Honours Multivariate Analysis

Continuous Assessment 4

Laylaa Jorge Abrahams (JRGLAY001)
Roche Witbooi (WTBROC002) Chris Eason (ESNCHR001)

Table of contents

Question 1

Suppose X Np(μ,Σ). Show that the maximum likelihood estimator of Σ is biased, and give the bias:

1. The following shows that the MLE of Σ is biased because of the added $\frac{n-1}{n}$ term

$$E[\hat{\Sigma}_{MLE}] = E[\frac{n-1}{n}S] = \frac{n-1}{n}E[S] = \frac{n-1}{n}\Sigma$$

2. The bias of the MLE of Σ

$$Bias(\hat{\Sigma}_{MLE}) = E[\hat{\Sigma}_{MLE}] - \Sigma$$

$$Bias(\hat{\Sigma}_{MLE}) = \frac{n-1}{n}\Sigma - \Sigma$$

$$Bias(\hat{\Sigma}_{MLE}) = -\frac{1}{n}\Sigma$$

Question 2

Consider again the Egyptian skull data from CA1, given in CA1.csv .

Examine the variables in period 2 for marginal and multivariate normality by creating the necessary QQ-plot(s) and chi-square plot(s).

Apply any statistical test to the univariate hypotheses and report a measure of the p-value. For the multivariate test, interpret the observed squared generalized distances.

