

análisis de outlier

Table of contents

Análisis de outlier	2
Análisis de outlier	2
Datos anatómicos de gatos domésticos.	2
<i>Descripción:</i> Los pesos del corazón y del cuerpo de muestras de gatos machos y hembras utilizados para experimentos <i>con digital</i> . Todos los gatos eran adultos y pesaban más de 2 kg. <i>Uso:</i> cats. <i>Formato:</i> Este marco de datos contiene las siguientes columnas:	2
Grafica de caja y bigote.	3
Grafico de dispersion.	4
Elipsoide de tolerancia basado en la distancia de Mahalanobis.	5

Analisis de outlier

Analisis de outlier

Datos anatómicos de gatos domésticos.

Descripción: Los pesos del corazón y del cuerpo de muestras de gatos machos y hembras utilizados para experimentos *con digital*. Todos los gatos eran adultos y pesaban más de 2 kg. **Uso:** cats. **Formato:** Este marco de datos contiene las siguientes columnas:

Sex: sexo: Factor con niveles “F” y “M”.

Bwt: peso corporal en kg.

Hwt: peso del corazón en g.

```
library(MASS)
data("cats")
head(cats)
```

	Sex	Bwt	Hwt
1	F	2.0	7.0
2	F	2.0	7.4
3	F	2.0	9.5
4	F	2.1	7.2
5	F	2.1	7.3
6	F	2.1	7.6

```
datos = data.frame(peso_corporal=log(cats$Bwt),
                    peso_corazon=log(cats$Hwt))
head(datos,10)
```

	peso_corporal	peso_corazon
1	0.6931472	1.945910
2	0.6931472	2.001480
3	0.6931472	2.251292
4	0.7419373	1.974081
5	0.7419373	1.987874

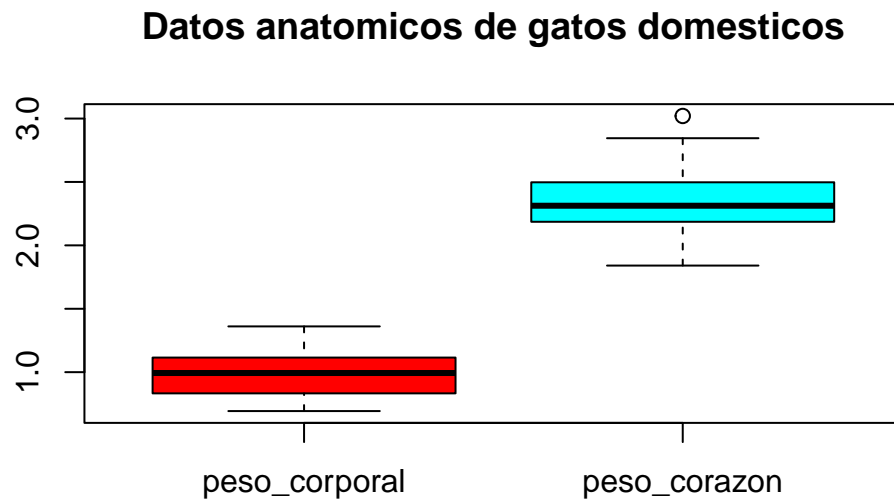
6	0.7419373	2.028148
7	0.7419373	2.091864
8	0.7419373	2.104134
9	0.7419373	2.116256
10	0.7419373	2.140066

```
summary(datos)
```

peso_corporal	peso_corazon
Min. :0.6931	Min. :1.841
1st Qu.:0.8329	1st Qu.:2.192
Median :0.9933	Median :2.313
Mean :0.9866	Mean :2.339
3rd Qu.:1.1068	3rd Qu.:2.495
Max. :1.3610	Max. :3.020

Grafica de caja y bigote.

