Tarea 2 Estadística Multivariada

Code

Hairo Ulises Miranda Belmonte

07 de Febrero del 2019

EJERCICIO 1

Demuestre que $\Sigma_{j=1}^{n}(x_j-\mu)(x_j-\mu)^{'}=\Sigma_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})^{'}+n(\bar{x}-\mu)(\bar{x}-\mu)^{'}$

$$\Sigma_{j=1}^{n}(x_j-\mu)(x_j-\mu)^{'}=$$

Sumando y restando $\bar{x}$

$$=\Sigma_{j=1}^{n}(x_j-\bar{x}+\bar{x}-\mu)(x_j-\bar{x}+\bar{x}-\mu)^{'}$$

Acomodando términos:

$$=\Sigma_{j=1}^{n}(x_j-\bar{x}+\bar{x}-\mu)(x_j-\bar{x}+\bar{x}-\mu)^{'}$$

$$=\Sigma_{j=1}^{n}[(x_j-\bar{x})(\bar{x}-\mu)][(x_j-\bar{x})^{'}+(\bar{x}-\mu)^{'}]$$

$$=\Sigma_{j=1}^{n}[(x_j-\bar{x})(x_j-\bar{x})^{'}+(x_j-\bar{x})(\bar{x}-\mu)^{'}+(\bar{x}-\mu)(x_j-\bar{x})^{'}+(\bar{x}-\mu)(\bar{x}-\mu)^{'}]$$

En donde los términos cruzados son cero; entonces:

$$=\Sigma_{j=1}^{n}[(x_j-\bar{x})(x_j-\bar{x})^{'}+(\bar{x}-\mu)(\bar{x}-\mu)^{'}]$$

$$=\Sigma_{j=1}^{n}(x_j-\bar{x})(x_j-\bar{x})^{'}+n(\bar{x}-\mu)(\bar{x}-\mu)^{'}$$

EJERCICIO 2

Demuestre que $\Sigma=\frac{1}{2b}B$ entonces $\frac{1}{|\Sigma|^b}e^{-tr(\Sigma^{-1}B)/2}\leq\frac{1}{B}^{2bpb}e^{-pb}$ cumple la igualdad.

Separando el problema en dos:

$$tr(\Sigma^{-1}B) = tr(2bB^{-1}B) = tr(2bI) = 2b$$

$$\frac{1}{|\Sigma|^b} = \frac{|B^{\frac{1}{2}}\Sigma^{-1}B^{\frac{1}{2}}|}{|B|}$$

$$= \frac{|\Sigma^{-1}B|}{|B|} = \frac{|2bB^{-1}B|}{|B|} = \frac{|2bI|}{|b|} = \frac{(2b)^p}{|B|}$$

Entonces;

$$= \left(\frac{(2b)^p}{|B|}\right)^b e^{-2bp/2}$$

$$= \frac{(2b)^{pb}}{|B|^b} e^{-bp}$$

EJERCICIO 3

Vea la razón del por qué la distancia generalizada se puede ver como una elipse. Justifique el resultado para $p = 2$

Sea; $0 < (distancia)^2 = X^{'}AX$ para $x \neq 0$, donde $X^{'}AX$ es definido positivo, y A es una matriz simétrica; por lo tanto, la distancia de X a un punto fijo, sea el μ, es: $(x-\mu)^{'}A(x-\mu)$. Si se expresa esta distancia en la raíz cuadrada, se tiene una interpretación geométrica en base a la descomposición espectral de A.

Suponga $p = 2$. Los puntos de $X^{'} = [X_1, X_2]^{'}$ a una distancia, sea “c”, satisface:

$$\begin{pmatrix} X_1 & X_2 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = c$$

$$\begin{pmatrix} a_{11}X_1 + a_{12}X_2 & a_{12}X_1 + a_{22}X_2 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = c$$

$$a_{11}X_1^2 + a_{12}X_1X_2 + a_{12}X_1X_2 + a_{22}X_2^2$$

$$a_{11}X_1^2 + 2a_{12}X_1X_2 + a_{22}X_2^2 = c^2$$

Sin embargo, por el teorema de la descomposición espectral:

$$A = \lambda_1 e_1 e_1^{'} + \lambda_2 e_2 e_2^{'}$$

entonces;

$$X^{'}AX = X^{'}(\lambda_1 e_1 e_1^{'} + \lambda_2 e_2 e_2^{'})X$$

$$= X^{'}\lambda_1 e_1 e_1^{'} X + X^{'}\lambda_2 e_2 e_2^{'} X$$

$$= \lambda_1 X^{'} e_1 e_1^{'} X + \lambda_2 X^{'} e_2 e_2^{'} X$$

$$= \lambda_1 (X^{'} e_1)^2 + \lambda_2 (X^{'} e_2)^2$$

de esta manera;

$$c^2 = \lambda_1 y_1^2 + \lambda_2 y_2^2$$

es una elipse, en $y_1 = x^{'} e_1$ y $y_2 = x^{'} e_2$ porque $\lambda_1, \lambda_2 > 0$ cuando $X^{'} AX > 0$.

donde, al despejar terminos:

$$c^2 = \lambda_1 (x_1^{'} e_1)^2$$

$$c = \sqrt{(}\lambda_1)(x_!^{'} e_1)$$

$$c\lambda_1^{-\frac{1}{2}} = (x_1^{'} e_1)$$

y sabemos que $y_1 = x_1^{'} e_1 = x e_1^{'}$; entonces:

$$c\lambda_1^{-\frac{1}{2}} = x e_1^{'}$$

multiplicando por la derecha e_1, y sabiendo que $e_1^{'} e_1 = 1$, entonces:

$$x_1 = c\lambda_1^{-\frac{1}{2}} e_1$$

De la misma forma se realiza para x_2

Por lo tanto, formas cuadraticas, o en nuestro caso, las raíz cuadrada de la distancia generalizada es;

$$c^2 = \frac{y_1^2}{\sqrt{\lambda}_1} + \frac{y_2^2}{\sqrt{\lambda}_2}$$

$$x_1 = c\sqrt{(}\lambda_1) e_1$$

y por otro lado:

$$x_2 = c\sqrt{(}\lambda_2) e_2$$

Pero sabemos que los valores propios para Σ^{-1} son $\frac{1}{\lambda_i}$; entonces:

$$x_1 = \frac{c}{\sqrt{\lambda}_1} e_1$$

$$x_2 = \frac{c}{\sqrt{\lambda_2}} e_2$$

EJERCICIO Extra en clase

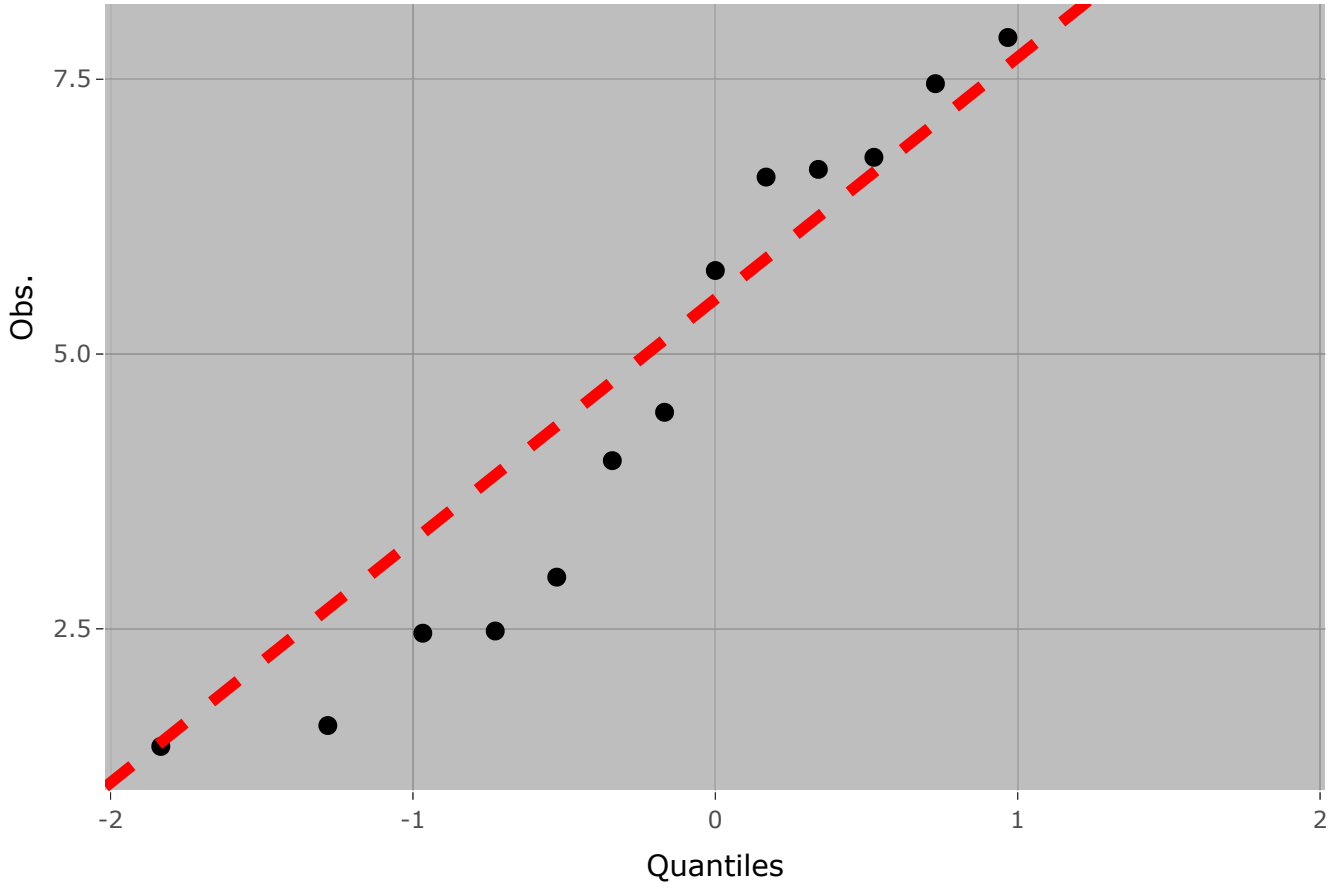
Realiza el QQ-plot con las observaciones que se encuentran en la diapositiva.

Code

Obs	adjProb	Quantile
1.43	0.0333333	-1.8339146
1.62	0.1000000	-1.2815516
2.46	0.1666667	-0.9674216
2.48	0.2333333	-0.7279133
2.97	0.3000000	-0.5244005
4.03	0.3666667	-0.3406948
4.47	0.4333333	-0.1678940
5.76	0.5000000	0.0000000
6.61	0.5666667	0.1678940
6.68	0.6333333	0.3406948
6.79	0.7000000	0.5244005
7.46	0.7666667	0.7279133
7.88	0.8333333	0.9674216
8.92	0.9000000	1.2815516
9.42	0.9666667	1.8339146

Code

Ejercicio extra. QQ-plot



Prueba de normalidad basada en la rectitud del QQ-plot

Code

Coeficiente de correlación de Pearson
0.975

para un $n = 15$, los puntos críticos de la prueba son; 0.9503 en $\alpha = 0.10$, 0.938 para $\alpha = 0.05$, y 0.9126 para $\alpha = 0.01$. Por lo tanto, no se tiene evidencia suficiente para rechazar la hipótesis nula de normalidad, a cualquier vaor mayor de α más grande de 0.01.

EJERCICIO 4

En climas nórdicos, las carreteras debe ser limpiadas de la nieve rápidamente después de una tormenta. Una de las medidas de la severidad de la tormenta es x1 = duración en horas, mientras que la efectividad de la limpieza de la nieve se puede cuantificar por x2 = horas de trabajo para limpiar la nieve. En la tabla inferior se muestran los resultados de 25 incidentes en Wisconsin.

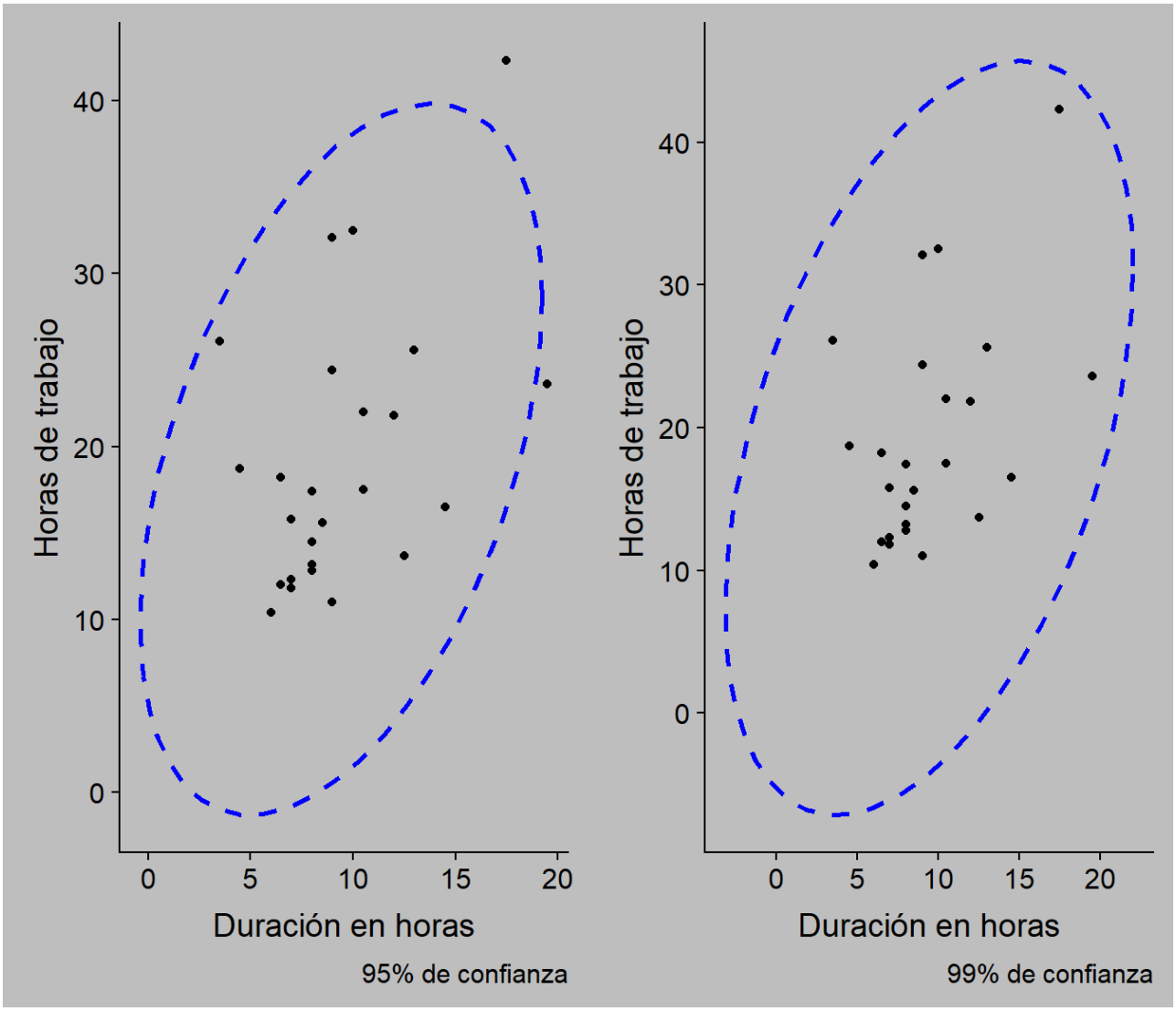
a. Detecte cualquier posible dato atípico mediante el diagrama de dispersión de las variables originales.

