Universidad de Granada
Escuela Internacional de Posgrado
Máster en Estadística Aplicada
Materia: Encuestas por Muestreo.
Alumno: Francisco Javier Márquez Rosales

Encustas por Muestreo:

Actividad Tema 4.

Noviembre, 2022

Ejercicio 1

Los datos del fichero adjunto (granjas.txt) corresponden a una muestra de granjas. La variable x es el área e y la producción.

Comprobar si dicha población sigue el siguiente modelo de superpoblación: yi=βxi+v(xi)uii siendo β un parámetro desconocido, v(x)=x1/2, y ui variables aleatorias i.i.d. con media cero (representa gráficamente los datos con cualquier paquete estadístico).

A partir de dicha muestra calcula una estimación apropiada para la producción media sabiendo que la superficie media de las 120 granjas es de 31.2 unidades

Solución

Iniciamos con la lectura y el procesamiento de los datos

```
data<-read.table("granjas.txt", header = TRUE)
head(data)
```

```
##    x     y
## 1 21 11703
## 2 22 26545
## 3 22 22509
## 4 23 30168
## 5 24 37274
## 6 24 41962
```

Luego obtenemos como ejercicio las defincionesnecesarias para la muestra

```
# valores de la muestra
n=27
N=120
media=31.2

# vector de probabilidades de inclusi´on para un m.a.s
pik=rep(n/N,n)
```

```
# matriz de probabilidades de inclusión para un m.a.s
pikl=matrix(n*(n-1)/(N*(N-1)),n,n)
```

Ahora, definimos el modelo de regresión de acuerdo a las indicaciones

```
# definici´on del modelo de superpoblación y = beta*xi + v(xi)uii
v=x**0.5
modelo=y~x+v
modelo
```

Y obtenemos el estimador de regresión a partir del modelo

```
tx=sum(x)
tv=sum(v)
total=c(tx,tv)
total
```

```
## [1] 814.0000 147.6059
```

```
#totales=c(tx1,tx2)

# cálculo del estimador de regresión 0
r=regest(formula=modelo, weights= 1/pik, Tx=total, pikl, n)
```

Obteniendo también los coeficientes

```
# estimador de regresión
t_ygreg=r$regest
med_ygreg=t_ygreg/N
med_ygreg
```

```
## [1] 132633.2
```

```
# coeficientes
```

```
coef=r$coefficients
coef
```

```
## x(Intercept)            xx            xv
##   160286.099     7862.303   -65839.265
```

De esta forma, ahora obtenemos la estimación apropiada para la producción media con base en el modelo, es decir, despejando el valor del coeficiente sabiendo que la superficie media de las 120 granjas es de 31.2 unidades.

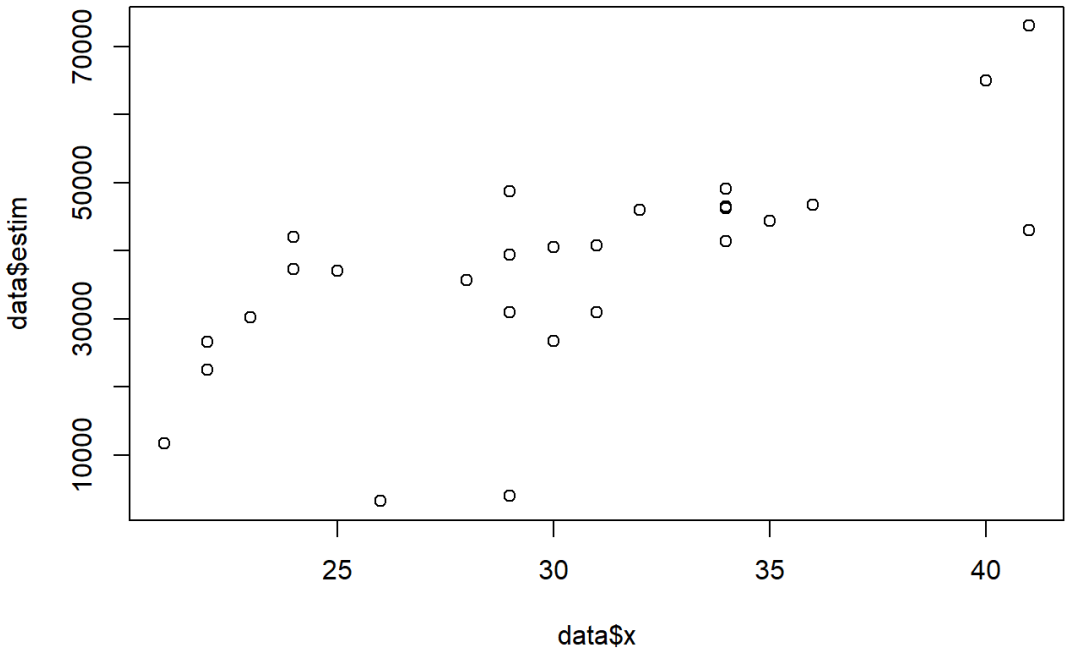
```
yestim<- 160286.099+7862.303*31.2-(65839.265/31.2)
yestim
```

```
## [1] 403479.7
```

Finalmente, representamos gráficamente los valores obtenidos a partir del modelo

```
data$estim<-r$y

plot(data$x,data$estim)
```


data$estim
10000
30000
50000
70000
25
30
35
40
data$x

Ejercicio 2

En el fichero grafico.doc se muestra el gráfico correspondiente a la población denominada Labor donde las variables consideradas son y, ingresos semanales, y x, horas semanales.

