

Computació i optimització en estadística

Practica SAS

Objetivo de la práctica:

En esta práctica se realizará un estudio sobre datos relacionados con mujeres que han tenido un embarazo en el último año en Cataluña. El objetivo de la presente práctica será estudiar esta información para conocer con más detalles las características de dicha población. Entre las preguntas planteadas está saber la edad en que las mujeres tienen un embarazo, qué características personales retrasan o adelantan el embarazo o qué factores están asociados a tener un embarazo prematuro.

Datos en estudio:

Los datos han sido extraídos del INE. Se ha filtrado los nacimientos pertenecientes a Cataluña y se han seleccionado aleatoriamente 6000 casos de mujeres embarazadas. Entre la información disponible se han seleccionado un total de 16 variables (1 identificador, 6 variables numéricas, 5 binarias y 4 categóricas).

- Variables numéricas: Edad madre, edad padre, peso nacido, número hijos en el embarazo, años de relación con la pareja, semanas de embarazo.
- Variables binarias (Sí/No): Cesárea, complicaciones parto, prematuridad parto, nacionalidad española de la madre y sexo del nacido (hombre/mujer).
- Variables categóricas: Nivel estudios madre, Tamaño municipio nacimiento madre, provincia nacimiento, lugar del parto).

1. Leer los datos que ha obtenido desde SAS para crear una base de datos en formato SAS.

Se ha realizado la importación de los datos con formato xls a partir del comando `proc import`.

2. Asignar etiquetas de variable y de valores a las variables de su archivo.

Se han asignado etiquetas a un conjunto de variables a partir de `proc format`. Como ejemplo se ha codificado el código que indica la provincia de nacimiento con las etiquetas de (Barcelona, Girona, Tarragona, Lleida).

3. Realizar un descriptivo para comprobar que no tenga valores de las variables fuera de rango u observaciones extremas que podamos considerar outliers. Indicar en el propio script si todo ha salido bien o no.

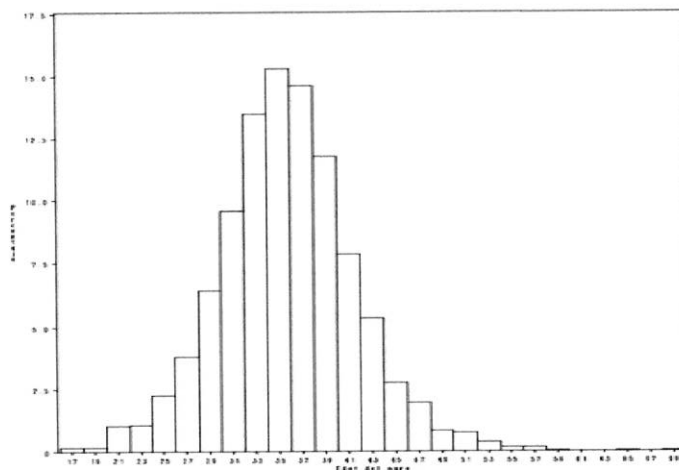
Se realiza la búsqueda de observaciones extremas a partir de inspección visual (histograma) y de las salidas sobre outliers que proporciona *proc univariate*. Se observan

valores extremos en variables como edad del padre (valores de más de 60 años) o del peso del recién nacido.

Si nos centramos en el peso de los hijos/hijas se observa como la media de peso es de 3222 gramos, pero existen valores extremos como pesos que no llegan a los 600 gramos o otros que pasan de los 5000:

Observaciones extremas			
Interior		Superior	
Valor	Observación	Valor	Observación
525	3847	4773	1841
535	5843	4800	2893
550	1588	4810	219
600	5876	5400	1789
620	428	5700	1337

Gráficamente también se pueden detectar valores altos en la edad de los padres:



Otro aspecto que se ha detectado es la existencia de valores faltantes (missing) en la variable de estudios de la madre.

4. Si encuentra algún tipo de error en el apartado anterior o bien observaciones muy extremas, decidir qué hacer con estas observaciones (eliminarlas? Convertirlas en missing? conservarlas?) Dejar constancia de lo que decida y hacerlo.

Los posibles valores extremos encontrados en el apartado anterior son casos que pueden ser factibles. No parece que sean errores en la entrada de datos sino más bien situaciones excepcionales que escapan de la normalidad de la situación. Consecuentemente se ha decidido no hacer ningún tipo de cambio en los datos.

