SEGUNDA PARTE:

ANALISIS DE VARIABLES MULTIPLES

Por:

Ing. Roberto Piol Puppio

E-Mail: rpiol@yahoo.com

Website: www.rpiol.com

CONTENIDO

I INTRODUCCION

1.0 En la práctica se observa que existe una relación entre dos o más variables, como por ejemplo la relación que existe entre el área de los terrenos y sus respectivos precios unitarios.

2.0 Lo ideal sería expresar esta relación mediante una expresión matemática, es decir hallar una ecuación que ligue las variables. Por lo tanto el problema reside en encontrar un modelo que se ajuste lo mejor posible a la muestra seleccionada.

3.0 Una vez encontrada la ecuación de la curva o modelo que más ajusta los datos obtenidos, se deberá calcular por algún modo una medida que indique la bondad del ajuste de la curva.

4.0 Sin embargo, la decisión del valor más representativo de una muestra de datos, está basada sobre la relación existente entre los valores que se conocen y los valores que se van a estimar, esto se conoce como “Estudio de Correlación”.

5.0 Se define como Regresión al estudio de la fuerza, consistencia o grado de asociación de la correlación de n variables independientes. El Análisis de Regresión determina la naturaleza de la correlación y permite realizar la correspondiente Predicción.

II CONCEPTOS BASICOS:. ANALISIS DE REGRESION SIMPLE

1.0 El problema de ajustar una curva a una serie de datos, consiste en primer término determinar la Familia de Curvas que mejor describe el fenómeno. Posteriormente realizada esta decisión se procederá a encontrar los parámetros de la curva correspondiente.

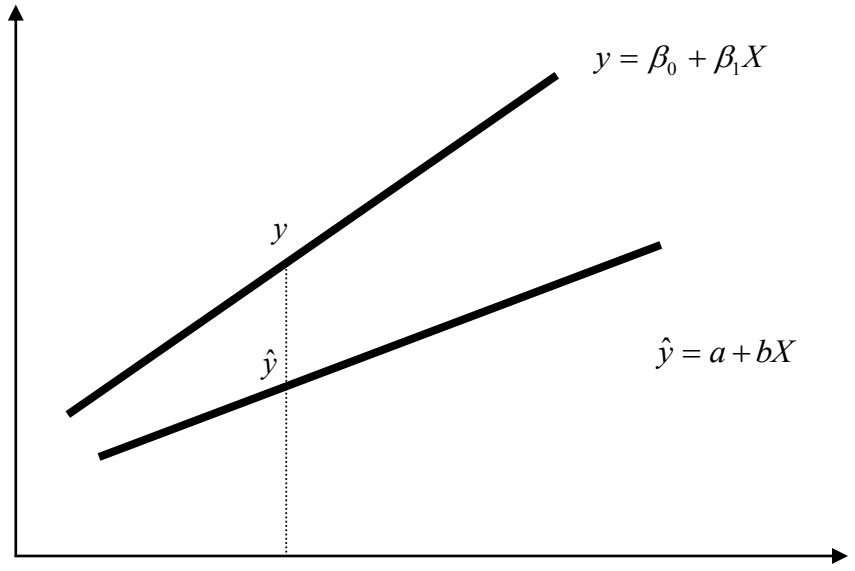
2.0 El Análisis de los Mínimos Cuadrados

2.1 En la siguiente gráfica se ha dibujado una curva (una línea recta en este caso) de una familia de curvas preseleccionadas y un grupo de datos.

2.2 Se han medido la diferencia entre la ordenada de cada punto y la función.

2.3 Una forma de seleccionar la curva que mejor representa el grupo de puntos, es elegir aquella que para la cuál sea menor el promedio de las diferencias de las ordenadas. Otra forma sería en hacer que tenga mínima la suma de las diferencias, tomadas en valor absoluto.

2.4 por lo tanto el Método de Ajuste de los Mínimos Cuadrados consiste en determinar los parámetros de una curva, de manera que la suma de los cuadrados de las diferencias mencionadas sea la menor posible.



3.0 LA RECTA DE REGRESION MINIMO CUADRATICA

3.1 El tipo mas sencillo de curva de aproximación en la línea recta cuya ecuación puede escribirse:

$$Y = a + b * X$$

3.2 La recta de aproximación por mínimos cuadrados del conjunto de puntos (x1,y1), (x2,y2)...(xn,yn) tienen las ecuaciones normales siguientes:

$$1.\sum Y = a * N + B * \sum X$$
$$2.\sum XY = a * \sum X^{2} + B * \sum X$$

3.3 Estas ecuaciones representan que la Suma del cuadrado de las desviaciones es mínima y se obtienen haciendo la primera derivada con respecto a A y la primera derivada con respecto a B igual a cero en la ecuación de la curva (recta) de mínimo cuadrado:

$$\sum (yi - a - b * xi)^2 = 0$$

$$(y - \hat{y})^2 = 0$$
$$\sum (y - \hat{y})^2 \equiv 0$$

$$\sum (y - a - b * x)^2 = 0$$

$$H_{(A,B)} = \sum (y - a - b * x)^2 = 0$$

$$\begin{cases} \dfrac{\partial}{\partial a} H_{(a,b)} = 0 \\ \dfrac{\partial}{\partial b} H_{(a.b)} = 0 \end{cases}$$

3.4 Resolviendo el sistema de ecuaciones en derivadas parciales anterior, se despejan los parámetros A y B de donde se obtienen sus respectivos valores:

$$a = \frac{\sum y \sum x^2 - \sum x \sum xy}{n * \sum x^2 - (\sum x)^2}$$

$$b = \frac{n * \sum xy - \sum x \sum y}{n * \sum x^2 - (\sum x)^2}$$

***EJEMPLO**:*

Se quieren actualizar una serie de valores (Precios Unitarios de Terrenos) en un período de tiempo de 18 meses a fin de calcular (predecir) cuál será el precio unitario (Bs/M2) en el futuro. Para eso se analizaron los libros de Registro del Municipio Autónomo correspondiente y se obtuvieron la siguiente serie de datos:

0	4	7	10	14	18	**x (meses)**
10.00	27.50	27.50	40.00	60.00	57.50	**y (Bs/M2)**

X: o sea la Variable Independiente, representa el tiempo transcurrido en meses desde la primera operación de compra venta hasta la más reciente (18 meses más tarde).

Y: o sea la Variable Dependiente, representa el precio unitario en Bs/M2 correspondiente a cada operación revisada.

n	x	y	x^2	xy
1	0	10.00	0	0.00
2	4	27.50	16	110.00
3	7	27.50	49	192.50
4	10	40.00	100	400.00
5	14	60.00	196	840.00
6	18	57.50	324	1,035.00
Sumatoria:	53	222.50	685	2,577.50

$$N = 6$$

$$a = \frac{(222.50 * 685) - (53 * 2{,}577.50)}{(6 * 685) - (53)^2} = 12.15$$

$$b = \frac{(6 * 2{,}577.50) - (53 * 222.50)}{(6 * 685) - (53)^2} = 2.82$$

Por lo tanto la ecuación de Correlación de la línea mínimo cuadrática de mejor ajuste será:

$$y = 12.15 + 2.82 * x$$

Ahora se puede predecir cuál será el comportamiento de la Variable Dependiente y (Precio Unitario) en función de la variable independiente x (Tiempo).

Si se desea saber cuál será el valor esperado a los 20 meses de haberse hecho la primera observación (o sea la fecha del avalúo), se obtendrá para X = 20

$$y = 12.15 + 2.82\ (20) = 68.57\ [Bs/M2\ ]$$

4.0 LA CURVA DE REGRESION EXPONENCIAL

La familia de rectas (y =a + b x) y las familias de curvas exponenciales ($y = a * b^x$), son las ecuaciones de correlación simple más utilizadas en la práctica.

4.2 Sin embargo se verá más adelante, el estudio de los métodos computarizados para la obtención de la familia de curvas de mejor ajuste en otros familias modelos también aplicables.

