APENDICE 2

El Análisis Factorial como Instrumento Decisorio ante los Problemas de Multicolinialidad en una Regresión.

Por

Ing. Roberto Piol Puppio
E-Mail: rpiol@yahoo.com
www.rpiol.com

I.- Introducción

En el texto de esta monografía, se ha hecho mucho hincapié en el problema que representa la Multicolinialidad entre Variables Independientes en una Regresión.

Como ya se indicó, un coeficiente de determinación alto (R^2), no es garantía para que la regresión exista.

Una alta correlación entre Dos (2) o mas variables independientes, lo afectan directamente.

Para la detección de un problema de Multicolinialidad en una regresión; se utiliza como instrumento: La Matriz de Correlación (o la matriz de covarianza). Interpretando a un coeficiente de correlación alto ($r>0.75$) entre dos variables independientes como señal de su presencia.

También en el texto se explicó que dos variables independientes autocorrelacionadas, no podían convivir juntas en una regresión. Por lo tanto una de las dos debía eliminarse.

La pregunta ante un problema de multicolinealidad, que una persona se hace es: ¿Cuál es la variable independiente que hay que eliminar del juego de datos referenciales?.

La respuesta a esta pregunta es: La menos significativa.

Pero, ¿Cómo se identifica esa variable “menos significativa” en una regresión?. No es fácil. Tampoco se puede deducir empíricamente de una simple observación a los datos.

Aquí es donde entra el “Análisis Factorial”. Este procedimiento estadístico, será utilizado para identificar la o las variables menos significativas de una regresión con problemas de Multicolinialidad,

II.- Conceptos Básicos

Se define como Análisis Factorial, al procedimiento estadístico que permite identificar un número de factores que representan la relación que existe entre un conjunto de variables independientes autocorrelacionadas entre si.

Por lo tanto, el procedimiento intenta agrupar las variables independientes autocorrelacionadas entre sí, de manera que las mismas tenga una correlación baja con el resto de las variables independientes.

De esta manera, identificaremos el grupo de variables independientes correlacionadas entre sí y podremos clasificarlas según su importancia; de manera que podamos eliminar la menos representativa sin perturbar (o con la mínima perturbación) a la serie de datos.

Otros de las ventajas del Método de Análisis Factorial, es el de reducir la número de variables independientes en un modelo de regresión, de tal manera de obtener otro modelo de regresión con menos variables independientes. Sin embargo, esto no forma parte del curso y nos centraremos en el problema de la Multicolinialidad.

El hecho de eliminar la Variable Independiente menos representativa, no implica necesariamente que el nuevo nivel de significación (R^2) del modelo de regresión aumente. Puede que la variable eliminada sea en realidad representativa en el modelo de regresión múltiple definitiva. En su lugar puede ser sustituida por la siguiente variable en orden de su representatividad.

Lo realmente importante es que solo una de las variables independientes de un factor compuesto por variables muy correlacionadas entre sí podrá quedar en la regresión. En caso de que esto no se cumpliera, seguiríamos teniendo problemas de multicolinialidad.

Es de hacer notar, que este procedimiento estadístico es valido para series grandes; mientras mas pequeña sea la serie, el método menos significativo será.

III.- El uso del paquete estadístico SPSS (versión 9), en el desarrollo del Análisis Factorial.

El paquete estadístico dedicado SPSS, por su facilidad y amigabilidad de sus comandos, es uno de los preferidos a nivel global.

En este curso, no se enseñará el manejo de dicho paquete; tan solo se explicará paso a paso el procedimiento.

El objetivo final será el de clasificar la variable (o variables) menos significativas dentro de un factor y eliminarla (o eliminarlas), a fin de resolver el problema de la multicolinialidad en una regresión.

Generalizando, los pasos para una Análisis Factorial son:

1. Generar la Matriz de Correlación
2. Extraer los factores de la Matriz, en base a los coeficientes de correlación de las variables
3. Rotar los factores con el fin de maximizar la relación entre las variables a algunos de los factores
4. Seleccionar Una (1) Variable Independiente por Factor.

Es de notar, que para lograr los enunciados anteriores es necesario tener nociones del manejo de un paquete estadístico dedicado. En este caso se usará el software SPSS versión 9. En este texto, se tratará de indicar paso a paso el procedimiento, sin embargo, esta monografía no es suficiente para el dominio de este procedimiento automatizado.

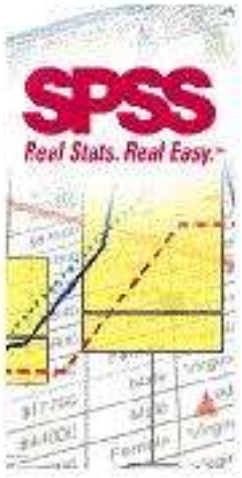
IV EL Análisis Factorial paso a paso:

1.- Preparación de la Data

En el editor de datos (DATA EDITOR) del SPSS, se debe crear la data a procesar. El software permite enterar directamente la data o importarla desde la Hoja de Cálculo Excel.

El siguiente ejemplo se corresponde a una serie de 20 referenciales de Casas Quintas en el sureste de Caracas. Las Variables seleccionadas son:

punit	Precio Unitario del inmueble
aterreno	Area del Terreno
aconstr	Area de Construcción
año	Año de construcción del inmueble
habitac#	Número de dormitorios
baños	Número de baños
vista	Inmuebles con vista a Caracas = 1. Con vista al los Valles del Tuy =0
fecha	Fecha[1] de protocolización de la compra-venta



Salida del Editor de Datos del SPSS:

[1] La fecha (Variable No Numérica), esta expresada en el formato de MS-Excel; donde cuenta los días transcurridos desde el 1ro de Enero del año 1900. Este formato permite a la Hoja de Cálculo expresar la fecha o bien como un número o bien como algunos de los formatos tradicionales: dd-mm-aa.