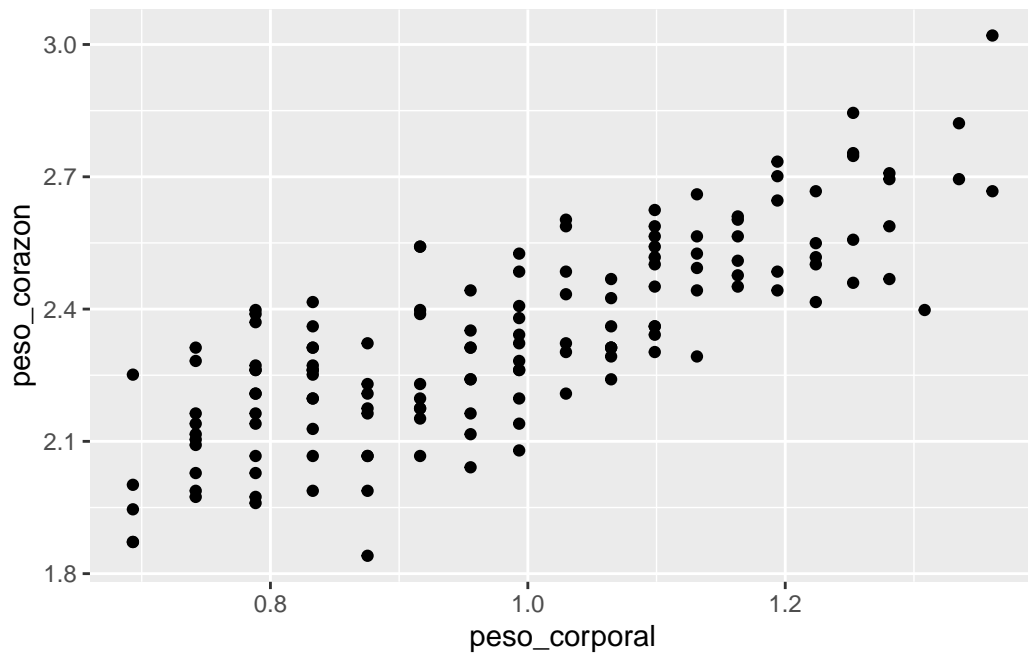
```
boxplot(datos ,main= "Datos anatomicos de gatos domesticos", col = rainbow(ncol(datos)))
```



En el diagrama de cajas y bigotes se observa que solamente hay un valor atípico univariado en el peso del corazón, en todo lo demás no hay valores atípicos, pero hace falta ver si hay valores atípicos en el entorno multivariante.

Grafico de dispersion.

```
library(ggplot2)
attach(datos)
ggplot(datos, aes(x=peso_corporal, y=peso_corazon))+ geom_point()
```



Al crear el grafico de dispersión se puede observar que hay dos valores atípicos multivariados. Pero no están tan separados de los demás datos por lo que debemos detectar los valores atípicos estimando correctamente la estructura de covarianza.

Elipsoide de tolerancia basado en la distancia de Mahalanobis.

```
cats.clcenter=colMeans(datos)
cats.clcov=cov(datos)

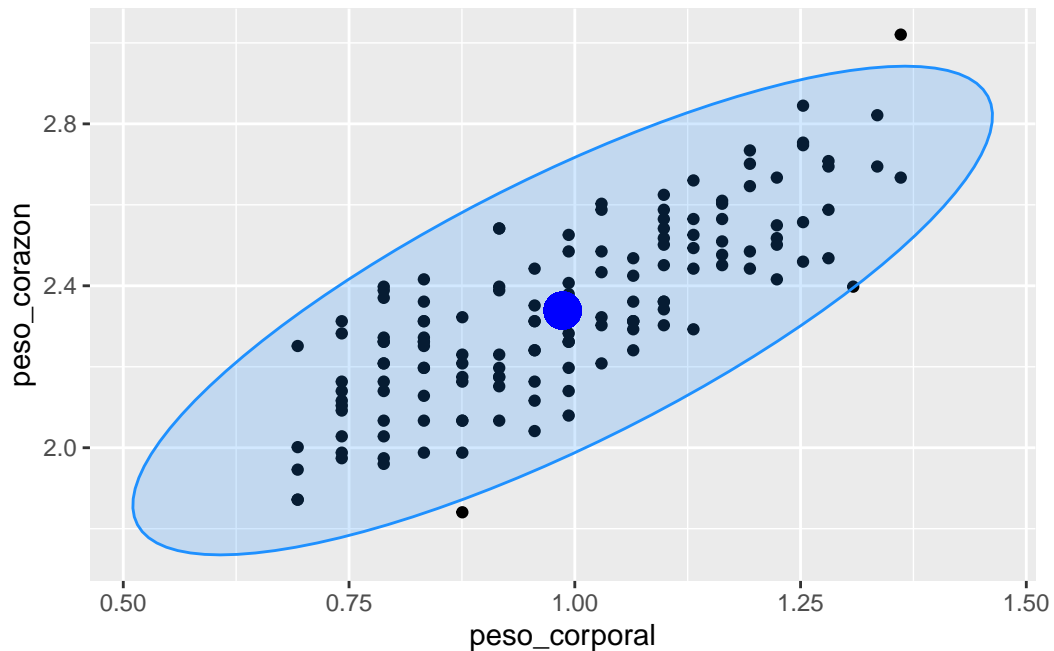
radio=sqrt(qchisq(0.975, df= ncol(datos)))

library(car)
```