Al tratarse de datos provenientes de organismos oficiales era esperable que tuvieran una cierta calidad.

5. Obtener un descriptivo completo de entre tres y cinco de las variables que le parezcan más relevantes.

Se realiza un estadístico completo de las variables edad de la madre, nivel educativo de la madre, peso y semanas de embarazo, a partir de los comandos *proc univariate*, *proc freq* y *proc means*.

Edad de la madre (*proc univariate*):

Procedimiento UNIVARIATE

Variable: Edat_mare (Edad de la mare)

Momentos			
N	6000	Sumar pesos	6000
Media	32.238	Observ suma	193428
Desviación std	5.32281362	Varianza	28.3334116
Asimetría	-0.3219774	Curtosis	0.10593076
SC no corregida	6405704	SC corregida	169972.136
Coeff. variación	16.5113029	Media error std	0.08871652

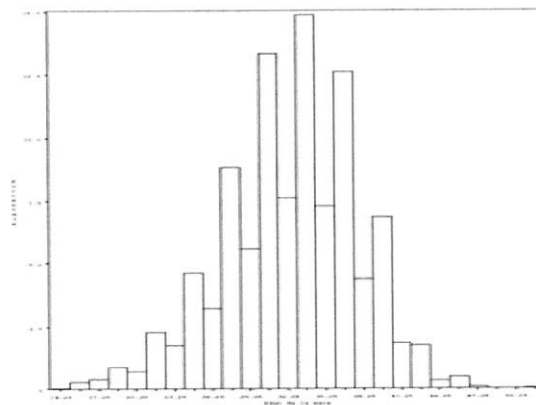
Medidas estadísticas básicas			
Ubicación		Variabilidad	
Media	32.23800	Desviación std	5.32281
Mediana	33.00000	Varianza	28.33341
Moda	32.00000	Rango	38.00000
		Rango intercuartil	7.00000

Tests para posición: Mu0=0			
Test	Estadístico		p valor
T de Student	t	469.1312	Pr > t <.0001
Signo	M	3000	Pr >= M <.0001
Puntuación con signo	S	9001500	Pr >= S <.0001

Cuantiles (Definición 5)	
Nivel	Cuantil
100% Máx	52
99%	43
95%	40
90%	39
75% Q3	38
50% Mediana	33
25% Q1	29
10%	25

Cuantiles (Definición 5)	
Nivel	Cuantil
5%	23
1%	18
0% Min	14

Observaciones extremas			
Inferior		Superior	
Valor	Observación	Valor	Observación
14	2783	47	4697
14	2589	47	5842
14	2169	48	2432
15	5542	51	3082
15	4808	52	5298



Peso y semanas de embarazo (proc means):

Procedimiento MEANS

Variable	Etiqueta	Media	Mediana	Varianza	Mínimo	Máximo	Asimetría	Kurtosis
Pes	Pes del Nado	3222.93	3260.00	293328.94	525.00	5700.00	-0.78	2.28
Setmanes_embaras	Num de setmanes de embaras	39.02	39.00	3.60	24.00	43.00	-2.17	9.24

Variable	Etiqueta	Pct 5	Cuantil superior
Pes	Pes del Nado	2270.00	3567.50
Setmanes_embaras	Num de setmanes de embaras	36.00	40.00

Nivel educativo (proc freq):

Procedimiento FREQ

Estudis	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
Cicle formatiu	1014	16.92	1014	16.92
ESO/Batxillerat	1780	29.70	2794	46.62
Sense estudis	805	13.43	3599	60.05
Universitat	2394	39.85	5993	100.00
Total de valores ausentes = 7				

Entre las variables estudiadas, se observa como la edad media de las mujeres al embarazo es de 32.2 años, se está de media 39 semanas de embarazo y el peso de la criatura es de 3222 gramos.

En cuanto a los estudios se observa como el 40% de las madres tienen estudios universitarios o superiores mientras que el 13% no tienen estudios.

