Ejercicio 2. Tema 3.

Amalia Romero Peñalver

Encuestas por muestreo. Aplicaciones económicas, sociales y medioambientales.

Enero 2021

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?

Al observar el gráfico se podría decir que el **modelo de razón**, con su respectivo estimador de razón, sería el más adecuado por varios motivos.

En primer lugar, existe una única variable auxiliar x. En un modelo general de regresión por ejemplo, podría haber más variables auxiliares.

En segundo lugar, la relación entre la variable de interés y y la variable auxiliar x muestra una recta de ajuste que pasa por el origen.

También se puede apuntar que la varianza de y es proporcional a x , es decir, que a medida que la variable horas semanales aumenta, la dispersión también aumenta.

Respecto al error, se puede esperar un valor elevado para aquellos valores que son mayores que 40, ya que a partir de ese valor es donde mayor dispersión de los datos hay. Así, cuanto más alejados estén los valores de la recta, mayor será el error. El error asociado al estimador de razón deberá ser:

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.

Se podría caer en el error de querer usar el estimador de Horvitz-Thompson pero no sería correcto ya que es un estimador que no hace uso de ningún tipo de información auxiliar, es decir, que se calcula utilizando únicamente la información obtenida en la muestra y los pesos de muestreo. Además, cuando el tamaño muestral es pequeño, no es un estimador adecuado ya que es muy inestable y su varianza puede ser muy grande en estos casos.

En este caso, se cuenta con una variable auxiliar, que representa las horas semanales de trabajo de la población objeto de estudio, que se quiere utilizar para la estimación. Por lo tanto, el estimador general de regresión y el estimador de razón, que es un tipo de estimador de regesión, son más adecuados.

El estimador general de regresión es un estimador que utiliza información de la variable auxiliar, en este caso x para estimar la variable y. Este estimador utiliza el modelo de regresión como un medio para conseguir un estimador consistente desde el punto de vista del diseño. Requiere que el muestreo sea aleatorio.

Para comenzar, es necesario cargar el paquete *sampling* en R, así como el fichero de datos con los que se va a trabajar.

```
library(sampling)
Labor <- read.table("C:/Users/HP/Desktop/T/Labor.txt", quote="\"", comment.char="")
```

Para calcular ambos estimadores es necesario definir; N, que va a ser igual a la longitud de cualquier vector de la matriz de datos y n, que es igual al tamaño de la muestra que se quiere sacar de la población mediante muestreo aleatorio simple. Seguidamente se definen los valores muestrales para la variable de interés y y la variable auxiliar x. También se saca la muestra mediante la función *sample* y se define $t.x$ como el total de la variable auxiliar.

```
N=length(Labor [,1])
N
```

```
## [1] 478
```

```
n=20
s=sample(N,n)
s
```

```
##  [1] 443 477 408  41 256 287 416 332 382  90 170  98  25  92 113 296 473 431 450
## [20] 218
```

```
x = Labor$V9[s]
y = Labor$V10[s]
t.x=sum(Labor$V9)
t.x
```

```
## [1] 18294
```

