ME731 - Métodos em Análise Multivariada – Distribuição Normal Multivariada III –

Prof. Carlos Trucíos
ctrucios@unicamp.br
ctruciosm.github.io

Instituto de Matemática, Estatística e Computação Científica,
Universidade Estadual de Campinas

Aula 06

Agenda I

1. Avaliando Normalidade
2. Outliers
3. Transformações

Avaliando Normalidade

Avaliando Normalidade

Avaliar Normalidade multivariada para $p > 2$ é uma tarefa dificil. Contudo, utilizando propriedades da distribuição Normal multivariada, podemos avaliar a falta de normalidade:

Avaliando Normalidade

Avaliar Normalidade multivariada para $p > 2$ é uma tarefa dificil. Contudo, utilizando propriedades da distribuição Normal multivariada, podemos avaliar a falta de normalidade:

- São as marginais Normais?
- São combinações lineares das marginais Normais?
- Gráficos de dispersão por pares de variáveis apresentam uma forma elíptica?

Avaliando Normalidade

Avaliar Normalidade multivariada para $p > 2$ é uma tarefa dificil. Contudo, utilizando propriedades da distribuição Normal multivariada, podemos avaliar a falta de normalidade:

- São as marginais Normais?
- São combinações lineares das marginais Normais?
- Gráficos de dispersão por pares de variáveis apresentam uma forma elíptica?

  Embora analisar de forma univariada ou bivariada não garante normalidade multivariada, a falta de Normalidade uni ou bi dimensional é suficiente para não termos Normalidade Multivariada.

Avaliando Normalidade: Q-Q plot

Q-Q (Quantile - Quantile) plot:

1. Ordenar os dados $x_{(1)}, \cdots, x_{(n)}$.
2. Calcular, para cada $x_{(j)}$, $p_{(j)} = (j - 1/2)/n$.
3. Calcular os quantis teóricos $q_{(1)}, \cdots, q_{(n)}$ em que

$$P(Z \leq q_{(j)}) = \frac{j - 1/2}{n}$$

4. Graficar $(q_{(1)}, x_{(1)}), \cdots, (q_{(n)}, x_{(n)})$.

   Se X for normalmente distribuida, esperamos que os quantis teóricos e amostrais estejam próximos (ou seja, estejam perto de uma linha reta).

Avaliando Normalidade: Q-Q plot

```
qq_norm <- function(x) {
  n <- length(x)
  x_ordered <- sort(x)
  j <- seq(1, n, 1)
  pj <- (j - 0.5)/n
  qj <- qnorm(pj)
  Qa <- as.vector(quantile(x, probs = c(0.25, 0.75)))
  Qt <- qnorm(c(0.25, 0.75))
  slope <- diff(Qa)/diff(Qt)
  inter <- Qa[1] - slope*Qt[1]
  plot(qj,x_ordered)
  abline(inter, slope)
}
```

Avaliando Normalidade: Q-Q plot

```
x <- rnorm(1000)
qq_norm(x)
x <- rchisq(1000, 2)
qq_norm(x)
qq_norm(-x)
x <- rt(1000,4)
qq_norm(x)
x <- rnorm(20)
qq_norm(x)
```

Avaliando Normalidade

Alguns testes de Normalidade univariada, que já devem ter visto em outras disciplinas, são:

- Kolmogorov-Smirnov,
- Shapiro-Wilk,
- Anderson-Darling,
- Cramer-von Mises,
- D'Agostino-Pearson,
- Jarque-Bera.

Avaliando Normalidade

Alguns testes de Normalidade univariada, que já devem ter visto em outras disciplinas, são:

- Kolmogorov-Smirnov,
- Shapiro-Wilk,
- Anderson-Darling,
- Cramer-von Mises,
- D'Agostino-Pearson,
- Jarque-Bera.

Na Lista 2 encontrará um exercício interessante para rever estes testes.

Avaliando Normalidade

Sabemos que se $\mathbf{X} \sim N_p(\mu, \Sigma)$, qualquer combinação linear das componentes terá uma distribuição normal univariada. Como é impossível testar todas as combinações lineares, uma prática comum é utilizar o autovetor associado ao maior autovalor de $\mathbf{S}$ e verificar a normalidade para essa combinação linear.