4.3 En este caso para correlacionar la muestra de datos obtenidas se estudiará una Ecuación Exponencial cuya expresión es:

$$y = a * b^x$$

4.4 Resolviendo el sistema de sus ecuaciones normales se obtienen las siguientes expresiones para los coeficientes a y b:

$$a = ant\log \frac{\sum \log y \quad \sum x^2 - \sum x \quad \sum x * \log y}{n * \sum x^2 - (\sum x)^2}$$

$$b = ant\log \frac{n * \sum x * \log y \quad - \sum x \quad \sum \log y}{n * \sum x^2 - (\sum x)^2}$$

EJEMPLO

En un caso similar al ejemplo anterior; se han obtenido el registro de operaciones de compra-venta de terreno en los últimos 20 meses:

x MESES	y Bs/M2
0	10
5	15
8	15
10	30
14	35
15	50
18	70
20	80

*En este caso **x** (la Variable independiente) seguirá siendo el tiempo (MESES) ý **y** (la variable dependiente) el Precio Unitario (Bs/M2).*

n	x	y	log y	x^2	x*log y
1	0	10	1.0000	0	0.0000
2	5	15	1.1761	25	5.8805
3	8	15	1.1761	64	9.4087
4	10	30	1.4771	100	14.7712
5	14	35	1.5441	196	21.6170
6	15	50	1.6990	225	25.4846
7	18	70	1.8451	324	33.2118
8	20	80	1.9031	400	38.0618
Sumatoria:	90	305	11.8205	1,334	148.4355

$$n = 8$$

$$\log A = \frac{(11.8205)*(1{,}334) - (90)*(148.4355)}{(8)*(1{,}334) - 90^2} = 0.9367$$

$$\log B = \frac{(8)*(148.4355) - 90*(11.8205)}{(8)*(1{,}334) - 90^2} = 0.0481$$

PERO AUN FALTAN CALCULAR LOS ANTILOGARITMOS

$$a = Antlog\ (0.9367) = 8.6437$$

$$b = Antlog\ (0.0481) = 1.1171$$

La ecuación de correlación será:

$$y = 8.6437 * 1.1171^{x}$$

En este ejercicio no solo se podrá predecir el valor unitario del terreno a la fecha del avalúo, sino también se podrá interpolar para meses en que no han existido operaciones de compra-venta o cualquier mes seleccionado:
Por ejemplo se podrá obtener el precio unitario para:

a) *Interpolar el valor unitario a los 12 meses después de la fecha de origen*

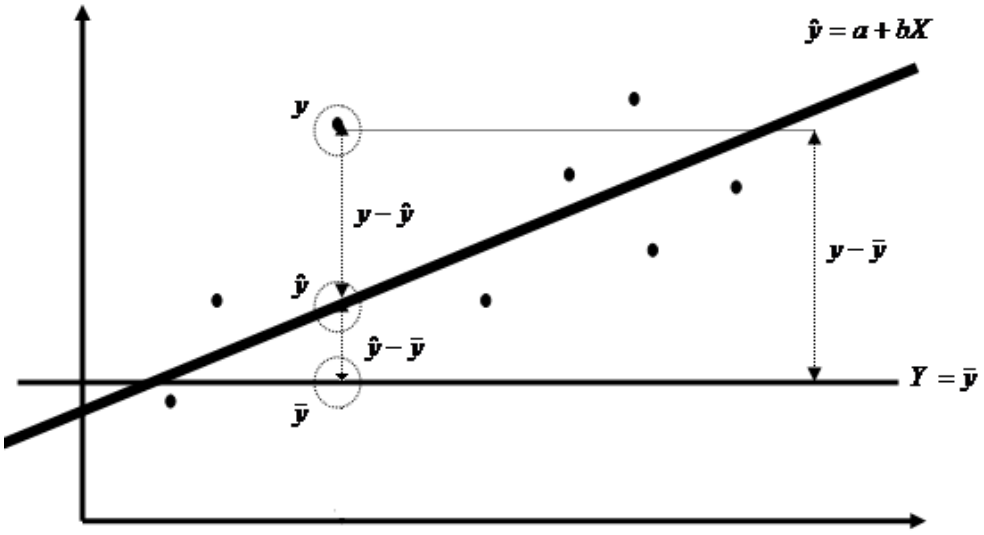
b) *ídem para 17 meses*

c) *Predecir el valor unitario a los 22 meses*

	x (meses)	y (Bs/M2)
a)	12	32.64
b)	17	56.79
c)	22	98.79

5.0 EL COEFICIENTE DE DETERMINACION

5.1 El Coeficiente de Determinación, mide la bondad del ajuste relativo de la curva de regresión. Indica la cantidad de variación en Y que se explica en la ecuación de regresión.



5.2 Desviación Total de Y

Es la diferencia entre el valor observado (datos) y el promedio de los valores observados:

$$DesviaciónTotal = y - \bar{y}$$

5.3 Desviación No Explicada

Corresponde al Error o Residual y se define como la diferencia entre el valor observado y el valor calculado:

$$Desviación\ No\ Explicada = y - \hat{y}$$

5.4 Desviación Explicada

Corresponde a la diferencia entre el valor calculado y el valor promedio:

$$Desviación\ Explicada = \hat{y} - \bar{y}$$

5.5 Relación entre los términos anteriores

Se cumplirá que:

Desviación Total = Desv. No Explicada + Desv. Explicada

$$(y - \bar{y}) = (y - \hat{y}) + (\hat{y} - \bar{y})$$

5.6 Dentro de la Teoría de los Mínimos Cuadrados que estamos utilizando, considerando que se eleven al cuadrado cada una de las desviaciones y sumando todos los valores correspondientes a los N datos u observaciones, se obtienen los siguientes Estadísticos:

a) **SCT** o Suma de Cuadrados Total

$$\sum (y - \bar{y})^2$$

b) **SCE** o Suma del Cuadrado del Error

$$\sum (y - \hat{y})^2$$

c) **SCR** o Suma del Cuadrado de la Regresión

$$\sum (\hat{y} - \bar{y})^2$$

5.7 De la misma manera anterior, se cumple la relación:

SCT = SCE + SCR

5.8 El Coeficiente de Determinación:

Se define como coeficiente de determinación:

$$R^2 = \frac{SCR}{SCT}$$

DESPEJANDO:

$$R^2 = 1 - \frac{SCE}{SCT}$$

DONDE EL COEFICIENTE DE DETERMINACION TOMA VALORES COMPRENDIDOS EN EL INTERVALO: **[0 , 1]**

5.9 Interpretación del Coeficiente de Determinación:

Un valor de $R^2 = 0.75$, debe interpretarse que el 75% de las variaciones de y, son explicadas por las variables y número de datos utilizados para calcular el modelo.

Se preferirá siempre el Modelo cuyo Coeficiente de Determinación sea lo más cercano a la unidad (1.00).

EJEMPLO:

Sean los siguientes datos correspondientes al ejemplo anterior:

x MESES	y Bs/M2
0	10
5	15
8	15
10	30
14	35
15	50
18	70
20	80

CALCULO DEL MODELO DE CORRELACION EXPONENCIAL:

$$y = 8.6437 * 1.1171^x$$

n	x	y	$\hat{y}$	$(\hat{y}-\bar{y})^2$	$(y-\bar{y})^2$
		Observada	Calculada	SCR	SCT
1	0	10	8.64	869.44	791.30
2	5	15	15.04	533.29	535.00
3	8	15	20.96	294.73	535.00
4	10	30	26.16	143.31	66.10
5	14	35	40.74	6.80	9.80
6	15	50	45.51	54.42	140.90
7	18	70	63.44	640.55	1,015.70
8	20	80	79.17	1,683.98	1,753.10
Sumatoria:				4,226.52	4,846.88

$$\bar{y} = 38.13 \quad (Bs./M2)$$

$$R^2 = \frac{4{,}226.52}{4{,}846.88} = 0.8720$$

$$R^2 = 87.20\%$$

Recalculando la misma data, pero esta vez suponiendo que el modelo de Correlación es Lineal, se obtiene:

CALCULO DEL MODELO DE CORRELACION LINEAL:

$$y = -2.5972 + 3.61975 * x$$

Y su correspondiente Coeficiente de Determinación:

$$R^2 = 0.86911$$

$$R^2 = 86.91\%$$

6.0 El Estadístico F (Test de Fischer)

El estadístico F corresponde una prueba o hipótesis para rechazar o aceptar la predicción de la correlación y así como el Coeficiente de Determinación nos ayuda a decidir entre varias curvas de regresión, el estadístico F nos dirá si los datos y variables tomadas son significativas o no; y es la forma de validar la ecuación o modelo de correlación.

Es precisamente el Estadístico F, quien indica la cantidad de datos o variables mínimas que se requieren para que la Regresión exista.

El Estadístico F, se compara con el valor de "F de prueba" (Fo), el cual se determina en la tabla que se anexa.

El valor de F será grande, cuando la regresión es significativa y obligatoriamente deberá ser mayor que Fo para que el modelo sea válido.

Si F es menor que Fo, deberán reestudiarse los datos ya que los datos y variables seleccionadas, no son suficientes o significativas para calcular un modelo de regresión que pueda predecir el comportamiento de la variable dependiente con relación a la independiente.

