STAT636 - Homework 2

Daniel Osorio - dcosorioh@tamu.edu
Department of Veterinary Integrative Biosciences
Texas A&M University

1. Find the maximum likelihood estimates of the 2×1 mean vector μ and the 2×2 covariance matrix Σ based on the random sample

$$X = \begin{bmatrix} 3 & 6 \\ 4 & 4 \\ 5 & 7 \\ 4 & 7 \end{bmatrix}$$

```
> X <- matrix(data = c(3,6,4,4,5,7,4,7), ncol = 2, byrow = TRUE)
> n <- nrow(X)
> X_hat <- 1/n * rep(1,n) %*% X
> X_hat

     [,1] [,2]
[1,]    4    6

> S_hat <- 1/n * (t(X) - drop(X_hat)) %*% t(t(X) - drop(X_hat))
> S_hat

     [,1] [,2]
[1,] 0.50 0.25
[2,] 0.25 1.50
```

2. Let $X_1, X_2, \ldots, X_{60}$ be a random sample of size $n = 60$ from a $N_6(\mu, \Sigma)$ population. Specify each of the following.

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

$$(X_1 - \mu)'\Sigma^{-1}(X_1 - \mu) \sim \mathcal{X}_6^2$$

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

$$\bar{X} \sim \mathcal{N}_6\left(\mu, \frac{1}{60}\Sigma\right)$$

$$\sqrt{n}(\bar{X} - \mu) \dot{\sim} \mathcal{N}_6(0, \Sigma)$$

   (c) The distribution of $n(\bar{X} - \mu)'\Sigma^{-1}(\bar{X} - \mu)$

$$n(\bar{X} - \mu)'\Sigma^{-1}(\bar{X} - \mu) \sim \mathcal{X}_6^2$$

   (d) The approximate distribution of $n(\bar{X} - \mu)'S^{-1}(\bar{X} - \mu)$

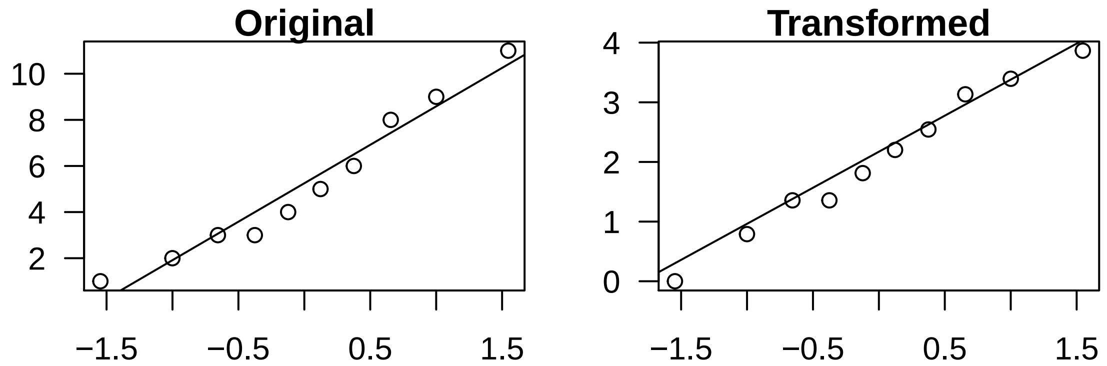
$$n(\bar{X} - \mu)'S^{-1}(\bar{X} - \mu) \dot{\sim} \mathcal{X}_6^2$$

3. Consider the `used_car` data. For each of 10 used cars, we have the numeric variables Age (age of the car) and Price (sale price of car, in $1,000s)

```
> used_car <- read.csv("used_cars.csv")
```