Code

x1	x2
12.5	13.7
14.5	16.5
8.0	17.4
9.0	11.0
19.5	23.6
8.0	13.2
9.0	32.1
7.0	12.3
7.0	11.8
9.0	24.4
6.5	18.2
10.5	22.0
10.0	32.5
4.5	18.7
7.0	15.8
8.5	15.6
6.5	12.0
8.0	12.8
3.5	26.1
8.0	14.5
17.5	42.3
10.5	17.5
12.0	21.8
6.0	10.4

x1	x2
13.0	25.6

Code



Como se puede ver en el cuadro de arriba, se realiza un gráfico de dispersión para las variables; horas de trabajo y duración en horas, en el cual, al gráficar una región de confianza al 99% (cuadro de la derecha), las observaciónes en conjunto se distribuyen como una normal bivariada; no obstante, si reducimos el nivel de confianza; i.e, al 95% de significancia, se puede observar dos datos atípicos.

b. Determine la potencia de la transformación λ_1 que convierte los valores de x_1 aproximadamente a normales. Construya el Q-Q plot de las observaciones transformadas.

Observaciones x_1 =Duración en horas.

Code

LVal	Lambda
-30.14	0.05
-30.14	0.06
-30.14	0.04
-30.14	0.07
-30.14	0.03
-30.14	0.08

Code

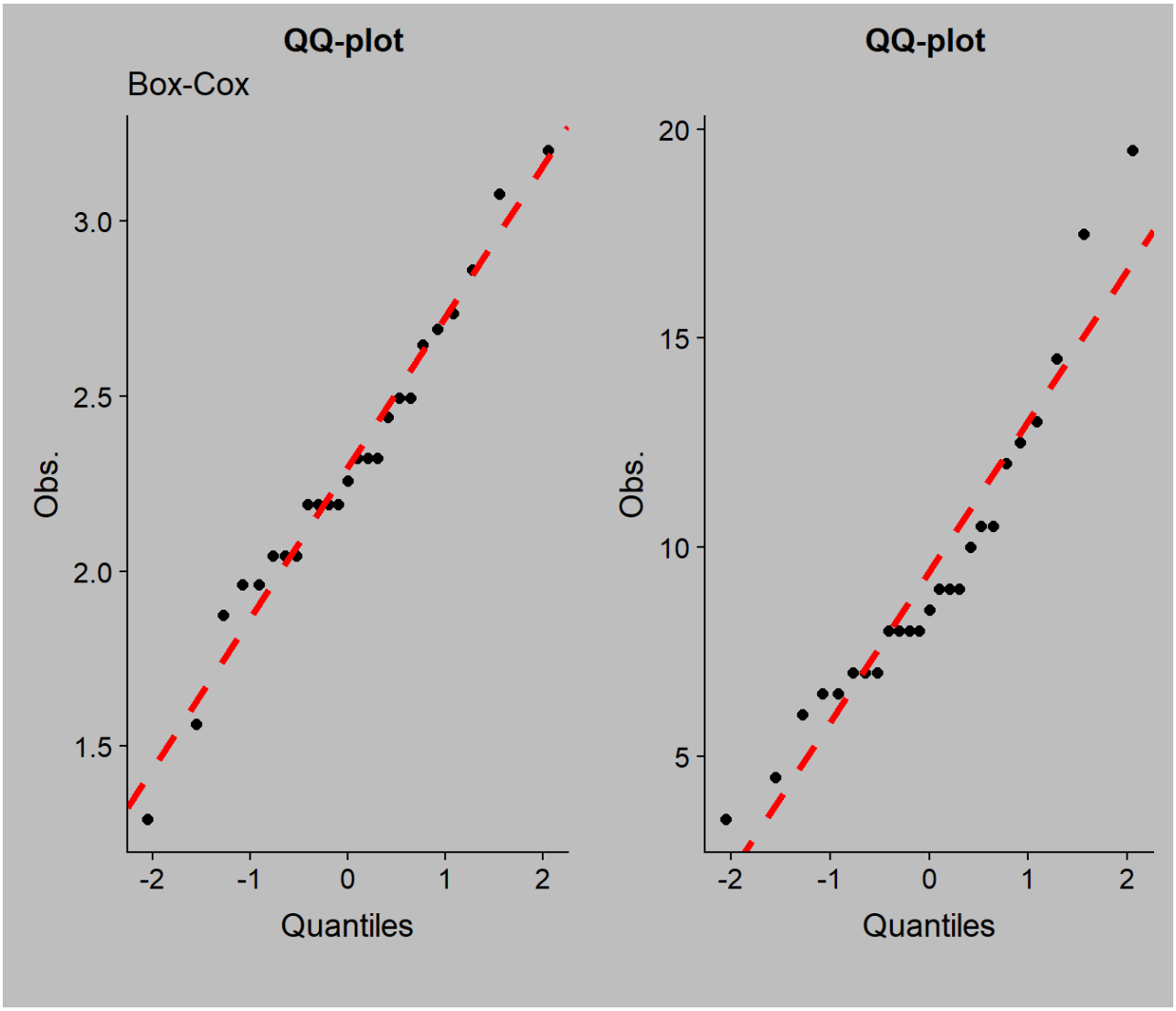
En el gráfico anterior se gráficaron los posibles valores del parámetro de transformación, en los cuales, al evaluarlos en la función a máximizar, se observa que el valor máximo se encuentra cuando $\gamma(\lambda) = -30.14$ y $\lambda = 0.05$.

A continuación, se realiza la transformación de los datos, propuesta por box-cox; cabe mencionar, que no se cuenta con observaciones de $x_1 = 0$, y por lo tanto, la transformación viene dada por:

$$x^{\lambda} = \frac{x^{\lambda-1}}{\lambda}$$

Realizando QQ-plot a la transformación de los datos

Code



Como se puede observar, bajo la tranformación realizada mediante la metodología de box-cox, la serie se ajusta mejor a la linea de 45°; i.e., podemos sugerir que bajo la transformación se considera una distribución normal.