Cálculo del Estadístico F:

$$F = \frac{\frac{SCR}{k}}{\frac{SCE}{n-(k+1)}}$$

DONDE:

k = Nro. de variables independientes
n = Nro. de observaciones

EJEMPLO:

En el ejemplo anterior, vamos a proceder a validar el modelo, el único dato faltante para calcula el Estadístico F, es SCE, sin embargo es fácilmente deducible partiendo de la relación:

$$SCT = SCE + SCR$$

$$SCT = 4{,}846.88$$

$$SCR = 4{,}226.52$$

$$SCE = SCT - SCR$$

$$SCE = 620.36$$

k = 1 (NRO: DE VARIABLES INDEPENDIENTES)

n = 8 (NRO: DE OBSERVACIONES)

$$F = \frac{\frac{4{,}226.52}{1}}{\frac{620.36}{6}} = 40.88$$

EN LA TABLA ANEXA SE PUEDE OBSERVAR QUE:

PARA: $k = 1$

Y PARA: $n - (k + 1) = 6$

SE OBTIENE: $Fo = 5.99$
(Para una Confianza del 95%)

40.88 >> 5.99

F > Fo *POR LO TANTO SE VALIDA LA REGRESION PARA UNA CONFIANZA DEL 95%*

Puntos Porcentuales de la Distribución F
Intervalo de Confianza: 95%

k (Número de Variables Independientes)

n - (k+1) Grados de Libertad

•	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	⟹
1	161.40	199.50	215.70	224.60	230.20	234.00	236.80	238.90	240.50	241.90	243.90	245.90	248.00	249.10	250.10	251.10	252.20	253.30	254.30
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.87	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.74	2.77	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.70	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65

7.0 MULTICOLINEALIDAD: La Matriz de Correlación

El problema de Multicolinealidad se presenta cuando entre las Variables Independientes existen relaciones lineales entre algunas de ellas; es decir las Variables Independientes están relacionadas entre sí, unas dependen de las otras.

Cuando se presenta el problema de multicolinealidad entre las variables independientes, el sistema de ecuaciones normales (que permitió obtener el valor de los coeficientes a, b, c..., n de la ecuación de regresión mínimo-cuadrática) no permite obtener una solución única para cada uno de los parámetros de la función de regresión.

El problema de la multicolinealidad afecta a la descripción del modelo de regresión múltiple, ya que significa que todos los datos se encuentran sobre una misma línea recta y por lo tanto no existe un plano óptimo en el sentido mínimo cuadrático; sino los infinitos que pasan por dicha recta.

La multicolinealidad en una serie de datos se mide a través de la Matriz de Correlación. La Matriz de correlación permite conocer la tendencia y magnitud de la relación lineal o asociación entre las variables independientes. El modelo de regresión se vuelve cada vez menos confiable a medida que aumenta la correlación entre dichas variables independientes.

La Matriz de Correlación tiene las siguientes características físicas:

a) Es una Matriz Unidad: La diagonal principal de la misma es la unidad (1.00).

b) Es una Matriz Simétrica: Ambos lados de la diagonal principal son antimétricos, de tal manera que si la matriz se “doblara” por la diagonal principal coincidirán los coeficientes.

Los Coeficientes de Correlación

Los Coeficientes de Correlación indican el grado y tipo de asociación entre las variables.

a) Si el coeficiente de correlación es positivo, indica que una de las variables está directamente relacionada con la otra.

b) Si el coeficiente de correlación es negativo, indica que una de las variables está inversamente relacionada con la otra.

c) La mantisa del coeficiente de correlación indica la magnitud de la relación entre las variables. En general se puede señalar que:

Cuando: 0.00 < r <= 0.30	La correlación es débil
Cuando: 0.30 < r <= 0.75	La correlación es media
Cuando: 0.75 < r <= 1.00	La correlación es fuerte

Se define que existe Multicolinealidad entre dos variables independientes cuando la correlación entre ambas es fuerte (r > 0.75).

Para solucionar el problema de multicolinealidad, se deberá eliminar de la regresión una de las dos variables independiente correlacionadas, ya que al estar una en función de la otra no permitirá una solución aceptable de la regresión mínimo-cuadrática.

El CRITERIO para el caso de un modelo de regresión múltiple donde dos (2) variables independientes estén altamente correlacionadas entre sí; es seleccionar cual de las dos Variable Independiente es la que tiene que salir del Modelo de Regresión.

Para esto se utiliza el procedimiento estadístico denominado “ANÁLISIS FACTORIAL”; el cual trata de agrupar aquellas variables que se encuentren muy relacionadas entre sí (r > 0.75) en un único factor, bajo el criterio de que las mismas a su vez estén poco correlacionadas (r < 0.75) con el resto de las variables independientes que no estén incluidas en ese factor; de tal manera que se logre pasar de un modelo inicial de “n” variables independientes a otro modelo con “n-1” variables independientes, eliminando de esta manera una de las dos variables correlacionadas.

Para utilizar la Técnica Estadística “Análisis Factorial”, se utilizan paquetes estadísticos dedicados, como lo son el SPSS, Statgraphics, etc.

El manejo de estos paquetes estadísticos, se sale del alcance de esta Monografía; ya que en la misma se utiliza como herramienta de desarrollo la Hoja de Cálculo Microsoft Excel (Versión 6.0 o superior).

Lo realmente importante, es que no pueden convivir (2) dos variables independientes correlacionadas entre sí en un modelo de regresión; una de las dos debe salir.

***Ejemplo**:*

La hoja de cálculo Excel, generó la siguiente matriz de correlación de una muestra de datos:

	CTE	AREA	EDAD	%COND	FECHA
CTE	1.00				
AREA	0.35	1.00			
EDAD	-0.41	-0.57	1.00		
%COND	0.25	0.50	-0.45	1.00	
FECHA	-0.95	-0.25	0.32	**-0.89**	1.00

En la Matriz de Correlación se observa:

A) La Diagonal Principal es la Unidad.

B) El Software utilizado solo mostró la parte inferior de la matriz de correlación, ya que la parte superior es antimétrica a esta.

C) Solamente existe una Correlación Fuerte e Inversa (Problema de Multicolinealidad) entre las variables independientes: FECHA y %COND.

***NOTA:** No se deberán tomar en cuenta los coeficientes de la primera columna CTE (Constante), ya que solo interesa conocer la correlación entre las Variables Independientes únicamente. De hecho, algunos de los principales paquetes estadísticos no la presentan al solicitar su impresión.*

Para poder obtener una solución aceptable al modelo de correlación mínimo-cuadrática, se deberán eliminar de los datos una de las dos variables que presentan problemas de multicolinealidad.

III USO DE LA HOJA DE CALCULO EXCEL EN LOS ANALISIS DE REGRESION SIMPLE

1.0 En esta Sección se orientará al uso de los microcomputadores para la solución de problemas de regresión simple aplicado a la materia de avalúos. En ningún momento pretende ser un curso de computación ya que únicamente se expondrán los métodos frecuentemente usados. El alumno deberá aplicar por su cuenta, en sus equipos y corriendo en su Hoja de Cálculo Excel, que forma parte del paquete "Microsoft Office" la metodología que aquí se explica.

2.0 —La Hoja de Cálculo Excel (Versión 6.0 o superior)

Con los conocimientos aquí suministrados es posible calcular a través de las funciones de regresión de la hoja de cálculo Excel:

- Los coeficientes de correlación
- El Estadístico F
- El Coeficiente de Determinación (R^2)
- SCR
- SCE
- Errores Estándar
- Otros factores

Lo importante está en ser cuidadosos en la entrada de los datos y obtener una salida presentable que sirva como anexo al avalúo donde se aplique este procedimiento.

***EJEMPLO**:*

Se desea obtener el valor de una vivienda rural de 80 M2 y 20 años de construida.

En la correspondiente Oficina de Registro Inmobiliario, se obtuvieron los siguientes grupos de datos correspondientes a viviendas rurales ubicadas en el mismo asentamiento agrícola:

REF	AREA M2	EDAD AÑOS	PRECIO Bs
A	80.00	10	100,000.00
B	80.00	13	80,000.00
C	80.00	16	64,800.00
E	80.00	19	50,000.00
D	80.00	24	41,000.00
F	80.00	30	32,500.00

Salta a la vista que debe existir una relación entre la Edad y el Precio de la vivienda rural ya que todas son de idéntica área y están ubicadas en el mismo parcelamiento. Por lo tanto obligatoriamente se debe deducir la forma en que se correlacionan ambas variables.

