

UNIVERSIDAD DE ARTEMISA

Carrera de Ingeniería Industrial
Estadística II

MSc. Ing. Rodolfo Pérez Léon
Email: rperezleon@uart.edu.cu

Temas
1. Análisis de Varianza.
2. Inferencias en Tablas de Contingencia
3. Análisis de correlación. y Simple
4. Análisis de Regresión. El Modelo de Regresión MRLG.
5. Verificación de los supuestos del Modelo (Homocedasticidad, No Autocorrelación, No Multicolinealidad)
6. Series de Tiempo

UNIVERSIDAD DE ARTEMISA

Sumario

Tema 1. Análisis de Varianza.

1.1. Introducción. Análisis de Varianza Supuestos del modelo.

1.2. Descomposición de la suma de cuadrados. Varianza del error estándar (Se). Tabla de Análisis de varianza. Prueba F.

1.3. EJERCITACIÓN.

Objetivos

Que el alumno se capaz de:

- Aplicar el análisis de varianza en la solución de problemas del perfil de la profesión.

- Interpretar salidas de un paquete de programas estadísticos sobre el análisis de varianza.

- Método:

Elaboración conjunta y trabajo independiente

- Medios:

Tiza, pizarra, libro de texto, Material Didáctico Digital (MDD).

- Forma de enseñanza:

Conferencia

- Bibliografía:

Probabilidades y Estadística para ingenieros

Materiales entregados por el profesor

1.1. Introducción. Supuestos del modelo.

Es muy frecuente en la vida que un profesional tenga que decidir si entre varios procesos o tratamientos algunos de ellos producen resultados diferentes.

La pregunta de interés del profesional será ¿existen diferencias significativas entre las medias de los tratamientos? Si comprueba que existen diferencias significativas, entonces le interesará saber ¿cuáles de las poblaciones o tratamientos comparados son diferentes?

El procedimiento para determinar si existen diferencias significativas entre MÁS de 2 poblaciones o grupos se llama Análisis de Varianza (ANOVA por Analysis of Variance, en inglés). Por ahora sólo hablaremos del ANOVA simple, de una vía ó factor (one way ANOVA), que se refiere a que se evalúa una sola causa de efecto (los tratamientos) a la comparación de medias de más de dos tratamientos. Vamos a llamar factor a una variable cualitativa que usaremos para designar a los grupos o tratamientos a comparar. Los niveles del factor serán el número de tratamientos o grupos.

Ejemplo 1. Un ingeniero desea comparar 4 máquinas que se dedican a fabricar juntas de gomas para ollas de presión. Para la comparación se analiza la resistencia a la flexión de las juntas, entonces son selecciona 4 juntas producidas de cada máquina y se mide la resistencia a la flexión de cada junta, obteniendo los siguientes resultados

Máquinas	Máquina No.1	Máquina No.2	Máquina No. 3	Máquina No.4
Resiste a la flexión Junta 1	4.3	4.0	2.0	3.5
Resiste a la flexión Junta 2	4.8	4.1	3.8	3.4
Resiste a la flexión Junta 3	3.9	3.7	3.6	4.0
Resiste a la flexión Junta 4	3.7	3.9	3.5	4.1

En el caso usaremos un análisis de varianza ANOVA simple de un factor (el valor de la característica que se mide resistencia a la flexión de las juntas) está influido por un solo efecto que en este caso serán la máquina que los produce que llamaremos (Tratamientos).
Si además de la resistencia a la flexión de las juntas se evaluarán otras variables como la elasticidad de la junta, entonces el Análisis sería Multivariado (Análisis de 2 factores).

El Modelo matemático se representa por la expresión siguiente:

$$y_{ij} = \mu + \sigma_1 + e_{ij}$$

y_{ij}= Valor que toma la característica aleatoria “y”

μ= Valor que temaría la característica “y” si no estuviera presente el tratamiento y el error aleatorio

σ_1= Efecto del tratamiento.

e_{ij} = Error de la Medición de y_{ij}

El subíndice “i”: los diferentes tratamientos en este caso serán las 4 máquinas.

El subíndice “j”: se designa a las veces que se repite el tratamiento en cada grupo, en este caso 4 juntas diferentes, se debe aclarar que en la clasificación simple no es necesario que todos los grupos tengan las mismas cantidades. simple no es necesario que todos los grupos tengan las mismas cantidades.

Supuestos del Modelo:

- Las “y_{ij}” siguen una Distribución Normal con Media ($\mu+\sigma_1$) y varianza σ^2 para todo “i; j”.

- Las observaciones estén incorrelacionadas o sean independiente.

Como se desea comprobar si alguno de los tratamientos (máquina) produce efectos diferentes entonces la hipótesis a comprobar son:

Prueba de Hipótesis.

Ho: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

Hi: al mejos una μ_i diferente

1.2. Descomposición de la suma de cuadrados total.

Tabla de Análisis de Varianza. Prueba F total.

$$y_{ij} = \mu + \sigma_1 + e_{ij}$$

y_{ij}= Valor que toma la característica aleatoria “y”

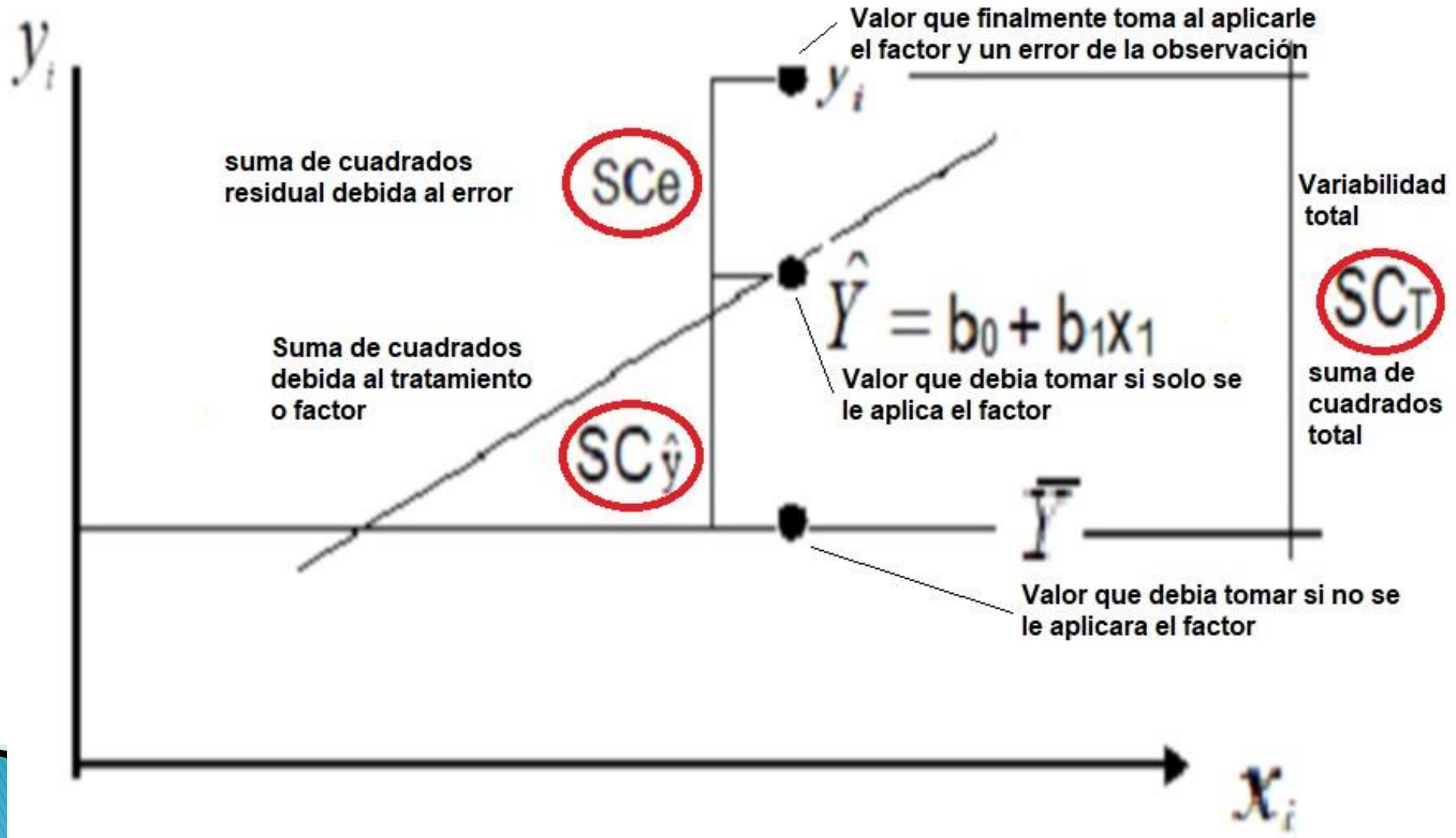
μ= Valor que temaría la característica “y” si no estuviera presente el tratamiento y el error aleatorio

σ_1= Efecto del tratamiento.

e_{ij} = Error de la Medición de y_{ij}

El valor de cada observación esta dado por estas causas

1.3. Descomposición de la Suma de Cuadrados. Cuadro ANOVA (Análisis de Varianza).



Causas de variación	Suma de cuadrados (SC)	Grados de libertad (GL)	Cuadrado medio (CM)	F(calc)	Pvalor
Debida a la Regresión	$SC\,\hat{y}$ $b_1(\Sigma xy - n\,\bar{Y}\,\bar{X})$	p – 1	$S\,\hat{y}^2$ $b_1(\Sigma xy - n\,\bar{Y}\,\bar{X}) / p\text{-}1$	S^2_{reg} / S_e^2	0.031
Residual o error	SC_e $SC_T - SC\,\hat{y}$	n – p	S_e^2 $SC_T - SC\,\hat{y} / n\text{-}p$ **Varianza del error estandar**		
Total	SC_T $\Sigma Y_i^2 - n\,\bar{Y}^2$	n - 1			

S_e^2: Este término se conoce como la **varianza del error estándar**

S_e : ?

Para la Prueba de Hipótesis acerca de las medias usaremos el test estadístico F. Este test contrasta la variabilidad entre los grupos con la variabilidad que será natural dentro de los grupos.

$$F = \frac{\text{variabilidad ENTRE las medias muestrales}}{\text{variabilidad DENTRO de las muestras}}$$

Para realizar la prueba se debe comparar la Razón F ó $F_{calculada}$ con la $F_{tabulada}$.

$F_{tabulada}$ se busca en la Tabla de Fisher: para

$$F_{(1-\alpha;\ p-1;\ n-p)}$$

α: Nivel de significación.

p: Cantidad de tratamientos.

n: Tamaño de la muestra

Si Razón $F > F_{(1-\alpha;\ p-1;\ n-p)}$. Se Rechaza Ho

Otra forma puede ser comparando el p-valor con el nivel de significación (α) y:

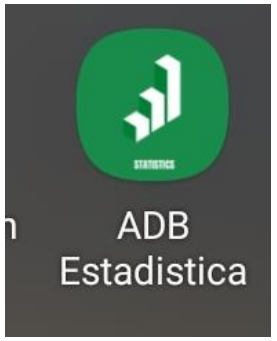
Si p-valor < (α) entonces se Rechaza Ho.

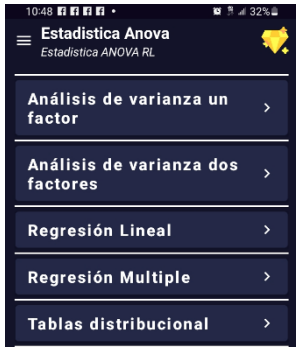
1.3. EJERCITACIÓN.

Ejemplo 1. Un ingeniero desea comparar 4 máquinas que se dedican a fabricar juntas de gomas para ollas de presión. Para la comparación se analiza la resistencia a la flexión de las juntas, entonces son selecciona 4 juntas producidas de cada máquina y se mide la resistencia a la flexión de cada junta, obteniendo los siguientes resultados

Máquina No1	Máquina No. 2	Máquina No. 3	Máquina No.4
4.3	4.0	2.0	3.5
4.8	4.1	3.8	3.4
3.9	3.7	3.6	4.0
3.7	3.9	3.5	4.1

Después de procesar la información con la apk ADB Estadística se obtuvieron los siguientes resultados:





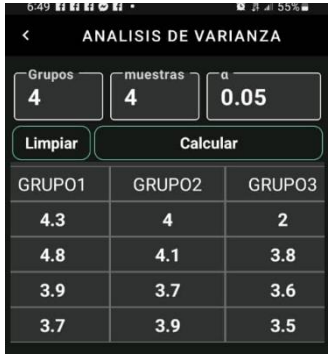


TABLA ANOVA

FUENTES DE VARIACION	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	VARIANZAS O MEDIA	ESTADISTICO TEST	p-value
ENTRE LOS GRUPOS	1.942	3	0.647	2.418	0.117
DENTRO DE LOS GRUPOS	3.212	12	0.268		
TOTAL	5.154	15			

Existirá diferencias significativas entre los resultados de las 4 máquinas con una confianza del 95%?

RESPUESTA:

Datos:

Tratamientos máquinas (p)= 4

n= 16

(1-α)= 0.95 entonces α= 0.05

1er Paso: Planteamiento de la Hipótesis.

H_o: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ No hay diferencias entre las resistenci a la flexión de las máquinas

H_i: al menos un $\mu_i \neq$ alguna es desigual

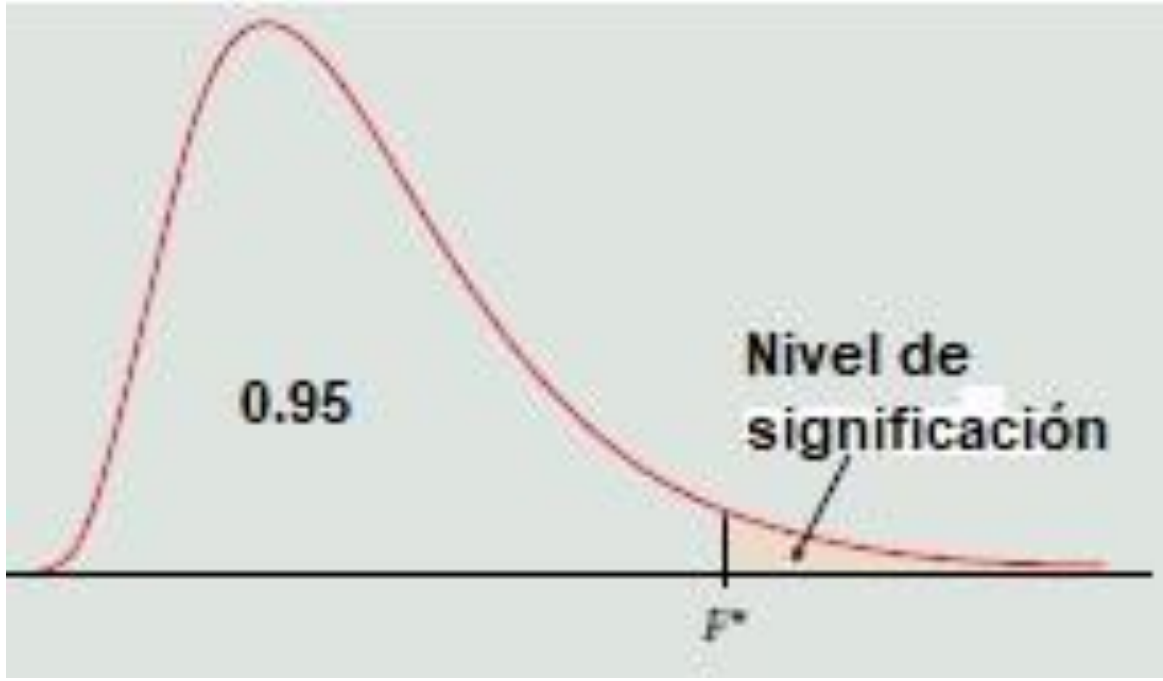
2Do Paso: Fijar el Nivel de significación α

α= 0.05

3er Paso: Determinar la Zona Crítica o Regla de Decisión

Si $F_{cal} > F_{(0.95;\ 3;\ 15)}(F_{crítico})$ entonces Rechazo H_o

ó Si $P_{valor} < \alpha$ entonces Rechazo H_o



4to Paso: Calcular el Estadístico Z_{cal}

$F_{(0.95;\ 3;\ 15)}=3.29$



5to Paso: Toma de decisión.

$F_{cal} > F_{(0.95;\ 3;\ 15)}(F_{crítico})$

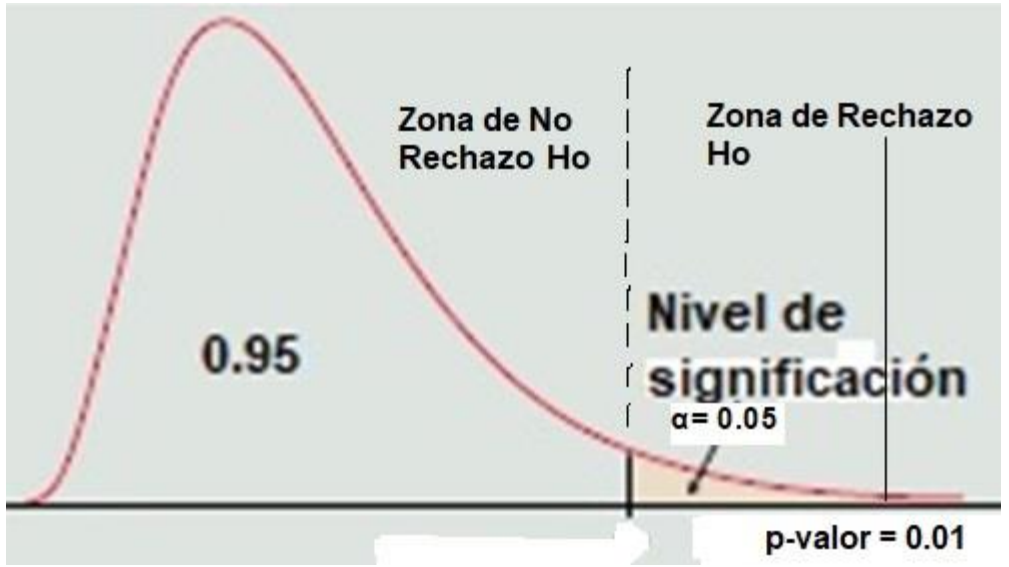
$2.18 < 3.29$ Entonces No Rechazo H_o

R/ No hay diferencias significativas en la resistencia a la flexión de las juntas producidas por las 4 máquinas

5to Paso: Toma de decisión.

Con p-valor

Si p-valor < α entonces Rechazo H_o



0.11>0.05 entonces No Entonces Rechazo H_o

R/ No hay diferencias significativas en la resistencia a la flexión de las juntas producidas por las 4 máquinas

Conclusiones:

- El análisis de varianza simple (oneway ó de un factor) analiza un solo efecto (los tratamientos) sobre la media de las muestras.
- El análisis de Varianza permite decidir si al menos uno de los tratamientos produce resultados diferentes a los demás, aunque no se pueda definir cual es el diferente, para ello sería necesario otras pruebas como la Prueba de Tukey que no se estudia en este curso.

ESTUDIO INDEPENDIENTE

ESTUDIO INDEPENDIENTE

Definir si hay diferencias significativas en la resistencia a flexión de las juntas que se producen en cada una de las 6 máquinas

Máquina

1	2	3	4	5	6
17.5	16.4	20.3	14.6	17.5	18.3
16.9	19.2	15.7	16.7	19.2	16.2
15.8	17.7	17.8	20.8	16.5	17.5
18.6	15.4	18.9	18.9	20.5	20.1

Muchas gracias por su atención

Prof: Rodolfo Pérez León
Email: rperezleon@uart.edu.cu

UNIVERSIDAD DE ARTEMISA

ESTADÍSTICA II

MSc. Rodolfo Pérez Léon
email: rperezleon@uart.edu.cu
rplrpl55@gmail.com

UNIVERSIDAD DE ARTEMISA

Tema 2. Prueba no paramétrica chi-cuadrado.

Sumario.

2.1. Introducción. Antecedentes históricos de la prueba Chi-cuadrado
2.2. Distribución Chi-Cuadrado. Características.
2.3. Prueba Chi-Cuadrado. Tablas de contingencia.
2.4. Prueba no paramétrica de independencia de variables.
2.5. Prueba de independencia hi-Cuadrado con App Estadística.
2.6. Prueba de homogeneidad o de proporcionalidad de varias muestras

Objetivos: Que el alumno se capaz de:

Aplicar las pruebas Chi-Cuadrado en la solución de ejercicios de la profesión.

Método: Expositivo.

Medios: Tiza y pizarra, Libro de texto, MDD.

Forma de enseñanza: Conferencia

Bibliografía:

- Probabilidades y Estadística para ingenieros. J.E Freund, I.R. Miller, R. Jonhnson.
- Materiales Didácticos Digitales entregados por el profesor

2.1. Introducción, antecedentes históricos

2.1. Introducción, antecedentes históricos de la distribución Chi-Cuadrado

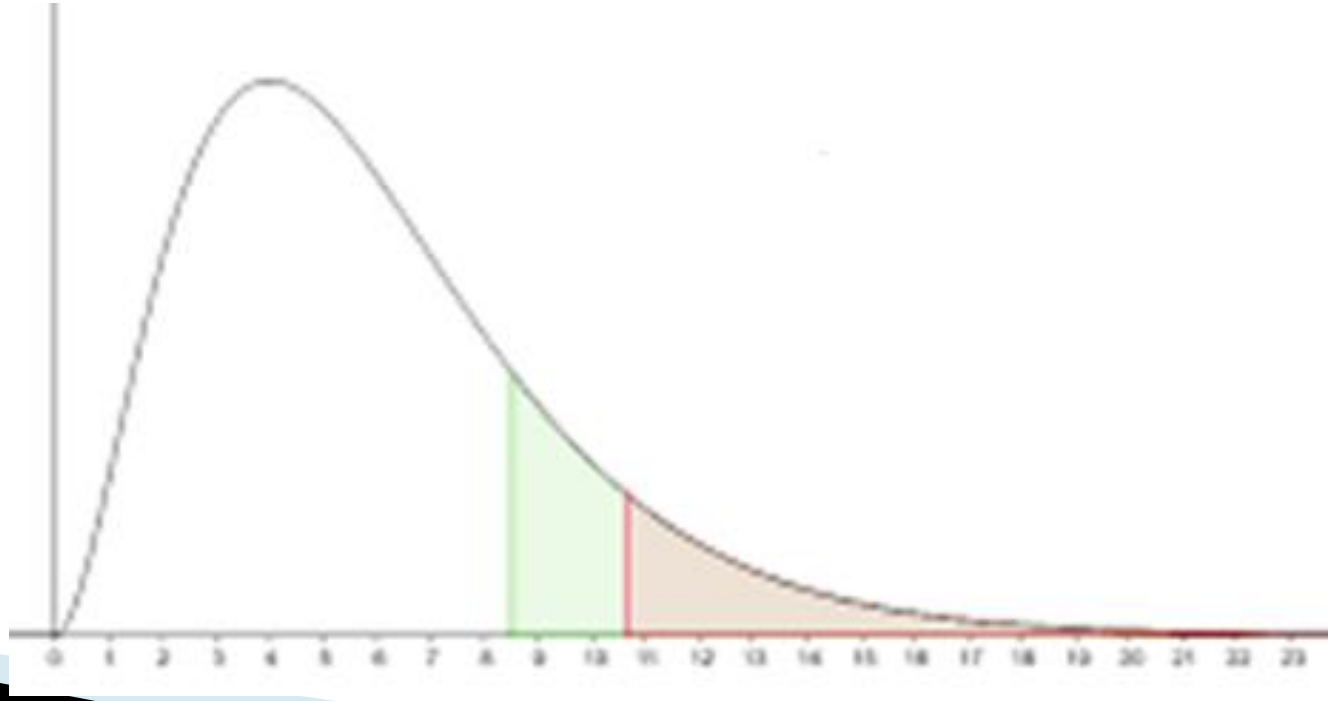
- El matemático Karl Pearson (1857−1936)
- Karl Pearson en 1900 propuso uno de los primeros Test Estadísticos que desde la óptica de las distribuciones de la probabilidad sirve para calcular si los resultados estadísticos de un experimento se alejan significativamente o no de los resultados esperados del modelo teórico, test que actualmente es conocido como el «Test Chi Cuadrado».
- se denota mediante la letra griega minúscula ji elevada al cuadrado (χ^2)

2.2. Distribución CHI CUADRADO. Características

- La distribución Chi-Cuadrado es una distribución asimétrica

- Sólo toma valores positivos y es asintótica con respecto al eje de las x positivas ($0 < \chi_2 < +\infty$)

- El área comprendida entre la curva y el eje de las x es 1 ó 100%



- Las pruebas Chi-Cuadrado se utilizan para determinar si las frecuencias observadas (O) en las categorías difieren significativamente de las frecuencias esperadas (E).

- No se utiliza en números solo se categoriza. (la prueba Chi-cuadrada no considera relaciones causales).

2.3. Prueba CHI CUADRADO
Tablas de Contingencia (TC)

¿QUÉ ES LA PRUEBA CHI CUADRADA?

- Es una prueba estadística para evaluar hipótesis acerca de la relación entre dos variables categóricas, también para evaluar si hay homogeneidad o igual proporción entre diferentes muestras de esas variables

- Se simboliza por $\chi 2$

- Hipótesis a probar : independencia o igual proporciones
  (H_0 : son independientes y H_1 : no son independientes)
  (H_0 : $p_1 = p_2 = \ldots p_n$ H_1 : $p_1, p_2, \ldots p_n$ no son iguales)

- Variables involucradas: 2, esta prueba no considera relaciones causales

- Nivel de medición de las variables: Nominal u ordinal organizadas en una Tabla de Contingencia

EJEMPLO DE UNA TABLA DE CONTINGENCIA 2x2

Dos variables : voto y sexo

Cada variable con dos categorías

		VOTO	
		candidato a	candidato b
sexo	masculino		
	femenino		

PROCEDIMIENTO

- Se calcula a través de una tabla de contingencia o tabulación cruzada.
- Es una tabla de dos dimensiones y cada dimensión contienen una variable
- Cada variable se subdivide en dos o más categorías.

Se utiliza

- **PRUEBA DE BONDAD DE AJUSTE:** compara la distribución de una muestra con la distribución teórica que se supone que describe a la población de la cual se extrajo.
- **PRUEBA DE INDEPENDENCIA DE DOS VARIABLES:** comprueba si dos o más variables que se miden en una misma muestras están o no relacionadas entre sí. Siempre hablamos de variables cualitativas.
- **PRUEBA DE HOMOGENEIDAD:** comprueba si dos o más muestras pertenecen a la misma población.

2.4. Pruebas no paramétrica de independencia de dos variables

Ejemplo 1. Independencia de dos variables (1 muestra)

Un taller de una industria azucarera fabrica piezas metálicas para la zafra y desea evaluar si existe una relación entre el turno de producción (mañana, tarde y noche) y la calidad de las piezas producidas. Para ello, **se seleccionó una muestra** de 1040 piezas y se clasificaron según el turno de fabricación y la calidad de las piezas, obteniendo la siguiente tabla de contingencia de 3x2:

Turnos categorías: mañana, tarde, noche.

Calidad de las piezas categorías: no defectuosas, defectuosas.

TABLA DE CONTINGENCIA

		calidad de las piezas	
		no defectuosas	**defectuosa**
turnos	**mañana**	180	100
	tarde	190	280
	noche	170	120

Prueba de hipótesis X^2: (Probar si hay relación entre las variables)

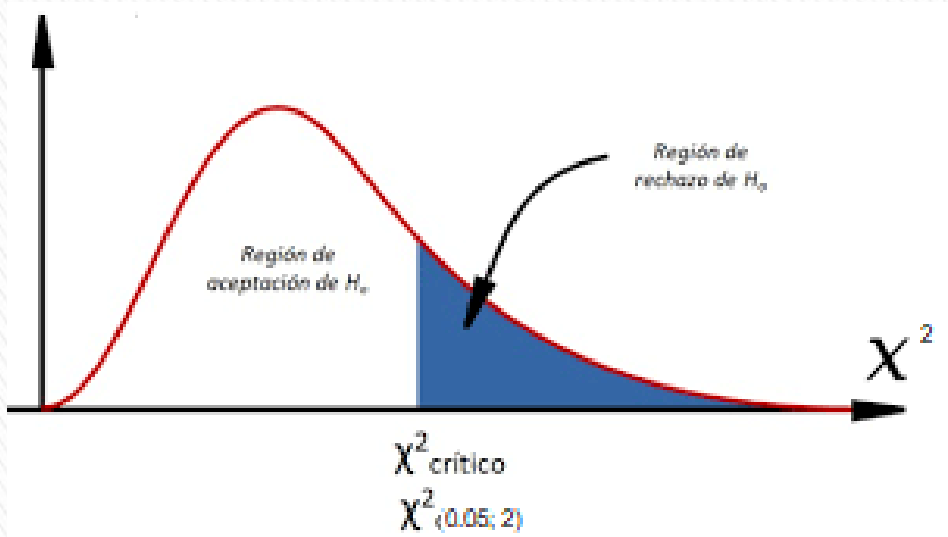
1. Planteamiento de la Hipótesis

H_0: No existe relación entre el turnos y la calidad de las piezas (son independiente).

H_1: Existe relación entre el turnos y la calidad de las pieza (no son independiente).

2. Nivel de significación

α = 0.05



3. Regla de decisión

Si $X^2_{calculada} > X^2_{tabla}$ se rechaza (H0)

4. Calcular $X^2_{calculada}$ y buscar $X^2_{critica}$

4. Calculo $X^2_{calculada}$

Tabla de frecuencias observadas (O):

	calidad de las piezas			TOTAL
		no defectuosas	**defectuosas**	
turnos	**mañana**	180	100	**280**
	tarde	190	280	**470**
	noche	170	120	**290**
	TOTAL	**540**	**500**	**1040**

La Chi-cuadrada es una comparación entre las tablas de **frecuencias observadas** y la denominada tabla de **frecuencias esperadas (**la tabla que esperaríamos encontrar si las variables fueran estadísticamente independientes o no estuvieran relacionadas**)**.

4. Calculo $X^2_{calculada}$

Tabla de frecuencias esperadas (E): Fe= Tf x Tc/ Tg

La **frecuencia esperada** de cada celda, casilla o recuadro, se calcula mediante la siguiente fórmula aplicada a la tabla de frecuencias observadas:

N = es el número total de frecuencias observadas.