1.- En función a dicha gráfica, ¿qué modelo de superpoblación supondrías para la población? ¿Qué se podría comentar acerca de la componente de error del modelo?

2.- Selecciona una muestra de tamaño 20 mediante muestreo aleatorio simple de la población (fichero labor.txt) y estima en base a ella la media de los ingresos semanales por trabajador mediante el estimador de razón y mediante el estimador de regresión.

Note que la columna 9 corresponde a la variable x (horas semanales) y la columna 10 a la variable y (ingresos semanales).

El ejercicio se puede realizar utilizando el paquete sampling, pero se valorará que el alumno desarrolle código propio en R.

1.- En función a dicha gráfica, ¿qué modelo de superpoblación supondrías para la población? ¿Qué se podría comentar acerca de la componente de error del modelo?

Por la forma del gráfico considero que sería apropiado un modelo de regresión múltiple G_{MR} en el caso especial de una variable auxiliar G_R de la forma $y_k=B_{xk}+e_k$. La forma de distribución de los puntos en el plano sugiere que la distribución del error del modelo sigue un comportamiento heterocedastico.

2.- Selecciona una muestra de tamaño 20 mediante muestreo aleatorio simple de la población (fichero labor.txt) y estima en base a ella la media de los ingresos semanales por trabajador mediante el estimador de razón y mediante el estimador de regresión.

Solución.

En primer lugar hacemos una lectura de la data y validamos los nombres de las variables, estructura y contenido.

```
data42<-read.table("Labor.txt", header = TRUE)
head(data42)
```

```
##   X1 X1.1 X1.2 X1.3 X22 X2 X2.1 X1.4 X40 X160 X52.6
## 1  1    1    1    2  53  4    2    1  40  224 48.97
## 2  1    1    1    3  20  2    2    2  40  164 56.15
## 3  1    1    1    4  19  1    2    1  40  134 36.25
## 4  1    1    1    5  24  2    2    1  40  146 60.03
## 5  1    1    2    6  28  3    1    1  40  320 50.42
## 6  1    1    2    7  32  3    1    1  40  300 45.67
```

```
str(data42)
```

```
## 'data.frame':    477 obs. of  11 variables:
##  $ X1   : int  1 1 1 1 1 1 1 1 1 ...
##  $ X1.1 : int  1 1 1 1 1 1 1 1 1 ...
##  $ X1.2 : int  1 1 1 1 2 2 2 3 3 3 ...
##  $ X1.3 : int  2 3 4 5 6 7 8 9 10 11 ...
##  $ X22  : int  53 20 19 24 28 32 42 74 40 17 ...
##  $ X2   : int  4 2 1 2 3 3 4 5 4 1 ...
##  $ X2.1 : int  2 2 2 2 1 1 1 1 1 1 ...
##  $ X1.4 : int  1 2 1 1 1 1 1 1 2 1 ...
##  $ X40  : int  40 40 40 40 40 40 48 40 20 18 ...
##  $ X160 : int  224 164 134 146 320 300 396 150 70 60 ...
##  $ X52.6: num  49 56.1 36.2 60 50.4 ...
```

Usamos ahora los valores requeridos para obtener la muestra solicitada

El total poblacional

```
N42=length(X1)
```

```
N42
```

```
## [1] 477
```

Y aplicando el paquete sampling obtenemos la muestra

```
library(sampling)
```

```
## Warning: package 'sampling' was built under R version 4.1.3
```

```
# tamaño de la muestra
n42=20
# selección de las unidades (trabajadores) que van a conformar la mues
tra
s42=sample(N42,n42)
# valores muestrales variable de interés y variable auxiliar
y42 = X160[s42]
x42 = X40[s42]
```

A continuación, logramos los cálculos para obtener el promedio con el estimador de razón basado en el total de Horvitz-Thompson

```
# total de la variable auxiliar
tx142=sum(X40)
total42=c(tx142)
# vector de probabilidades de inclusión para un m.a.s
pik42=rep(n42/N42,n42)
# cálculo del estimador de Horvitz-Thompson para el total poblacional
t_42=HTestimator(y42,pik42)

#obtencion del promedio poblacional mediante el estimador de razon
med_42=t_42/N42
med_42
```

De esta forma obtenemos la media de los ingresos semanales por trabajador mediante el estimador de razón.

```
##      [,1]
## [1,]  278
```

Ahora, obtendremos la estimación por Regresión.

Iniciamos definiendo el modelo

```
# definición del modelo de superpoblación y42 = beta0 + beta1*x42
modelo42=y42~x42

# matriz de probabilidades de inclusión para un m.a.s
pikl42=matrix(n42*(n42-1)/(N42*(N42-1)),n42,n42)

# cálculo del estimador de regresión 0
# considerando el modelo lineal y42 = beta0 + beta1*x42
r42=regest(formula=modelo42, weights= 1/pik42, Tx=total42, pikl42, n42
)
```

Luego obtenemos el estimador de Regresión con la siguiente sintaxis.

```
# estimador de regresi´on
t_ygreg42=r42$regest
med_ygreg42= t_ygreg42/N42
med_ygreg42
```

y finalmente, la media de los ingresos semanales por trabajador mediante el estimador de regresión.

```
## [1] 259.7213
```