x}$**,** por q variáveis de qualidade de carcaça (peso da carcaça fria, peso do contrafilé, do filé mignon etc.)

O vetor de médias e a matriz de covariâncias do vetor aleatório particionado $\boldsymbol{v} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{bmatrix}$ podem ser expressos como:

$$\boldsymbol{\mu} = E(\boldsymbol{v}) = E\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{bmatrix} = \begin{bmatrix} E(\boldsymbol{y}) \\ E(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{bmatrix} \tag{3.30}$$

$$\boldsymbol{\Sigma} = cov(\boldsymbol{v}) = cov\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix} \tag{3.31}$$

Em que:

- O subvetor $\boldsymbol{\mu}_y = [E(y_1), E(y_2), \ldots, E(y_p)]'$ contem as médias de $y_1, y_2, \ldots, y_p$ enquanto $\boldsymbol{\mu}_x$ contem as médias das variáveis x's.
- $\boldsymbol{\Sigma}_{yy} = cov(\boldsymbol{y})$ é uma matriz $p \times p$ de covariâncias de $\boldsymbol{y}$ contendo as variâncias de $y_1, y_2, \ldots, y_p$ na diagonal principal e as covariâncias σ_{ij} de cada y_i com cada y_j $(i \neq j)$ fora da diagonal, ou seja:

$$\boldsymbol{\Sigma_{yy}} = \begin{bmatrix} \sigma_{y_1}^2 & \sigma_{y_1 y_2} & \cdots & \sigma_{y_1 y_p} \\ \sigma_{y_2 y_1} & \sigma_{y_2}^2 & \cdots & \sigma_{y_2 y_p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{y_p y_1} & \sigma_{y_p y_2} & \cdots & \sigma_{y_p}^2 \end{bmatrix}$$

- $\boldsymbol{\Sigma}_{xx} = cov(\boldsymbol{x})$ é uma matriz $q \times q$ de covariâncias de $x_1, x_2, \ldots, x_q$.
- $\boldsymbol{\Sigma}_{yx} = cov(\boldsymbol{y}, \boldsymbol{x})$ é uma matriz $p \times q$ que contém as covariâncias entre cada y_i e cada x_j, ou seja:

$$\boldsymbol{\Sigma}_{yx} = E[(\boldsymbol{y} - \boldsymbol{\mu}_y)(\boldsymbol{x} - \boldsymbol{\mu}_x)'] = \begin{bmatrix} \sigma_{y_1 x_1} & \sigma_{y_1 x_2} & \cdots & \sigma_{y_1 x_q} \\ \sigma_{y_2 x_1} & \sigma_{y_2 x_2} & \cdots & \sigma_{y_2 x_q} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{y_p x_1} & \sigma_{y_p x_2} & \cdots & \sigma_{y_p x_q} \end{bmatrix} \quad (3.32)$$

$\boldsymbol{\Sigma}_{xy}$ é a transposta de $\boldsymbol{\Sigma}_{yx}$

Note a diferença entre

$$cov\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{x}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx}\\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx}\end{bmatrix} \qquad \text{em (3.31)}$$

e

$$cov(\boldsymbol{y}, \boldsymbol{x}) = \boldsymbol{\Sigma}_{yx} \qquad \text{em (3.32).}$$

Usamos a mesma notação "cov" de três maneiras distintas:

(1) $cov(y_i, y_j) = \sigma_{ij}$ é um escalar

(2) $cov(\boldsymbol{y}) = \boldsymbol{\Sigma}_{yy}$ é uma matriz quadrada, $p \times p$, simétrica e positiva definida.

(3) $cov(\boldsymbol{y}, \boldsymbol{x}) = \boldsymbol{\Sigma}_{yx}$ é uma matriz retangular $p \times q$

3.6. FUNÇÕES LINEARES DE VETORES ALEATÓRIOS

Situação: Em muitas aplicações trabalharemos com combinações lineares das variáveis $y_1, y_2,\ldots, y_p$ de um vetor aleatório.