ATERRENO	ACONSTR	AÑO	HABITAC	BAÑOS	VISTA	FECHA	PUNIT
291.00	256.00	1,978	4	3	0	36,758	180,000,000
252.30	158.06	1,970	2	2	0	36,784	120,000,000
283.63	181.52	1,969	3	2	0	36,856	120,000,000
255.75	170.50	1,969	2	2	0	36,854	143,000,000
340.90	152.00	1,958	2	2	0	36,857	100,000,000
270.00	320.00	1,970	5	4	0	36,848	140,000,000
395.25	161.72	1,972	2	2	0	36,823	165,000,000
390.48	157.91	1,975	2	2	0	36,882	178,000,000
390.00	157.91	1,975	2	2	0	36,882	185,500,000
388.24	157.91	1,975	2	2	0	36,877	185,500,000
529.10	157.91	1,975	2	2	0	36,877	190,000,000
384.00	157.91	1,975	2	2	0	36,877	185,500,000
306.93	187.00	1,972	3	2	0	37,082	168,000,000
341.00	158.96	1,983	2	2	1	37,208	280,000,000
399.75	289.08	1,980	4	4	1	37,193	280,000,000
315.00	330.00	1,980	5	4	1	37,255	250,000,000
414.00	261.00	1,975	4	3	0	37,242	205,000,000
179.00	161.31	1,978	2	2	0	36,906	150,000,000
183.60	166.16	1,971	2	2	0	37,251	150,000,000
300.00	177.10	1,958	2	2	0	36,916	70,000,000

2.- Acceso al la Subrutina de Análisis Factorial (FACTOR ANÁLISIS):

Una vez cargados los datos en el Editor de Datos (DATA EDITOR), en la Barra de Menú seleccione:

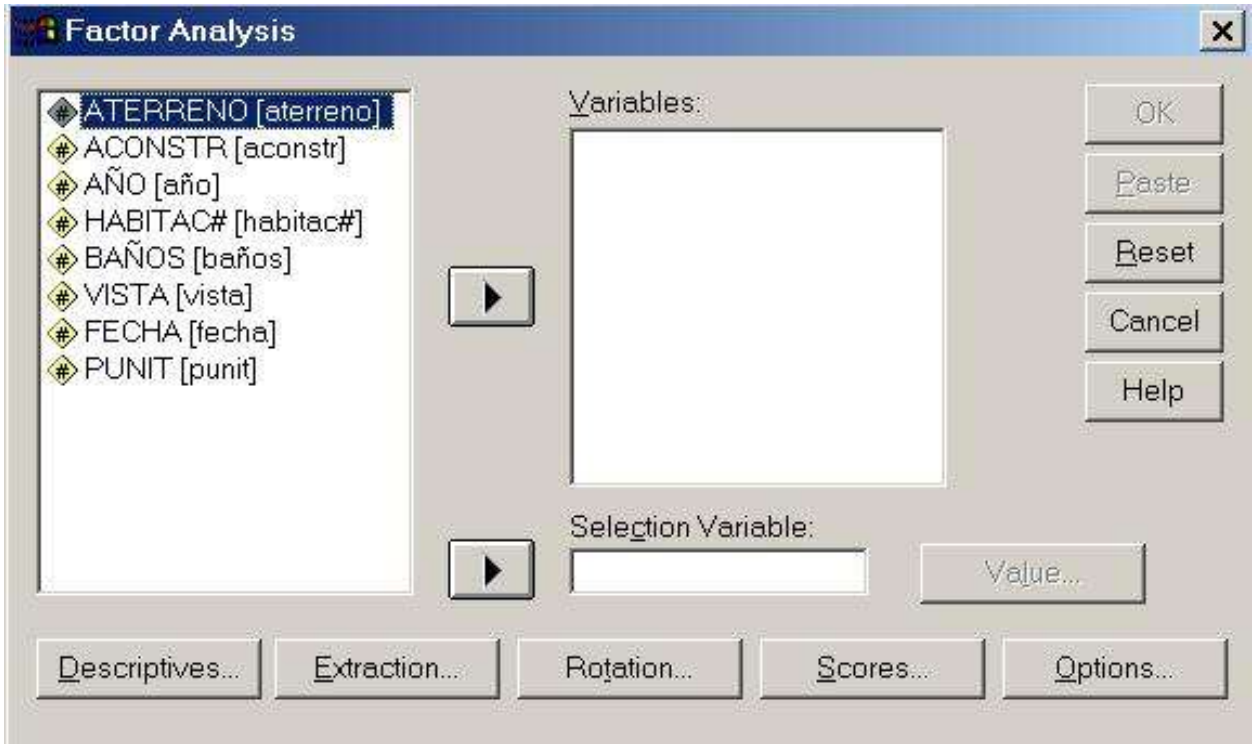
Analyze
Data Reduction
Factor

Estos comandos presentaran la caja de diálogo principal del Análisis Factorial (FACTOR ANÁLISIS), que tendrá la forma siguiente[2]:

[2] Nótese que la caja de dialogo tiene Dos (2) ventanillas verticales. En la ventanilla izquierda el software presentó todas las variables de nuestra serie en estudio. También fíjense los Cinco Botones en la parte baja de la caja:

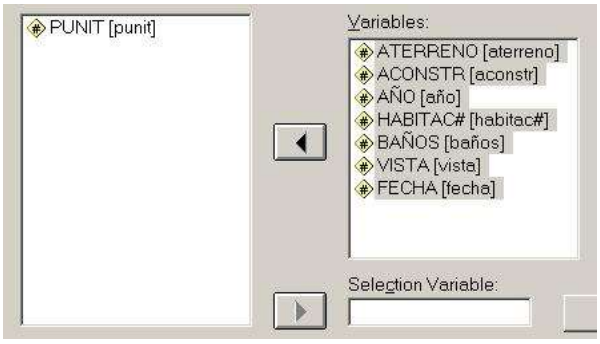
Descriptives **Extraction** **Rotation** **Scores** **Options**

Estos Cinco (5) botones conforman la configuración del Análisis Factorial y su uso es fundamental para la correcta salida del programa.