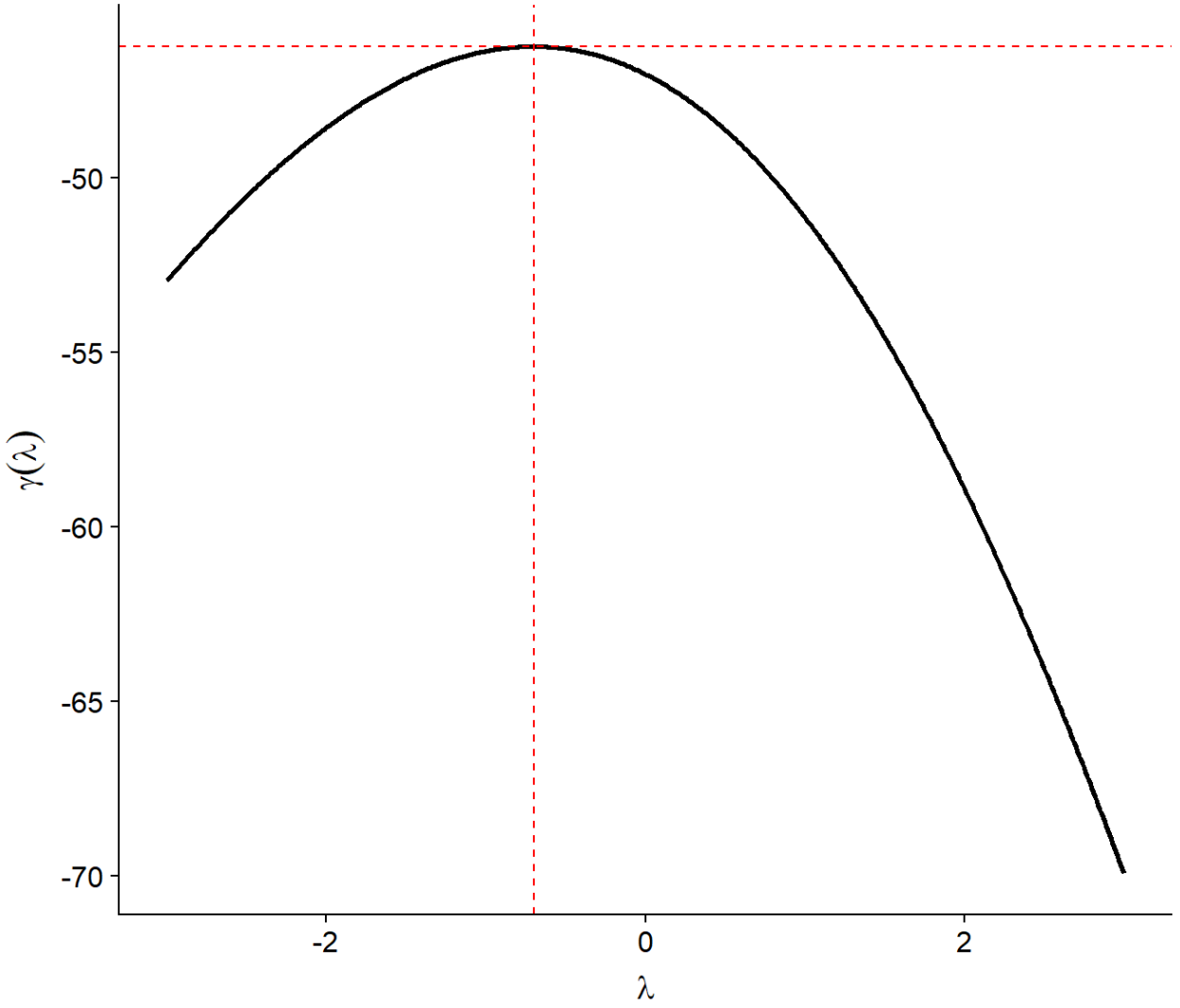
c. Determine la potencia de la transformación que convierte los valores de x_2 aproximadamente a normales. Construya el Q-Q plot de las observaciones transformadas.

Observaciones x_2 = Horas de trabajo.

Code

LVal	Lambda
-46.23	-0.70
-46.23	-0.71
-46.23	-0.69
-46.23	-0.72
-46.23	-0.68
-46.23	-0.73

Code

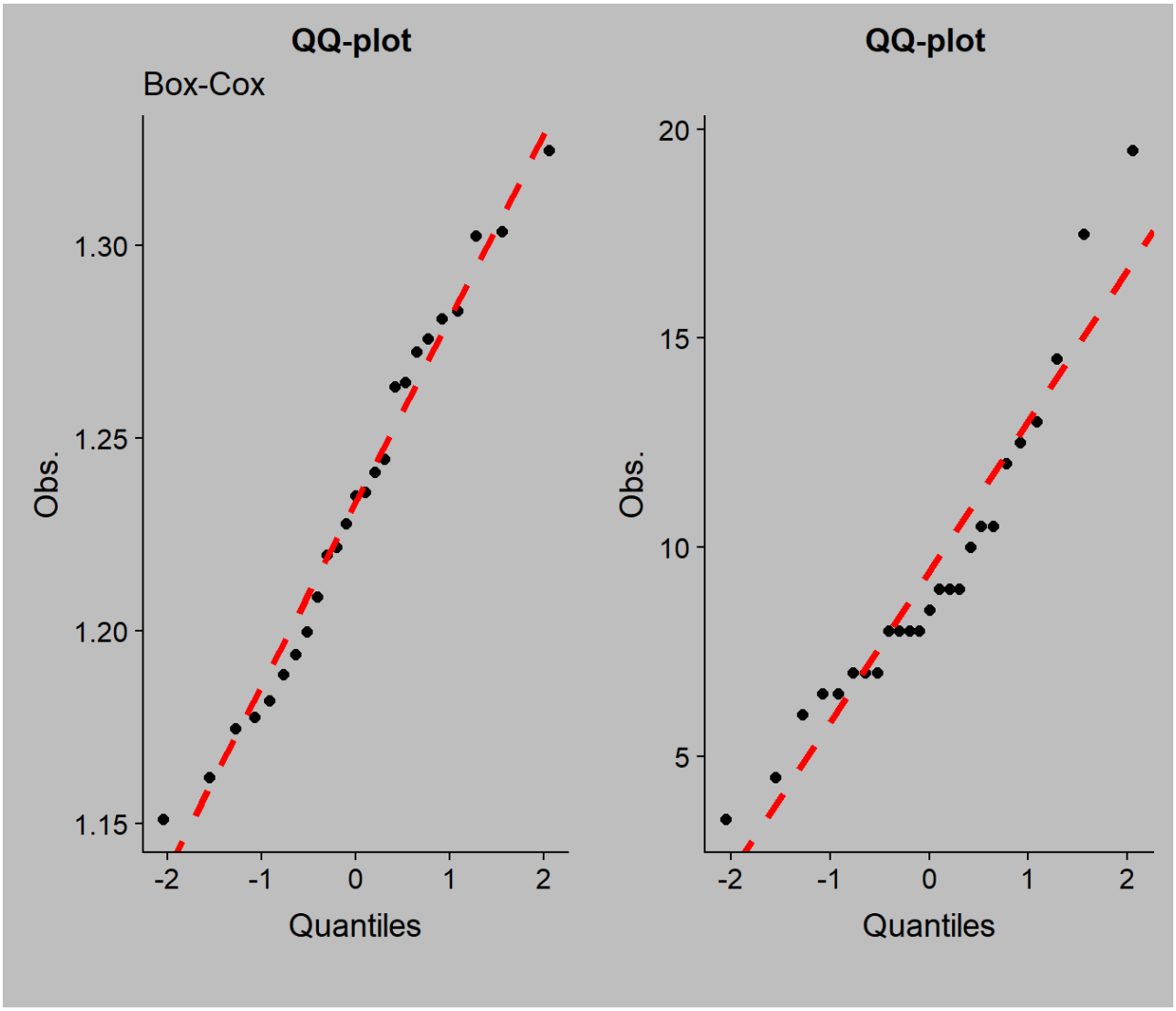


En el gráfico anterior se observan los posibles valores del parámetro de transformación, en los cuales, al evaluarlos en la función a máximizar, se observa que el valor máximo se encuentra cuando $()=-46.23 $ y $\lambda = -0.70$.

Ahora, realizamos la transformación antes mencionada a las observaciones x_2, horas trabajadas.

