ESTADISTICA ESPAÑOLA
Núm. 95, 1982, págs. 113 a 120

Criterio de detección de outliers en modelos probabilísticos tipo Pareto

por JOAQUIN MUÑOZ GARCIA
Y ANTONIO PASCUAL COSTA
Departamento de Estadística e
Investigación Operativa.
Universidad de Sevilla

RESUMEN

Partiendo de una población tipo Pareto con origen de rentas θ conocido, se encuentra un criterio para detección de outliers. Se estudia a continuación el caso de origen de rentas desconocido, dándose al final un criterio para la detección de varias observaciones outliers. Se incluye el correspondiente programa de ordenador.

Palabras clave: Distribución de Pareto, outliers, detección de outliers.

INTRODUCCION

Un outliers es una observación o un conjunto de observaciones que «parecen» ser inconsistentes con el resto del conjunto de datos.

Aunque son muchas las definiciones que sobre el concepto se han dado (véase Muñoz, 1980), lo que caracteriza a una observación outliers es el «impacto» que produce en el estadístico cuando va a analizar los datos. Parece evidente que la presencia de outliers en un conjunto de datos puede conducirnos a errores en nuestro intento de hacer inferencias acerca de la población de la que proceden, de ahí que la presencia de outliers plantee un problema fundamental en el análisis de datos. Un procedimiento para la resolución de este problema consiste en encontrar reglas de decisión para detectar dichas observaciones.

En el presente trabajo se dan procedimientos para la detección de outliers en poblaciones tipo Pareto, que como se sabe es un modelo probabilístico que se ajusta adecuadamente a distribuciones de frecuencias de rentas observadas en la realidad.

Sea X una variable aleatoria que se distribuye según una ley de Pareto de parámetros θ y a (Pa(θ, a)), por lo que su función de densidad es:

$$f(x) = \frac{a\theta^a}{x^{a+1}}; \quad a > 0, x \geq \theta \text{ y } \theta > 0$$

Se sabe que el estimador de máxima verosimilitud para la constante de Pareto a, conocido el origen de rentas θ viene dado para una muestra aleatoria simple de tamaño n por:

$$\hat{a} = n\left[\sum_{i=1}^{n} \ln\left(\frac{X_i}{\theta}\right)\right]^{-1}$$

Johnson Kotz (1970)

Y además se verifican las siguientes proposiciones, cuyas demostraciones son inmediatas.

Proposición 1

$\hat{a}$ es un estimador invariante por transformaciones de escala:

$$\hat{a} = n\left[\sum_{i=1}^{n} \ln\left(\frac{X_i'}{\theta'}\right)\right]^{-1}$$

siendo

$X_i' = X_i O_x \quad \forall i,\ i = 1, 2, \ldots, n$ y $\theta' = \theta O_x$, siendo O_x la transformación realizada.

Proposición 2

Si la variable aleatoria X se distribuye según Pa(θ, a), entonces la variable:

$$2a\,Y$$

con

$$Y = \ln \frac{X}{\theta}$$

se distribuye según una ley gamma de parámetros 1 y 1/2, G(1, 1/2).

Proposición 3

Dada una muestra aleatoria simple de tamaño $n(X_1, X_2, \ldots, X_n)$ procedente de una población $Pa(\theta, a)$ el estadístico.

$$\sum_{i=1}^{n} 2a\,Y_i = \frac{2an}{\hat{a}}$$

con

$$Y_i = \ln \frac{X_i}{\theta}$$

se distribuye según una ley $\chi^2(2n)$.

2. DISTRIBUCION DEL ESTADISTICO BASICO

Consideramos como estadístico básico para la detección de outliers.

$$T_i = \frac{n-1}{n} \cdot \frac{\hat{a}_n}{\hat{a}_{n-1}^{(i)}}$$

donde $\hat{a}_n$ es el estimador de máxima verosimilitud de a obtenido a partir de una muestra aleatoria simple de tamaño n y $\hat{a}_{n-1}^{i}$ es el mismo estimador de máxima verosimilitud, pero obtenido a partir de $n-1$ de las n observaciones anteriores en las que hemos suprimido la observación i-ésima.

Del estadístico T_i podemos afirmar que sus valores muestrales no van a depender del parámetro a y por la proposición 1, que es invariante mediante transformaciones de escala.

Teorema 1

Bajo la hipótesis de que no existen outliers, el estadístico T_i se distribuye según una ley Beta de parámetros $(n-1, 1)$.