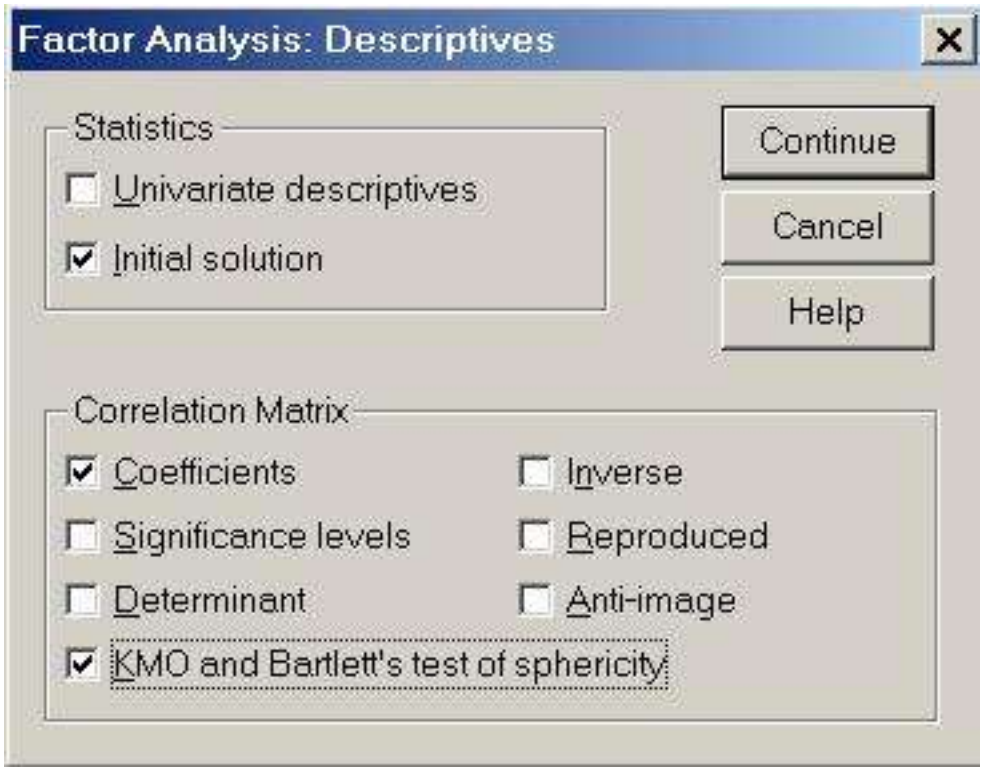
Ilumínese con el ratón únicamente las Variables Independientes de la ventanilla izquierda y por medio de la flecha central (>) pásense a la ventanilla derecha (denominada **Variables:**).

A continuación, configúrese cada uno de los Cinco (5) botones que conformarán la salida (OUTPUT) de la subrutina Análisis Factorial (FACTOR ANÁLISIS):



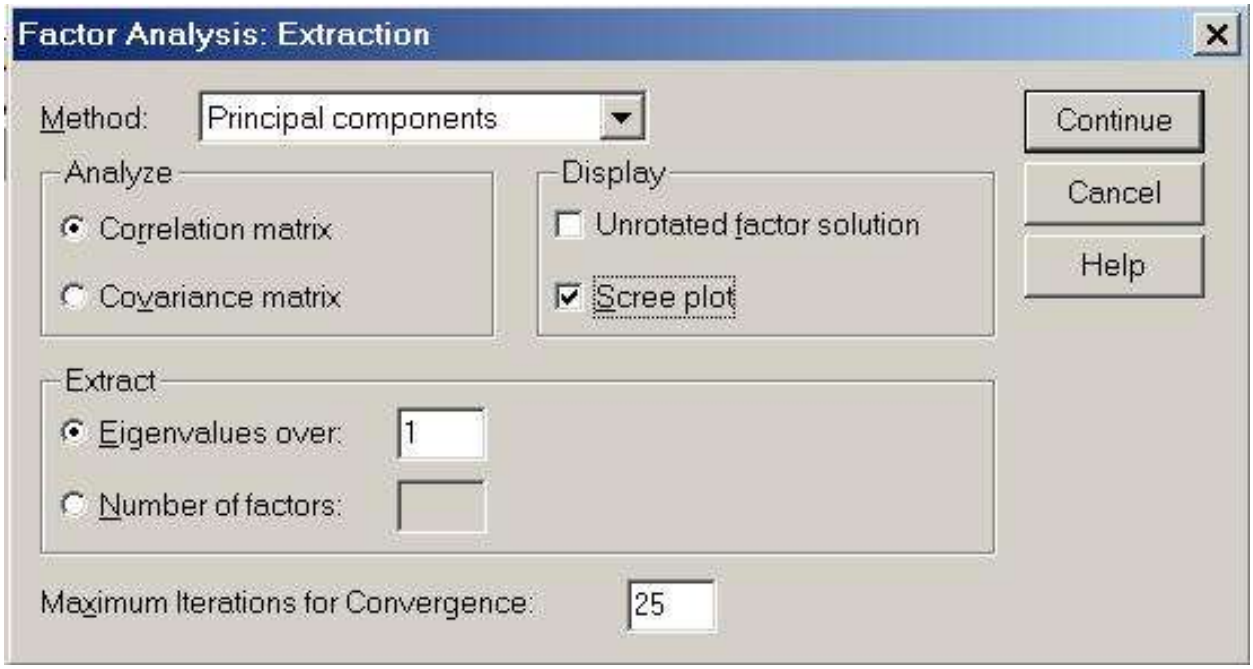
2.1.- Configuración del botón **Descriptives**:

La caja de diálogo del botón **Descriptives**, debe estar configurado de la siguiente manera:



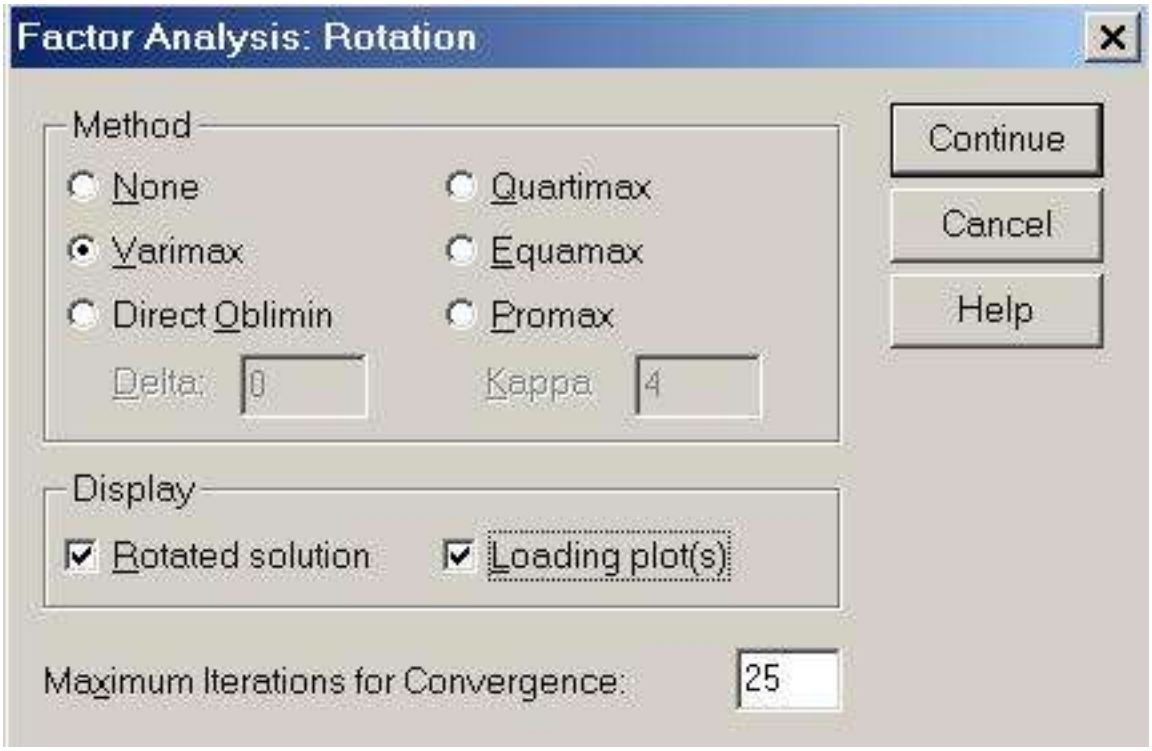
2.2.- Configuración del botón **Extraction**:

La caja de diálogo del botón **Extraction**, debe estar configurado de la siguiente manera:



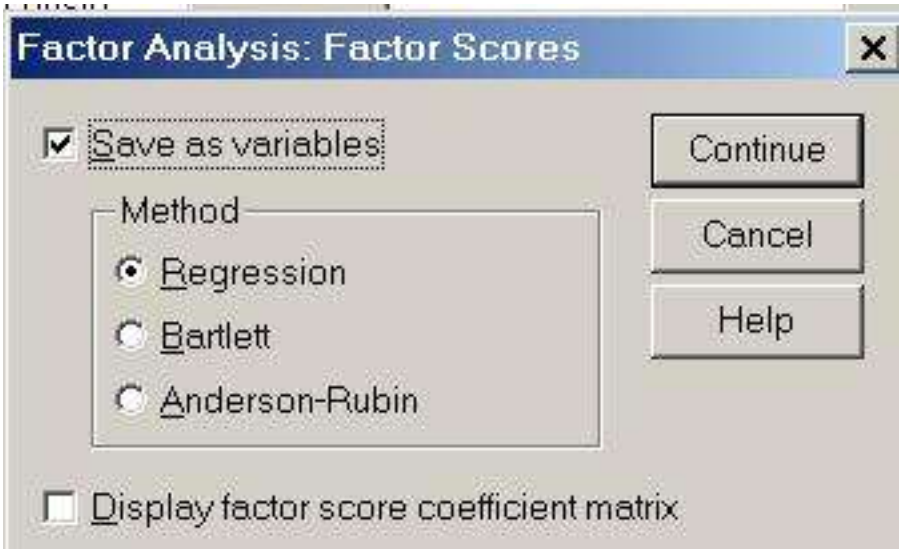
2.3.- Configuración del botón **Rotation**:

La caja de diálogo del botón **Rotation**, debe estar configurado de la siguiente manera:



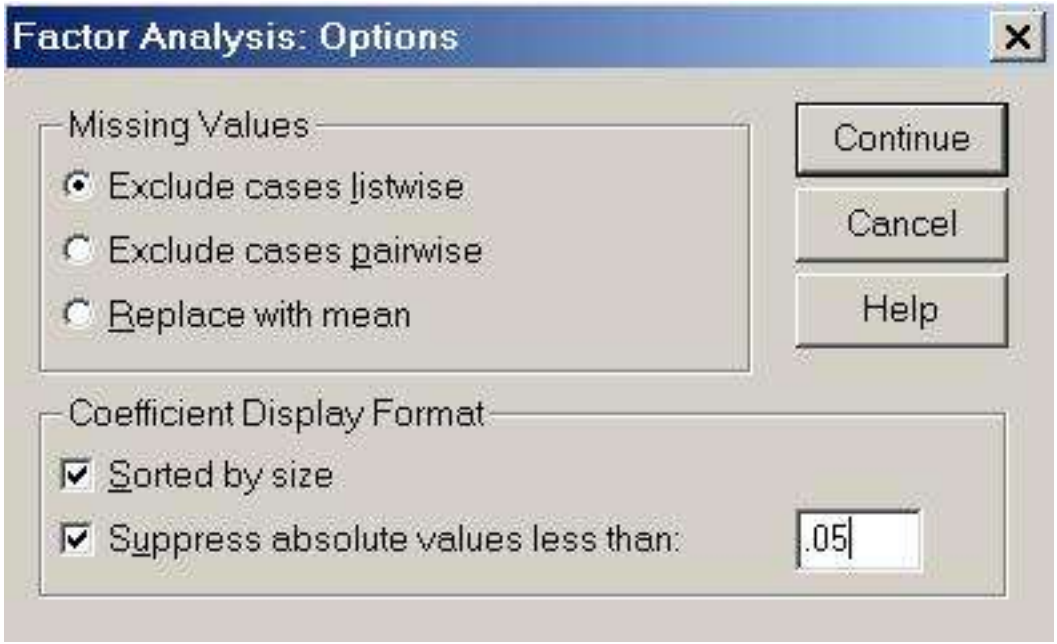
2.4.- Configuración del botón **Scores**:

La caja de diálogo del botón **Scores**, debe estar configurado de la siguiente manera:



2.5.- Configuración del botón **Options**:

La caja de diálogo del botón **Options**, debe estar configurado de la siguiente manera:



3.- Interpretación de la Salida (OUTPUT) de la Subrutina Análisis Factorial (FACTOR ANÁLISIS):

3.1.- La Matriz de Correlación:

La primera salida del software es la matriz de correlación:

Correlation Matrix

	ATERRENO	ACONSTR	AÑO	HABITAC#	BAÑOS	VISTA	FECHA
ATERRENO	1.000	-.037	.197	-.042	.025	.110	-.008
ACONSTR	-.037	1.000	.284	.978	.977	.462	.379
AÑO	.197	.284	1.000	.277	.325	.546	.385
HABITAC#	-.042	.978	.277	1.000	.930	.385	.334
BAÑOS	.025	.977	.325	.930	1.000	.534	.368
VISTA	.110	.462	.546	.385	.534	1.000	.649
FECHA	-.008	.379	.385	.334	.368	.649	1.000

Obsérvese que existe una correlación muy alta entre las variables:

HABITC# - ACONSTR :	0.978
BAÑOS – ACONSTR:	0.977
BAÑOS – HABITAC#:	0.930

Como puede observarse, existe un problema de Multicolinialidad en la serie y por lo tanto solo una de las tres variables: HABITC# - ACONSTR – BAÑOS debe quedar para que la regresión exista.

3.2 Tests KMO y de Bartlett:

Para que sea significativo el Análisis Factorial, el test KMO (Kaiser – Meyer – Olkin) debe ser **> 0.5**.

El test de esfericidad de Bartlett, indica que la matriz de correlación no sea una matriz identidad[3].

El nivel de significancia (sig.) debe ser **< 0.05** (mientras mas se aproxime a cero (0) mejor).

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.625
Bartlett's Test of Sphericity	Approx. Chi-Square	131.902
	df	21
	Sig.	.000

[3] Se define como Matriz Identidad, aquella que todos sus elementos son Cero (0) menos la diagonal principal que es Uno (1), por ejemplo:

1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

3.3.- Las Comunalidades (COMMUNALITIES):

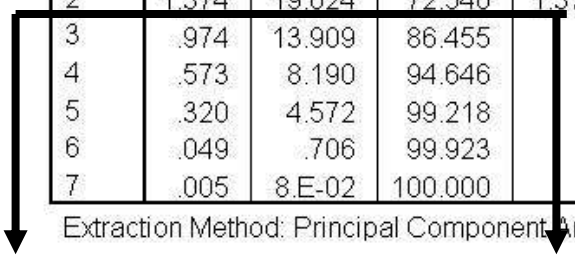
La tabla de Comunalidades, muestra la proporción de la varianza de cada variable explicada por los factores extraídos.

3.4.- La Varianza Total Explicada (TOTAL VARIANCE EXPLAINED)

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.705	52.922	52.922	3.705	52.922	52.922	3.125	44.645	44.645
2	1.374	19.624	72.546	1.374	19.624	72.546	1.953	27.901	72.546
3	.974	13.909	86.455						
4	.573	8.190	94.646						
5	.320	4.572	99.218						
6	.049	.706	99.923						
7	.005	8.E-02	100.000						

Extraction Method: Principal Component Analysis.