6. Definir agrupamientos (realizar una categorización) de una o más de las variables numéricas originales.

Se ha decidido realizar un agrupamiento para la edad de la madre. De esta manera conseguiremos una variable categórica a partir de la información de una variable numérica. Los criterios de categorización han sido los siguientes.

$$\begin{aligned}
 \text{Edad} < 20 &= \text{"Menos de 20"} \\
 20 \leq \text{Edad} < 28 &= \text{"Entre 20 y 27"} \\
 28 \leq \text{Edad} < 36 &= \text{"Entre 28 y 35"} \\
 36 \leq \text{Edad} < 43 &= \text{"Entre 36 y 42 años"} \\
 \text{Edad} > 42 &= \text{"Más de 42 años"}
 \end{aligned}$$

7. Utilizar el lenguaje IML para calcular las correlaciones entre las variables numéricas.

Se realiza la correlación de las 6 variables numéricas en estudio a partir del coeficiente de correlación de Pearson (viene por defecto en SAS) y Spearman. Se observa como las variables que presentan una mayor correlación son el peso con las semanas de embarazo, y la edad del padre y la madre.

Pearson correlation matrix

corr							
	ID	Numero_nascuts	Setmanes_embaras	Edat_mare	Anys_relacio	Edat_pare	Pes
ID	1	0.0098238	-0.001368	-0.010068	-0.000176	-0.010993	0.0030964
Numero_nascuts	0.0098238	1	-0.325885	0.0770928	0.0472634	0.0562652	-0.344667
Setmanes_embaras	-0.001368	-0.325885	1	-0.056893	-0.018364	-0.033823	0.5519546
Edat_mare	-0.010068	0.0770928	-0.056893	1	0.3353544	0.6117227	-0.052383
Anys_relacio	-0.000176	0.0472634	-0.018364	0.3353544	1	0.2396503	-0.005081
Edat_pare	-0.010993	0.0562652	-0.033823	0.6117227	0.2396503	1	0.0018191
Pes	0.0030964	-0.344667	0.5519546	-0.052383	-0.005081	0.0018191	1

Spearman correlation matrix

spearman							
	ID	Numero_nascuts	Setmanes_embaras	Edat_mare	Anys_relacio	Edat_pare	Pes
ID	1	0.007564	0.0079915	-0.012988	0.0036956	-0.01274	0.0080472
Numero_nascuts	0.007564	1	-0.26938	0.0747291	0.0386242	0.0579918	-0.281105
Setmanes_embaras	0.0079915	-0.26938	1	-0.054759	-0.012077	-0.03865	0.4191272
Edat_mare	-0.012988	0.0747291	-0.054759	1	0.302616	0.6078416	-0.0298
Anys_relacio	0.0036956	0.0386242	-0.012077	0.302616	1	0.2240214	0.0149436
Edat_pare	-0.01274	0.0579918	-0.03865	0.6078416	0.2240214	1	0.0096389
Pes	0.0080472	-0.281105	0.4191272	-0.0298	0.0149436	0.0096389	1

8.Hacer algún tipo de análisis de algunas variables continuas, cruzándolas con alguna/s variables categóricas o bien con la/las variables agrupadas en el apartado anterior, debidamente etiquetadas.

Se realiza un modelo de regresión logística para estudiar qué variables están asociadas con que el embarazo sea prematuro. La variable respuesta es binaria (Prematuro/a tiempo) y se utilizan como predictoras un total de 8 variables (categóricas, binarias y numéricas). Los predictores considerados fueron seleccionados en base a aquellas que consideramos como posibles causantes o posibles variables confusoras. Las podemos ver todas en las siguientes tablas que contienen los resultados.

Procedimiento LOGISTIC	
Información del modelo	
Conjunto de datos	PRAC.DADES
Variable de respuesta	Prematur
Número de niveles de respuesta	2
Modelo	logit binario
Técnica de optimización	Puntuación de Fisher

Número de observaciones lei	6000
Número de observaciones usa	5993

Perfil de respuesta		
Valor ordenado	Prematur	Frecuencia total
1	Prematur	446
2	A termini	5547

La probabilidad modelada es Prematur='Prematur'.