Demostración

Podemos expresar T_1 en la forma:

$$T_1 = \frac{n-1}{n}\frac{\hat{a}_n}{\hat{a}_{n-1}} = \frac{2a\sum_{j=2}^{n} \ln\left(\frac{X_j}{\theta}\right)}{2a\sum_{j=1}^{n} \ln\left(\frac{X_j}{\theta}\right)} = \sum_{j=2}^{n} Z_j$$

con

$$Z_j = \frac{2a \ln\left(\frac{X_j}{\theta}\right)}{2a \sum_{j=1}^{n} \ln\left(\frac{X_j}{\theta}\right)} \qquad j = 1, 2, \ldots, n$$

donde se ha supuesto que la observación suprimida es la primera.

De la expresión anterior se deduce que:

$$Z_1 = \frac{2a \ln\left(\frac{X_1}{\theta}\right)}{2a \ln\left(\frac{X_1}{\theta}\right) + 2a \sum_{j=2}^{n} \ln\left(\frac{X_j}{\theta}\right)}$$

En virtud de la proposición 2, el numerador se distribuye según una ley $\chi^2(2)$, por la proposición 3, $2a \sum_{j=2}^{n} \ln (X_i/\theta)$ se distribuye según la ley $\chi^2(2(n-1))$ y es independiente de la variable $2a \ln (X_1/\theta)$.

Por tanto, Z_1 se distribuye según una ley Beta, Be$(1, n-1)$. Y al ser $T_1 = 1 - Z_1$, se deduce trivialmente que T_1 sigue una ley Be$(n-1, 1)$, es decir, el estadístico T posee una distribución libre.

3. CRITERIO PARA LA DETECCION DE UN OUTLIERS

Para la detección de un outliers vamos a considerar los estadísticos:

$$T_i = \frac{n-1}{n} \cdot \frac{\hat{a}_n}{\hat{a}_{n-1}^{(i)}}$$

para $i = 1, 2, \ldots, n$.

Si todos los elementos de la muestra de tamaño n pertenecen a la misma población, los estadísticos $T_i = 1, 2, \ldots, n$ serán próximos a la unidad.

Por ello, el estadístico que proponemos para detectar un outliers será:

$$\min_i T_i$$

Para determinar la región crítica hacemos la siguiente acotación:

$$P[\min_i T_i \leq t] = P\left[\bigcup_{i=1}^{n} A_i\right] \leq \sum_{i=1}^{n} P(A_i) = n\,P(A_i)$$

siendo

$$A_i = [T_i \leq t]$$

con $i = 1, 2, \ldots, n$.

Y para un nivel de significación α determinamos t de forma que:

$$P[T_i \leq t] = \frac{\alpha}{n}$$

es decir, t es el percentil de orden α/n de una ley $Be(n - 1, 1)$.

En definitiva, si $T_k = \min_i T_i \leq t$, para $i = 1, 2, \ldots, n$, podemos afirmar con un coeficiente de confianza superior o igual a $(1 - \alpha)$ que la observación k-ésima es outliers.

Hasta ahora se ha supuesto que el origen de rentas θ es conocido, en el caso en que dicho origen no sea conocido, utilizaremos el mismo método, sólo que sustituiremos θ por su estimador de máxima verosimilitud.

$$\hat{\theta} = \min_{1 \leq i \leq n} X_i$$

aunque para mantener la eficiencia del método habrá que disponer de una muestra suficientemente grande.

A continuación damos un criterio a emplear para detectar más de un outliers y así poder evitar el efecto de enmascaramiento citado por Tietjen y Moore (1972).

4. CRITERIO PARA DETECTAR MAS DE UN OUTLIERS

Para detectar s outliers $s > 1$, calcularemos $\binom{n}{s}$ estadísticos de la forma:

$$T_{i_1 i_2, \ldots, i_s} = \frac{n - s}{n} \frac{\hat{a}_n}{\hat{a}_{n-s}^{(i_1 i_2, \ldots, i_s)}}$$

siendo $(i_1, i_2, \ldots, i_s)$ una de las $\binom{n}{s}$ permutaciones posibles $(1, 2, \ldots, n)$ y donde cada uno de los estadísticos se distribuye según una ley beta $Be(n - s, s)$.

El estadístico a utilizar para la detección de s outliers será:

$$\min_{(i_1, i_2, \ldots, i_s)} T_{i_1, i_2, \ldots, i_s}$$

y diremos que las observaciones $(k, l, \ldots, v)$ son s outliers para un nivel de significación α si:

$$T_{k,l,\ldots,v} = \min_{(i_1, \ldots, i_s)} T_{i_1, i_2, \ldots, i_s} \leq t_s$$

donde t_s va a ser el percentil del orden $\binom{n}{s}$ correspondiente a una ley beta Be$(n - s, s)$.

A continuación damos el programa Fortran que emplearíamos para detectar uno y dos outliers en una muestra de tamaño 19.

