San Antonio 10 de julio de 2009

RESOLVIENDO EL PROBLEMA DE IMPUTACIÓN USANDO SAS.

Andrés E Reyes Polanco[1].

Palabras claves: datos faltantes, imputación, imputación simple con SAS, imputación múltiple con SAS, proc mi, proc mianalyze.

INTRODUCCIÓN.

La siguiente monografía tiene como finalidad abordar el problema de datos faltantes en un archivo de datos, usando como herramienta computacional el software SAS.

Un problema común en la recolección de datos primarios es la presencia de datos perdidos que puede explicarse por varios motivos, por ejemplo, en una encuesta sobre el ingreso familiar pueden darse casos en donde la persona se niegue a dar la información o simplemente no se encuentre la persona adecuada que pueda suministrarla. Otro caso es aquel donde se encuentran valores atípicos. Hay dos formas de ver la presencia de valores atípicos: uno desde el punto de vista probabilístico y el otro como consecuencia de errores humanos. Desde el punto de vista probabilístico, se puede definir como valor atípico, aquel que según la población en estudio, la probabilidad de ocurrencia de ese valor es casi nula y sin embargo ocurrió. Pero cuando ese no es el caso sino por motivos distintos a la ley de probabilidad, entonces debe eliminarse convirtiéndose en un valor perdido. Por tanto en un archivo de datos debe detectarse los valores perdidos y aquellos atípicos productos de errores humanos que deben considerarse como valores faltantes cuando no es posible corregirlos.

Las soluciones dadas a este problema son muy variadas, una de ellas es desechar en el análisis estadístico aquellas variables que para al menos una observación no existe el dato correspondiente. Esto ocurre cuando se está aplicando técnicas de análisis de datos tales como regresión múltiple, componentes principales, análisis de correspondencia entre otros, en el caso de regresión múltiple se requiere que todas las variables explicativas contengan todas

[1] Profesor Asociado de la Universidad Central de Venezuela.

los datos de las observaciones, en los casos de reducción de dimensionalidad como componentes principales todas las observaciones de todas las variables deben tener sus datos presentes. La otra solución, en el caso de análisis univariante, es trabajar con los datos existentes y hacer las estimaciones de los parámetros con los datos disponibles en cada variable, obviando la presencia de valores faltantes.

Resolver el problema por alguna de las dos formas anteriores trae consecuencias a los resultados inferenciales, entre ellos que los estimadores dejan de ser eficientes y insesgados.

Otra forma de solucionar el problema es usar imputación, esto es, sustituir los datos faltantes por valores obtenidos mediante alguna técnica estadística. Estas técnicas se divide en dos grandes grupos: imputación simple e imputación múltiple. Al mismo tiempo, tiene que considerarse la escala de medición de las variables para aplicar una de las técnicas adecuadas y la estructura de los datos perdidos.

En el caso en que las variables que se van imputar son métricas (escala de intervalo o escala de razón), la imputación simple consiste en sustituir el dato perdido de cada observación por un valor ajustado, obtenido bien sea la media aritmética condicional de las observaciones existentes en cada una de las variables, o bien por regresión estándar o regresión estocástica y finalmente en el caso de muestreo aleatorio, por el método de máxima verosimilitud (Gómez García et al 2006; Medina F et al 2007). De este tipo de imputación se recomienda la regresión estocástica o el método de máxima verosimilitud empleando el algoritmo EM (Dempster N et al 1977). Cuando las variables no son métricas (escala nominal o escala ordinal) se emplea el ajuste obtenido por la regresión logística, probit o la función discriminante.

Una situación especial de datos faltantes y que acarrea ciertas dificultades es cuando se está construyendo un modelo en donde se combinan variables semicontinuas con otras variables métricas y categóricas tales como los modelos del tipo Tobit (Shafer, J. L et al 1999).

La otra técnica estadística es la imputación múltiple desarrollada originalmente por Rubin y continuada por varios autores (Rubin 1977, 1996; Roderick et al 1988; Daniel et al 1991, Zhang 2003), para ello se obtiene un conjunto de datos en forma vectorial mediante simulación (tantos vectores como datos perdidos o faltantes hay) y a cada dato perdido se le asigna los valores de uno de los vectores obtenidos.

La imputación múltiple trae como consecuencia la multiplicación del archivo, si se decide replicar la simulación, por ejemplo, seis veces, al final tendremos seis matrices de datos.

De la misma forma como en el caso de imputación simple, en la múltiple también se toma en cuenta la escala de medición. Todas estas técnicas están orientadas a observaciones resultantes de aplicar técnicas de muestreo. Sin embargo, algunas de estas técnicas son extensivas a otras situaciones siempre que se hagan los cambios pertinentes, tales como datos almacenados en archivos para ser utilizados en la minería de datos o datos censales.

Se han propuesto limitaciones en cuanto al uso de estos métodos de imputación, por una parte lo relacionado con el porcentaje de datos faltantes y, por otra, el uso que se dará de los datos en la toma de decisiones. Para Medina F et al (2007) un porcentaje de datos faltantes en las variables bajo estudio que ronde el 25%, la imputación debería descartarse si además su aplicación es para definir políticas públicas. Por tanto, lo recomendable es hacer revisitas y en el peor de los casos repetir la investigación. Además de estos dos, hay una limitación que tiene una importancia especial, esta se refiere al diseño de la muestra que se ha aplicado para seleccionar las unidades de la muestra. ¿Se trata de un diseño en donde las unidades tienen la misma probabilidad de selección o no?

Otro punto importante que no debe obviarse es la estructura que da origen a los datos perdidos que son básicamente dos: aleatoria o arbitrario como así llama SAS a esta estructura y, monótona.

Ningún método en particular es una solución final, de tal forma, que habrá que seleccionar varios métodos y decidir cuál o cuáles son los más adecuados en cada caso. La mayoría de los métodos están orientados a problemas de diseños de muestras aleatorias y los supuestos suelen restringirse al supuesto de distribución normal multivariante o a otros adicionales considerando cuál es la estructura de los datos perdidos, si es un patrón aleatorio o un patrón monótono, si los valores perdidos dependen o no de las observaciones de otras variables o de las propias.

La siguiente monografía a parte de la introducción se divide como sigue: la segunda parte se refiere a la la distribución y estructura de los datos faltantes, en la tercera los tipos de imputación, la cuarta se hace una breve exposición del fundamento teórico de la imputación simple y algunos métodos, el quinto se da la exposición teórica del algoritmo EM, en la sexta, las ventajas y desventajas de la imputación simple, en la séptima hace una breve exposición del fundamento teórico de la imputación múltiple, en el octavo punto se hace la exposición del

análisis en conjunto y finalmente se hacen algunos comentarios, que ayudan a definir una metodología de imputación. En los puntos anteriores se dan ejemplos utilizando el software SAS.

II.-DISTRIBUCIÓN Y ESTRUCTURA O PATRON DE LOS DATOS FALTANTES.

El conocimiento de la estructura de datos faltantes, junto con la escala de medición (métrica o no métrica) es fundamental para la aplicación del método de imputación. La presencia de datos faltantes responden a dos tipos de categorías: presencia por causa "arbitraria" o aleatoria y presencia con un patrón monótono.

Antes de hablar de la estructura de los datos perdidos o de los atípicos que pueden considerarse como tales, daremos la notación que se empleará durante el desarrollo de esta monografía. Dado la matriz de datos: $X = \|x_{ij}\|$ de p variables y n observaciones, particionemos la matriz en variables con todos los datos observados y las variables con datos faltantes:

$$X = \|X_{obs} | X_{per}\|$$

Consideremos una nueva matriz:

$R = \|r_{ij}\|$

Tal que

$$r_{ij} = 1 \Leftrightarrow x_{ij} = obs$$
$$r_{ij} = 0 \Leftrightarrow x_{ij} = no - obs$$

Esto es, cuando existe la observación se le asigna uno y en caso contrario cero.

El problema consiste en estimar el parámetro θ ; $\theta \in \Omega$; sabiendo que hay datos faltantes, reflejados en la matriz R en donde su distribución de probabilidad está gobernado por el parámetro ξ.

La distribución conjunta de X y R con parámetros: θ y ξ es:

$$P(X, R; \theta, \xi) = P(X/\theta)P(R/X, \zeta)$$

La forma de la distribución condicional de R dado X es por definición:

$$P(R/X, \xi)$$

De aquí se originan los siguientes casos, estudiados por Rubin (1976), Zhang P. (2003) entre otros:

- Si el mecanismo que da origen a los datos perdidos es independiente de las respuestas, es decir, es independiente tanto de los datos observados como de los perdidos, entonces el mecanismo es MCAR (Missing Completely At Random). Esto es: $P(R/X;\xi) = P(R/\xi,)$

- Cuando el mecanismo es independiente de los datos perdidos pero no de los observados, entonces este se define como MAR (Missing At Random). Esto es: $P(R/X,\xi) = P(R/\xi, X_{obs}))$

- Por último, tenemos el MNAR que se da cuando el mecanismo de datos perdidos no depende de las observaciones existentes en la variable considerada ni los existentes en las otras variables. Esto es: $P(R/X;\xi) \neq P(R/\xi, X_{obs})$

De los dos primeros casos el menos restrictivo es el MAR y este supuesto es de donde parte todas las aplicaciones de imputación de SAS en los procedimientos MI y MIANALYZE. La hipótesis MCAR es más restrictiva puesto que supone la independencia X y R. Y finalmente la última hipótesis es la más general pero poco importante en la práctica.

El MAR se puede ilustrar con el siguiente ejemplo, supongamos que tenemos tres variables: X_1, X_2 y X_3

Además, las dos primeras variables tienen todas sus observaciones mientras la tercera tiene datos perdidos, entonces, la probabilidad de que un individuo tenga un dato perdido en X_3 depende de las observaciones de X_1 y X_2 y no de las observaciones de X_3. .

Little. R(1988) ha propuesto un test diferente a los propuestos por otros autores tales como Dixon que utilizan la prueba t de student, para determinar si el comportamiento de los datos perdidos siguen un MCAR. Otro mecanismo es usar los test de aleatoriedad como el test de las rachas.
Las dos estructuras posibles que se presentan los datos perdidos son la aleatoria o la monótona. Para cada uno de estos casos hay métodos para resolver los datos faltantes.

IIa.-Aleatoria o arbitrario.

La estructura aleatoria es tal como se ilustra en la siguiente figura:

X1	X2	X3	X4	X5	X6	X7	X8	X9
.	x	x		x	.	x	x	x
.	x	x	x	x	x	x	x	x
x	.	x	x	x	x	x	,	x
x	x	x	.	x	x	x	x	.
x	x	.	x	x	.	x	x	x
.	x	.	x	.	x	x	x	x

En donde las x representan datos observados y los . , datos faltantes. Se puede observar, que en este caso no se ve ningún patrón, los datos perdidos suponen un comportamiento aleatorio.

IIa1.-Modelo del comportamiento aleatorio.

Retomemos la distribución conjunta $P(X,R;\theta,\xi)$ como se recordará $X=\left\|X_{obs}\middle|X_{per}\right\|$

Entonces, podemos reescribir la función conjunta como: $P(X_{obs},X_{per},R;\theta,\xi)$

Ahora si queremos estimar $\theta\in\Omega$ partiendo de la función de verosimlitud, encontramos que tiene una parte no observada dada por**:** X_{per} por tanto**,** esta función no puede ser evaluada directamente. Pero sabemos que la función de verosimilitud dado los datos observados es proporcional a la distribución marginal: $P(X_{obs},R;\theta,\xi)$

$$L(\theta,\xi/X_{obs},R)\propto P(X_{obs},R;\theta,\xi)$$

Donde, por definición la distribución marginal es:

$$P(X_{obs},R;\theta,\xi)=\int_{R^k}P(X,R;\theta,\xi)dX_{per}=\int_{R^k}P(X/\theta)P(R/X,\zeta)dX_{per}$$

Ahora considerando los casos de MCAR y MAR, y obtenemos los siguientes resultados:

Si es MCAR, entonces: $P(X_{obs}, R; \theta, \xi) = P(R; \xi)P(X_{obs}; \theta)$ por ser independiente tantos de los datos perdidos como de los observados de la variable de interés o de cualquiera otra del conjunto de datos.

Si es MAR, entonces: $P(X_{obs}, R; \theta, \xi) = P(R / X_{obs}; \xi)P(X_{obs}; \theta)$. En este caso la verosimilitud se ha factorizado en dos partes, correspondiendo cada una a un parámetro. Si los parámetros θ y ξ aportan poca información entre sí, entonces la inferencia de θ se puede lograr considerando solamente la distribución: $P(X_{obs}; \theta)$

Esto es:

$L(\theta / X_{obs}) \propto P(X_{obs}; \theta)$

IIb.-Monótona.

En el cuadro siguiente se ilustra la ocurrencia de datos faltantes con una estructura monótona. Formalmente los datos perdidos tienen una estructura monótona si x_{ij} es un dato perdido entonces x_{ik} también es un dato perdido para $\forall k > j$

y1	y2	X1	X2	X3	X4
.x	x	x	x	x	x
x	x	x	x	x	
x	x	x	x		
x	x	x	.		
x	x	.			

Una estructura de este tipo se presenta en muestras provenientes de triales clínicos, datos de panel como el estudio del índice de precio al consumidor, seguimiento de cohortes estudiantiles etc. Se puede observar además que los tamaños de valores observados tienen un comportamiento monótono decreciente.

IIb1.- Modelo del comportamiento monótono.

El modelo en este caso siguiendo a Little (1995) es como sigue: se puede observar en el cuadro anterior que las variables y1,y2 tienen todas las observaciones, mientras las variables x1…x4 no tienen todas las observaciones siguiendo el patrón dado anteriormente. Entonces, podemos generalizar el modelo de la siguiente manera: $Y_1, Y_2 ... Y_r$ son r variables con todas sus observaciones, adicionalmente algunas de estas variables pueden representar el tiempo y, $X_1, X_2 ... X_p$ representan las p variables con algunas variables con datos perdidos. Cada observación es un vector fila tal como: $z_i = (y_{i1}, y_{i2} ... y_{ir} \left| x_{i1}, x_{i2} ... x_{ip} \right.)$. Por el momento nos ocuparemos de las variables $X_1, X_2 ... X_p$ cuyos valores pueden estar presente o no y que podemos escribirlas como en el punto anterior como: $X = (X_{obs} \left| X_{per} \right.)$

Consideremos ahora, similar al modelo aleatorio visto anteriormente una variable indicadora r_i tal que: $r_i = 0$ si la observación tiene todos los datos de todas las variables y $r_i = k$ si existen observaciones hasta la variable X_{k-1} y en las restantes desde k hasta p donde los valores están ausentes.

Cada $x_{i,j}$ puede representarse su valor esperado como: $E(x_{i,j} / \beta_j) = \beta_{i0} + \beta_{ij} t_{ij}$ que es un modelo de regresión en el tiempo donde los parámetros son aleatorios a diferencia de la regresión estándar. Por otra parte se asume que las observaciones de los individuos son independientes entre sí. Por tanto, considerando este último supuesto podemos encontrar el estimador de $\theta \in \Theta$ partiendo de la función de verosimilitud, dada como:

$$L(\theta / X_{obs}, R) = const \prod_{i=1}^{n} \int P(x_{obs,i}, x_{per,i}, r_i, \beta_j / y_i, \theta) dx_{per,i} d\beta_j$$

IIc.-Máxima verosimilitud.

Vistos las dos estructuras anteriores (arbitraria y monótona) ha quedado establecido el método de máxima verosimilitud como un método para estimar los parámetros de observaciones con datos perdidos. La intuición nos indica que los pasos para obtener un estimador por este método deben ser:

1. Maximizar la función de verosimilitud con las observaciones que tengan todos los datos en todas las variables, para obtener las estimaciones de los parámetros. Un ejemplo es el modelo de regresión.

2. Predecir los valores faltantes dado las estimaciones obtenidas en el punto anterior. Por ejemplo los valores predicho por el modelo de regresón.

3. Maximizar nuevamente la función de verosimilitud con las observaciones que tienen todos los datos en todas las variables más las observaciones cuyos datos se han completado con la predicción.

Los puntos dos y tres se iteran hasta que se logre la convergencia, entendiendo por tal, aquella situación en donde la estimación del parámetro en el paso t: $\hat{\theta}^{t}$ y el obtenido en el paso t+1: $\hat{\theta}^{t+1}$ difieran entre sí en valor absoluto a una cantidad arbitrariamente pequeña. $\left|\hat{\theta}^{t} - \hat{\theta}^{t+1}\right| < \varepsilon$

Un algoritmo que logra la convergencia eficientemente es el algoritmo EM (Expectation-Maximization), desarrollado por Dempster. A.P., N.M. Lair and D.B. Rubin (1977) y que veremos más adelante.

III.-TIPOS DE IMPUTACIÓNES.

Se han desarrollado diferentes tipos de imputación tanto para datos muestrales como censales, para este último caso se ha hecho popular el método aplicado en Canadá y E.E.U.U conocido como **CANCEIS** desarrollado por Bankier, (Bor-Chung Chen) pero igualmente se ha aplicados los métodos de imputación simple empleados en el muestreo tales como el método de regresión múltiple (Todd R. Williams) que en el caso de datos censales sólo se requiere estar atento a la presencia de colinealidad y al valor del coeficiente de determinación. Tanto para datos proveniente de una muestra o de un censo es posible combinar el análisis de conglomerado y la imputación. Veremos más adelante que SAS tiene un procedimiento para cumplir con este último propósito.
En el caso de investigación por muestreo, los tipos de imputación son: simple y múltiple, bajo una estructura aleatoria o monótona. Por otra parte la imputación puede ser univariante o multivariante. Los métodos propuestos para imputación múltiple de variables continuas generalmente están construidos bajo el supuesto de normal multivariante. Si existe poca asimetría y kurtosis, tal que la distribución bajo estudio no difiere fuertemente de la distribución normal puede emplearse el método de máxima verosimilitud que es robusto cuando no se cumple estrictamente el supuesto de normalidad.