A continuación se muestran el peso de cada variable:

Tipo 3 Análisis de efectos			
Efecto	DF	Chi-cuadrado de Wald	Pr > ChiSq
Sexe	1	0.0135	0.9076
Estudis	3	32.9448	<.0001
Edat_mare	1	3.3321	0.0679
Edat_pare	1	0.9105	0.3400
Nacionalitat	1	9.6344	0.0019
Numero_nascuts	1	387.6248	<.0001
Provincia	3	0.4992	0.9191
Anys_relacio	1	0.3678	0.5442

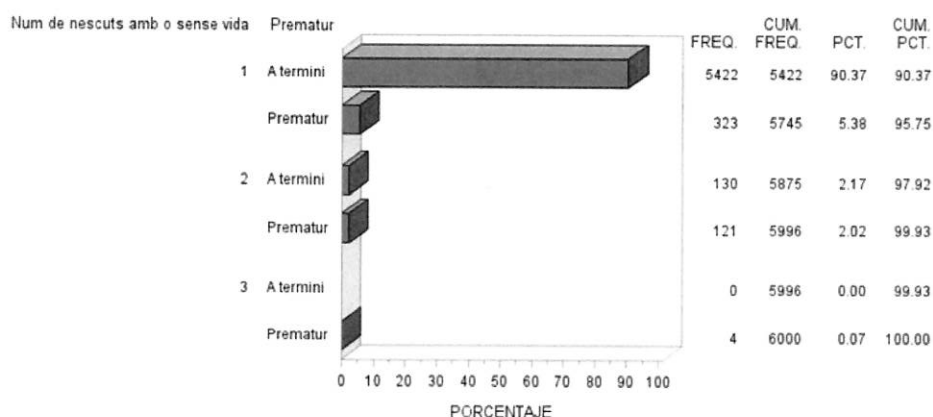
Y las odds-ratios de las comparaciones realizadas:

Estimadores de cocientes de disparidad;				
Efecto		Estimador del punto	Límites de confianza al 95% de Wald	
Sexe	Dona vs Home	1.012	0.824	1.244
Estudis	Cicle formatiu vs ESO/Batxillerat	0.832	0.601	1.151
Estudis	Sense estudis vs ESO/Batxillerat	1.995	1.444	2.756
Estudis	Universitat vs ESO/Batxillerat	0.757	0.581	0.987
Edat_mare		1.025	0.998	1.053
Edat_pare		0.989	0.966	1.012
Nacionalitat	No vs Si	0.643	0.486	0.850
Numero_nascuts		16.028	12.160	21.126
Provincia	Girona vs Barcelona	0.900	0.623	1.299
Provincia	Lleida vs Barcelona	0.934	0.584	1.493
Provincia	Tarragona vs Barcelona	0.926	0.652	1.314
Anys_relacio		0.992	0.968	1.017

Se observa cómo la variable que más significativamente se asocia con que un embarazo sea prematuro es el número de hijos que se tienen en el embarazo. A más hijos se tengan en el vientre más aumenta la odds de que el parto sea prematuro (OR=16, 12.1 - 21.1 IC95%). El nivel de estudios también marca las probabilidades de embarazo prematuro, las mujeres sin estudios son las que tienen mayores odds de embarazo prematuro. En sentido opuesto, se observa cómo variables como el sexo de la criatura, la provincia o los años que la pareja lleva de relación no parecen estar asociadas con la prematuridad de manera significativa.

A partir de estos resultados se ha decidido visualizar la relación entre las variables número de hijos en el embarazo y prematuridad. En el siguiente gráfico mostramos las proporciones de prematuridad condicionadas por el número de bebés en el embarazo:

Proporciones de prematuridad condicionando por número de bebés



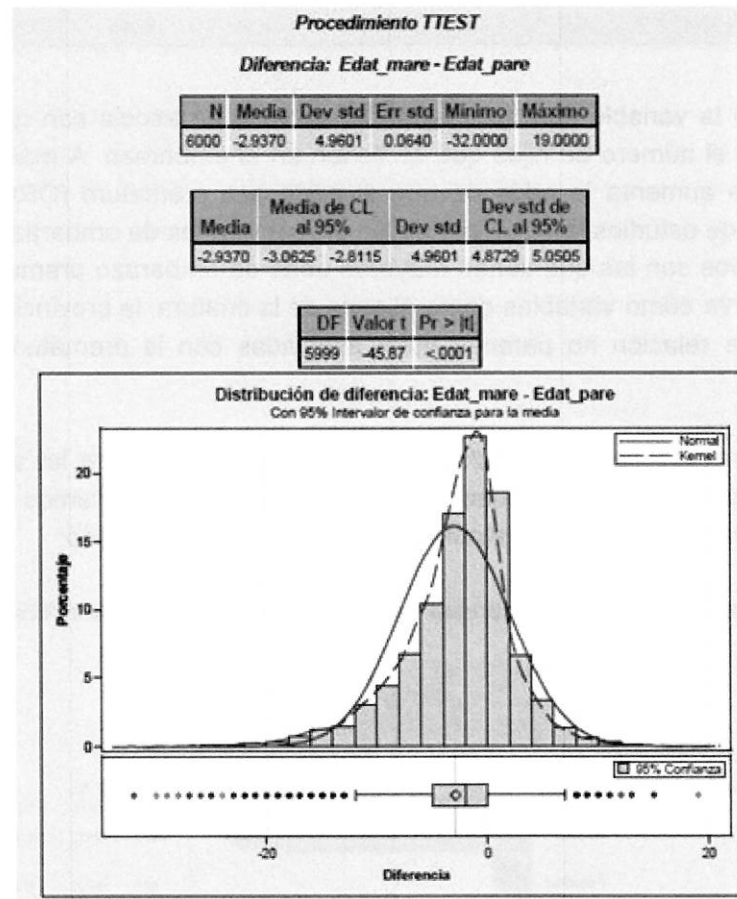
9.Hacer otros análisis que os puedan interesar. Indicar previamente a la ejecución de los PROCESOS, en forma de comentario, cuál es el objetivo del análisis que desea realizar.