Seja $\boldsymbol{a} = [a_1, a_2,\ldots, a_p]'$ um vetor de constantes. Uma combinação linear dos y_i's usando os a_i's como coeficientes pode ser escrita como:

$$z = \boldsymbol{a}'\boldsymbol{y} = a_1y_1+a_2y_2 +\ldots+ a_py_p \qquad (3.33)$$

3.6.1 Média de uma função linear de vetor aleatório

Perceba que: Se $\boldsymbol{y}$ é um vetor aleatório p-dimensional ⇒ a combinação linear $z = \boldsymbol{a}'\boldsymbol{y}$ é uma variável aleatória univariada.

Teorema 3.6A. Se $\boldsymbol{a}$ é um vetor $p\times 1$ de constantes e $\boldsymbol{y}$ é um vetor $p\times 1$ de variáveis aleatórias, então a média de $z = \boldsymbol{a'y}$ é dada por

$$\mu_z = E(\boldsymbol{a}'\boldsymbol{y}) = \boldsymbol{a}'E(\boldsymbol{y}) = \boldsymbol{a}'\boldsymbol{\mu} \qquad (3.34)$$

Caso mais geral: Se tivermos k combinações lineares de $\boldsymbol{y}$ com coeficientes constantes, a_{ij}:

$$\begin{aligned} z_1 &= a_{11}y_1 + a_{12}y_2 + \ldots + a_{1p}y_p = \boldsymbol{a}_1{}'\boldsymbol{y} \\ z_2 &= a_{21}y_1 + a_{22}y_2 + \ldots + a_{2p}y_p = \boldsymbol{a}_2{}'\boldsymbol{y} \\ &\vdots \\ z_k &= a_{k1}y_1 + a_{k2}y_2 + \ldots + a_{kp}y_p = \boldsymbol{a}_k{}'\boldsymbol{y} \end{aligned}$$

Essas k funções lineares podem ser escritas na forma matricial como:

$$\boldsymbol{z} = \boldsymbol{A}\boldsymbol{y} \qquad (3.35)$$

em que

$$\boldsymbol{z} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_k \end{bmatrix} \qquad \boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1^t \\ \boldsymbol{a}_2^t \\ \vdots \\ \boldsymbol{a}_k^t \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kp} \end{bmatrix} \text{ é } k \times p$$

Mais comum:

Trabalhar com $k \leq p$ combinações, em que as linhas de **A** são linearmente independentes.

Neste caso, $posto(\boldsymbol{A}) = k$, ou seja, $\boldsymbol{A}$ tem posto linha completo.

Note que:

Se $\boldsymbol{y}$ é um vetor aleatório ⇒ cada $z_i = \boldsymbol{a}_i{}'\boldsymbol{y}$ é uma variável aleatória univariada ⇒ $\boldsymbol{z} = [z_1, z_2, \ldots, z_k]'$ é um **vetor aleatório** $k \times 1$.

Teorema 3.6B. Supondo que $\boldsymbol{y}$ é um vetor aleatório, $\mathbf{X}$ é uma matriz aleatória, $\boldsymbol{a}$ e $\boldsymbol{b}$ são vetores de constantes e $\mathbf{A}$ e $\mathbf{B}$ são matrizes de constantes. Então, assumindo que as matrizes e os vetores em cada produto sejam conformes, temos:

$$(i)\ E(\mathbf{A}\boldsymbol{y}) = \mathbf{A}E(\boldsymbol{y}) \tag{3.36}$$

$$(ii)\ E(\boldsymbol{a}'\mathbf{X}\boldsymbol{b}) = \boldsymbol{a}'E(\mathbf{X})\,\boldsymbol{b} \tag{3.37}$$

$$(iii)\ E(\mathbf{AXB}) = \mathbf{A}E(\mathbf{X})\mathbf{B} \tag{3.38}$$

Corolário 1. Se $\boldsymbol{A}$ é uma matriz $k\times p$ de constantes, $\boldsymbol{b}$ é um vetor $k\times 1$ de constantes e $\boldsymbol{y}$ é um vetor aleatório $p\times 1$, então:

$$E(\mathbf{A}\boldsymbol{y} + \boldsymbol{b}) = \mathbf{A}E(\boldsymbol{y}) + \boldsymbol{b} \tag{3.39}$$