Antes de decidir que método de imputación debe usarse, es importante incorporar un análisis exploratorio de datos que permita visualizar a que estructura responden los datos perdidos y cuales pueden considerarse como atípicos. Dentro de las posibles herramientas que pueden acompañar al proceso de imputación están las técnicas de reducción de dimensionalidad y de conglomerados; las primeras permiten detectar valores atípicos que tendrán que eliminarse según el caso como ya indicamos inicialmente y el uso de conglomerados permitirá afinar la imputación. Los conglomerados deben formarse con aquellas variables que están asociadas con la variable objeto de imputación y el método y número de conglomerados dependerán del investigador.

IV. IMPUTACIÓN SIMPLE.

La imputación simple consiste en asignar a cada valor perdido uno y solamente un valor partiendo de los datos observados. Una práctica muy común es sustituir los datos perdidos por la media de los valores observados de las variables que presentan esta situación, obviamente esta es la peor forma de imputación porque afecta la variabilidad inherente de las variables. Mejor que esta solución, es definir por algún método varios conglomerados y asignar a los valores perdidos la media de acuerdo al conglomerado que pertenece la observación o la media de los valores de los individuos más cercanos a la observación a imputar dentro del conglomerado. En el caso de estar utilizando el software SAS, como ya indicamos, él tiene un procedimiento asociado a conglomerado no jerárquico que permite hacer la imputación dentro de cada conglomerado.

La otra solución es, tomando en cuenta la escala de medición de la variable cuyos valores perdidos se va a imputar y, construir modelos tales como regresión, regresión estocástica, regresión logit, discriminante y emplear los valores predicho para efectuar la imputación. Finalmente se tiene los métodos de imputación que se basan en el principio de máxima verosimilitud. A continuación se desarrollan brevemente cada uno de los modelos, ilustrando los mismos con ejemplos desarrollado con SAS.

Los procedimientos de SAS que emplearemos para realizar las imputaciones en esta monografía son:

Procedimiento	***Escala***	***Modelo***	***Objetivo***
PROC MEANS	Métrica	Estadística básica	Análisis Previo
PROC FREQ	Métrica y no Métrica	Estadística Básica	Análisis Previo
PROC REG	Métrica	Regresión estándar y estocástica	Imputación
PROC LOGISTIC	No métrica	Regresión Logística	Imputación
PROC DISCRIM	No métrica	Discriminante no paramétrico	Imputación
PROC FASTCLUS	Métrica	Conglomerado no jerárquico	Detección de valores atípico e Imputación
PROGRAMAS SAS	Métricas y no métricas	Manejo de archivos y variables	Determinar valores atípicos e imputar

IVa.-Imputación mediante regresión.

Supongamos que queremos imputar un valor faltante de la variable Y que está correlacionada con las variables $X_1, X_2 ... X_p$ y cuyas observaciones están completas. Consideremos el modelo de regresión:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... \beta_p X_p + \varepsilon$$

Podemos emplear la notación matricial para expresar el modelo de regresión como:

$$Y = X\beta + \varepsilon$$

$$E(Y / X) = X\beta$$

Los supuestos generales del modelo de regresión lineal estándar son**:**

1.- Y es el vector de variables aleatorias observables de n componentes.

2.- X es una matriz de $n.(p+1)$ observaciones de rango completo $(p+1)$**.** Esto significa que las columnas de los datos asociados a las variables son linealmente independientes. En otras palabras, cada una contiene una información no redundante.

3.- β es un vector columna de $(p+1)$ parámetros desconocidos.

4.- $E(\varepsilon / X) = 0_n$ vector columna nulo de n componentes

5. $Var(\varepsilon / X) = \sigma^2 I_{nn}$ (matriz de varianza-covarianza) donde se asume que todas las variables tienen la misma dispersión y, esto se llama homoscedasticidad (homo significa igual y scedasticidad significa variabilidad) y, además no hay covariación entre los ε_i .

6.- El vector ε se distribuye como una normal multivariante $N(0_n, \sigma^2 I_{nn})$**,** asumida esta ley de probabilidad, se puede estimar los parámetros β y σ^2 por el método de máxima verosimilitud y, efectuar los contrastes de hipótesis y la construcción de intervalos.

El estimador de mínimos cuadrados es: $\beta^* = (X^T X)^{-1} X^T Y$ que coincide con el de máxima verosimilitud bajo el supuesto de normalidad esférica de ε .

Para usar la regresión en el problema de imputación debe cumplirse que las variables explicativas deben tener todas sus observaciones sin datos faltantes, se asume además, en el caso de observaciones proveniente de una muestra, que el modelo propuesto cumple con los supuestos generales del modelo de regresión estándar, esto es, ausencia de colinealidad, homoscedásticidad, normalidad etc. En el caso de datos censales sólo interesa la ausencia de colinealidad y se espera que el coeficiente de determinación sea lo más alto posible. Una vez validado el modelo propuesto se procede a imputar cada observación por su correspondiente valor predicho por el modelo. Esto es:

$$Y_i^* = \beta_0 + \beta_1^* X_{i1} + \beta_2^* X_{i2} + ... + \beta_p^* X_{ip}$$

Si se selecciona un modelo de regresión para realizar la imputación debe tenerse presente la importancia que tiene una buena especificación. Esto es, seleccionar las variables explicativas relevantes y tener presente el signo esperados de los parámetros.

Veamos un ejemplo usando SAS.

Ejemplo 1

Para desarrollar este ejemplo tomamos la base que se anexa al final de esta monografía. La variable que se quiere imputar es la variable ingreso y se supone que la misma se puede explicar según la siguiente lista de variables.

La lista de variables es:

Puntos=valoración como cliente

Ingreso=ingreso mensual declarado por el cliente.

Ttrabajo=tiempo trabajando.

Carga=carga familiar

Ahorro=monto promedio mensual en la cuenta de ahorro.

La variable que se quiere imputar es ingreso, por tanto, es la variable respuesta.

Antes de proceder a imputar, debe chequearse que las variables explicativas no tengan valores perdidos, de lo contrario algunos valores serán imputados y otros no. Para ello usaremos el procedimiento means de SAS para conocer cuales variables tienen datos perdidos.

```
PROC MEANS DATA=SASUSER.bancarios3 mean std n nmiss min max;
VAR  puntos ingreso ttrabajo carga ahorro;
 attrib _all_ label='';
RUN;
```

The SAS System

The MEANS Procedure

Variable	Mean	Std Dev	N	N Miss	Minimum	Maximum
PUNTOS	12.6373626	3.5008459	91	0	10.0000000	20.0000000
INGRESO	4916.38	4707.85	87	4	1200.00	30000.00
Ttrabajo	11.4725275	9.7369179	91	0	1.0000000	60.0000000
Carga	1.8901099	1.8040113	91	0	0	6.0000000
AHORRO	34677.80	149629.47	91	0	0	950000.00

En este caso, las variables explicativas tienen todas las observaciones completas, así que se espera poder imputar los cuatro valores perdidos de la variable ingreso. A continuación empleamos el procedimiento REG.

```
PROC REG DATA=SASUSER.bancarios3 corr;
```

```
MODEL ingreso= ttrabajo carga ahorro puntos/p;
OUTPUT OUT=reyes1 P=ingresoaj;
RUN;
DATA SASUSER.bancarios3;
SET reyes1;
MISSING P;
DO;
if ingreso =P then ingreso=ingresoaj2;
end;
run;
```

Los resultados fundamentales de esta corrida se presentan a continuación:

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1342384701	335596175	48.82	<.0001
Error	82	563709434	6874505		
Corrected Total	86	1906094134			

Root MSE	2621.92778	R-Square	0.7043
Dependent Mean	4916.37931	Adj R-Sq	0.6898
Coeff Var	53.33046		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-3787.50486	1238.18236	-3.06	0.0030
Ttrabajo	Ttrabajo	1	32.29553	32.61875	0.99	0.3250
Carga	Carga	1	530.68836	194.97772	2.72	0.0079
AHORRO	AHORRO	1	0.01238	0.00219	5.65	<.0001
PUNTOS	PUNTOS	1	545.07050	105.39217	5.17	<.0001

Con esta información más las relacionadas con la colinealidad y los de los residuos debe validarse el modelo antes de usarlo en la imputación. Este análisis no lo mostraremos por falta de espacio. En este ejemplo, aparte de procedimiento de regresión hay un programa en código de SAS que hace la imputación: Una vez corrido el programa vemos si se disminuyó el número de datos perdidos, para ello aplicamos nuevamente el procedimiento means y obtenemos:

The SAS System

The MEANS Procedure

Variable	Mean	Std Dev	N	N Miss	Minimum	Maximum
PUNTOS	12.6373626	3.5008459	91	0	10.0000000	20.0000000
INGRESO	4903.75	4640.33	91	0	1200.00	30000.00
Ttrabajo	11.4725275	9.7369179	91	0	1.0000000	60.0000000
Carga	1.8901099	1.8040113	91	0	0	6.0000000
AHORRO	34677.80	149629.47	91	0	0	950000.00

De esta forma se resolvió el problema de imputación. Aquí no termina el problema, previa la validación del modelo es necesario estudiar si los valores imputados están dentro del rango admisible.

Cuando el número de observaciones a imputar aumenta, el resultado puede conducir a sobre estimar las correlaciones entre las variables y el coeficiente de determinación.

IVb.-Imputación mediante regresión estocástica.

La regresión estocástica, en el caso de imputación simple, consiste en añadir al valor ajustado por regresión un componente aleatorio. De esta forma se contribuye a disminuir la presencia de sesgo. Concretamente es sumar a las observaciones ajustadas las cantidades obtenidas de la distribución normal de media cero y desviación estándar obtenida en el modelo original de la regresión. Entonces el modelo es:

$$Y_i^* = \beta_0 + \beta_1^* X_{i1} + \beta_2^* X_{i2} + ... + \beta_p^* X_{ip} + \xi$$

$$\xi \triangleright N(0, s)$$

Donde $s = \sqrt{\sum_{i=1}^{n} (Y_i - Y_i^*)^2 / n - p}$

Un problema que puede presentarse es la no existencia de valores ajustado para la variable que se quiere imputar por ausencia de uno o varios valores de las variables explicativas y sin embargo como se le está añadiendo a los valores ajustado los valores de una distribución normal, puede darse que se impute a los valores ausentes con valores que salen del rango de la variable imputada, esto obligaría a truncar la distribución. Esta situación se presenta en el siguiente ejemplo, donde la matriz de datos se amplío en el número de observaciones comparada con el ejemplo anterior.

Ejemplo 2.

Consideremos el modelo anterior pero con una mayor cantidad de observaciones.

En primer lugar emplearemos el procedimiento; PROC MEANS, en el se puede observar que las variables explicativas no tienen todas las observaciones. Si se aplica solamente el procedimiento de regresión no todos los valores perdidos del variable ingreso se pueden imputar. Sin embargo al usar la regresión estocástica se logra ese acometido pero con el problema de que pueden darse valores negativos.

La pregunta es ¿qué ocurre si se censura o trunca la distribución? En cualquiera de los dos casos seleccionados: truncar o censurar se sabe el efecto que tiene tanto en el valor esperado como en la varianza de la distribución.

Analicemos los datos disponibles en este archivo ampliado mediante el procedimiento PROC MEANS.

```
PROC MEANS DATA= sasuser.banca mean std n nmiss min max;
VAR  puntos ingreso ttrabajo carga ahorro;
 attrib _all_ label='';
RUN;
```

The SAS System

The MEANS Procedure

Variable	Mean	Std Dev	N	N Miss	Minimum	Maximum
PUNTOS	13.1583333	3.7461371	120	1	10.0000000	20.0000000
INGRESO	5222.54	5030.95	114	7	700.0000000	30000.00
Ttrabajo	10.9495798	10.0741730	119	2	1.0000000	60.0000000
Carga	1.6637931	1.7641057	116	5	0	6.0000000
AHORRO	45791.08	160219.22	120	1	0	950000.00

Se puede observar que todas las variables explicativas tienen datos perdidos.

Para realizar la imputación emplearemos el procedimiento regresión más un programa que añade al ajuste la variable aleatoria.

En nuestro caso empleamos el siguiente programa en código SAS.

```
PROC REG DATA=SASUSER.banca corr;
MODEL ingreso= ttrabajo carga ahorro puntos/p;
OUTPUT OUT=reyes1 P=ingresoaj;
RUN;
data sasuser.bancarios3;
SET reyes1;
missing p;
DO;
NORMAL=2622*RANNOR(1);
if ingreoaj=p then ingresoaj=0;
else;
if ingreso=p then
ingreso=ingresoaj+normal;
else;
OUTPUT;
END;
PROC PRINT NOOBS;
RUN;
PROC MEANS DATA= sasuser.bancarios3 mean std n nmiss min max;
VAR  puntos ingreso ttrabajo carga ahorro;
 attrib _all_ label='';
RUN;
```

El resultado final usando el procedimiento means es:

The SAS System

The MEANS Procedure

Variable	Mean	Std Dev	N	N Miss	Minimum	Maximum
PUNTOS	13.1583333	3.7461371	120	1	-10.	20.0000000
INGRESO	4965.63	4998.50	121	0	787.9367050	30000.00
Ttrabajo	10.9495798	10.0741730	119	2	1.0000000	60.0000000
Carga	1.6637931	1.7641057	116	5	0	6.0000000
AHORRO	45791.08	160219.22	120	1	0	950000.00

Los dos modelos visto hasta ahora son aplicables sólo en el caso que la variable a imputar es una variable métrica.

En los casos donde las variables a imputar son categóricas u ordinales los modelos más apropiados son: regresión logística y discriminante no paramétrico.

En ambos casos de regresión vistos, los resultados siempre serán mejores cuando la estructura de los datos perdidos es monótona tal como la definimos IIb.1 donde las variables Y hacen el papel de regresores siempre y cuando caben en la especificación.

IVc.-Imputación mediante regresión logística.

Este método está relacionado con el modelo de regresión lineal estándar y es aplicable para el caso de varios grupos sin embargo, sólo veremos el caso de dos grupos por tanto, asumimos que la pertenencia a uno de los dos grupos es una variable aleatoria con distribución de Bernoulli. Consideremos entonces la variable aleatoria Y que toma los valores 0 y 1, con probabilidad p y q respectivamente. Estas probabilidades responden a un número de factores observables que denotamos por el vector **x.** De este supuesto, se deriva la idea que p se puede expresar como una función logística:

$p = \exp x\beta/(1+\exp x\beta)$, como $p+q=1$ entonces, $q = 1/(1+\exp x\beta)$

Por tanto, podemos escribir la probabilidad que pertenezca al primer grupo como:

$P(y=1) = p = \exp x\beta/(1+\exp x\beta)$ y que pertenezca al segundo grupo es $P(y=0) = q = 1/(1+\exp x\beta)$. Para simplificar podemos plantearnos el establecer la relación que existe entre la pertenencia a un grupo o a otro y, esto se logra tomando logaritmo de la división de las dos probabilidades:

$\ln\big(P(y=1)/P(y=0)\big) = x\beta$

De la misma forma que en el modelo de regresión estándar habrá que estimar: el valor ajustado, esto es: el valor Y^* que rara veces dará los valores 0 o 1, pero sí valores comprendidos entre esos dos valores; para ello habrá que estimar previamente el parámetro β y hacer los contrastes correspondientes**.** La estimación del parámetro β se puede realizar mediante el método de máxima verosimilitud.

Igualmente al caso del modelo de regresión, deberá validarse el modelo antes de su aplicación en el problema de imputación en el caso de datos proveniente del muestreo, y en el caso de censo la bondad de ajuste.

Para resolver el problema de imputación usando SAS, hay que tomar en cuenta que él da siempre como ajuste una sola categoría acompañada con la probabilidad correspondiente. El valor de esta probabilidad facilitará el problema de imputación.

Veamos el siguiente ejemplo:

Ejemplo 3

Tomando la misma base de datos, determinemos si hay datos perdidos en la categoría sexo, para ello empleamos el procedimiento: PROC FREQ.

```
PROC FREQ DATA=SASUSER.Banca1;
TABLE nombre sexo estatura_ peso ingreso carga ttrabajo/missprint;
RUN;
```

SEXO				
SEXO	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	9	.	.	.
f	52	46.43	52	46.43
m	60	53.57	112	100.00

Frequency Missing = 9

Se puede observar, que hay nueve observaciones que no tienen asignado el sexo, por tanto, usaremos el PROC LOGISTIC para obtener los valores ajustado.
Las variables explicativas son estatura peso y edad.

Ahora aplicaremos el procedimiento PROC LOGISTIC más el código para hacer la imputación:

```
 proc logistic data=SASUSER.banca;
model SEXO=estatura_  peso EDAD;
OUTPUT OUT=SASUSER.BANCARIOSA  P=SEXOAJ ;
title ´IMPUTACIÓN´;
run;
DATA SASUSER.BANCA1;
SET SASUSER.BANCA;
MISSING P;
DO;
```

```
if  SEXO=P AND SEXOAJ<0.5 then SEXO = 'm';
ELSE;
if  SEXO=P AND SEXOAJ>0.5 then SEXO = 'f';
ELSE;
end;
run;
```

NB:Los resultados del modelo deben verificarse antes de realizar la imputación. Una vez realzada la imputación, verificamos la misma mediante el procedimiento; PROC FREQ.

```
PROC FREQ DATA=SASUSER.Banca1;
TABLE nombre sexo estatura_ peso ingreso carga ttrabajo/missprint;
RUN;
```

SEXO				
SEXO	Frequency	Percent	Cumulative Frequency	Cumulative Percent
f	52	42.98	52	42.98
m	69	57.02	121	100.00

IVd.-Imputación mediante análisis discriminante no paramétrico.

