Trabajo Multivariado

Parcial 2

Ana María López - Pedro Pablo Villegas

Octubre, 2017

INTRODUCCIÓN

El supuesto de normalidad esta siempre presente en los análisis estadisticos sea univariado o multivariado. Métodos univariados como: análisis de varianza (ANOVA), regresión lineal, entre otros, se basan en la distribución normal, adicionalmente técnicas estadísticas multivariantes como: análisis multivariado de varianza (MANOVA), análisis de componentes principales (PCA), análisis de discriminantes y otros, se basan en el supuesto de normalidad multivariante para hacer inferencias. Estas suposiciones requieren un conjunto de datos sobre los cuales una prueba estadística de la significancia sea aproximadamente distribuido de manera normal. Por lo tanto es importante contar con técnicas que permitan comprobar este supuesto. Para muestras grandes no es de gran preocupación por el teorema del límite central, el cual indica que la distribución de las medias sigue aproximadamente una distribución normal.(Oppong and Agbedra 2016)

Si bien se cuentan con técnicas para poder determinar si se cumple o no la hipótesis de normalidad, es importante conocer que se debe hacer cuando esta hipótesis es rechazada, para estos casos existen tipos de transformación de datos que puede llegar a hacer estos aproximadamente normales, de tal manera que los métodos estándar sean aplicables (D. Peña and Peña 1986). Un tipo de transformación es el enfoque de Box y Cox, estos modificaron la familia de transformaciones sugerida por Tukey (1957), teniendo en cuenta la discontinuidad en $\lambda = 0$ tal que:

$$y_i^{\lambda} = \begin{cases} (y_i^{\lambda} - 1)/\lambda; \lambda \neq 0 \\ log(y_i); \lambda = 0 \end{cases}$$

Esta transformación es valida para $y_i > 0$, si se tienen observaciones negativas se han realizado otras modificaciones a la transformación de Box y Cox.(Sakia 1992)

En el presente trabajo se realizan unos ejercicios que abarcan problema de evaluar la normalidad en casos univariados y multivariados, usando herramientas de análisis de graficos Q-Q, pruebas de Shapiro, transformaciones de poder Box Cox para transformar datos no normales a normales, entre otros.

EJERCICIOS PROPUESTOS

Punto 4.28

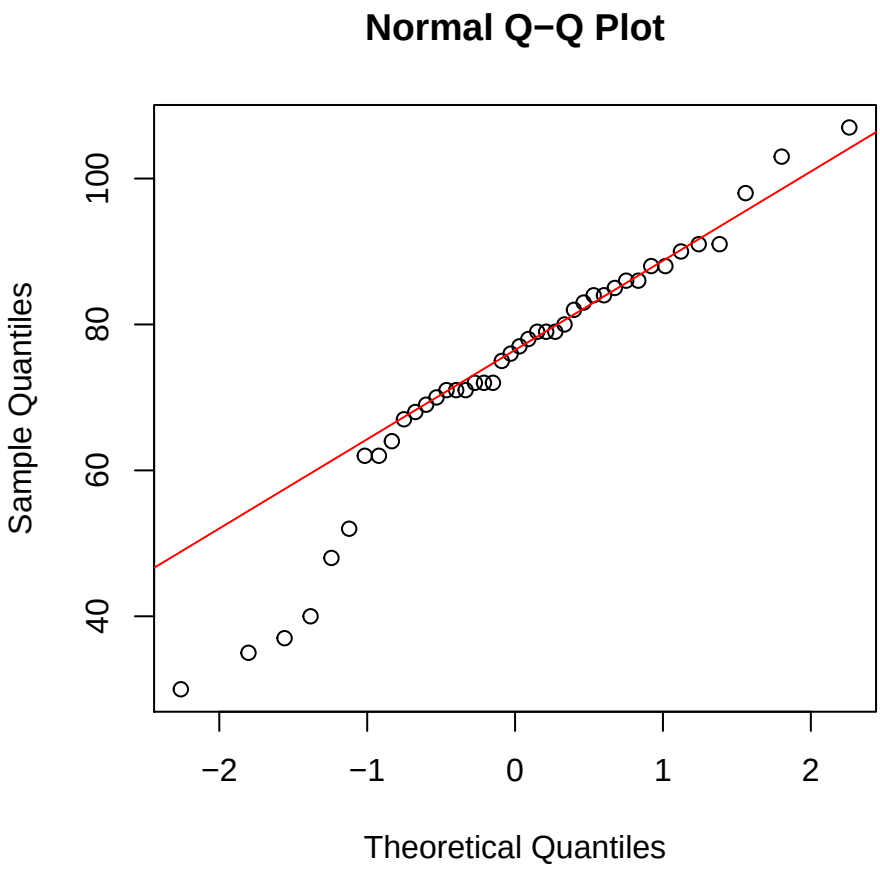
Considere los datos dados en la tabla 1.5 de polución del aire. Construya una grafica Q-Q para las medidas de radiación solar y lleve a cabo un test de normalidad basado en el coeficiente de correlacion r_Q. Defina $\alpha = 0.05$ y use la entrada correspondiente a $n = 40$ en la tabla 4.2.

Table 1.5 Air-Pollution Data

Wind (x_1)	Solar radiation (x_2)	CO (x_3)	NO (x_4)	NO_2 (x_5)	O_3 (x_6)	HC (x_7)
8	98	7	2	12	8	2
7	107	4	3	9	5	3
7	103	4	3	5	6	3
10	88	5	2	8	15	4
6	91	4	2	8	10	3
8	90	5	2	12	12	4
9	84	7	4	12	15	5
5	72	6	4	21	14	4
7	82	5	1	11	11	3
8	64	5	2	13	9	4
6	71	5	4	10	3	3
6	91	4	2	12	7	3
7	72	7	4	18	10	3
10	70	4	2	11	7	3
10	72	4	1	8	10	3
9	77	4	1	9	10	3
8	76	4	1	7	7	3
8	71	5	3	16	4	4
9	67	4	2	13	2	3
9	69	3	3	9	5	3
10	62	5	3	14	4	4
9	88	4	2	7	6	3
8	80	4	2	13	11	4
5	30	3	3	5	2	3
6	83	5	1	10	23	4
8	84	3	2	7	6	3
6	78	4	2	11	11	3
8	79	2	1	7	10	3
6	62	4	3	9	8	3
10	37	3	1	7	2	3
8	71	4	1	10	7	3
7	52	4	1	12	8	4
5	48	6	5	8	4	3
6	75	4	1	10	24	3
10	35	4	1	6	9	2
8	85	4	1	9	10	2
5	86	3	1	6	12	2
5	86	7	2	13	18	2
7	79	7	4	9	25	3
7	79	5	2	8	6	2
6	68	6	2	11	14	3
8	40	4	3	6	5	2

Source: Data courtesy of Professor G. C. Tiao.

Se toman los datos de la radiación solar los cuales son todos positivos y se construye la grafica Q-Q teniendo el siguiente resultado:



En el grafico Q-Q se evidencia que los datos podrían no provenir de una distribución normal, los puntos de la izquierda sugieren observaciones atípicas, pues estan muy lejos del resto de los datos y los puntos de la derecha tambien se alejan de la linea, estos puntos extremos también nos sugieren unas colas más pesadas que la distribución normal, por lo tanto con este grafico podría sugerirse que los datos de la medida de la radiación solar no se distribuyen normal.

Table 4.2 Critical Points for the Q–Q Plot Correlation Coefficient Test for Normality

Sample size	Significance levels α		
n	.01	.05	.10
5	.8299	.8788	.9032
10	.8801	.9198	.9351
15	.9126	.9389	.9503
20	.9269	.9508	.9604
25	.9410	.9591	.9665
30	.9479	.9652	.9715
35	.9538	.9682	.9740
40	.9599	.9726	.9771
45	.9632	.9749	.9792
50	.9671	.9768	.9809
55	.9695	.9787	.9822
60	.9720	.9801	.9836
75	.9771	.9838	.9866
100	.9822	.9873	.9895
150	.9879	.9913	.9928
200	.9905	.9931	.9942
300	.9935	.9953	.9960

Para confirmar esto se calcula el coeficiente de correlación r_Q el cual da como resultado $r_Q = 0.9693$, este valor es menor comparado con el punto critico de la forma correspondiente a la muestra de tamaño $n = 40$ de la tabla 4.2 el cual es 0.9726, por lo tanto este test sugiere rechazar la hipotesis nula de normalidad con un nivel de significancia $\alpha = 0.05$.

```
## [1] 0.9693258
```

Otro test que puede emplearse es el test de Shapiro, este nos arroja un $p - value$ que nos permite definir si cumple o no la hipotesis de normalidad, en este caso el test de Shapiro-Wilk nos da un $p - value = 0.0262$ el cual es menor que el nivel de significancia $\alpha = 0.05$, por lo tanto nuevamente obtenemos que se debería rechazar la hipotesis de normalidad.

```
##
##  Shapiro-Wilk normality test
##
## data:  data
## W = 0.93883, p-value = 0.02601
```

Punto 4.29

Dado los datos de la polución del aire de la tabla 1.5, examine los pares $X_5 = NO_2$ y $X_6 = O_3$ para normalidad bivariada.