3.6.2 Variâncias e covariâncias de funções lineares de vetores aleatórios

Teorema 3.6C. Se $\boldsymbol{a}$ é um vetor $p\times1$ de constantes e $\boldsymbol{y}$ é um vetor aleatório $p\times1$ com matriz de covariâncias $\boldsymbol{\Sigma}$, então a variância de $z = \boldsymbol{a'y}$ é um escalar calculado por:

$$\sigma_z^2 = var(z) = var(\boldsymbol{a'y}) = \boldsymbol{a'\Sigma a} \qquad (3.40)$$

Prova:

$$\begin{aligned} var(z) &= var(\boldsymbol{a}'\boldsymbol{y}) \\ &= E[\boldsymbol{a}'\boldsymbol{y} - E(\boldsymbol{a}'\boldsymbol{y})]^2 && \text{mas } E(\boldsymbol{a}'\boldsymbol{y}) = \boldsymbol{a}'\boldsymbol{\mu} \\ &= E[\boldsymbol{a}'\boldsymbol{y} - \boldsymbol{a}'\boldsymbol{\mu}]^2 \\ &= E[\boldsymbol{a}'(\boldsymbol{y} - \boldsymbol{\mu})]^2 \\ &= E[\boldsymbol{a}'(\boldsymbol{y} - \boldsymbol{\mu})\boldsymbol{a}'(\boldsymbol{y} - \boldsymbol{\mu})] && \text{mas } \boldsymbol{a}'(\boldsymbol{y} - \boldsymbol{\mu}) = (\boldsymbol{y} - \boldsymbol{\mu})'\boldsymbol{a} \end{aligned}$$

Então:

$$
\begin{aligned}
var(z) &= E[\boldsymbol{a}'(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{a}] && \text{pelo Teorema 3.6B}(ii)\\
&= \boldsymbol{a}'E[(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})']\boldsymbol{a} && E[(\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})']=cov(\boldsymbol{y})\\
&= \boldsymbol{a}'cov(\boldsymbol{y})\boldsymbol{a} = \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}
\end{aligned}
$$

Portanto:

$$var(z) = var(\boldsymbol{a}'\boldsymbol{y}) = \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}$$

Ilustrando para $p = 3$ temos:

$$var(\boldsymbol{a'y}) = var(a_1y_1+a_2y_2+a_3y_3)$$

$$= \boldsymbol{a'\Sigma a} = [a_1 \;\; a_2 \;\; a_3]\begin{bmatrix}\sigma_1^2 & \sigma_{12} & \sigma_{13}\\ \sigma_{21} & \sigma_2^2 & \sigma_{23}\\ \sigma_{31} & \sigma_{32} & \sigma_3^2\end{bmatrix}\begin{bmatrix}a_1\\ a_2\\ a_3\end{bmatrix}$$

$$= a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + a_3^2\sigma_3^2 + 2a_1a_2\sigma_{12} + 2a_1a_3\sigma_{13} + 2a_1a_2\sigma_{23}$$

Exemplo: Seja $\boldsymbol{y} = [y_1, y_2, y_3]'$ um vetor aleatório 3×1. Calcular a variância da soma das 3 variáveis, $z =$, sabendo que

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix}$$

Resolução: Podemos escrever $z = \sum_{i=1}^{3} y_i = [1 \;\; 1 \;\; 1] \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \boldsymbol{a'y}$. Por (3.40) temos que:

$$var(z) = [1 \;\; 1 \;\; 1] \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$= \sigma_1^2 + \sigma_2^2 + \sigma_3^2 + 2\sigma_{12} + 2\sigma_{13} + 2\sigma_{23}$$

Corolário 1. Se $\boldsymbol{a}$ e $\boldsymbol{b}$ são vetores $p\times 1$ de constantes, então:

$$cov(\boldsymbol{a}'\boldsymbol{y}, \boldsymbol{b}'\boldsymbol{y}) = \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{b} \tag{3.41}$$

Teorema 3.6D. Sejam $\boldsymbol{z} = \mathbf{A}\boldsymbol{y}$ e $\boldsymbol{w} = \mathbf{B}\boldsymbol{y}$, onde $\mathbf{A}$ é uma matriz $k\times p$ de constantes, $\mathbf{B}$ é uma matriz $m\times p$ de constantes e $\boldsymbol{y}$ é um vetor aleatório $p\times 1$ com matriz de covariâncias $\boldsymbol{\Sigma}$. Então:

$$(i)\ cov(\boldsymbol{z}) = cov(\mathbf{A}\boldsymbol{y}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' \tag{3.42}$$

$$(ii)\ cov(\boldsymbol{z}, \boldsymbol{w}) = cov(\mathbf{A}\boldsymbol{y}, \mathbf{B}\boldsymbol{y}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{B}' \tag{3.43}$$

- Geralmente $k \leq p \Rightarrow$ a matriz $\mathbf{A}$, $k\times p$, é de posto linha completo $\Rightarrow$ pelo Corolário 1 do Teorema 2.6B, $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$ é **positiva definida** (desde que a matriz $\boldsymbol{\Sigma}$ seja positiva definida).