Se considerará como variable independiente X [años] y como variable dependiente Y [Bs.], ya que la variable "Area" es constante.

La salida de la Hoja de Cálculo Excel podrá ser parecida a la siguiente:

-3,276.90418	122,552.21130
538.90723	10,692.47725
0.90238	8,876.98709
36.97426	4.00000
2,913,604,733.82473	315,203,599.50860

Donde:

a =	122,552.21130
b =	-3,276.90418
R^2 =	0.90238
F =	36.97426
SCR =	2,913,604,733.82473
SCE =	315,203,599.50860

Viendo los resultados de la salida de la hoja de cálculo, el Modelo de Correlación Lineal será el siguiente:

Y = 122,552,211.30 - 3,276.90418 * X

Sustituyendo X=20 años (Edad del Inmueble) se obtiene:

*Y = 122,552,211.30 - 3,276.90418 * (20)* *[Bs.]*

Y = 57,014.13 ***[Bs.]***

y el Valor del inmueble será Bs. 57,014.13

3.0 El uso de los Paquetes Estadísticos en los Informe de Avalúos

La mayoría de los paquetes estadísticos son complejos, difíciles de usar, caros y la mayoría de los datos que nos suministran no nos interesa en absoluto al momento de hacer un avalúo.

Sin embargo, su utilización cada día es mayor y su versatilidad nos permite llegar en forma extremadamente rápidas a resultados.

Existen en el mercado una gran variedad de Paquetes Estadísticos mas o menos complejos para cada tipo de Sistema Operativo (DOS, Windows, McIntosh, Linux, Unix, OS2, etc.),

Sin embargo, en los últimos años, los paquetes integrados tales como MS-Office (Excel), Lotus Smart Suite (Lotus 123) y Smart Office (Q-Pro), han mejorado sus aplicaciones estadísticas, de tal forma que se han transformado en los preferidos de los usuario.

IV USO DE LA HOJA DE CALCULO EXCEL EN LOS ANALISIS DE REGRESION MULTIPLE

1.0 La mayoría de los casos en la vida real, para poder predecir la variación de una variable, no se hace en función de una sola variable independiente (Precio Unitario vs. Area, por ejemplo); sino mas bien son VARIAS las variables que son necesarias para predecir un comportamiento o fenómeno.

2.0 En este caso solamente se estudiará el caso de REGRESION LINEAL MULTIPLE[1], es decir una variable estará explicada en función de otras en forma lineal:

$$Y = A + B\ X1 + C\ X2 + D\ X3 + ... + M\ Xn$$

[1] También se estudiará la Regresión Logarítmica Múltiple de la forma:

$$y = a * b^{X1} * c^{X2} * ...m^{Xn}$$

la cual puede ser linealizada y resuelta como un caso especial de la Regresión Lineal Múltiple:

$$\log y = \log a + X1 * \log b + X2 * \log c + ...Xn * \log m$$

Sustituyendo, la función se tranforma en:

$$y' = a' + b' * X1 + c' * X2 + ...m' * Xn$$

Que es una función lineal, cuya solución es:

$$y = e^{\log a + X1 * \log b + X2 * \log c + ...Xn * \log m}$$

3.0 Como ejemplo, en el caso del avalúo de un apartamento, se deberían considerar las siguientes variables:

Como "Variable Dependiente":

Y: PRECIO UNITARIO DE APARTAMENTOS COMPARABLES

Como "Variables Independientes":

X1: AREA EN M2 DE LOS REFERENCIALES
X2: NUMERO DE HABITACIONES POR APARTAMENTO
X3: NUMERO DE PUESTOS DE ESTACIONAMIENTO
X4: EDAD DE LOS REFERENCIALES
X5: PORCENTAJE DE CONDOMINIO QUE PAGA CADA REFERENCIAL
X6: PISO EN QUE SE UBICA EL APARTAMENTO
X7: NUMERO DE APARTAMENTOS QUE TIENE EL EDIFICIO
Y quizás otras no menos importantes

y así sucesivamente se pueden estudiar las todas las diferentes variables que son posibles de medir u obtener, que ayudarían a explicar el fenómeno, que en este caso sería LA VARIACION DEL PRECIO UNITARIO DE APARTAMENTOS.

4.0.- La metodología que se utiliza en la correlación lineal múltiple es similar o más bien la misma que la que hemos estudiado en la correlación lineal simple. La dificultad está en obtener los parámetros del modelo, la cual sin el computador u ordenador, se hace muy engorroso o prácticamente imposible cuando se superan las tres variables independientes, ya que habría que resolver el sistema de ecuaciones normales a través de matrices y determinantes.

***NOTA 1**: Se explicará la implementación de las Técnicas de Regresión Múltiple, como "Metodología Valuatoria", por la vía del ejemplo.*

***NOTA 2**: Se utilizará en esta monografía, la Hoja de Cálculo de Uso General: Microsoft Excel, la cual, entre su funciones estadísticas, posee: Regresión Lineal Múltiple y Regresión Exponencial Múltiple.*

5.0.-Implementación de las Técnicas de Regresión Múltiple como Métodología Valuatoria

Es común observar una relación entre dos o más variables cuando se analizan una serie de “Inmuebles Referenciales” para una zona o región determinada.

Por ejemplo, analizando los Precios Unitarios y las Areas de Terreno; en estas dos variables parece existir una relación inversa de proporcionalidad; ya que aparentemente: A mayor área de terreno, se observa menor precio unitario.

Lo ideal sería expresar estas relaciones mediante una expresión algebraica que sea capaz de interrelacionar las variables entre sí. Sin embargo, es casi imposible encontrar una función que se ajuste perfectamente a la serie de datos estudiados, por lo tanto se deberá buscar el “Modelo de Mejor Ajuste” que indique la tendencia de las diferentes variables consideradas en una Serie.

Se deberá entonces acudir a Métodos Estadísticos complejos, a fin de poder determinar la Ecuación o Modelo que permitirá obtener “La Tendencia” en términos generales de una Serie de Datos, en virtud del incremento o disminución que tendrá una variable en función de la otra u otras.

Estos Métodos Estadísticos, entre otros son:

- La Regresión Simple: Trata de correlacionar dos (2) Variables (una Dependiente y una Independiente)
- La Regresión Múltiple: Trata de correlacionar Una (1) Variable Dependiente y “n” Variables Independientes.

6.0.- Reglas Básicas en la implementación de las técnicas de Regresión Múltiple, como Proceso Valuatorio

6.1.- Se considerará siempre como Variable Dependiente, el Precio Unitario (sin corregir) de una serie de referenciales, y deberá siempre estar expresada en Bs/M2 (Unidad Monetaria / Area).

6.2.- Las Variables Independientes numéricas, tales como el área del terreno, el área de construcción, la edad del inmueble etc., podrán ser enteradas libremente en las ecuaciones de correlación.

6.3.- Otras Variables, que no puedan ser expresadas algebraicamente tal como el tiempo transcurrido entre la protocolización y la fecha del avalúo, deberán ser transformada a una expresión numérica; una vez obtenida la expresión numérica podrán ser enteradas en las ecuaciones de correlación.

***NOTA IMPORTANTE**: Las variables No Numéricas o Cualitativas, no pueden formar parte de la Regresión, ya que para poder transformarla en "Variables Numéricas", habría que recurrir a establecer un criterio (casi siempre empírico) desvirtuando así la técnica eminentemente objetiva, donde no entra para nada el criterio del Profesional Tasador.*

6.4.- Se presenta a veces el caso, de que no es posible obtener todas las variables de un referencial por diversas razones, siendo principalmente: La información incompleta del inmueble en el Documento Protocolizado en la Oficina de Registro Inmobiliario. En estos casos, alguno de los Software Estadístico podrá generar automáticamente la predicción de la variable o variables faltantes, permitiendo continuar el proceso de correlación.

***NOTA IMPORTANTE**: En caso de utilizar MS-Excel o un paquete similar, habría que descartar al Referencial o Comparable, ya que en una Hoja de Cálculo, si una celda está "en blanco", la asumiría como cero (0.00). Cosa que afectaría el resultado de la regresión.*

6.5.- Preparación y entrada de los datos a correlacionar: Hay que tener especial cuidado en la trascripción de los datos dentro de la hoja de cálculo o Programa Estadístico. Se ha comprobado que la mayoría de las veces los errores ocurren por una o varias equivocaciones en la trascripción de la data.