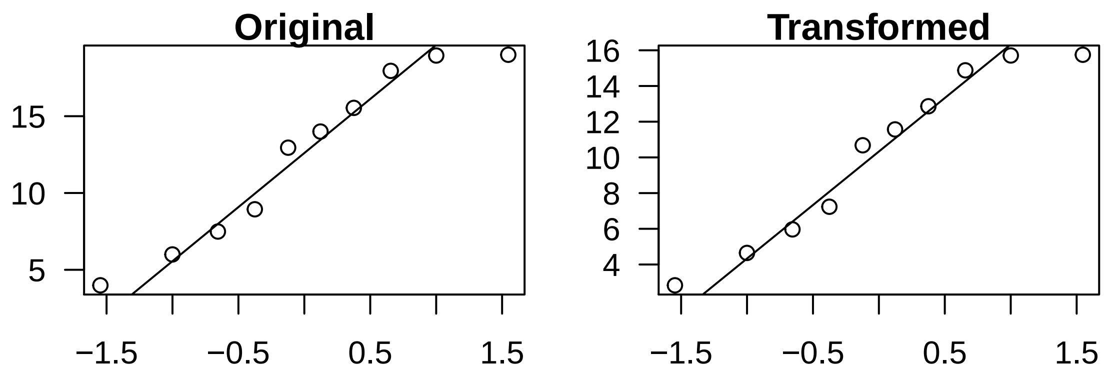
(a) Determine the power transformation $\hat{\lambda}_1$ that makes the x_1 values approximately normal. Construct a Q-Q plot for the transformed data.

```
> lambda <- as.numeric(car::powerTransform(used_car[,1])$lambda)
> lambda
[1] 0.3708906
> par(mfrow = c(1,2), mar=c(2.5,2.5,1,1))
> qqnorm(used_car[,1], main = "Original", las=1)
> qqline(used_car[,1])
> transformed <- ((used_car[,1] ^ lambda) - 1)/lambda
> qqnorm(transformed, main = "Transformed", las=1)
> qqline(transformed)
```



(b) Determine the power transformation $\hat{\lambda}_2$ that makes the x_2 values approximately normal. Construct a Q-Q plot for the transformed data.

```
> lambda <- as.numeric(car::powerTransform(used_car[,2])$lambda)
> lambda
[1] 0.9361967
> par(mfrow = c(1,2),mar=c(2.5,2.5,1,1))
> qqnorm(used_car[,2], main = "Original", las=1)
> qqline(used_car[,2])
> transformed <- ((used_car[,2] ^ lambda) - 1)/lambda
> qqnorm(transformed, main = "Transformed", las=1)
> qqline(transformed)
```



(c) Determine the power transformations $\hat{\lambda}' = \left[\hat{\lambda}_1, \hat{\lambda}_2\right]$ that make the $[x_1, x_2]$ values approximately multivariate normal. Compare the results with those from above.

```
> car::powerTransform(used_car)
```

```
Estimated transformation parameters
      Age     Price
1.2732157 0.0310405
```

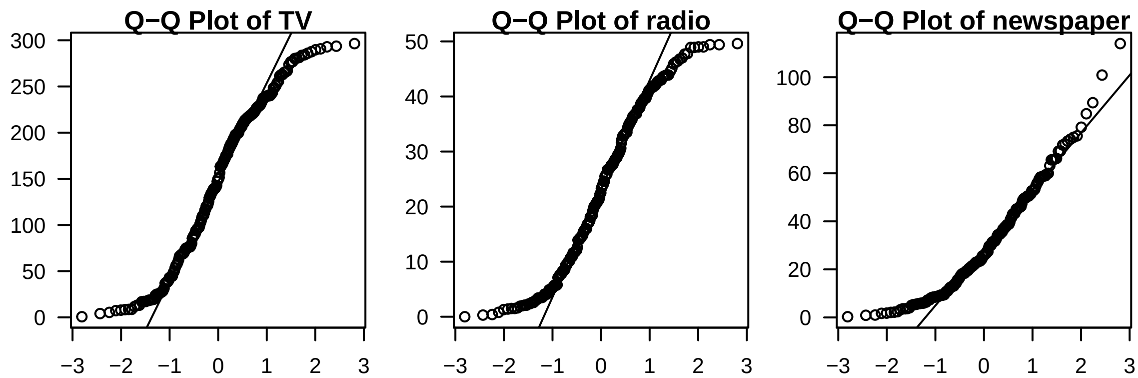
The values of λ required to approximate to the multinormal distribution are different to those computed independently for each variable.