Fe = (marginal de la fila)(marginal de columna) / N.

		calidad de las piezas		
		no defectuosas	**defectuosas**	Marginal de filas
turnos	**mañana**	**(280) (540) / 1040**	**(280) (500) / 1040**	**280**
	tarde	**(470) (540) / 1040**	**(470) (500) / 1040**	**470**
	noche	**(290) (540) / 1040**	**(290) (500) / 1040**	**290**
marginal de columnas		**540**	**500**	**1040**

4. Calculo $X^2_{calculada}$

Frecuencia observada:

		calidad de las piezas		TOTAL
		no defectuosas	defectuosas	
turnos	mañana	180	100	**280**
	tarde	190	280	**470**
	noche	170	120	**290**
	TOTAL	**540**	**500**	**1040**

Frecuencia esperada:

		calidad de las piezas		TOTAL
		no defectuosas	defectuosas	
turnos	mañana	145,4	134,6	**280**
	tarde	244,0	226,0	**470**
	noche	150,6	139,4	**290**
	TOTAL	**540**	**500**	**1040**

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Donde:
O i: frecuencia observada en cada celda
E i: frecuencia esperada en cada celda

Calculo $X^2_{calculada}$

$$X^2 = \frac{\sum (O - E)^2}{E}$$

Celda	O	E	O-E	$(O-E)^2$	$(O-E)^2 / E$
Mañana / no defectuosa	180	145,4	34,6	1197,16	8,23
tarde/ no defectuosas	190	244,4	-54,4	2959,36	12,11
noche / no defectuosas	170	150,6	19,4	376,36	2,50
mañana / defectuosa	100	134,6	-34,6	1197,16	8,69
tarde / defectuosa	280	226,0	54,0	2916,00	12,80
noche / defectuosa	120	139,4	-19,4	376,36	2,70
				X^2 =	**47,33**

4. Buscar $X^2_{crítica}$

Se comienza por calcular los grados de libertad.

G.L. = (Nº de filas - 1)(Nº de columnas - 1).

Para el ejemplo: **Nº de filas** = 3 y **Nº de columnas** = 2; entonces G.L. = (3-1)(2-1) = 2.

Luego, acudimos a la “tabla de distribución de Chi-cuadrado”, eligiendo nuestro nivel de significación (α = 0,05).

TABLA 3-Distribución Chi Cuadrado χ^2

P = Probabilidad de encontrar un valor mayor o igual que el chi cuadrado tabulado, ν = Grados de Libertad

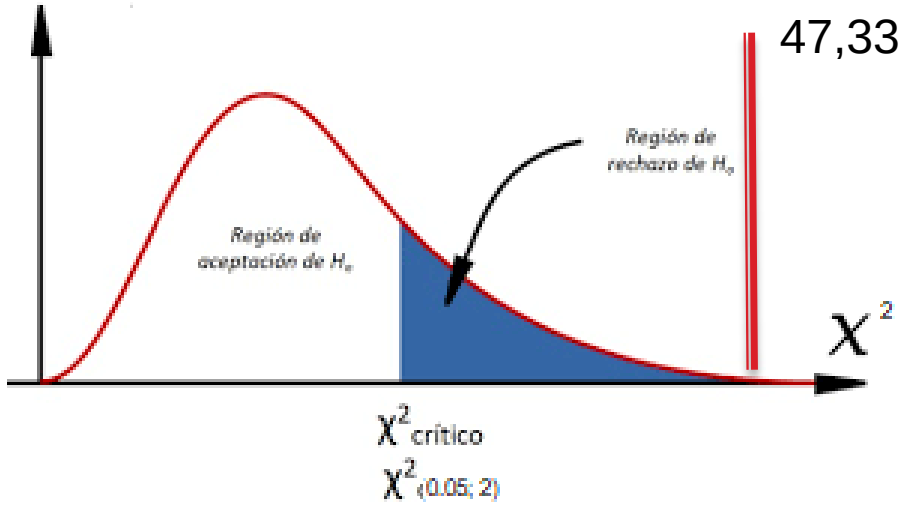
ν/p	0,001	0,0025	0,005	0,01	0,025	0,05	0,1	0,15	0,2	0,25	0,3	0,35
1	10,8274	9,1404	7,8794	6,6349	5,0239	3,8415	2,7055	2,0722	1,6424	1,3233	1,0742	0,8735
2	13,8150	11,9827	10,5965	9,2104	7,3778	5,9915	4,6052	3,7942	3,2189	2,7726	2,4079	2,0996
3	16,2660	14,3202	12,8381	11,3449	9,3484	7,8147	6,2514	5,3170	4,6416	4,1083	3,6649	3,2831
4	18,4662	16,4238	14,8602	13,2767	11,1433	9,4877	7,7794	6,7449	5,9886	5,3853	4,8784	4,4377
5	20,5147	18,3854	16,7496	15,0863	12,8325	11,0705	9,2363	8,1152	7,2893	6,6257	6,0644	5,5731
6	22,4575	20,2491	18,5475	16,8119	14,4494	12,5916	10,6446	9,4461	8,5581	7,8408	7,2311	6,6948
7	24,3213	22,0402	20,2777	18,4753	16,0128	14,0671	12,0170	10,7479	9,8032	9,0371	8,3834	7,8061
8	26,1239	23,7742	21,9549	20,0902	17,5345	15,5073	13,3616	12,0271	11,0301	10,2189	9,5245	8,9094
9	27,8767	25,4625	23,5893	21,6660	19,0228	16,9190	14,6837	13,2880	12,2421	11,3887	10,6564	10,0060
10	29,5879	27,1119	25,1881	23,2093	20,4832	18,3070	15,9872	14,5339	13,4420	12,5489	11,7807	11,0971
11	31,2635	28,7291	26,7569	24,7250	21,9200	19,6752	17,2750	15,7671	14,6314	13,7007	12,8987	12,1836
12	32,9092	30,3182	28,2997	26,2170	23,3367	21,0261	18,5493	16,9893	15,8120	14,8454	14,0111	13,2661
13	34,5274	31,8830	29,8193	27,6882	24,7356	22,3620	19,8119	18,2020	16,9848	15,9839	15,1187	14,3451
14	36,1239	33,4262	31,3194	29,1412	26,1189	23,6848	21,0641	19,4062	18,1508	17,1169	16,2221	15,4209
15	37,6978	34,9494	32,8015	30,5780	27,4884	24,9958	22,3071	20,6030	19,3107	18,2451	17,3217	16,4940
16	39,2518	36,4555	34,2671	31,9999	28,8453	26,2962	23,5418	21,7931	20,4651	19,3689	18,4179	17,5646
17	40,7911	37,9462	35,7184	33,4087	30,1910	27,5871	24,7690	22,9770	21,6146	20,4887	19,5110	18,6330
18	42,3119	39,4220	37,1564	34,8052	31,5264	28,8693	25,9894	24,1555	22,7595	21,6049	20,6014	19,6993

Aplicación:

Para el nivel de significación de α = **0,05** y **g.l. = 2**, el X^2 de tabla es **5,9915** (ver tabla).

$$X^2_{caculado} = 47{,}33$$

$$X^2_{tabla} = 5{,}9915$$



5,9915

Prueba de hipótesis X^2: (Probar si hay relación entre las variables)

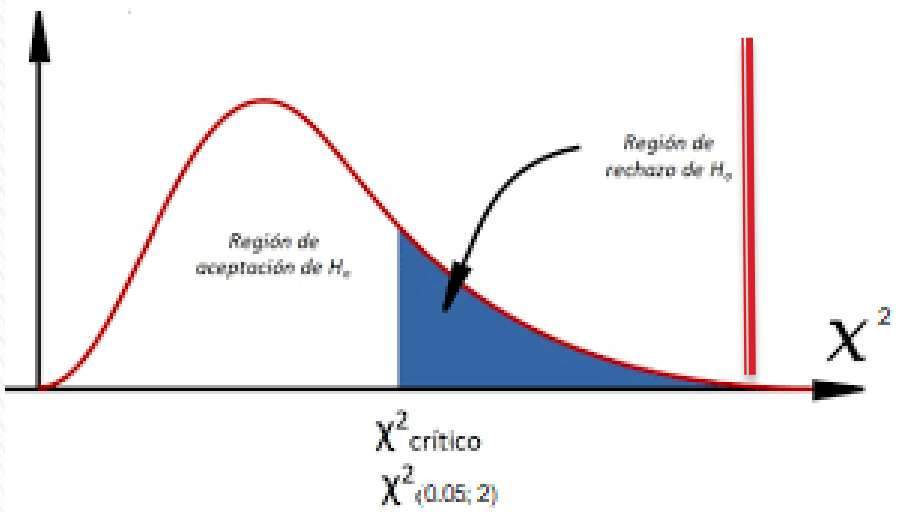
Planteamiento de la Hipótesis

H_0: No existe relación entre el turnos y la calidad de las piezas.

H_1: Existe relación entre el turnos y la calidad de las piezas.

Nivel de significación

$\alpha = 0.05$



Regla de decisión

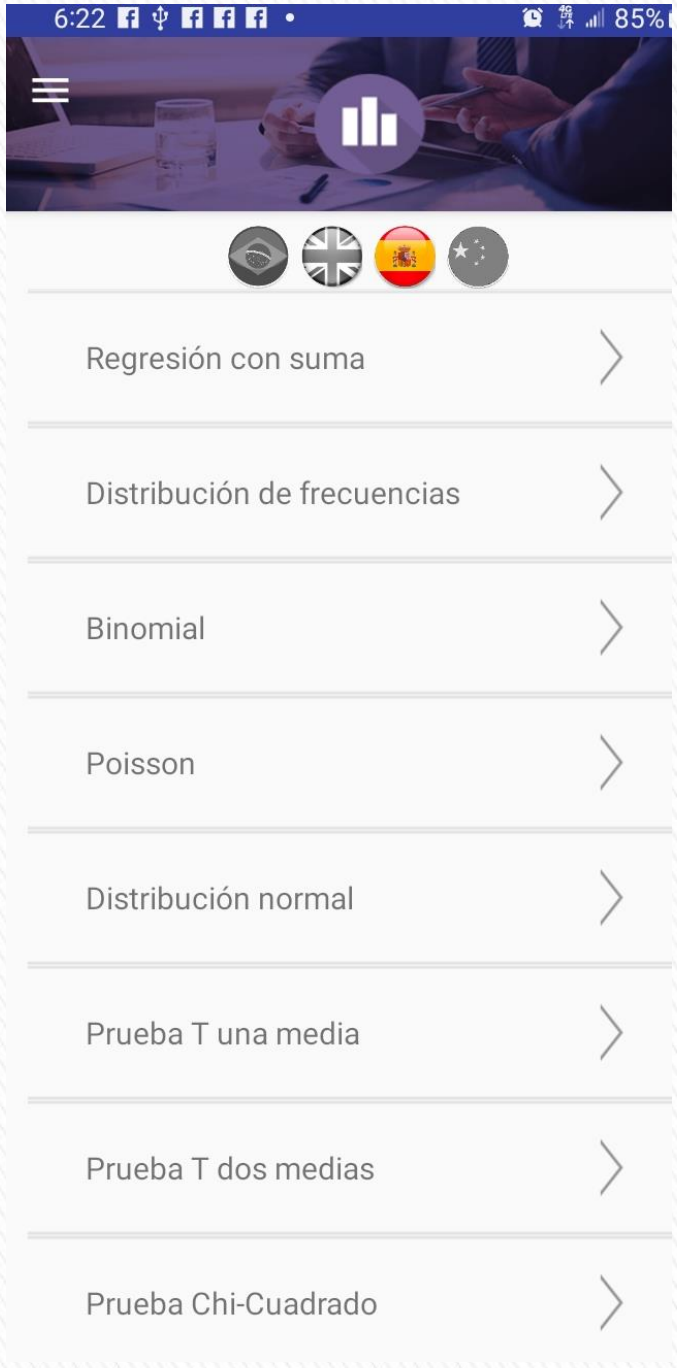
Si $X^2_{calculada} > X^2_{tabla}$ se rechaza (H_0)

Calculo $X^2_{calculada}$

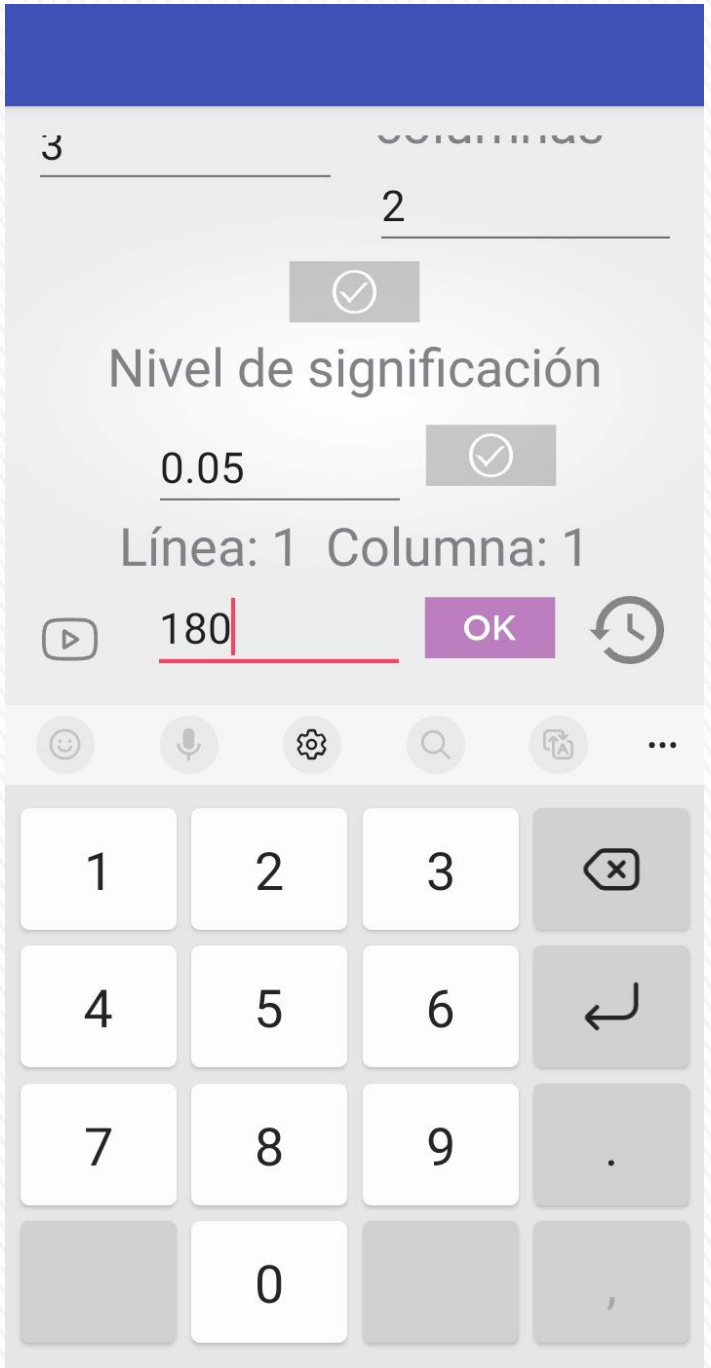
5,9915

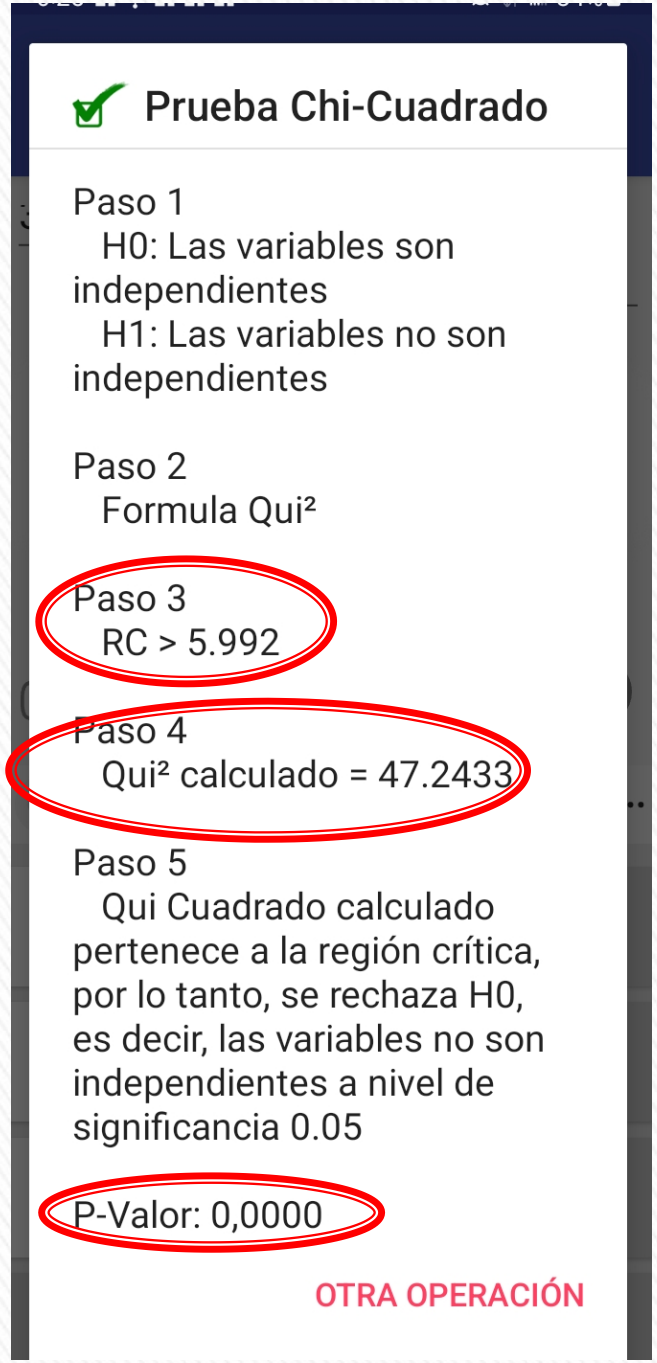
47.33 > 5,9915 Se rechaza Ho R/ Existe relación entre las variables

2.5. Prueba de independencia CHI CUADRADO con la App ESTADÍSTICA


6:22
85%
Regresión con suma
Distribución de frecuencias
Binomial
Poisson
Distribución normal
Prueba T una media
Prueba T dos medias
Prueba Chi-Cuadrado


Nº de líneas
3
Nº de columnas
2
Nivel de significación
0.05
OK
Valores obtenidos
1 2 3
4 5 6
7 8 9
0


3
2
Nivel de significación
0.05
Línea: 1 Columna: 1
180
OK
1 2 3
4 5 6
7 8 9
0


Prueba Chi-Cuadrado
Paso 1
H0: Las variables son independientes
H1: Las variables no son independientes
Paso 2
Formula Qui²
Paso 3
RC > 5.992
Paso 4
Qui² calculado = 47.2433
Paso 5
Qui Cuadrado calculado pertenece a la región crítica, por lo tanto, se rechaza H0, es decir, las variables no son independientes a nivel de significancia 0.05
P-Valor: 0,0000
OTRA OPERACIÓN

Prueba de hipótesis X^2: (Probar si hay relación entre las variables)

Planteamiento de la Hipótesis

H_0: No existe relación entre el turnos y los niveles de educación.

H_1: Existe relación entre el turnos y niveles de educación.

Nivel de significación

$\alpha = 0.05$

Regla de decisión

Si Pvalor < α rechazo (H_0)

Toma de decisión

0.000 **<** 0.05 **entonces** se rechaza (H_0),

R/ Se acepta que hay relación entre el **turnos y el calidad de las piezas**.

2.6. Prueba Chi- Cuadrado de homogeneidad o de proporción de varias muestras.

Ejemplo 2. Prueba de proporción de varias muestras

Un taller de una industria azucarera fabrica piezas metálicas para la zafra y desea evaluar si existe una relación entre los turno de producción (mañana, tarde y noche) y la calidad de las piezas producidas. Para ello, **se seleccionó 3 muestras** (una de cada turno) y se clasificaron las piezas según la calidad de ellas (no defectuosas y no defectuosas), obteniendo la siguiente tabla de contingencia de 3x2:

Turnos categorías: mañana, tarde, noche.

Calidad de las piezas categorías: no defectuosas, o defectuosas.

TABLA DE CONTINGENCIA

		calidad de las piezas	
		no defectuosas	**defectuosa**
turnos	**mañana**	180	100
	tarde	190	280
	noche	170	120

Prueba de hipótesis X^2: (Probar si hay relación entre las variables)

Planteamiento de la Hipótesis

H_0: $p_1 = p_2 = p_3$ (la proporciones de piezas defectuosas son iguales por turnos)

H_1: p_1, p_2, p_3 diferentes (las proporciones de piezas defectuosas no son iguales por turnos)

Nivel de significación

$\alpha = 0.05$

Regla de decisión

Si Pvalor < α rechazo (H_0)

Toma de decisión

0.000 **<** 0.05 **entonces** se rechaza (H_0),

R/ La proporción de piezas defectuosos no son iguales en cada turno.

Conclusiones

CONCLUSIONES

- **La prueba Chi-Cuadrado** es una prueba no paramétrica para **variables categóricas.**
- **PRUEBA DE BONDAD DE AJUSTE:** compara la distribución de una muestra con la distribución teórica que se supone que describe a la población de la cual se extrajo.
- **PRUEBA DE INDEPENDENCIA DE DOS VARIABLES:** comprueba si dos o más variables que se miden en una misma muestras están o no relacionadas entre sí. Siempre hablamos de variables cualitativas.
- **PRUEBA DE HOMOGENEIDAD:** comprueba si dos o más muestras pertenecen a la misma población o si sus proporciones son iguales o no

CONCLUSIONES

- En la prueba de independencia para construir la TC, al emplear una sola muestra, solo se conoce el total general pero no se conoce de antemano los marginales ni de las filas ni de las columnas.
- Mientras que en la prueba de Homogeneidad o de proporciones como son varias muestras se conocen los marginales que coinciden con el tamaño de las muestras
- Mientras mayor sea la diferencia entre los valores observados y los esperados, mayor será el valor de $\chi 2$ y aumentará por tanto la probabilidad de rechazar la hipótesis de nula
- No se puede usar la prueba de $\chi 2$ cuando el valor esperado en alguna celda es menor que 5.

Estudio independiente

Un taller de la industria deportiva desea evaluar si existe una relación entre la materia prima (MP1, MP2 y MP3) empleada en la fabricación de pelotas de béisbol y su calidad (cumplen la norma y no cumplen la norma). Para ello, evalúan 500 pelotas de cada materia prima y se conforma la siguiente tabla de contingencia. Habrá igual proporción de pelotas que cumplen la norma entre cada materia prima.:

TABLA DE CONTINGENCIA

		calidad de las pelotas de béisbol	
		cumplen la norma	**no cumplen la norma**
Materia Prima	**MP 1**	485	15
	MP 2	480	20
	MP 3	475	25

Muchas gracias por su atención

UNIVERSIDAD DE ARTEMISA

Estadística II

MSc. Rodolfo Pérez Léon
email: rperezleon@uart.edu.cu
rplrpl55@gmail.com

INGENIERÍA INDUSTRIAL

Asignatura: Estadística II **Año:** 3er.

Tema 3. Análisis de Correlación

3.1. Introducción. Análisis de Correlación

3.2. Coeficiente de Correlación de Pearson

3.3. Prueba de hipótesis de independencia de variables

3.4. Coeficiente de Correlación de Spearman

3.5. Coeficiente de Correlación de Kendall.

Objetivos: Que el alumno se capaz de:

Calcular los coeficientes de correlación, de Pearson, Spearman, y Kendall, utilizando la forma algebraica o aplicaciones digitales estadísticas, interpretando sus resultados.

Método: Elaboracion Conjunta.

Medios: Tiza y pizarra, Libro de texto, MDD, telef. moviles.

Forma de enseñanza: Clase Encuentro

- **Probabilidades y Estadística para ingenieros. J.E Freund, I.R. Miller, R. Jonhnson.**
- **Materiales Didácticos Digitales entregados por el profesor**

UNIVERSIDAD DE ARTEMISA

3.1. Introducción. Análisis de Correlación

3.1. Introducción. Análisis de Correlación

Hasta ahora nuestro estudio se centró en el análisis de una solo variable, de manera, que a este tipo de estudio se le conoce como **análisis univariado**, pero en la mayoría de los problemas que se presentan en la vida surge la necesidad de observar el comportamiento de 2 ó más variables y estudiar la relación que existe entre ellas.

Ejemplo:

Relación entre el aumento de la temperatura y la emisisón de CO_2.

Relación entre el rendimiento laboral de un empleado con el tiempo dedicado a la capacitación.

Relación entre los gastos mensuales de una familia y los ingresos mensuales de ella.

Relación entre la demanda de un producto y la calidad él.

Para realizar este tipo de estudio existen **el análisis multivariado**, que cuenta con diversas técnicas estadísticas, una de ellas es:

Análisis de Correlación.

Este análisis mide la fuerza o grado de asociación que existe entre las variables. El **análisis de correlación** intenta medir la fuerza de tales relaciones entre dos variables por medio de un solo numero denominado **coeficiente de correlación** que mide tanto la fuerza como la dirección de la relación. En el estudio se abordan tres coeficientes de correlación: Pearson, Spearman, Kendall

UNIVERSIDAD DE ARTEMISA

3.2. Coeficiente de correlación de Pearsón (rho)

3.2. Coeficiente de correlación de Pearsón (rho), evalúa la relación lineal entre dos variables continuas. Una relación es lineal cuando un cambio en una variable se asocia con un cambio proporcional en la otra variable. Por ejemplo, usted puede usar una correlación de Pearson para evaluar si las temperaturas en las instalaciones de producción están asociados con los accidentes laborales en ellas

Se parte de los supuestos siguientes:

- Los datos siguen una distribución Normal.
- Son independientes unos de otros.
- No existen datos anómalos

Expresión matemática:

$$r_{xy} = \frac{n\sum X * Y - (\sum X)(\sum Y)}{\left[\sqrt{n\sum X^2 - (\sum X)^2}\right]\left[\sqrt{n\sum Y^2 - (\sum Y)^2}\right]}$$

Parámetro (Población)	Estadístico (Muestra)	Valores del coeficiente	Grado de la asociación	Ejemplo
ρ	r	-1	Fuerte asociación negativa	Ahorro personales de salario mensual y gastos personales
		0	No hay asociación	Ahorro personales de salario mensual y color del pelo
		1	Fuerte Asociación positiva	Calidad de un producto y demanda del producto

Valor r	Fuerza de la correlación
0.0 < 0.1	no hay correlación
0.1 < 0.3	poca correlación
0.3 < 0.5	correlación media
0.5 < 0.7	correlación alta
0.7 < 1	correlación muy alta

Ejemplo 1

Suponga que quiere analizar si existe relación entre la velocidad de una línea de producción (X) y el número de defectos por lote (Y). Se recolectaron los siguientes datos:

X: velocidad de una línea de producción

Y: número de defectos por lotes

Velocidad (m/min	Defectos (Y)	XY	X^2	Y^2
10	5	50	100	25
12	7	84	144	49
15	9	141	225	81
18	10	180	324	100
20	15	300	400	225
75	46	749	1193	480

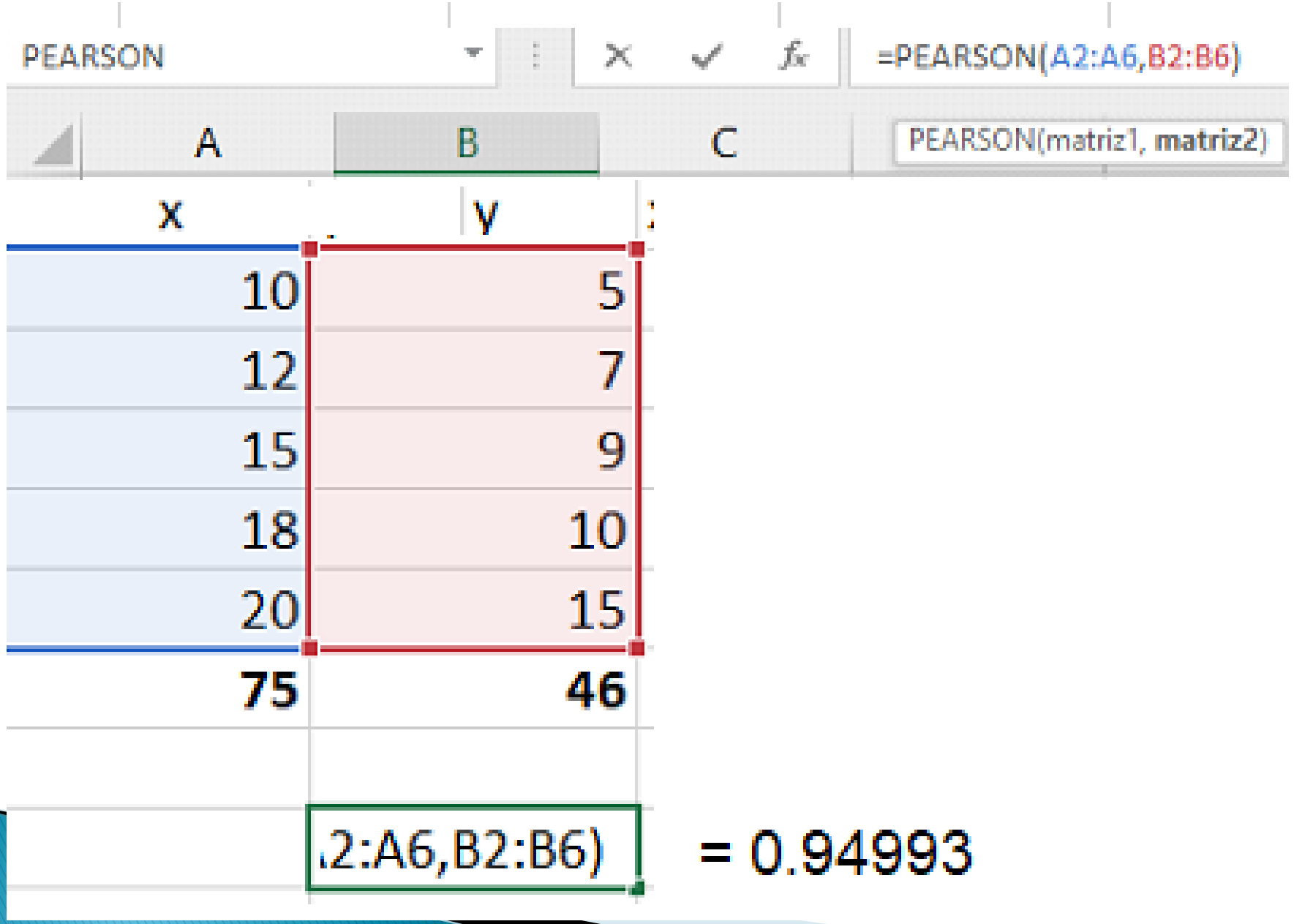
UNIVERSIDAD DE ARTEMISA

Solución: Algebraica

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\sum x^2 - \left(\sum x\right)^2} * \sqrt{n\sum y^2 - \left(\sum y\right)^2}}$$

$$r = \frac{5\sum 749 - \left(\sum 75\right)\left(\sum 46\right)}{\sqrt{5\sum 1193 - \left(\sum 75\right)^2} * \sqrt{5\sum 480 - \left(\sum 46\right)^2}} = 0.9499$$

Solución empleando Excel



Solución empleando aplicaciones digitales

Correlaciones

		VARX	VARY
VARX	Correlación de Pearson	1	0,940
	Sig. (bilateral)		,000
	N	5	5

SPSS

R/ Existe una fuerte correlación positiva entre la velocidad de la línea de producción y el número de defectos por lotes.

Nota: es importante señalar que la correlación entre variables no significa causalidad. Ejemplo cada año se gasta más en ciencia y esto no es la causa que también aumenta la temperatura promedio del planeta.

Pero esa relación se da para la muestra de datos seleccionada, pero si quisiéramos ver si sucede igual a toda la población tendríamos que hacer una prueba de hipótesis de correlación

UNIVERSIDAD DE ARTEMISA

3.3.

Prueba de hipótesis de independencia de variables

UNIVERSIDAD DE ARTEMISA

3.3. Prueba de Hipótesis de independencia (no relación de variables)

1er Paso: Planteamiento de la Hipótesis.