Avaliando Normalidade

Sabemos que se $\mathbf{X} \sim N_p(\mu, \Sigma)$, qualquer combinação linear das componentes terá uma distribuição normal univariada. Como é impossível testar todas as combinações lineares, uma prática comum é utilizar o autovetor associado ao maior autovalor de $\mathbf{S}$ e verificar a normalidade para essa combinação linear.

Se os dados forem normalmente distribuidos, ao fizermos gráficos de dispersão por pares, as nuvens de pontos deveriam ter forma elíptica,

Avaliando Normalidade

Caso multivariado:

Sabemos que se $\mathbf{X} \sim N_p(\mu, \Sigma)$, então

$$(\mathbf{X} - \mu)'\Sigma^{-1}(\mathbf{X} - \mu) \sim \chi^2_p.$$

Assim, podemos construir um gráfico Q-Q mas utiizando a distribuição χ^2.

Avaliando Normalidade

Caso multivariado:

Sabemos que se $\mathbf{X} \sim N_p(\mu, \Sigma)$, então

$$(\mathbf{X} - \mu)'\Sigma^{-1}(\mathbf{X} - \mu) \sim \chi^2_p.$$

Assim, podemos construir um gráfico Q-Q mas utiizando a distribuição χ^2.

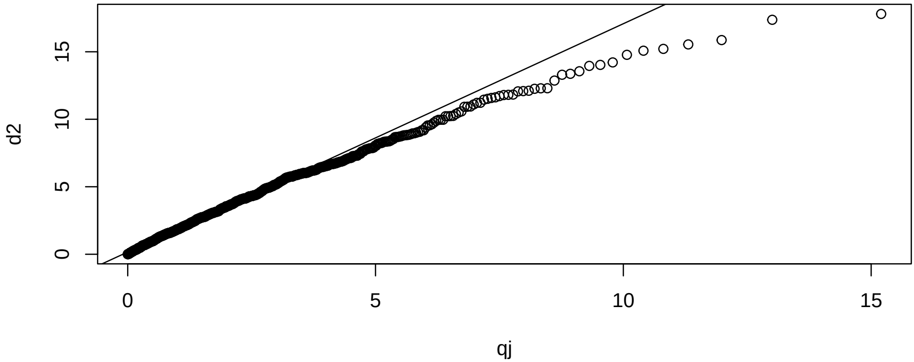
1. Calcular $d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})'S^{-1}(\mathbf{x}_j - \bar{\mathbf{x}})$ e ordená-los em forma crescente: $d^2_{(1)}, \cdots, d^2_{(n)}$.
2. Calcular os quantis teóricos $q_{\chi^2,p}\left(\frac{j-1/2}{n}\right)$.
3. Graficar $\left(q_{\chi^2,p}(\frac{j-1/2}{n}), d^2_{(j)}\right)$.

Avaliando Normalidade

```
qq_chi <- function(x) {
  n <- nrow(x)
  p <- ncol(x)
  d2 <- sort(mahalanobis(x, center = TRUE, cov = cov(x)))
  j <- seq(1, n, 1)
  qj <- qchisq((j - 0.5)/n, p)
  Qa <- as.vector(quantile(d2, probs = c(0.25, 0.75)))
  Qt <- qchisq(c(0.25, 0.75), p)
  slope <- diff(Qa)/diff(Qt)
  inter <- Qa[1] - slope*Qt[1]
  plot(qj, d2)
  abline(inter, slope)
}
```

Avaliando Normalidade

```
set.seed(246)
x <- mvtnorm::rmvnorm(1000, sigma = matrix(c(1, 0.3, 0.3, 1), 2))
qq_chi(x)
```



Avaliando Normalidade

Teste de Mardia (1970)

- Um dos testes mais conhecidos e utilizados.
- Baseia-se nos coeficiente multivariados de assimetria e curtose propostos por Mardia (1970).
- Se $\mathbf{X}$ e $\mathbf{Y}$ são *iid* com $\mathbf{X} \sim (\mu, \Sigma)$, então os coeficientes de assimetria e curtose multivariados são dados por:

$$\beta_{1,p} = \mathbb{E}[(\mathbf{X} - \mu)'\Sigma^{-1}(\mathbf{Y} - \mu)]^3 \quad e \quad \beta_{2,p} = \mathbb{E}[(\mathbf{X} - \mu)'\Sigma^{-1}(\mathbf{X} - \mu)]^2$$

- No caso da distribuição Normal multivariada, $\beta_{1,p} = 0$ e $\beta_{2,p} = p(p+2)$.
- O teste de Mardia testa se de fato $\beta_{1,p} = 0$ e $\beta_{2,p} = p(p+2)$.