Realizando QQ-plot a la transformación de los datos

Code



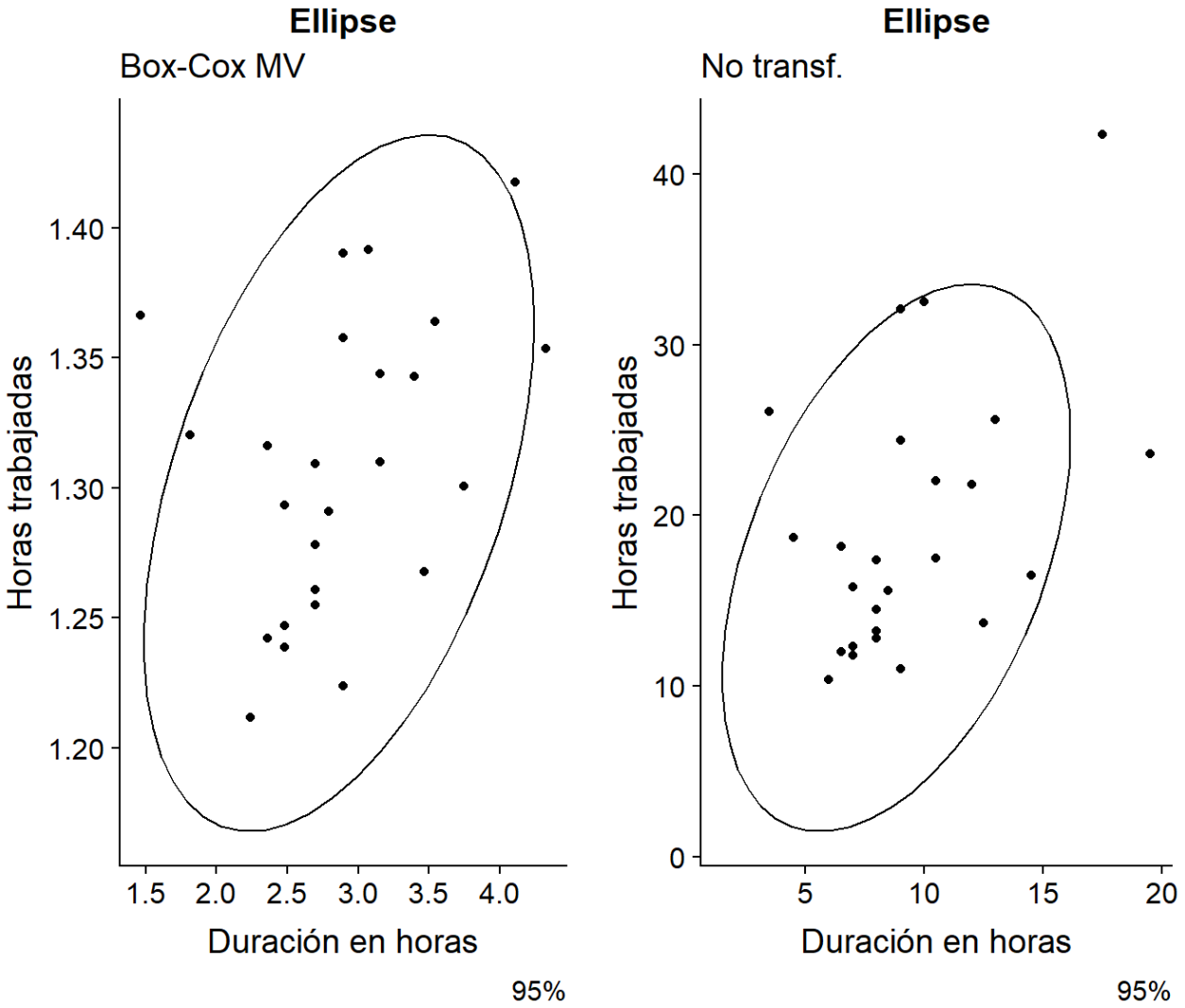
Al igual que x_1, la variable x_2, horas de trabajo, bajo la transformación propuesta por boxcox, se observa que sigue una distribución normal.

d. Determine la potencia de la transformación que convierte las observaciones bivariadas en aproximadamente normales.

Code

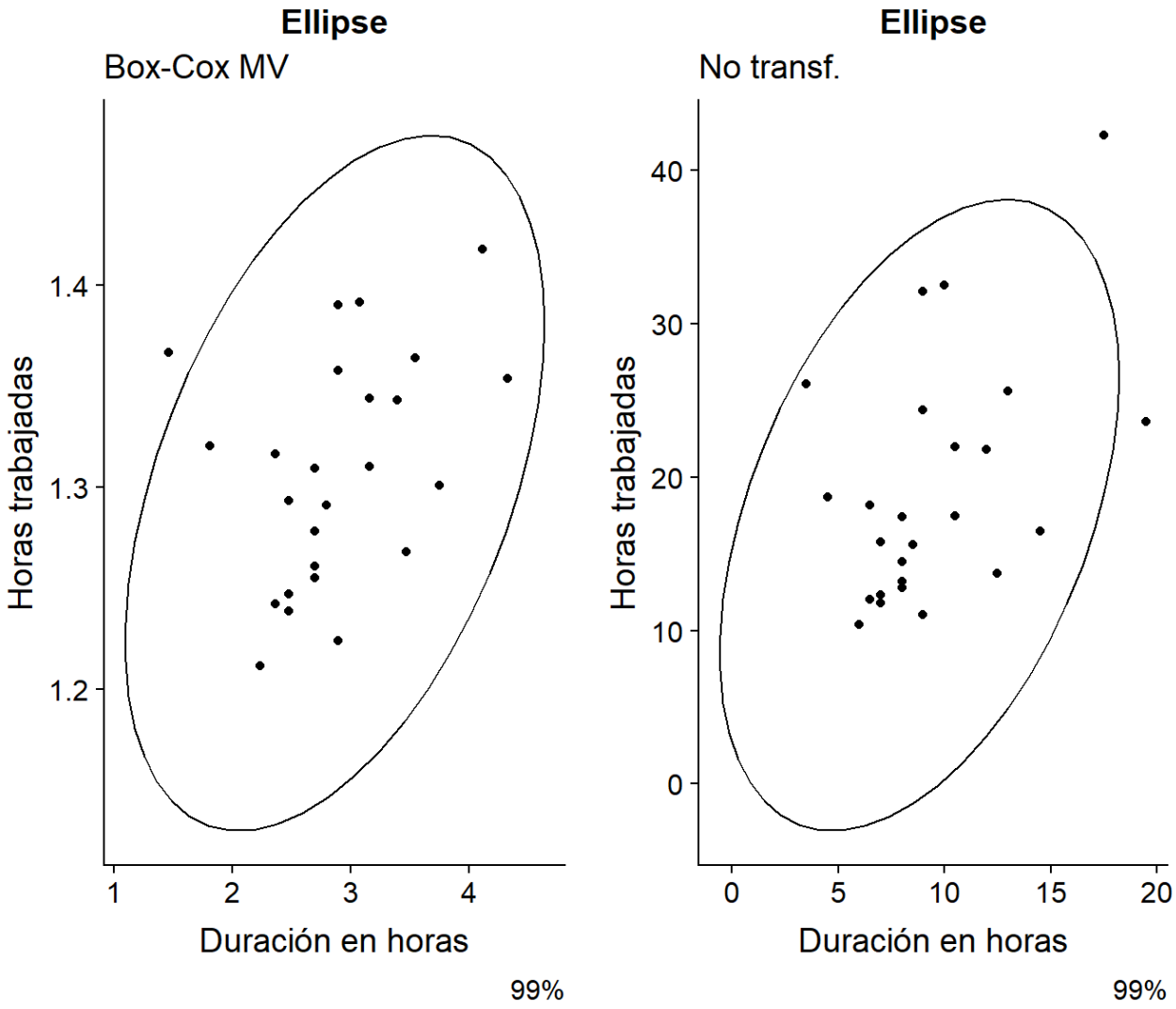
	Parámetro de Transformación
x1	0.2396
x2	-0.6417

Code



Se utilizaron los parámetros de transformación de las series univariadas para iniciar el algoritmo; i.e., el box-cox bivariado. Asimismo, se dibuja un intervalo de confianza bivariado - elipse-, al 95% de significancia. Lo que podemos observar, en el grafíco anterior , es que bajo una transformación, el conjunto de observaciones sigue teniendo valores fuera de la región de confianza; sin embargo, la distancia -eucllidina-, entre las observaciones que salen de la región y un posible centroide, podría eser menor.