4. Consider the `advertising` data. For each of 200 strategies, we have three numeric variables that influence the sales: `TV`, `radio`, and `Newspaper`.

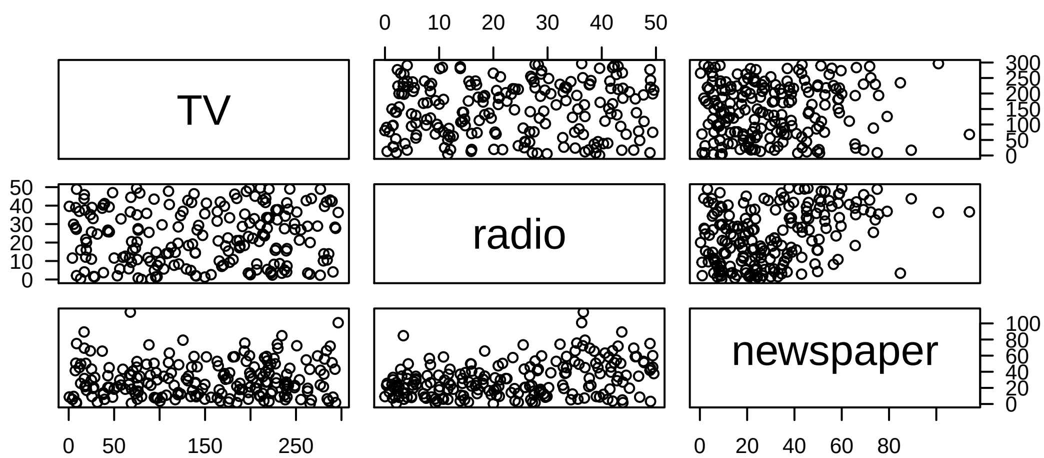
```
> advertising <- read.csv("advertising.csv", row.names = 1)
```

(a) Construct univariate Q-Q plots for each of the three variables. Also make the three pairwise scatterplots. Does the multivariate normal assumption seem reasonable?

```
> par(mfrow = c(1,3), mar=c(2.5,2.5,1,1))
> out <- sapply(colnames(advertising[,1:3]), function(x){
+   qqnorm(advertising[,x], main = paste0("Q-Q Plot of ",x), las = 1)
+   qqline(advertising[,x])
+ })
```




```
> par(mar=c(1,1,1,1))
> plot(advertising[,1:3], las = 1)
```



The multivariable normal distribution does not look reasonable for this dataset because of not all the variables are independently normal

(b) Determine the 95% confidence ellipsoid for μ. Where is it centered? What are its axes and corresponding half-lengths?

(c) Compute 95% T2 simultaneous confidence intervals for the three mean components.

(d) Compute 95% Bonferroni simultaneous confidence intervals for the three mean components.

(e) Carry out a Hotelling's T^2 test of the null hypothesis H0 : $\mu' = [150.0, 20.0, 30.0]$ at $\alpha = 0.05$. What is the test statistic, critical value, and the p-value? What is your conclusion regarding H0?

(f) Is $\mu' = [150.0, 20.0, 30.0]$ inside the 95% confidence ellipse you computed in part (b)? Is this consistent with your findings in part (e)? Hint: It should be.

(g) Use the bootstrap to test the same null hypothesis as in part (e), now using this as your test statistic

$$\Lambda = \left(\frac{|S|}{|S_0|}\right)^{n/2},$$

where

$$S = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})(x_j - \bar{x})'$$

is the sample covariance matrix, and

$$S_0 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})(x_j - \bar{x})'$$

is the sample covariance matrix under the assumption that H_0 is true. So that all our answers match, first do `set.seed(2)`, and use $B = 500$ bootstrap iterations. What is the p-value?

```
>
```