En diferentes situaciones se tiene un conjunto de datos conformado por varias observaciones o individuos, cada una de ella con p mediciones clasificadas de acuerdo a un determinado criterio. Esta clasificación puede deberse a que las observaciones responden a un muestreo estratificado o, son el resultado de aplicar algún algoritmo de conglomerado o simplemente, la clasificación se hizo siguiendo algún criterio especial. Por ejemplo, una institución financiera puede tener sus clientes divididos en tres categorías: clientes con muy bajo riesgo, es decir, alta solvencia, con riesgo moderado o solvencia aceptable y con alto riesgo o insolvencia. En el criterio para hacer esta clasificación posiblemente emplearon varios índices financieros y otras variables. Entonces el problema que se presenta es dado que se tienen varios grupos previamente definidos ¿A cuál grupo debe clasificarse los nuevos clientes que solicitan un crédito?

A este tipo de pregunta responde el análisis discriminante. Hay varios métodos de análisis discriminantes que parten de diferentes supuestos y criterios, pero todos buscan minimizar el error de una mala clasificación. En la práctica se pueden emplear diferentes métodos tal como ocurre en el problema de pronóstico y, seleccionar aquel que tenga menor probabilidad de cometer el error de mala clasificación. Los métodos se clasifican en probabilistico, entre ellos están el de

máxima verosimilitud, bayesiano y, los que parten del concepto de distancia, entre ellos están el análisis discriminante lineal de Fisher y el discriminante cuadrático.

Además se pueden clasificar de acuerdo a los supuestos sobre las distribuciones de las poblaciones como paramétricos y no paramétricos, y también, de acuerdo a la naturaleza de las variables en: análisis discriminante para variables continuas y análisis discriminante para variables discretas.

Para el propósito de esta parte de la monografía aplicaremos el análisis discriminante no paramétrico que se basa en el concepto de núcleo.

Consideremos las poblaciones G_s y dado una nueva observación x, para decidir a cuál de las poblaciones debe asignarse esta nueva observación, construiremos la siguiente regla:

La nueva observación se clasifica en la población G_t si la cantidad

$$\sum_{s=1}^{k} \pi_s f_s(x,\Theta) c(t/s); x \in R^p$$

Se minimiza para t, donde π_s y $f_s(x;\Theta)$ son respectivamente la probabilidad a priori, la función de densidad de la población correspondiente y $c(t/s)$ es el costo de clasificar la observación en la población t cuando proviene de la población s.

Consideremos una muestra aleatoria: $x_i \quad i = 1,2....n_t$ de la población G_t y consideremos una nueva observación x proveniente de la población G_t con función de densidad desconocida $f(x;\Theta)$ estimemos esta función mediante:

$$f_t^*(x;\Theta) = \frac{1}{n_t} \sum_{i=1}^{n_t} K_t(x - x_i); x \in R^p$$

Donde K está definida para el vector z p-dimensional y cumple en primer lugar que es una función no negativa y además:

$$\int_{R^p} K_t(z) dz = 1$$

Esta función se llama núcleo[2] y para cada familia de leyes de probabilidad se define el núcleo tal como en los siguientes casos:

Núcleo uniforme:

[2] El concepto estimaciones de núcleo deriva del concepto de distribución acumulada empírica. Ver Hand, D.J (1981) Discrimination and Clasfication. Jhon Wiley pag:24-31

$K_t(z) = \frac{1}{v_r(t)}$ si $z^t V_t^{-1} z \le r^2$ y cero en otro caso, donde $v_r(t)$ es el volumen del elipsoide: $\{z : z^t V_t^{-1} z \le r^2\}$

Núcleo Normal de $\mu = 0_p$ y varianza $r^2 V$

$K_t(z) = \frac{1}{c_0(t)} \exp(-0{,}5 z^t V_t^{-1} z / r^2)$ si $z^t V_t^{-1} z \le r^2$ y cero en otro caso.

Donde $c_0(t) = (2\pi)^{p/2} r^p |V_t|^{1/2}$

Aparte de estos núcleos hay el núcleo de Epanechnikov, núcleo doblemente ponderado y triplemente ponderado. Estos tienen la siguiente forma:

Núcleo de Epanechnokov:

$K_t(z) = c_1(t)(1 - z^t V_t^{-1} z / r^2)$; si $z^t V_t^{-1} z \le r^2$ y cero en caso contrario.

Donde: $c_1(t) = (1 + p/2) v_r(t)$

Núcleo doblemente ponderado:

$K_t(z) = c_2(t)(1 - z^t V_t^{-1} z / r^2)^2$; si $z^t V_t^{-1} z \le r^2$ y cero en caso contrario.

Donde $c_2(t) = (1 + p/4) c_1(t)$

Núcleo triplemente ponderado:

$K_t(z) = c_3(t)(1 - z^t V_t^{-1} z / r^2)^3$ si $z^t V_t^{-1} z \le r^2$ y cero en caso contrario

Donde $c_3(t) = (1 + p/6) c_2(t)$

V es una métrica previamente seleccionada y r es el radio cuya determinación es más importante que la selección del núcleo que se empleará (Khattree R y Naik D. (2000))

La selección de la métrica puede ser:

$V_t = S$ se considera la matriz de varianza covarianza del total de observaciones

$V_t = diagonalS$ se toma la diagonal de la matriz de varianza covarianza del total de observaciones

$V_t = S_t$ se considera la matriz de varianza covarianza de las observaciones del grupo t

$V_t = diagonalS_t$ se toma la diagonal de la matriz de varianza covarianza de las observaciones del grupo t

$V_t = I$ matriz identidad

La selección del valor de r depende de la experiencia del investigador, un valor alto se obtiene una densidad estimada alsada, con un valor bajo conduce a una estimación “aserrada”.

En todo caso, en el análisis discriminante lo importante es reducir la probabilidad de mala clasificación.

Hay varios métodos para estimar la probabilidad de mala clasificación, una de ellas es emplear las observaciones que se posee en ambos grupos y aplicar el método sobre estas observaciones para determinar la proporción de observaciones mal clasificadas tal como se presenta en el siguiente cuadro:

	Perteneciente a G_1	Perteneciente G_2
Clasificada en G_1	n_{11} (frecuencia de bien clasificadas)	n_{12}(frecuencia de mal clasificadas)
Clasificada en G_2	n_{21}(frecuencia de mal clasificadas)	n_{22} (frecuencia de bien clasificadas)

Por tanto, el estimador de $P(G_i / G_j)$ es $P^*(G_i / G_j) = n_{ij} / n$.

Ahora veremos un ejemplo de la aplicación del discriminante no paramétrico.

Ejemplo 4

Queremos saber, cuántos datos perdidos hay en el archivo que venimos analizando de las variables: carga (carga familiar) y ver si podemos usar como variables discriminante la edad, el ahorro y puntos.

Lo primero es aplicar el procedimiento de SAS:PROC MEANS.

```
proc means data=sasuser.banca1 SUM mean std n nmiss min max;
var carga edad ahorro puntos;
  attrib _all_ label='';
run;
```

The SAS System

The MEANS Procedure

Variable	Sum	Mean	Std Dev	N	N Miss	Minimum	Maximum
Carga	193.0000000	1.6637931	1.7641057	116	5	0	6.0000000
EDAD	3666.00	31.3333333	10.5241922	117	4	18.0000000	80.0000000
AHORRO	5494930.00	45791.08	160219.22	120	1	0	950000.00
PUNTOS	1579.00	13.1583333	3.7461371	120	1	10.0000000	20.0000000

El procedimiento de SAS que emplearemos para aplicar la imputación es el discriminante no paramétrico es PROC DISCRIM.

Una primera corrida con las tres variables seleccionadas no arrojó un resultado deseable porque a igual que el modelo de regresión las observaciones de las variables predictoras, en este caso discriminantes, deben poseer el menor número de valores perdidos, lo ideal es que todas las observaciones en las variables discriminantes estén completa, este no fue el caso, por tanto se seleccionó sólo la variable puntos que está relacionada de forma no lineal con las restantes.

```
proc discrim data =SASUSER.banca1   testdata=SASUSER.banca1
testout=SASUSER.banca1
method=npar kernel=biw r=0.002;
class CARGA;
var   puntos ;
title ´Analisis discriminante´;
run;
data SASUSER.banca2;
set SASUSER.banca1;
missing p;
do;
if carga=p  then carga =_into_3;
else;
end;
run;


proc means data=sasuser.banca2 SUM mean std n nmiss min max;
var puntos carga edad ahorro;
  attrib _all_ label='';
run;
```

´Analisis discriminante´

The MEANS Procedure

Variable	Sum	Mean	Std Dev	N	N Miss	Minimum	Maximum
PUNTOS	1579.00	13.1583333	3.7461371	120	1	10.0000000	20.0000000
Carga	207.0000000	1.7107438	1.7721833	121	0	0	6.0000000
EDAD	3666.00	31.3333333	10.5241922	117	4	18.0000000	80.0000000
AHORRO	5494930.00	45791.08	160219.22	120	1	0	950000.00

IVe.-Imputación por conglomerados.

Consideremos una matriz de datos **X** con p variables y n observaciones, el problema es determinar si es posible agrupar las n observaciones de acuerdo a los valores de las p variables en varios grupos lo más homogéneos posible dentro de ellos y los mas separados posible entre los grupos. Esto se llama formar cluster o conglomerados de individuos. No existe para este problema una única solución, esto es, existe un gran número de algoritmos que fijado el número de conglomerados que se desean, puede ocurrir que al utilizarlos todos den diferentes conglomerados en cuanto a los individuos que los integran. Tampoco existe un criterio que indique cuantos grupos se pueden formar. Por tanto, el número de grupos y el algoritmo a emplear para obtener esos grupos dependen en buena medida de la experiencia que tiene el investigador sobre el problema.

Los algoritmos para definir conglomerados de individuos se dividen en jerárquicos y no jerárquicos.

1.-El no jerárquico consiste en definir previamente k puntos como centro de gravedad o semillas y luego se calculan las distancias de los individuos a estas semillas agrupándolos de acuerdo a su cercanía con los mismos, una vez obtenidos estos conglomerados se recalculan los centro de gravedad y las distancias y se obtienen nuevos conglomerados, el algoritmo se repite hasta tener grupos o conglomerados estables, esto es, ya no es posible redefinir nuevos conglomerados. Este procedimiento se llama de k medias. Un inconveniente de este algoritmo es que la escogencia de las semillas es en cierto sentido arbitrario, de tal forma que dos investigadores pueden llegar a conglomerados muy diferentes en cuanto a los elementos que los integran, una variante del algoritmo anterior es el desarrollado por E. Diday conocido como nube dinámica.

Estos tipos de algoritmos no siempre están disponibles en los software de estadística, entre los que los ofrecen está el software SAS que facilita el uso de varios procedimientos.

2.-El otro grupo de algoritmo se denomina jerárquico, y está basado en definir la distancia entre los puntos tomando en cuenta diferentes criterios, que no son otra cosa que la forma de definir distancias entre conjuntos (distancia de Hausdorff).

Ejemplo 5.

Consideremos ahora del mismo archivo varias variables a imputar: peso, ingreso, edad, años de estudios, carga familiar, tiempo de trabajo, ahorro y puntos de los clientes:

Para el problema de imputación se ha seleccionado un método no jerárquico que está disponible en SAS y permite hacer la imputación. El procedimiento en SAS es PROC FASTCLUS.

Como siempre, haremos un análisis univariante de cada variable de las variables de interés, que en este caso son:

Peso, ingreso, edad, años de estudios, carga familiar, tiempo de trabajo, ahorro y puntos de los clientes:

```
proc means data=sasuser.bancar SUM mean std n nmiss min max;
var peso ingreso edad A_OS_est  carga ttrabajo ahorro puntos;
  attrib _all_ label='';
run;
```

Variable	Sum	Mean	Std Dev	N	N Miss	Minimum	Maximum
PESO	8440.00	72.1367521	10.0240280	117	4	51.0000000	98.0000000
INGRESO	600370.00	5220.61	5008.88	115	6	700.0000000	30000.00
EDAD	3666.00	31.3333333	10.5241922	117	4	18.0000000	80.0000000
A_OS_EST	1761.00	14.7983193	4.9224147	119	2	2.0000000	20.0000000
Carga	201.0000000	1.6750000	1.7497299	120	1	0	6.0000000
Ttrabajo	1320.00	10.9090909	9.9966661	121	0	1.0000000	60.0000000
AHORRO	5494930.00	45791.08	160219.22	120	1	0	950000.00
PUNTOS	1579.00	13.1583333	3.7461371	120	1	10.0000000	20.0000000

Se puede observar en la tabla anterior que casi todas las variables consideradas tienen datos perdidos. Para resolver el problema, empleamos el procedimiento: PROC FASTCLUS:

```
proc fastclus data=sasuser.bancarios3 maxclusters=3 out=banc  outseed=puntos
IMPUTE;
```

```
var  peso ingreso edad A_OS_est  carga ttrabajo ahorro puntos;
run;
```

Para ver el resultado de la imputación aplicamos nuevamente el procedimiento: PROC MEANS.

```
proc means data=banc SUM mean std n nmiss min max;
var peso ingreso edad A_OS_est  carga ttrabajo ahorro puntos;
  attrib _all_ label='';
run;
```

Variable	Sum	Mean	Std Dev	N	N Miss	Minimum	Maximum
PESO	8733.05	72.1739965	9.8654860	121	0	51.0000000	98.0000000
INGRESO	627252.75	5183.91	4884.72	121	0	700.0000000	30000.00
EDAD	3789.06	31.3145708	10.3478041	121	0	18.0000000	80.0000000
A_OS_EST	1790.54	14.7978498	4.8812236	121	0	2.0000000	20.0000000
Carga	202.6140351	1.6744962	1.7424329	121	0	0	6.0000000
Ttrabajo	1320.00	10.9090909	9.9966661	121	0	1.0000000	60.0000000
AHORRO	5506552.19	45508.70	159580.48	121	0	0	950000.00
PUNTOS	1591.80	13.1553574	3.7306392	121	0	10.0000000	20.0000000

Se puede observar en la última tabla, que todas las variables tienen ya todas las observaciones completas y los rangos permanecen iguales a la tabla original.

Cabe hacer algunas observaciones sobre el uso de conglomerados en la imputación:

1. El uso del análisis de conglomerados puede resultar útil antes de aplicar algún método de imputación porque él permite estudiar si hay valores atípicos que se agrupan en uno o varios conglomerados.
2. La selección del método y el número de conglomerados dependerá del conocimiento que se tenga de las características de los datos.
3. La facilidad que da SAS con el procedimiento factclus de imputar los datos faltantes para cada conglomerado tiene como limitación que usa un método no jerárquico para agrupar observaciones con variables numéricas.
4. Una vez determinado el método para obtener los conglomerados, se puede aplicar alguno de los métodos de imputación a cada conglomerado, esto siempre y cuando el número de observaciones por conglomerado sea lo suficientemente grande.

5. Debemos tener presente que al definir conglomerados para emplearlo en la imputación esta variabilidad se descompone como sigue:

$T = \sum_{i=1}^{n} \left\| X_i - \mu^* \right\|^2$ Variabilidad total de la masa de puntos.

$W_k = \sum_{i}^{nk} \left\| X_{ik} - \mu_k^* \right\|^2$ Variabilidad de la masa de datos dentro del grupo g.

$E = \sum_{j=1}^{g} \left\| \mu_j^* - \mu^* \right\|^2$ Variabilidad de la masa de puntos entre los g grupo.

El criterio es maximizar E que es equivalente a minimizar $\sum_k W_k$.

V.-Estimación de Máxima Verosimilitud con el algoritmo EM.

Este algoritmo fue desarrollado por los autores: Dempster. A.P., N.M. Lair and D.B. Rubin (1977) para imputar datos faltantes. Partimos que se tiene una muestra aleatoria de p variables n observaciones y queremos estimar el parámetro: $\theta \in \Theta$, recordemos que la matriz de datos puede descomponerse como: $X = (X_{obs} \left| X_{per} \right.)$. La probabilidad conjunta de X se puede expresar como sigue: $P(X_{obs}, X_{per} / \theta) = P(X_{obs} / \theta) P(X_{per} / \theta)$.

Tomando logaritmo y despejando obtenemos:

$$\log P(X_{obs}, / \theta) = \log P(X_{obs} X_{per} / \theta) - \log P(X_{per} / \theta)$$

El miembro derecho $\log P(X_{obs}, / \theta)$ es la función soporte de los datos observados, la función de verosimilitud construida con ella: $L(\theta \left| X_{obs} \right.)$ y su maximización nos conduce al estimador máximo verosímil del parámetro $\theta \in \Theta$, $\log P(X_{obs} X_{per} / \theta)$ es la función soporte si hubiésemos observado los datos completos y con ella podemos obtener la función de verosimilitud: $L(\theta \left| X_{obs}, X_{per} \right.)$ y finalmente $\log P(X_{per} / \theta)$ nos facilita la probabilidad de los datos ausentes conocida la muestra y el parámetro.

$$L(\theta \left| X_{obs} \right.) = L((\theta \left| X_{obs}, X_{per} \right.) - \log P(X_{per} / \theta)$$

El algoritmo EM emplea la función $L(\theta \left| X_{obs}, X_{per} \right.)$ para obtener el estimador de $\theta \in \Theta$. Este algoritmo se hace en dos pasos:

1. El primer paso cosiste, partiendo de un valor inicial como estimador, obtener el valor esperado de las funciones de los valores perdidos que aparecen en la función: $L(\theta|X_{obs}, X_{per})$, con respecto a la distribución de X_{per} dados los datos observados: X_{obs} y el valor inicial $\hat{\theta}^{(t)}$. Consideremos que el resultado de este primer paso que se denomina paso E se puede expresar como: $L^*(\theta / X_{obs}) = E_{X_{per}\ \hat{\theta}^{(t)}}(L(\theta / X_{obs}, X_{per})$. Con este resultado se sustituyen los valores perdidos por alguna función $h(X_{per})$ dado $\hat{\theta}^{(t)}$. Ahora tendremos la función de verosimilitud $L(\theta|X_{obs}, X_{ajust})$ donde los valores de X_{per} han sido sustituidos por los valores ajustados: X_{ajust}.