Avaliando Normalidade

Teste de Mardia (1970)

Baseia-se nas generalizações multivariadas de assimetria e curtose. Sejam

$$A_p = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^3 \quad e \quad K_p = \frac{1}{n}\sum_{i=1}^{n} d_{ii}^2,$$

os estimadores dos coeficientes de assimetria e curtose multivariada, em que $d_{ij} = (\mathbf{X}_i - \bar{\mathbf{X}})' S^{-1} (\mathbf{X}_j - \bar{\mathbf{X}})$. Então, sob H_0 (normalidade multivariada) e quando $n \to \infty$,

$$\frac{n}{6} A_p \sim \chi^2_{\frac{1}{6}p(p+1)(p+2)} \quad e \quad \sqrt{n}\frac{K_p - p(p+2)}{\sqrt{8p(p+2)}} \sim N(0,1)$$

Demostração: ver Mardia (1970, 1974).

Avaliando Normalidade

```
mvnormalTest::mardia(x)
```

```
## $mv.test
##            Test Statistic p-value Result
## 1      Skewness    6.8239  0.1455    YES
## 2      Kurtosis    0.4011  0.6883    YES
## 3 MV Normality      <NA>    <NA>    YES
##
## $uv.shapiro
##    W      p-value UV.Normality
## V1 0.9986 0.6183  Yes
## V2 0.9983 0.4416  Yes
```

`$mv.test` é o teste de Mardia. Rejeitamos H_0 se ambos os p-valores são $>$ 0.05.

Avaliando Normalidade

- Ebner e Henze (2020) recomendam o uso dos testes *BHEP* (Baringhaus-Henze-Epps-Pulley), *DEH* (Doerr-Ebner-Henze) e *Energy test* por serem computacionalmente *rápidos* e apresentarem bons resultados (em termos de poder do teste).
- Se o tempo de processamento/poder computacional não for um problema, Ebner e Henze (2020) sugerem o uso do teste *HJM* (Henze-Jimenes-Gamero-Meintanis). No meu computador (Core i5, 8Gb RAM) o método deu problema de memória.
- Não entraremos em detalhes de como os testes funcionam, mas nos conformaremos com saber que existem e como são utilizados.
- Todos estes testes estão implementados no pacote `mnt`.

Avaliando Normalidade

```
mnt::test.BHEP(x, MC.rep = 1000)
mnt::test.DEHT(x, MC.rep = 1000)
mnt::test.SR(x, MC.rep = 1000)
mnt::test.HJM(x, MC.rep = 1000)
```

Rejeitamos H_0 se o valor da estatística de Teste for maior do que o valor crítico.

Outliers

Outliers

- Outliers tem um efeito negativo na estimação de parâmetros.

Outliers

- Outliers tem um efeito negativo na estimação de parâmetros.

Detectar outliers

- Faça gráficos de pontos ou boxplots.
- Faça diagramas de dispersão.
- Padronize os dados $z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{s_{jj}}}$, $i = 1, \cdots, n$ e $j = 1, \cdots, p$ e identifique valores inusuais.
- Calcule $(\mathbf{x}_i - \bar{\mathbf{x}})' S^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$ e identifique valores inusuais.

Transformações

Transformações

O que fazer se os dados não tem distribuição Normal?