Loading required package: carData

```
ellipse.cl=data.frame(ellipse(center = cats.clcenter,
                              shape =cats.clcov,radius = radio,
                              segments = 100,draw = FALSE))
colnames(ellipse.cl)=colnames(datos)
ggplot(data=datos,mapping=aes(x=peso_corporal, y=peso_corazon))+geom_point()+
  geom_polygon(data=ellipse.cl,color= "dodgerblue",fill= "dodgerblue",
              alpha=0.2)+ geom_point(aes(x=cats.clcenter[1],
                                          y=cats.clcenter[2]),
                                     color="blue", size=6)
```

Warning in geom_point(aes(x = cats.clcenter[1], y = cats.clcenter[2]), color = "blue", : All i Please consider using `annotate()` or provide this layer with data containing a single row.



Se observa que hay dos datos que son atipicos en la grafica, esto se debe a que hay dos especies de gatos que tienen medidas diferentes a los demas.

```
library(robustbase)
cats.mcd=covMcd(datos)
cats.mcd
```

```
Minimum Covariance Determinant (MCD) estimator approximation.
Method: Fast MCD(alpha=0.5 ==> h=73); nsamp = 500; (n,k)mini = (300,5)
Call:
covMcd(x = datos)
Log(Det.): -9.444
```

```
Robust Estimate of Location:
peso_corporal  peso_corazon
      0.9828      2.3340

Robust Estimate of Covariance:
           peso_corporal  peso_corazon
peso_corporal      0.03292      0.03151
peso_corazon      0.03151      0.04812
```

```
cats.mcd$center #Estimación robusta de ubicación
```

```
peso_corporal  peso_corazon
      0.9828105      2.3340114
```

```
cats.mcd$cov
```

```
           peso_corporal peso_corazon
peso_corporal    0.03292397  0.03151348
peso_corazon     0.03151348  0.04811881
```

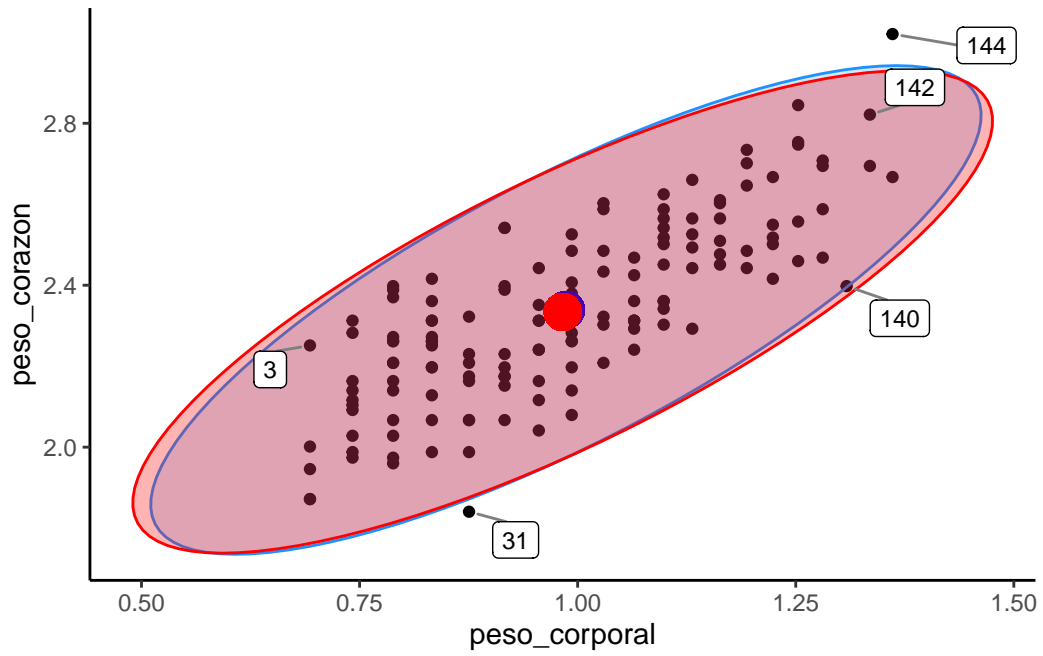
```
#Construimos el elipsoide de tolerancia robusto.
ellipse.mcd=data.frame(ellipse(center=cats.mcd$center,
                                shape=cats.mcd$cov,
                                radius=radio,
                                segments=100,draw=FALSE))
colnames(ellipse.mcd)=colnames(datos)
fig2 = ggplot(data=datos,mapping=aes(x=peso_corporal, y=peso_corazon),
              label=row.names(datos))+geom_point()+
  geom_polygon(data=ellipse.cl, color= "dodgerblue",fill= "dodgerblue",
              alpha=0.2)+geom_point(aes(x=cats.clcenter[1],
                                         y= cats.clcenter[2]),
                                   color="blue", size=6)+
  geom_polygon(data=ellipse.mcd,color= "red", fill= "red",alpha=0.3)+
  geom_point(aes(x=cats.mcd$center[1], y=
                 cats.mcd$center[2]),color="red", size=6)
```

```
library(ggrepel)
fig2 + geom_label_repel(aes(label= row.names(datos)),size= 3,
                        box.padding=0.5,point.padding = 0.3,
                        segment.color = 'grey50')+theme_classic()
```