2. El segundo paso consiste en maximizar la función: $L(\theta|X_{obs}, X_{ajust})$ y obtener el estimador MV $\hat{\theta}^{(t+1)}$ y con este estimador se regresa al paso 1.

El algoritmo termina cuando dos estimadores sucesivos difieran en una cantidad muy pequeña.

El desarrollo teórico de este algoritmo parte de que la ley de probabilidad pertenece a la familia de distribuciones del tipo exponencial[3] donde en el caso de distribuciones de variables continuas un ejemplo es la distribución normal.

El algoritmo EM se emplea también para realzar la imputación múltiple, en el caso de SAS éste está incluido en el procedimiento MI que veremos en su oportunidad.

Aparte de esta propuesta existe la de Li-Chun Zhang (Likelihood Imputation. Statistics Norway. Oslo pag 147-152). Esta propuesta parte de considerar la estructura latente de la imputación. Una estructura latente es aquella donde la observación *x* es observada mediante un estadístico $y = Y(x)$, entonces, x se

---

[3] Consideremos la familia $\{f_X(x;\theta); \theta \in \Theta\}$ donde $\Theta = \{\theta; \gamma < \theta < \delta\}$ si:
$f_X(x;\Theta) = \exp[p(\theta)K(\theta) + S(x) + q(\theta)]; a < x < b$ entonces se dice que la f.d es miembro de la clase exponencial.
Si además cumple con:
1.- $p(\theta)$ es una función continua no trivial de $\theta, \gamma < \theta < \delta$
2.- $dK(\theta)/d\theta \neq 0$ y $S(x)$ una función continua de x en $a < x^* < b$

Entonces es un caso regular de la clase exponencial para f.d continua. En el caso discreto se logra cambiando de los valores de $x = a_1, a_2 .....$ y $K(x)$ es una función no trivial de x en el conjunto $\{x : x = a_1, a_2 .....\}$

dice que es latente e y su manifestación. Una imputación x^* se llama latente si satisface $y = Y(x^*)$ y además, verosímilmente consistente con su manifestación si la estructura latente es ignorable, esto es x^* conduce a la misma inferencia que y. Habrá que definir una medida de discrepancia entre el estimador obtenido con x^* y el obtenido con y, y encontrar un valor x^* de imputación que minimice este error. La medida que propone Li-Chun Zhang es la verosimilitud residual asociada al parámetro θ que se define como:

$$L_r(\theta\ , x^* / y) = L(\theta, x^*) / L_1(\theta; y)$$

Este método que no está incluido en la versión 9.0 de SAS e igual que el EM, es aplicable al caso de imputación múltiple.

VI.-VENTAJAS Y DESVENTAJAS DE LA IMPUTACIÓN SIMPLE.

Las ventajas y desventajas de la imputación simple se pueden resumir como sigue:

1. Seleccionado y aplicado un método adecuado de imputación, se dispondrá de información para hacer comparaciones entre las variables tales como el cálculo de los coeficientes de correlación, empleo de modelos lineales, análisis multivariante de datos entre otros.

2. Cuando se emplea regresión como método de imputación puede presentarse sesgo en la varianza y el coeficiente de determinación. El método se mejora si se definen conglomerados y, a cada conglomerado se aplica la imputación por regresión o cualquier otro método, dependiendo de la escala de medición de la variable con datos faltantes que hay imputar (si todos los conglomerados tienen suficientes observaciones).

3. Por otra parte, la imputación simple tiene la desventaja que no reduce la incertidumbre de cual valor es el más adecuado para emplearlo como imputación.

4. En el caso de usar como modelo de imputación: la regresión lineal, logística o discriminante, los valores de las variables exógenas deben estar completa para cada observación que tiene un valor perdido en la variable respuesta.

5. En el caso de datos censales se requiere que el soporte computacional tenga una buena capacidad de memoria de procesamiento. El software

SAS indica cual debe ser para cada caso. En principio no hay limitaciones en cuanto al número de variables ni de observaciones.

VII. IMPUTACIÓN MÚLTIPLE.

La imputación múltiple consiste en imputar a cada valor perdido m veces con valores obtenidos mediante métodos de simulación. Esto quiere decir, que se simularan tantas muestras aleatorias de tamaño m como valores perdidos existan en la matriz de datos. Por tanto, al operar de esta forma se tendrán m matrices con las observaciones completas.

El procedimiento que emplearemos del software SAS es el PROC MI. Este procedimiento parte de la hipótesis de que la distribución de los datos faltantes cumple con el supuesto de MAR, esto es:

$$P(R/X,\xi) = P(R/X_{obs})$$

Este procedimiento tiene los siguientes métodos[4]:

Estructura	***Escala de la variable a imputar***	***Método***
Monótono	continua	1. Regresión 2. puntuación de la propensión
Monótona	Ordinal	Regresión logística
Monótona	Nominal	Discriminante
Arbitraria		MCMC

En general, se puede resumir los pasos de la imputación múltiple como sigue:

1. Estudiar la estructura de los datos perdidos, si es monótona o arbitraria y además, considerar la escala de medición de las variables con datos faltantes.

[4] SAS versión 9.0

2. Seleccionar el método de imputación, por ejemplo regresión o MCMC de acuerdo al resultado del punto dos.

3. Generar aleatoriamente un número determinado de valores que se usaran en la imputación considerando la distribución de los datos perdidos dados los observados y la distribución a priori de los parámetros que se desea estimar y así generar un pool por ejemplo de matrices de varianzas covarianzas, o de vectores de parámetros de un modelo de regresión o bien un pool de vectores de madias.

4. Realizar la imputación de los datos perdidos. Se efectuarán m imputaciones por cada valor perdido usando un resultado de la selección aleatoria del pool generado en el punto anterior.

5. Una vez realizada la imputación, cada matriz de datos obtenida se analiza desde el punto de vista estadístico, tal como el uso de modelos de dependencia o causales o bien modelos de interdependencia.

6. En este último paso se integran las matrices obtenidas para realzar el análisis conjunto.

VIIa- Efecto del análisis conjunto en la inferencia.

La imputación múltiple tiene consecuencia en la variabilidad de los estimadores de los parámetros cuando se realiza el análisis conjunto. Una variabilidad dentro de cada matriz de datos y una variabilidad entre las matrices resultante de las m imputaciones. Consideremos el hecho que estamos interesados en estimar $\theta \in \Theta$ y sea θ^* el estimador de ese parámetro con estimador de varianza: $\text{var}^*(\theta^*)$. En el caso de imputación múltiple por cada imputación tendremos m estimadores, es decir $\{\theta_1^*, \theta_2^* ... \theta_m^*\}$ y con estimadores de las varianzas $\{\text{var}^*(\theta_1^*), \text{var}^*(\theta_2^*) ... va^* r(\theta_m^*)\}$. El estimador de $\theta \in \Theta$ considerando las m imputaciones es:

$$\theta_m^* = \sum_{J=1}^{m} \theta_j^* / m$$

Y la varianza total es:

$$Var_T^*(\theta^*) = \sum_{j=1}^{m} \text{var}^*(\theta_j^*) / m + (1 + \frac{1}{m}) B_m$$

Donde $B_m = \frac{1}{m-1}\sum_{j=1}^{m}\left(\theta_j^* - \theta_m^*\right)^2$

Entonces, la variación total se descompone en la variación dentro de las imputaciones representada por el primer sumando y la variación entre las imputaciones dadas por el último sumando. El primer sumando lo denotamos por U^*, esto es:

$$U^* = \sum_{j=1}^{m} Var^*(\theta^*)/m$$

Rubin(1988) propone un estadístico que permite hacer inferencia estadística a cerca del parámetro $\theta \in \Theta$ con el conjunto integrado de imputaciones. Este estadístico es:

$$(\theta - \theta^*)Var_T^*(\theta^*)^{-\frac{1}{2}}$$

Este estadístico tiene una distribución de referencia aproximada a una distribución t con $v = (m-1)(1+\left[(1+m^{-1})B_m/U^*\right]^{-1})^2$ grados de libertad. El Procedimiento MI de SAS usa una corrección de los grados de libertad propuesta por el mismo Rubin en una publicación posterior.

La razón: $r = B_m/U$ es un estimador de la cantidad poblacional de $(1-\gamma)/\gamma$ donde γ es la fracción de pérdida de información de $\theta \in \Theta$ por la presencia de datos perdidos.(Rubin 1988). Cuando m es grande y r pequeño entonces, aumenta los grados de libertad de la distribución t, y el estadístico tiende a una distribución normal. El SAS usa una pequeña modificación de r, su valor es: $r = (1+m^{-1})B_m/U$, y considera este valor como el incremento relativo en la varianza debido a la no repuesta, además emplea en su salida otro estadístico alternativo dado por:

$$\lambda^* = (r + 2/(v+3))/(r+1)$$

Este último estadístico mide la información perdida del parámetro $\theta \in \Theta$. Ambos estadísticos dan cuenta de la incertidumbre debida a datos perdidos en las observaciones de las variables.

Un indicador adicional está dado como:

$$RE = (1+\lambda/m)^{-1}$$

Esto mide la eficiencia relativa de usar un número finito de imputaciones. Con estos tres estadísticos es posible comparar los resultados de diferentes métodos de imputación. Entre dos imputaciones se seleccionará aquella que tenga un valor menor de r.

Con la información anterior se puede construir el intervalo de confianza para $\theta \in \Theta$ y hacer contraste de hipótesis. En la salida de SAS se encuentra ambos resultados.

Ahora, siguiendo el trabajo de Rubin (Rubin 1988), consideremos que $\theta \in \Theta$ es un vector de parámetros con k componentes, entonces $\{\theta_1^*, \theta_2^* ... \theta_m^*\}$ son m vectores de estimadores consecuencias de las imputaciones y $\{\text{var}^*(\theta_1^*), \text{var}^*(\theta_2^*) ... va^* r(\theta_m^*)\}$ son m matrices de varianzas covarianzas. Igualmente $Var_T^*(\theta^*)$ es una matriz de varianza covarianza. El estadístico en este caso tiene la forma:

$$(\theta - \theta^*) Var_T^*(\theta^*)^{-1} (\theta - \theta^*)^T$$

Este estadístico tiene una distribución F con k y v grados de libertad, en donde v ahora es:

$$v = (m-1)(1+\left[(1+m^{-1})B_m U^{*-1}\right]^{-1})^2$$

Y la razón r es:

$$r = traza(B_m U^{*-1})/k \text{ .}$$

VIIb.-IMPUTACIÓN EN ESTRUCTURA ARBITRARIA.

Como vimos al inicio de este punto, en el caso que los datos perdidos tengan una estructura monótona, SAS proporciona varios métodos en el procedimiento MI y, en el caso de estructura aleatoria o arbitraria MI emplea el algoritmo MCMC. Veremos en primer lugar este último algoritmo, pero antes daremos algunos conceptos previos.

Cadena de Markov.

Consideremos el proceso $\{X(t), t \in T\}$ donde T es el conjunto índice que puede ser

continuo $t \geq 0$ o discreto $t = 0,1,2...$ Al conjunto de valores posibles del proceso se llama espacio de estado, si este es un número finito o infinito numerable, entonces el espacio de estado es discreto en caso contrario continuo.

Un proceso estocástico es un proceso de Markov si:

$$P\{X(t_n) \leq x_n / X(t_1) \leq x_1 .....X(t_{n-1}) \leq x_{n-1}\} = P\{X(t_n) \leq x_n / X(t_{n-1}) \leq x_{n-1}\}$$

Esto es: el futuro sólo depende del presente y no de los valores pasados. Si el espacio de estado es discreto, entonces el proceso se llama cadena de Markov con parámetro discreto o continuo[5]. Para el propósito de esta monografía su notación es MC.

Una cadena de Markov tiene una probabilidad de transición estacionaria si la:

$$P\{X(t) \leq x_n / X(t_0) \leq x_{n-1}\}$$

depende de t y t_0 vía la diferencia: $|t - t_0|$

Monte Carlo.

Es un método que permite generar muestras aleatorias de una determinada distribución partiendo de números seudoaleatorios comprendidos en el inérvalo $[0,1]$.

MCMC.

El es un algoritmo que permite general muestras aleatorias virtuales de una distribución de probabilidad $F_X(x;;\Theta)$ mediante el método de Monte Carlo. Con esto se origina una cadena de Markov con probabilidad de transición estacionaria. Un problema asociado de generar por simulación una cadena de Markov es la selección de la mejor probabilidad de transición y esto puede resolverse empleando el algoritmo MCMC. Este método permite generar muestras de la distribución a posteriori, cuando el cálculo de la misma es dificultoso.

[5] Parzen E(1972)
Procesos Estocásticos Cap. 6. Pág. 221
Paraninfo-Madrid

Se asume que $F_X(x;;\Theta)$ pertenece a una familia de distribuciones, y en el caso de imputación se asume que se trata una distribución normal multivariante.

Partiremos del concepto de distribución condicional de un parámetro θ utilizando el teorema de Bayes:

$$P(\theta / X) = P(\theta)P(X / \theta) / \int P(\theta)P(X / \theta)d\theta .$$

Ahora consideremos que los datos perdidos siguen un mecanismo MAR y escribimos:

$$P(X_{per} / X_{obs}) = \int P(X_{per} / X_{obs}, \theta)P(\theta / X_{obs})d\theta$$

En general $P(\theta / X_{obs})$ es intratable desde el punto de vista computacional pero cuando los datos perdidos pueden simularse tomando en cuenta los datos observados se puede implementar el algoritmo MCMC para obtener: $P(\theta / X_{obs}, X_{per})$, esto es, la distribución a posteriori.

El algoritmo se ejecuta en dos etapas.

1. Esta primera etapa empieza con el estimador del vector de media y de la matriz de covarianza, en el caso de una normal multivariante. Anotemos en general al valor inicial del parámetro por θ^n. Generamos independientemente las primeras imputaciones X_{per}^{n+1} partiendo de $P(X_{per} / X_{obs}, \theta^n)$.

2. En esta etapa se obtiene un nuevo estimador θ^{n+1} de $P(\theta / X_{obs}, X_{per}^{n+1})$, con este nuevo estimador se regresa a la primera etapa y así sucesivamente hasta obtener una cadana de Markow:

   $$\left\{(\theta^n, X_{per}^{n})n = 1,2...\right\}$$

Esta cadena converge en probabilidad a $P(X_{per}, \theta / X_{obs})$.

Se puede estudiar la convergencia del MCMC por varios procedimientos, uno de ellos es usar la función de autocorrelación de las medias, varianzas o covarianzas:

$$r_k = \sum_{k=1}^{n-k}(z_k - z_k^*)(z_{k+t} - z_{k+t}^*) / \sqrt{\sum_{k=1}^{n}(z_k - z_k^*)^2 \sum_{k=1}^{n}(z_{k+t} - z_{k+t}^*)^2}$$

Y mediante su representación gráfica ver la convergencia requerida.

Para el caso de la distribución a priori el procedimento MI tiene varias opciones, una es asumir que las probabilidades son equiprobables, otra es que son proporcionales al tamaño de la muestra en el grupo, otra es jeffreys que por defecto toma el valor 0,5 y finalmente rige, que por defecto toma el valor uno.

Ejemplo 6.

Emplearemos la base de datos de siempre e imputaremos las variables: peso, edad y estatura empleando el algoritmo MCMC. En este ejemplo el algoritmo se inicia con el algoritmo EM. Se representará en este ejemplo solamente la función de autocorrelación de las medias obtenidas de la variable peso, pero igual se puede hacer con sus varianzas.

```
proc means data=sasuser.bancariosa SUM mean std n nmiss min max;

Var  peso edad estatura_ ;
run;
```

The SAS System
The MEANS Procedure

Variable	Label	Sum	Mean	Std Dev	N	N Miss	Minimum	Maximum
PESO	PESO	8175.00	72.3451327	9.8468359	113	8	51.0000000	98.0000000
EDAD	EDAD	3666.00	31.3333333	10.5241922	117	4	18.0000000	80.0000000
ESTATURA_	ESTATURA	198.5800000	1.7118966	0.0819959	116	5	1.5400000	1.8900000

```
proc mi data=sasuser.bancariosa seed=21355417 nimpute=6 mu0=70 30 1.70
out=letty ;
    mcmc chain=multiple displayinit initial=em(itprint) acfplot(mean(peso));
    var peso edad estatura_;
  run;
```

The SAS System
The MI Procedure

Model Information	
Data Set	SASUSER.BANCARIOSA
Method	MCMC
Multiple Imputation Chain	Multiple Chains

Model Information	
Initial Estimates for MCMC	EM Posterior Mode
Start	Starting Value
Prior	Jeffreys
Number of Imputations	6
Number of Burn-in Iterations	200
Seed for random number generator	21355417

El cuadro anterior muestra las característica del procedimiento MI; el método de imputación en este caso es MCMC que utiliza el algoritmo EM para obtener los valores iniciales de los estimadores. Indica además que hace seis imputaciones (m=6), el número de iteraciones (200) y el valor asignado a la semilla para empezar la simulación (21355417). Cuando no se especifica el número de imputaciones entonces hace cinco imputaciones por defecto.