EJEMPLO:

1.- Descripción General del Inmueble

El inmueble objeto de este avalúo está representado por una casa identificada como 6-3, Manzana 6, que forma parte del Sector identificado como "Aragua", ubicado en el Conjunto Residencias Venezuela, Urbanización Coche. Caracas

2.- Area de Terreno y Construcción:

Según el Documento de Propiedad y las mediciones efectuadas en el propio inmueble:

AREA APROXIMADA:

TERRENO:	650,00	M2
CONSTRUCCION:	260,00	M2

3.- Referenciales o Comparables de casas en la Urbanización Coche:

REF	DIRECCION	CASA #	VENDEDOR / COMPRADOR	PRECIO Bs.	AREA TERRENO M2	AREA CONSTR. M2
A	AV.INTERCOMUNAL	NP2-2-2	C.MICHEL J.USECHE	72.000	540	280
B	CONJ.PQUE.EL VALLE	12-2	L.TORRES M.RIVAS	56.000	740	200
C	AV.SUCRE MZ.D	504-A	M.YANES I.HIDALGO	77.000	890	320
E	CONJ.LOS SAMANES	D-32-E	O.MONSALVEA.ALDAN	68.000	800	280
D	CONJ.LA FLORESTA	A-42	G.PARRA F.SERRANO	90.000	620	400
F	URB.RADIOCARACAS	0905	I.CABALLEROR.NOVELL	44.000	700	120

4.- Entrada de las data en la Hoja de Cálculo:

REF	Y	X1	X2
A	72.000	540	280
B	56.000	740	200
C	77.000	890	320
E	68.000	800	280
D	90.000	620	400
F	44.000	700	120

5.- Salida de la Función de Regresión Múltiple Lineal:

165.372261	-6.29879967	28237.7054
6.46513335	5.00378553	4112.36483
0.99549555	1397.56268	#N/A
331.503709	3	#N/A
1294973789	5859544.34	#N/A

6.- Interpretación de los Resultados:

a =	28237.7054
b =	-6.29879967
c =	165.372261
R^2 =	0.99549555
F =	331.503709
SCR =	1294973789
SCE =	5859544.34

7.- Cálculo de la Matriz de Correlación:

	Cte.	X1	X2
Cte.	1	-	-
X1	-	1	-
X2	-	-0.07	1

8.- Interpretación de la Matriz de Correlación:

No existe problemas de MULTICOLINIALIDAD entre las Variables Independientes: Area de Terreno (X1) y Area de Construcción (X2).

9.- Cálculo de Fo para la validación del Modelo de Regresión Lineal Múltiple:

k =	2
n-(k+1) =	3
Fo =	9.55

Utilizando la tabla de Puntos de Porcentaje de la Distribución F para una confianza del 95%, que se anexa a este informe, obtenemos el valor de Fo = 9.5 Valor que satisface el CRITERIO:

F >> Fo

10.- Conclusión: Se concluye que existe una regresión conjunta entre las variables incluidas en el Modelo y por lo tanto se puede afirmar que:

Modelo de Correlación Múltiple Lineal:

y = 28,237.7054 - 6.29879967 * X1 + 165.372264 * X2

Sustituyendo:

X1=	650.00	M2
X2=	260.00	M2

VALOR DE LA CASA (y) = 67,140.27 Bs.

7.0 Aplicación General de la Metodología de Correlación Múltiple

7.1 En el ejemplo anterior, se obtuvo un Coeficiente de Determinación Alto, permitiendo lograr la solución del Modelo de Correlación Múltiple Lineal.

7.2 Sin embargo, en el campo de la valuación de inmuebles, la realidad es otra; debido a la alta dispersión de los datos referenciales obtenidos y a la falta de sinceridad en la Protocolización de los Documentos de Compra-Venta, es poco probable obtener un Coeficiente de Determinación alto al aplicar esta Metodología en la primera corrida.

7.3 Para tratar de solventar este problema, se ha establecido un procedimiento que permite determinar cuál es el problema que impide que exista la cohesión entre los datos referenciales. Este procedimiento se puede enunciar de la siguiente manera:

- Determinar cuál es el modelo de mejor ajuste
- Determinar la existencia de Multicolinealidad entre las Variables Independientes
- Determinar la existencia de Valores Atípicos
- Validar la Regresión

7.3.1 Determinación del Modelo de Mejor Ajuste.

La mayoría de las Hojas de Cálculo, Paquetes Estadísticos y algunas calculadoras científicas tienen la opción de ofrecer varios modelos o familias de curvas; pero las mismas se limitan al caso de correlación simple únicamente.

Para el caso de Correlación Múltiple, la situación es invertida: Muy pocos softwares permiten el estudio de Correlación Múltiple No Lineal (de manera directa).

Quizás la única Hoja de Cálculo que tiene un modelo de regresión múltiple exponencial, además del modelo lineal, es el MS-EXCEL versión 6.0 o superior.

Conocido lo anterior, es muy poco o nada lo que pueda hacerse sin contar con varios modelos de regresión múltiple en función de buscar el modelo de regresión que mejor se ajuste a los datos, o sea el que posea un Coeficiente de Determinación significativo.

7.3.2 Determinación de la existencia de Multicolinealidad entre las Variables Independientes.

El caso de la Multicolinealidad, se estudió con detalle en las páginas anteriores. Para el caso de Correlación Múltiple, la aplicación de la Matriz de Correlación, permite determinar la existencia de Variables Independientes que están en función de otras, obligando a la eliminación de una de las variables correlacionadas.

Es importante de señalar, que la existencia de Multicolinealidad entre Variables Independientes, debe verificarse, aún si el Coeficiente de Determinación del Modelo de Regresión Múltiple sea cercano a 1.0, ya que este hecho no necesariamente implica la inexistencia de problemas de Multicolinealidad en la regresión.

7.3.3 Determinación de la existencia de Valores Atípicos.

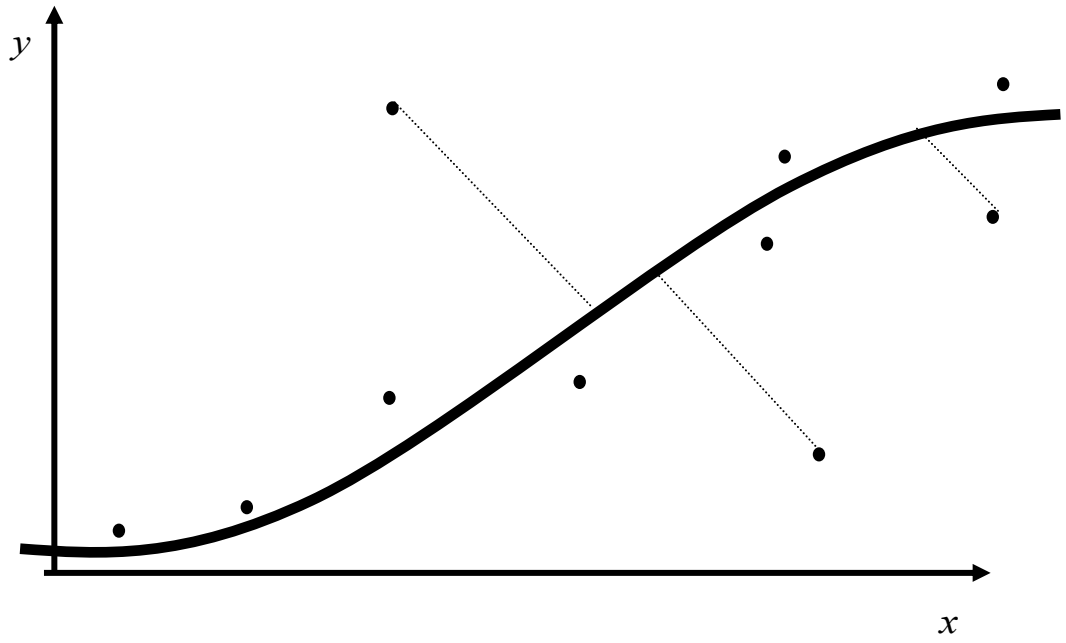
Se definen como “Valores Atípicos”, aquellos valores que no perteneciendo a la serie estudiada, forman parte de la muestra recolectada.

En un sistema de registro inmobiliario insincero u ofertas engañosas de la prensa inmobiliaria especializada, donde una gran cantidad de operaciones de compra-venta de inmuebles no están sujetas a la realidad, es muy común la presencia de “Valores Atípicos” en la Serie de datos referenciales estudiada.