QQ - plots

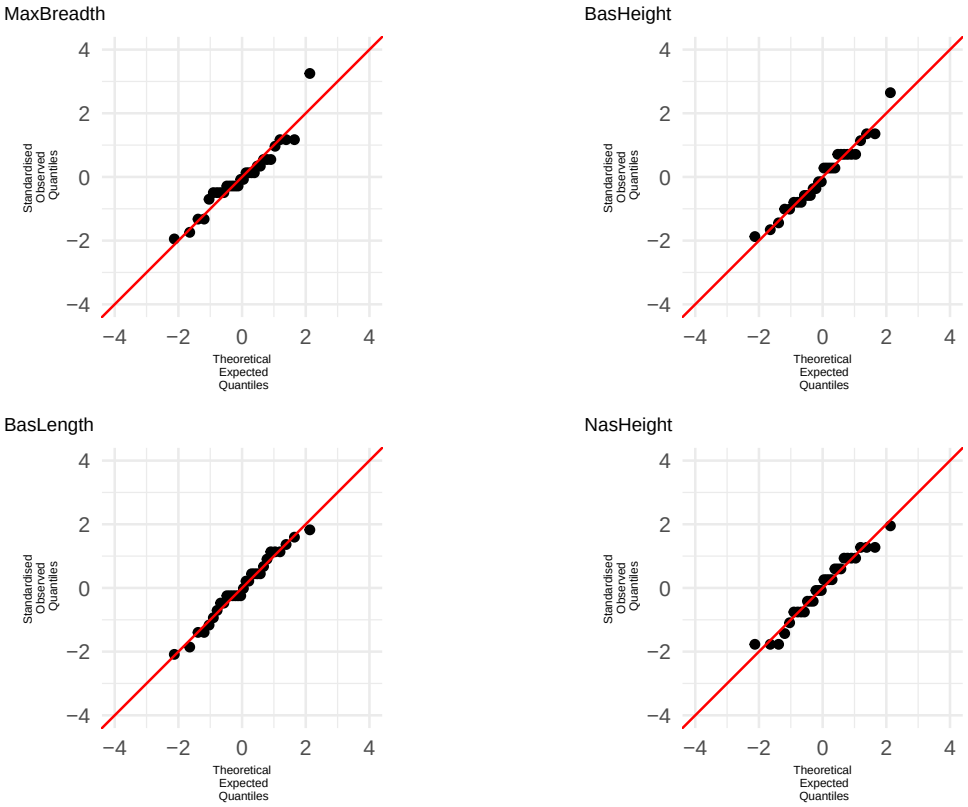


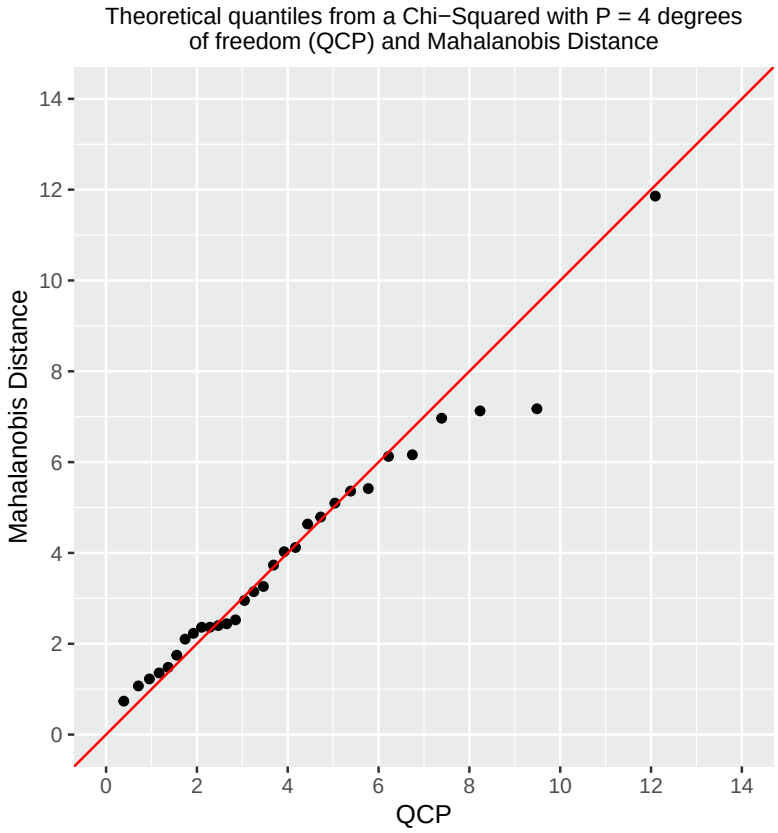
Table 1: Shapiro Wilk Test

	P Value
MaxBreadth	0.04
BasHeight	0.58
BasLength	0.75
NasHeight	0.38

Interpretation:

- Shapiro Wilk test Null hypothesis (H0): The data is normally distributed. Alternative hypothesis (H1): The data is not normally distributed.
- MaxBreadth P_value $= 0.04 < 0.05$, meaning we reject the null hypothesis and conclude at the 5% significance level that MaxBreadth is not normally distributed
- BasHeight, BasLength and NasHeight all have P_valeus > 0.05, meaning we faily to reject the null hypothesis and conclude at the 5% significance level that BasHeight, BasLength and NasHeight are normally distributed

Chi-square plot



Critical value : 3.36.

Proportion of points where mahalanobis distance is smaller then the critical value : 53.33%.

Interpretation

53.33 % of points have a mahalanobis distance smaller than the critical value, indicating a slight deviation from the expected 50%. However, the deviation is not large, and there is no

significant evidence against the normality assumption. It is important to note that since the sample size is 30, which is relatively small, only severe deviations will indicate a lack of fit.

Appendix

```
suppressPackageStartupMessages({ library(dplyr) })

dat <- read.csv("CA1.csv")

period_2 <- dat |> filter(TimePeriod ==2)
period_2 <- scale(period_2[,1:4])
```

QQ plot

```
library(knitr)
library(ggplot2)
library(patchwork)

# qqplot
p_val <- matrix(nrow=4,ncol = 1)
plots <- list()

for(i in 1:ncol(period_2)){
  result <- qqnorm(period_2[, i], main=colnames(period_2)[i], plot.it = FALSE)
  df <- data.frame(x = result$x, y = result$y)
  plots[[i]] <- ggplot()+
    geom_point(data = df, aes(x=x,y=y), color = "black")+
    geom_abline(slope = 1, intercept = 0, color = "red") +
    ggtitle(colnames(period_2)[i]) +
    coord_fixed(ratio = 1) +
    scale_x_continuous(
      breaks = seq(-4, 4, by = 2),
      limits = c(-4, 4)
    )+
    scale_y_continuous(
      breaks = seq(-4, 4, by = 2),
      limits = c(-4, 4)
    ) +
    labs(
      y = "Standardised \nObserved\n Quantiles",
      x = "Theoretical\n Expected\n Quantiles"
    ) +
    theme_minimal() +
    theme(
```

```
      axis.title.x = element_text(size = 5),
      axis.title.y = element_text(size = 5),
      plot.title = element_text(size = 8),
      )

  p_val[i] <- shapiro.test(period_2[,i])$p.value
}
wrap_plots(plots, nrow=2) +
  plot_layout(nrow = 2, ncol = 2, widths = c(2, 2), heights = c(2, 2))
```

```
rownames(p_val) <- colnames(period_2)
colnames(p_val) <- "P Value"

kable(round(p_val, 2), caption = "Shapiro Wilk Test")
```

Chi-Squared Plots

```
library(ggplot2)
# chi-squared plots
d2 <- mahalanobis(period_2, colMeans(period_2), cov= cov(period_2))
d2 <- sort(d2)
qcp <-  qchisq((1:nrow(period_2) - 0.5)/nrow(period_2), ncol(period_2))

df1 <- data.frame(x=qcp, y=d2)
ggplot()+
  geom_point(data = df1, aes(x=x, y=y), color = "black")+
  geom_abline(intercept = 0, slope = 1, color="red") +
  labs(
    title = "Theoretical quantiles from a Chi-Squared with P = 4 degrees
of freedom (QCP) and Mahalanobis Distance",
    x = "QCP",
    y = "Mahalanobis Distance"
  ) +
  scale_x_continuous(
    breaks = seq(0, 14, by = 2),
    limits = c(0, 14)
  ) +
  scale_y_continuous(
    breaks = seq(0, 14, by = 2),
    limits = c(0, 14)
  ) +
```

```
  coord_fixed(ratio = 1) +
  theme(plot.title = element_text(size = 10, hjust = 0.5))
```

```
critical <- qchisq(0.5, df = 4)
proportion <- mean(d2 < critical)*100
```