La tabla de Varianza Total Explicita (TOTAL VARIANCE EXPLAINED), muestra todos los Factores extraíbles ordenados de acuerdo a su Valor Propio (EIGENVALUES).

Si se observa la columna identificada "Total " (por nosotros como "A"); se puede notar que solamente en Dos (2) Factores su Valor Propio (EIGENVALUES) es mayor que 1.00.

Todos los demás factores no son significativos y por lo tanto serán ignorados.

Obsérvese en la columna identificada "Cumulative %" (por nosotros como "B"), que los Dos (2) factores seleccionados suman el 72.546% de la varianza (52.922% + 19.624%).

3.5.- La Rotación de la Estructura de los Factores

El objetivo de la rotación de la estructura de los factores, es la de obtener un claro esquema para su correcta interpretación de la relación entre las variables y los factores extraídos.

El método de rotación de mayor uso en este tipo de análisis, es el denominado "Varimax"; y consiste rotar los ejes en cualquier dirección, sin cambiar la localización relativa de los factores extraídos, hasta obtener una claro esquema de la posición de las variables independientes en relación a los factores extraídos.

Rotated Component Matrix[a]

	Component	
	1	2
ATERRENO	-.207	.459
ACONSTR	.980	.148
AÑO	.174	.776
HABITAC#	.967	9.84E-02
BAÑOS	.948	.218
VISTA	.390	.778
FECHA	.322	.676

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

La tabla denominada "Matriz de Componentes Rotados" (ROTATED COMPONENT MATRIZ); indican la correlación existente entre cada una de las variables independientes y su correspondiente factor:

FACTOR 1:	ACONSTR	0.980	Fuerte/directa
	HABITAC#	0.967	Fuerte/directa
	BAÑOS	0.948	Fuerte/directa
FACTOR 2:	VISTA	0.778	Fuerte/directa
	AÑO	0.776	Fuerte/directa

Nótese que la matriz está ordenada bajo el criterio del grado de correlación de las variables independientes con respecto al "Factor Extraído"; de manera que sea fácil identificar las variables independientes incluidas en cada Factor.

Pero, regresando nuevamente a la Matriz de Correlación de la serie observamos:

Correlation Matrix

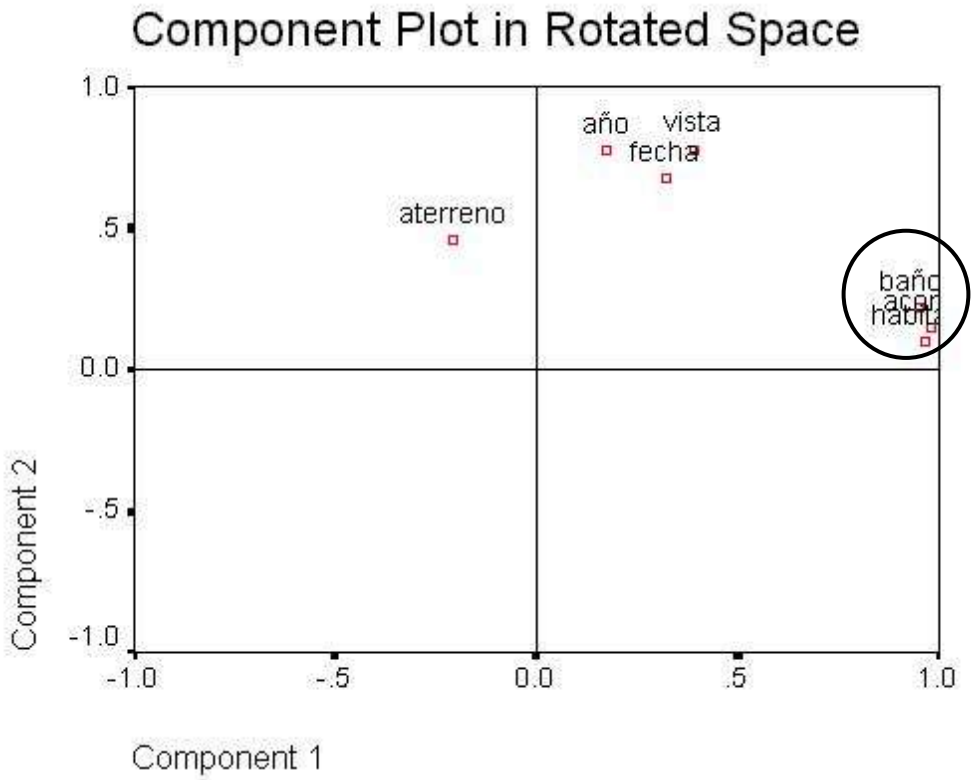
	ATERRENO	ACONSTR	AÑO	HABITAC#	BAÑOS	VISTA	FECHA
ATERRENO	1.000	-.037	.197	-.042	.025	.110	-.008
ACONSTR	-.037	1.000	.284	.978	.977	.462	.379
AÑO	.197	.284	1.000	.277	.325	.546	.385
HABITAC#	-.042	.978	.277	1.000	.930	.385	.334
BAÑOS	.025	.977	.325	.930	1.000	.534	.368
VISTA	.110	.462	.546	.385	.534	1.000	.649
FECHA	-.008	.379	.385	.334	.368	.649	1.000

Las variables autocorrelacionadas entre sí (que generan los problema de multicolinealidad) son únicamente: **ACONST HABITAC# BAÑOS**

Como se puede observar, esas variables son precisamente las mismas que conforman el FACTOR 1. Por lo tanto **solamente una de esas tres variables podrá existir en la regresión** y las demás deben ser excluidas, porque si así no se hiciera el problema de multicolinialidad persistiría en la serie.