H_o: $\rho = 0$ las variables son independientes (no hay relación entre las variables)

H_i: $\rho \neq 0$ las variables son dependientes (hay relación entre las variables)

2Do Paso: Fijar el Nivel de significación α

α= 0.05

3er Paso: Regla de decisión

Si pvalor < α Rechazo Ho

4to Paso: Toma de decisión.

0.000 < 0.05 entonces se rechaza Ho.

R/ Existe una correlación entre las variables velocidad de la línea de producción y el número de defectos por lotes, con una confianza del 95%

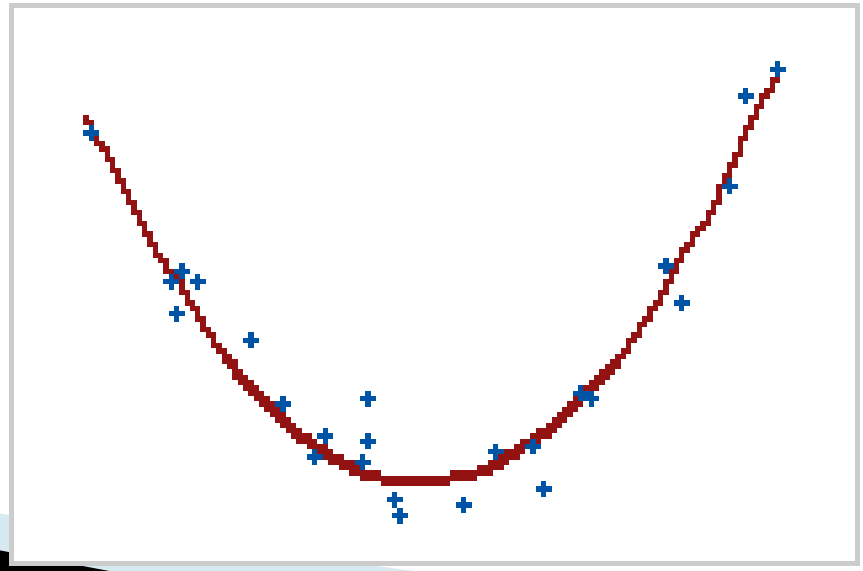
UNIVERSIDAD DE ARTEMISA

3.4. Coeficiente de correlación de Spearman (r_s)

3.4. Coeficiente de correlación de Spearman (r_s).

<u>Características de su empleo</u>

- Mientras que el coeficiente de correlación de Pearson solo miden relaciones lineales en cambio el coeficiente de correlación de Spearman (r_s) mide la relación monótona (no necesariamente lineal)



- Puede medir la relación entre dos variables cuantitativa y también puede medir esa relación entre una variable cuantitativa y otra cualitativa.
- Los dato no tiene que seguir una distribución Normal.
- Ideal para datos ordinales o no normales.
- No se afecta ante datos anómalos.

Expresión Matemática

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Pasos para el cálculo del coeficiente de correlación de Spearman (s_r).

1. Recoger los datos.
2. Asignar rangos.
3. Calcular la diferencia de rangos (d).
4. Elevar la diferencia (d) al cuadrado.
5. Aplicar la expresión matemática o fórmula.

Ejemplo 2

Se quiere evaluar si el nivel de experiencia de operadores (X) está relacionado con el tiempo de ensamblaje (Y). Los datos son:

Experiencia	Tiempo	Rango X	Rango Y	Diferencia de rango (d)	d^2
Bajo	25	1	3	-2	4
Media	20	2	2	0	0
Alto	15	3	1	2	4
				total	8

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$r_s = 1 - \frac{6\sum 8}{3(3^2-1)} = -1$$

R/ Existe una correlación negativa perfecta entre la experiencia de los operadores y el tiempo de ensamblaje, a medida que aumenta la experiencia disminuye el tiempo de ensamblaje

Ejemplo 3.

Analizar si existe una relación entre el número de horas que un grupo de estudiantes universitarios dedica al estudio y sus notas finales en un examen.

1. Recoger los datos:

Estudiante	Horas de Estudio (X)	Nota Final (Y)
A	4	78
B	6	85
C	2	72
D	8	88
E	5	80

2. Convertir los Datos en Rangos:

Estudiante	Horas de Estudio (X)	Rango de X	Nota Final (Y)	Rango de Y
A	4	2	78	2
B	6	4	85	4
C	2	1	72	1
D	8	5	88	5
E	5	3	80	3

3. Calcular las Diferencias de Rangos:

4. Elevar al cuadrado las Diferencias de Rango:

Estudiante	Rango de X	Rango de Y	d_i	$(d_i)^2$
A	2	2	0	0
B	4	4	0	0
C	1	1	0	0
D	5	5	0	0
E	3	3	0	0

5. Aplicar la formula: $$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$r_s = 1 - \frac{6 \times 0}{5(5^2-1)}$$

$$r_s = 1 - \frac{0}{120} = 1$$

R/ Existe una relación perfecta positiva entre las horas de estudio de los estudiantes y las notas que obtienen

UNIVERSIDAD DE ARTEMISA

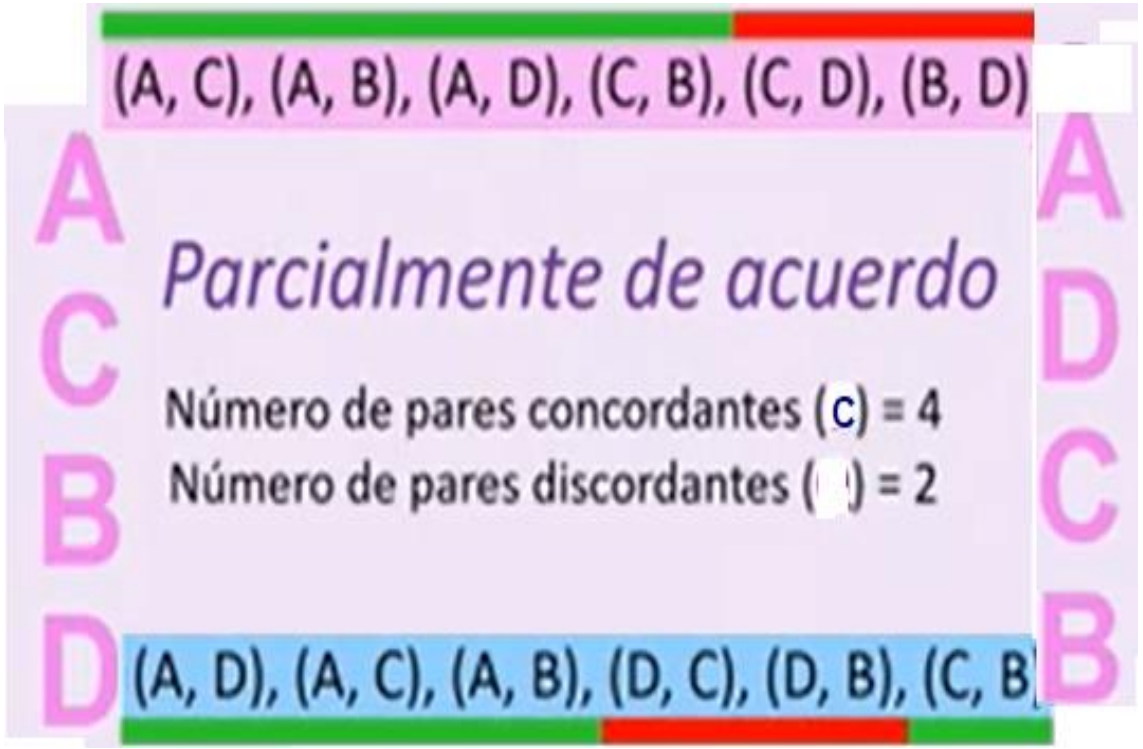
3.5.

Coeficiente de correlación de Kendall (Ƭ) (tau)

3.5 Coeficiente de correlación de Kendall (tau), comúnmente conocido como coeficiente τ de Kendalll (con la letra griega τ, tau), es utilizado para medir **la asociación ordinal entre dos cantidades medidas**. Por lo que los datos no tienen que estar normalmente distribuidos.

Por ejemplo, dos inspectores diferentes examinan 45 Tv para determinar los defectos encontrados en las pantallas. ¿Concuerdan consistentemente los inspectores en su clasificación?

Ideal para medir la similitud entre dos conjuntos de ranking ejemplo, dos evaluadores determinan la calidad de 4 marcas de equipos (A,B,C,D) y la ordenan según su ranking.



Podemos calcular la **tau de Kendall** con esta **fórmula**:

$$\tau = \frac{C - D}{C + D}$$

Donde **C** es el número de pares concordantes

y **D** es el número de pares discordantes

T = (4-2) / (4+2) = 2 / 6 = 0.33

R/ Existe una cierta concordancia entre las evaluaciones de ambos evaluadores.

R/ Existe una cierta concordancia entre las evaluaciones de ambos evaluadores.

El Coeficiente de Kendal es muy similar que el de Spearman, pero se deberia preferir
en el caso que se disponen de muy pocos datos y hay muchos empates de rangos.

CONCLUSIONES

CONCLUSIONES

- *Pearson:* Ideal para relaciones lineales en datos paramétricos (ej: control de calidad). Los datos distribuidos normalmente, variable cuantitativas

- *Spearman:* Para tendencias no lineales (ej: rankings de eficiencia). Los datos no necesariamente distribuidos normalmente, variable cuantitativas y también cualitativas

- *Kendall:* Útil con muestras pequeñas o empates (ej: logística). Variables ordinales.

UNIVERSIDAD DE ARTEMISA

ESTUDIO INDEPENDIENTE

ESTUDIO INDEPENDIENTE.

En un proceso de fabricación, se cree que a mayor temperatura (X), mayor es la resistencia (Y) del material. Calcule e interprete el coeficiente de correlación de Pearson y de Spearman con los datos siguientes:

(X): 100, 120, 150, 180 °C

(Y): 50, 55, 60, 65 MPa

UNIVERSIDAD DE ARTEMISA

MUCHAS GRACIAS

Prof Aux. Rodolfo Pérez León
email: rperezleon@uart.edu.cu
rplrpl55@gmail.com

UNIVERSIDAD DE ARTEMISA

ESTADISTICA II

MSc. Rodolfo Pérez Léon
email: rperezleon@uart.edu.cu
rplrpl55@gmail.com

UNIVERSIDAD DE ARTEMISA

Tema 4. Modelo de Regresión Lineal General. Sumario.

4.1. Introducción. Análisis de Regresión. El Modelos de Regresión Lineal General (MRLG).

4.2. El Modelo de Regresión Lineal General (MRLG). Supuestos del modelo.

4.3. Cálculo de los estimadores por MMCO, en su forma algebraica. Ecuación de Regresión. Propiedades de los estimadores MCO. El Coeficiente de Regresión. Pronósticos o estimación a partir de la Ecuación de Regresión.

UNIVERSIDAD DE ARTEMISA

Tema 4. Modelo de Regresión Lineal General.

Sumario.

4.4. Tabla Anova. Coeficiente de Determinación R^2, interpretación.

4.5. Pruebas F_{total} para la verificar la bondad de ajuste de la curva con un nivel de confianza.

UNIVERSIDAD DE ARTEMISA

Objetivos: Que el alumno se capaz de:

Formular un modelo de regresión interpretando sus resultados.

Método: Expositivo.

Medios: Tiza y pizarra, Libro de texto, MDD.

Forma de enseñanza: Conferencia

Bibliografía:

- Estadística pag. 243-247 Dra Caridad Bustillo.
- Materiales Didácticos Digitales entregados por

4.1.

Introducción. Análisis de Regresión. El Modelos de Regresión Lineal General (MRLG)

UNIVERSIDAD DE ARTEMISA

Estadistica II

4.1. Introducción. Análisis de Regresión, clasificación de los Modelos de Regresión.

Hasta ahora nuestro estudio se centró en el análisis de una solo variable, de manera, que a este tipo de estudio se le conoce como **análisis univariado**, pero en la mayoría de los problemas que se presentan en la vida surge la necesidad de observar el comportamiento de 2 ó más variables y estudiar la relación que existe entre ellas.

UNIVERSIDAD DE ARTEMISA

Estadistica II

Ejemplo:

- *La relación que existe entre los gastos mensuales de una familia y los ingresos mensuales de ella.*
- *Las emisiones de CO2 y la temperatura ambiental.*
- *Demanda de un producto contra el precio de él.*
- *Rendimiento académico de un estudiante con el tiempo dedicado al estudio.*

UNIVERSIDAD DE ARTEMISA

Estadística II



UNIVERSIDAD DE ARTEMISA

Análisis de Correlación

EL análisis de Correlación mide la fuerza o grado de asociación que existe entre las variables. Para ello se emplea el coeficiente de **CORRELACIÓN DE PEARSÓN (r).**

UNIVERSIDAD DE ARTEMISA

Análisis de Correlación

La expresión matemática del coeficiente de correlación de Pearsón es:

$$r = \frac{\sum xy - \left(\sum x\right)\left(\sum y\right)/n}{\sqrt{\sum x^2 - \left(\sum x\right)^2/n}\sqrt{\sum y^2 - \left(\sum y\right)^2/n}}$$

UNIVERSIDAD DE ARTEMISA

Análisis de Correlación

Parámetro (Población)	Estadístico (Muestra)	Valores del coeficiente	Grado de la asociación	Ejemplo
ρ	r	-1	Fuerte asociación negativa	Ahorro personales de salario mensual y gastos personales
		0	No hay asociación	Ahorro personales de salario mensual y color del pelo
		1	Fuerte Asociación positiva	Ahorro personales salario mensual y salario mensual

UNIVERSIDAD DE ARTEMISA

Análisis de Regresión

En este Curso vamos a centrarnos en el estudio de otra técnica estadística que es:

Análisis de Regresión.

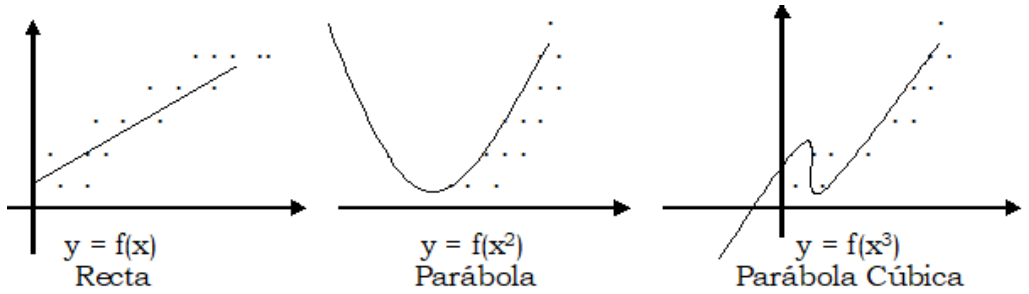
Permite determinar un modelo matemático, que represente la relación entre dos o más variables.

Con este modelo se pueden realizar pronósticos, por eso es considerada una herramienta prospectiva.

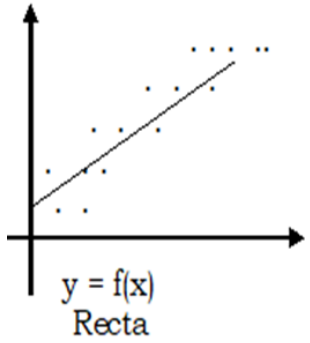
UNIVERSIDAD DE ARTEMISA

Modelo de Regresión Lineal General (MRLG).

Los Modelos de regresión se clasifican en modelos de regresión lineales y no lineales

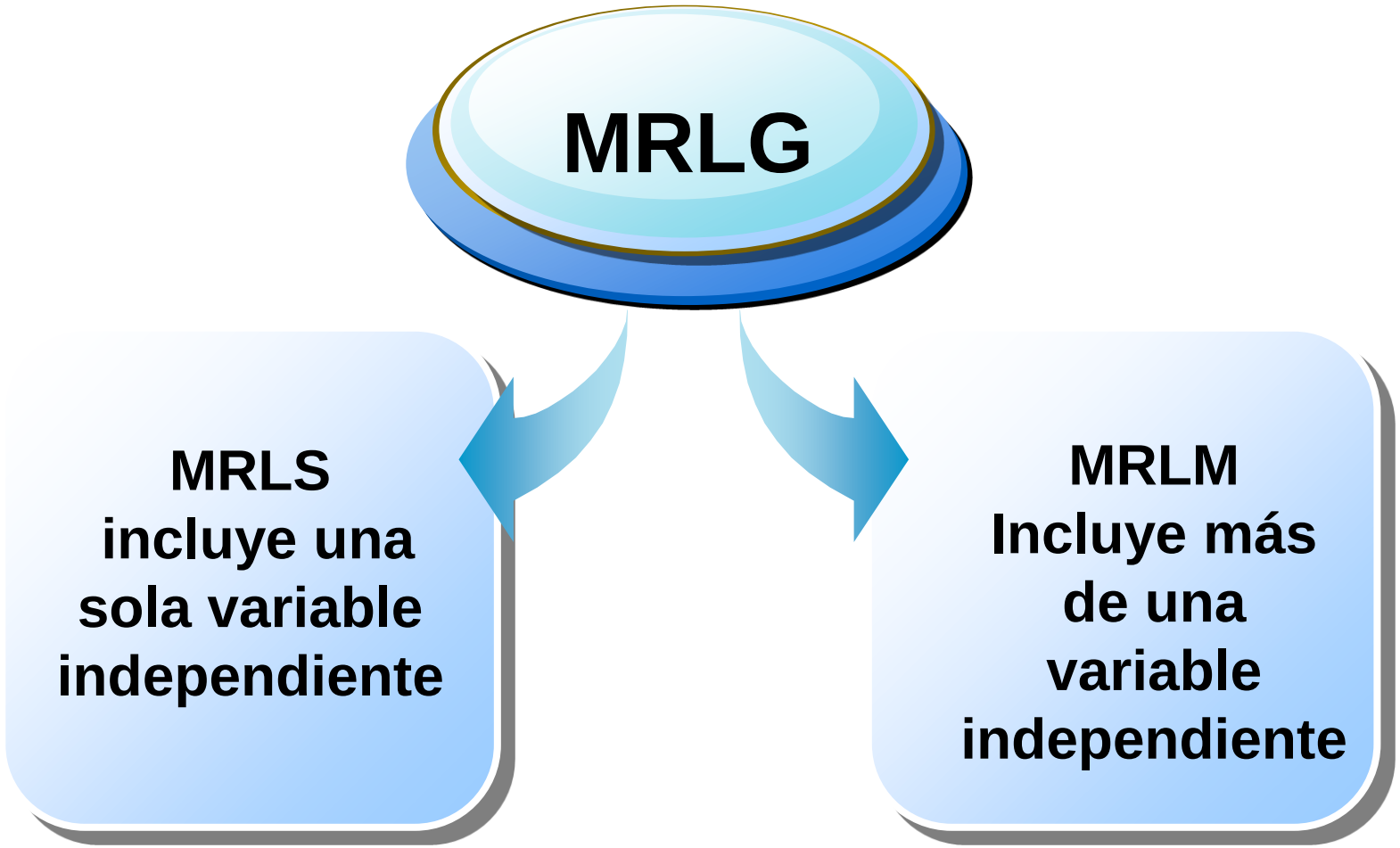


El tema aborda el Modelo de Regresión Lineal General (MRLG). Lineal porque la relación entre las variables dependiente y las independientes sigue una línea recta.



UNIVERSIDAD DE ARTEMISA

Modelo de Regresión Lineal General (MRLG).



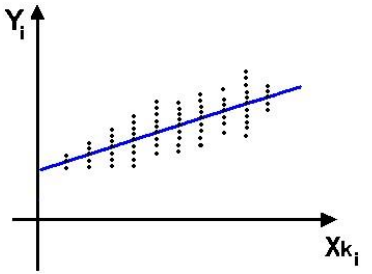
4.2.

El Modelo de Regresión Lineal Simple (MRLS). Supuestos del modelo.

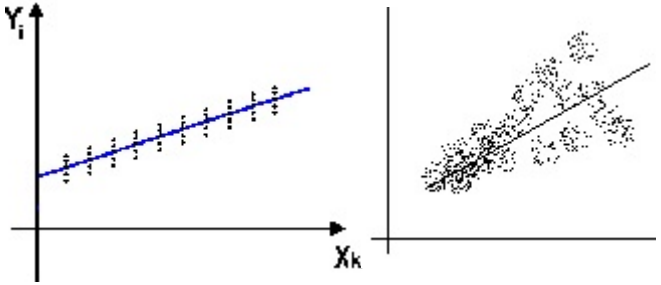
UNIVERSIDAD DE ARTEMISA

Supuestos del MRLG

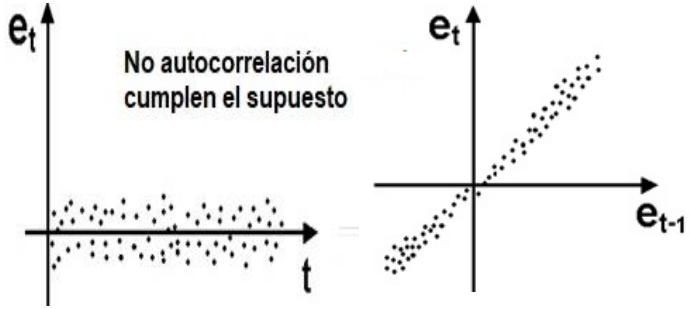
LA MEDIA DE LOS $E_i = 0$



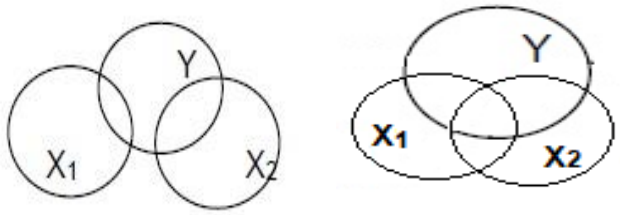
HOMOCEDASTICIDAD



NO AUROCORRELACIÓN



NO MULTICOLINEALIDAD



4.3. Cálculo de los estimadores por MMCO, en su forma algebraica ó por paquetes estadísticos

UNIVERSIDAD DE ARTEMISA

Modelo de Regresión Lineal Simple MRLS

Esta clase se centrará en los (MRLS) el que se expresa matemáticamente como:

$$Y = \beta_0 + \beta_1 X_1 + E_i$$

Donde

Y : es la variable dependiente.

β_0 y β_1: son parámetros.

X_1: variable independiente

Ei: es el error aleatorio o residuo, también en algunas bibliografías se representa por U_i.

UNIVERSIDAD DE ARTEMISA

Modelo de Regresión Lineal Simple MRLS

Ejemplo:

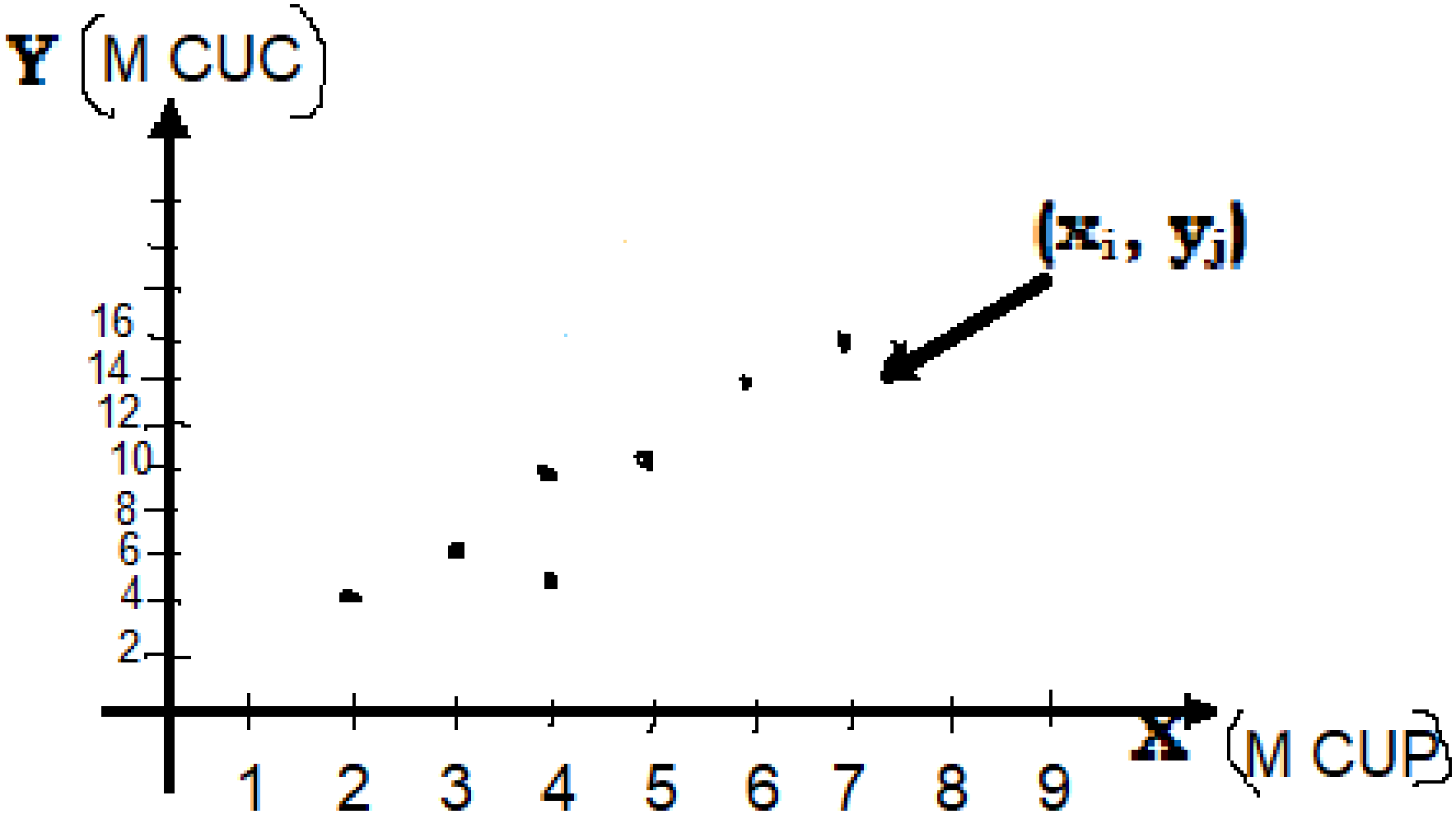
Sea Y: los costos de producción de una fábrica (en MCUP)

X: los gastos de materia prima de la fábrica (en MCUP).

Y	4	7	10	11	16	8	14
X	2	3	4	5	7	4	6

UNIVERSIDAD DE ARTEMISA

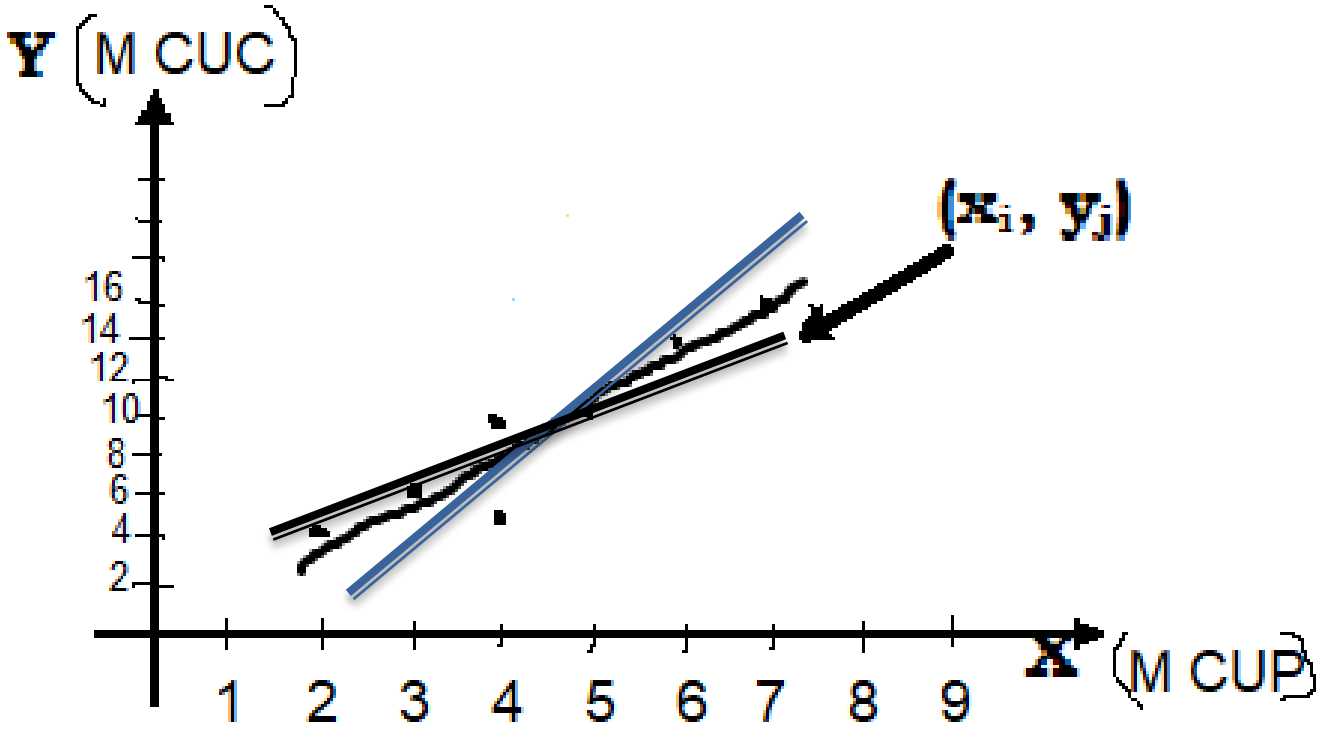
Diagrama de dispersión



UNIVERSIDAD DE ARTEMISA

Modelo de Regresión Lineal Simple MRLS

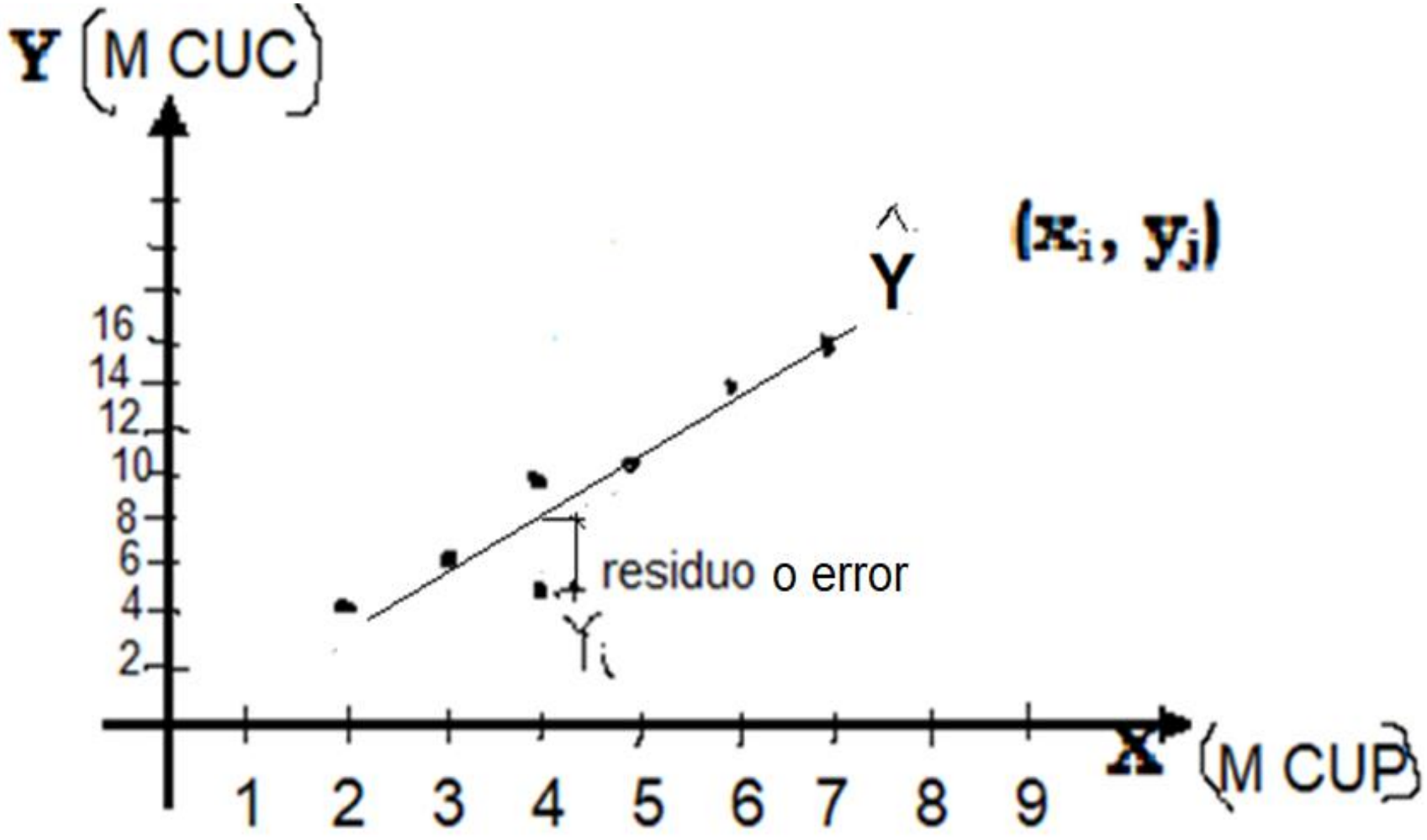
Por esta nube de puntos se pueden pasar infinitas rectas