- Calcule la distancia estadistica $(x_j - \bar{x})' S^{-1} (x_j - \bar{x})$, $j = 1, 2, ..., 42$, donde $x_j' = [x_{j5}, x_{j6}]$.
  Se realiza el calculo de la distancia estadistica mahalanobis, obteniendo el siguiente resultado:

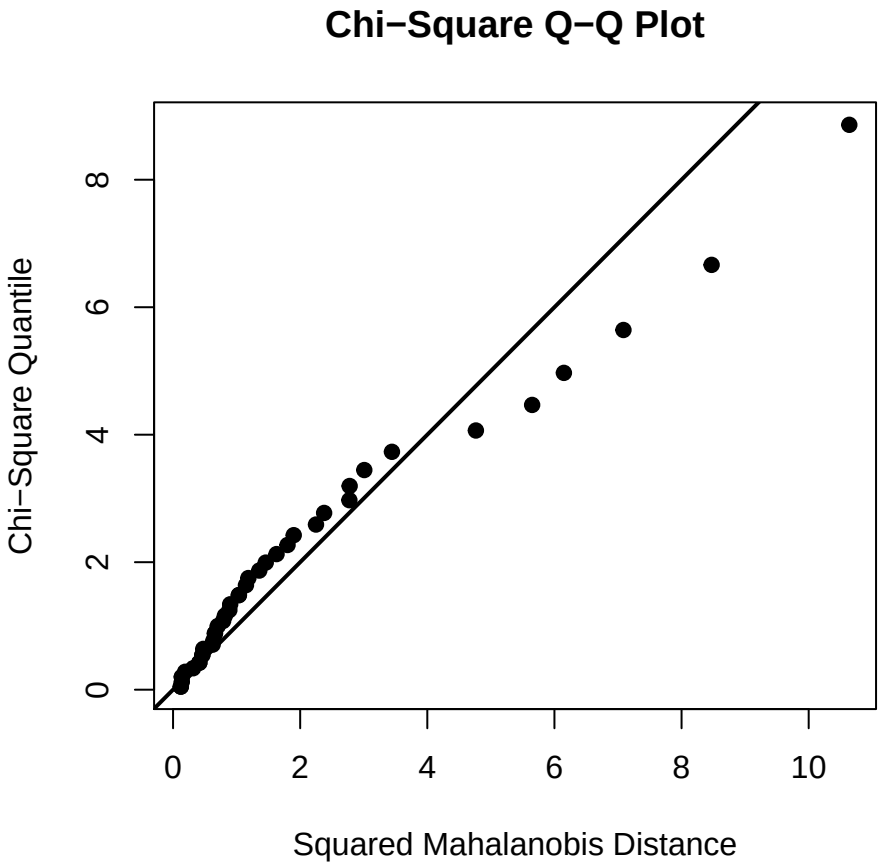
```
##  [1]  0.1224973  0.1224973  0.1379719  0.1388339  0.1388339  0.1901188
##  [7]  0.3159498  0.4135364  0.4135364  0.4606524  0.4606524  0.4760726
## [13]  0.6228096  0.6370218  0.6592206  0.6592206  0.7032485  0.7874152
## [19]  0.8162468  0.8856041  0.8987982  1.0360061  1.0360061  1.1471939
## [25]  1.1848895  1.3566301  1.4584229  1.6282902  1.8013611  1.8984708
## [31]  2.2488867  2.3770610  2.7741416  2.7782596  3.0089122  3.4437748
## [37]  4.7646873  5.6494392  6.1488606  7.0857237  8.4730649 10.6391792
```

- Determine la proporción de observaciones $x'_j = [x_{j5}, x_{j6}]$, $j = 1, 2, ..., 42$, que caen entre el 50% del contorno de probabilidad de una distribución normal bivariada.
  Se realiza el calculo de la proporción de observaciones que caen en el 50% del contorno de probabilidad de una distribución bivariada, el resultado obtenido es:

```
## [1] 0.6190476
```

61.90% de las observaciones cae entre el 50% del contorno de probabilidad de una distribucion normal bivariada, se espera que las observaciones sean muy cercanas al 50%, sin embargo hay un 11.9% porciento más de observaciones, podríamos pensar que no son normal bivariada, para afirmar esto se construirá el grafico χ^2.

- Construya un grafico χ^2 de las distancias ordenadas del primer punto.



En el gráfico χ^2 se evidencia que los datos no siguen un patrón de linea recta, adicional podemos evidenciar datos atipicos, es decir que nuevamente nos encontramos con que los datos no distribuyen normal bivariado. Por ultimo el Royston test nos confirma el resultado.

```
##   Royston's Multivariate Normality Test
## ---------------------------------------------
```

```
##   data : data
##
##   H       : 17.10542
##   p-value : 0.0001930844
##
##   Result  : Data are not multivariate normal.
## --------------------------------------------
```

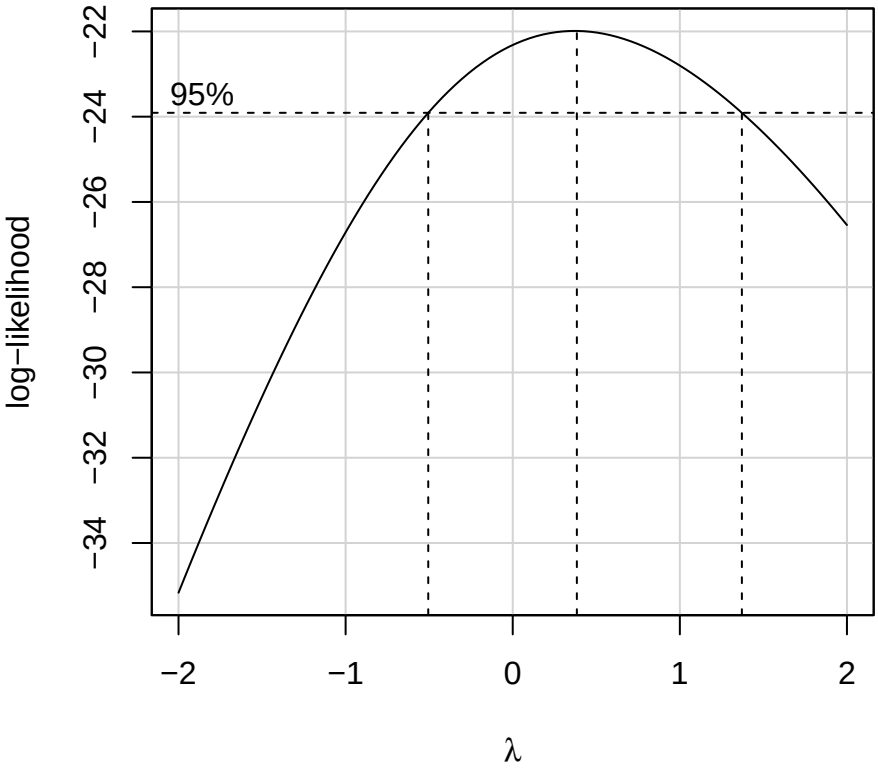
Punto 4.30

Considere los datos del carro usado del ejercicio 4.26.

```
##  [1]  1  2  3  3  4  5  6  8  9 11
```

```
##  [1] 18.95 19.00 17.95 15.54 14.00 12.95  8.94  7.49  6.00  3.99
```

- Determine la transformación de poder para $\widehat{\lambda}_1$ que hace los valores x_1 aproximadamente normales. Construya una grafica Q-Q para los datos transformados.
  Para el calculo del λ usamos la transformación de poder Box Cox, la cual nos permite graficamente mirar que valor de λ podría usarse para realizar la transformación de los datos.

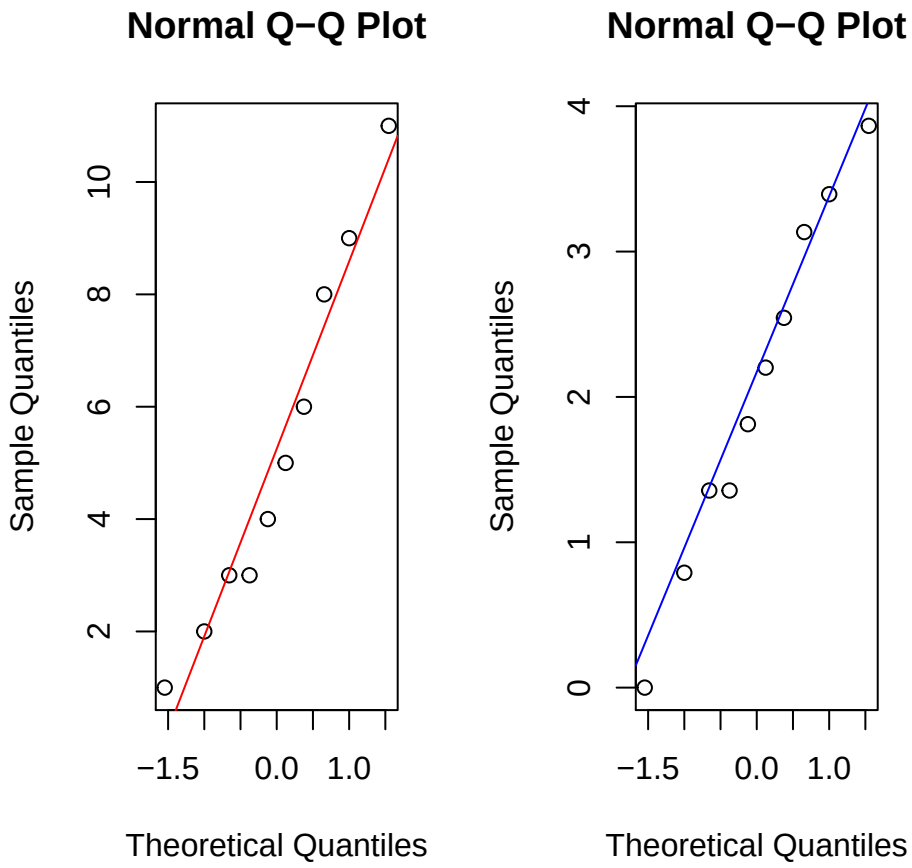


El valor de λ que sugiere la transformación de poder es $\lambda_1 = 0.3709$, por lo tanto se procede a realizar la transformacion de los datos usando este valor y teniendo en cuenta que es diferente de cero.

