ST437/537 – HW #02

Arnab Maity

Due date: January 24, 2019

Instructions

Please follow the instructions below when you prepare and submit your assignment.

- **Include a cover-page** with your homework. It should contain

i. Full name,
ii. Course#: ST 437/537 and
iii. HW-#
iv. Submission date

- Assignments should be submitted in class on the date specified ("due date").
- Neatly typed or hand-written solution on standard letter-size papers (stapled on the top-left corner) should be submitted. **All R code/output should be well commented, with relevant outputs highlighted.**
- **Always staple (upper left corner) your homework before coming to class. Ten percent points will be deducted otherwise.**
- When you solve a particular problem, do not only give the final answer. Instead **show all your work** and the steps you used (with proper explanation) to arrive at your answer to get full credit.

Problems

Soleve the following problems. You may use `R` for these problems unless I specifically instruct otherwise.

1. (10 points) Consider the [lumber stiffness data] (../data/T4-3.DAT) used in class to assess multivariate normality. Load the data in R and remove the two rows that might be outliers.

a. Perform univariate **Shapiro-Wilk** tests and multivariate **Royston** tests on the new dataset.
b. Create a new chi-square plot.
c. Discuss the results you fond in part (a) and (b).

2. (20 points) Consider the `skulls` dataset in the `HSAUR3` package in `R` you considered in HW #1. You will first need to install the package in `R` to access the dataset. Use `?skulls` command to get more details on the data. A snapshot of the data is shown below.

```
library(HSAUR3)
```

```
## Loading required package: tools
```

```
head(skulls)
```

```
##     epoch  mb  bh  bl nh
## 1 c4000BC 131 138  89 49
## 2 c4000BC 125 131  92 48
## 3 c4000BC 131 132  99 50
## 4 c4000BC 119 132  96 44
## 5 c4000BC 136 143 100 54
## 6 c4000BC 138 137  89 56
```

Consider only the `c4000BC` epoch.

a. Create a pairs-plot. Do you see any unusual patterns?
b. Create chi-square plot and overlay a 45 degrees digonal line (e.g. a line with zero intercept and slope 1; the function `abline()` will be useful). Comment on the plot. [Hint: Do not over-interpret a single point.]
c. Create a new dataset where you add the z-scores for each variable as well as the sample Mahalanobis distances as columns (as was done in class). Consider the two data points with highest distance values (these are not necessarily outliers), and comment on their z-scores.
d. Perform univariate **Shapiro-Wilk** tests and multivariate **Royston** tests on the dataset.

3. (20 points) Perform the following tasks related to bvariate normal distribution. You need to install the R package `mnormt`, and then load the package by calling `librart(mnormt)` command. The function `rmnorm()` in the `mnormt` library generates random samples from a multivariate normal. Use `?rmnorm()` to open the documentation and read how to use it.

a. Generate 100 data points from a bivariate normal distribution such that $E(X_1) = 1$, $E(X_2) = 2$, $var(X_1) = 1$, $var(X_2) = 2$ and $cov(X_1, X_2) = 1$. [Hint: you need to write iut the mean vector μ and covariance matrix Σ to use them in the R function.]

b. Make a scatter plot of the generated data, and overlay data ellipses (50% and 95%). What pattern would you expect to see in the scatterplot?

c. Now consider a new variable $Y = X_1 + X_2$. Explicitly write down the distribution of Y; clearly specify the distribution and its parameters to get full credit.

d. From the sample you generated in (a), compute observations of Y, draw a histogram of these values, and overlay the PDF of the distribution you obtained in part (c) on the histogram. What pattern do you expact to see? Does you expectation match with what you see in the plot? Explain. [Hint: the function `dnorm()` can be used to compute the PDF of a normal distribution.]

4. (20 points) Answer the following questions.

a. Suppose X and Y are two random variables with means μ_X and μ_Y, respectively, and variances σ_X^2 and σ_Y^2, respectively. Recall that the covariance and correlation coefficient of two variables is defined as

$$cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] \text{ and } cor(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y}.$$

Define the standardized variables

$$Z_1 = \frac{X - \mu_X}{\sigma_X} \text{ and } Z_2 = \frac{Y - \mu_Y}{\sigma_Y}.$$

Show that

$$cov(Z_1, Z_2) = cor(X, Y),$$

that is, covariance between two standardized variables is the same as correlation coefficient between the two original variables.

For your information: the same result applies to multivariate data. Specifically, the covariance matrix of the standarized data is the same as the correlation matrix of the original data.

b. In this part we will verify the result in part (a) numerically. To do so generate 100 data points from a multivariate normal distribution with

$$\mu = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 3 \end{pmatrix}.$$

- Using the generated data points, first compute the sample correlation matrix $\boldsymbol{R}$.
- Then standardize each variable, and using the standardized variables, compute their sample covariance matrix.

Do these matrices match? Explain.