Code

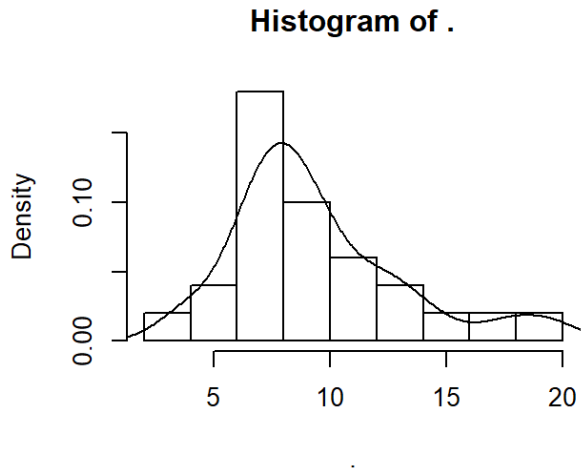


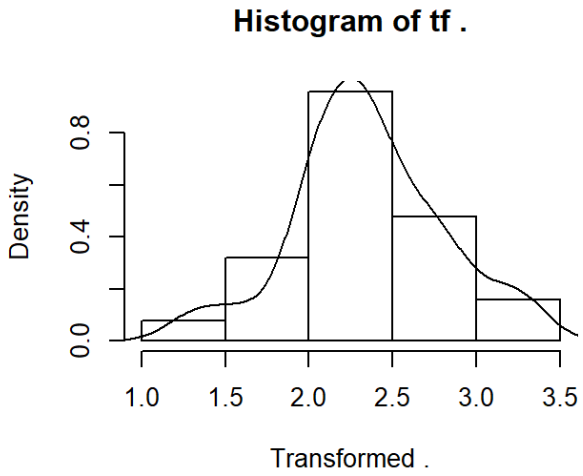
Al incrementar al 99% de significancia la elipse, podemos ver como bajo la transformación se distribuye normal bivariada. En conclusión, la transformación ayuda de manera adecuada para aproximar la distribución de x_1 y x_2, a una normal bivariada.

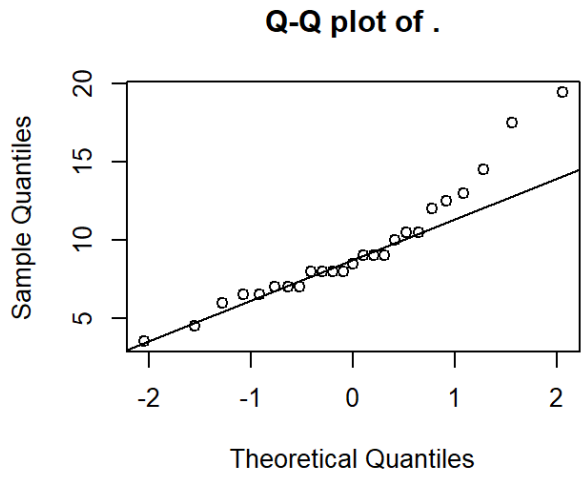
Lo anterior se realizó programando la transformación de box-cox, esto para la estimación univariado; de este modo, a continuación se presenta el mismo ejercicio del inciso (b)-(d), pero utilizando la paqueteria boxcoxnc; para los casos univariados.

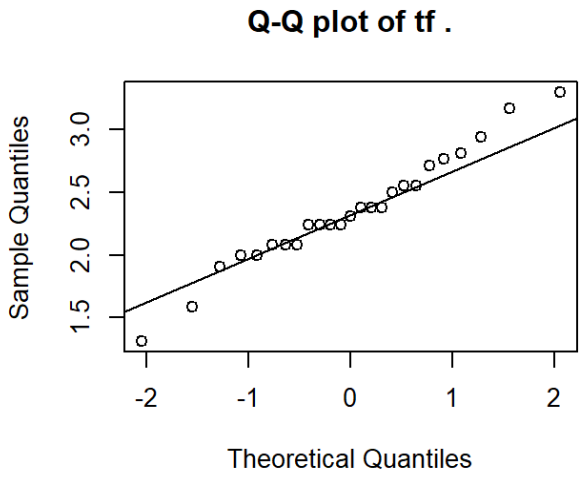
Para x_1: Duración en horas

Code







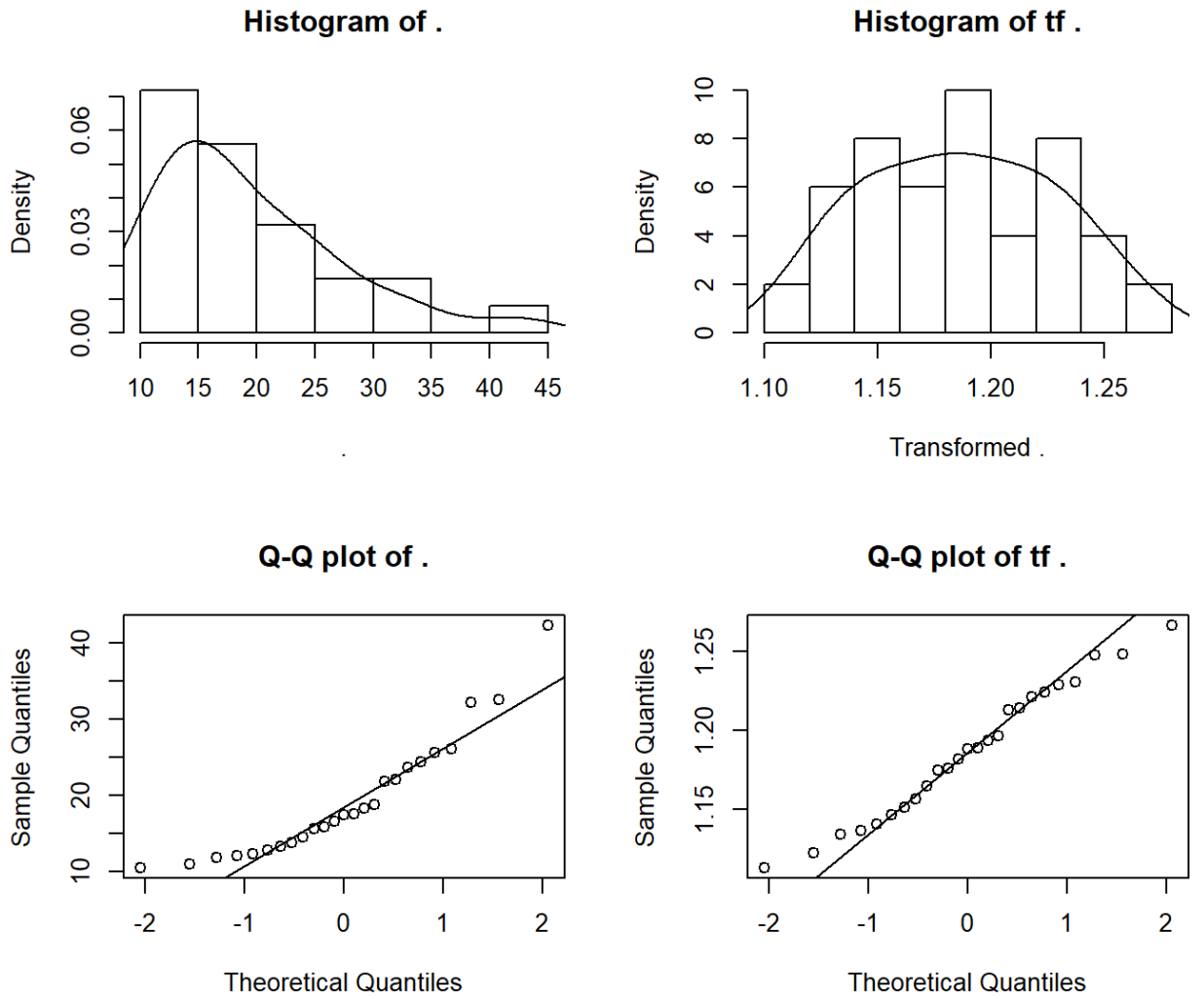



```
##
##   Box-Cox power transformation
## ------------------------------------------------------------------
##   data : .
##
##   lambda.hat : 0.07
##
##
##   Shapiro-Wilk normality test for transformed data (alpha = 0.05)
## ------------------------------------------------------------------
##
##   statistic  : 0.9783
##   p.value    : 0.8496
##
##   Result     : Transformed data are normal.
## ------------------------------------------------------------------
```