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr bnd Wald Upr Bnd
## Y1    0.3709           1      -0.5462       1.2879
##
## Likelihood ratio tests about transformation parameters
##                              LRT df      pval
```

```
## LR test, lambda = (0) 0.6604358  1 0.4164061
## LR test, lambda = (1) 1.6254078  1 0.2023394
```

Se construyen los graficos Q-Q de la distribucion de x_1 (Gráfico de la izquierda) y de la distribución transformada $\frac{x_1^{\lambda_1}-1}{\lambda_1}$ por ser $\lambda \neq 0$ (Gráfico de la derecha).



En los datos transformados las observaciones se aproximan mas a la linea, sin embargo no es una diferencia significativa, se realiza la prueba de Shapiro para los datos no transformados:

```
##
##  Shapiro-Wilk normality test
##
## data:  x1
## W = 0.94805, p-value = 0.6454
```

Adicional a esto se realiza este mismo test con los datos transformados:

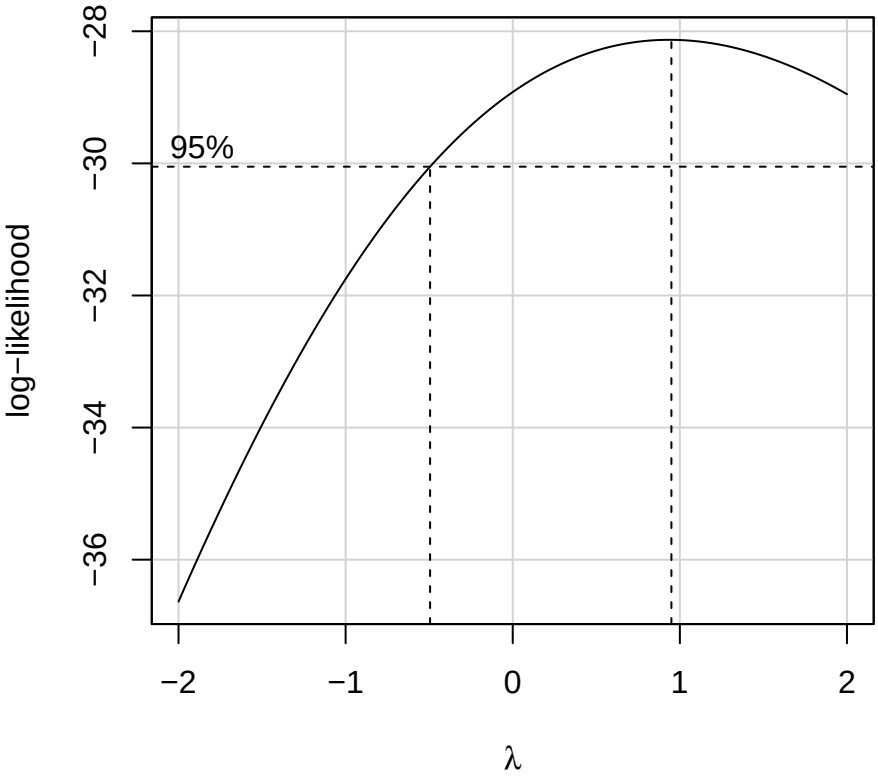
```
shapiro.test((x1^lambda1-1)/lambda1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  (x1^lambda1 - 1)/lambda1
## W = 0.97961, p-value = 0.963
```

De los resultados podemos concluir que transformar los datos no sería necesario, ya que encontramos en la prueba de shapiro que se puede aceptar la hipotesis nula de normalidad en ambos casos, teniendo un $p-value = 0.963$ para los datos transformados y un $p-value = 0.6465$ para los datos no transformados.

- Determine la transformación de poder para $\widehat{\lambda}_2$ que hace los valores x_2 aproximadamente normales. Construya una grafica Q-Q para los datos transformados.

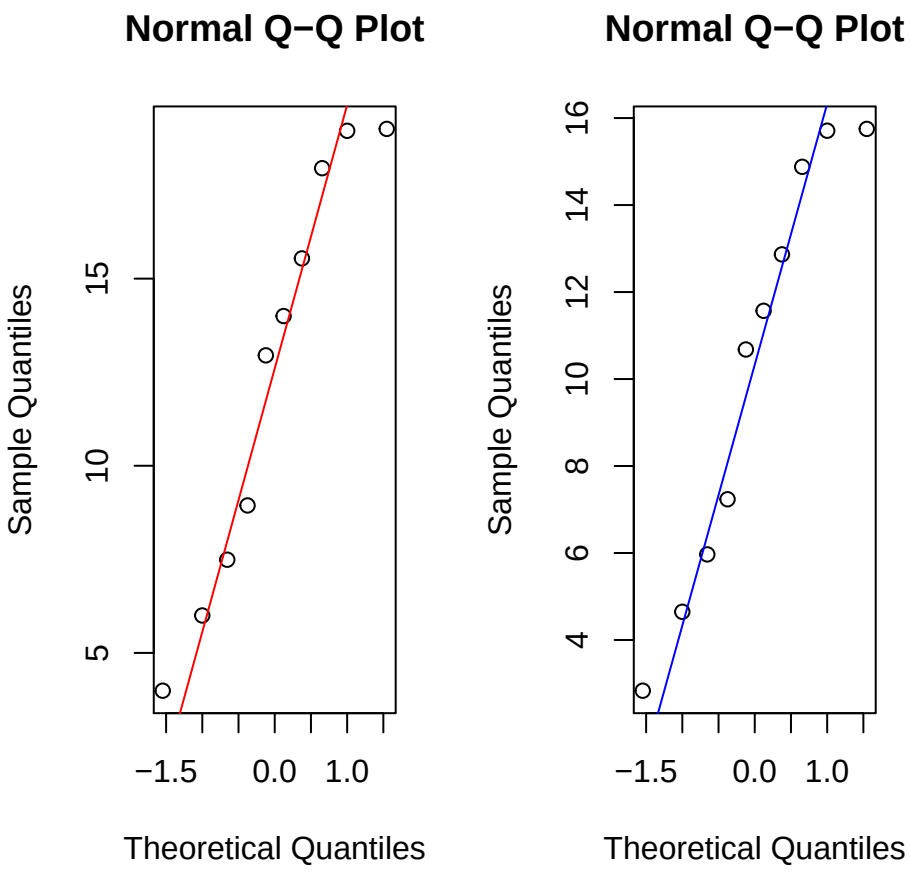
Para x_2 realizamos el mismo procedimiento del punto anterior.



Esta función nos devuelve información del máximo λ, siendo en esta caso $\lambda_2 = 0.9362$

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr bnd Wald Upr Bnd
## Y1    0.9362           1      -0.5933       2.4657
##
## Likelihood ratio tests about transformation parameters
##                                LRT df      pval
## LR test, lambda = (0) 1.581370085  1 0.2085635
## LR test, lambda = (1) 0.006638004  1 0.9350650
```

Este $\lambda_2 = 0.9362$ es muy cercada a 1, asi que no se necesita realizar ninguna transformación a los datos x_2 ya que estos podrían ser aproximadamente normales. Sin embargo se construye el grafico Q-Q para x_2 con los datos normales y transformados y no se encuentra ninguna diferencia.



Adicional se realiza la prueba de Shapiro para los datos brutos y los transformados:

```
shapiro.test(x2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x2
## W = 0.91801, p-value = 0.3406
```

```
shapiro.test((x2^lambda2-1)/lambda2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  (x2^lambda2 - 1)/lambda2
## W = 0.91804, p-value = 0.3409
```

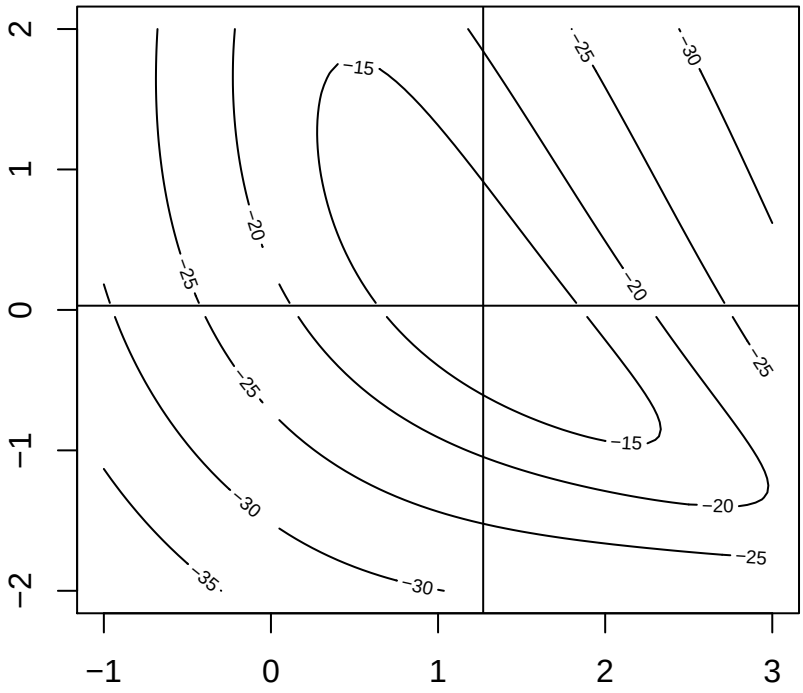
En este caso no se evidencia cambio significativo en la prueba por lo cual se concluye que no es necesario realizar una transformación de los datos.

- Determine la transformación de poder para $\widehat{\lambda}' = [\widehat{\lambda}_1, \widehat{\lambda}_2]$ que hace los valores $[x_1, x_2]$ normal conjunta usando (4-40). Compare los resultados obtenidos en los puntos 1 y 2. Realizamos el mismo procedimiento de los puntos anteriores con la variante de trabajar con una multivariada.

```
## bcPower Transformations to Multinormality
##    Est Power Rounded Pwr Wald Lwr bnd Wald Upr Bnd
## x1    1.2732           1       0.6707       1.8758
## x2    0.0310           0      -0.7008       0.7629
```

```
##
## Likelihood ratio tests about transformation parameters
##                                LRT df         pval
## LR test, lambda = (0 0) 21.937754  2 1.722969e-05
## LR test, lambda = (1 1)  5.881454  2 5.282731e-02
```

Tener las distribuciones marginales normales, no significa que la distribución conjunta sea de forma normal. Por la transformacion de Box-Cox, encontramos que los $\widehat{\lambda}' = [1.2732, 0.0310]$ son diferentes a los hallados en los puntos anteriores. Si consideramos el gráfico de contorno, se puede encontrar que $(0.3709, 0.9362)$ también caen en la región superior, esto significa que $(0.3709, 0.9362)$ podría ser una buena opción también.



Punto 4.34

Examine los datos sobre el contenido mineral oseo de la tabla 1.8 para marginale y bivariada normalidad.