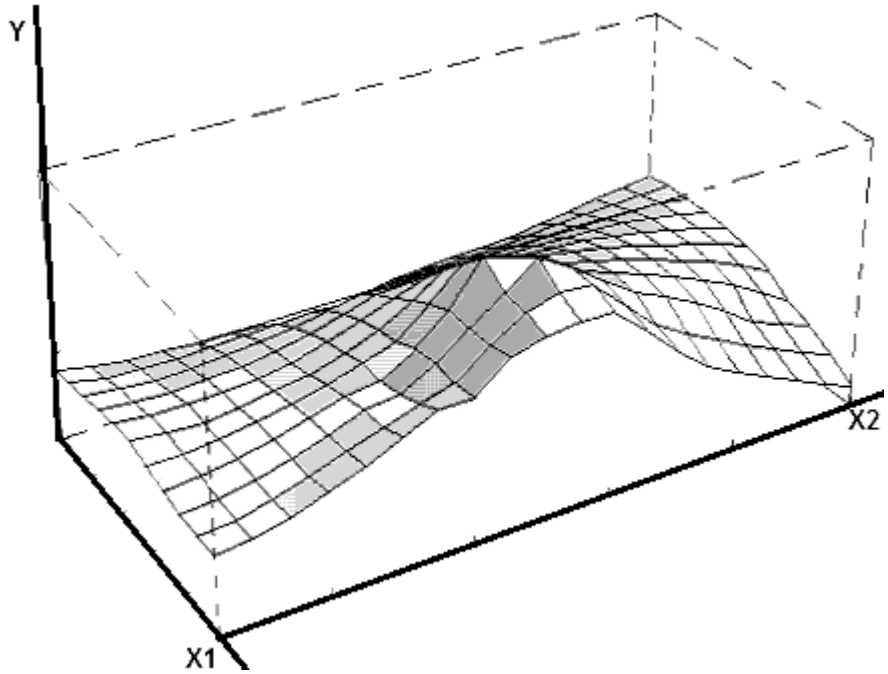
En la estadística de Regresión, se define como “Residuo” o “Residual”, a la diferencia entre los valores observados en la serie y los valores calculados o estimados de la regresión:

$$R = y - \hat{y}$$

Gráficamente, para un Modelo de Correlación Simple, se puede observar que existen valores (x , y) muy cercanos a la curva de regresión, mientras que otros están muy alejados.



En el caso de Correlación Múltiple, donde no se habla de curvas de regresión, sino mas bien de Planos de Regresión, si se correlacionan Tres (3) variables; es muy difícil representar gráficamente los Valores Observados en relación con el Plano de Regresión para sistemas de Tres Variables:



Sin embargo, es imposible la representación gráfica cuando existen mas de tres variables, ya que estaríamos fuera del espacio convencional; por eso se habla de Hiperplanos de Regresión, que aunque no pueden ser representados gráficamente (o físicamente), existen matemáticamente.

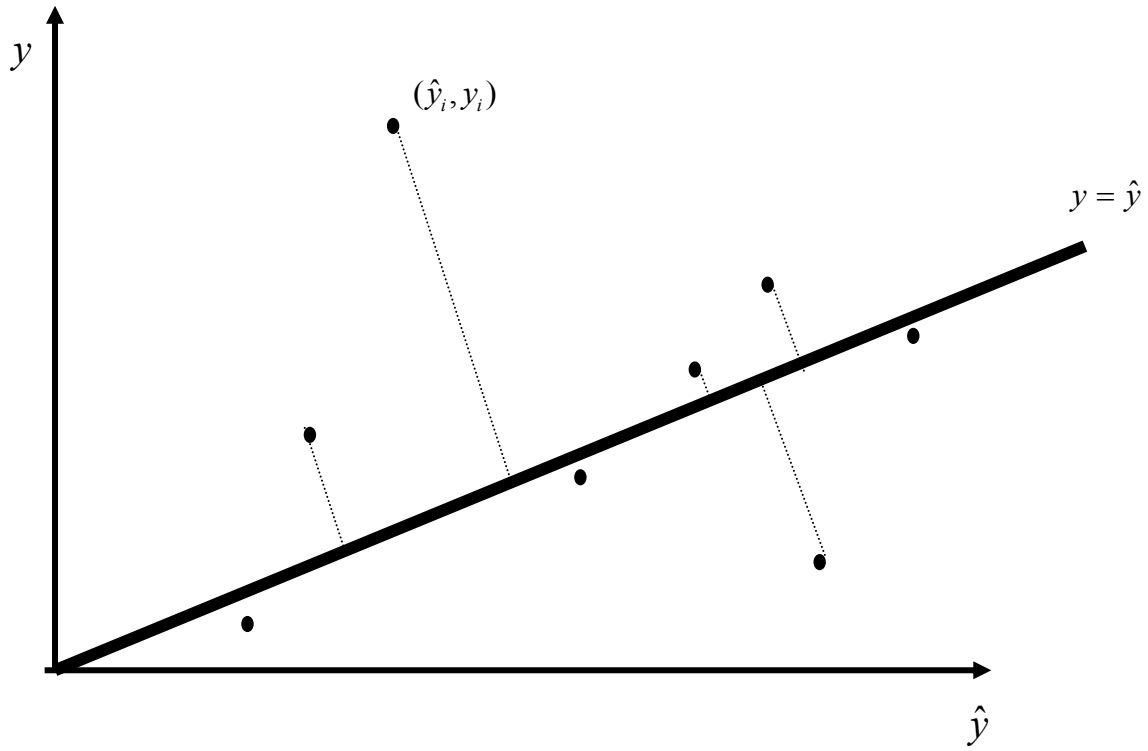
La representación de los valores atípicos en planos o hiperplanos de correlación, es posible de visualizar mediante el siguiente procedimiento:

a) Se define el Eje de las Ordenadas (Y) para representar los valores observados (y), (En el caso de avalúos: el Precio Unitario).

b) Se define el Eje de las Abscisas (X) para representar los valores calculados o estimados $(\hat{y})$.

c) Se define una recta bisectriz $y = \hat{y}$, que corta el plano XY en dos semiplanos.

d) Se plotean los pares ordenados $(\hat{y}, y)$ (Valor Calculado , Valor Observado); la distancia perpendicular de cada punto a la recta bisectriz definirá a los valores atípicos, que serán los más alejados a esa recta bisectriz.



Los valores que más alejados de la curva, plano o hiperplano de regresión, son los que se definirán como "Valores Atípicos".

Estos datos, que por definición no pertenecen a la Serie estudiada, deberán ser eliminados a fin de obtener un mejor ajuste en la regresión (un Coeficiente de Determinación ($R^{2)}$ mas alto).

El problema se presenta en determinar cuantos valores atípicos hay que eliminar de la serie, cuidando a su vez, no alterar sustancialmente el fenómeno estudiado (comportamiento del mercado en nuestro caso)

Si se eliminaran todos los valores atípicos de la serie, mas bien estaríamos "forzando" a unos datos a que encajen en un modelo, y lo que realmente se busca: Es el modelo que "mejor se ajuste (explique)" los datos de la muestra seleccionada.

7.3.4.- Procedimiento Analítico para la detección de Valores Atípicos

Analíticamente, se consideran Valores Atípicos, aquellos datos cuyos residuos $(y - \hat{y})$, se alejen más de un determinado (k) número de veces de la Recta Bisectriz $(y = \hat{y})$ precitada en el punto anterior.

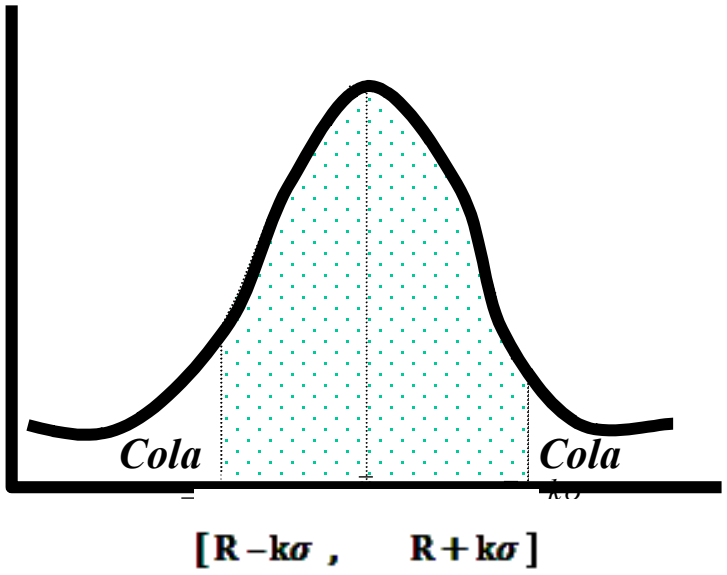
Para poder aplicar este procedimiento se deberá suponer que los Residuos $(y - \hat{y})$ de los referenciales se distribuye de manera "Normal"[2].