UNIVERSIDAD DE ARTEMISA

Modelo de Regresión Lineal Simple MRLS

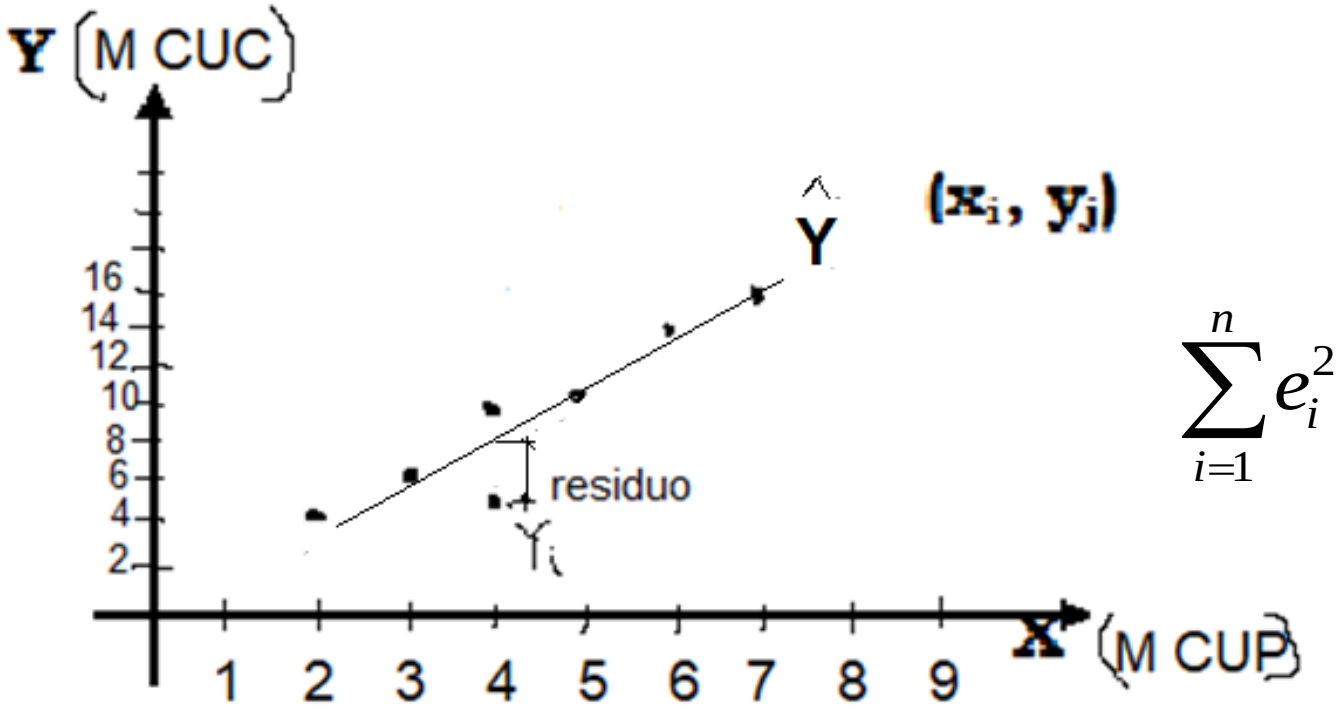
Se trata de encontrar aquella recta que haga mínima la suma al cuadrado de los residuos.



UNIVERSIDAD DE ARTEMISA

Modelo de Regresión Lineal Simple MRLS

El método de los Mínimos cuadrados ordinarios (MCO) hace mínima el cuadrado de los residuos



$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

UNIVERSIDAD DE ARTEMISA

Modelo de Regresión Lineal Simple MRLS

Y proporciona los estimadores mínimos cuadrados (b_o y b_1) de los parámetros (β_1 y β_2) para conformar la ecuación de la recta

$$Y = \beta_0 + \beta_1 X_1 + E_i$$

Parámetros

$$\hat{Y} = b_0 + b_1 X$$

Estimadores

UNIVERSIDAD DE ARTEMISA

Estimadores MCO

Expresión matemática de los estimadores MCO (algebraica)

$$b_1 = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X_i^2 - \left(\sum X\right)^2}$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

Regresemos al problema anterior

X_i	Y_i	XY	X^2	Y^2
2.0	4.0	8.0	4.0	16.0
3.0	7.0	21.0	9.0	49.0
4,0	10.0	40.0	16.0	100.0
5.0	11.0	55.0	25.0	121.0
7.0	16.0	112.0	49.0	256.0
4.0	5.0	20.0	16.0	25.0
6.0	14.0	84.0	36.0	196.0
31.0	67.0	340.0	155.0	763.0

X_i	Y_i	XY	X^2	Y^2
31.0	67.0	340.0	155.0	763.0

$$b_1 = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X_i^2 - \left(\sum X\right)^2}$$

$b_1 = 7(340) - 31(67) \;/\; 7(155) - (31)^2 =$

$43.29 / 17.71 = 2.44$

$$b_0 = \bar{Y} - b_1\bar{X}$$

$= (67 / 7) - 2.44(31 / 7) = -1.21$

UNIVERSIDAD DE ARTEMISA

Estimadores MCO

$$\hat{Y} = \beta_0 + \beta_1 X_1 + E_i$$

$$\hat{Y} = b_0 + b\, x_1 + e_i$$

$$\hat{Y} = -1.21 + \underline{2.44}\, X_1$$

UNIVERSIDAD DE ARTEMISA

Estimadores MCO

Propiedades de los estimadores Mínimos Cuadrados Ordinarios (MCO).

- Todo estimadores MCO es Insesgado de βj.
- Los estimadores MCO son de varianza mínima (eficientes).
- Si el error aleatorio (residuo e_i) se distribuye normalmente son también de máxima verosimilidad o sea Mejor Estimador lineal Insesgado (MELI).

UNIVERSIDAD DE ARTEMISA

Coeficiente de regresión

b_1 (**COEFICIENTE DE REGRESIÓN**): es el estimador de β_1 y su interpretación es:

$$\hat{Y} = \quad -1.21 + \underline{\mathbf{2.44}}\ X_1$$

a medida que aumenta en una unidad la variable X_1 aumentará en 2.44 unidades la variable Y.

UNIVERSIDAD DE ARTEMISA

Estimadores MCO

su interpretación en el contexto del ejercicio será:

$$\hat{Y} = -1.21 + \underline{2.44}\ X_1$$

a medida que aumenta en mil CUP los gastos de materia prima aumentarán en 2.44 miles de CUP los costos de producción.

UNIVERSIDAD DE ARTEMISA

Ecuación de regresión

Estimación o Pronósticos con el MRLG.

Este Modelo sirve para realizar pronósticos.

Ejemplo: Cuál será el gasto de producción cuando los gastos de materia prima sean de 8 MCUP

$$\hat{Y} = \text{-1.21} + \underline{2.44}\ X_1$$

$$\hat{Y} = \text{-1.21} + \underline{2.44}\ (8)$$

$$\hat{Y} = 18.31 \text{ MCUP}$$

R/ Cuándo los gastos de materia prima sean de 8 MCUP los gastos de producción serán de 18.31 CUP

4.4. Tabla ANOVA. Coeficiente de Determinación R^2.

UNIVERSIDAD DE ARTEMISA

Tabla ANOVA

Tabla de Análisis de Varianza.

Causas de variación	Suma de cuadrados (SC)	Grados de libertad (GL)	Cuadrado medio (CM)	F(calc)	p-valor
Debida a la Regresión	$SC\,\hat{y}$ $b_1(\Sigma xy - n\,\bar{Y}\,\bar{X})$	p – 1	$S\,\hat{y}^2$ $b_1(\Sigma xy - n\,\bar{Y}\,\bar{X}) / p\text{-}1$	S^2_{reg} / S_e^2	
Residual o error	SC_e $SC_T - SC\,\hat{y}$	n – p	S_e^2 $SC_T - SC\,\hat{y} / n\text{-}p$		
Total	SC_T $\Sigma Y_i^2 - n\,\bar{Y}^2$	n - 1	**Varianza del error estandar**		

S_e^2 : Este término recordemos se conoce como la **varianza del error estándar .**

Entonces (**Se**) es el error estándar de la estimación también conocido como error típico (**Et**)

UNIVERSIDAD DE ARTEMISA

Coeficiente de Determinación R^2

El Coeficiente de Determinación R^2 Interpretación

$$R^2 = SC_y / SC_t$$

El Coeficiente de Determinación (R^2) nos indica el % de la variación de la (Y) que es explica por las variable (X_i)

UNIVERSIDAD DE ARTEMISA

Coeficiente de Determinación R^2

R^2	Indica
$R^2 \geq 85\ \%$	Buen ajuste
$65\ \% \leq R^2 < 85\ \%$	Ajuste Aceptable
$R^2 < 65\ \%$	La ecuación no es buena
$R^2 = 0$	No hay relación entre las variables

UNIVERSIDAD DE ARTEMISA

Coeficiente de Determinación R^2

Ejemplo $R^2 = 97.5\%$

Interpretación: El 97.5% de la variación de los costos de producción quedan explicados por la variable gastos de materia prima.

Los paquetes estadísticos ofrecen además un valor de R^2 ajustado que se emplea para comparar la bondad de ajustes entre distintos modelos con diferentes variables y diferente números de ellas.

4.5. Pruebas F_{total} para la verificar la bondad de ajuste de la curva con un nivel de confianza

UNIVERSIDAD DE ARTEMISA

Prueba F total

Pruebas F_{total} para la verificar la bondad de ajuste de la curva.

Como se estudió anteriormente el R^2 solo ofrece una idea de la bondad de ajuste de la curva, Para tener confianza en el modelo se realiza la Prueba F_{total} con la que si ofrece confianza por lo general continuaremos trabajando con un nivel de confianza $(1- \alpha) = 0.95$

UNIVERSIDAD DE ARTEMISA

Prueba F total

1er Paso Planteamiento de la Hipótesis

H_0: $\beta_1 = 0$ (la ecuación no es de buen ajuste)

H_1: $\beta_1 \neq 0$ (la ecuación es de buen ajuste)

2do Paso Determinación del nivel de significación

$\alpha = 0.05$

3er Regla de decisión

Si $P_{valor} < \alpha$ se rechaza H_0 para el α dado.

4to. Paso Toma de la Decisión

Si $P_{valor} < \alpha$ se rechaza H_o concluyendo:

R/ La ecuación es de buen ajuste para una confianza del 95 %.

CONCLUSIONES

UNIVERSIDAD DE ARTEMISA

Conclusiones

- El Análisis de Regresión permite determinar un modelo matemático que representa la relación entre dos o más variables, aunque en esta clase solo tratamos con una variable independiente.

- El modelo matemático del MRLS es:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + E_i$$

- Debe cumplir un conjunto de supuestos para que tengan validez prospectiva.

UNIVERSIDAD DE ARTEMISA

Conclusiones

- El Método de los Mínimos Cuadrados Ordinarios (MMCO) permite formular la ecuación de MRLS, calculando los estimadores MCO de los parámetros.

$$b_1 = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X_i^2 - \left(\sum X\right)^2}$$

$$b_0 = \overline{Y} - b_1\overline{X}$$

UNIVERSIDAD DE ARTEMISA

Conclusiones

- Las pruebas F_{total} se realiza para verificar la bondad de ajuste de una curva (Modelo).

- El Coeficiente de Determinación **(R^2)** indica la **variación** de la **variable dependiente** que **queda explicada** por las **variables independientes**.

UNIVERSIDAD DE ARTEMISA

Conclusiones

Tabla de Análisis de Varianza.

Causas de variación	Suma de cuadrados (SC)	Grados de libertad (GL)	Cuadrado medio (CM)	F(calc)	p-valor
Debida a la Regresión	$SC\,\hat{y}$ $b_1(\Sigma xy - n\,\bar{Y}\,\bar{X})$	$p - 1$	$S\hat{y}^2$ $b_1(\Sigma xy - n\,\bar{Y}\,\bar{X}) / p\text{-}1$	S^2_{reg} / S_e^2	
Residual o error	SC_e $SC_T - SC\,\hat{y}$	$n - p$	S_e^2 $SC_T - SC\,\hat{y} / n\text{-}p$ **Varianza del error estandar**		
Total	SC_T $\Sigma Y_i^2 - n\,\bar{Y}^2$	$n - 1$			

$$R^2 = SC_y / SC_t$$

p: cantidad de variables
n: cantidad de observaciones

Estudio Independiente:

Conforme la tabla del Análisis de Varianza para el ejercicio desarrollado.

Entregar en la próxima clase con una presentación acorde al nivel universitario

MUCHAS GRACIAS

Prof Aux. Rodolfo Pérez León
email:
rperezleon@uart.edu.cu
rplrpl55@gmail.com

UNIVERSIDAD DE ARTEMISA

ESTADÍSTICA II

MSc. Rodolfo Pérez Léon
email: rperezleon@uart.edu.cu
rplrpl55@gmail.com

UNIVERSIDAD DE ARTEMISA

Tema 4. Modelo de Regresión Lineal General. Sumario.

4.6. Cálculo de los estimadores por MCO por la vía matricial. Matriz Varianza Covarianza.

4.7. Análisis de Regresión con el empleo de la aplicación Excel. Interpretación de los resultados.

4.8. Análisis de Regresión con Apk "Estadística Anova". Interpretación de los resultados.

4.9. Análisis de Regresión con el paquete estadístico SPSS.

4.10. Estimación de los intervalos de confianza para el parámetro (βj) y estimación de Y_k.

UNIVERSIDAD DE ARTEMISA

Objetivos: Que el alumno se capaz de:

Formular un modelo de regresión a travez del método matricial, interpretando sus resultados.

Método: Expositivo.

Medios: Tiza y pizarra, Libro de texto, MDD.

Forma de enseñanza: Conferencia

Bibliografía:

- Materiales Didácticos Digitales entregados por el profesor

4.6.

Cálculo de los estimadores por MCO, por la vía matricial. Matriz varianza covarianza

UNIVERSIDAD DE ARTEMISA

Cálculo de Estimadores por vía matricial

Ejemplo:

Sea Y: los costos de producción de una fábrica (en CUP)

X: los gastos de materia prima de la fábrica (en CUP).

Y	4	7	10	11	16	8	14
X	2	3	4	5	7	4	6

UNIVERSIDAD DE ARTEMISA

Cálculo de Estimadores por vía matricial

$$\hat{y} = b\,x\,; \qquad b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = (x'x)^{-1}\,x'y$$

1er Paso Plantear matriz Columna y matriz Datos

Matriz columna $y = \begin{bmatrix} 4 \\ 7 \\ 10 \\ 11 \\ 16 \\ 5 \\ 14 \end{bmatrix}_{7x1}$ $x = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 7 \\ 1 & 4 \\ 1 & 6 \end{bmatrix}_{7x2}$ Matriz Datos

UNIVERSIDAD DE ARTEMISA

Cálculo de Estimadores por vía matricial

2 do Paso determinar inversa de la transpuesta de X por X

$$(x'x)^{-1}$$

a) Transpuesta de $x'x$

$$x'x = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} = \begin{bmatrix} 7 & 31 \\ 31 & 155 \end{bmatrix}$$

b) Determinante de $x'x$

$$x'x = 7(155) - (31)^2 = |1085 - 961| = 124 \neq 0$$

Entonces Tienen inversa

UNIVERSIDAD DE ARTEMISA

Cálculo de Estimadores por vía matricial

c) Inversa de la transpuesta de X por X

$$(x'x)^{-1} = 1 / x'x \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{bmatrix}$$

$$(x'x)^{-1} = \begin{bmatrix} \frac{155}{124} & \frac{-31}{124} \\ \frac{-31}{124} & \frac{7}{124} \end{bmatrix} = \begin{bmatrix} 1,25 & -0,25 \\ -0,25 & 0,056 \end{bmatrix}$$

UNIVERSIDAD DE ARTEMISA

3er Paso determinar la transpuesta de X por Y

$$x'y = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum x_i y_i \end{bmatrix} = \begin{bmatrix} 67 \\ 340 \end{bmatrix}$$

4to Paso calcular los estimadores MCO

$$b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = (x'x)^{-1} x'y$$

$$b = \begin{bmatrix} 1,25 & -0,25 \\ -0,25 & 0,056 \end{bmatrix} \begin{bmatrix} 67 \\ 340 \end{bmatrix} = \begin{bmatrix} 1,25(67) + (-0,25)(340) \\ -0,25(67) + 0,056(340) \end{bmatrix} = \begin{bmatrix} 83,75 - 85 \\ 16,75 + 19,19 \end{bmatrix} = \begin{bmatrix} -1,25 \\ 2,44 \end{bmatrix}$$

Ecuación de regresión $\hat{Y} = b_0 + b_1 X = -1,25 + 2,44 X$

UNIVERSIDAD DE ARTEMISA

Matriz varianza covarianza

Matriz varianza covarianza

$$\hat{V}(b) = \begin{bmatrix} \hat{V}(b_0) & Cov(b_0\ b_1) \\ Cov(b_0\ b_1) & \hat{V}(b_1) \end{bmatrix}$$

Con esta matriz se puede obtener el error estándar de los estimadores, necesarios para estimar el intervalo de confianza de ellos, como se verá más adelante

UNIVERSIDAD DE ARTEMISA

Matriz varianza covarianza

Varianza del error estándar por la Inversa de la traspuesta de x por x, anteriormente fue calculada

$$\hat{V}(b) = S_e^2 (x'x)^{-1}$$

$$\hat{V}(b) = S_e^2 \cdot \begin{bmatrix} 1,25 & -0,25 \\ -0,25 & 0,056 \end{bmatrix}$$

Varianza del estimador b_0

$$\hat{V}(b) = \begin{bmatrix} \hat{V}(b_0) & Cov(b_0\, b_1) \\ Cov(b_0\, b_1) & \hat{V}(b_1) \end{bmatrix}$$

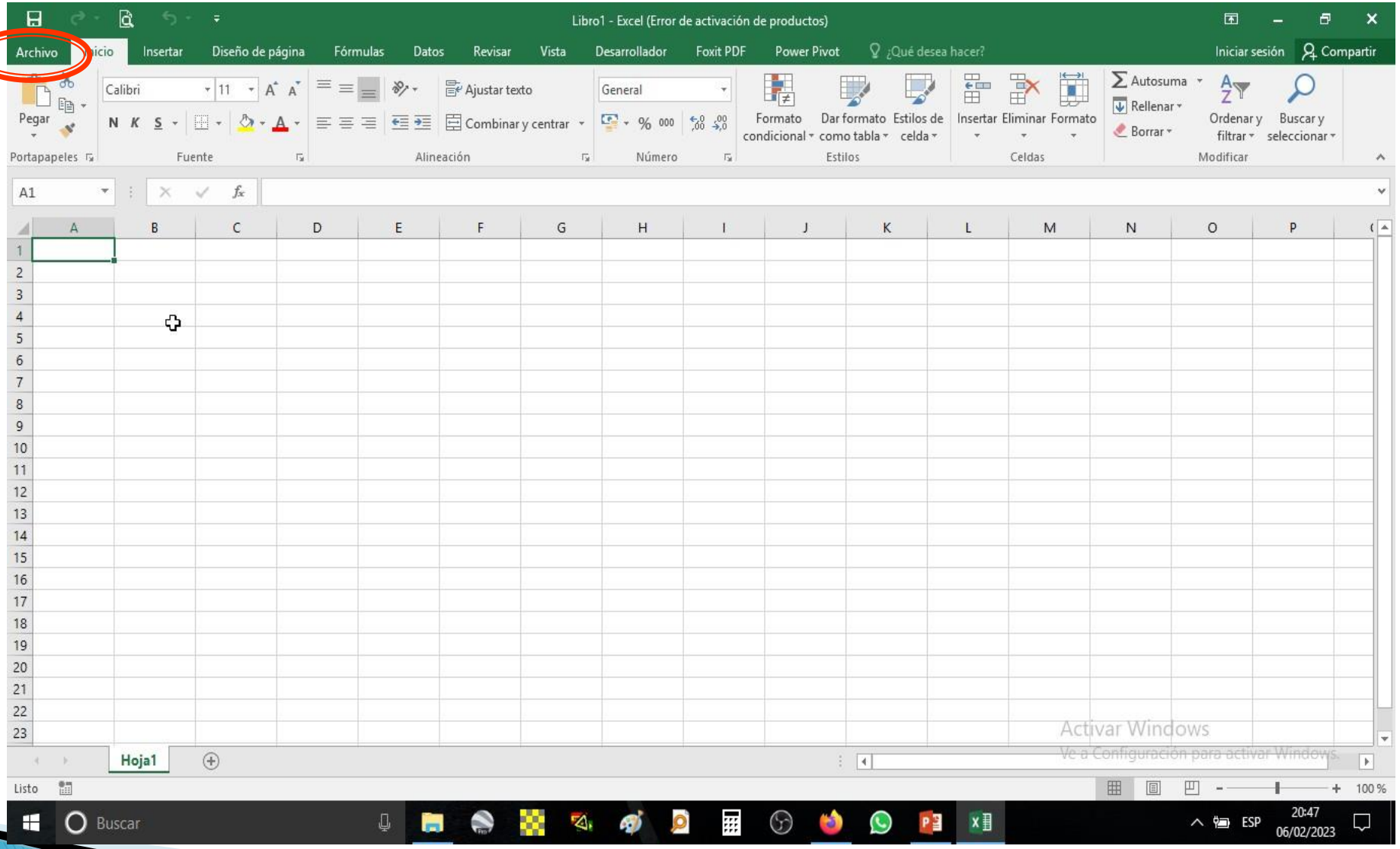
Varianza del estimador b_1



4.7.

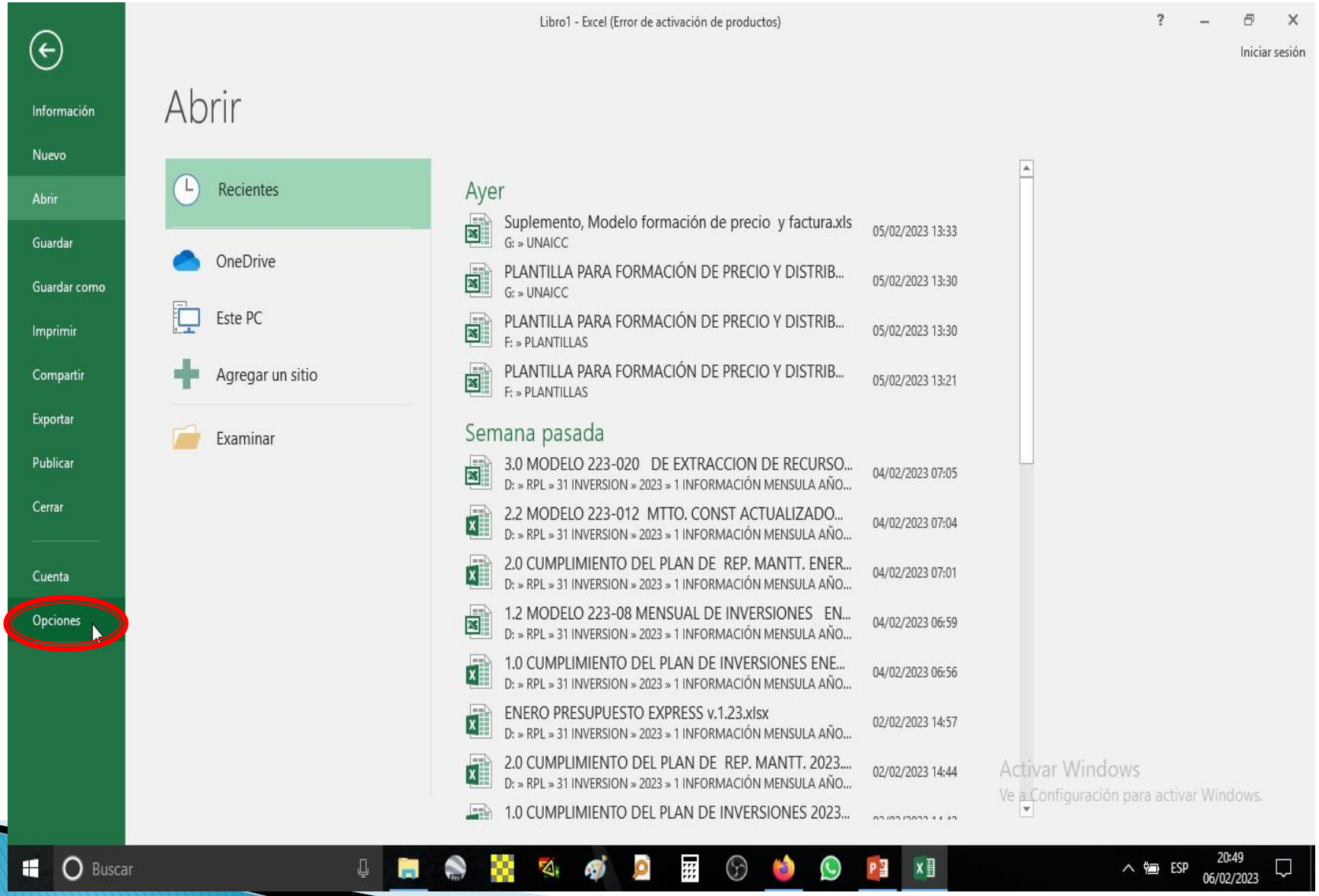
Análisis de Regresión con el empleo de la aplicación Excel. Interpretación de los resultados.