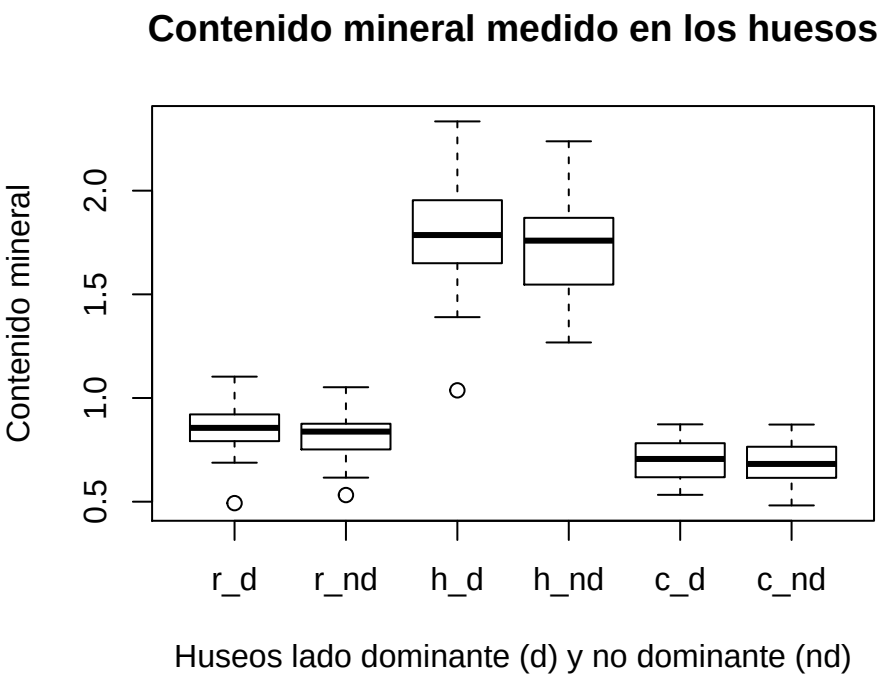
Sabemos que si un vector aleatorio X distribuye normal multivariado, cualquier subgrupo o combinación de los elementos de este, distibuyen normal, incluso sus distribuciones univariadas son normales. Por lo que la estrategia para evaluar la normalidad de los datos de la tabla 1.8 será evaluar primero la normalidad de las distribuciones marginales y luego la normalidad de las distribuciones bi-variadas.

Si esto se da hay fuertes evidencias de que el vector distribuye multivariado aunque esto no siempre es cierto pero los casos en los que no, no son comunes.

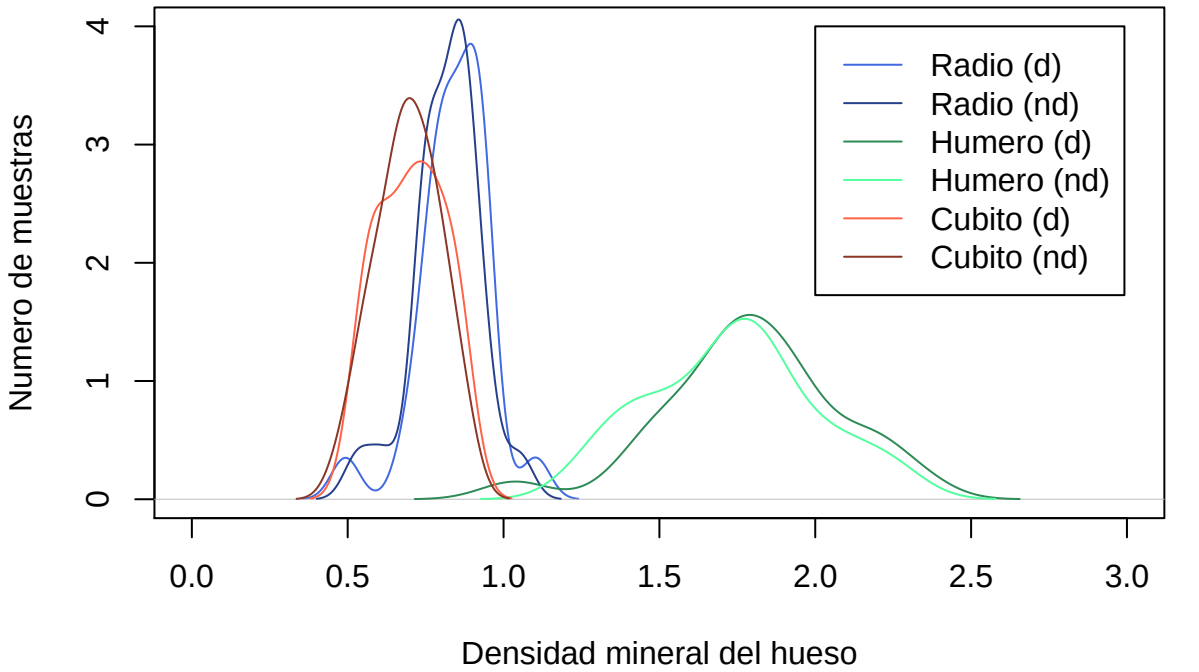
Se inicia el analisis exploratorio multivariado.

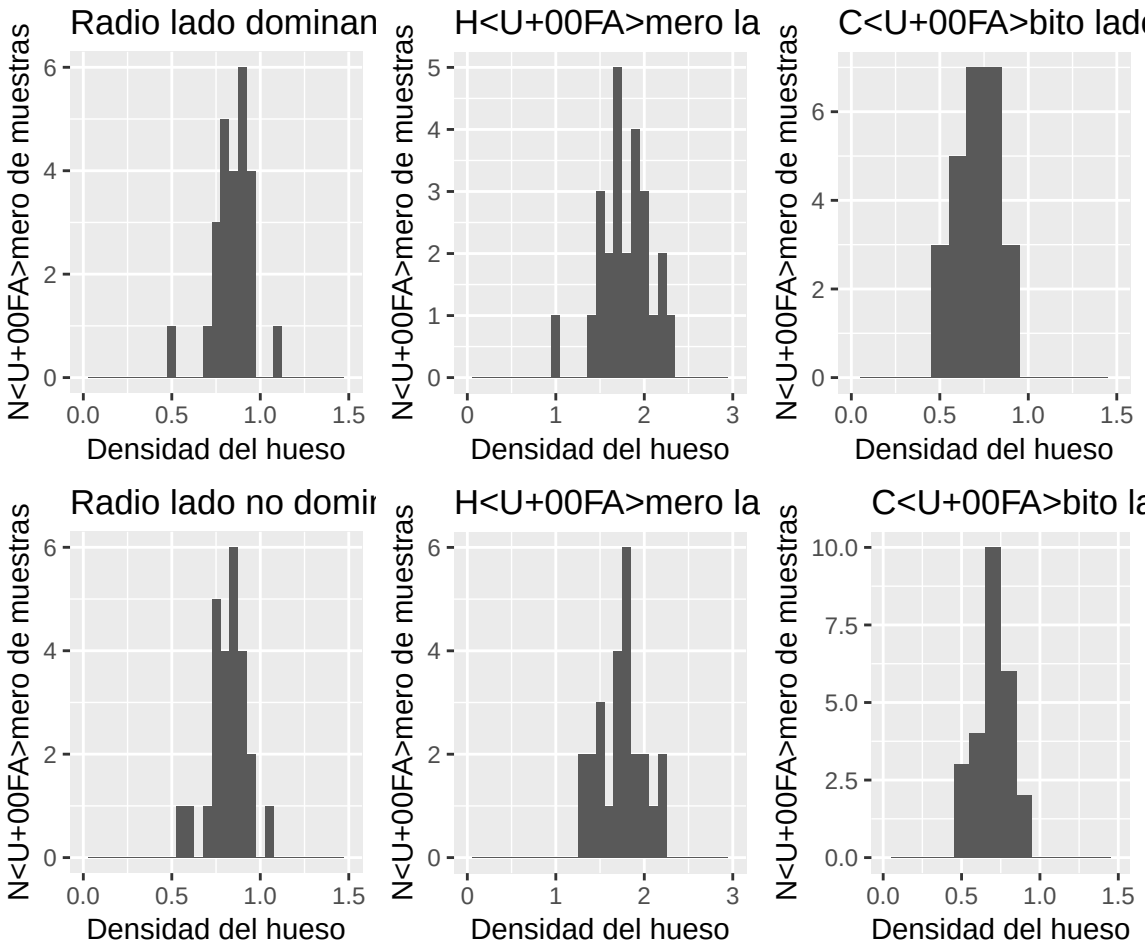
```
## Observations: 25
## Variables: 6
## $ r_d  <dbl> 1.103, 0.842, 0.925, 0.857, 0.795, 0.787, 0.933, 0.799, 0...
## $ r_nd <dbl> 1.052, 0.859, 0.873, 0.744, 0.809, 0.779, 0.880, 0.851, 0...
## $ h_d  <dbl> 2.139, 1.873, 1.887, 1.739, 1.734, 1.509, 1.695, 1.740, 1...
## $ h_nd <dbl> 2.238, 1.741, 1.809, 1.547, 1.715, 1.474, 1.656, 1.777, 1...
## $ c_d  <dbl> 0.873, 0.590, 0.767, 0.706, 0.549, 0.782, 0.737, 0.618, 0...
## $ c_nd <dbl> 0.872, 0.744, 0.713, 0.674, 0.654, 0.571, 0.803, 0.682, 0...
```

```
##       r_d              r_nd             h_d             h_nd      
##  Min.   :0.4930   Min.   :0.5320   Min.   :1.037   Min.   :1.268  
##  1st Qu.:0.7920   1st Qu.:0.7520   1st Qu.:1.650   1st Qu.:1.547  
##  Median :0.8560   Median :0.8380   Median :1.786   Median :1.759  
##  Mean   :0.8438   Mean   :0.8183   Mean   :1.793   Mean   :1.735  
##  3rd Qu.:0.9210   3rd Qu.:0.8760   3rd Qu.:1.954   3rd Qu.:1.869  
##  Max.   :1.1030   Max.   :1.0520   Max.   :2.334   Max.   :2.238  
##       c_d              c_nd       
##  Min.   :0.5330   Min.   :0.4820  
##  1st Qu.:0.6180   1st Qu.:0.6150  
##  Median :0.7060   Median :0.6820  
##  Mean   :0.7044   Mean   :0.6938  
##  3rd Qu.:0.7820   3rd Qu.:0.7650  
##  Max.   :0.8730   Max.   :0.8720
```