Una vez acordado la hipótesis anterior, se deberá calcular aquella desviación estándar que cumpla con la condición: Todo dato ubicado fuera del rango $[R - k\sigma \ , \ R + k\sigma]$, tenga una "Probabilidad" (p) que tienda a cero (0).

Donde la probabilidad (p) se calcula:

$$p = \frac{1}{n}$$

Siendo "n" el número de datos de la serie de referenciales seleccionados.



Aquellos datos, cuyos residuos se ubiquen debajo de las dos "colas", se consideran atípicos.

[2] EL Concepto de Normalida de una distribución, se explicó detalladamente en la Primera Parte (Análisis de una sola variable) de esta monografía.

Para conocer el inicio de cada una de las colas, debemos calcular en número de desviaciones estándar $(k\sigma)$ mas allá de las cuales la probabilidad (p) sea inferior que:

$$(\frac{1}{n})$$

La función que genera el coeficiente (k), se denomina: "**Distribución Normal Estándar Inversa**" (IDF) y se calcula por medio de una subrutina presente en la hoja de cálculo Excel dentro de las funciones estadísticas[3]

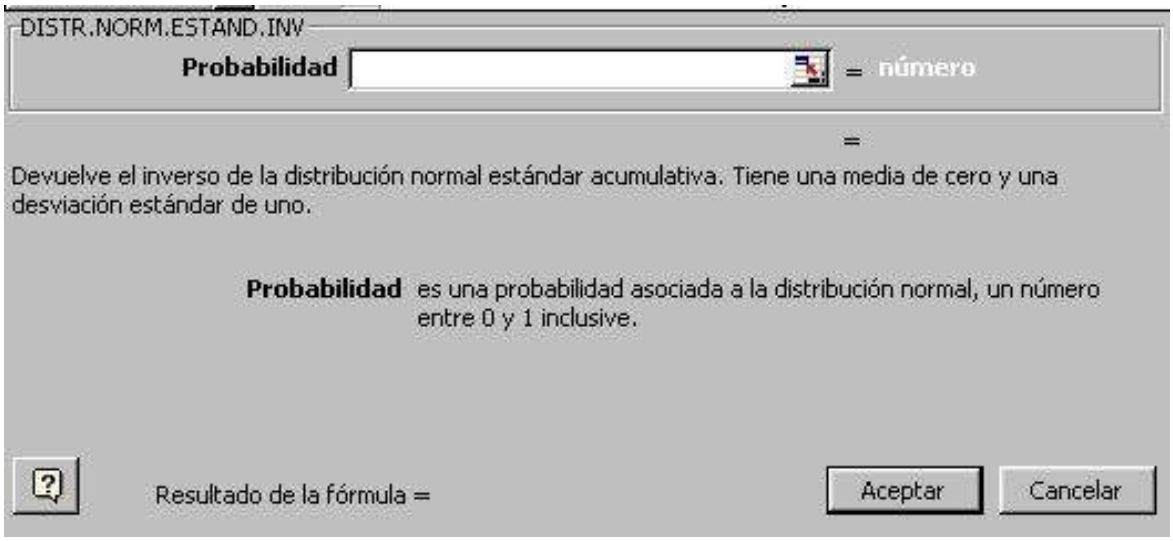
Ahora bien, habiendo calculado los residuos $(y - \hat{y})$ de todos los datos de la serie de referenciales:

Se definirán como "**Valores Atípicos**" todos aquellos datos que cumplan con la condición de que el Valor Absoluto de su residuo, se aleje $(k\sigma)$ veces del valor observado (y).

$$|y - \hat{y}| \quad \geq \quad |k\sigma|$$

Estos Valores Atípicos, serán eliminados de la serie de referenciales; y se volverá a correr la Regresión Múltiple con los datos remanentes.

[3] Para tener acceso a esta función: Clic sobre **fx** **Estadísticas** **DISTR.NORM.ESTAND.INV**.



La caja de diálogo solicitara que se entere la probabilidad (p) y la salida de la subrutina será el valor del coeficiente (k).

7.3.5.- Método Empírico:

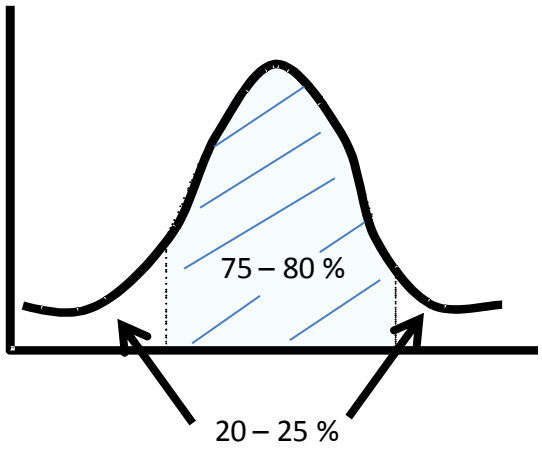
El “Método Empírico”, se basa en suponer:

a) Que los Residuos $(y - \hat{y})$ se distribuyen en forma Normal

b) Que debajo de las colas se ubica de un 20 a 25 % de los Residuos

c) Que debajo de la campana se ubica de un 75 a 80 % de los Residuos

Distribución Empírica de Los Residuos



El método, aconseja que el máximo de datos o valores atípicos que pueden ser eliminados de una serie, sin que la misma se altere sustancialmente, es de un 20 a 25% de los valores.

Adicionalmente se recomienda, para garantizar la integridad de los datos; que la totalidad de los valores atípicos de una serie no deben ser eliminados de una sola vez, sino por lo menos en dos corridas:

a) Primero se elimina el 50% de los valores atípicos o menos

b) Se vuelve a correr el software de Regresión Múltiple con los datos restantes, se chequea el R^2 y si no es satisfactorio

c) Se elimina el 50% restante y se vuelve a correr el software con el remanente que quedó de la serie.

7.3.6.- Validación: de la Regresión

Una vez eliminados los Valores Atípicos de la serie, se deberá comprobar si el Número de Datos y Variables Independientes que quedan en el modelo cumplen con el Test de Fisher (Estadístico F o Prueba F).

Para esto se vuelve a correr la serie de datos remanentes. De la salida del software se ubicará en valor del Estadístico F y se comparará con el Fo (F de prueba); que deberán cumplir con el criterio que F >>Fo para poder validar la regresión.

7.3.7.- Alcance del Método:

Por supuesto, todas las recomendaciones vistas en el Apartado 7.3, no garantizan la seguridad de poder determinar y validar la regresión múltiple.

Si no es posible determinar el modelo satisfactorio para explicar el fenómeno estudiado (comportamiento de los precio, en nuestro cas); no queda otro camino que el de realizar la valoración por la metodología clásica de Comparación o Mercado, ajustando los referenciales a las correcciones y criterios del Profesional Tasador.

EJEMPLO

Durante el último trimestre, se registraron los siguientes inmuebles, correspondientes a Apartamentos de una misma urbanización de interés social:

REF	P.UNIT. Bs/M2 Y	AREA M2 X1	EDAD EDIF. AÑOS X2
1	98,750.0	100.00	7
2	124,500.0	75.00	1
3	119,500.0	80.00	3
4	87,250.0	110.00	12
5	108,250.0	90.00	7
6	75,000.0	65.00	3
7	97,500.0	100.00	13
8	145,000.0	60.00	1
9	133,250.0	65.00	8
10	70,500.0	125.00	16
11	100,750.0	90.00	19
12	150,000.0	100.00	8

a) Utilizando la hoja de cálculo MS-Excel, se presenta la Salida de la subrutina correspondiente a la regresión múltiple lineal, de donde se obtiene:

* *Coeficientes del modelo de regresión múltiple lineal:*
* *Estadístico F*
* *Coeficiente de Determinación R^2*
* *Desviación Estándar de la regresión*
* *Otros datos estadísticos importantes*

NOTA: El paquete utilizado solo tiene definido algoritmos para el cálculo de Regresión Múltiple Lineal y Regresión Múltiple Exponencial (mal llamada por los desarrolladores de Microsoft como "Logarítmica"); por lo tanto no será posible determinar otro modelo de correlación diferente a los anteriores que explique mejor el fenómeno.