(MRLG) en la aplicación Excel



UNIVERSIDAD DE ARTEMISA

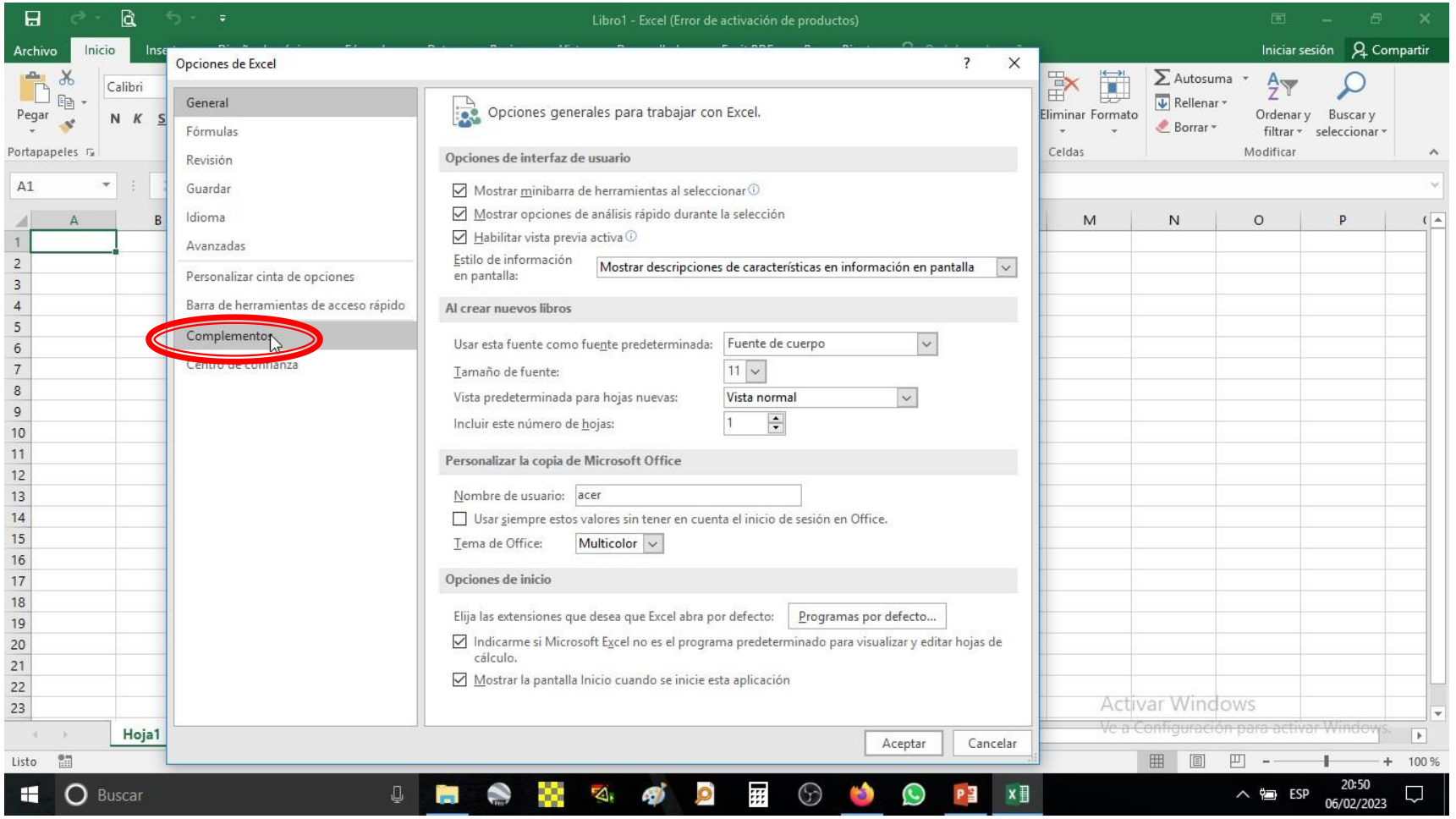
(MRLG) en la aplicación Excel



UNIVERSIDAD DE ARTEMISA

(MRLG) en la aplicación Excel

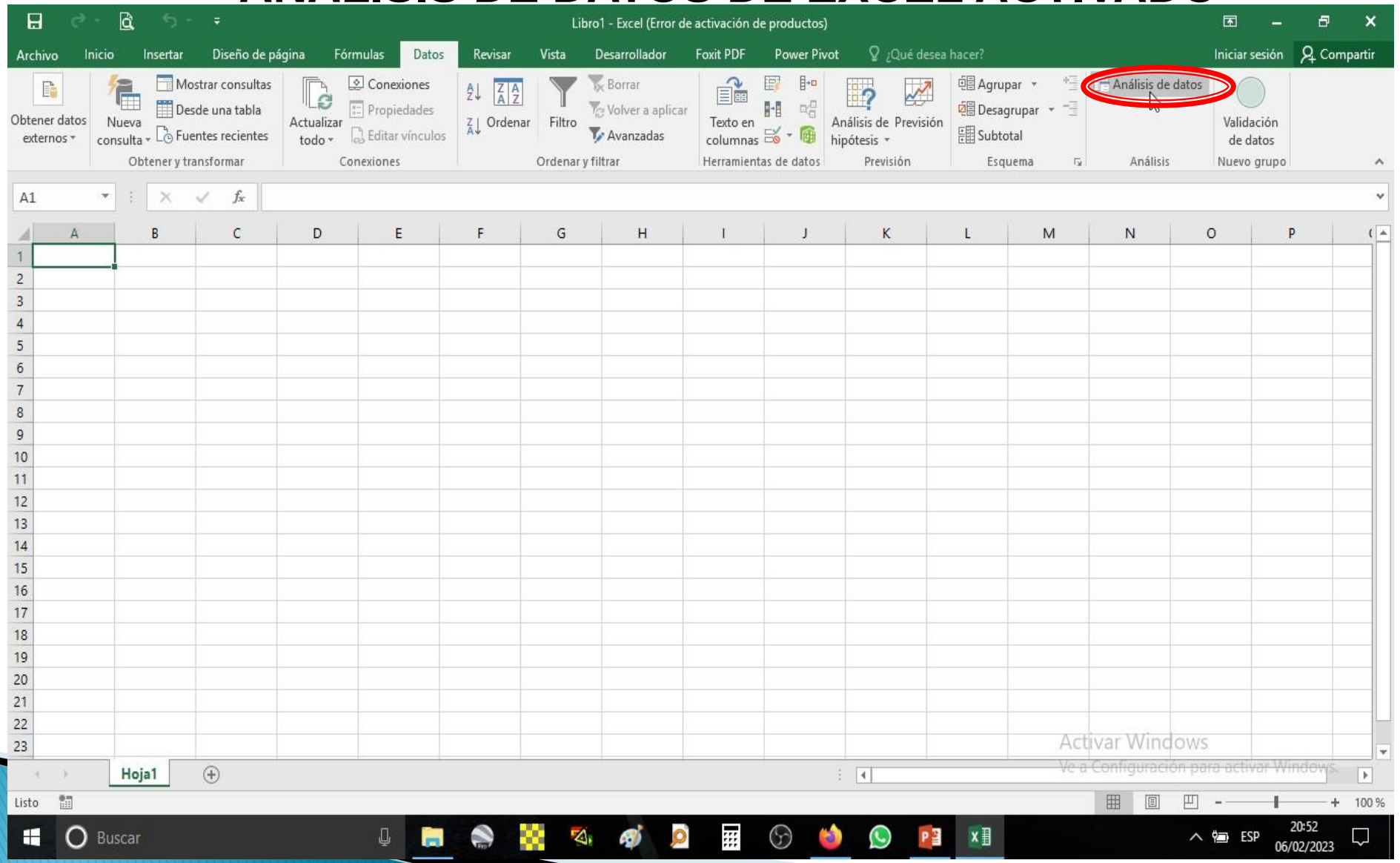
ACTIVAR OPCIÓN DE COMPLEMENTO DE EXCEL



UNIVERSIDAD DE ARTEMISA

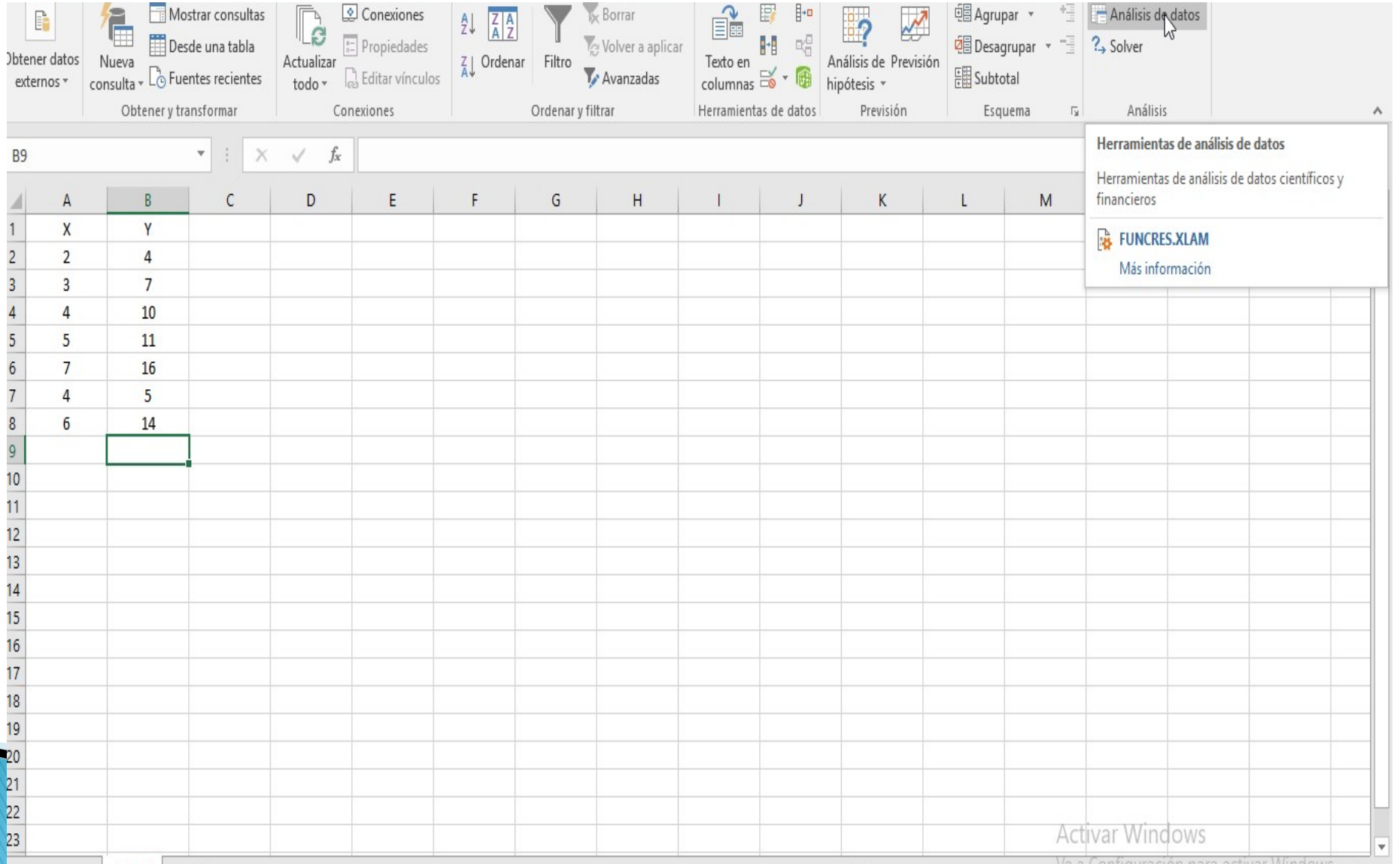
(MRLG) en la aplicación Excel

ANÁLISIS DE DATOS DE EXCEL ACTIVADO



(MRLG) en la aplicación Excel

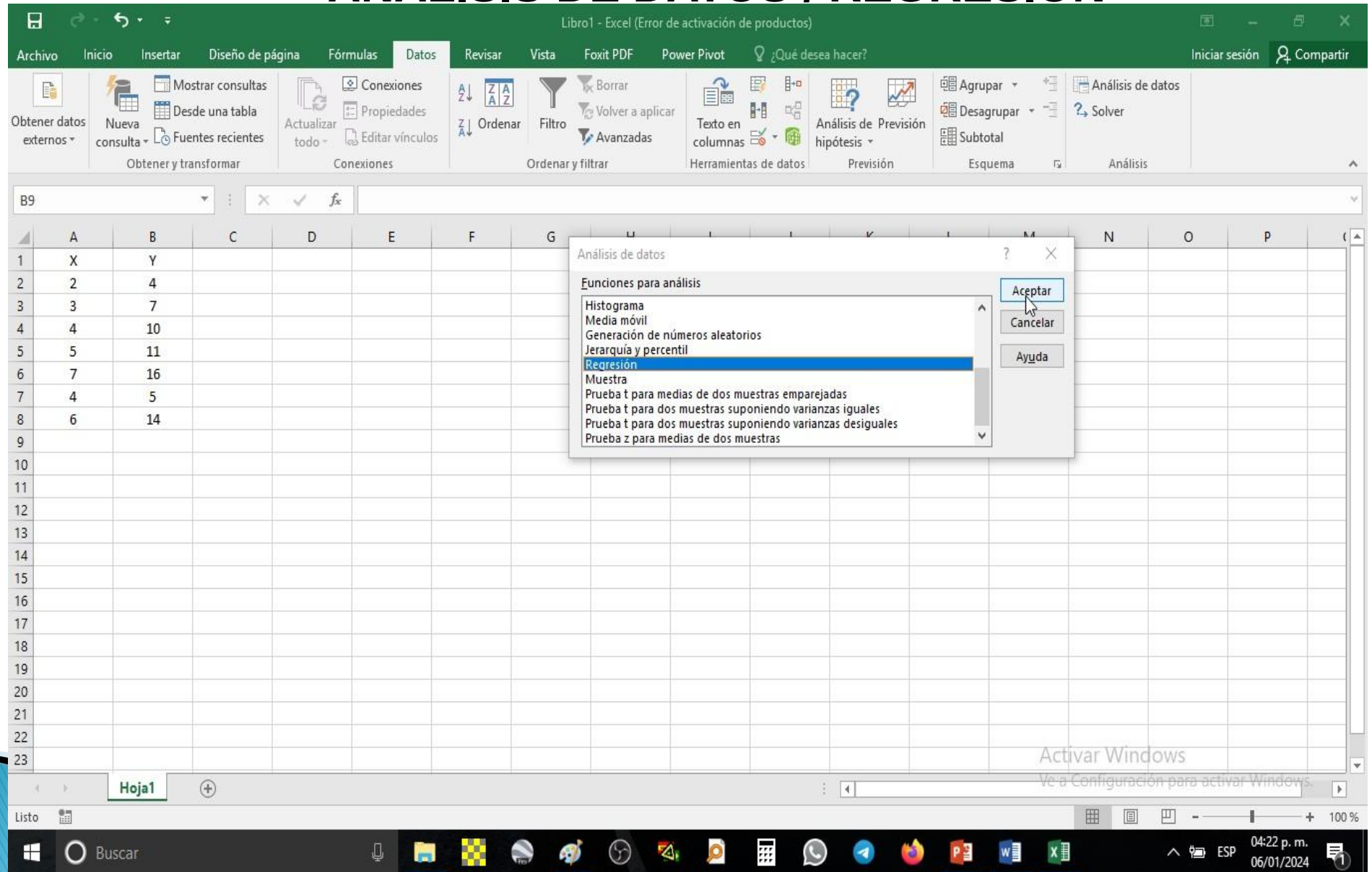
IMTRODUCIR DATOS EN EXCEL



UNIVERSIDAD DE ARTEMISA

(MRLG) en la aplicación Excel

ANÁLISIS DE DATOS / REGRESIÓN

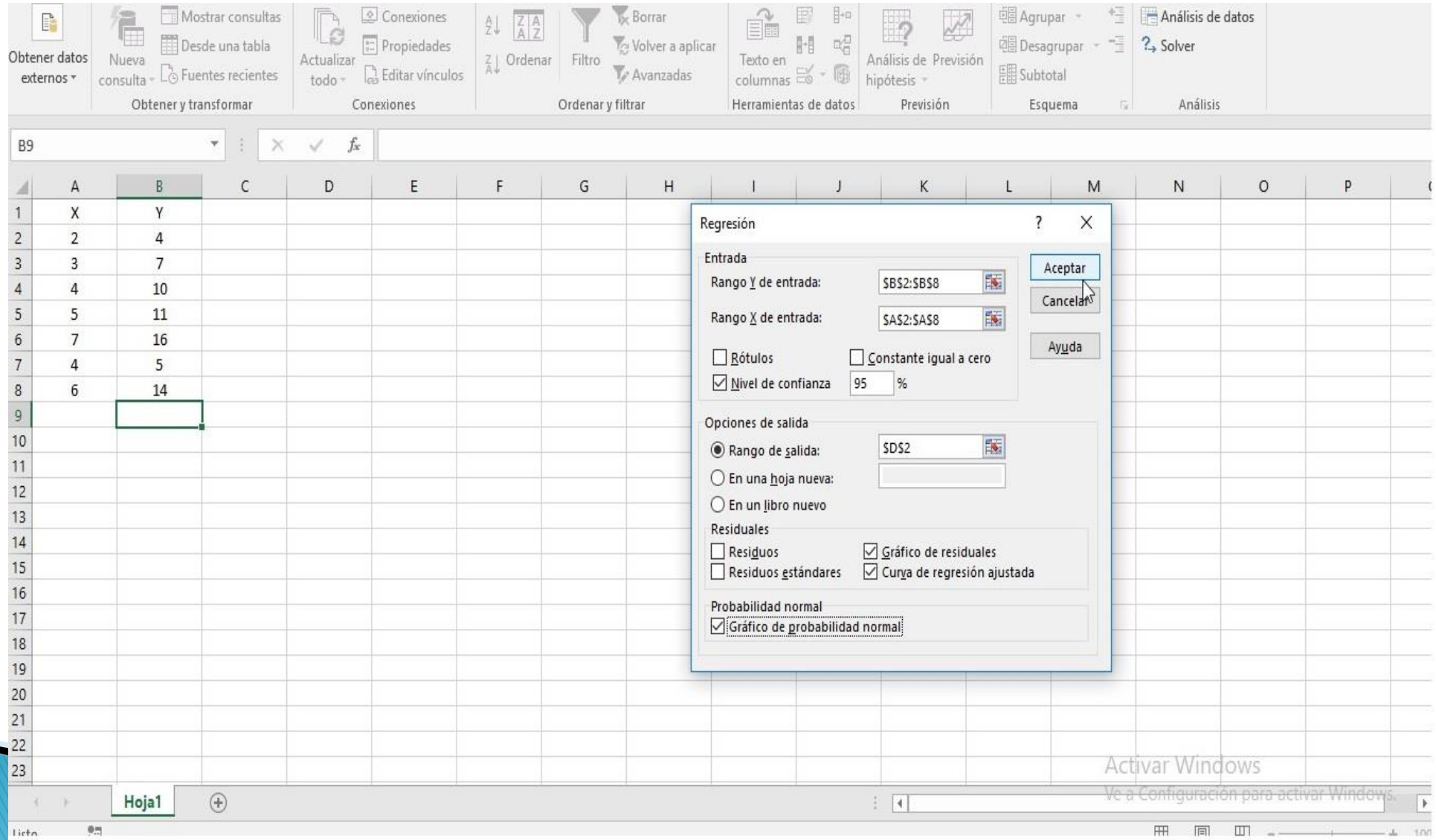


	A	B
1	X	Y
2	2	4
3	3	7
4	4	10
5	5	11
6	7	16
7	4	5
8	6	14

UNIVERSIDAD DE ARTEMISA

(MRLG) en la aplicación Excel

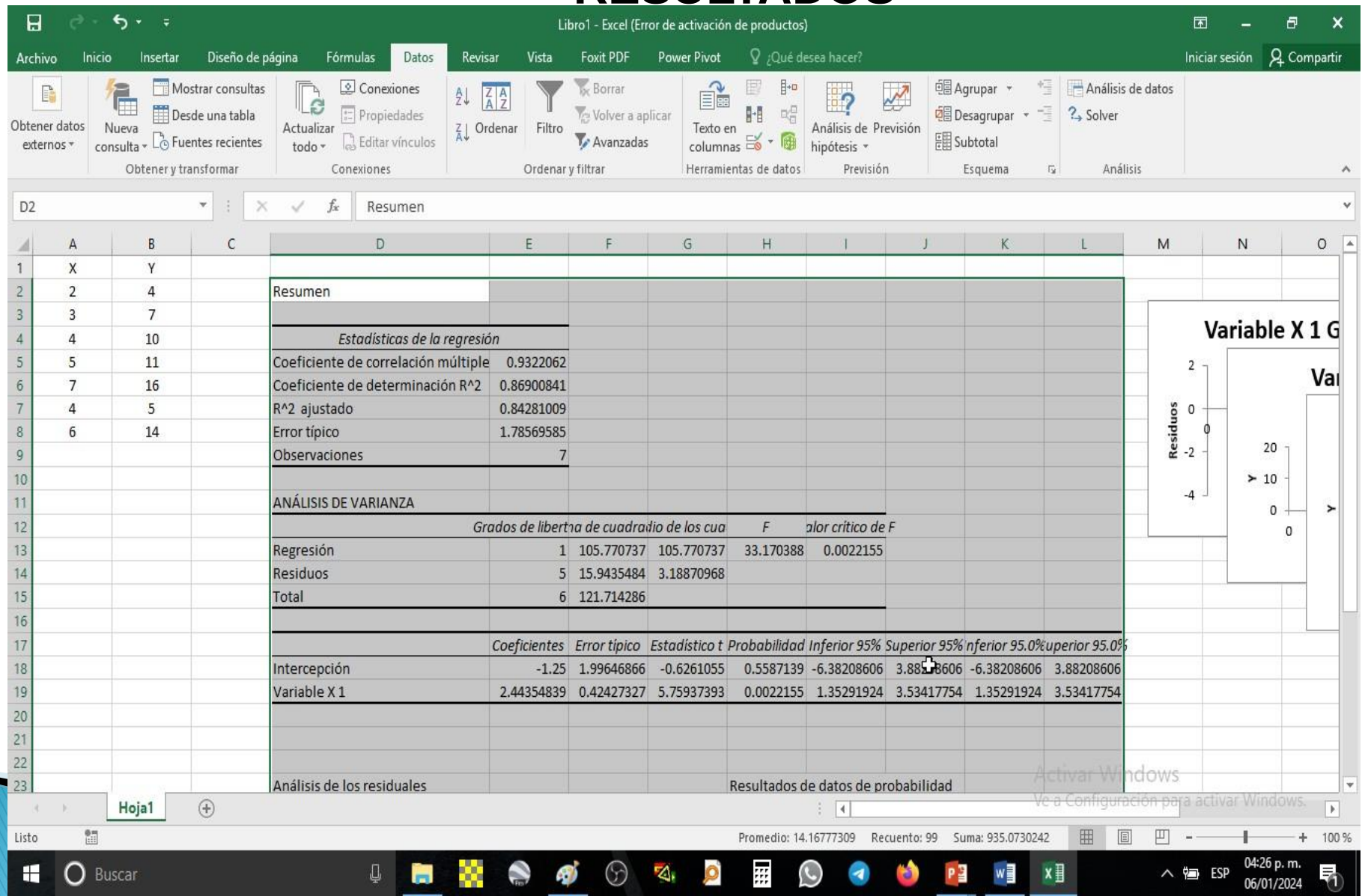
DEFINIR RANGO DE LAS VARIABLES



	A	B
1	X	Y
2	2	4
3	3	7
4	4	10
5	5	11
6	7	16
7	4	5
8	6	14

(MRLG) en la aplicación Excel

RESULTADOS



Libro1 - Excel (Error de activación de productos)

	A	B	C	D	E	F	G	H	I	J	K	L
1	X	Y										
2	2	4		Resumen								
3	3	7										
4	4	10		*Estadísticas de la regresión*								
5	5	11		Coeficiente de correlación múltiple	0.9322062							
6	7	16		Coeficiente de determinación R^2	0.86900841							
7	4	5		R^2 ajustado	0.84281009							
8	6	14		Error típico	1.78569585							
9				Observaciones	7							
10												
11				ANÁLISIS DE VARIANZA								
12					*Grados de liberta*	*a de cuadrad*	*io de los cua*	*F*	*alor crítico de F*			
13				Regresión	1	105.770737	105.770737	33.170388	0.0022155			
14				Residuos	5	15.9435484	3.18870968					
15				Total	6	121.714286						
16												
17					*Coeficientes*	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*nferior 95.0%*	*uperior 95.0%*
18				Intercepción	-1.25	1.99646866	-0.6261055	0.5587139	-6.38208606	3.88208606	-6.38208606	3.88208606
19				Variable X 1	2.44354839	0.42427327	5.75937393	0.0022155	1.35291924	3.53417754	1.35291924	3.53417754
20												
21												
22												
23				Análisis de los residuales				Resultados de datos de probabilidad				

Activar Windows
Ve a Configuración para activar Windows.

Hoja1

Listo

Promedio: 14.16777309 Recuento: 99 Suma: 935.0730242 100 %

Buscar

ESP 04:26 p. m. 06/01/2024

UNIVERSIDAD DE ARTEMISA

(MRLG) en la aplicación Excel

RESULTADOS

Estadísticos de la regresión	
Coeficiente de correlación múltiple	0.9322062
Coeficiente de determinación R^2	0.86900841
R^2 ajustado	0.84281009
Error típico	1.78569585
Observaciones	7

ANÁLISIS DE VARIANZA

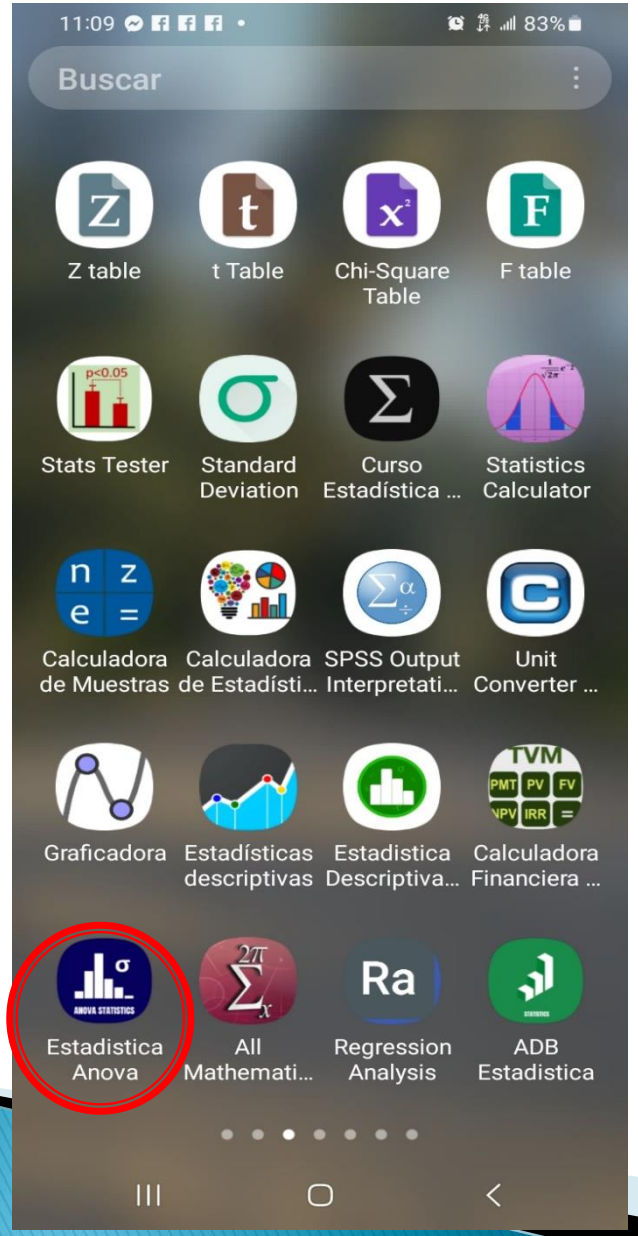
	Grados de liberta	*a de cuadrad*	*io de los cua*	*F*	*alor crítico de*
Regresión	1	105.770737	105.770737	33.170388	0.0022155
Residuos	5	15.9435484	3.18870968		
Total	6	121.714286			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*nferior 95.0%*	*uperior 95.0%*
Intercepción	-1.25	1.99646866	-0.6261055	0.5587139	-6.38208606	3.88208606	-6.38208606	3.88208606
Variable X 1	2.44354839	0.42427327	5.75937393	0.0022155	1.35291924	3.53417754	1.35291924	3.53417754

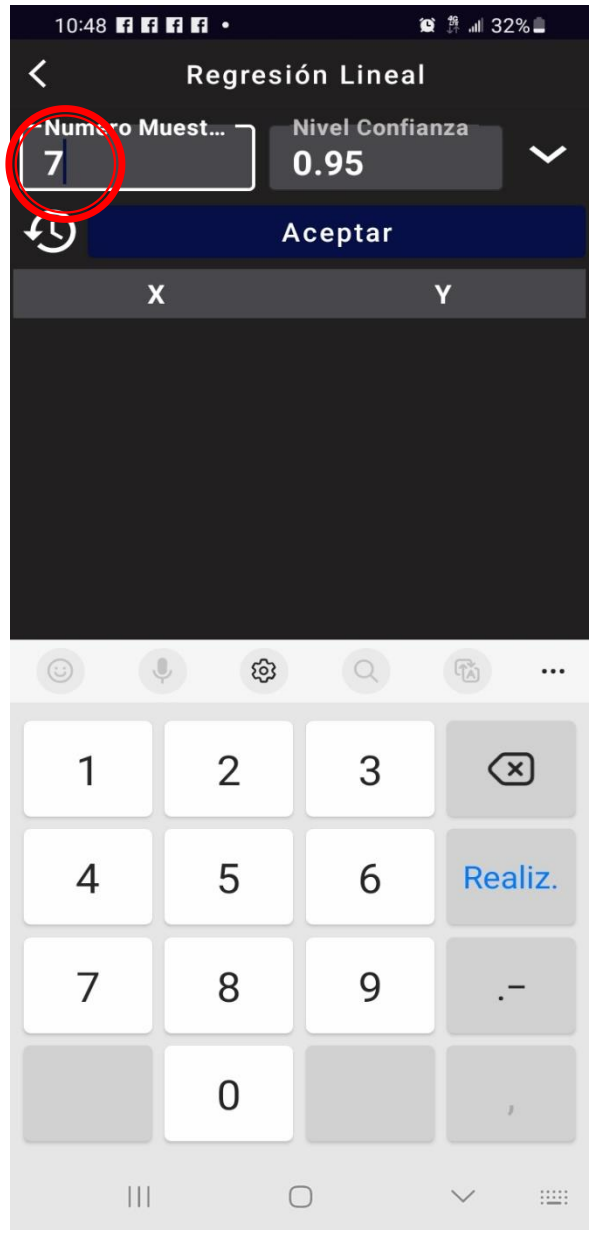


4.8. Análisis de Regresión con Apk “Estadística Anova”. Interpretación de los resultados.

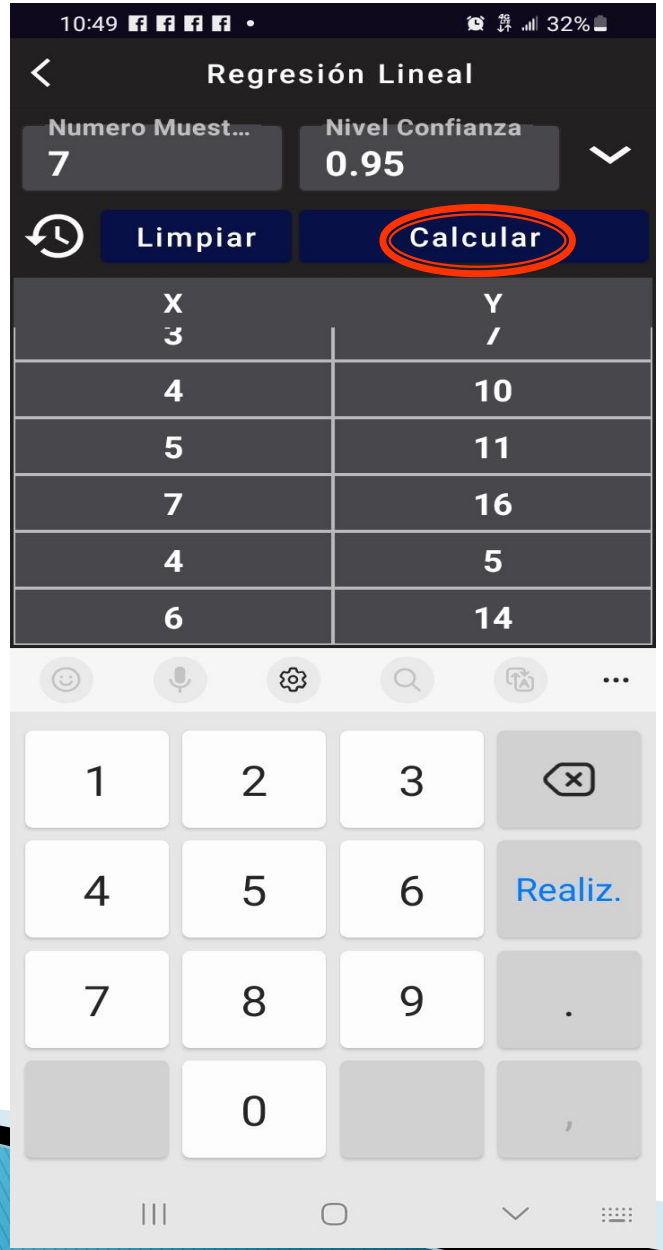
(MRLS) en la Apk Estadística Anova

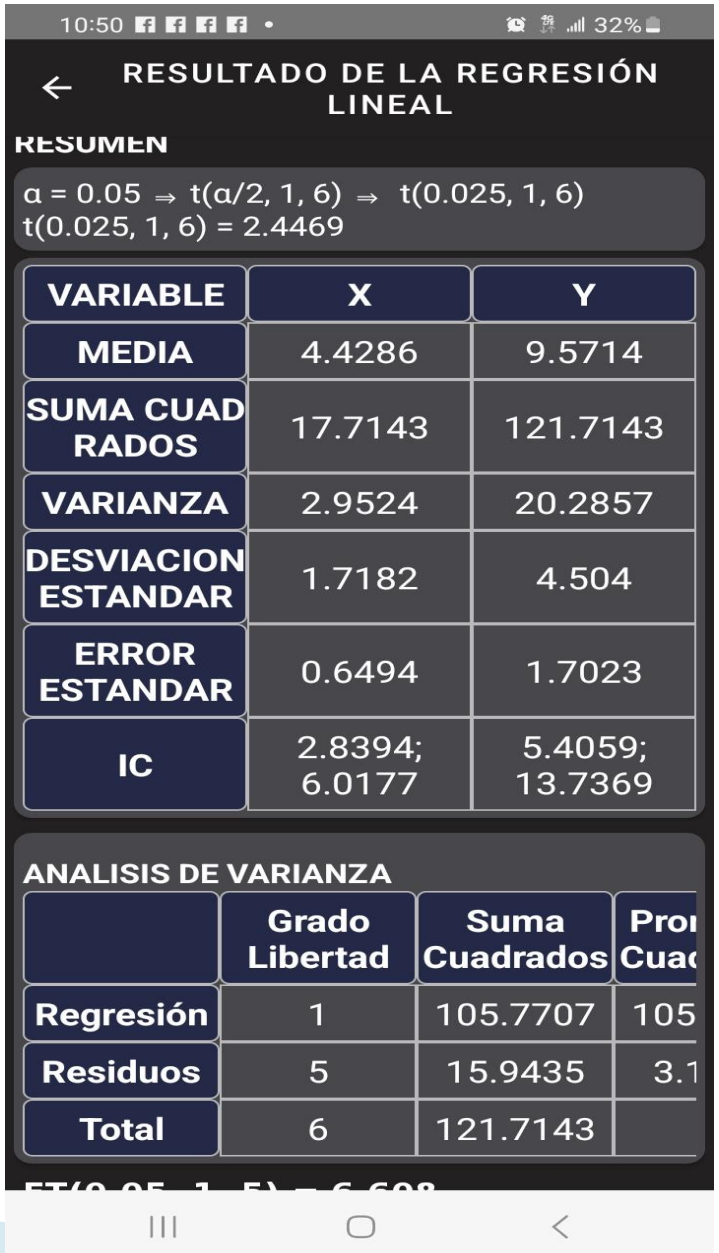






(MRLS) en la Apk Estadística Anova





UNIVERSIDAD DE ARTEMISA

(MRLS) en la Apk Estadística Anova

Tabla de análisis de varianza

	Grado Libertad	Suma Cuadrados	Promedio Cuadrados	F	p-value
Regresión	1	105.7707	105.7707	33.1704	0.0029
Residuos	5	15.9435	3.1887		
Total	6	121.7143			

Coeficiente de determinación

$R^2 = 0.869$

Coeficiente de correlación de Pearson

$r = 0.9322$

UNIVERSIDAD DE ARTEMISA

(MRLS) en la Apk Estadística Anova

Resultados del modelo de regresión

RESULTADO DE LA REGRESIÓN LINEAL

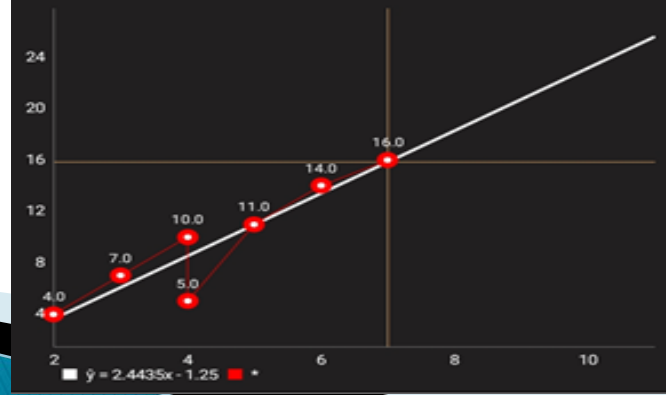
RESULTADOS DEL MODELO DE LA REGRESIÓN LINEAL

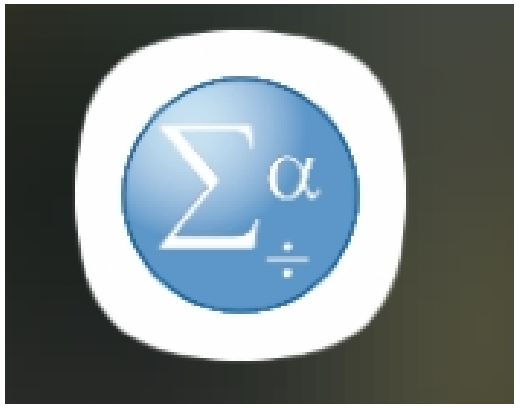
t(α/2, 1, 5) → t(0.025, 1, 5)
t(0.025, 1, 5) = 2.5706

VAR	COEF	ERROR ESTANDAR	INFERIOR 5.0%	SUPERIOR 5.0%
INTERCEPCIÓN (b)	-1.25	1.9965	-6.3821	3.8821
VAR X1 (m)	2.4435	0.4243	1.3529	3.5342

$$\hat{y} = 2.4435x - 1.25$$

R = 0.9322 ➡ R^2 = 0.869
R: 0.9322: Existe una alta correlación entre las dos variables.
R^2%: 86.9008%: El modelo lineal es adecuado para explicar la variación de Y explicado por X.