3.6.- Representación Gráfica

La representación gráfica de los Factores Extraídos, nos permiten una mas fácil comprensión de las variables incluidas en cada Factor:



Nótese que las variables con un factor de correlación mas cercano a 1.00 (y por lo tanto mas correlacionados con el FACTOR 1 (Eje X) son:

ACONST
HABITAC#
BAÑOS

4.0.- La Selección de la Variable Independiente

Ya se definió en el punto anterior que de las Tres (3) variables independientes que se encuentran correlacionadas entre si, solo una podrá entrar en el modelo de regresión múltiple.

Si volvemos a observar la tabla "Matriz de Componentes Rotados" (ROTATED COMPONENT MATRIX):

Component Matriz

	Component 1	2
BAÑOS	.931	-.284
ACONSTR	.923	-.361
HABITAC#	.888	-.397
VISTA	.726	.480
FECHA	.616	.425
AÑO	.538	.586
ATERRENO	4.917E-02	.501

Extraction Method: Principal Component Analysis.
a 2 components extracted.

Observaríamos que la variable independiente "BAÑOS", tiene el coeficiente de correlación mas alto. Pero, estos coeficientes de correlación son los correspondientes entre la variable BAÑOS y el FACTOR 1.

Por lo tanto, no necesariamente es esta la variable que va a quedar en el modelo de regresión múltiple.

El criterio para aceptar la variable que va a quedar en la regresión múltiple, debemos buscarla en la Matriz de Correlación de la serie:

En este caso debemos solicitar al software SPSS la Matriz de Correlación incluyendo la Variable Dependiente (PUNIT)[4]

4.1.- Cálculo de la Matriz de Correlación incluyendo la Constante

Para obtener la Matriz de Correlación incluyendo la constante; debemos regresar al menú principal de SPSS:

ANALYZE DATA REDUCTION FACTOR

y al obtener la caja de diálogo principal marcar todas las variables (dependientes e independientes), cuidando que en la parte superior de la ventanilla derecha sea inicializada por la variable dependiente (PUNIT):

Cerciorarse que dentro de la configuración del boton DESCRIPTIVES, se marque el recuadro “Coeficients” para poder obtener la salida de la matriz de la siguiente forma:

Correlation Matrix

	Cte / PUNIT	ATERRENO	ACONSTR	AÑO	HABITAC#	BAÑOS	VISTA	FECHA
PUNIT / Cte	1.000	.409	.379	.876	.339	.448	.780	.603
ATERRENO	.409	1.000	-.037	.197	-.042	.025	.110	-.008
ACONSTR	.379	-.037	1.000	.284	.978	.977	.462	.379
AÑO	.876	.197	.284	1.000	.277	.325	.546	.385
HABITAC#	.339	-.042	.978	.277	1.000	.930	.385	.334
BAÑOS	.448	.025	.977	.325	.930	1.000	.534	.368
VISTA	.780	.110	.462	.546	.385	.534	1.000	.649
FECHA	.603	-.008	.379	.385	.334	.368	.649	1.000

Si analizamos la primera columna redefinida como **Cte.**, podremos inferir la correlación que existe entre cada variable independiente con la variable dependiente (regresión).

Podemos observar, que la correlación mas alta es 0.876 (AÑO – Cte).

Sin embargo, lo que nos interesa a nosotros es seleccionar la Variable Independiente que quedará en la regresión entre ACONSTR – HABITAC# - BAÑOS (que generan nuestros problemas de multicolinialidad).

De la matriz de correlación observamos los siguientes coeficientes de correlación:

Cte – ACONST	r= 0.379
Cte – HABITAC#	r= 0.339
Cte – BAÑOS	r= 0.448

Obsérvese que el coeficiente de correlación mas alto corresponde a la variable independiente BAÑOS.

En teoría, este sería la variable independiente que quedaría dentro del modelo de regresión; mientras que las variables ACONST y HABITAC# tendrían que salir para darle solución al problema de multicolinialidad de la serie.

4.2.- Comprobación de los Resultados

Para comprobar la hipótesis anterior; correremos tres veces el modelo de regresión lineal múltiple; utilizando para cada corrida una de las tres variables diferentes:

Serie 1	***Serie 2***	***Serie 3***
PUNIT	PUNIT	PUNIT
ACONST	**HABITAC#**	**BAÑOS**
ATERRENO	ATERRENO	ATERRENO
AÑO	AÑO	AÑO
VISTA	VISTA	VISTA
FECHA	FECHA	FECHA

4.3.- Regresión Lineal Múltiple con SPSS

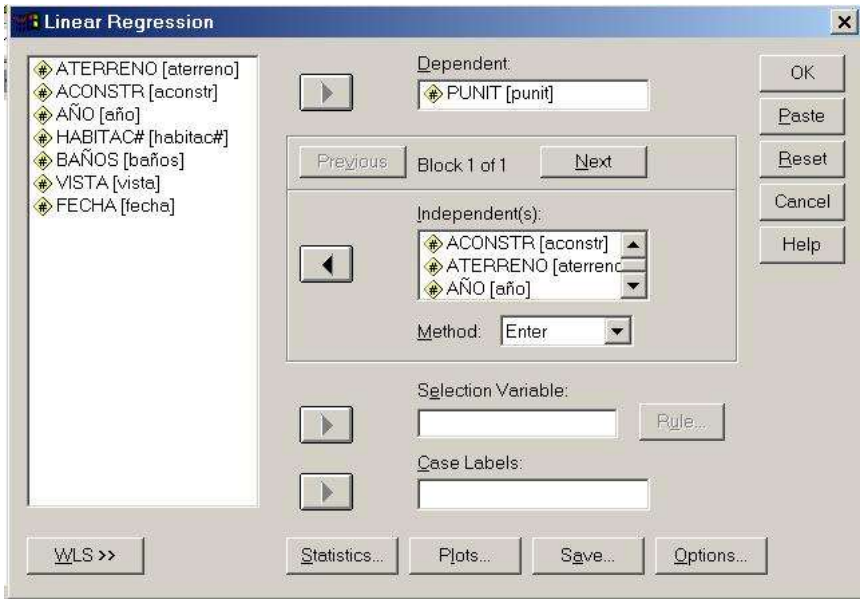
Para correr los modelos anteriores, utilizaremos la subrutina del paquete SPSS denominado “Regresión Lineal Múltiple”.