Warning in geom_point(aes(x = cats.clcenter[1], y = cats.clcenter[2]), color = "blue", : All points have the same color. Please consider using `annotate()` or provide this layer with data containing a single row.

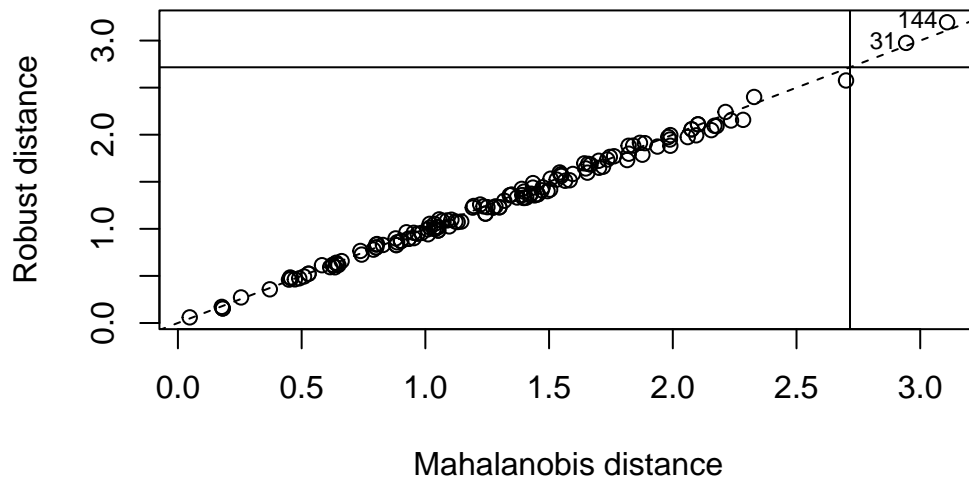
Warning in geom_point(aes(x = cats.mcd\$center[1], y = cats.mcd\$center[2]), : All aesthetics have the same value. Please consider using `annotate()` or provide this layer with data containing a single row.

Warning: ggrepel: 139 unlabeled data points (too many overlaps). Consider increasing max.overlaps



```
plot(cats.mcd, which="dd")
```


Distance–Distance Plot



```
row.names.data.frame(cats)
```

```
[1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12"
[13] "13" "14" "15" "16" "17" "18" "19" "20" "21" "22" "23" "24"
[25] "25" "26" "27" "28" "29" "30" "31" "32" "33" "34" "35" "36"
[37] "37" "38" "39" "40" "41" "42" "43" "44" "45" "46" "47" "48"
[49] "49" "50" "51" "52" "53" "54" "55" "56" "57" "58" "59" "60"
[61] "61" "62" "63" "64" "65" "66" "67" "68" "69" "70" "71" "72"
[73] "73" "74" "75" "76" "77" "78" "79" "80" "81" "82" "83" "84"
[85] "85" "86" "87" "88" "89" "90" "91" "92" "93" "94" "95" "96"
[97] "97" "98" "99" "100" "101" "102" "103" "104" "105" "106" "107" "108"
[109] "109" "110" "111" "112" "113" "114" "115" "116" "117" "118" "119" "120"
[121] "121" "122" "123" "124" "125" "126" "127" "128" "129" "130" "131" "132"
[133] "133" "134" "135" "136" "137" "138" "139" "140" "141" "142" "143" "144"
```

Se observa que los dos datos atipicos son el 31 y 144.

Se ve que todos los datos van siguiendo una linea pero los datos atipicos estan muy alejados de los demas , si siguen la misma linea , pero mucho mas arriba.