El siguiente cuadro da el patrón de los datos perdidos. Se puede observar que el patrón es arbitrario (según la terminología de SAS) o aleatorio. Además, muestra seis grupos, en el primero todas las observaciones tienen sus datos completos y en las tres variables consideradas y representa el 88,43%, junto a esta información están las medias de las tres variables: peso, edad y estatura y así sucesivamente. El segundo grupo hay tres observaciones que tienen los datos completos en peso y edad pero faltan en estatura, por tanto tienen el cálculo de las medias de estas dos primeras variables y no así de la otra, representan el 2,48% de las observaciones.Y así con los demás grupos.

Missing Data Patterns								
						Group Means		
Group	PESO	EDAD	ESTATURA_	Freq	Percent	PESO	EDAD	ESTATURA_
1	X	X	X	107	88.43	72.224299	31.672897	1.711402
2	X	X	.	3	2.48	79.333333	34.666667	.
3	X	.	X	3	2.48	69.666667	.	1.713333
4	.	X	X	5	4.13	.	24.600000	1.728000
5	.	X	.	2	1.65	.	25.000000	.

Missing Data Patterns								
						Group Means		
Group	PESO	EDAD	ESTATURA_	Freq	Percent	PESO	EDAD	ESTATURA_
6	.	.	X	1	0.83	.	.	1.680000

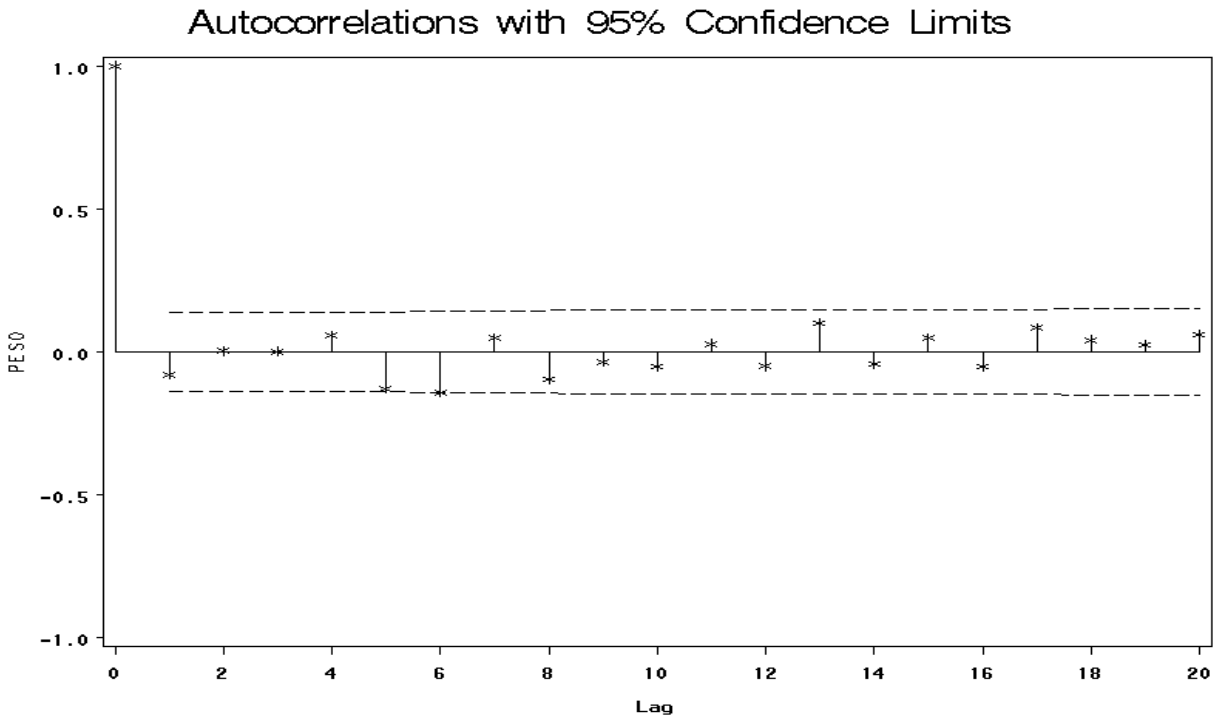
Las salidas siguientes indican el resultado de las cinco iteraciones que efectúa SAS para obtener las estimaciones del vector de media y la matriz de varianza covarianza bajo el supuesto de una distribución normal multivriante. Al final se da el vector de media y la matriz de varianza covarianza que se empleará en MCMC para hacer las imputaciones. Empieza indicando los valores iniciales partiendo del algoritmo EM que servirán para la aplicación de MCMC.

EM (Posterior Mode) Iteration History					
Iteration	-2 Log L	-2 Log Posterior	PESO	EDAD	ESTATURA_
0	808.524725	824.745971	72.264541	31.310627	1.712312
1	808.709716	824.540678	72.264529	31.310626	1.712312
2	808.727199	824.540161	72.264527	31.310626	1.712312
3	808.728142	824.540158	72.264474	31.310590	1.712312
4	808.728199	824.540158	72.264464	31.310586	1.712312

EM (Posterior Mode) Estimates				
TYPE	_NAME_	PESO	EDAD	ESTATURA_
MEAN		72.264464	31.310586	1.712312
COV	PESO	92.554540	27.014294	0.255195
COV	EDAD	27.014294	106.109014	0.092030
COV	ESTATURA_	0.255195	0.092030	0.006437

Initial Parameter Estimates for MCMC				
TYPE	_NAME_	PESO	EDAD	ESTATURA_
MEAN		72.264464	31.310586	1.712312
COV	PESO	92.554540	27.014294	0.255195
COV	EDAD	27.014294	106.109014	0.092030
COV	ESTATURA_	0.255195	0.092030	0.006437

El siguiente gráfico corresponde a a la función de autocorrelación de la sucesión de medias de la variable peso. Ella nos permite estudiar la convergencia del proceso generado. Así como se hace con las medias de esta variable se puede hacer con el resto de las variables que se están imputando. Además se pueden realizar gráficos de la función de autocorrelación de las varianzas y covarianzas.



En los cuadros siguientes se tiene la información de la descomposición de la varianza en: dentro de las imputaciones y, entre las imputaciones. Por ejemplo para la variable peso, la varianza dentro de las imputaciones:

$U^* = \sum_{j=1}^{m} Var^*(\theta^*)/m$ es 0,79 y entre las imputaciones: $B_m = \frac{1}{m-1}\sum_{j=1}^{m}\left(\theta_j^* - \theta_m^*\right)^2$ es: 0,059 por otra parte, da los valores del incremento relativo en la varianza: $r = \left(1+m^{-1}\right)B_m/U$, para la variable peso es:0,087337 . Además se tiene la información para construir los intervalos de confianza al 95% y efectuar los contrastes de hipótesis, con sus respectivos valores p.

En este ejemplo se especificó como hipótesis nula:

$$H_o = (\mu_1, \mu_2, \mu_3) = (70{,}30{,}1.70)$$

Al nivel de signiicación 0,05 solamente es significativa la media de la variable peso.

Multiple Imputation Variance Information

Variable	Variance			DF	Relative Increase in Variance	Fraction Missing Information
	Between	Within	Total			
PESO	0.059838	0.799338	0.869149	95.227	0.087337	0.082686
EDAD	0.063713	0.931349	1.005681	97.658	0.079811	0.075929
ESTATURA_	0.000003769	0.000055842	0.000060240	98.004	0.078741	0.074963

Multiple Imputation Parameter Estimates

Variable	Mean	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Mu0	t for H0: Mean=Mu0	Pr > |t|
PESO	72.30	0.9322	70.44999	74.15150	95.227	71.90	72.587320	70.00	2.47	0.0154
EDAD	31.24	1.0028	29.25305	33.23342	97.658	30.86	31.528727	30.00	1.24	0.2180
ESTATURA_	1.71	0.0077	1.69626	1.72707	98.004	1.70	1.714165	1.700	1.50	0.1361

Ahora aplicamos el procedimiento means para obtener las estimaciones con los datos completos:

```
proc means data=letty SUM mean std n nmiss min max;

Var  peso edad estatura_ ;
run;
```

The SAS System

The MEANS Procedure

Variable	Label	Sum	Mean	Std Dev	N	N Miss	Minimum	Maximum
PESO	PESO	52490.34	72.3007459	9.8032029	726	0	51.0000000	98.0000000
EDAD	EDAD	22682.59	31.2432375	10.5815480	726	0	4.4211762	80.0000000
ESTATURA_	ESTATURA	1242.67	1.7116630	0.0819358	726	0	1.4059014	1.8900000

Este último resultado ayudará en el análisis de los datos, para ello se compara los estadísticos originales con los obtenidos después de la imputación. Usando la sentencia BY y como variable: _imputation_ se obtiene los diferentes estadísticos de interés por cada imputación como veremos más adelante.

Ejemplo 7.

Consideremos los datos del ejemplo anterior pero usando otra modalidad del procedimiento MI y tomando en cuenta las variables peso edad e ingreso. En este ejemplo veremos además la posibilidad de estudiar cada imputación por separado.

```
proc means data=sasuser.banca1 SUM mean std n nmiss min max;
Var  peso edad ingreso ;
run;
```

The SAS System

The MEANS Procedure

able	Label	Sum	Mean	Std Dev	N	N Miss	Minimum	Maxi
O	PESO	8175.	72.345132	9.84683	113	8	51.00	98.000
D	EDAD	3666.	31.333333	10.52419	117	4	18.00	80.000
RESO	INGRESO	595370	5222.54	5030.95	114	7	700.00	3000

```
proc mi data= SASUSER.BANCARIOS3 seed=899603 out=oute1 mu0=70 30
1.70;
Var  peso edad ingreso puntos ;
run;
```

En el programa se asume la imputación múltiple con MCMC por defecto, el modelo posterior se obtiene mediante el algoritmo EM, con la opción out=oute1, guarda el

nuevo archivo. Como no se fijó el número de imputaciones, por defecto hace cinco. La interpretación de la salida es similar al ejemplo anterior.

The SAS System
The MI Procedure

Model Information	
Data Set	SASUSER.BANCARIOS3
Method	MCMC
Multiple Imputation Chain	Single Chain
Initial Estimates for MCMC	EM Posterior Mode
Start	Starting Value
Prior	Jeffreys
Number of Imputations	5
Number of Burn-in Iterations	200
Number of Iterations	100
Seed for random number generator	899603

Missing Data Patterns										
							Group Means			
Group	PESO	EDAD	INGRESO	PUNTOS	Freq	Percent	PESO	EDAD	INGRESO	PUNTOS
1	X	X	X	X	110	90.91	72.418182	31.754545	4988.934568	13.154545
2	X	.	X	X	3	2.48	69.666667	.	1514.410988	11.333333
3	.	X	X	X	6	4.96	.	24.666667	7652.552679	14.666667
4	.	X	X	.	1	0.83	.	25.000000	299.758080	.
5	.	.	X	X	1	0.83	.	.	1300.000000	10.000000

EM (Posterior Mode) Estimates					
TYPE	_NAME_	PESO	EDAD	INGRESO	PUNTOS
MEAN		72.218585	31.236063	4965.629005	13.135474
COV	PESO	91.966344	26.908499	1353.220909	2.095881

EM (Posterior Mode) Estimates					
TYPE	_NAME_	PESO	EDAD	INGRESO	PUNTOS
COV	EDAD	26.908499	105.091839	15682	10.535722
COV	INGRESO	1353.220909	15682	23795227	13738
COV	PUNTOS	2.095881	10.535722	13738	13.357001

Multiple Imputation Variance Information						
Variable	Variance			DF	Relative Increase in Variance	Fraction Missing Information
	Between	Within	Total			
PESO	0.060445	0.80260	0.87514	91.29	0.09037	0.086017
EDAD	0.049559	0.91259	0.97206	100.4	0.06516	0.062932
PUNTOS	0.0000830	0.115927	0.116027	117.94	0.000859	0.000859

Multiple Imputation Parameter Estimates										
Variable	Mean	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Mu0	t for H0: Mean=Mu0	Pr > |t|
PESO	72.148	0.93549	70.29042	74.0067	91.291	71.87702	72.470346	70.0000	2.30	0.0239
EDAD	31.377	0.98593	29.42193	33.3338	100.41	31.07764	31.667139	30.00000	1.40	0.1653
PUNTOS	13.129	0.34062	12.45468	13.8037	117.94	13.11698	13.139143	0	38.54	<.0001

Ahora veremos el efecto de la imputación múltiple.

```
proc means data=oute1 SUM mean std n nmiss min max;
Var  peso edad ingreso puntos;
run;
```

The SAS System

The MEANS Procedure

Variable	Label	Sum	Mean	Std Dev	N	N Miss	Minimum	Maximum
PESO	PESO	43649.8	72.1485711	9.8244939	605	0	49.9144704	98.0000000
EDAD	EDAD	18983.6	31.3778969	10.4753105	605	0	8.5049288	80.0000000
INGRESO	INGRESO	3004205.5	4965.63	4981.92	605	0	-787.9367050	30000.00
PUNTOS	PUNTOS	7943.18	13.1292169	3.7328786	605	0	8.1556927	20.0000000

Para analizar mejor los resultados por separado, pediremos las mismas medidas estadísticas por cada imputación, para ello aplicamos el siguiente código en SAS.

```
proc means data=oute1 SUM mean std n nmiss min max;
 by _imputation_;
Var  peso edad ingreso puntos;
run;
```

En los cuadros siguientes se presentan las estimaciones obtenidas por cada una de las imputaciones. Con esto se ilustra que se puede realizar un análisis por separado de cada una de ellas. Por ejemplo especificar un modelo de regresión y comparar los resultados para cada imputación.

The SAS System

The MEANS Procedure

Imputation Number=1

Variable	Label	Sum	Mean	Std Dev	N	N Miss	Minimum	Maximum
PESO	PESO	8703.04	71.9259521	9.9764338	121	0	51.0000000	98.0000000
EDAD	EDAD	3788.96	31.3137549	10.3979505	121	0	18.0000000	80.0000000
INGRESO	INGRESO	600841.11	4965.63	4998.50	121	0	-787.9367050	30000.00
PUNTOS	PUNTOS	1588.00	13.1239258	3.7496462	121	0	8.9950210	20.0000000

Imputation Number=2

Variable	Label	Sum	Mean	Std Dev	N	N Miss	Minimum	Maximum
PESO	PESO	8768.91	72.4703462	9.7924335	121	0	51.0000000	98.0000000
EDAD	EDAD	3812.59	31.5089772	10.4926862	121	0	18.0000000	80.0000000
INGRESO	INGRESO	600841.11	4965.63	4998.50	121	0	-787.9367050	30000.00
PUNTOS	PUNTOS	1589.54	13.1366981	3.7380791	121	0	10.0000000	20.0000000

Imputation Number=3

Variable	Label	Sum	Mean	Std Dev	N	N Miss	Minimum	Maximum
PESO	PESO	8741.21	72.2414383	9.9482309	121	0	49.9144704	98.0000000
EDAD	EDAD	3831.72	31.6671386	10.5327092	121	0	18.0000000	80.0000000
INGRESO	INGRESO	600841.11	4965.63	4998.50	121	0	-787.9367050	30000.00
PUNTOS	PUNTOS	1587.16	13.1169892	3.7581148	121	0	8.1556927	20.0000000

Imputation Number=4								
Variable	Label	Sum	Mean	Std Dev	N	N Miss	Minimum	Maximum
PESO	PESO	8739.60	72.2280896	9.6873156	121	0	51.0000000	98.0000000
EDAD	EDAD	3789.96	31.3219648	10.4886777	121	0	16.3367523	80.0000000
INGRESO	INGRESO	600841.11	4965.63	4998.50	121	0	-787.9367050	30000.00
PUNTOS	PUNTOS	1589.84	13.1391431	3.7364632	121	0	10.0000000	20.0000000

Imputation Number=5								
Variable	Label	Sum	Mean	Std Dev	N	N Miss	Minimum	Maximum
PESO	PESO	8697.12	71.8770292	9.8663363	121	0	50.1406527	98.0000000
EDAD	EDAD	3760.40	31.0776487	10.6279980	121	0	8.5049288	80.0000000
INGRESO	INGRESO	600841.11	4965.63	4998.50	121	0	-787.9367050	30000.00
PUNTOS	PUNTOS	1588.65	13.1293284	3.7441144	121	0	9.6487349	20.0000000

Si queremos un modelo de regresión por cada imputación donde puntos (importancia del cliente) esté en función de la edad y el ingreso, aplicamos el siguiente código:

```
proc reg data=oute1 ;
 by _imputation_;
Var  puntos edad ingreso ;
model puntos=edad ingreso;
run;
```

Por razones, de espacio sólo mostraremos algunos resultados de la imputación uno y de la cinco.