Podemos observar que los valores del estadístico de transformación, es de 0.07, y en el caso de cuando se programó, erá de 0.05, lo cual no realiza muchos cambios en nuestras conclusiones.

Para x_2: Horas trabajadas

Code



```
##
##   Box-Cox power transformation
## ------------------------------------------------------------------
##   data : .
##
##   lambda.hat : -0.74
##
##
##   Shapiro-Wilk normality test for transformed data (alpha = 0.05)
## ------------------------------------------------------------------
##
##   statistic  : 0.9747
##   p.value    : 0.7633
##
##   Result     : Transformed data are normal.
## ------------------------------------------------------------------
```

Para este caso, observamos que el valor del parámetro de transformación estomado es el mismo al que programamos.

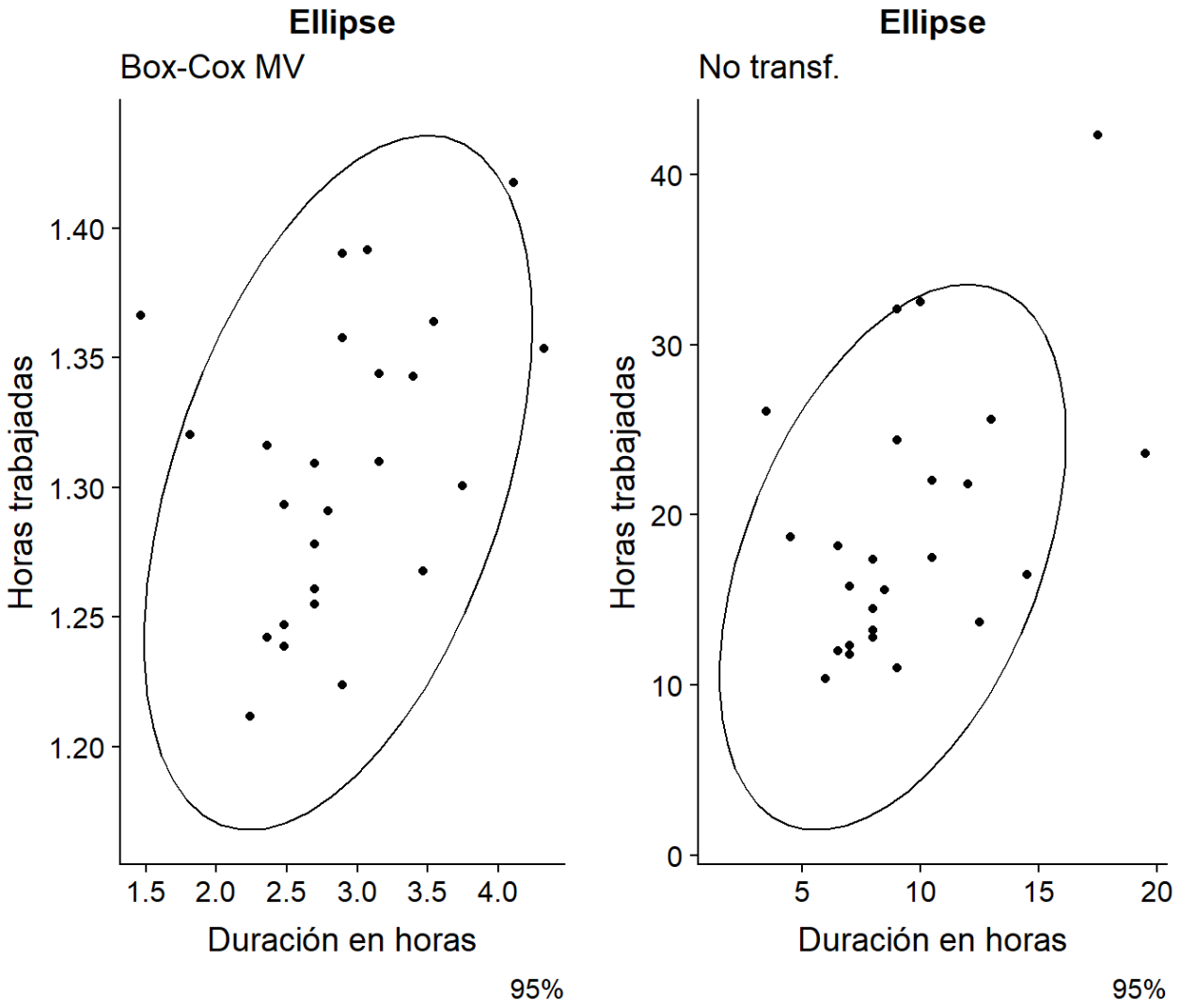
Realizando la versión bivariada con los parámetros de la función boxcoxnc.

Code

	Parámetro de Transformación
x1	0.2396
x2	-0.6417

Realizando la transformación bivariada Nota: se utilizan los parámetros de transformación de los casos univariados para inicializar las iteraciones; esto de acuerdo a la literatura.

Code



Como podemos observar, obtuvimos los mismos resultados utilizando la funciones boxcoxnc, que programando la función para estimar los parámetros de formas univariada, y después pasarlo a la función, que si utilizamos libreria, de box.cox multivariado.

EJERCICIO 5

Para p y n fijos, genérese una muestra de tamaño N de una ley $T^2(p,n)$ de Hotelling. Para esto construya una función que tome como entradas los valores de p, N, y utilice un generador de números aleatorios gaussianos. Represénte los resultados mediante un histograma, y haga pruebas para diferentes valores de entrada