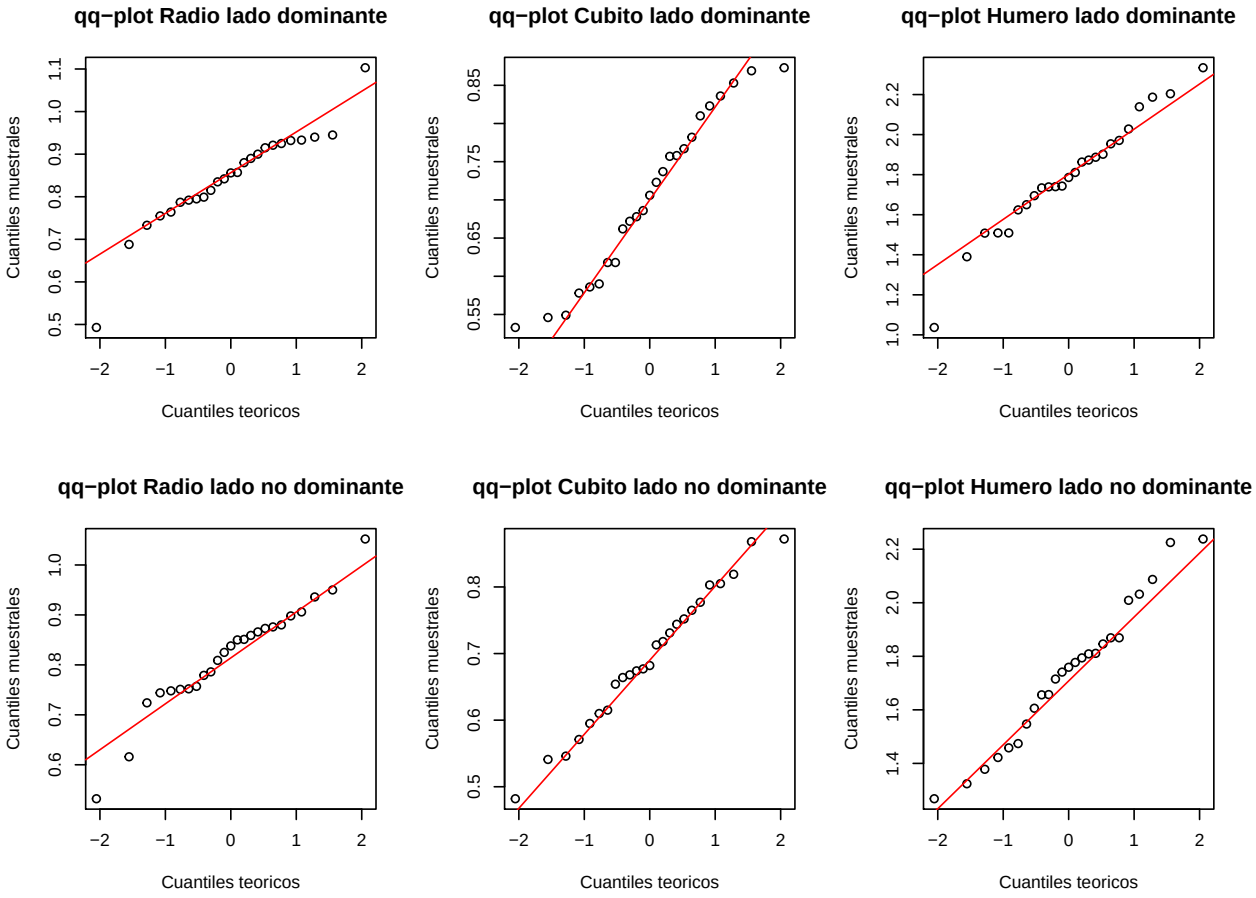


Densidades estimadas de los contenidos minerales en los huesos



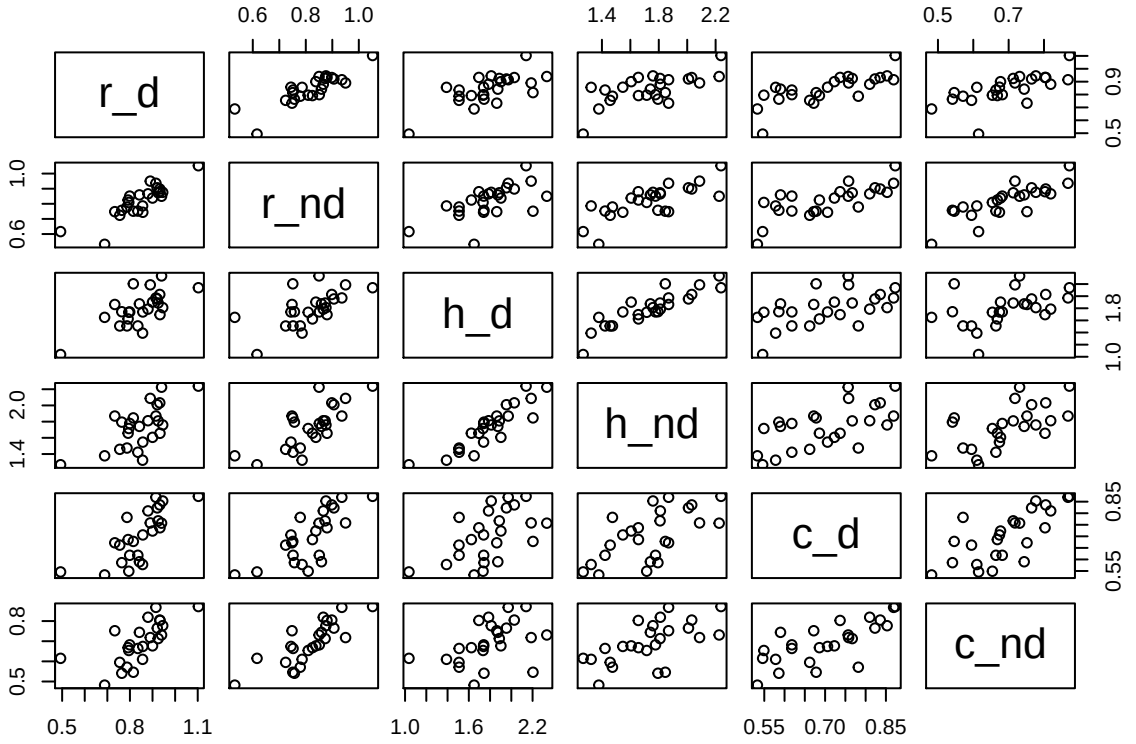


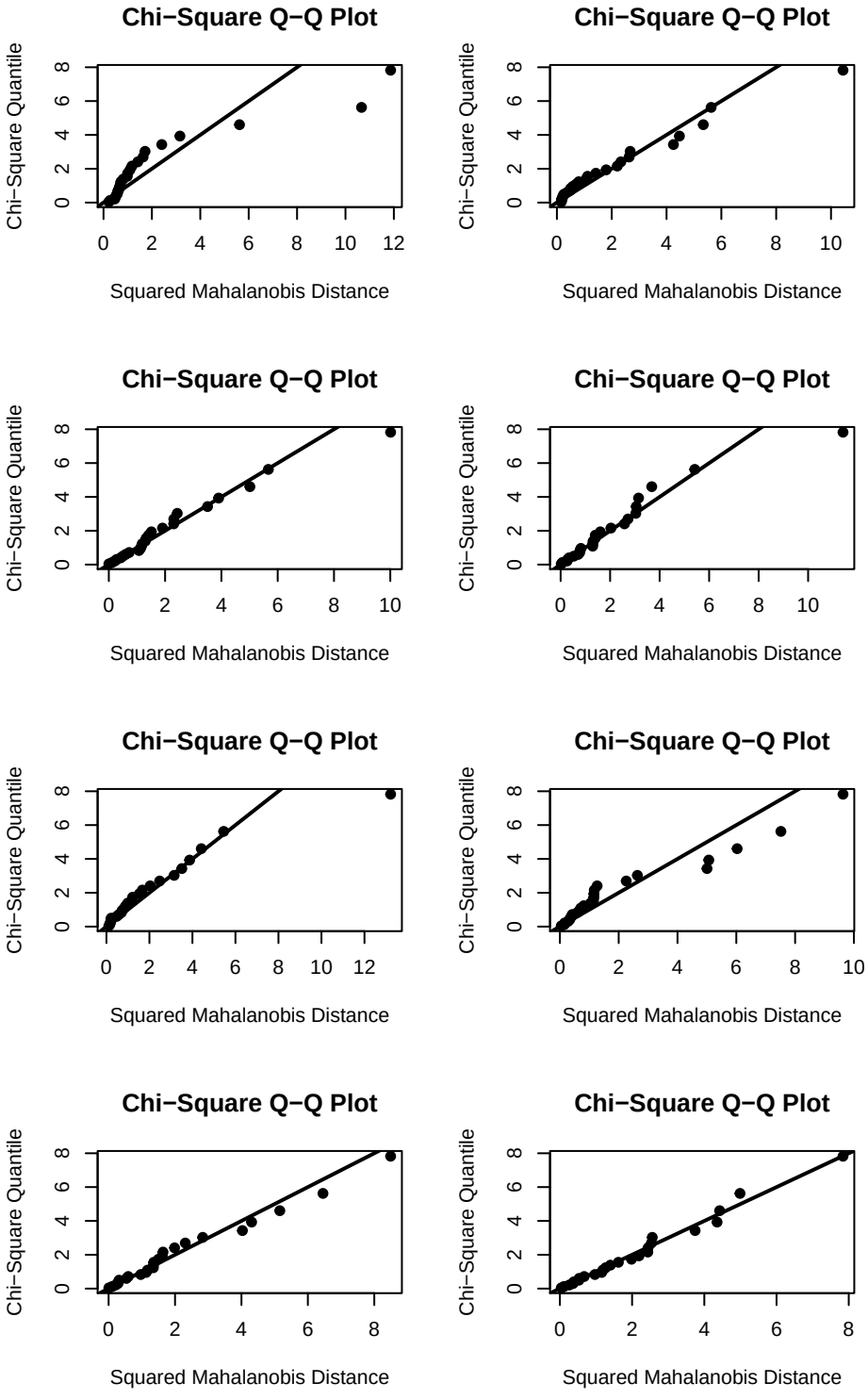
Ahora realizaremos un análisis de normalidad univariable (marginales), con herramientas graficas:

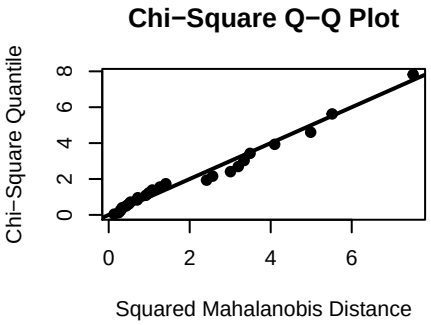


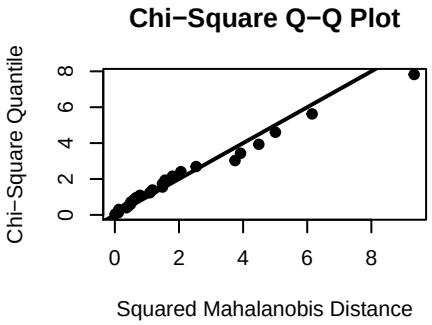
De los diagnósticos (gráficos qqplot) y las pruebas formales (prueba de Shapiro-Wilk podemos concluir con un 95% de confianza que todas las seìs variables distribuyen normal univariado. El tamaño de muestra no es muy grande podemos considerarlo un tamaño medio por lo que el resultado del valor p de la prueba de shapiro-wilk para la población Radio dominante lo aceptamos así sea muy cerca al 0.05 ya que al ser un tamaño de muestra pequeño o medio los resultados son muy ajustados, posiblemente las observaciones extremas distorsionan la prueba.)