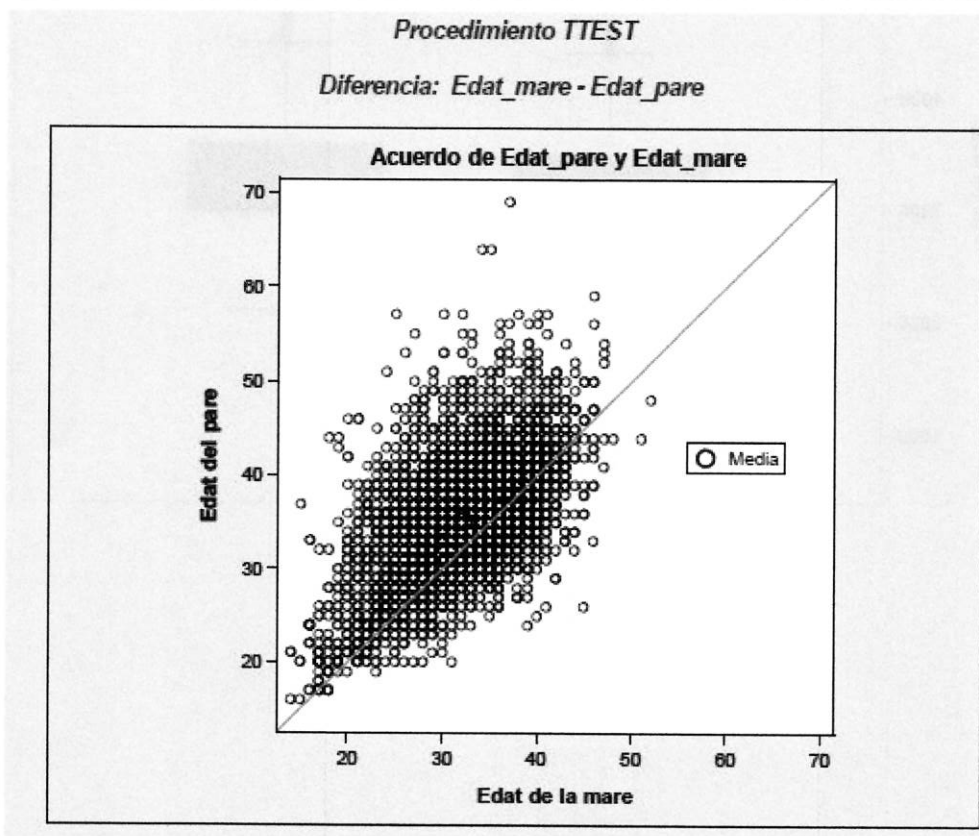
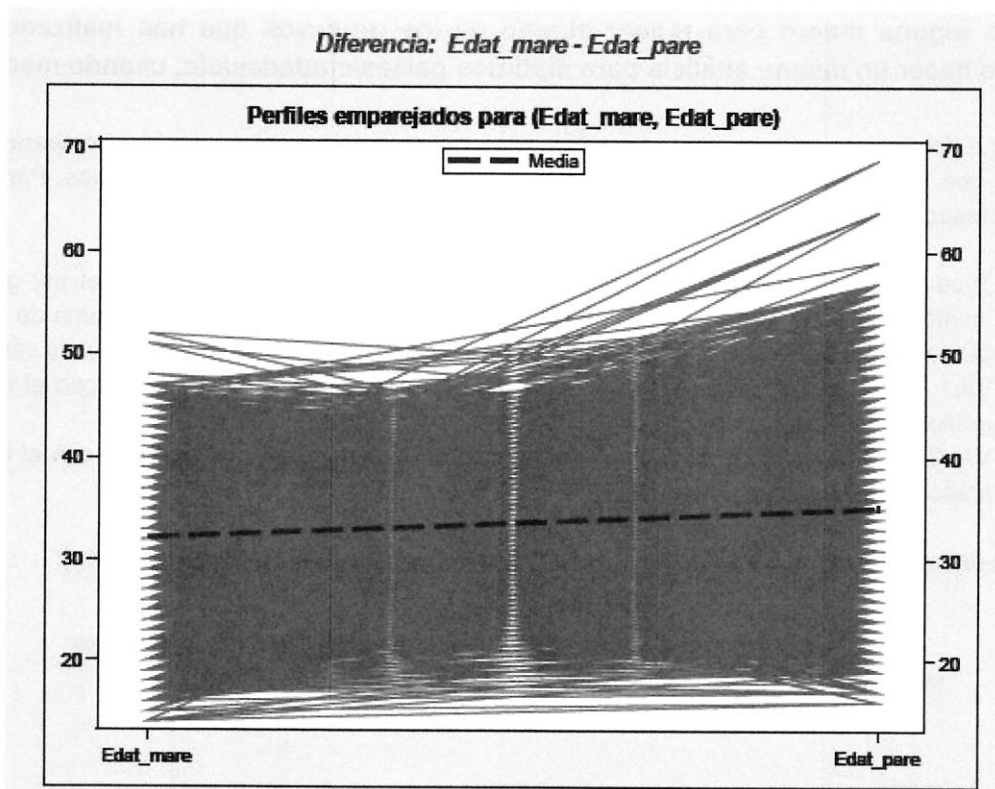
Se quiere estudiar si la edad de los hombres es mayor que el de las mujeres en el momento del embarazo, (o de manera más general, si hay diferencias entre la edad del hombre y la mujer entre las parejas que tienen hijos). La hipotesi nula es que consideramos es que no hay diferencias.

Para realizar dicho estudio se crea realizamos un t-test para muestras apareadas. Las hipótesis del test son la siguientes:

$$H_0 : \text{Diferencia de edad} = 0$$

$$H_1 : \text{Diferencia de edad} \neq 0$$





Y se observa como la edad de los hombres es estadísticamente mayor, con una diferencia de 2.9 años (2.8 - 3.1 IC95%, p-valor <0.001).

10. Crea alguna macro para repetir alguno de los procesos que has realizado, por ejemplo hacer un mismo análisis para distintos países/ciudades/etc, usando macros

Por último nos ha parecido que podría ser útil una macro que nos permitiera realizar comparaciones gráficas con comparaciones dos a dos entre variables cualquiera de la base de datos. Para ello, hemos creado la macro %CompareVars que toma los siguientes argumentos:

- Type: tipo de variables que vamos a introducir. "Contbycat" realizará boxplot por grupos, "contbycont" realizará un gráfico de dispersión y "catbycat" realizará un diagrama de barras 3D con información sobre los porcentajes marginales condicionando por la segunda variable.
- Var1: variable que irá en el eje x (el tipo de la variable tiene que coincidir con el tipo de gráfico especificado o dará error).
- Var 2: variable que irá en el eje y (el tipo de la variable tiene que coincidir con el tipo de gráfico especificado o dará error).

Ejemplo de gráfico obtenido mediante la ejecución de %Comparevars(contbycat,pes,sexe):