The SAS System

The REG Procedure

Model: MODEL1

Dependent Variable: PUNTOS PUNTOS

Imputation Number=1					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1010.48878	505.24439	88.10	<.0001
Error	118	676.69281	5.73468		
Corrected Total	120	1687.18159			

Root MSE	2.39472	**R-Square**	0.5989
Dependent Mean	13.12393	**Adj R-Sq**	0.5921
Coeff Var	18.24698		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	9.75147	0.69351	14.06	<.0001
EDAD	EDAD	1	0.01762	0.02213	0.80	0.4276
INGRESO	INGRESO	1	0.00056807	0.00004603	12.34	<.0001

The SAS System

The REG Procedure

Model: MODEL1

Dependent Variable: PUNTOS PUNTOS

Imputation Number=5

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1005.49130	502.74565	87.66	<.0001
Error	118	676.71586	5.73488		
Corrected Total	120	1682.20715			

Root MSE	2.39476	**R-Square**	0.5977
Dependent Mean	13.12933	**Adj R-Sq**	0.5909
Coeff Var	18.23978		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	9.88699	0.67536	14.64	<.0001
EDAD	EDAD	1	0.01338	0.02176	0.62	0.5397
INGRESO	INGRESO	1	0.00056920	0.00004627	12.30	<.0001

VIIc.-IMPUTACIÓN EN ESTRUCTURA MONÓTONA.

En el caso de estructura monótona, en la matriz X_{per} de datos perdido, el número de observaciones presentes en cada variable va disminuyendo, si denotamos por n_j el número de observaciones de la variable j-ésima, la estructura monótona es: $n_1 \geq n_2 \geq n_3 \geq ...n_q$.

Este tipo de estructura tiene asociadas varias formas para imputar datos perdidos, entre estos métodos están: mcmc, método de puntuación de la propensión (propensity score method), método de modelo predictivo, regresión, regresión logística, análisis discriminante. El método seleccionado depende de la escala de medición de la variable a imputar.

Vic1.-MCMC.

El algoritmo MCMC puede emplearse igualmente cuando se asume que la estructura de los datos perdidos es monótona. Veamos un ejemplo:

Ejemplo 8..

En este ejemplo se trabaja con las variables estatura, edad, peso y carga.

```
proc mi data= SASUSER.BANCA1 seed=1305417 out=censo;
 mcmc impute=monotone;
Var estatura_ edad peso carga;
run;

proc means data=censo SUM mean std n nmiss min max;

Var  peso edad estatura_ ;
run;
```

En el siguiente cuadro se indica el método de imputación de forma similar que en los ejemplos anteriores.

The SAS System
The MI Procedure

Model Information	
Data Set	SASUSER.BANCA1
Method	Monotone-data MCMC
Multiple Imputation Chain	Single Chain
Initial Estimates for MCMC	EM Posterior Mode
Start	Starting Value
Prior	Jeffreys
Number of Imputations	5
Number of Burn-in Iterations	200
Number of Iterations	100
Seed for random number generator	1305417

La interpretación de este cuadro es similar a la realizada en ejemplos anteriores.

Missing Data Patterns										
							Group Means			
Group	ESTATURA_	EDAD	PESO	Carga	Freq	Percent	ESTATURA_	EDAD	PESO	Carga
1	X	X	X	X	103	85.12	1.709515	31.572816	71.883495	1.757282
2	X	X	X	O	4	3.31	1.760000	34.250000	81.000000	.
3	X	X	.	X	5	4.13	1.728000	24.600000	.	1.400000
4	X	.	X	X	2	1.65	1.725000	.	64.500000	1.500000
5	X	.	X	O	1	0.83	1.690000	.	80.000000	.
6	X	.	.	X	1	0.83	1.680000	.	.	1.000000
7	.	X	X	X	3	2.48	.	34.666667	79.333333	0.333333
8	.	X	.	X	2	1.65	.	25.000000	.	0

EM (Posterior Mode) Estimates					
TYPE	_NAME_	ESTATURA_	EDAD	PESO	Carga
MEAN		1.712673	31.292709	72.284935	1.658813
COV	ESTATURA_	0.006388	0.095087	0.253138	-0.016100
COV	EDAD	0.095087	105.094039	26.367238	5.217372
COV	PESO	0.253138	26.367238	91.558947	-1.181805
COV	Carga	-0.016100	5.217372	-1.181805	2.948626

Al aplicar el prodimientos means de SAS sobre el conjunto de datos formado por las matrices obtenemos lo siguiente:

The MEANS Procedure

Variable	Label	Sum	Mean	Std Dev	N	N Miss	Minimum	Maximum
PESO	PESO	43713.85	72.2542948	9.8893394	605	0	47.7538847	98.0000000
EDAD	EDAD	18875.24	31.1987495	10.4549990	605	0	14.1303912	80.0000000
ESTATURA_	ESTATURA	1035.83	1.7121191	0.0814816	605	0	1.5400000	1.8900000

VIIc2. Método de regresión.

Este método se basa, como en el caso de imputación simple de un modelo de regresión, pero con la diferencia fundamental es que partiendo del modelo ajustado se obtiene nuevos modelos por simulación y estos últimos se emplean para realzar cada una de las m imputaciónes. Entonces, consideremos el modelo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

La estimación de los parámetros es: $\beta^* = (\beta_0{}^*, \beta^*{}_1, \beta^*{}_2 ... \beta^*{}_k)$ y la estimación de la matriz de varianza covarianza es: $\sigma^{*2}(X_{obs}^T X_{obs})^{-1}$ (se asume que los regresores tienen todas sus observaciones, esto es: $n_j = n_{jobs} \forall j : j = 1,2...k$) donde: $\sigma^{*2} = (Y_{obs} - X_{obs}\beta^*)^T (Y_{obs} - X_{obs}\beta^*)/(n_{obs} - k)$

Como se asume que $\varepsilon \approx N(0, \sigma^2 I_{nn})$ entonces $\beta^* \approx N(\beta, \sigma^2 (X_{obs}^T X_{obs})^{-1})$. La distribución condicional de cada parámetro es:

1.- $\sigma^2 / X \approx (Y_{obs} - X_{obs}\beta^*)^T (Y_{obs} - X_{obs}\beta^*) / \mathrm{X}^2{}_{nobs-k}$ g.l

2.- $\beta / \sigma^2, X \approx N(\beta^*, \sigma^2 (X_{obs}^T X_{obs})^{-1})$

Ahora para cada imputación $j = 1,2...m$, efectuamos los siguientes pasos:

1.-Obtenemos por simulación el vector de parámetros de los regresores y la varianza σ_j^2 en base a las distribuciones de los mismos:

$$\sigma_{+j}^2 = (Y_{obs} - X_{obs}\beta^*)^T (Y_{obs} - X_{obs}\beta^*)(n_j - k - 1) / \mathrm{X}_{n_j - k - 1}^2$$

$$\beta_+ = (\beta_{+0}, \beta_{+1}, \beta_{+2}...\beta_{+k}) = (\beta_0^*, \beta^*_1, \beta^*_2...\beta^*_k) + V_{hj}^T Z$$

Donde V_{hj} es una matriz tal que $(X^T X)_j^{-1} = V_{hj}^T V_{hj}$ y Z es un vector de k+1 componentes de variables aleatorias normales independientes.

2.-Imputamos los valores perdidos mediante el modelo:

$$Y_j^* = \beta_{+0} + \beta_{+1} X_1 + \beta_{+2} X_2 + ... + \beta_{+k} X_k + z\sigma_{+j}$$

Donde z se distribuye como una normal: N(0,1).

Ejemplo 9.

En este ejemplo usaremos otro archivo cuyos datos perdidos en algunas de sus variables tiene una estructura monótona. En este caso imputaremos por regresión la variable estudio tomando como variable regresora el ingreso.

```
proc means data=SASUSER.monotono1 SUM mean std n nmiss min max;

Var  estudios;
run;
```

Podemos notar que la variable ingreso tiene seis observaciones perdidas.

The SAS System
The MEANS Procedure

Analysis Variable : estudios estudios						
Sum	Mean	Std Dev	N	N Miss	Minimum	Maximum
926.0000000	12.3466667	4.0486679	75	6	4.0000000	19.0000000

El código en SAS en este caso es:

```
proc mi data=SASUSER.monotono1 round=.1
seed=13951639 out=andres mu0=14;
```

```
monotone reg(ingreso/ details)
regpmm(estudios= ingreso/ details);
Var   ingreso estudios;
run;
```

The SAS System
The MI Procedure

Model Information	
Data Set	SASUSER.MONOTONO1
Method	Monotone
Number of Imputations	5
Seed for random number generator	13951639

Monotone Model Specification	
Method	Imputed Variables
Regression(PMM)	estudios

Missing Data Patterns						
					Group Means	
Group	ingreso	estudios	Freq	Percent	ingreso	estudios
1	X	X	75	92.59	9329.333333	12.346667
2	X	.	6	7.41	8500.000000	.

Los resultados de la regresión se presentan a continuación: En primer lugar el valor de los parámetros con los datos observados y de ellos deriva los parámetros correspondientes a las cinco imputaciones. Debe tenerse presente que las estimaciones de los parámetros obtenidas por simulación no tienen porque coincidir con las estimaciones de los parámetros que se obtienen una vez imputados los datos.

Regression Models for Monotone Predicted Mean Matching Method							
Imputed Variable	Effect	Obs Data	Imputation				
			1	2	3	4	5
estudios	Intercept	-0.00634	-0.061122	-0.038597	0.089822	0.113304	-0.174762
estudios	ingreso	0.58036	0.710866	0.722847	0.630699	0.572589	0.482806

De la misma forma que en los ejemplos anteriores se dan los valores de la varianza entre y dentro de las imputaciones y la total.

Multiple Imputation Variance Information						
Variable	Variance			DF	Relative Increase in Variance	Fraction Missing Information
	Between	Within	Total			
estudios	0.001448	0.192838	0.194576	77.256	0.009010	0.008969

Multiple Imputation Parameter Estimates										
Variable	Mean	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Mu0	t for H0: Mean=Mu0	Pr > |t|
estudios	12.308	0.4411	11.430	13.186	77.256	12.2592	12.345679	14.0000	-3.83	0.0003

Para ver el resultado de la imputación aplicamos el procedimiento de siempre**:**

```
proc means data=andres SUM mean std n nmiss min max;
Var  estudios;
run;
```

The SAS System

The MEANS Procedure

Analysis Variable : estudios estudios						
Sum	Mean	Std Dev	N	N Miss	Minimum	Maximum
4985.00	12.3086420	3.9327326	405	0	4.0000000	19.0000000

VIIc3. Método de puntuación de la propensión.

Consideremos la matriz ya vista R en donde $r_{ij} = 0$ representa la observación perdida que puede ser imputada mediante el modelo logístico. Ahora consideremos la observación x_{ij} y las observaciones $x_{i1}, x_{i2}...x_{ij-1}$; se llama puntuación de la propensión a:

$$s_{ij} = P(r_{ij} = 0 / x_{i1}, x_{i2}...x_{ij-1})$$

Para resolver el problema de la imputación formulamos el siguiente modelo:

1. $\ln(s_{ij} /(1-s_{ij})) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_{j-1} X_{j-1}$

Una vez obtenidos los valores de los estimadores de los parámetros, obtenemos los valores ajustados de las puntuaciones de propensión

2. $\hat{s}_{ij} = \exp(\beta^*_0 + \beta^*_1 X_1 + \beta^*_2 X_2 + ... + \beta^*_{j-1} X_{j-1}) /(1 + \exp(\beta^*_0 + \beta^*_1 X_1 + \beta^*_2 X_2 + ... + \beta^*_{j-1} X_{j-1}))$

Luego todas las observaciones son asignadas dentro de g estratos de acuerdo a los cuantiles de las puntuaciones de propensión ajustadas. Dentro cada estrato se crea un pool de "donadores" aplicando el Bootstrap Bayesiano[6], esto consiste en que en cada estrato se obtiene una muestra aleatoria con reemplazo de los valores observados, de esta muestra se selecciona aleatoriamente una cantidad igual a los valores perdidos para ser asignados a los mismos. Esto se repite m veces, así resulta que cada observación perdida se le ha asignado m valores aleatoriamente.

Es importante considerar que el modelo relaciona el indicador de valores perdidos r_{ij} con $X_1, X_2, ..X_{j-1}$ y no los valores perdidos de X_j con los valores de las variables $X_1, X_2, ..X_{j-1}$. Por tanto, aunque el modelo anterior puede ser un

[6] Bootstrap Bayesiano. Cada muestra bootstrat se selecciona aleatoriamente con reemplazamiento siguiendo los siguientes pasos:

1. Se selecciona n números aleatorios de una distribución uniforme definida en el intervalo $[0,1]$ y se ordenan de menor a mayor $a_o, a_1, a_2...a_n$ y tal que $a_o = 0$ y $a_n = 1$
2. De la muestra bootstrat $X^b = \{x_1^b, x_2^b...x_n^b\}$ se selecciona una muestra de tamaño n con probabilidades: $a_1 - a_0, a_2 - a_1, ...a_n - a_{n-1}$, esto es , se selecciona n veces un número uniforme u y se tona a x_i si $a_{i-1} < u \leq a_i$

buen ajuste entre r_{ij} y $X_1, X_2, ..X_{j-1}$ no necesariamente lo es para X_j y $X_1, X_2, ..X_{j-1}$.

Ejemplo 10.

Para este ejemplo emplearemos el archive “monótono1” Las variables a imputar son estudios (años de escolaridad) y edad.

```
proc mi data=sasuser.monotono1 seed=899603 simple out=ccc2 mu0=30 14 28;
    monotone propensity;
    var ingreso estudios edad;
  run;
```

Las tablas siguientes dan cuenta del método que se está empleando, el número de imputaciones, la semilla y la estructura de los datos perdidos.

Model Information	
Data Set	SASUSER.MONOTONO1
Method	Monotone
Number of Imputations	5
Seed for random number generator	899603

Monotone Model Specification	
Method	Imputed Variables
Propensity(Groups= 5)	estudios edad

Missing Data Patterns								
						Group Means		
Group	ingreso	estudios	edad	Freq	Percent	ingreso	estudios	edad
1	X	X	X	71	87.65	9322.535211	12.338028	26.704225
2	X	X	.	4	4.94	9450.000000	12.500000	.
3	X	.	.	6	7.41	8500.000000	.	.

En este ejemplo se ha empleado la opción “simple” para obtener directamente los estadísticos básicos y la matriz de correlación de los datos originales (antes de la imputación).

Univariate Statistics							
						Missing Values	
Variable	N	Mean	Std Dev	Minimum	Maximum	Count	Percent
ingreso	81	9268	5624	1200	23000	0	0.00
estudios	75	12.34667	4.04867	4.00000	19.00000	6	7.41
edad	71	26.70423	4.42523	21.00000	38.00000	10	12.35

Pairwise Correlations			
	ingreso	estudios	edad
ingreso	1.000000000	0.579918531	-0.012100314
estudios	0.579918531	1.000000000	0.129163708
edad	-0.012100314	0.129163708	1.000000000

Las siguientes tablas tienen la misma interpretación que los ejemplos anteriores.

Multiple Imputation Variance Information						
	Variance					
Variable	Between	Within	Total	DF	Relative Increase in Variance	Fraction Missing Information
estudios	0.013900	0.205424	0.222105	65.536	0.081200	0.077699
edad	0.020698	0.243595	0.268433	61.519	0.101963	0.096388

Multiple Imputation Parameter Estimates										
Variable	Mean	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Mu0	t for H0: Mean=Mu0	Pr > |t|
estudios	12.261	0.47128	11.32066	13.202	65.536	12.098765	12.395062	14.00	-3.69	0.0005
edad	26.711	0.51810	25.67527	27.746	61.519	26.555556	26.901235	28.00	-2.49	0.0156

Vic4.-Regresión logística.

La regresión logística se emplea para imputar datos categóricos. Consideremos el caso: $p_{ij} = P(r_{ij} = 1/ x_{i1}, x_{i2}...x_{ij-1})$ es la probabilidad de que r_{ij} tome el valor 1 dadas las observaciones de las variables $X_1, X_2,..X_k$ El modelo logit es:

$$\log it(p_j) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

Una vez obtenidos los estimadores $\beta^* = (\beta_0{}^*, \beta^*{}_1, \beta^*{}_2...\beta^*{}_k)$ y la matriz de varianza covarianza $\sigma^{*2}(X_{obs}^T X_{obs})^{-1}$ obtenemos para cada imputación $j = 1,2...m$ tal como en el caso de regresión lo siguiente:

1. $\beta_+ = (\beta_{+0}, \beta_{+1}, \beta_{+2}...\beta_{+k}) = (\beta_0{}^*, \beta^*{}_1, \beta^*{}_2...\beta^*{}_k) + V_{hj}^T Z$

Estimamos la probabilidad de que la variable a imputar $Y_j = 1,$ tome el valor 1, esto es: $Y_j = 1,$ mediante la relación conocida

2. $p_j = \exp(\mu_j)/(1+\exp(\mu_j))$

La imputación es para cada j:

3. $Y_j^* = \beta_{+0} + \beta_{+1} X_1 + \beta_{+2} X_2 + ... + \beta_{+k} X_k$.

Ejemplo 11.

Consideremos el archivo que estamos usando con los últimos ejemplos y estudiaremos tres variables: sexo, tipo de vehículo y tipo de vivienda.

