Middle East Technical University
Department of Statistics
2020-21 Fall Semester
STAT 467 Multivariate Analysis

November 26, 2020, Thursday

RECITATION 4

Assessing the assumptions of Normality

The MV Normal distribution plays an important role in the most of the statistical analysis. Therefore, it is important to check this assumption at times. We'll mainly concentrate on checking univariate or bivariate normality rather than multivariate normality since it is quite difficult to check multivariate normality.

Multivariate Normality -> Univariate Normality (Reverse is not true.)

Therefore, the control of the univariate normality can never be sure that we have not missed some feature that is revealed only in higher dimensions. Yet many types of non-normality are often reflected in marginal or bivariate distribution. **In general, people check univariate and bivariate normality and assume multivariate normality if univariate and bivariate normality holds.**

Testing Bivariate Normality

If $X_1, .. X_n$ is a random sample from $N_2(\mu, \Sigma)$, then we know that the set of bivariate outcomes x satisfy.

$$(X \quad -\mu)' \, \Sigma^{-1}(X \quad -\mu) \quad \leq \; \chi^2_{p=2,a}$$ where α is significance level.

If the observations were from normal distribution, then at least 50% of them would satisfy this condition.

Chi-square or Gamma Plot

Can be used for high dimensional case where $p \geq 2$.

If the plot resembles a straight line, it indicates the normality.

Detecting Outliers

Outliers are observations that are too large or too small relative to rest.

For univarite case, i.e $p = 1$

1. Visual Inspection (Line graph, histogram,etc)

2. Box Plot

For p >1, case

1) Look at p marginals. If $X_i \sim N(\mu_i, \sigma_i^2)$ for i=1,…,p. Then $Z_{ij=}\frac{X_{ij}-\bar{X}_i}{S_i} \sim N(0,1)$. If $|Z_{ij}|$ is too large, i.e greater than 3, then the observation may be outlier.

2) Multivariate approach: Draw chi-square plot and if you see unusually large values, which would be the points fartherst from the origin, may be outlier.

Depending upon the nature of the outliers and objectives of the investigation, outliers may be deleted or approximately weighted in a subsequent analysis.

What to do if the data are not normal?

- **Transformation**
- **Discart the outliers**
- **Apply nonparametric methods**
- **Increase the sample size, n.**
- **Generate bootsrap samples.**

Transformations

In practice,

- For data skewed to the right: $y^* = y^{1/2}, y^{\frac{1}{4}}, \log(y), y^{-1}, \ldots.$
- For data skewed to the left: $y^* = y^2, y^3, \ldots.$
- For counts: $y^* = \sqrt{y}$
- For proportions: Logit transformation
- For correlations: Fisher correlation transformation

Since these transformations do not guarantee normality, you must check that the distribution of y^* is approximately normal.

QUESTIONS

1. Let $X_1, .. X_{20}$ be a random sample of size 60 from an $N_6(\mu, \Sigma)$ population. Specify each of the following completely.

a) The distribution of $(X_1 - \mu)' \, \Sigma^{-1}(X_1 - \mu)$

b) The distribution of $\bar{X}$ and $\sqrt{n}(\bar{X} - \mu)$

c) The distribution of $(n-1)S$

2. The world's 10 largest companies yield the following data

	x_1(Sales)	**x_2 (Profit)**
Citigroup	108.28	17.05
General Electric	152.36	16.59
American Intl Group	95.04	10.91
Bank of America	65.45	14.14
HSBC	62.97	9.52
ExxonMobil	263.99	25.33
Royal Dutch/Shell	265.19	18.54
BP	285.06	15.73
ING Group	92.01	8.10
Toyota Motor	165.68	11.13

Determine the proportion of the obeservations falling within the estimated 50% probability contour of a bivariate normal distribution.

QUIZ 4

Please send the name of the movie that you love to ozancan@metu.edu.tr until tomorrow at 23.59.