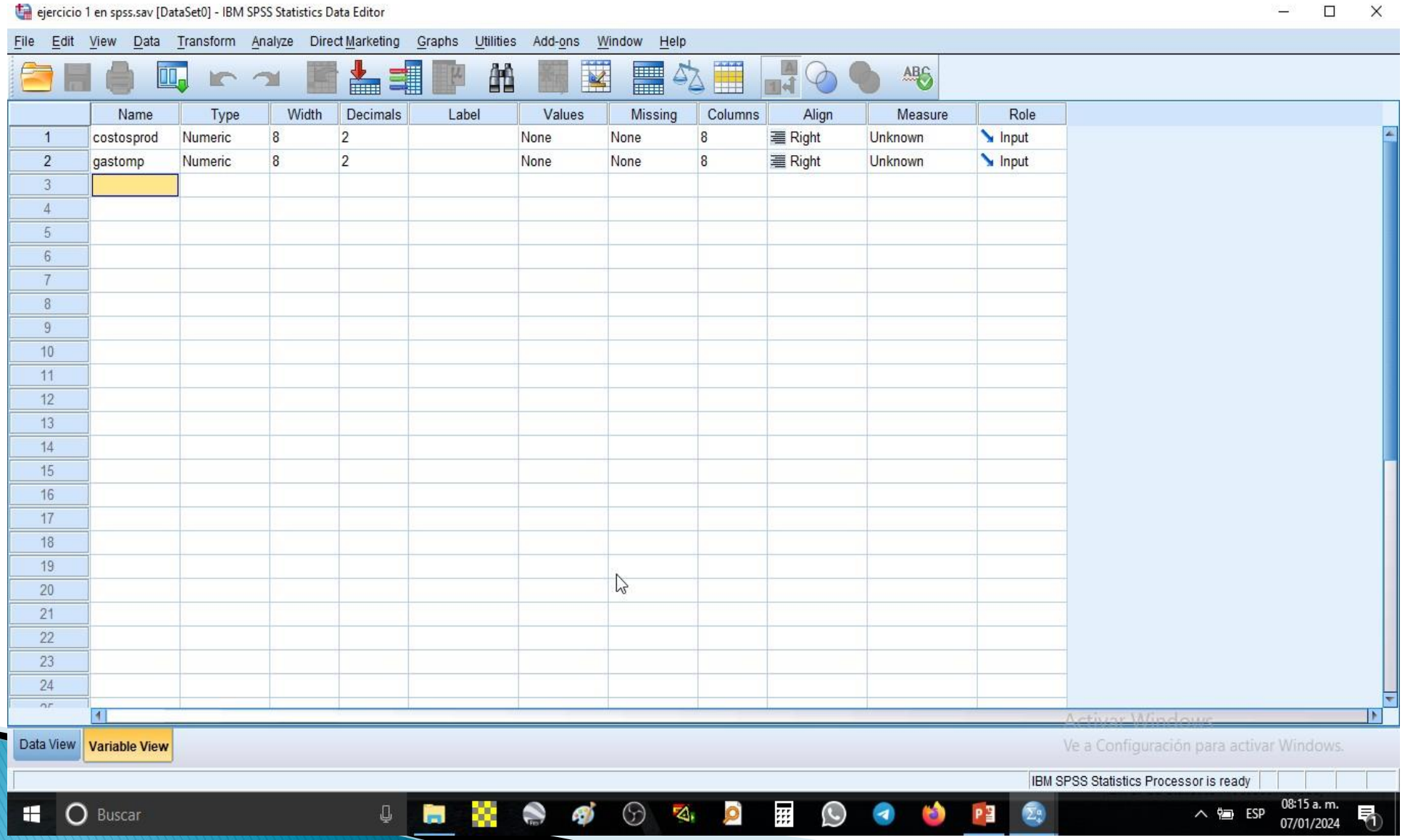


4.9. Análisis de Regresión con el paquete estadístico SPSS.

UNIVERSIDAD DE ARTEMISA

(MRLS) en software SPSS

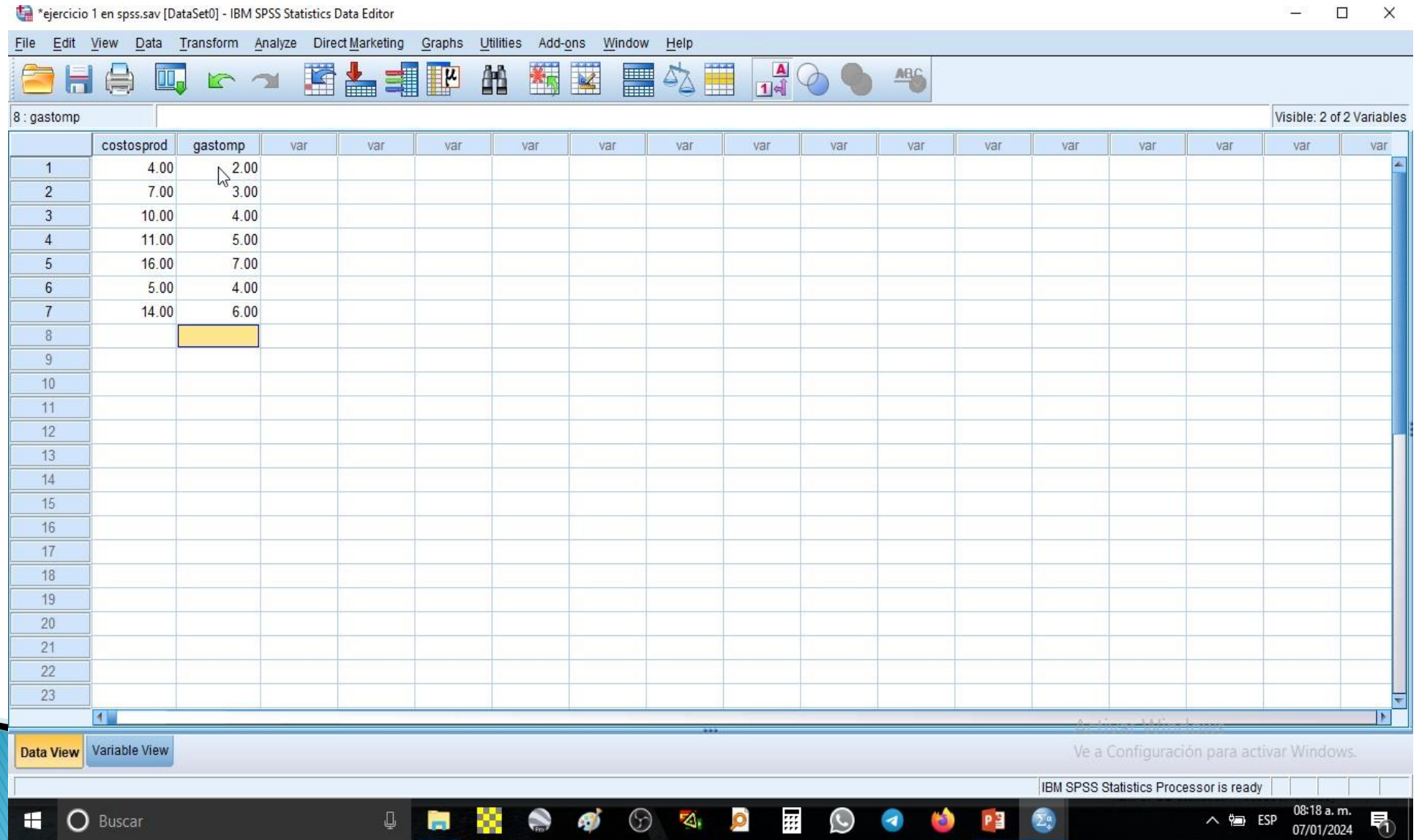
Declarar las variables



UNIVERSIDAD DE ARTEMISA

(MRLS) en software SPSS

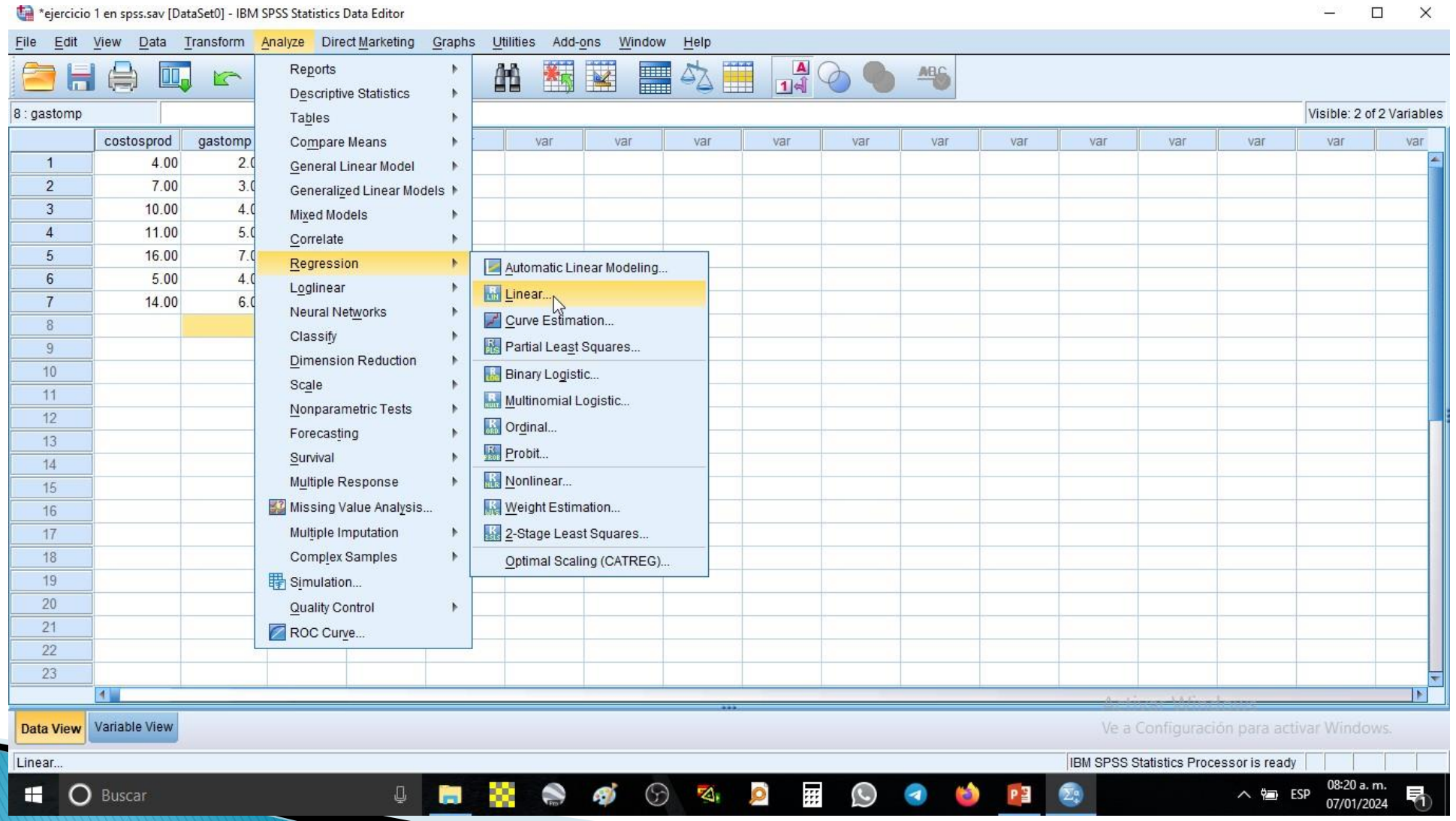
Introducir los datos



UNIVERSIDAD DE ARTEMISA

(MRLS) en software SPSS

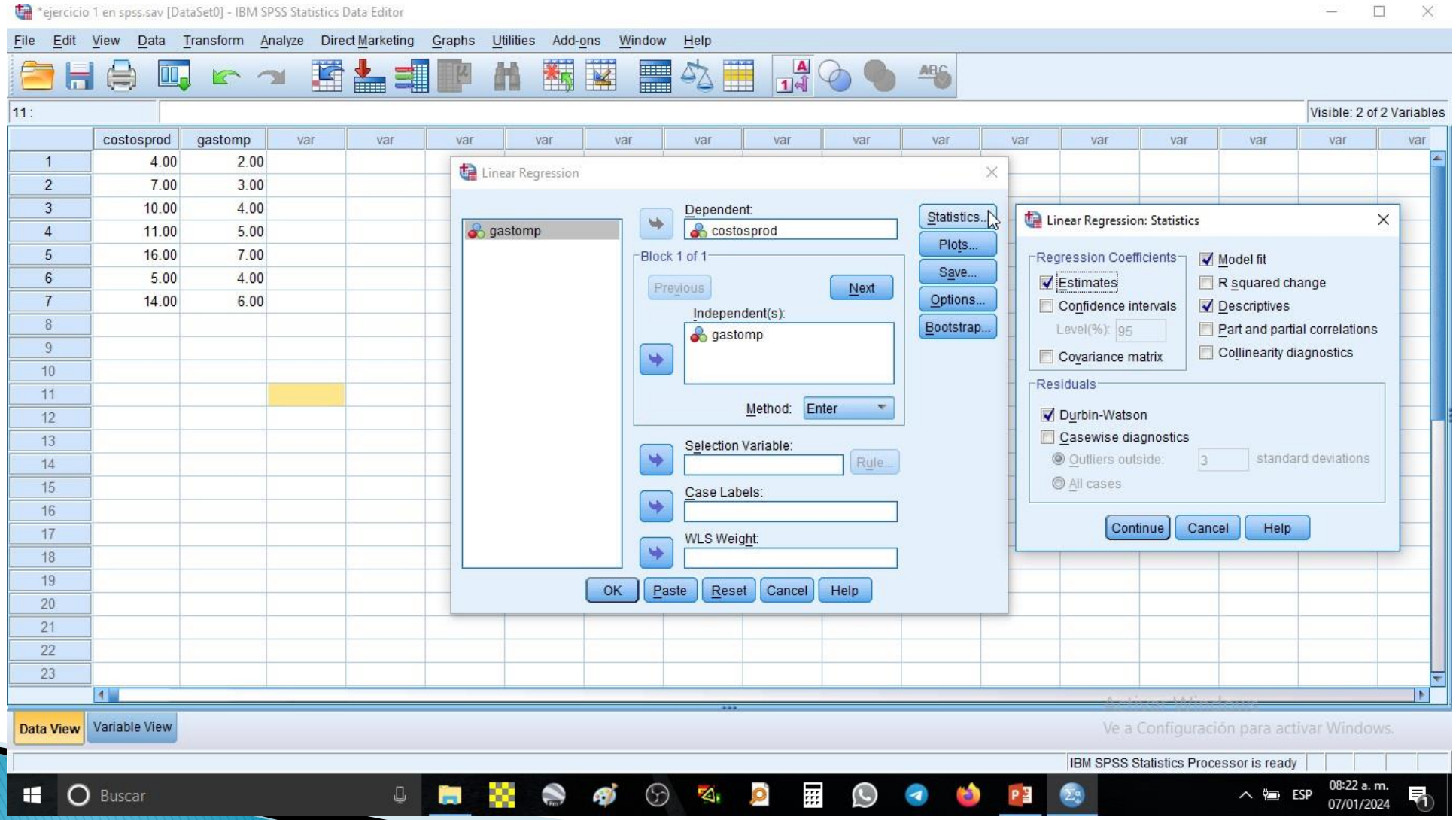
Análisis de Regresión



UNIVERSIDAD DE ARTEMISA

(MRLS) en software SPSS

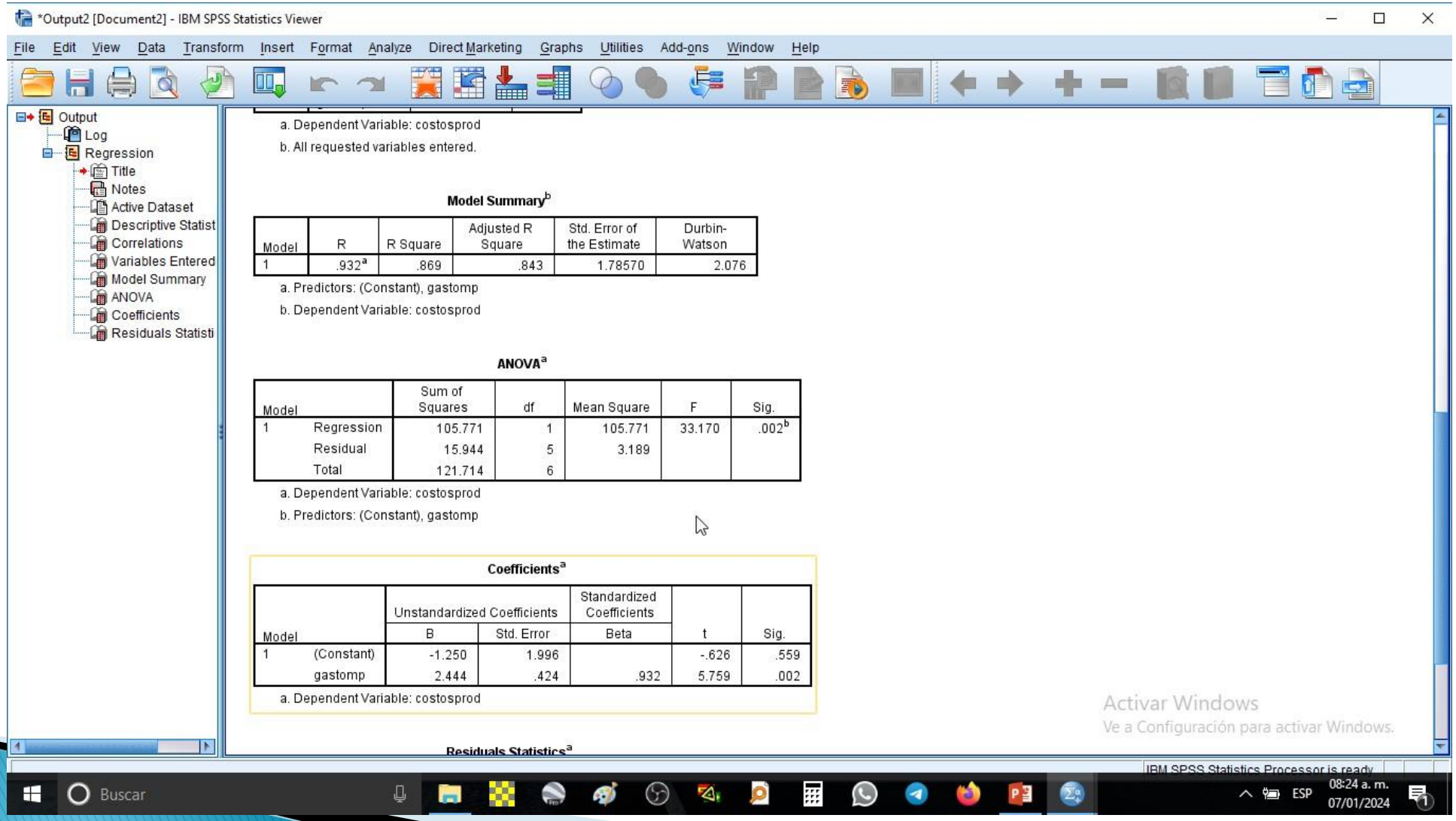
Definir las variables dependiente e independiente



UNIVERSIDAD DE ARTEMISA

(MRLS) en software SPSS

Resultados



UNIVERSIDAD DE ARTEMISA

(MRLS) en software SPSS

Resultados

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.250	1.996		-.626	.559
	gastomp	2.444	.424	.932	5.759	.002

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	105.771	1	105.771	33.170	.002[b]
	Residual	15.944	5	3.189		
	Total	121.714	6			

a

Model Summary[b]

b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.932[a]	.869	.843	1.78570	2.076

a. Predictors: (Constant), gastomp

b. Dependent Variable: costosprod

4.10

Estimación de los intervalos de confianza para el parámetro (βj) y estimación de Y_k.

UNIVERSIDAD DE ARTEMISA

Intervalo de confianza para los parámetros β_j

El intervalo de confianza para los parámetros β_j del MRLS.

$$\text{Luego:} \quad I(\beta_j) = b_j \pm d$$

$$I(\beta_j) = b_j - d \leq \beta_j \leq b_j + d$$

UNIVERSIDAD DE ARTEMISA

Intervalo de confianza para los parámetros β_j

$$I(\beta_j) = b_j \pm d$$

Donde “d ” es el error máximo permisible en la estimación del β_j

“d” se obtiene de manera similar al error máximo permisible en la estimación

$$d = t_{(1-\alpha/2);(n-p)}\ Es\ (b_j)$$

$t_{(0.975);\ (n-p)}$ Donde n: cantidad de observaciones (7),
p: cantidad de variables (2)

UNIVERSIDAD DE ARTEMISA

Intervalo de confianza para los parámetros β_j

$$I(\beta_j) = b_j \pm d$$

$$d = t_{(1-\alpha/2);(n-p)}\ Es\ (b_j)$$

$Es_{(bj)}$: el error estándar se puede obtener de los resultados de las aplicaciones informáticas.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.250	1.996		-.626	.559
	gastomp	2.444	.424	.932	5.759	.002

a. Dependent Variable: costosprod

UNIVERSIDAD DE ARTEMISA

Intervalo de confianza para los parámetros β_j

$Es_{(bj)}$: el error estándar también se puede obtener de la matriz varianza covarianza.

Varianza del estimador b_0

$$\hat{V}(b) = \begin{bmatrix} \hat{V}(b_0) & Cov(b_0\ b_1) \\ Cov(b_0\ b_1) & \hat{V}(b_1) \end{bmatrix}$$

Varianza del estimador b_1

Hallando la raíz cuadrada a la varianza de los estimadores

UNIVERSIDAD DE ARTEMISA

Intervalo de confianza para los parámetros β_j

$$b_j - d \leq \beta_j \leq b_j + d$$

Interpretación del intervalo $I(\beta_1)$.

En la medida que aumenta en 1 unidad la variable independiente la variable dependiente variará entre esos valores

UNIVERSIDAD DE ARTEMISA

Intervalo de confianza para la estimación

El cálculo de los intervalos de confianza se de la estimación realiza con las siguientes expresiones matemáticas:

$$I(\hat{Y}\, k_i) = \hat{Y}\, k \pm d$$

$$d = t_{(1-\alpha/2);(n-p)} \sqrt{S_e^2}$$

$Es_{(bj)}$: el error estándar del modelo se puede obtener de la Tabla de Anova, hallando la raíz cuadrada a la varianza Se^2

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	105.771	1	105.771	33.170	.002[b]
	Residual	15.944	5	3.189		
	Total	121.714	6			

CONCLUSIONES

UNIVERSIDAD DE ARTEMISA

Conclusiones

- El Análisis de Regresión se puede realizar a través de la vía algebraica, vía matricial, aplicaciones y software informáticas y hasta de Apk para teléfonos móviles.

Estudio Independiente:

Clase Práctica.

Con la información que se dispone sobre el ejercicio anterior responda:

Responda:

a) Formule la ecuación de regresión del Modelo e interprete el coeficiente de Regresión b_1.

b) Verificar la Bondad de ajuste del Modelo con una confianza del 95 %.

c) Interpretar el Coeficiente de Determinación (R^2) en el contexto.

c) Calcular el intervalo de confianza para los parámetro β_1 con una confianza del 95 %. Interpretar su significado en el contexto.

d) Estime con una confianza del 95 % el costos de la producción cuando los gastos de materia prima son de 3 MP. Interpretar su significado

MUCHAS GRACIAS

Prof Aux. Rodolfo Pérez León
email:
rperezleon@uart.edu.cu
rplrpl55@gmail.com

UNIVERSIDAD DE ARTEMISA

ESTADÍSTICA II

MSc. Rodolfo Pérez Léon
email: rperezleon@uart.edu.cu
rplrpl55@gmail.com

CLASE PRÁCTICA 1.

TEMA 1. El Modelo de Regresión Lineal General (MRLG).

EJERCICIO 1.

Sea Y: los gastos de producción de una fábrica (en Miles CUP)

X: los gastos de materia prima de la fábrica (en Miles CUP).

Y	**4.0**	**7.0**	**10.0**	**11.0**	**16.0**	**5.0**	**14.0**
X	2.0	3.0	4.0	5.0	7.0	4.0	6.0

Resultados obtenidos del software SPSS

Parámetro	Estimador	Error Estand.	Statistic	p-value
CONSTANT	-1. 25	1.9964	-0.6261	0.5587
X1	2.4435	0.4242	5.7594	0.0222

Tabla Anova (Análisis de varianza)

Fuente	S.Cuadrados	GL	C.Medio	F-Ratio	p-value
Model	105.7707	1	5.7707	33.1703	0.0214
Residual	15.9435	5	3.1887		
Total (Corr.)	121.7142	6			

R^2 = 86.90% R^2(ajustado) = 84.28% Error típico= 1.7857

Durbin-Watson statistic = 1.82225 P valor Prueba K-S = 0.0004

Prueba de Bartlett: 1.58149 P. Valor de la Prueba de Bartlett = 0.000568

Responda:

a) Formule la ecuación de regresión del Modelo e interprete el coeficiente de Regresión b_1.

b) Verificar la Bondad de ajuste del Modelo con una confianza del 95 %.

c) Interpretar el Coeficiente de Determinación (R^2) en el contexto.

c) Calcular el intervalo de confianza para los parámetro β_1 con una confianza del 95 %. Interpretar su significado en el contexto.

d) Estime con una confianza del 95 % el costos de la producción cuando los gastos de materia prima son de 3 MP. Interpretar su resultado

RESPUESTAS

a) Formule la ecuación de regresión del Modelo e interprete el coeficiente de Regresión b_1.

Para formular la ecuación de regresión nos auxiliaremos de la salida del paquete estadístico; debemos tener presente que cada paquete estadístico aporta sus salidas especificas (información, respuesta) pero de forma general siempre contienen la información fundamental para el análisis.

Para formular la ecuación de regresión tomemos de la tabla de Parámetros los valores de los estimadores b_0 y b_1.

Parámetro	Estimador	Error Estand.	Statistic	p-value
CONSTANT	-1. 25	1.9964	-0.6261	0.5587
X1	2.4435	0.4242	5.7594	0.0222

b_0

b_1

$$\hat{Y} = \beta_0 + \beta_1 X_1 + E_i$$

$$\hat{Y} = b_0 + b_1 x_1 + e_i$$

$$\hat{Y} = -1.25 + 2.4435\ X_1$$

R/ Ecuación de regresión

R/ El coeficiente de regresión b_1 se interpreta así:

En la medida que aumente en 1.0 MCUP los gastos de materia prima de la fábrica, aumentará en 2.443 MCUP los costos de producción de la fábrica.

b) Verificar la bondad de ajuste del Modelo con una confianza del 95 %.

Para verificar la bondad de ajuste del modelo se realiza la prueba F_{total} que es otra prueba de hipótesis y se sigue el mismo procedimiento.