FORTRAN IV

```
0001        OPEN(UNIT=2,NAME='SALAR.DAT',TYPE='UNKNOWN')
0002        DIMENSION XX(19),X(19),T1(19),V1(19),V2(19)
0003        PRINT 40
0004        DO 1 I=1,19
0005        READ(2,100) XX(I)
0006        X(I)=XX(I)*200
0007        PRINT 110, X(I)
0008   1    CONTINUE
0009        THETA=XMIN1(X,19)
0010        SUMD=SUMLOG(X,19,THETA)
0011        DO 2 I=1,19
0012        CALL REST1(X,19,I,V1)
0013        SUMN=SUMLOG(V1,18,THETA)
0014        T1(I)=SUMN/SUMD
0015   2    CONTINUE
0016        T1MIN=XMIN1(T1,19)
0017        T2MIN=1
0018        DO 3 I=1,18
0019        DO 3 J=I+1,19
0020        CALL REST2(X,19,I,J,V2)
0021        SUMN=SUMLOG(V2,17,THETA)
0022        TT=SUMN/SUMD
0023        T2MIN=AMIN1(T2MIN,TT)
0024   3    CONTINUE
0025        PRINT 120, T1MIN
0026        PRINT 130,T2MIN
0027        STOP
        C   ------------------------------------------------------------
        C                          FORMATOS
        C   ------------------------------------------------------------
0028   40   FORMAT(20X,'OBSERVACIONES:'/)
0029  100   FORMAT(2X,F5.0)
0030  110   FORMAT(20X,F9.0)
0031  120   FORMAT(20X,'VALOR MINIMO DE T(I):',2X,F14.4)
```

```
0032  130  FORMAT(20X,'VALOR MINIMO DE T(I1,I2):'2X,F10.4)
0033       END
```

FORTRAN IV. V02.1-10

```
0001       FUNCTION SUMLOG(X1,N1,TH)
0002       DIMENSION X1(N1)
0003       SUMLOG=0
0004       DO 10 II=1,N1
0005       SUMLOG=SUMLOG+ALOG(X1(II)/TH)
0006  10   CONTINUE
0007       RETURN
0008       END
```

FORTRAN IV

```
0001       SUBROUTINE REST1(X1,N1,L,A)
0002       DIMENSION X1(N1),A(N1)
0003       TYPE 30, L
0004  30   FORMAT(1X,I3)
0005       IF(L.EQ.1) GO TO12
0007       DO 10 L1=1, L-1
0008       A(L1)=X1(L1)
0009       TYPE 20, A(L1)
0010  20   FORMAT(3X,F14.0)
0011  10   CONTINUE
0012       IF(L.EQ.19) RETURN
0014  12   DO 11 L1=L+1,N1
0015       A(L1-1)=X1(L1)
0016       TYPE 20, A(L1-1)
0017  11   CONTINUE
0018       RETURN
0019       END
```

FORTRAN IV

```
0001       SUBROUTINE REST2(X1,N1,LI,LJ,A)
0002       DIMENSION X1(N1),A(N1)
0003       TYPE 30, LI,LJ
0004  30   FORMAT(1X,2I3)
0005       IF(LI.EQ.1) GO TO 13
0007       DO 10 L=1,LI-1
0008       A(L)=X1(L)
0009       TYPE 20, A(L)
0010  20   FORMAT(3X,F12.0)
0011  10   CONTINUE
0012       IF(LI.EQ.18) RETURN
0014  13   IF((LI+1).EQ. LJ) GO TO 14
0016       DO 11 L=LI+1,LJ-1
0017       A(L-1)=X1(L)
0018       TYPE 20, A(L-1)
0019  11   CONTINUE
0020       IF(LJ.EQ.19) RETURN
0022  14   DO 12 L=LJ+1,N1
0023       A(L-2)=X1(L)
0024       TYPE 20, A(L- 2)
```

```
0025  12  CONTINUE
0026      RETURN
0027      END
C         CALCULO MINIMO REAL DE UN CONJUNTO DE N
          OBSERVACIONES REALES
C         FUNCTION XMIN1(X,N)
          DIMENSION X(N)
          DO1
          11=1,N
          XMIN1=AMIN1(XMIN1,X(11))
1         CONTINUE
          RETURN
          END
```

BIBLIOGRAFIA

JOHNSON, N. L. and KOTZ, S.: *Distributions in Statistics: Continuous univariate distributions 1*. Wiley, 1970.

MUÑOZ, J.: *Algunas técnicas sobre detección de outliers*. Public. Universidad de Sevilla, 1980.

TIETJEN, G. L. and MOORE, R. H.: «Some Grubbs - Type Statistics for the detection of several outliers». *Technometrics* (14), 1972.

SUMMARY

Starting from a Pareto type population with known minimum income θ in this paper a criterium for detection of outliers is derived. The unknowm minimum income case is also studied, showing a criterium for detection of some outliers observations. The corresponding computing program is included.

Key words: Pareto distribution, outliers, detection of outliers.

AMS, Subject classification 62 F 35.