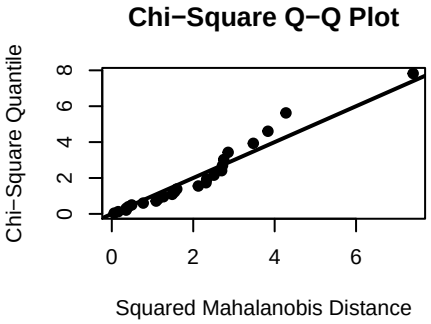
Ahora realizaremos un analisis de la normalidad bivariable:

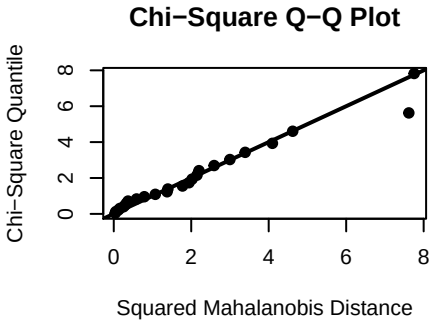

r_d
r_nd
h_d
h_nd
c_d
c_nd
0.6 0.8 1.0
1.4 1.8 2.2
0.5 0.7
0.5 0.9
0.6 1.0
1.0 1.8
1.4 2.0
0.55 0.85
0.5 0.8
0.5 0.8 1.1
1.0 1.6 2.2
0.55 0.70 0.85

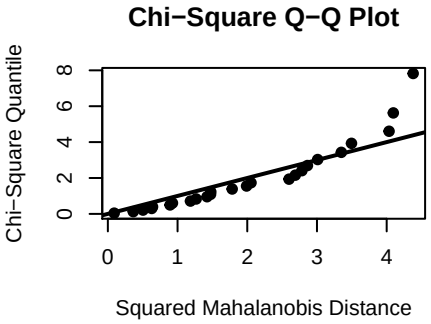

Chi−Square Q−Q Plot
Chi−Square Quantile
Squared Mahalanobis Distance
Chi−Square Q−Q Plot
Chi−Square Quantile
Squared Mahalanobis Distance
Chi−Square Q−Q Plot
Chi−Square Quantile
Squared Mahalanobis Distance
Chi−Square Q−Q Plot
Chi−Square Quantile
Squared Mahalanobis Distance
Chi−Square Q−Q Plot
Chi−Square Quantile
Squared Mahalanobis Distance
Chi−Square Q−Q Plot
Chi−Square Quantile
Squared Mahalanobis Distance
Chi−Square Q−Q Plot
Chi−Square Quantile
Squared Mahalanobis Distance
Chi−Square Q−Q Plot
Chi−Square Quantile
Squared Mahalanobis Distance

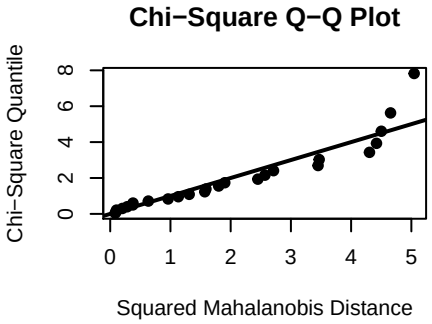

Chi−Square Q−Q Plot
Chi−Square Quantile
0
2
4
6
8
0
2
4
6
Squared Mahalanobis Distance

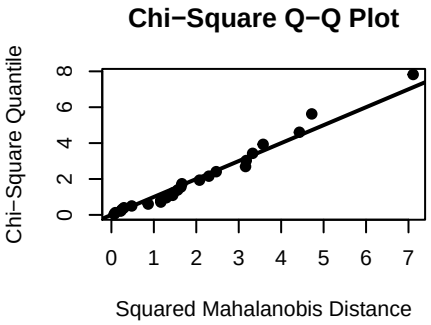

Chi−Square Q−Q Plot
Chi−Square Quantile
0
2
4
6
8
0
2
4
6
8
Squared Mahalanobis Distance

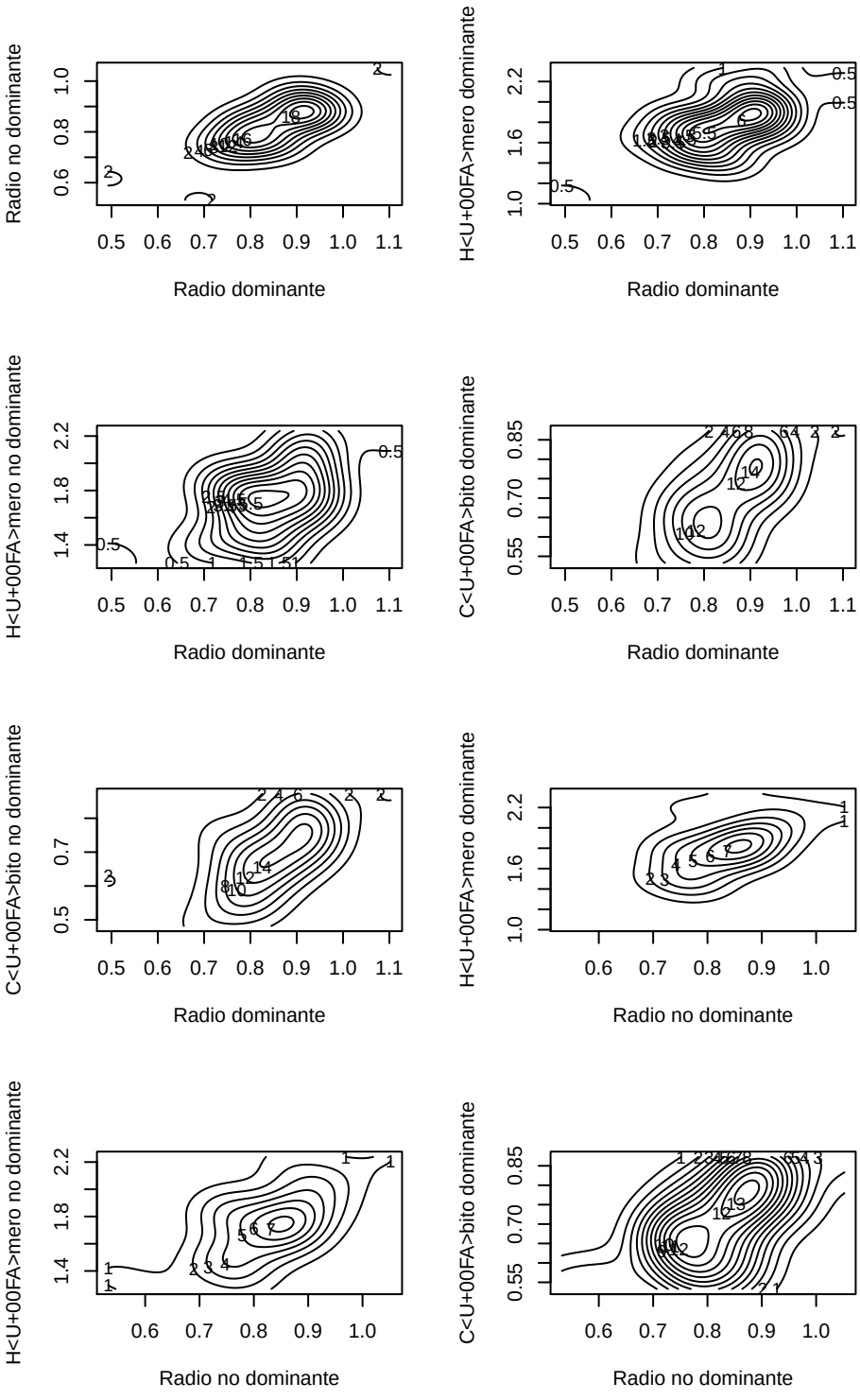

Chi−Square Q−Q Plot
Chi−Square Quantile
0
2
4
6
8
0
2
4
6
Squared Mahalanobis Distance