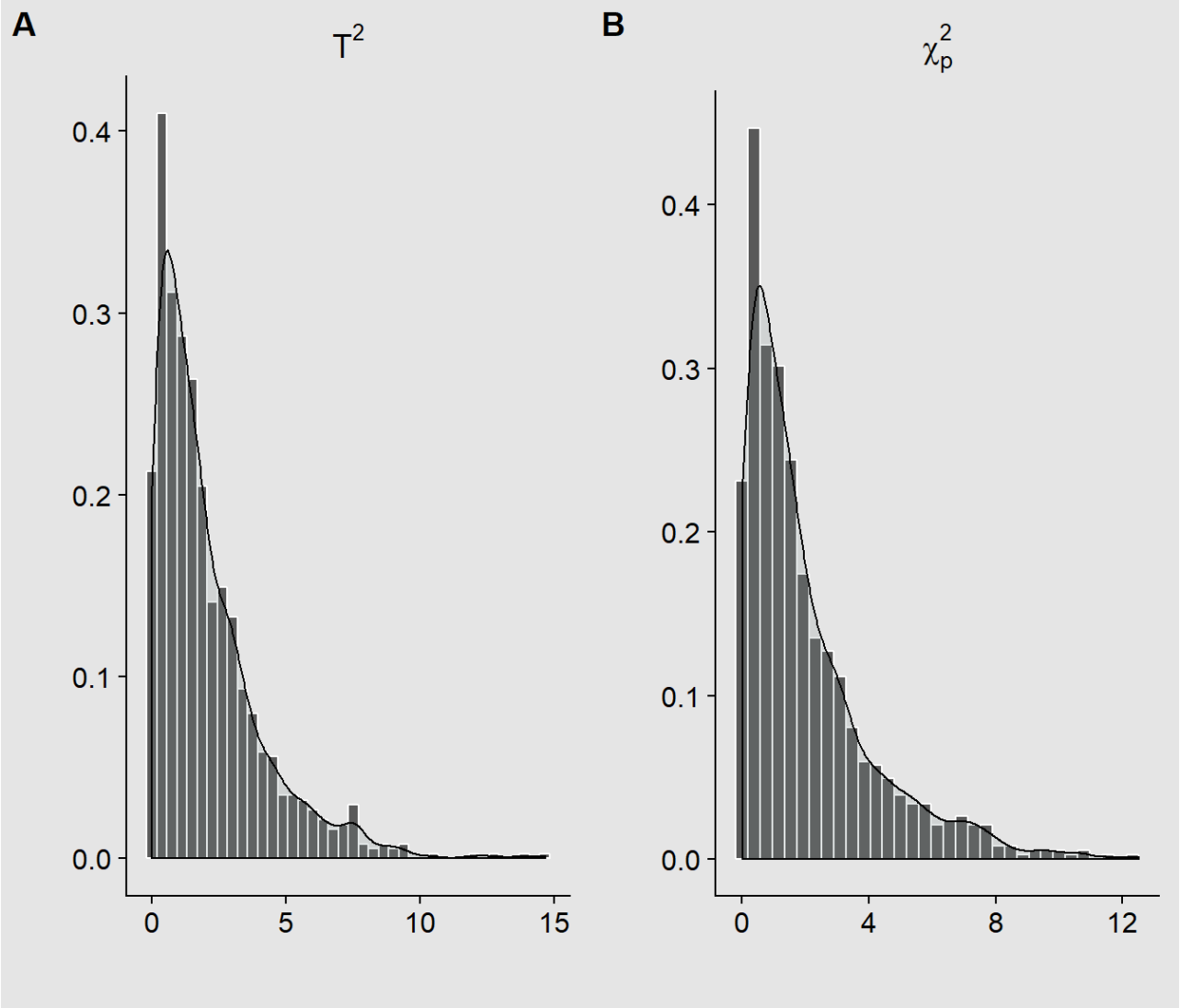
La función se encuentra a continuación. Como se sabe, si $x \sim N_p(\mu, \Sigma)$ y $(n-1)S \sim W_p(S|\Sigma)$; entonces, la distribución de la variable escalar $T^2 = (x-\mu)' S^{-1}(x-\mu)$, tiene una distribución Hotelling con p y $n-1$ grados de libertad $T^2 \sim T^2(p, n-1)$. Asimismo, bajo el TLC, sabemos que S (converge en probabilidad); entonces, T^2 converge a la distancia de Mahalandas, y sabemos que esa distancia se aproxima a una χ^2_p cuando n.

Code

De esta manera, utilizamos la función generamos observaciones de una normal bivariada, con $n = 1000$, y la contrastamos contra una χ^2_p.

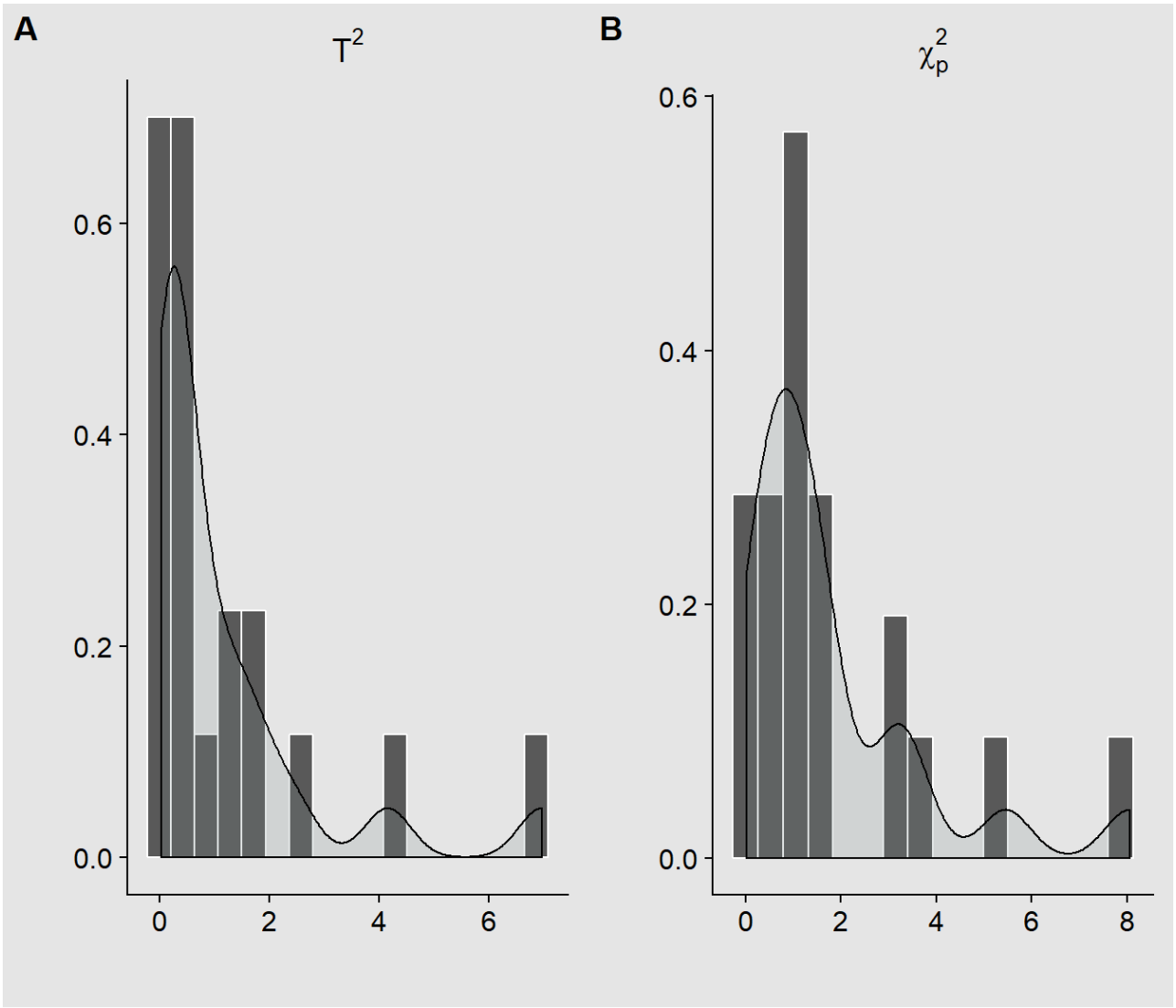
Code

Code



Ahora realizamos lo anteriorpero con $n = 20$, en el cual el TLC no tendrá efecto alguno.

Code


A
T^2
0.6
0.4
0.2
0.0
0
2
4
6
B
χ^2_p
0.6
0.4
0.2
0.0
0
2
4
6
8