Transformações: Algumas transformações univariadas úteis

Dados	Transformação
Asimétricos	$\log(y)$ ou $\sqrt{y}$
Dados de contagem	$\sqrt{y}$
Proporções	$1/2 \log\left(\frac{y}{1-y}\right)$

Transformações: Algumas transformações univariadas úteis

Dados	Transformação
Asimétricos	$\log(y)$ ou $\sqrt{y}$
Dados de contagem	$\sqrt{y}$
Proporções	$1/2 \log\left(\frac{y}{1-y}\right)$

Outra transformação útil quando $x > 0$ é a transformação Box-Cox:

$$x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{se } \lambda \neq 0, \\ \log(x) & \text{se } \lambda = 0. \end{cases} \tag{1}$$

Note que $\lim_{\lambda \to 0} \frac{x^\lambda - 1}{\lambda} = 0.$

Transformações

Utilizaremos os dados do exemplo 4.10 do Johnson & Wichern (2007) que estão disponíveis aqui.

```
radiation <- read.csv("./datasets/radiation_data.csv")
tseries::jarque.bera.test(radiation$close)
```

```
##
##  Jarque Bera Test
##
## data:  radiation$close
## X-squared = 11.939, df = 2, p-value = 0.002555
```

Transformações

```r
lambda1 <- car::powerTransform(radiation$close)$lambda
close_boxcox <- (radiation$close^(lambda1) - 1)/lambda1
tseries::jarque.bera.test(close_boxcox)
```

```
##
##  Jarque Bera Test
##
## data:  close_boxcox
## X-squared = 0.19123, df = 2, p-value = 0.9088
```

Transformações

```
lambda1 <- car::powerTransform(radiation$close)$lambda
close_boxcox <- (radiation$close^(lambda1) - 1)/lambda1
tseries::jarque.bera.test(close_boxcox)
```

```
##
##  Jarque Bera Test
##
## data:  close_boxcox
## X-squared = 0.19123, df = 2, p-value = 0.9088
```

Note que mesmo que as x não sejam todas positivas, sempre podemos fazer $\underbrace{x + c}_{x*} > 0$.

Transformações

Como é escolhido λ?

Transformações

Como é escolhido λ?

Assumindo que existe λ tal os dados transformados atingem normalidade, utilizamos o método de máxima verossimilhança para encontrar qual o valor de λ que é mais provavel de ter dado origem aos dados observados.

Transformações

Como é escolhido λ?

Assumindo que existe λ tal os dados transformados atingem normalidade, utilizamos o método de máxima verossimilhança para encontrar qual o valor de λ que é mais provavel de ter dado origem aos dados observados.

λ é escolhido tal que maximize

$$\begin{aligned}
\log[f_{X_1,\cdots,X_n}(x_1,\cdots,x_n)] &= \log[f_{X_1^{(\lambda)},\cdots,X_n^{(\lambda)}}(u(x_1),\cdots,u(x_n))|J|] \\
&= \log\left[f_{X_1^{(\lambda)},\cdots,X_n^{(\lambda)}}(x_1^{(\lambda)},\cdots,x_n^{(\lambda)})\left|\frac{\partial x_i^{(\lambda)}}{\partial x_j}\right|\right] \\
&\propto -\frac{n}{2}\log\left[\frac{1}{n}\sum_{i=1}^{n}(x_i^{\lambda}-\bar{x_i^{\lambda}})^2\right]+(\lambda-1)\sum_{i=1}^{n}\log(x_j).
\end{aligned}$$

Transformações

O que fazer quando temos dados multivariados?

1. Aplicar a transformação a cada uma das variáveis.
2. Calcular os $\lambda_1, \cdots, \lambda_p$ de forma que maximizem

$$l(\lambda_1, \cdots, \lambda_p) = -\frac{n}{2}\log(\mathbf{S}(\lambda)) + \sum_{j=1}^{p}\left[(\lambda_j - 1)\sum_{i=1}^{n}\log(x_{i,j})\right],$$

em que $\mathbf{S}(\lambda)$ é a matriz de covariância de $\mathbf{x}_i^{\lambda} = (\frac{x_{i1}^{\lambda_1}-1}{\lambda_1}, \cdots, \frac{x_{ip}^{\lambda_p}-1}{\lambda_p})'$

Transformações

O que fazer quando temos dados multivariados?

1. Aplicar a transformação a cada uma das variáveis.
2. Calcular os $\lambda_1, \cdots, \lambda_p$ de forma que maximizem

$$l(\lambda_1, \cdots, \lambda_p) = -\frac{n}{2}\log(\mathbf{S}(\lambda)) + \sum_{j=1}^{p}\left[(\lambda_j - 1)\sum_{i=1}^{n}\log(x_{i,j})\right],$$

em que $\mathbf{S}(\lambda)$ é a matriz de covariância de $\mathbf{x}_i^{\lambda} = (\frac{x_{i1}^{\lambda_1}-1}{\lambda_1}, \cdots, \frac{x_{ip}^{\lambda_p}-1}{\lambda_p})'$

Utilizar a segunda opção é mais complexo e, na prática, os resultados não são muito melhores.