Para realizar la prueba es necesario la información del Análisis de varianza.

Análisis de varianza

Fuente	S.Cuadrados	GL	C.Medio	F-Ratio	p-value
Model	105.7707	1	5.7707	33.1703	0.0214
Residual	15.9435	5	3.1887		
Total	121.7142	6			

1er Paso Planteamiento de la Hipótesis

H_0: $\beta_1 = 0$ (la ecuación no es de buen ajuste)
H_1: $\beta_1 \neq 0$ (la ecuación es de buen ajuste)

2do Paso Determinación del nivel de significación α

$1-\alpha = 0.95$ entonces $\alpha = 0,05$ (5%)

3er Paso Regla de decisión:

Si $P_{valor} < \alpha$ Se rechaza H_0

4to. Paso Toma de la Decisión

0,0214 < 0,05 se rechaza H_o concluyendo que:

R/ La ecuación es de buen ajuste con una confianza del 95 %.

c) Interpretar el Coeficiente de Determinación (R^2) en el contexto.

Entre la información que aporta el paquete estadístico esta que $R^2 = 86.90\%$ y se interpreta así:

R/ El 86.90% de la variación en los costos de producción de la fábrica quedan explicados por los gastos de matéria prima.

d) Calcular el intervalo de confianza para el parámetro β_1 con una confianza del 95 %. Interpretar su significado en el contexto.

$$I(\beta_j) = b_j \pm d$$

$$d = t_{(1-\alpha/2);(n-p)}\ Es\ (b_j)$$

El estimador b_1 y su error estándar Es (b_j) lo encontramos en la tabla de los Parámetros.

Parámetro	Estimador	Error Estand.	Statistic	p-value
CONSTANT	-1. 25	1.9964	-0.6261	0.5587
X1	2.4435	0.4242	5.7594	0.0222

Estimador b_1 Es (error estandar de b_1)

$$I(\beta_j) = b_j \pm d$$

$$d = t_{(1-\alpha/2);(n-p)}\ Es\ (b_j)$$

Para el percentil $t_{(1-\alpha/2);(n-p)}$ se tienen los valores de los grados de libertad (n-p) de la tabla Análisis de varianza que es de 5.

Para una confianza de 95% entonces $\alpha = 0.05$ y $(1-\alpha/2)= 0.975$ quedando el percentil

$$t_{(1-\alpha/2);(n-p)} = t_{(0.975);(5)} = 2.57 \text{ (Table t-Student)}$$

$$d= 2.57(0.4242) = 1.09$$

$$I(\beta_1) = b_1 \pm d$$

$$I(\beta_1) = 2.4435 \pm 1.09$$

$$I(\beta_1) = b_1 - d \leq \beta_1 \leq b_1 + d$$

$$2.4435 - 1.09 \leq \beta_1 \leq 2.4435 + 1.09$$

$$1.3535 \leq \beta_1 \leq 3.5335$$

R/ El parámetro β_1 se encontrará entre 1.3535 y 3.5335 con una confianza del 95%

e) Estime con una confianza del 95 % el costo de la producción cuando los gastos de materia prima son de 3.0 MP. Interpretar el resultado.

$$I\left(\hat{Y}_{k_i}\right) = \hat{Y}_k \pm d$$

$$d = t_{(1-\alpha/2);(n-p)} \sqrt{S_e^2}$$

Comenzamos por evaluar la ecuación para X = 3.0

$$\hat{Y} = -1.25 + 2.4435\ x_1$$

$$\hat{Y} = -1.25 + 2.4435\ (3.0)$$

$$\hat{Y} = 6.08$$

Posteriormente se calcula $d = t_{(1-\alpha/2);(n-p)} \sqrt{S_e^2}$

El paquete estadístico ofreció el error típico de la estimación (S_e) = 1.7857 pero también se podia calcular con la raiz cuadrada de la varianza del error estandar.

$$\sqrt{S_e^2} = \sqrt{3.1887} = 1.7857$$

Entonces se calcula “d” el error máximo admisible

$$d = t_{(0.975);(5)} S_e$$

$$d = 2.57\ (1.7857)$$

$$d = 4.59$$

Finalmente se puede calcular el intervalo

$$I(\hat{Y}\,k_i) = \hat{Y}\,k \pm d$$

$$I(\hat{Y}\,k_i) = 6.08 \pm 4.59$$

$$(\hat{Y}\,k_i) - d \leq (\hat{Y}\,k_i) \leq (\hat{Y}\,k_i) + d$$

$$6.08 - 4.59 \leq (\hat{Y}\,k_i) \leq 6.08 + 4.59$$

$$1.49 \leq (\hat{Y}\,k_i) \leq 10.67$$

R/ Se estima que cuando el gasto de material prima sea de 3.0 MCUP el costo de producción de la fábrica esté entre 1.49 y 10.67 MCUP con una confianza de 95%.

Estudio Independiente.

Se incrementaron las observaciones en el ejercicio anterior, realice el nuevo análisis de regresión aplicando un paquete estadístico.

Y	4.0	7.0	10.0	11.0	16.0	5.0	14.0	4.0	7.0	10.0	11.0	16.0	5.0	14.0	10.0	5.0
X	2.0	3.0	4.0	5.0	7.0	4.0	6.0	2.0	3.0	4.0	5.0	7.0	4.0	6.0	5.0	3.0

MUCHAS GRACIAS

Prof Aux. Rodolfo Pérez León
email: rperezleon@uart.edu.cu
rplrpl55@gmail.com



UNIVERSIDAD DE ARTEMISA

ESTADÍSTICA II

MSc. Rodolfo Pérez Léon
Email: rperezleon@uart.edu.cu

Tema 5. SUPUESTOS DEL MRLG.

5.1.1. **Normalidad**. Verificación del supuestos.
5.1.2. **Homocedasticidad**. Verificación del supuesto. Consecuencias.
5.1.3. **No Autocorrelación**, sus efectos. Pruebas para verificar el supuesto.
5.1.4. **No Multicolinealidad**, sus efectos, métodos para detectarla.
5.2. Especificaciones de un Modelo

Objetivos: Que el alumno sea capaz de:

- Identificar los supuestos del Modelo de regresión Lineal General, sus causas y consecuencias.
- Verificar los supuestos básicos del Modelo, Normalidad, Homocedasticidad, No autocorrelación y No multicolinealidad mediante el análisis de los residuos obtenidos por las salidas de programas informáticos

Los Residuos

El análisis de los residuos $e_i = y - \hat{y}$ podría considerarse una de las partes más importantes, así como la que más informa, al estudiar las ecuaciones de regresión, es por ello que se hace necesario una vez que se ha estimado la ecuación de regresión, estudiar los residuos.

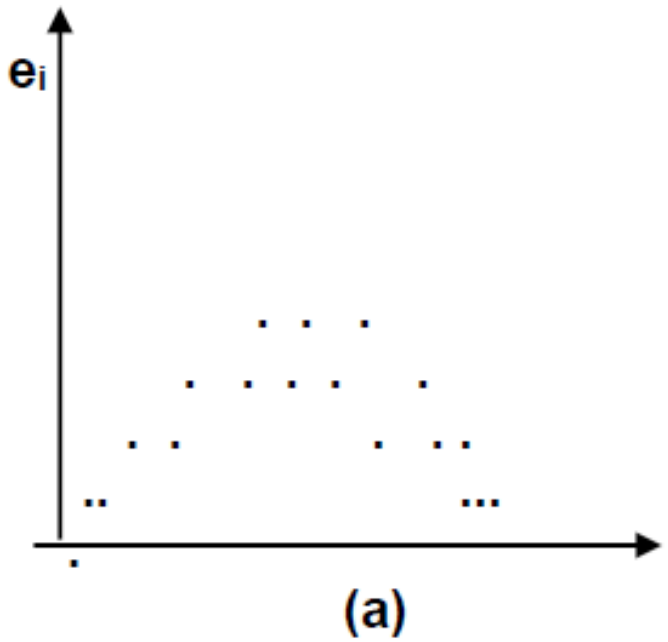
Este estudio puede indicar múltiples situaciones en los datos, como es:

- Si la relación funcional elegida no es la adecuada
- Si existen observaciones atípicas
- Si existen cambios estructurales
- Si se omitieron variables en el modelo etc
- Si se viola algunos de los supuestos del modelo
  - ✓Si los u_i no están distribuidas normalmente
  - ✓Si la varianza (σ^2) de u_i no es constante.
  - ✓Cómo que las u_i están correlacionadas en

Esto se podría ver Gráficamente:

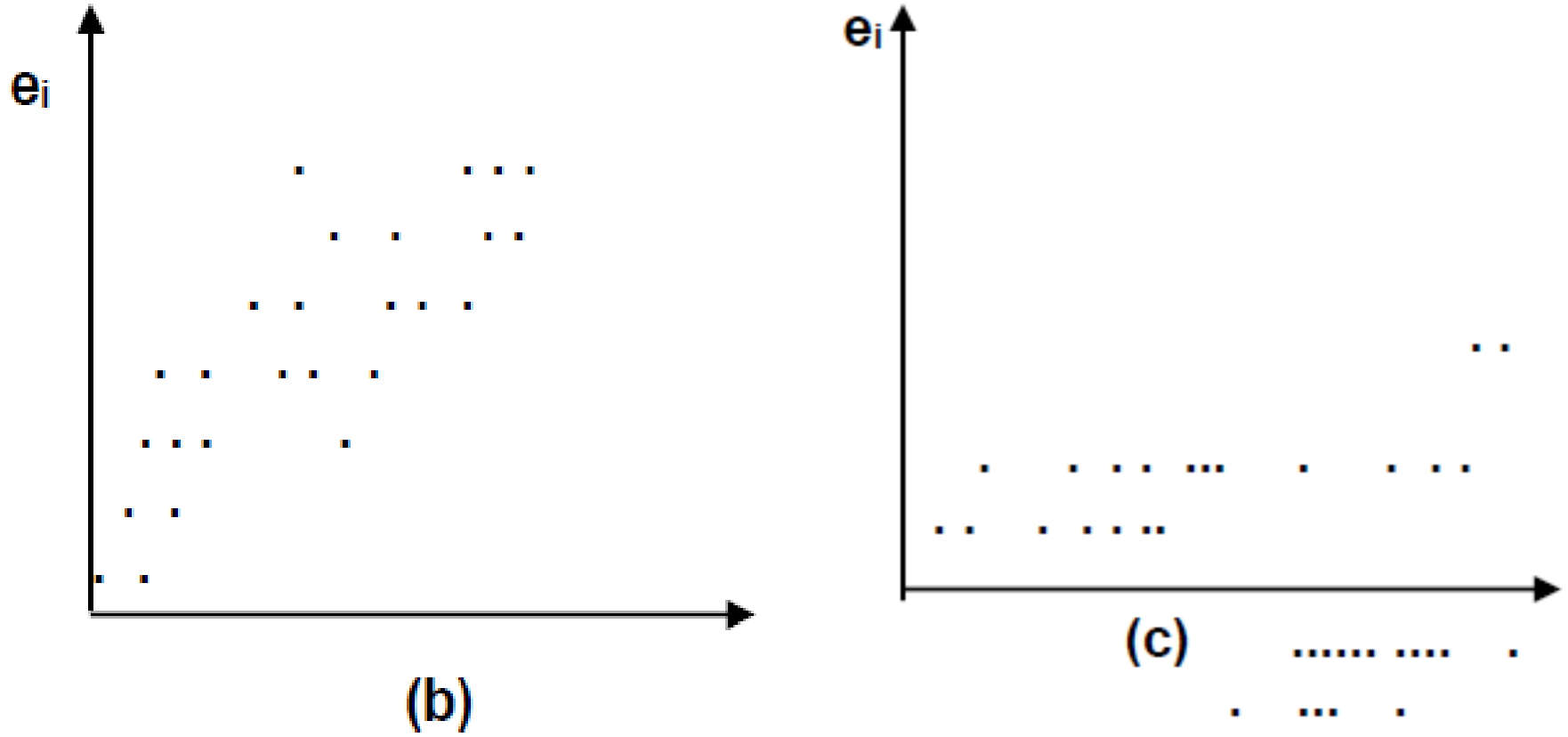
Gráficos de e_i contra x_i (donde lo que se considera como v.i. representa la abscisa en este caso "x" y e_i la ordenada)

$$e_i = y_i - \hat{y}_i \quad siendo \quad i = 1, 2, \cdots n \quad pares$$

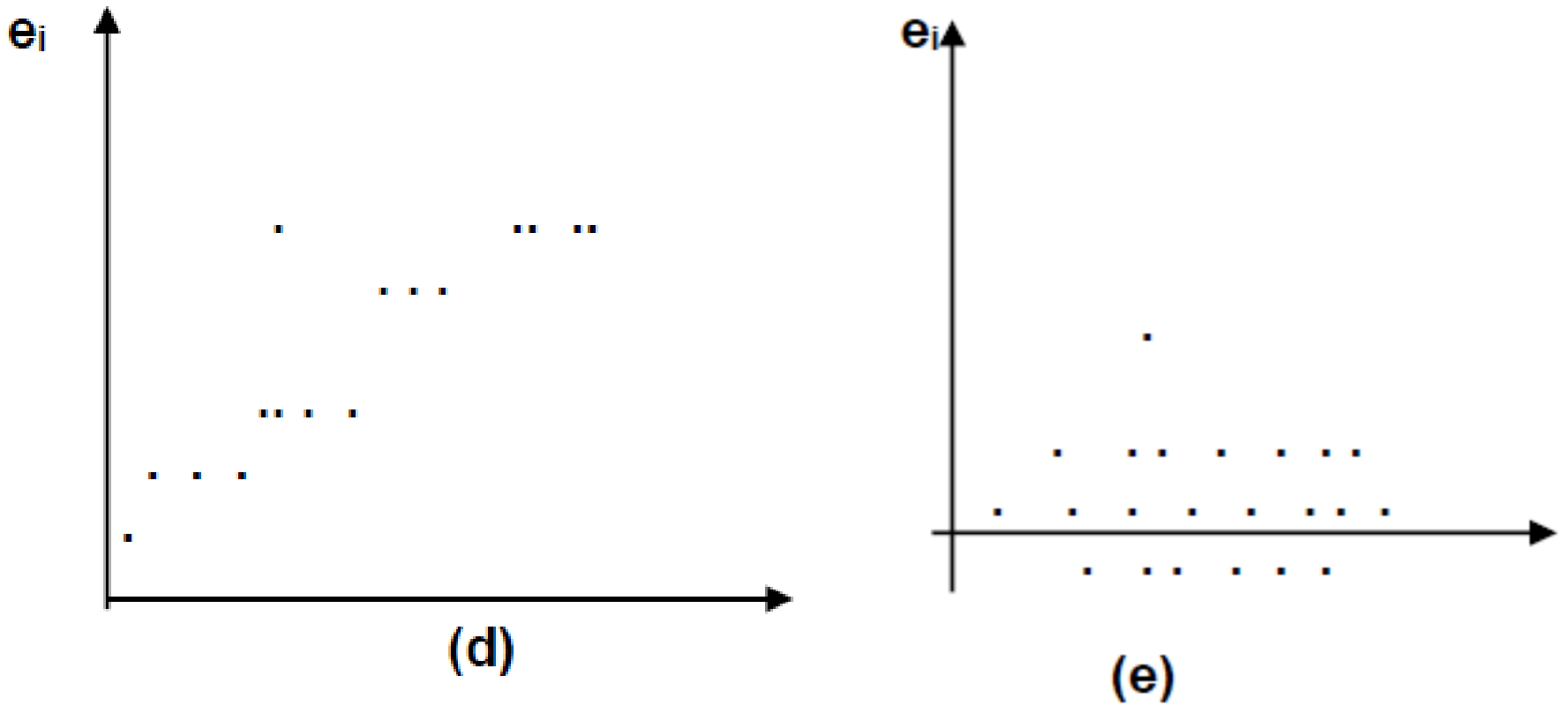


Si se analizan estos gráficos, se observa lo siguiente:

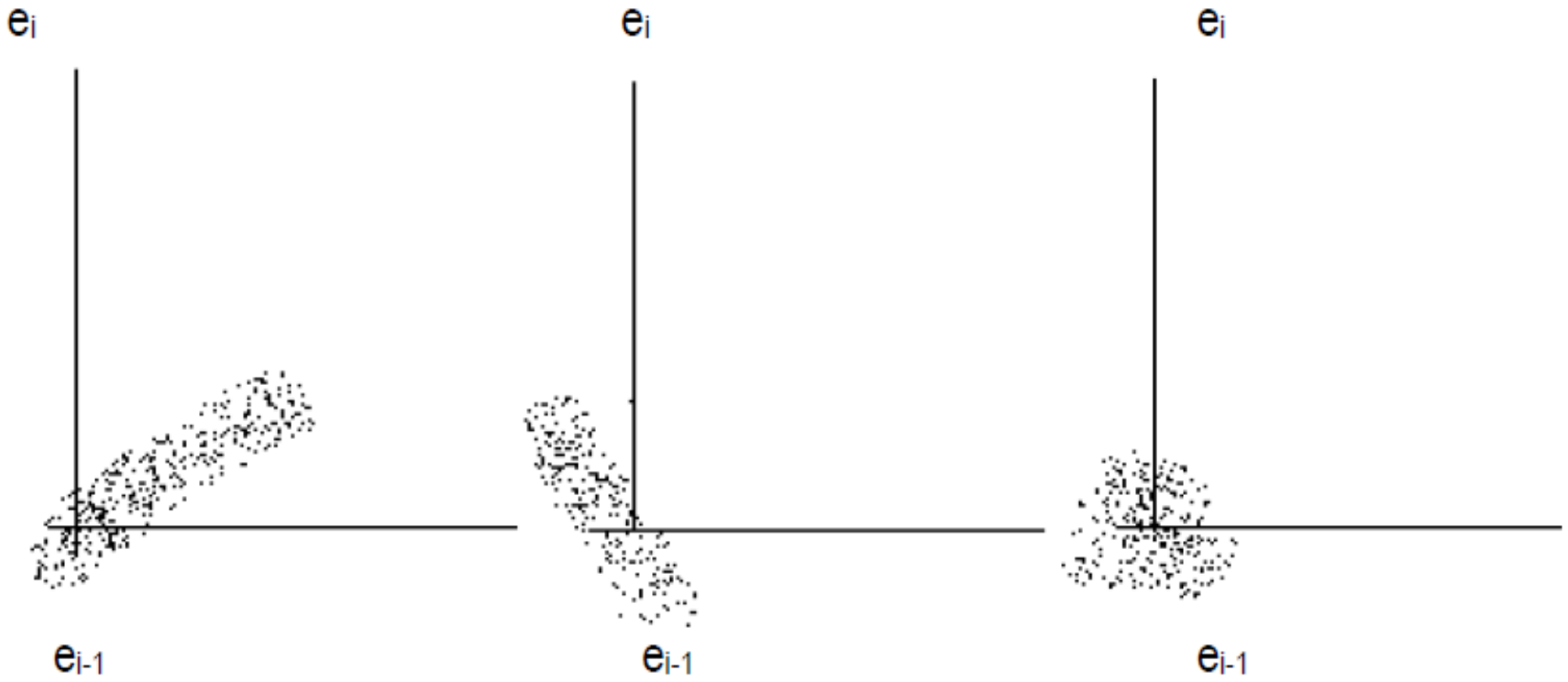
a) Se debe revisar el modelo seleccionado, este gráfico sugiere una elección incorrecta de la relación funcional



b) y c) indica que es posible un cambio estructural



d) y e) sugiere que es posible la existencia de un valor atípico



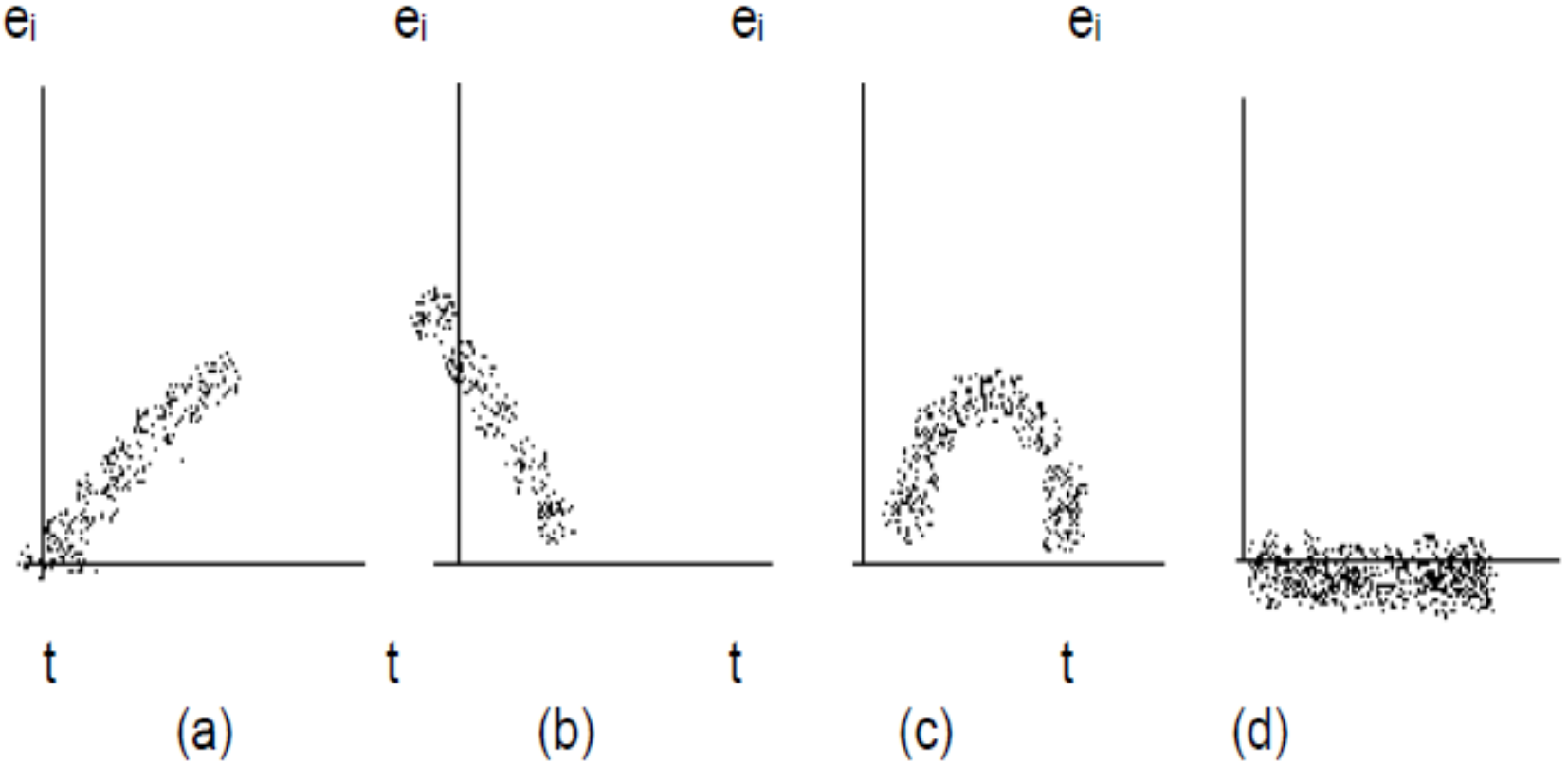
(1) (2) (3)

Posible autocorrelación:

1. Positiva de orden k
2. Negativa de orden k
3. No autocorrelación

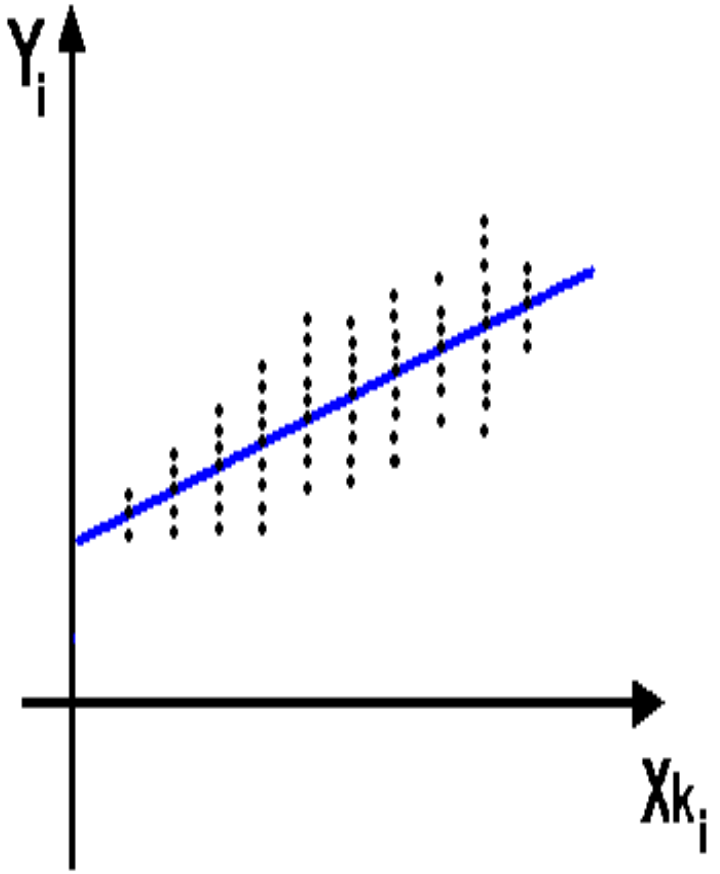
También se podrían presentar Gráficos de "e_i" contra t (t , e_i)

Siempre y cuando los datos hayan sido tomados secuenciados en el tiempo



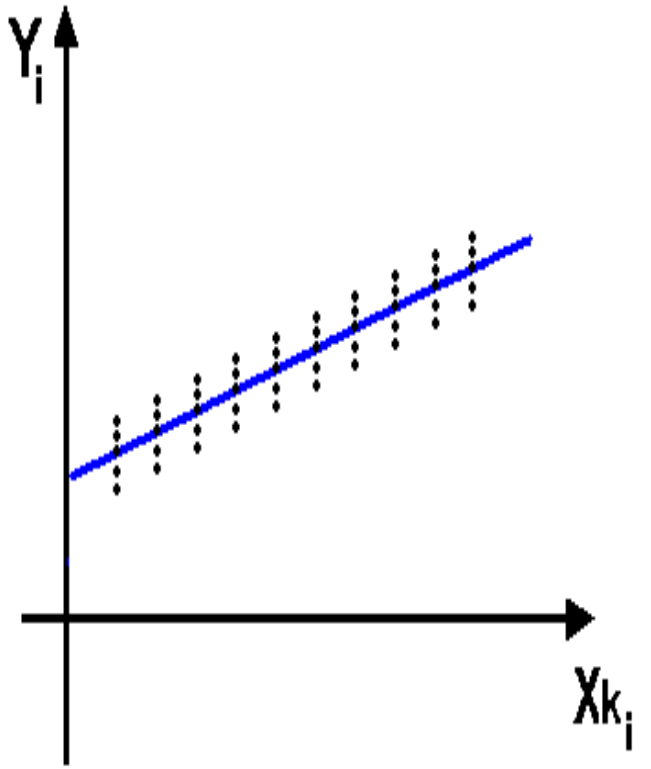
Supuestos del MRLG

1) **La sumatoria de los errores aleatorios debe tener media igual a cero $\sum e_i = 0$. Seguir una distribución Normal.**



Supuestos del MRLG

2) Las varianzas de los errores aleatorios deben ser constantes $\sigma_i^2 = \sigma j^2$ (Homocedasticidad).





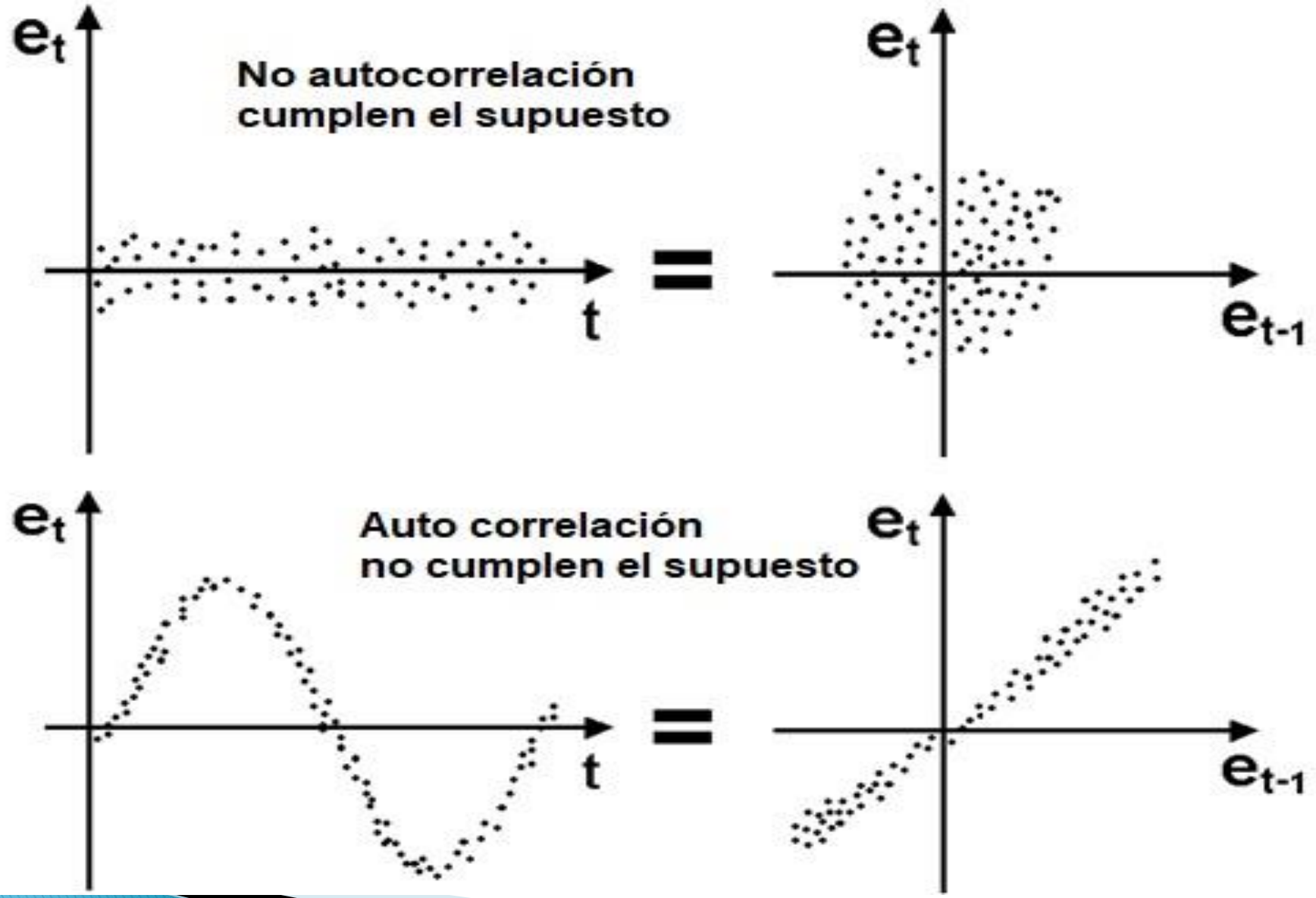
Supuestos del MRLG

3) Los errores aleatorios deben ser independientes entre sí (No Autocorrelación).