The SAS System

The FREQ Procedure

sexo				
sexo	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	2	.	.	.
f	41	51.90	41	51.90
m	38	48.10	79	100.00

Frequency Missing = 2

vehiculo				
vehiculo	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	2	.	.	.
0	17	21.52	17	21.52
1	44	55.70	61	77.22
2	15	18.99	76	96.20
3	3	3.80	79	100.00

Frequency Missing = 2

casa				
casa	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	3	.	.	.
1	31	39.74	31	39.74
2	20	25.64	51	65.38
3	12	15.38	63	80.77
4	15	19.23	78	100.00

Frequency Missing = 3

Imputaremos la variable sexo que es una variable medida en escala nominal y además, los datos perdidos tienen un patrón monótono. La imputación se efectúa mediante el modelo de regresión logístico donde la variable exógena es el ingreso

```
proc mi data=sasuser.monotono1 seed=1305417 out=andres4;
    class sexo;
    monotone logistic( sexo= ingreso / details);
    var ingreso sexo ;
  run;
```

The SAS System

The MI Procedure

Model Information	
Data Set	SASUSER.MONOTONO1
Method	Monotone
Number of Imputations	5
Seed for random number generator	1305417

Monotone Model Specification	
Method	Imputed Variables
Logistic Regression	sexo

Missing Data Patterns					
					Group Means
Group	ingreso	sexo	Freq	Percent	ingreso
1	X	X	79	97.53	9435.443038
2	X	.	2	2.47	2650.000000

Logistic Models for Monotone Method							
			Imputation				
Imputed Variable	Effect	Obs-Data	1	2	3	4	5
sexo	Intercept	0.08679	-0.028048	0.132174	0.453390	-0.105369	0.515954
sexo	ingreso	-0.41945	-0.668273	-0.750187	-0.394864	-0.189273	-0.415463

El resultado de la imputación es:

The SAS System
The FREQ Procedure

sexo				
sexo	Frequency	Percent	Cumulative Frequency	Cumulative Percent
f	213	52.59	213	52.59
m	192	47.41	405	100.00

VIc5.-Discriminante.

La imputación mediante discriminante se emplea para datos medidos en escala nominal con dos o más categorías y para ello se emplea la sentencia CLASS en el código SAS. Se asume que dentro de cada grupo formado por las categorías de tamaño n_i las variables: $X_1, X_2, ..X_k$ tienen aproximadamente una distribución normal multivariante y con matriz de varianza covarianzas iguales. Para aplicar el algoritmo de imputación múltiple se genera un poul de matrices de varianza covarianza para cada grupo S_i y los respectivos vectores de medias μ_i^*, de sus distribuciones posteriores dadas por:

$$\Sigma / X \approx W^{-1}(n-G,(n-1)S)$$

Donde W^{-1} es la función inversa de la distribución de Wishart[7].

$$\mu_i \ /(\Sigma, \mu_i^*) \approx N(\mu_i^*, \frac{1}{n_i}\Sigma)$$

[7] La distribución de Wishart es la generalización de la distribución univariante gamma y se define como: Dada la matriz V de orden pxp simétrica y definida positiva. La matriz aleatoria V se dice que sigue una distribución p dimensional no singula de Wishart si su función de densidad es:

$f_V(v;\Theta) = \frac{c|V|^{(n-p-1)/2}}{|\Sigma|^{n/2}} \exp(\frac{1}{2} tr\Sigma^{-1/2}V)$ donde $V > 0, \Sigma > 0$ y:

$c = \left[2^{np/2} \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma\left(\frac{n+1+j}{2} \right) \right]^{-1}$ (James Press S.(1972) Applied Multivariate Analysis Cap 5 pag 100 y siguientes. HRW New York)

La probabilidad a priori de que un individuo pertenezca a algún grupo puede ser igual o proporcional. Estas son las dos opciones más frecuentes, sin embargo SAS añade a estas dos opciones las siguientes: JEFFREYS y RIDGE.

Los pasos de la imputación son los siguientes:

1. Seleccionar aleatoriamente una matriz S del pool de matrices obtenidas de la distribución $\Sigma / X \approx W^{-1}(n-G,(n-1)S)$.
2. Seleccionar para cada grupo un vector de media μ_i^* y la matriz de varianza covarianza S_i
3. Definir la probabilidad a priori.
4. Con los vectores de medias y matrices de covarianza obtenido en el punto 2 obtener:

$$p_i(x) = \exp(-d^2(x,G_i)) / \sum(1+\exp(-d^2(x,G_i))$$

$$d^2(x;G_i) = (x-\mu_i^*)^T S_i^{-1}(x-\mu_i^*) - 2\log P(G_i)$$

5. Realizar la imputación de acuerdo a la siguiente regla, se selecciona un número aleatoriamente de una distribución uniforme $[0,1]$ la observación a imputar se le asigna uno si $p_i(x)$ es menor que este valor y dos si es igual o mayor pero menor que $p_1(x)+p_2(x)$ y así sucesivamente.

Ejemplo 12:

Repetiremos el caso expuesto en el ejemplo anterior pero usando discriminante.

```
proc mi data= sasuser.monotono1 seed=6755407 nimpute=6 out=pp5;
    class sexo;
    monotone reg(ingreso)
          discrim( sexo= ingreso  / details);
    var ingreso sexo;
  run;
```

The SAS System

The MI Procedure

Model Information

Model Information	
Data Set	SASUSER.MONOTONO1
Method	Monotone
Number of Imputations	6
Seed for random number generator	6755407

Monotone Model Specification	
Method	Imputed Variables
Discriminant Function	sexo

Missing Data Patterns					
					Group Means
Group	ingreso	sexo	Freq	Percent	ingreso
1	X	X	79	97.53	9435.443038
2	X	.	2	2.47	2650.000000

En este cuadro se presenta las simulaciones de los estimadores de las funciones discriminantes.

Group Means for Monotone Discriminant Method								
			Imputation					
sexo	Variable	Obs-Data	1	2	3	4	5	6
f	ingreso	-0.15995	0.074226	-0.430940	-0.049204	-0.142344	0.022796	0.008625
m	ingreso	0.23451	0.274778	0.639525	0.320342	0.474172	0.035468	0.209158

VII. Análisis conjunto.

Para hacer el análisis conjunto, se emplea el procedimiento MIANALIZE y, se ha debido seleccionar previamente algunos de los métodos de imputación múltiple con el procedimiento MI, guardando los resultados en un archivo temporal o permanente. Con este archivo se efectúa un análisis de cada imputación para estudiar por separado cada una de ellas. Luego se usa el procedimiento MIANALIZE

Ejemplo 13.

Con el siguiente ejemplo se ilustra los pasos para hacer el análisis conjunto. En primer lugar se selecciona las variables y el método de imputación que se empleará. Las varables a imputar son: estatura, peso, edad, ingreso y puntos. El método de imputación es MCMC bajo una estuctura monótona. Las matrices obtenidas se guardan en un archivo temporal o permanente. Sobre este archivo se definen los procedimientos de SAS que de acuerdo a los intereses del investigador seleccione. En este ejemplo se usó el procedimiento PROC UNIVARIATE por cada imputación por separado. Finalmente se empléo el procedimiento MIANALIZE cuyos resulados comentaremos más adelante.

```
**********************************;
*paso 1;
proc mi data=sasuser.bancarios3  seed=17655417 out=letty;
mcmc impute=monotone;
var estatura_  peso edad ingreso puntos;
run;
*paso 2;


proc univariate data=LETTY ;
     var estatura_ PESO  edad ingreso puntos;
     output out=LETTY1 mean=estatura_ PESO  edad ingreso puntos
                    stderr=Sestatura_  SPESO Sedad Singreso Spuntos;
     by _Imputation_;
   run;
*paso 3;
proc mianalyze data=LETTY1  theta0= 1.70 78 34;

 var  estatura_ PESO  edad ingreso puntos;
stderr Sestatura_  SPESO Sedad Singreso Spuntos;
run;
```

El siguiente cuadro muestra el método de imputación empleado: MCMC para estructura monótona, usando como valores iniciales de estimación los aportados por el algoritmo EM.

Model Information	
Data Set	SASUSER.BANCARIOS3
Method	Monotone-data MCMC
Multiple Imputation Chain	Single Chain
Initial Estimates for MCMC	EM Posterior Mode
Start	Starting Value
Prior	Jeffreys
Number of Imputations	5
Number of Burn-in Iterations	200
Number of Iterations	100
Seed for random number generator	17655417

El cuadro indica la estructura de los datos perdidos divididos en siete grupos. El primer grupo tiene las bservaciones completas y representa el 88,43%, por tanto tiene las medias de todas las variables. El segundo grupo tiene obsevaciones que le faltan datos en la variable edad, por tanto se presenta las medias de todas las variables menos en la variable edad y así sucesivamente.

Missing Data Patterns													
								Group Means					
Group	ESTATURA_	PESO	EDAD	INGRESO	PUNTOS	Freq	Percent	ESTATURA_	PESO	EDAD	INGRESO	PUNTOS
1	x	x	X	X	X	107	88.43	1.7114	72.2242	31.672897	5016.661	13.14
2	x	x	.	X	X	3	2.48	1.7133	69.6666	.	1514.410	11.33
3	x	.	X	X	X	5	4.13	1.7280	.	24.600000	7983.063	14.40
4	x	.	.	X	X	1	0.83	1.680000	.	.	1300.000	10.00
5	.	x	X	X	X	3	2.48	.	79.333333	34.666667	4000.000	13.33
6	.	.	X	X	X	1	0.83	.	.	25.000000	6000.000	16.00
7	.	.	X	X	O	1	0.83	.	.	25.000000	299.758	.

El siguiente cuadro muestra el vector de media y la matriz de varianza covarianza obtenido por el algoritmo EM que sirve de base para obtener el correspondiente pool y realizar las imputaciones.

EM (Posterior Mode) Estimates						
TYPE	_NAME_	ESTATURA_	PESO	EDAD	INGRESO	PUNTOS
MEAN		1.712511	72.258382	31.230341	4965.629005	13.135467
COV	ESTATURA_	0.006335	0.252567	0.092845	-39.018795	-0.007208
COV	PESO	0.252567	91.039765	26.461928	1298.504102	1.957934
COV	EDAD	0.092845	26.461928	104.226848	15593	10.499021
COV	INGRESO	-39.018795	1298.504102	15593	23607863	13629
COV	PUNTOS	-0.007208	1.957934	10.499021	13629	13.251515

Una vez realzadas las imputacines se aplica el procedimiento UNIVARIATE para obtener diferentes estadísticos por cada variable en cada imputación de lo que sólo se muestra una pequeña parte de la salida completa

The UNIVARIATE Procedure
Variable: ESTATURA_ (ESTATURA)
Imputation Number=1

Moments			
N	121	Sum Weights	121
Mean	1.71140912	Sum Observations	207.080504
Std Deviation	0.08112245	Variance	0.00658085
Skewness	-0.187646	Kurtosis	-0.2957201
Uncorrected SS	355.189166	Corrected SS	0.78970225
Coeff Variation	4.74009693	Std Error Mean	0.00737477

Basic Statistical Measures			
Location		Variability	
Mean	1.711409	Std Deviation	0.08112
Median	1.710000	Variance	0.00658
Mode	1.720000	Range	0.35000
		Interquartile Range	0.12000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	232.0628	Pr > |t|	<.0001
Sign	M	60.5	Pr >= |M|	<.0001
Signed Rank	S	3690.5	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	1.89

Quantiles (Definition 5)	
Quantile	**Estimate**
99%	1.89
95%	1.84
90%	1.80
75% Q3	1.79
50% Median	1.71
25% Q1	1.67
10%	1.57
5%	1.56
1%	1.54
0% Min	1.54

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
1.54	102	1.84	63
1.54	62	1.84	117
1.54	42	1.84	119
1.56	101	1.89	5
1.56	98	1.89	118

Finalmente se analiza los resultados de las cinco imputaciones como sigue: La primera información relevante es la descomposición de la varianza total de cada varable en variación debida a la existente entre imputaciones y la debida a la variación dentro de las imputaciones. La otra información al incremento relativo y la fración de información perdida de cada uno de los parámetros

The MIANALYZE Procedure

Model Information	
Data Set	WORK.LETTY1
Number of Imputations	5

Multiple Imputation Variance Information

Parameter	Variance			DF	Relative Increase in Variance	Fraction Missing Information
	Between	Within	Total			
estatura_	0.000001175	0.000056262	0.000057672	6697.5	0.025051	0.024730
PESO	0.014042	0.798875	0.815726	9373.5	0.021093	0.020867
edad	0.005422	0.910534	0.917040	79476	0.007145	0.007119
ingreso	0	206488	206488	.	0	.
puntos	0	0.116946	0.116946	.	0	.

Luego se obtiene las estimaciones de los parámetros con error estándar, los límtes de confianza al 95%, los valores máximos y mínimos y los contrastes de hipótesis.

Multiple Imputation Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
estatura_	1.712153	0.007594	1.69727	1.72704	6697.5	1.710940	1.713293	1.700000	1.60	0.1096
PESO	72.429647	0.903176	70.65923	74.20007	9373.5	72.352295	72.624238	78.000000	-6.17	<.0001
edad	31.350578	0.957622	29.47364	33.22751	79476	31.267681	31.468388	34.000000	-2.77	0.0057
ingreso	4965.629005	206488	.	.	.	4965.629005	4965.629005	0	.	.
puntos	13.158333	0.116946	.	.	.	13.158333	13.158333	0	.	.

Ejemplo 14.

Este ejemplo es simlar al anterior, lo que varía es que en vez de analizar las estimaciones de los parámetros básicos, se analiza los parámetros de un modelo

de regresión tal cual como se especifica en la sentencia model del procedimento REG. $peso = \beta_0 + \beta_1 estatura + \beta_2 edad + \varepsilon$

```
 proc mi data=sasuser.bancarios3  seed=17655417 out=letty;
mcmc impute=monotone;
var estatura_  peso edad ingreso puntos;
run;
proc reg data=letty outest=outreg covout noprint;
     model peso=estatura_  edad;
     by _Imputation_;
  run;

  proc print data=outreg(obs=8);
    var _Imputation_ _Type_ _Name_
       Intercept estatura_  edad;
  run;
 proc mianalyze data=outreg edf=28;
    modeleffects Intercept estatura_  edad;
  run;
```

A continuación se muestra sólo dos conjuntos de estimaciones de los parámetros: vectores de medias y de sus matrices de varianza y covarianza del total de cinco que se obtiene del programa al fijar cinco imputaciones por defecto.

REG Model Coefficients and Covariance matrices

Obs	_Imputation_	_TYPE_	_NAME_	Intercept	ESTATURA_	EDAD
1	1	PARMS		10.436	32.468	0.20265
2	1	COV	Intercept	318.987	-185.496	-0.02608
3	1	COV	ESTATURA_	-185.496	110.302	-0.10448
4	1	COV	EDAD	-0.026	-0.104	0.00654
5	2	PARMS		2.980	36.640	0.21251
6	2	COV	Intercept	270.539	-157.154	-0.03224
7	2	COV	ESTATURA_	-157.154	93.537	-0.09159
8	2	COV	EDAD	-0.032	-0.092	0.00600

REG Model Coefficients and Covariance matrices

The MIANALYZE Procedure

Model Information	
Data Set	WORK.OUTREG
Number of Imputations	5

Multiple Imputation Variance Information						
Parameter	Variance			DF	Relative Increase in Variance	Fraction Missing Information
	Between	Within	Total			
Intercept	28.688674	299.017899	333.444309	22.105	0.115132	0.107986
estatura_	8.993077	103.215453	114.007145	22.518	0.104555	0.098687
edad	0.000388	0.006371	0.006837	23.735	0.073156	0.070327

El cuadro siguiente presenta las estimaciones de los parámetros del modelo de regresión con sus correspondientes errores estándar, los límites de confianza etc.

Multiple Imputation Parameter Estimates										
Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
Intercept	3.425	18.2604	-34.434	41.284	22.105	-4.493107	10.4356	0	0.19	0.853
estatura_	36.665	10.6774	14.551	58.779	22.518	32.467568	40.8932	0	3.43	0.002
edad	0.198	0.0826	0.027	0.369	23.735	0.171915	0.2200	0	2.40	0.024

El modelo obtenido es $peso = 3.4253 + 36.6658 estatura + 0{,}1985 edad$. El contraste de hipótesis para cada parámetro es: $H_0 : \beta_j = 0$ y $H_1 : \beta_j \neq 0$.

Con el procedimiento MYANALIZE pueden plantearse test más generales, con la sentencia TEST. Por ejemplo para el modelo lineal general se puede establecer los siguientes conjuntos de hipótesis.

$H_0 : L\beta = c$

$H_0 : L\beta \neq c$

Donde L es la matriz de coeficientes asociados a los parámetros, β es el vector de parámetros asociados a un modelo lineal y finalmente c es un vector de constantes.

Ejemplo 15.
Consideremos el problema anterior donde el modelo es:
$peso = \beta_0 + \beta_1 estatura + \beta_2 edad + \varepsilon$
El contraste es:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 36 \\ 0{,}5 \end{bmatrix}$$

Aplcando el procedimiento MIANALIZE y usando la sentencia TEST codificamos el siguiente programa:

```
proc mianalyze data=outreg edf=28;
     modeleffects Intercept estatura_  edad;
          test  Intercept=0/bcov tcov wcov;
 test estatura_ =36/ bcov tcov wcov;
test edad=0.5/bcov tcov wcov;
   run;
```

El programa es similar al anterior pero se le añadió varias sentencias TEST para realizar el contraste propuesto, por tanto sólo se comentará de la nueva salida el resultado de aplcar dicha sentencia.
Para comprender la salida seleccionada de este programa basta con comentar el primer caso.
La primera salida da el primer vector fila de la columna de la matriz L y el primer componente del vector c.

The SAS System

The MIANALYZE Procedure

Test: Test 1

Test Specification				
	L Matrix			
Parameter	Intercept	estatura_	edad	C
TestPrm1	1.000000	0	0	0

La construcción de un test trae consecuencias en la variación dentro y entre imputaciones y dentro de las imputaiones tal como muestra en el cuadro siguiente.

Multiple Imputation Variance Information						
Parameter	Variance			DF	Relative Increase in Variance	Fraction Missing Information
	Between	Within	Total			
TestPrm1	28.688674	299.017899	333.444309	22.105	0.115132	0.107986

El cuadro siguiente da la siguiente información: La hipótesis nula es $H_0 : \beta_0 = 0$ y la hipótesis altena: $H_1 : \beta_0 \neq 0$. La estimación obtenida del parámetro es $\beta_0^* = 3{,}42539$, además muestra el valor de la estimación del error estándar, el intevalo de confianza y el resultado del test.