Transformações

```
lambda2 <- car::powerTransform(radiation$open)$lambda
open_boxcox <- (radiation$open^(lambda2) - 1)/lambda2
tseries::jarque.bera.test(open_boxcox)
```

```
##
##  Jarque Bera Test
##
## data:  open_boxcox
## X-squared = 0.085604, df = 2, p-value = 0.9581
```

Transformações

```
new_data <- cbind(close_boxcox, open_boxcox)
mvnormalTest::mardia(new_data)$mv.test
```

```
##             Test Statistic p-value Result
## 1       Skewness    4.0126  0.4043    YES
## 2       Kurtosis    1.3245  0.1853    YES
## 3 MV Normality      <NA>    <NA>    YES
```

```
mvnormalTest::mardia(radiation)$mv.test
```

```
##             Test Statistic p-value Result
## 1       Skewness   28.9993       0     NO
## 2       Kurtosis    5.2331       0     NO
## 3 MV Normality      <NA>    <NA>     NO
```

Transformações

```r
lambda <- car::powerTransform(radiation)$lambda
close_boxcoxM <- (radiation$close^(lambda[1]) - 1)/lambda[1]
open_boxcoxM <- (radiation$open^(lambda[2]) - 1)/lambda[2]
new_dataM <- cbind(close_boxcoxM, open_boxcoxM)
mvnormalTest::mardia(new_dataM)$mv.test
```

```
##             Test Statistic p-value Result
## 1       Skewness    4.1895   0.381    YES
## 2       Kurtosis    1.3753   0.169    YES
## 3 MV Normality      <NA>    <NA>    YES
```

Transformações

```r
lambda <- car::powerTransform(radiation)$lambda
close_boxcoxM <- (radiation$close^(lambda[1]) - 1)/lambda[1]
open_boxcoxM <- (radiation$open^(lambda[2]) - 1)/lambda[2]
new_dataM <- cbind(close_boxcoxM, open_boxcoxM)
mvnormalTest::mardia(new_dataM)$mv.test
```

```
##             Test Statistic p-value Result
## 1       Skewness    4.1895   0.381    YES
## 2       Kurtosis    1.3753   0.169    YES
## 3 MV Normality      <NA>    <NA>    YES
```

As transformações sempre funcionam? Não!

Transformações: Outras transformações úteis

- Manly(1971): permite valores negativos e funciona bem em casos unimodeis assimetricos.
- John and Draper(1980): permite valores negativos e funciona bem em casos simétricos.
- Yeo and Johnson (2000)
- etc.

Transformações: Outras transformações úteis

- Manly(1971): permite valores negativos e funciona bem em casos unimodeis assimetricos.
- John and Draper(1980): permite valores negativos e funciona bem em casos simétricos.
- Yeo and Johnson (2000)
- etc.

Raymaekers & Rousseeuw (2021) apresentam uma recente e interessante discussão ao respeito.

Exemplo

Utilizando o seguinte dataset disponível aqui:

- Verifique a normalidade (univariada) para as variáveis `Indep`, `Supp`, `Benev`, `Conform`, `Leader`.
- Utilize as cinco variáveis de forma conjunta e verifique a normalidade multivariada.
- Se necessário, faça transformações para atingir normalidade.

V1(independence), V2(support), V3(benevolence), V4(conformity), V5(leadership), V6(gender) e V7(socieconomic status)

Referências

- Mardia, K. V., Kent, J. T., & Bibby, J, M. (1979). Multivariate Analysis. Academic Press. Capítulo 2.
- Ebner, B., & Henze, N. (2020). Tests for multivariate normality—a critical review with emphasis on weighted L^2-statistics. Test, 29(4), 845-892.
- Raymaekers, J., & Rousseeuw, P. J. (2021). Transforming variables to central normality. Machine Learning, 1-23..
- Mardia, K. V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. Sankhyā: The Indian Journal of Statistics, Series B, 115-128.
- Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. Biometrika, 57(3), 519-530.