Desde el Editor de Datos (DATA EDITOR), accionaremos los siguientes comandos:

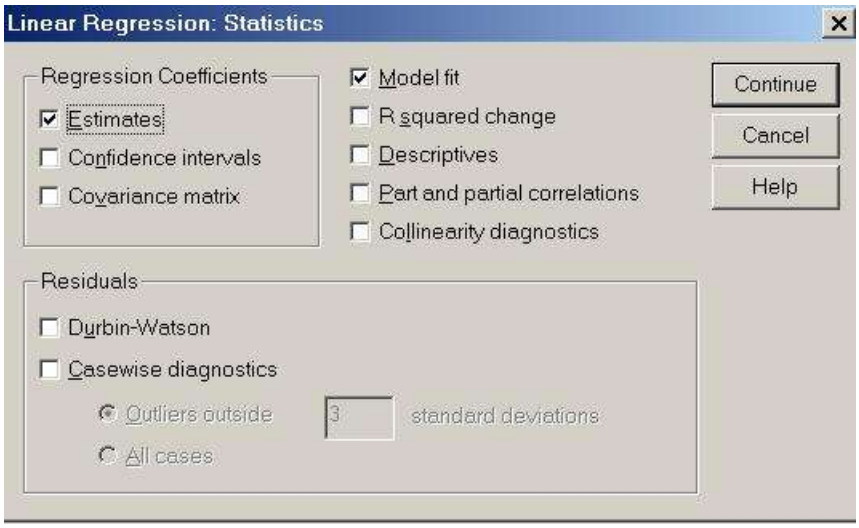
Analize
Regresión
Lineal

Y se activará la caja de diálogo correspondiente al módulo de regresión lineal múltiple.

Se selecciona como variable dependiente PUNIT y las variables independientes señaladas en la “Serie 1”:



Hacer Clic sobre el botón “Estadísticas” (STADISTICS) y configurar la caja de diálogo de la siguiente forma:



Clic en el botón “Contínue” y el SPSS lo devolverá al menú principal de regresión múltiple lineal y Clic en el botón “OK”.

El software correrá la regresión de la ***Serie 1*** y la salida del Resumen del Modelo (MODEL SUMMARY) será:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.986[a]	.972	.962	1.0E+07

a. Predictors: (Constant), FECHA, ATERRENO, ACONSTR, AÑO, VISTA

De igual manera se correrá la ***Serie 2***; y su resultado será:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.986[a]	.972	.962	1.0E+07

a. Predictors: (Constant), FECHA, ATERRENO, HABITAC#, AÑO, VISTA

Repetimos el procedimiento para la ***Serie 3***:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.986[a]	.973	.963	1.0E+07

a. Predictors: (Constant), BAÑOS, ATERRENO, FECHA, AÑO, VISTA

4.4.- Resumen de los Resultados:

Serie	R^2	R^2 adj.	Variable
Serie 1	0.972	0.962	ACONSTR
Serie 2	0.972	0.962	HABITAC#
Serie 3	0.973	0.963	BAÑOS

Como se puede observar, el modelo que mejor explica el fenómeno es la Serie 3; por lo tanto la variable independiente BAÑOS, es la queda en la regresión múltiple y las otras dos (ACONSTR y HABITAC#) saldrán. Quedando de esta manera comprobada la hipótesis planteada en el punto anterior.

5.- Conclusión

5.1.- Salida del software

El modelo de regresión que explica el comportamiento de los precios unitarios de casas en el suroeste de Caracas será:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.2E+10	1.1E+09		-10.632	.000
	ATERRENO	167660.8	29215.32	.261	5.739	.000
	AÑO	4907495	453633.5	.582	10.818	.000
	VISTA	4.6E+07	1.0E+07	.310	4.480	.001
	FECHA	54085.88	18519.80	.171	2.920	.011
	BAÑOS	1735790	3759077	.024	.462	.651

a. Dependent Variable: PUNIT

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5.4E+16	5	1.1E+16	99.172	.000[a]
	Residual	1.5E+15	14	1.1E+14		
	Total	5.5E+16	19			

a. Predictors: (Constant), BAÑOS, ATERRENO, FECHA, AÑO, VISTA

b. Dependent Variable: PUNIT

Fo = 2.39

F >> Fo

5.2.- Modelo

El modelo de regresión lineal múltiple, quedará de la siguiente forma:

$$y = -1.12 * 10^{10} + 167660.8 * X1 + 4907495 * X2 + 4.6 * 10^{7} * X3 + 54085.88 * X4 + 17357.9 * X5$$

Donde:

X1: Area del terreno
X2: Año de construcción de la casa
X3: Vista a la ciudad de caracas
X4: Fecha de protocolización
X5: Números de baños

Revisión: Febrero-2011