Es preciso obtener tanto un vector *pik* como una matriz *pikl*, ambos de probabilidades de inclusión para una muestra aleatoria simple que van a ser necesarios para calcular ambos estimadores.

```
pik=rep(n/N,n)
pik
```

```
##  [1] 0.041841 0.041841 0.041841 0.041841 0.041841 0.041841 0.041841 0.041841
##  [9] 0.041841 0.041841 0.041841 0.041841 0.041841 0.041841 0.041841 0.041841
## [17] 0.041841 0.041841 0.041841 0.041841
```

```
pikl = matrix(n*(n-1)/(N*(N-1)),n,n)
pikl
```

```
##              [,1]        [,2]        [,3]        [,4]        [,5]        [,6]
##  [1,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##  [2,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##  [3,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##  [4,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##  [5,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##  [6,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##  [7,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##  [8,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##  [9,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [10,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [11,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [12,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [13,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [14,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [15,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [16,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [17,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [18,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [19,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [20,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##              [,7]        [,8]        [,9]       [,10]       [,11]       [,12]
##  [1,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##  [2,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##  [3,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##  [4,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##  [5,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##  [6,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##  [7,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##  [8,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##  [9,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [10,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [11,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [12,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [13,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [14,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [15,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [16,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [17,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [18,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [19,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [20,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##             [,13]       [,14]       [,15]       [,16]       [,17]       [,18]
##  [1,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##  [2,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##  [3,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##  [4,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##  [5,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##  [6,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##  [7,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##  [8,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##  [9,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [10,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [11,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
```

```
## [12,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [13,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [14,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [15,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [16,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [17,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [18,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [19,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
## [20,] 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623 0.001666623
##             [,19]       [,20]
##  [1,] 0.001666623 0.001666623
##  [2,] 0.001666623 0.001666623
##  [3,] 0.001666623 0.001666623
##  [4,] 0.001666623 0.001666623
##  [5,] 0.001666623 0.001666623
##  [6,] 0.001666623 0.001666623
##  [7,] 0.001666623 0.001666623
##  [8,] 0.001666623 0.001666623
##  [9,] 0.001666623 0.001666623
## [10,] 0.001666623 0.001666623
## [11,] 0.001666623 0.001666623
## [12,] 0.001666623 0.001666623
## [13,] 0.001666623 0.001666623
## [14,] 0.001666623 0.001666623
## [15,] 0.001666623 0.001666623
## [16,] 0.001666623 0.001666623
## [17,] 0.001666623 0.001666623
## [18,] 0.001666623 0.001666623
## [19,] 0.001666623 0.001666623
## [20,] 0.001666623 0.001666623
```

Para calcular el **estimador de regresión**, se define el modelo que es y~x y se aplica la función *regest* que realiza el cálculo.

```
modelo=y~x
reg=regest(formula=modelo, weights =  1/pik, Tx=t.x, pikl, n)
```

Esta función muestra varios parámetros del estimador de regresión. En este caso, es necesario utilizar *$regest* que da el total de las características de los ingresos semanales por trabajador, es decir, de la variable de interés y.

El modelo de regresión sirve para encontrar la expresión matemática que permite estimar tanto los coeficientes de regresión, *$coefficients*, como la eficiencia del estimador de regresión de la población total. Este último, normalmente se compara con la del estimador de Horvitz-Thompson pero, como ya se ha mencionado, cuando se tiene un tamaño muestral muy pequeño no es conveniente debido a su inestabilidad que puede hacer que su varianza sea muy grande. Aún así, en este caso, el estimador de Horvitz-Thompson da un valor no muy alejado al del estimador general de regresión. Lo mismo se puede decir de la media.

```
reg$coefficients
```

```
## x(Intercept)           xx
##    -5.438037    10.416345
```

```
t_ypi=HTestimator(y,pik)
t_ypi
```

```
##           [,1]
## [1,] 182619.9
```

```
mean(t_ypi/N)
```

```
## [1] 382.05
```

El estimador de regresión de la muestra va a permitir obtener la media de los ingresos semanales por trabajador.

```
t_reg=reg$regest
t_reg
```

```
## [1] 187957.2
```

```
mean(t_reg/N)
```

```
## [1] 393.216
```

A continuación, se utliza la función pertinente para obtener el **estimador de razón**. Esta función muestra directamente el total de las características de los ingresos semanales por trabajador y va a permitir sacar la media de esta variable de interés.

```
rat=ratioest(y,x,Tx=t.x, pik)
mean(rat/N)
```

```
## [1] 393.0593
```

Ya que se tiene la población de la variable de interés y, se puede obtener la media real de los ingresos semanales por trabajador que va a permitir la comparación.

```
mean(Labor$V10)
```

```
## [1] 294.5983
```

En este caso, ambos estimadores resultan de conveniencia similar. No se ha podido optar por uno concreto debido, posiblemente, a que la muestra no es de gran tamaño. Al estar trabajando en *R* y con el procesador de documentos *Rmarkdown*, cada vez que se genera el documento PDF, el valor de las medias varía debido a que la muestra lo hace. Se ha preferido no dejar este valor fijo, ya que este resultado es un buen indicador de la resolución del ejercicio planteado. Por lo tanto, la mejor opción para estimar la media de los ingresos semanales por trabajador será la que más se aproxime a la media real que tiene un valor de 294.5983.