- Se $k > p$, então pelo Corolário 2 do Teorema 2.6B, **AΣA'** é **positiva semidefinida**. Neste caso, **AΣA'** ainda é uma matriz de covariâncias, mas é **singular** e não pode ser usada como numerador ou denominador da densidade normal multivariada.

- Note que **AΣB'** é uma matriz retangular $k \times m$ contendo as covariâncias de z_i com cada w_j, isto é, $cov(\boldsymbol{z}, \boldsymbol{w})$ contem $cov(z_i, w_j)$, $i = 1,\ldots, k, j = 1,\ldots, m$. Essas km covariâncias podem ser calculadas individualmente por (3.41).

Corolário 1. Se $\boldsymbol{b}$ é um vetor $k \times 1$ de constantes, então:

$$cov(\mathbf{A}\boldsymbol{y} + \boldsymbol{b}) = \mathbf{A\Sigma A'} \qquad (3.44)$$

Exercícios: Problemas das páginas 83 a 85 do livro texto.

```
* ------------------------------------------------------------ ;
* Exemplo de cálculo de matriz de Covariâncias e de Correlações ;
* amostrais;
* ------------------------------------------------------------ ;
* Weights of Cork Boring (in Centigrams) in Four Directions for ;
* 28 trees;
* Applied Multivariate Statistics with SAS Software;
* Khattree & Naik(2003) - p. 11;

proc iml;
y1 = {72,60,56,41,32,30,39,42,37,33,32,63,54,47, 91,56,79,81,78,46,
39,32,60,35,39,50,43,48};
y2 = {66,53,57,29,32,35,39,43,40,29,30,45,46,51, 79,68,65,80,55,38,
35,30,50,37,36,34,37,54};
y3 = {76,66,64,36,35,34,31,31,31,27,34,74,60,52,100,47,70,68,67,37,
34,30,67,48,39,37,39,57};
y4 = {77,63,58,38,36,26,27,25,25,36,28,63,52,43, 75,50,61,58,60,38,
37,32,54,39,31,40,50,43};

Y = y1||y2||y3||y4;
create Cork var {North East South West}; * Cria SASdataset Cork;
append from Y;
Close Cork;
```

```
p = ncol(Y);
n = nrow(Y);
In = I(n);
jn = j(n,1,1);
Jnn = J(n,n,1);
Sigma = (1/(n-1))*t(Y)*(In-(1/n)*Jnn)*Y;
D = sqrt(diag(Sigma));
corr = inv(D)*Sigma*inv(D);
Verifica = D*corr*D;
title 'Matriz de variâncias e covariâncias amostrais utilizando proc
iml';
print ,,Sigma[format=8.4],, 'Matriz de correlações:' ,,
corr[format=8.5],, Verifica[format=8.4];

mi = (1/n)*t(jn)*y; * Calcula vetor de médias;
print 'Vetor de médias:' mi[format=5.2],,;

DM2 = j(n,1,0); * Inicia cálculo da distância de Mahalanobis;
i=1;
do while (i<=n);
yi= Y[i,];
DM = (yi-mi)*inv(Sigma)*t(yi-mi);
```

```
DM2[i] = DM;
i=i+1;
end;

rank = rank(DM2);   * Calcula a ordem de cada valor do vetor DM2;
print
'-----------------------------------------------------------------',
'Distância de Mahalanobis de cada ponto (y) ao vetor de médias(mi)',
'-----------------------------------------------------------------';
print ,,Y ' ' DM2[format=8.4] ' ' rank;
quit;

proc corr cov data=cork;
 title 'Matriz de variâncias e covariâncias utilizando proc corr';
 var north east south west;
run;
```