1) Modelo de Regresión Múltiple Lineal:

-1093.6993	-408.311899	154186.929
1756.58857	515.22433	37628.882
0.26938572	24397.5164	#N/A
1.65920076	9	#N/A
1975241361	5357149264	#N/A

De igual forma, se llama desde la hoja de cálculo la Subrutina de regresión múltiple logarítmica de donde se obtiene la salida siguiente:

2) Modelo de Regresión Múltiple Logarítmica:

0.99075842	0.99589442	164973.071
0.01655038	0.00485438	0.354535
0.26847521	0.22987059	#N/A
1.65153456	9	#N/A
0.17453579	0.47556441	#N/A

Tal como se observa, los Coeficientes de Determinación (R^2) para cada uno de los modelos es:

	Regresión Lineal	*0.26938572*	*26.94%*
	Regresión Logarítmica	*0.26847521*	*26.85%*

Sin embargo, el Coeficiente de Determinación del Modelo Lineal es ligeramente superior al del Modelo Logarítmico.

Por lo tanto, se elegirá al Modelo de Regresión Lineal, por tener el Coeficiente de Determinación más alto.

No obstante, el modelo lineal solo explica algo mas del 26% del fenómeno a ser estudiado. Por lo tanto, a esta altura del ejemplo, la correlación entre las variables no existe.

3) Existen por lo menos Dos (2) posibles causas de este bajo Coeficiente de Determinación (R^2):

a) La existencia de Multicolinealidad entre las Variables Independientes

b) La Existencia de Valores Atípicos en la serie de referenciales seleccionados.

4) Para determinar si existen problemas de Multicolinealidad entre las variables independientes, se presenta la salida del software, que corresponde a la Matriz de Correlación:

	CTE	X1	X2
CTE	1.0000	-o-	-o-
X1	-o-	1.0000	-o-
X2	-o-	**0.6975**	1.0000

Se puede apreciar en la Matriz de Correlación, que entre las Variables Independientes AREA (X1) y EDAD (X2), existe una correlación MEDIA y DIRECTA, no detectándose problemas de Multicolinealidad entre las Variables Independientes.

Habiendo descartado problemas de multicolinialidad, se presume que el bajo Coeficiente de Determinación calculado en el modelo, es consecuencia de la presencia de Valores Atípicos en la serie de referenciales. Por lo tanto se procederá a calcular los Residuos para determinar dichos valores.

5) Cálculo del valor de los Residuos:

REF	VALORES OBSERVADOS Y	V.CALCUL. $\hat{Y}$	RESIDUOS
1	98,750	105,700	-6,950
2	124,500	122,470	2,030
3	119,500	118,241	1,259
4	87,250	96,148	-8,898
5	108,250	109,783	-1,533
6	75,000	124,366	**-49,366**
7	97,500	99,138	-1,638
8	145,000	128,595	**16,405**
9	133,250	118,897	14,353
10	70,500	85,649	-15,149
11	100,750	96,659	4,091
12	150,000	104,606	**45,394**

6) Se procede a determinar los Valores Atípicos:

***NOTA IMPORTANTE**: En este ejemplo se emplearán los Dos (2) Métodos "Analítico" y "Empírico", para demostrar su aplicación.*

Sin embargo, es de hacer notar al estudiante, que debido a un "bug" en la programación de la función "Regresión Exponencial" (llamada erróneamente en Excel como "Logarítmica"); el resultado de la salida de los elementos:

- *"Desviación Estándar de la Regresión"*
- *"Desviaciones Parciales de las Variables"*
- *"SCR"*
- *"SCE"*

No son correctos; y al no poder obtener σ (Desviación Estándar de la Regresión), no se puede calcular el intervalo $k\sigma$ y por lo tanto no se puede aplicar el "Método Analítico" a los resultados de la "Regresión Exponencial".

a) Procedimiento Analítico:

a-1) Se calcula la probabilidad $p = \frac{1}{n}$

a-2) Se calcula (k) a través de la función de la hoja de cálculo Excel: "Distribución Normal Estándar Inversa" (IDF)

a-3) Se obtiene la Desviación Estándar de la Regresión de la salida de la Regresión Múltiple Lineal $(\sigma = 24397.5164)$

a-4) Se calcula el factor $k\sigma$

p =	0.08333333
k =	-1.38299583
σ =	24397.5164
$k\sigma$ =	-33741.6635

a-5) Se procede a determinar cuáles son los datos que el Valor Absoluto de su residuo es mayor a $(k\sigma)$ *:*

REF	VALORES OBSERVADOS Y	V.CALCUL. ^ Y	RESIDUOS	PROCE-DIMIENTO ANALITICO	
1	98,750	105,700	-6,950		
2	124,500	122,470	2,030		
3	119,500	118,241	1,259		
4	87,250	96,148	-8,898		
5	108,250	109,783	-1,533		
6	75,000	124,366	**-49,366**	$\lvert y - y\rvert \geq k\sigma$	Valor Atípico
7	97,500	99,138	-1,638		
8	145,000	128,595	**16,405**		
9	133,250	118,897	14,353		
10	70,500	85,649	-15,149		
11	100,750	96,659	4,091	$\lvert y - y\rvert \geq k\sigma$	
12	150,000	104,606	**45,394**		Valor Atípico

Donde $\lvert k\sigma \rvert = 33741.665$

a-6) Se concluye que los datos "6" y "12", son Valores Atípicos, y por lo tanto podrán eliminarse y volver a correr el software con los datos restantes para determinar el modelo de regresión.

b) Comprobamos a través del Método Empírico:

a-1) Siendo 12 los datos de la serie de referenciales, el 25% de los mismos serán 3 datos, que es el número máximo de Valores Atípicos que se pueden eliminar. Ordenados de Mayor a menor estos serán:

ORDEN	REF	RESIDUOS
A	6	-49,366
B	12	45,394
C	8	16,405

a-2) Sin embargo, el procedimiento indica que por lo menos hay que eliminar estos Valores Atípicos en Dos (2) corridas como mínimo.

a-3) Por lo tanto, se eliminarán primeramente los referenciales "6" y "12" y se volverá a correr las subrutinas antes explicadas.

NOTA: *Sin aún persistiera el problema, se procedería a eliminar el referencial "8" y volver a llamar las subrutinas de regresión de la Hoja de Cálculo.*

a-4) Una vez eliminados los referenciales "6" y "12", la nueva serie de referenciales a procesar será:

REF	P.UNIT. Bs/M2 Y	AREA M2 X1	EDAD EDIF. AÑOS X2
1	98,750.0	100.00	7
2	124,500.0	75.00	1
3	119,500.0	80.00	3
4	87,250.0	110.00	12
5	108,250.0	90.00	7
7	97,500.0	100.00	13
8	145,000.0	60.00	1
9	133,250.0	65.00	8
10	70,500.0	125.00	16
11	100,750.0	90.00	19

a-5) Llamando nuevamente la subrutina, llegamos a la nueva salida, donde se indican los Coeficientes del Nuevo Modelo de Regresión Lineal, el nuevo Coeficiente de Determinación y el valor actualizado del Estadístico F.

-536.611263	**-983.999794**	**201261.5**
142.676756	43.7999005	3263.63483
0.99412436	1942.10911	#N/A
592.179888	7	#N/A
4467153736	26402514.5	#N/A

a-6) El coeficiente de Determinación 0.99412436, indica una excelente correlación de los Diez (10) datos remanentes.

7) Para validar la regresión procedemos a calcular el estadístico Fo en la tabla anexa:

# DE REFERENCIALES:	10
# DE VARIABLES INDEPENDIENTES:	2
GRADOS DE LIBERTAD [n-(k+1)]:	7
Fo =	4.74
F =	592.179888
F >> Fo	

8) Con esto queda validado el modelo de correlación múltiple, el cual queda expresado de la siguiente manera:

y = 201261.5 - 983.999794 * X1 - 536.311263*X2

Rev.: Ene/2011

BIBLIOGRAFIA

DESEDA, L., Estadística Aplicada a la Valuación. Editorial Akros. 1996

NOVALES, A., Econometría. Segunda Edición. McGraw-Hill. 1993

FERNANDEZ, A., Ejercicios de Econometría. Primera Edición. Mac Graw Hill. 1993

GREENE, W., Análisis Econométrico. Tercera edición. Editorial Prentice Hall.1999

GUAJARATI, D., Econometría., cuarta edición. Editorial Mc Graw Hill. 2004

MADDALA, G.S., Introducción a la Econometría. Segunda edición. Editorial Prentice Hall.1996

PINDICK, R., RUBINFIELD, D., Econometría: Modelos y pronósticos. Mc Graw Hill. 1998

PIOL, R. Estadística Aplicada a la Valuación Inmobiliaria. Parte II. Análisis de Variables Múltiples. Sociedad de Ingeniería de Tasacion de Venezuela. 1999