Multiple Imputation Parameter Estimates										
Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	C	t for H0: Parameter=C	Pr > |t|
TestPrm1	3.425390	18.260458	-34.4340	41.28479	22.105	-4.493107	10.435661	0	0.19	0.8529

La información siguiente es la misma que la vista anteriormente. Si en el test hubiese estados varios parámetros (una sóla sentencia) tendríamos las matrices de covarianzas descompuestas en la asociada a dentro de las imputaciones y entre imputaciones. Pero los tres test están formado cada uno por un solo parámetro.

Within-Imputation Covariance Matrix	
	TestPrm1
TestPrm1	299.0178993

Between-Imputation Covariance Matrix	
	TestPrm1
TestPrm1	28.68867435

Total Covariance Matrix	
	TestPrm1
TestPrm1	333.4443085

Los comentarios anteriores son extensivos a las siguientes saldas

The SAS System
The MIANALYZE Procedure
Test: Test 2

Test Specification				
	L Matrix			
Parameter	Intercept	estatura_	edad	C
TestPrm1	0	1.000000	0	36.000000

Multiple Imputation Variance Information						
	Variance					
Parameter	Between	Within	Total	DF	Relative Increase in Variance	Fraction Missing Information
TestPrm1	8.993077	103.215453	114.007145	22.518	0.104555	0.098687

Multiple Imputation Parameter Estimates										
Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	C	t for H0: Parameter=C	Pr > |t|
TestPrm1	36.665801	10.677413	14.55170	58.77991	22.518	32.467568	40.893262	36.000000	0.06	0.9508

Within-Imputation Covariance Matrix	
	TestPrm1
TestPrm1	103.2154532

Between-Imputation Covariance Matrix	
	TestPrm1
TestPrm1	8.993076824

Total Covariance Matrix	
	TestPrm1
TestPrm1	114.0071454

The SAS System
The MIANALYZE Procedure
Test: Test 3

Test Specification				
Parameter	L Matrix			C
	Intercept	estatura_	edad	
TestPrm1	0	0	1.000000	0.500000

Multiple Imputation Variance Information						
Parameter	Variance			DF	Relative Increase in Variance	Fraction Missing Information
	Between	Within	Total			
TestPrm1	0.000388	0.006371	0.006837	23.735	0.073156	0.070327

Multiple Imputation Parameter Estimates										
Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	C	t for H0: Parameter=C	Pr > |t|
TestPrm1	0.198543	0.082684	0.027791	0.369295	23.735	0.171915	0.220084	0.500000	-3.65	0.0013

Within-Imputation Covariance Matrix	
	TestPrm1
TestPrm1	0.0063705870

Between-Imputation Covariance Matrix	
	TestPrm1
TestPrm1	0.0003883718

Total Covariance Matrix

	TestPrm1
TestPrm1	0.0068366332

Compare las variaciones entre y dentro obtenido en este ejemplo y el anterior.

Ejemplo 16.

En este ejemplo veremos una variante del caso anterior, en donde se hace el contraste de los parámetros en una sóla sentencia:

```
proc mianalyze data=outreg edf=28;
     modeleffects Intercept estatura_  edad;
          test  Intercept=0/bcov tcov wcov;
 test Intercept=0, estatura_ =36,edad=22.5 / bcov tcov wcov;

   run;
```

La diferencia entre esta corrida y la anterior es la forma que presenta los resultados con respecto a los test.

The SAS System
The MIANALYZE Procedure

Model Information	
Data Set	WORK.OUTREG
Number of Imputations	5

Multiple Imputation Variance Information						
Parameter	Variance			DF	Relative Increase in Variance	Fraction Missing Information
	Between	Within	Total			
Intercept	28.688674	299.017899	333.444309	22.105	0.115132	0.107986
estatura_	8.993077	103.215453	114.007145	22.518	0.104555	0.098687
edad	0.000388	0.006371	0.006837	23.735	0.073156	0.070327

Multiple Imputation Parameter Estimates										
Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
Intercept	3.425390	18.260458	-34.4340	41.28479	22.105	-4.493107	10.435661	0	0.19	0.8529
estatura_	36.665801	10.677413	14.5517	58.77991	22.518	32.467568	40.893262	0	3.43	0.0023

Multiple Imputation Parameter Estimates										
Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
edad	0.198543	0.082684	0.0278	0.36930	23.735	0.171915	0.220084	0	2.40	0.0245

The SAS System
The MIANALYZE Procedure
Test: Test 1

Test Specification				
Parameter	L Matrix			C
	Intercept	estatura_	edad	
TestPrm1	1.000000	0	0	0

Multiple Imputation Variance Information						
Parameter	Variance			DF	Relative Increase in Variance	Fraction Missing Information
	Between	Within	Total			
TestPrm1	28.688674	299.017899	333.444309	22.105	0.115132	0.107986

Multiple Imputation Parameter Estimates										
Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	C	t for H0: Parameter=C	Pr > |t|
TestPrm1	3.425390	18.260458	-34.4340	41.28479	22.105	-4.493107	10.435661	0	0.19	0.8529

Within-Imputation Covariance Matrix	
	TestPrm1
TestPrm1	299.0178993

Between-Imputation Covariance Matrix	
	TestPrm1
TestPrm1	28.68867435

Total Covariance Matrix

	TestPrm1
TestPrm1	333.4443085

La siguiente matriz L corresponde al segundo test en donde se especifica la hipótesis nula test Intercept=**0**, estatura_ =**36**,edad=**22.5.** Esto es: se ha colocado en una sóla sentencia el contraste para cada uno de los parámetros. De esta forma, a diferencia del ejemplo anterior, se muestra la matriz completa de una vez.

The SAS System
The MIANALYZE Procedure
Test: Test 2

Test Specification				
	L Matrix			
Parameter	Intercept	estatura_	edad	C
TestPrm1	1.000000	0	0	0
TestPrm2	0	1.000000	0	36.000000
TestPrm3	0	0	1.000000	22.500000

La información que sigue de los cuadros ya se conoce, la interpretación de los cuadros es la misma de los ejemplos anteriores. Como ya se indicó con anterioridad, ella permite comparar sus resultados con los obtenidos con otro método de imputación.

Multiple Imputation Variance Information						
	Variance					
Parameter	Between	Within	Total	DF	Relative Increase in Variance	Fraction Missing Information
TestPrm1	28.688674	299.017899	333.444309	22.105	0.115132	0.107986
TestPrm2	8.993077	103.215453	114.007145	22.518	0.104555	0.098687
TestPrm3	0.000388	0.006371	0.006837	23.735	0.073156	0.070327

Multiple Imputation Parameter Estimates										
Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	C	t for H0: Parameter=C	Pr > |t|
TestPrm1	3.425390	18.260458	-34.4340	41.28479	22.105	-4.493107	10.435661	0	0.19	0.8529
TestPrm2	36.665801	10.677413	14.5517	58.77991	22.518	32.467568	40.893262	36.000000	0.06	0.9508
TestPrm3	0.198543	0.082684	0.0278	0.36930	23.735	0.171915	0.220084	22.500000	-269.72	<.0001

Los cuadros siguientes pueden ser útiles para comparar varios métodos de imputación, tomando en cuenta la variación dentro y entre las imputaciones dado los contrastes de hipótesis.

Within-Imputation Covariance Matrix			
	TestPrm1	TestPrm2	TestPrm3
TestPrm1	299.0178993	-173.6538787	-0.0321848
TestPrm2	-173.6538787	103.2154532	-0.0978463
TestPrm3	-0.0321848	-0.0978463	0.0063706

Between-Imputation Covariance Matrix			
	TestPrm1	TestPrm2	TestPrm3
TestPrm1	28.68867435	-15.95484906	-0.04693224
TestPrm2	-15.95484906	8.99307682	0.02004129
TestPrm3	-0.04693224	0.02004129	0.00038837

Total Covariance Matrix			
	TestPrm1	TestPrm2	TestPrm3
TestPrm1	319.0816452	-185.3058477	-0.0343444
TestPrm2	-185.3058477	110.1410875	-0.1044116
TestPrm3	-0.0343444	-0.1044116	0.0067980

OBSERVACIONES GENERALES.

Ahora haremos algunas observaciones a manera de síntesis de lo visto en cuanto al problema de imputación se refiere, sin limitarnos al software empleado.

- En primer lugar no siempre las técnicas existentes de imputación son las mejores soluciónes cuando se puede recuperar información con revisitas en el caso de datos faltantes o, se pueden corregir los valores atípicos producto de errores humanos.

- El recuperar información o la corrección de lo mismos pueden estar sujeto a los costos asociados a estas actividades y al tiempo previso, esto último es al horizonte donde la información recabada es pertinente.

- Una limitación importante que es común a cualquier técnica que se quiera emplear para imputar los datos faltantes, es la proporción de los mismos en las observaciones obtenidas en la investigación. Como se indicó en su momento la importancia de ese porcentaje está ligado al uso que se hará a la información.

- Toda imputación debe estar presedida por el uso de métodos estadísticos que den cuenta de la presencia tanto de datos perdidos como atípicos, estos métodos en algunos casos podrán dar alguna luz sobre las causas de la presencia de los mismos.

- Los datos que se detecten como atípicos deberán considerarse como perdidos siempre y cuando su presencia no se justifique desde el punto de viata de la ley de la población de origen.

- Algo similar al caso de los datos atípicos, puede ocurrir con distribuciones censuradas. Thomas Büttner y Susanne Rässler (2006) presentan un trabajo donde utilizan la imputación sobre datos censurados, concretamente sobre censura establecida a los salarios. Ellos emplean la técnica MCMC con algunas modificaciones apropiadas para el caso que estudian.

- Para imputar los datos perdidos debe considerarse varios elementos: la escala de medición de la variable que será imputada y la escala de medición de aquellas que formaran parte del modelo estadístico que servirá para realizar la imputación.

- Todo modelo, tales como regresión, discriminante etc que se esté empleando para realizar la imputación, debe ser plenamente validados de acuerdo a los criterios propios de cada modelo.

- Aunado con los dos puntos anteriores está la estructura o patrón que siguen la presencia de datos perdidos: monótona o arbitraria. Además debe considerarse el comportamiento aleatorio de los datos perdidos. En la mayoría de los métodos de imputación y en todos en el caso de SAS versión 9.0 se asume que el comportamiento probabilístico es MAR.

- Decidir entre usar imptación simple o imputación múltiple está ligado a varios aspectos: uno de ellos es el tipo de investigación que se realiza y el otro a la disponibilidad de memoria ran que tienen los equipos. En el caso de una base de datos de más de 100000 observaciones, si no se posee equipos de PC con una alta capacidad de procesamiento, posiblemente se tenga problema a la hora de aplicar imputación multiple.

- Un método prometedor dentro la imputación simple es aquel que esté construidos sobre los vecinos más próximos, en donde la distancia es multivariante. Esto se puede lograr usando la idea de conglomerados. El problema está en decidir cual de los diferentes métodos de construcción de conglomerados seleccionar y cual es el número de conglomerado a elegir. La respuesta queda en mano del investigador.

- De acuerdo a la bibliografía consultada para realizar esta breve monografía y alguna experiencia que tiene el que suscribe sobre este tópico, el uso de la imputación múltiple reduce el efecto incertidumbre sobre la presencia de datos faltantes y, abre un abanico para facilitar el análisis de los datos. Se puede analizar cada imputación por separado y en su conjunto.

- En todo caso, es recomendable usar diferentes herramientas estadística para efectuar la imputación. Tan válida puede ser una imputación simple con regresión logística o discriminante como una múltiple usando MCMC. La decisión final dependerá de los expertos que adelantan el estudio.

- Hay que diferenciar claramente el modelo que se emplea para realizar la imputacón, obtendo para cada imputación por simulación y, el mismo modelo que sirve para hacer el análisis de los datos completados con las imputaciones.

- Los métodos de imputación vistos aquí no son los únicos pero si son de los más populares. Se ha empleado otras técnicas como regression tree para hacer imputaciones sobre datos no métricos.

- De los métodos de imputación que ofrece SAS y que parecieran los más versátiles son para el caso de imputación simple el que está contruido con el análisis de conglomerado no jerarquico y que se resuelve con el procedimiento FATCLUS y de imputación múltiple el MCMC que sirve tanto para datos con estructura arbitraria como monótona.

- En la selección del método de imputación múltiple hay que tomar en cuenta varios elementos. El primero tiene que ver con el concepto de imputación razonable, esto quire decir: si la imputación coincide con el valor que espera el investigador dada la naturaleza de los datos. Los otros criterios son: el resultado de la descomposición de la varianza total en dentro de imputaciónes y entre imputaciones, además, del incremento relativo de la varianza y la fracción de pérdida de información por la presencia de datos perdidos.

- El procedimento MIANALIZE permite construir cualquier modelo, si se acompaña con el código adecuado, y ver la descomposición de la varianza, los contrastes de hipótesis, la construcción de intervalos asociado a las estimaciones de los parámetros.

- En el caso de imputación quedan por resolver varios problemas. Uno de ellos es contar con test que permitan diferenciar entre un MAR y cualquier otro. Una posible solución es usar el test no paramétrico de las rachas.

- Finalmente, una vez imputado los datos por cualquiera de los métodos que se seleccione siempre se tendrá la ventaja sobre datos no imputados, bien sea para construir modelos multivariantes o para otros fines. Claro, tomando en cuenta las limitaciones de cada caso.

BIBLIOGRAFÍA.

Andrieu et al (2005).
On the efficiency of adaptive MCMC algorithms.
portal.acm.org/citation.cfm?id=1190150

Allison Paul D.
Imputation categorical data.
University of Pennsylvania, Philadelphia, PA

Barnes H el al (2006)
Multiple imputation of missing data in the 2001 South African Census.
Center for the analysis of South African social Policy. University of Oxfort.

Bor-Chung Chen.
CANCEIS Experiments of Edit and Imputation with 2006 Census Test Data.
Statistical Research Division U.S. Census Bureau
Washington, DC 20233 Report.

Büttner T and Rässler S (2006)
Multiple Imputation of Right-Censored Wages in the
German IAB Employment Register Considering Heteroscedasticity.
q2008.istat.it/sessions/paper/28Buttner.pdf

Cea D´Ancona M.A (2004)
Análisis Multivariante.-Teoría y Práctica en la Investigación Social.
Editoral Síntesis. Madrid. España

Chernozhukov, V et al (2003)
An MCMC approach to classical estimation.
Journal of Econometrics 115 pag 293-346

Cohen M. (1997)
The Bayesian Bootstrap and Multiple Imputation for uniqual probability sample designs. www.anstat. Org

Dempster. A.P., N.M. Lair and D.B. Rubin (1977)
Maximum Likelihood from Incomplete Data via EM Algorithm.
Jounal of the Royal Statistical Society, Serie B, 39.

Fay Robert. E (1996)
Alternative Paradigms for the Analysis of Imputed Survey Data.
JASA Vol 91 Nº 434 pag 490-498

Greene. W,H (2000)
Análisis Econométrico.
Prentice Hall.-Tercera Edición. Madrid. España.

Gomez García J et al (2006)
Métodos de Inferencia Estadística con datos faltantes. Estudio de simulación sobre los efectos de las estimaciones.
Estadística Española Vol 48. Num 162, pags 241 a 270.

Gujarati, D.N (2004)
Econometría 4ª Edición.
McGraw Hill-México.

Jagannathan G et al(2008)
Privacy-preserving imputation of missing data.
www.sciencedirect.com

Khattree R y Naik D.(2000)
Multivarate Data Reduction and Discrimination with SAS software.
SAS institute. Inc.

Little. R(1976)
Iference abaut means incomplete multivariate data.
Biometrika. 63,3 pag:593-604

Little. R(1988)
A Test of Missing Completely at Randon for Multivariate Data With Missing Values.
JASA Vol 83 Nº 404 pag. 1198-1202

Little. R(1995)
Modeling the Drop-Out Mechanism in Repeated-Measures Studies.
JASA, Vol 70, Nº 431.

Littell. R; Stroup. W,W: Freund. R.J (2002)
SAS for Linear Models
SAS institute. Inc.

Medina F et al (2007)
Imputación de datos: teoría y práctica.
CEPAL.

Peña D (2002)
Análisis de Datos Multivarientes. McGraw Hill.

Rao J.N.K (1996)
On variance Estimation with Imputed Survey Data.
JASA Vol 91 Nº 434 pag 499-506.

Rubin D.B (1976).
Inference and missing data.
Biometrika 63, pag 581-592

____(1988)
An Overview of Multiple Imputation.
JASA.

____ (1996)
Multiple Imputation after 18+ years.
JASA Vol 91 Nº 434 pag 473-489.

Shafer, J.L and Olsen, M.K (1999)
Modeling and imputation of semicontinous survey variables:
The Pennsylvania State University.

_____(2002)
Dealing with Missing Data
Res. Lett. Inf. Math. Sci. (2002) **3**, 153-160

SAS/STAT (1999)
Version 8 Vol 1-2 y 3.
SAS institute. Inc.

Todd R. Williams.
Imputing person age for the 2000 census short form: A Model Base Approach.
Bureau of the Census, twilliam@census.govE

Tutorial SAS 9.0,
SAS institute. Inc.

Yang C. Yuan.
Multiple Imputation for Missing Data: Concepts and New Development
SAS Institute. Inc. www.**sas**.com

Zhang L (1998)
Likelihood Imputation.
Statistics Norway. Oslo pag 147-152
www.jstor.org/stable/4616510

Zhang P. (2003)
Multiple Imputation: Theory and Method.
International Statistical Review. Pag 581-592