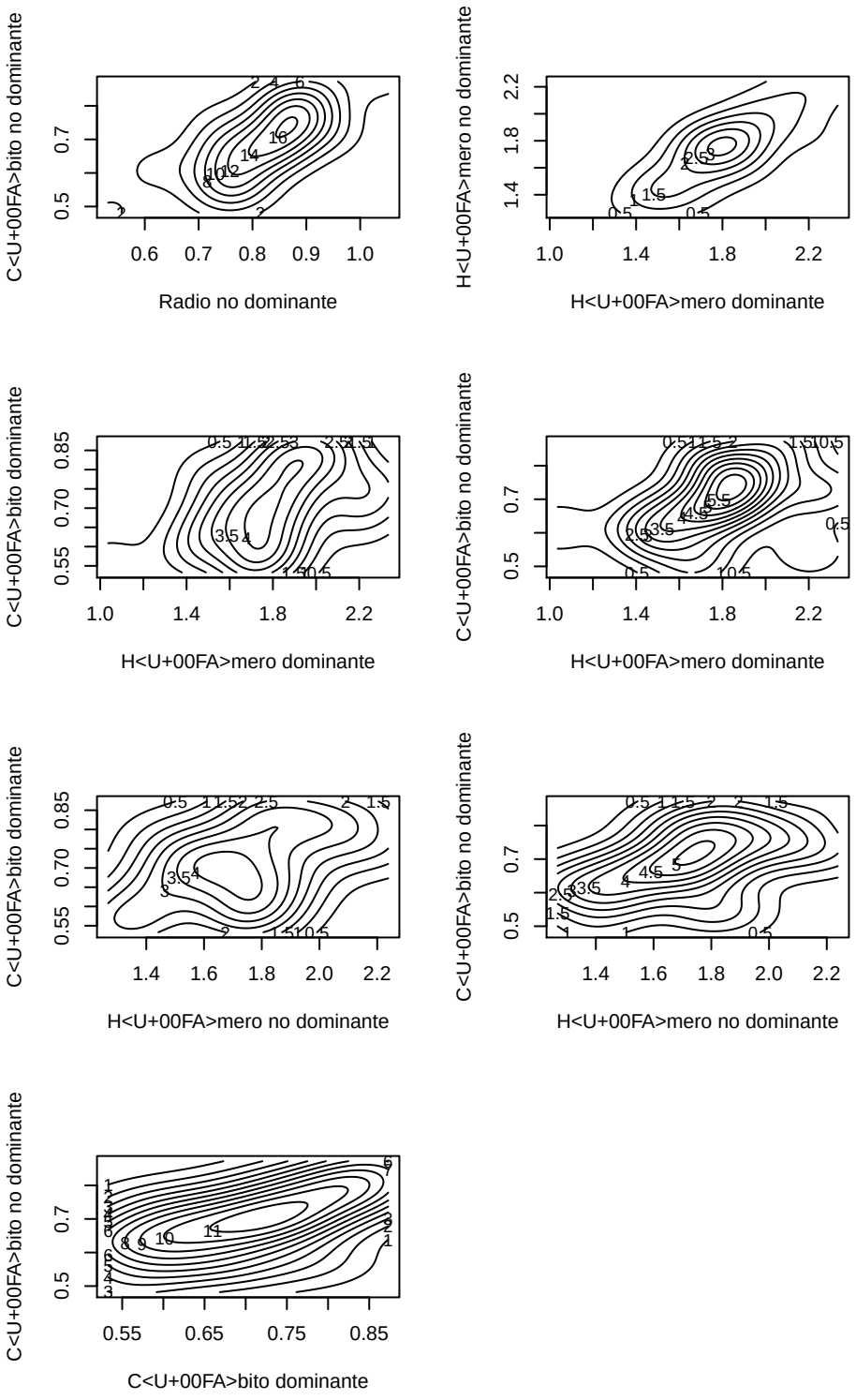

Chi−Square Q−Q Plot
Chi−Square Quantile
0
2
4
6
8
0
2
4
6
8
Squared Mahalanobis Distance


Chi−Square Q−Q Plot
Chi−Square Quantile
0
2
4
6
8
0
1
2
3
4
Squared Mahalanobis Distance


Chi−Square Q−Q Plot
Chi−Square Quantile
0
2
4
6
8
0
1
2
3
4
5
Squared Mahalanobis Distance

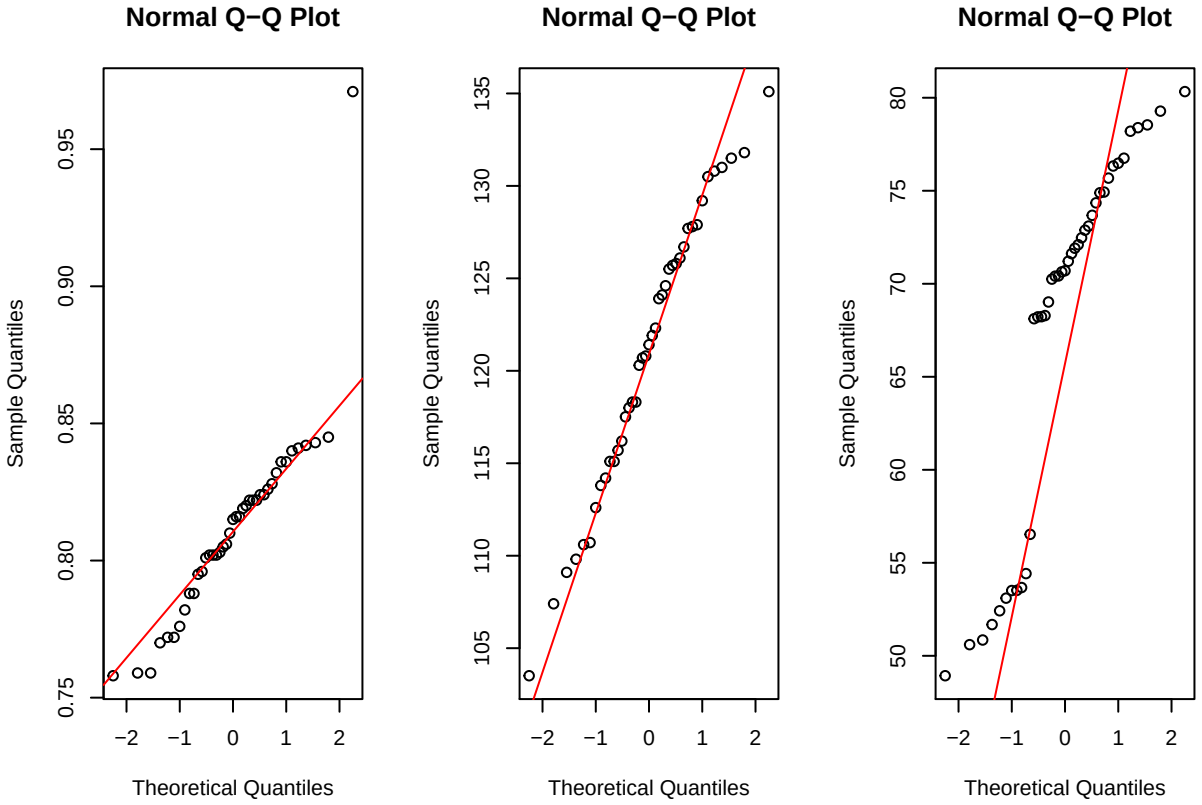

Chi−Square Q−Q Plot
Chi−Square Quantile
0
2
4
6
8
0
1
2
3
4
5
6
7
Squared Mahalanobis Distance

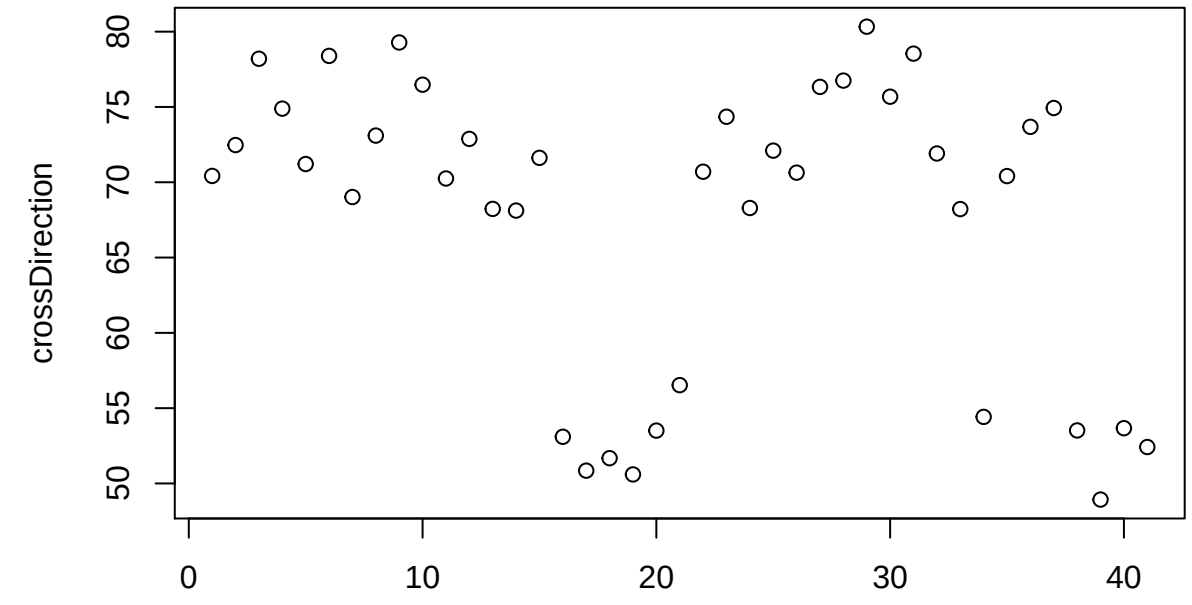

Radio no dominante
Radio dominante
H<U+00FA>mero dominante
H<U+00FA>mero no dominante
C<U+00FA>bito dominante
C<U+00FA>bito no dominante