STAT467 / Recitation 4

Answer Key

1) a) $X_1 \sim N(\mu_1, \sigma_1^2)$ In matrix notation, $x^2 = x'x$

$(X_1-\mu)' \Sigma^{-1} (X_1-\mu)$ $X_1 - \mu_1 \sim N(0, \sigma_1^2)$

$(X_1-\mu_1)^2 . \underset{\hookrightarrow \sigma_1^2}{\Sigma^{-1}}$ (one dimensional) $z = \frac{X_1-\mu_1}{\sigma_1} \sim N(0,1)$

$\frac{(X_1-\mu_1)^2}{\sigma_1^2} \sim ?$ $z^2 = \frac{(x_1-\mu_1)^2}{\sigma_1^2} \sim \chi^2(1)$ #

b) $\bar{X} \sim N_6(\mu, \Sigma/n) \xrightarrow{n=20} \bar{X} \sim N_6(\mu, \Sigma/20)$

$\sqrt{n}(\bar{X}-\mu) \rightarrow \bar{X}-\mu \sim N_6(0, \Sigma/n)$

$\sqrt{n}(\bar{x}-\mu) \sim N_6(0, \Sigma)$

$(n=20)$

$\sqrt{20}(\bar{x}-\mu) \sim N_6(0, \Sigma)$

c) $(n-1)S$ is distributed as wishart random matrix with d.o.f $n-1$ since $n=20$

19S " " " " " " " " 19.

Example 4.12 (Checking bivariate normality) Although not a random sample, data consisting of the pairs of observations (x_1 = sales, x_2 = profits) for the 10 largest companies in the world are listed in Exercise 1.4. These data give

$$\bar{\mathbf{x}} = \begin{bmatrix} 155.60 \\ 14.70 \end{bmatrix}, \qquad \mathbf{S} = \begin{bmatrix} 7476.45 & 303.62 \\ 303.62 & 26.19 \end{bmatrix}$$

so

$$\mathbf{S}^{-1} = \frac{1}{103{,}623.12} \begin{bmatrix} 26.19 & -303.62 \\ -303.62 & 7476.45 \end{bmatrix}$$

$$= \begin{bmatrix} .000253 & -.002930 \\ -.002930 & .072148 \end{bmatrix}$$

From Table 3 in the appendix, $\chi_2^2(.5) = 1.39$. Thus, any observation $\mathbf{x}' = [x_1, x_2]$ satisfying

$$\begin{bmatrix} x_1 - 155.60 \\ x_2 - 14.70 \end{bmatrix}' \begin{bmatrix} .000253 & -.002930 \\ -.002930 & .072148 \end{bmatrix} \begin{bmatrix} x_1 - 155.60 \\ x_2 - 14.70 \end{bmatrix} \leq 1.39$$

is on or inside the estimated 50% contour. Otherwise the observation is outside this contour. The first pair of observations in Exercise 1.4 is $[x_1, x_2]' = [108.28, 17.05]$. In this case

$$\begin{bmatrix} 108.28 - 155.60 \\ 17.05 - 14.70 \end{bmatrix}' \begin{bmatrix} .000253 & -.002930 \\ -.002930 & .072148 \end{bmatrix} \begin{bmatrix} 108.28 - 155.60 \\ 17.05 - 14.70 \end{bmatrix}$$
$$= 1.61 > 1.39$$

and this point falls outside the 50% contour. The remaining nine points have generalized distances from $\bar{\mathbf{x}}$ of .30, .62, 1.79, 1.30, 4.38, 1.64, 3.53, 1.71, and 1.16, respectively. Since four of these distances are less than 1.39, a proportion, .40, of the data falls within the 50% contour. If the observations were normally distributed, we would expect about half, or 5, of them to be within this contour. This difference in proportions might ordinarily provide evidence for rejecting the notion of bivariate normality; however, our sample size of 10 is too small to reach this conclusion. (See