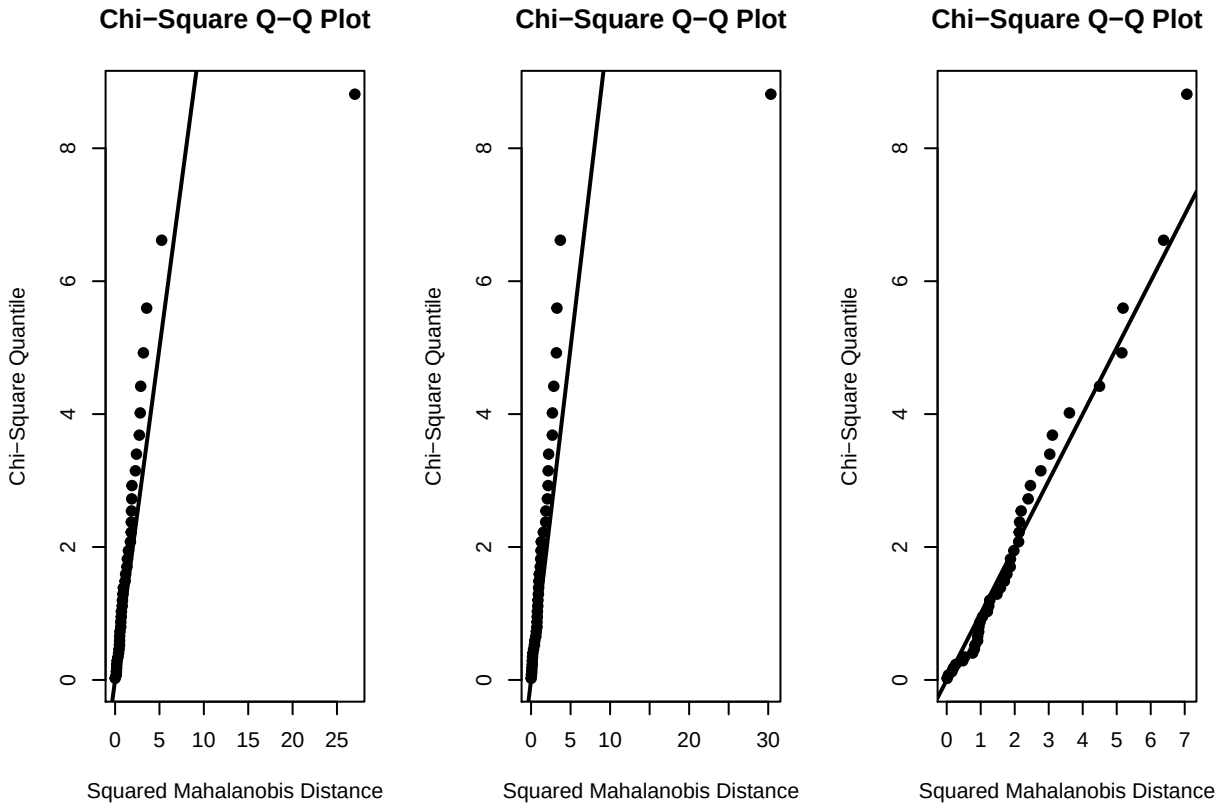


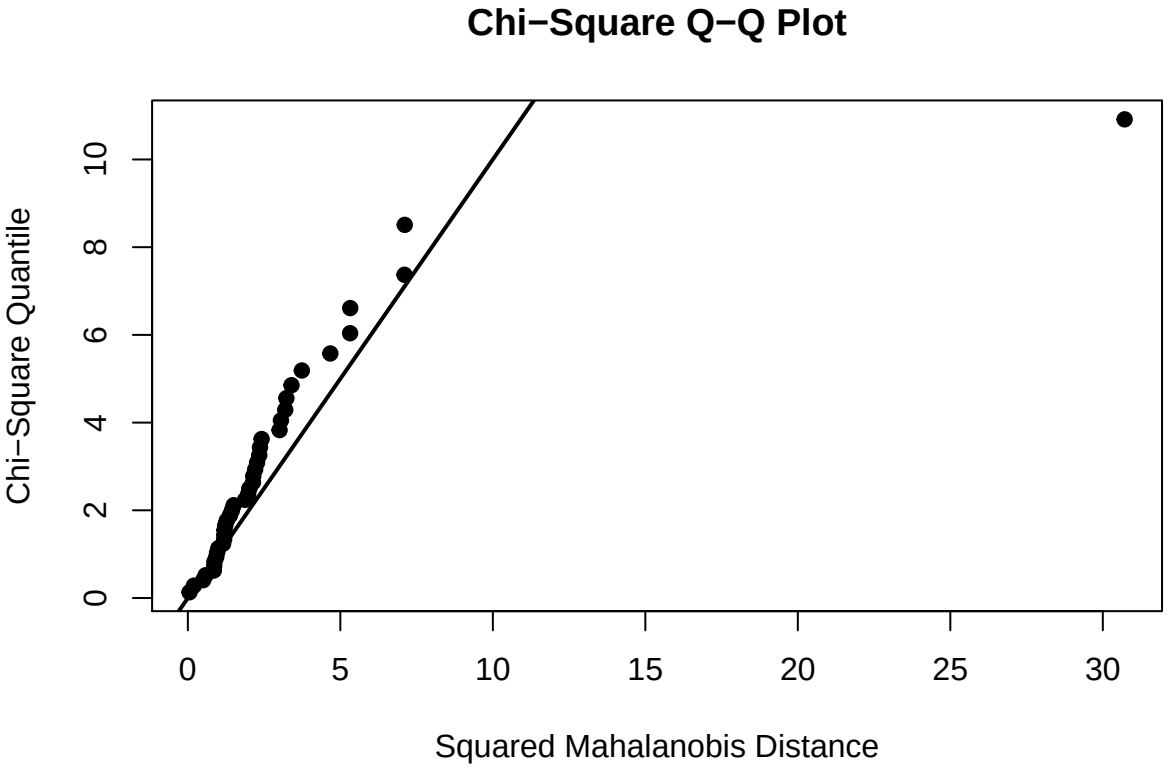
Para evaluar la normalidad bivariada primero graficamos los pares de variables como diagnóstico gráfico y observar una distribución elipsoide de los datos pero la gráfica no es muy clara en muchos casos por lo que procedimos entonces a aplicar la prueba de normalidad de Mardia donde se prueba la kurtosis y la asimetría y se crean pruebas de hipótesis donde la hipótesis nula es que los datos distribuyen normales bivariados, se rechaza dicha hipótesis si los valores p calculados son menores que un alpha del 0.05. De los datos calculados en normalidad.bivariada vemos que las distribuciones Radio_dominante-Radio_no_dominante y Radio_dominante-Cúbito_no_dominante tienen valores p tanto para kurtosis y asimetría son menores del

0.05 y junto con el gráfico qq-plot chicuadrado establecemos que no son normales bivariados. Adicionalmente calculamos el valor p para muestras pequeñas ya que apenas contamos con 25 observaciones que puede ser un valor no tan grande. Con el valor p para muestras pequeñas ratificamos lo analizado para las distribuciones adio_dominante-Radio_no_dominante y Radio_dominante-Cúbito_no_dominante pero aparece la distribución Radio_no_dominante-Húmero_dominante pero al analizar los valore p para kurtosis y asimetría vemos que son bastante altos por lo que no rechazamos la hipótesis de normalidad bivariada. ### Punto 4.35 Examine los datos de las medidas de la calidad del papel de la tabla 1.2 para marginales y normalidad multivariada.



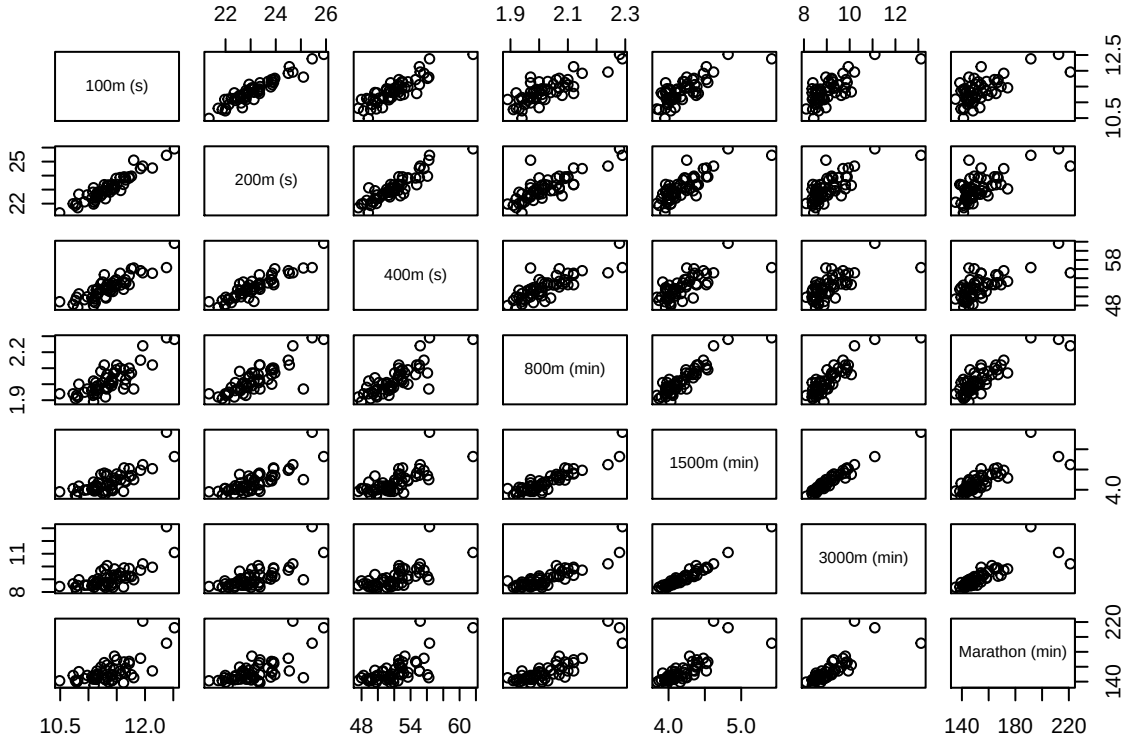

crossDirection
50
55
60
65
70
75
80
0
10
20
30
40
Index


Chi−Square Q−Q Plot
Chi−Square Quantile
Squared Mahalanobis Distance
Chi−Square Q−Q Plot
Chi−Square Quantile
Squared Mahalanobis Distance
Chi−Square Q−Q Plot
Chi−Square Quantile
Squared Mahalanobis Distance



Punto 4.36

Examine los datos sobre records de trayectorias nacional de mujeres de la tabla 1.9 para marginales y normalidad multivariada.



REFERENCIAS

Oppong, Felix Boakye, and Senyo Yao Agbedra. 2016. "Assessing Univariate and Multivariate Normality, a Guide for Non-Statisticians." *Mathematical Theory and Modeling* 6 (2).

Peña, Daniel, and Juan I. Peña. 1986. "Un Contraste de Normalidad Basado En La Transformación Box-Cox." *Estadística Española*, no. 110: 33–46.

Sakia, R. M. 1992. "The Box-Cox Transformation Technique: A Review." *The Statistician*. doi:10.2307/2348250.