AUTOCORRELACIÓN: cuando el término de error en un período está correlacionado con el término de error en otro período que puede ser el período anterior o entre dos valores cualesquiera

Por ejemplo si se presenta una huelga que afecta la producción en un trimestre, no existe razón para creer que esta interrupción afectará la producción del próximo trimestre.

Supuestos del MRLG



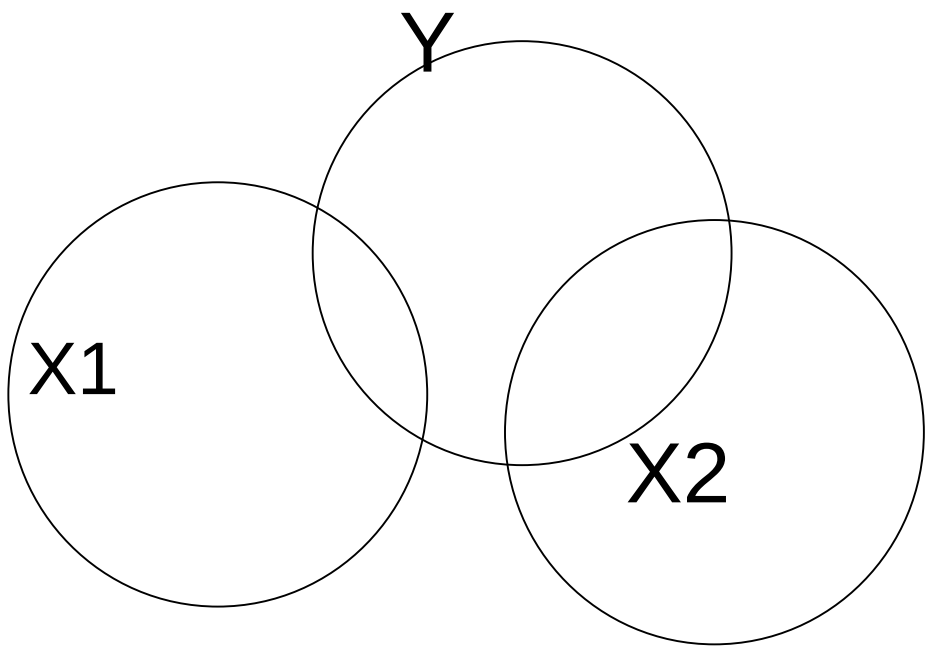
Supuestos del MRLG

4) No debe existir multicolinealidad entre las variables (No Multicolinealidad)

Se entiende por Multicolinealidad el efecto, como consecuencia de una fuerte interrelación entre dos variables independientes, ya que esto pudiera provocar que es difícil entonces conocer los efectos que cada variable tiene de forma individual sobre la variable dependiente.

Supuestos del MRLG

4) No debe existir multicolinealidad entre las variables (No Multicolinealidad)

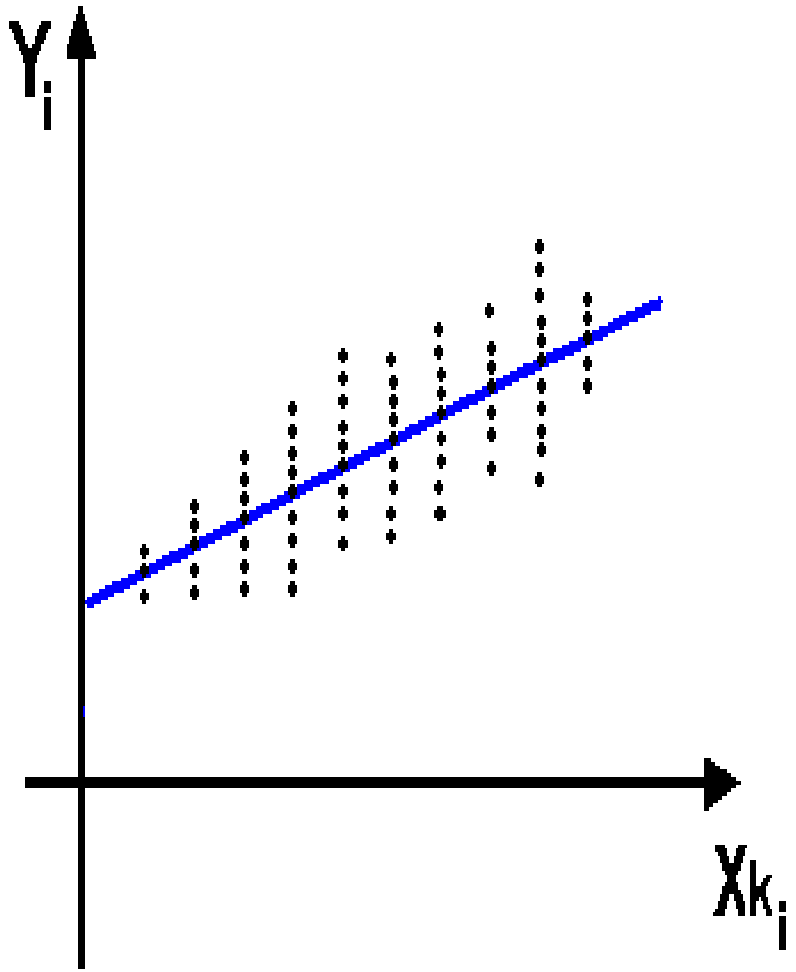


Supuesto de Media Nula. $\bar{U} = 0$ ó $\forall i$ $E(U_i) = 0$

Es un supuesto no verificable en la práctica porque su verificación partiría del trabajo con la e_i y por propiedades del método empleado MCO, la $\Sigma e_i = 0$ por tanto $e_i = 0$ no obstante si se cumple los demás supuestos y además el análisis del modelo arroja que es bueno es porque el supuesto de media nula se cumple.

Supuesto de Media Nula. $\bar{U} = 0$ ó $\forall i$ $E(U_i) = 0$

$\Sigma e_i = 0$ por tanto $e_i = 0$.



1.Supuesto de Normalidad

$U_i \sim N(0;\sigma)$

Este supuesto es necesario para garantizar la validez de las pruebas **t y F** aunque si el tamaño de la muestra es suficientemente grande su violación no es de gran importancia.

1.Supuesto de Normalidad $\forall i \; U_i \sim N(0;\sigma)$

Para muestras pequeñas en caso de violarse pueden de todas formas explicarse dichas pruebas estando consiente de que no necesariamente son del todo válida, y la decisión tomada es mejor si el estadígrafo de prueba se aleja suficientemente del valor crítico o $P_{valor} >> \alpha$

La normalidad se puede probar por varias pruebas estadísticas:

- Jarque y Bera,
- Kolmogorov-Smirnov.

En nuestro caso estudiamos la prueba Kolmogorov-Smirnov con SPSS.

Prueba de Kolmogorov-Smirnov

Es una prueba no paramétrica que no requiere supuesto sobre la distribución de la población. Es una prueba más potente y práctica que la Chí Cuadrado ($\chi 2$).

Planteamiento de las hipotésis

Ho $u_i \rightarrow N(0 , \sigma)$ Hay Normalidad

$H_i \ u_i \not\rightarrow N(0 , \sigma)$ No hay Normalidad

Región critica

Si Pvalor prueba K-S < α Entonces rechazo Ho

No se cumple el supuesto de Normalidad

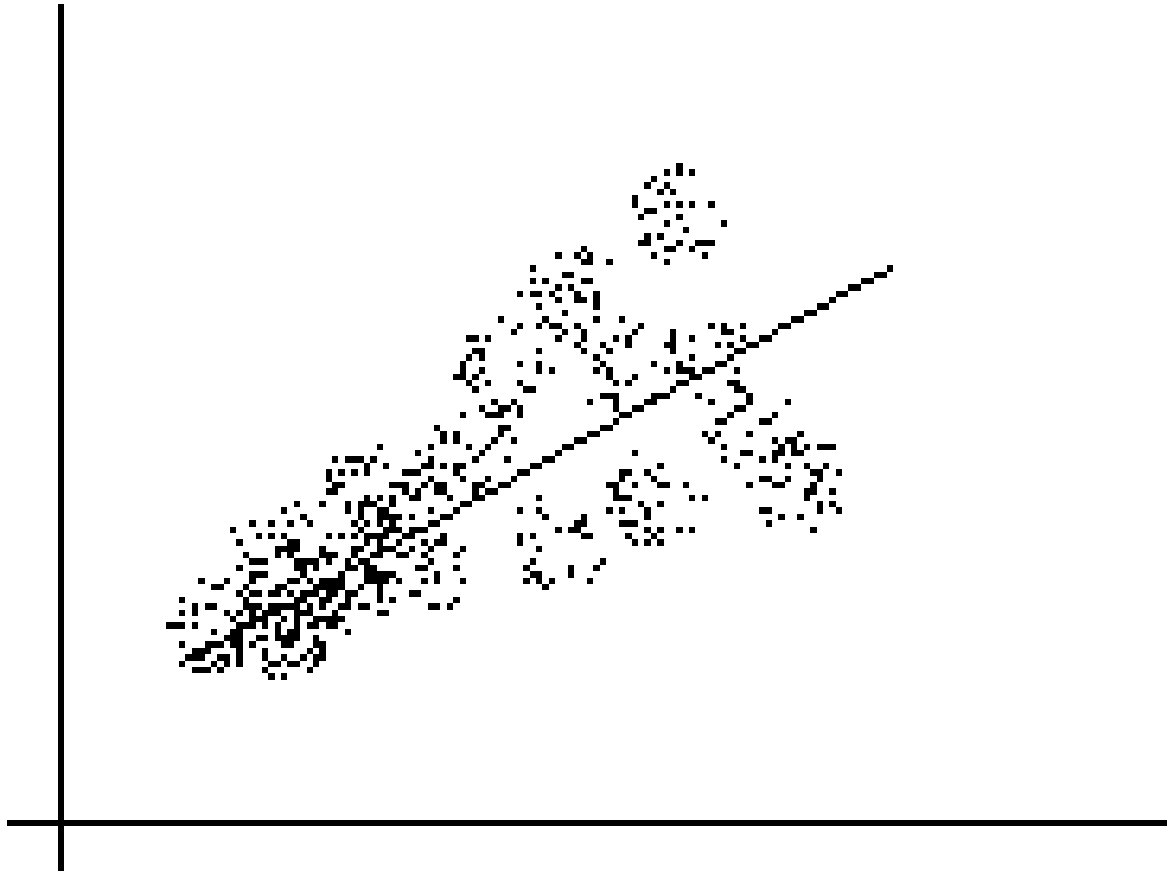
3. Supuesto de Homocedasticidad o igualdad de Varianza (V)

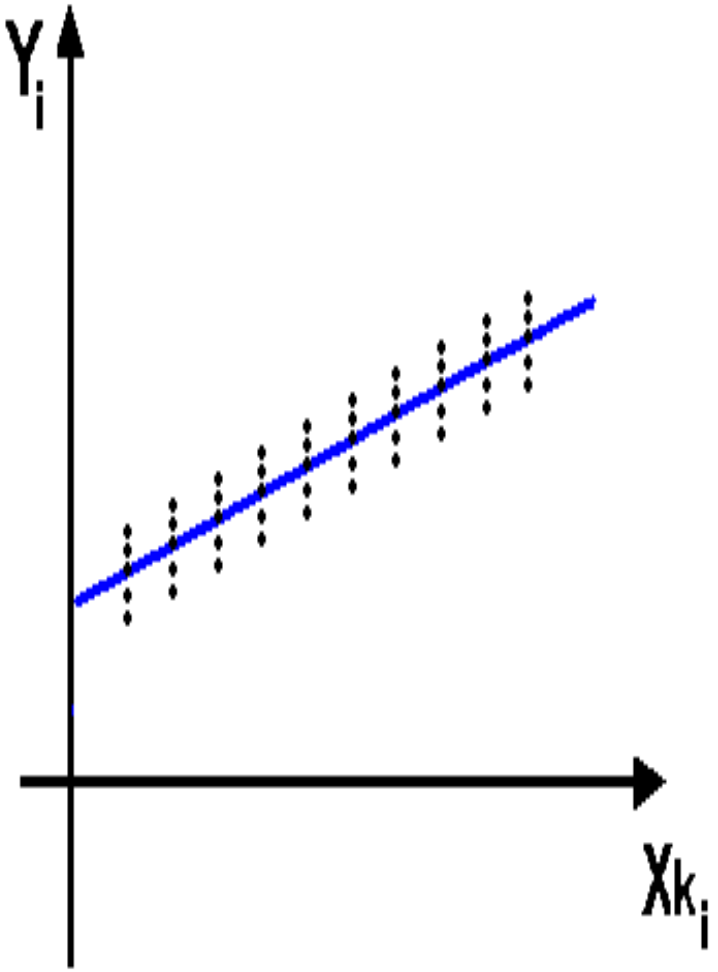
Según el modelo de regresión lineal, las perturbaciones aleatorias deben tener varianza constante a lo largo de todas las observaciones, es decir todas tienen la misma varianza:

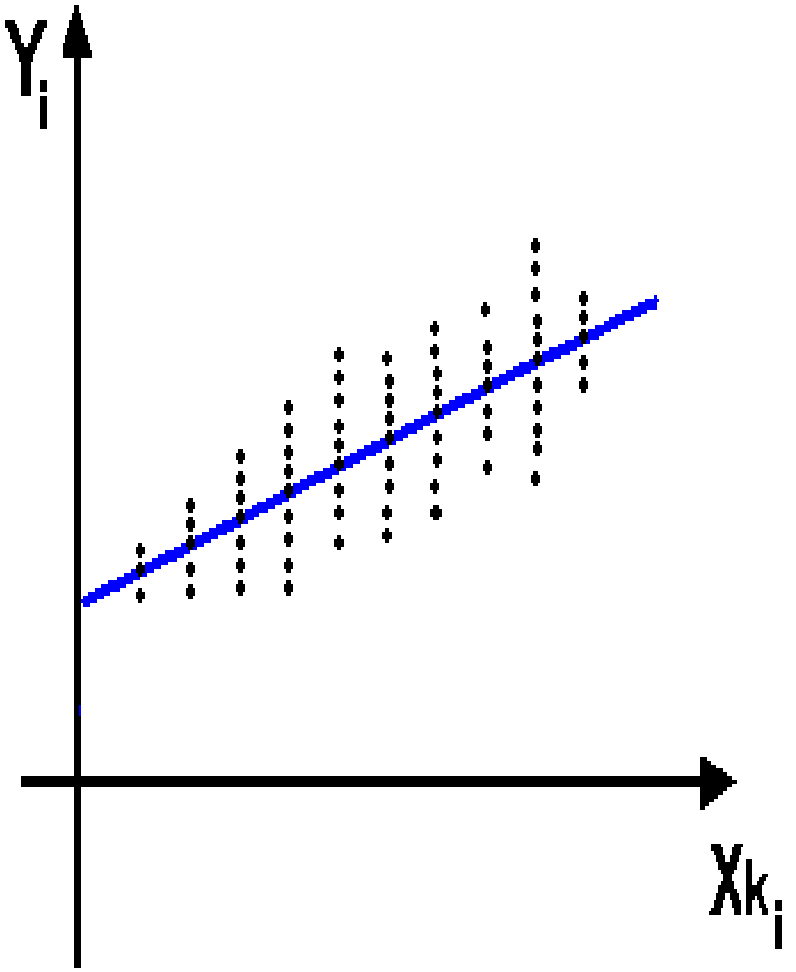
$$\forall_i \; \mathbf{V}(u_i) = \sigma^2 \quad \text{ó} \quad \forall_i \; \mathbf{E}(u_i^2) = \sigma^2$$

En general se dice que existe homocedasticidad cuando la varianza de los residuos es constante y que hay heteroscedasticidad cuando no lo es.

Heterocedasticidad







Homocedasticidad ***Heterocedasticidad***

Las razones para el incumplimiento del supuesto de Homoscedasticidad o la existencia de Heterocedasticidad, es muy variada:

- Se puede deber al propio modelo económico
- Por transformaciones que se efectúan sobre la variable
- Si se omiten variables relevantes en el modelo

- Por el tipo de dato que se utilice

- Cuando la información muestral de que se dispone consta de datos agregados procedentes de distintas submuestras

- Cuando se dispone de los promedios de las variables en distintas submuestras

Propiedades de los estimadores MCO del modelo de regresión en presencia de heterocedasticidad (no se cumple el supuesto)

- ✓Siguen siendo insesgados, pues $E(b_i) = \beta_i$.
- ✓Siguen siendo consistentes, o sea, a medida que crece el tamaño de la muestra se acercan a los verdaderos valores.
- ✓Ya no son eficientes, pues su varianza se hace grande.

Consecuencias de la heterocedasticidad al aplicar MCO

- ✓ La varianza residual de la muestra $(s_R)^2$ es un estimador sesgado de σ^2.

- ✓ La varianza de los estimadores $V(b_j)$ es sesgada por lo que se estará sobreestimando o subestimando esa varianza y no se sabe si el sesgo es positivo o negativo.

Consecuencias de la heterocedasticidad al aplicar MCO

- ✓ Los intervalos de confianza son más anchos de lo necesario, pues $\beta i = bi + d(b_i)$ y $d(b_i) = t(_{1-\alpha/2\ (n-k)}) s(bi)$

- ✓ Las pruebas “t” y “F” dejan de ser válidas.

- ✓ El R^2 puede sobreestimar la bondad del ajuste.

Consecuencias de la heterocedasticidad al aplicar MCO

- ✓Se puede concluir que la heterocedasticidad no invalida las propiedades de insesgadez y de consistencia de los estimadores MCO; pero no son eficientes, ni siquiera cuando las muestras son grandes, por lo que la falta de eficiencia le resta credibilidad a los procedimientos de pruebas de

¿Cómo detectar la Heterocedasticidad?

No existen reglas fijas y seguras para detectarla, sino unas cuantas normas muy generales, y esto se debe a que sólo se conoce la varianza poblacional, cuando se conoce la totalidad de la población. Y esto se puede decir que es una excepción antes que una regla en las investigaciones económicas.

Entre estas reglas pueden encontrarse:

Debido a la naturaleza del problema

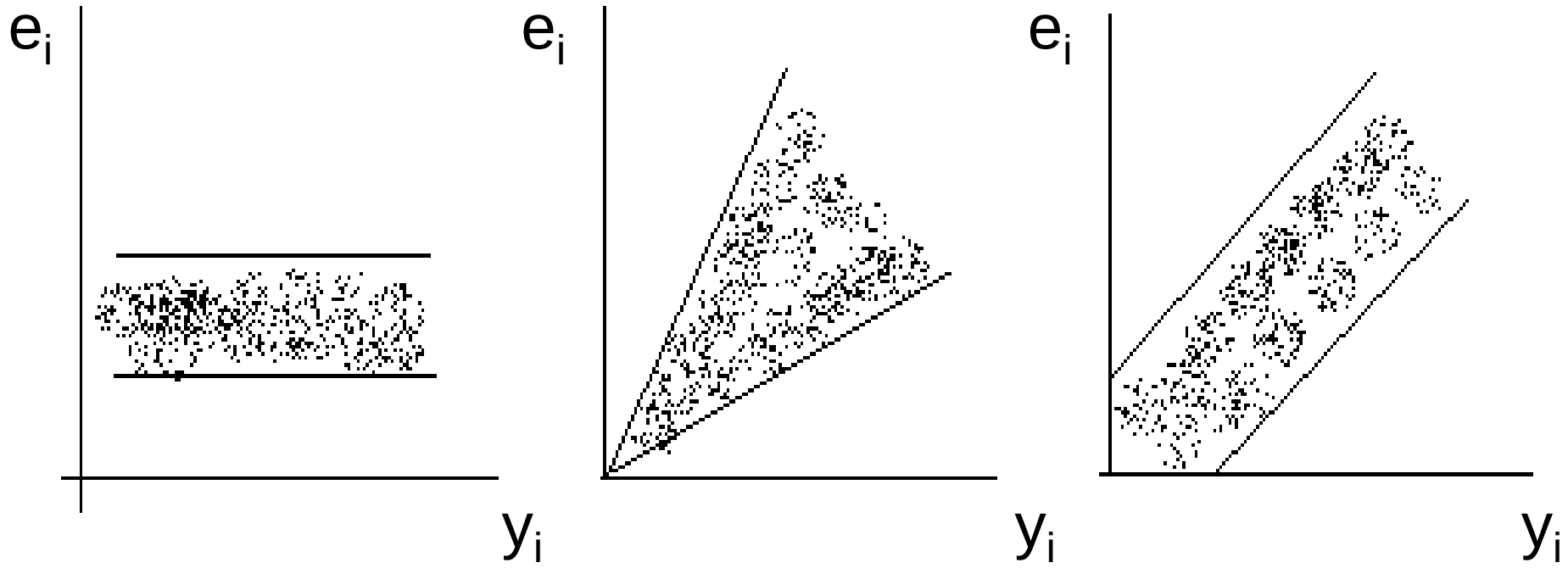
En las series de corte transversal que involucran unidades heterogéneas, la heterocedasticidad puede ser la regla general en lugar de la excepción. Por ejemplo en un análisis de Corte Transversal que considere los gastos de inversión con relación a las ventas, la tasa de interés etc. Puede presentarse heterocedasticidad si se consideran empresas pequeñas, medianas y grandes

Método gráfico

Este método ayuda mucho para tener una idea previa de la situación que se presenta; sobre todo cuando no se tiene información a priori que es lo más común.

Se obtienen los residuos estimados al cuadrado, para ver si presentan algún patrón sistemático. Estos gráficos se podrán hacer ploteando los pares

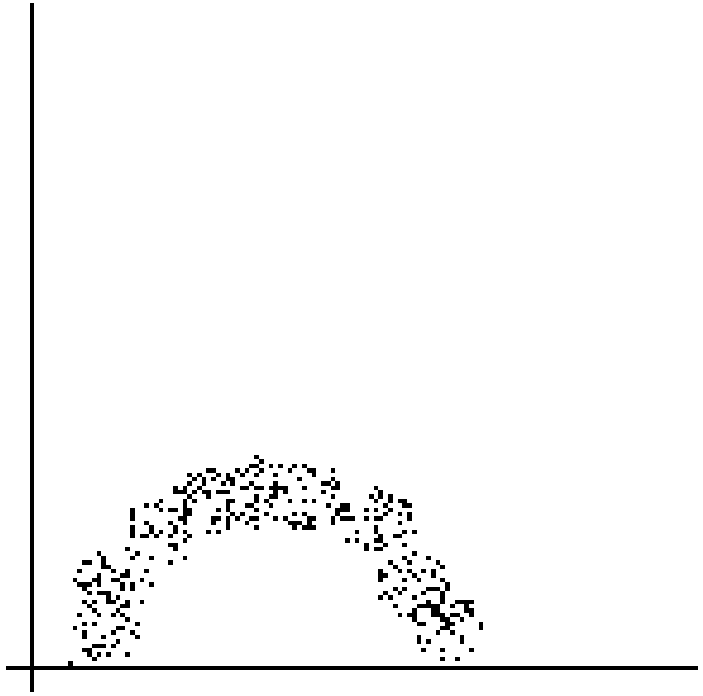
$$(\hat{y}, e_i^2)$$

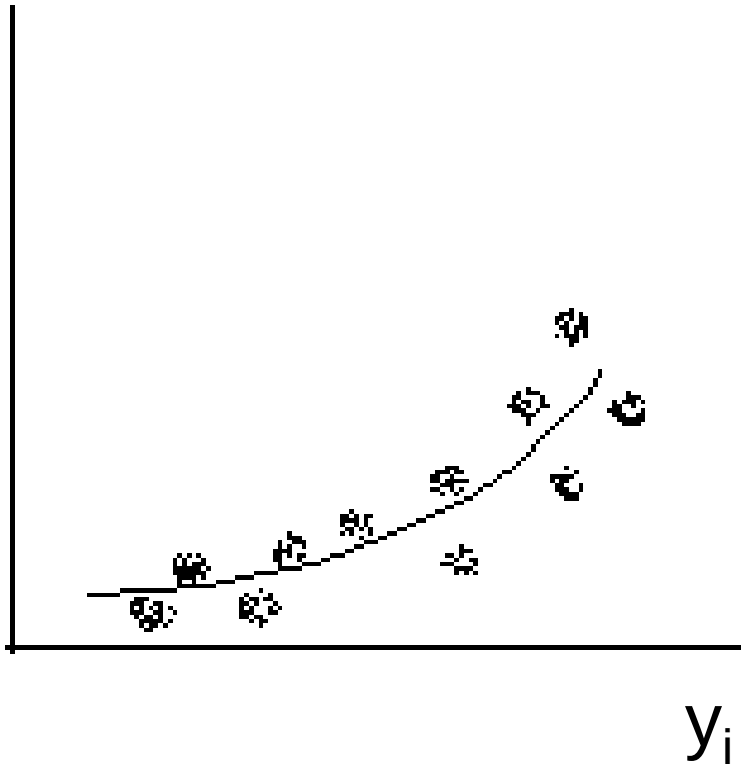


(a) No hay heterocedasticidad, ya que no hay un patrón sistemático

(b) Hay heterocedasticidad ya que crece (e_i^2) al crecer (y)

(c) Hay heterocedasticidad presentando una relación lineal





d) y e) Hay heterocedasticidad presentando una relación cuadrática.

Para comprobar la Heterocedasticidad existen diferentes Pruebas:

- Prueba de Park
- Prueba de White
- Prueba de Breusch-Pagan
- Prueba de Goldfeld-Quandt
- Prueba de Glejser
- **Prueba de Bartlett**

Prueba de Bartlett

Hipótesis y estadígrafo de prueba

Las hipótesis son:

H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_m^2$ Hay homocedasticidad

H_1: Alguna σ^2 diferente Hay heterocedasticidad

<u>Región critica</u>

Si Pvalor prueba Bartlett < α Entonces Rechazo Ho

No se cumple el supuesto de Homocedasticidad

3. Supuesto de No Autocorrelación.

$$\forall_{ij} = \mathbf{Cov}\ (U_i\ ;\ U_j) = 0 \qquad E = (U_i\ ;\ U_j) = 0$$

Presencia de Autocorrelación:

- ✓ Cuando se trabaja con datos de corte temporal, es decir, con series de tiempo.
- ✓ Cuando se han excluido del modelo variables que son necesarias.
- ✓ Cuando la forma funcional del modelo no es la correcta.
- ✓ Cuando se incluyen variables retardas en el modelo (fenómeno de la telaraña).
- ✓ Cuando se incluye la variable dependiente en forma retardada como una variable

Por tanto, se puede decir que se denomina AUTOCORRELACIÓN: cuando el término de error en un período está correlacionado con el término de error en otro período que puede ser el período anterior o entre dos valores cualesquiera de “U_T” con independencia del número de períodos intercalados entre ambos.

Por ejemplo si se presenta una huelga que afecta la producción en un trimestre, no existe razón para creer que esta interrupción afectará la producción del próximo trimestre, ó que en un análisis de los gastos de consumo familiares sobre el ingreso familiar, no se espera que el efecto de un aumento en el ingreso de una familia sobre sus gastos de consume afecte los gastos de consumo de otra familia.

En fin si existe dependencia, entonces existe autocorrelación ó correlación serial, esto es,

$$E\left(u_i u_j\right) \neq 0 \quad \forall\, i \neq j$$

Se plantea que la auto correlación se asocia fundamentalmente con series temporales, mientras que la heterocedasticidad se relaciona con datos transversales.

La autocorrelación se puede presentar por diferentes situaciones:

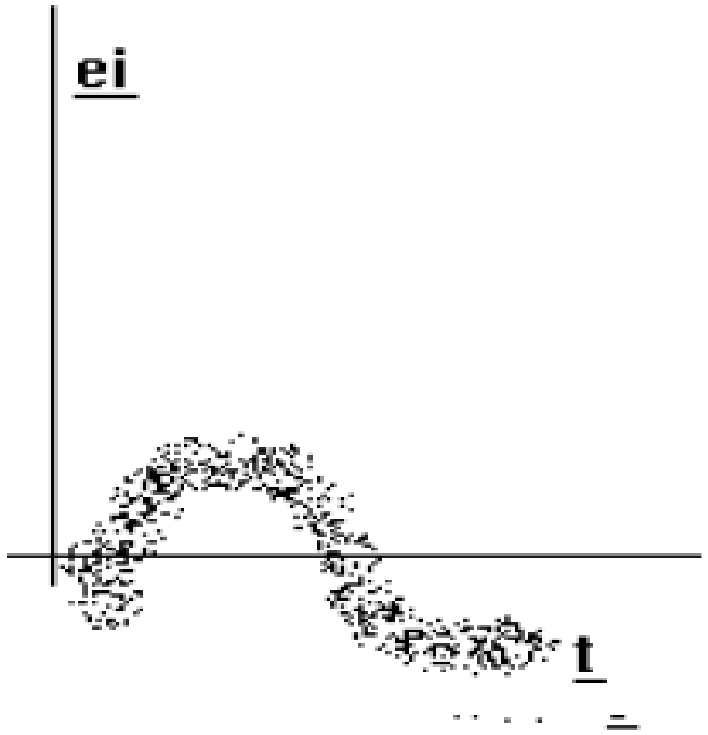
- Cuando está presente en la relación funcional, las series de tiempo
- Cuando se han excluido del modelo, variables que son necesarias
- Cuando la forma funcional del modelo no es la correcta
- Cuando se incluyen variables retardas en el modelo(fenómeno de la telaraña)

La autocorrelación se puede presentar por diferentes situaciones:

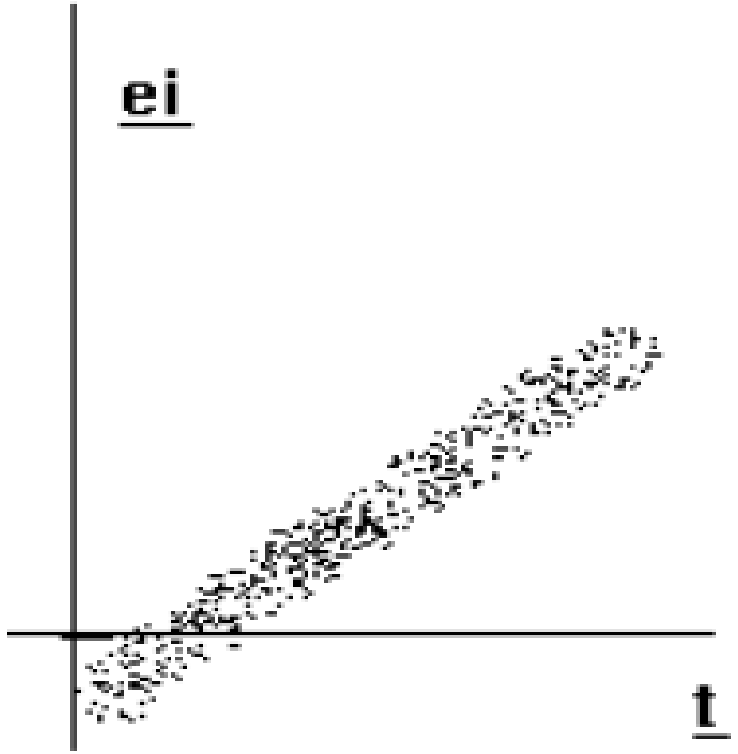
- Cuando se incluye como variable explicatoria la variable dependiente en forma retardada
- Cuando en la serie hay existencia de ciclos y tendencia
- Al manipular los datos.
- Cuando hay errores de medición.

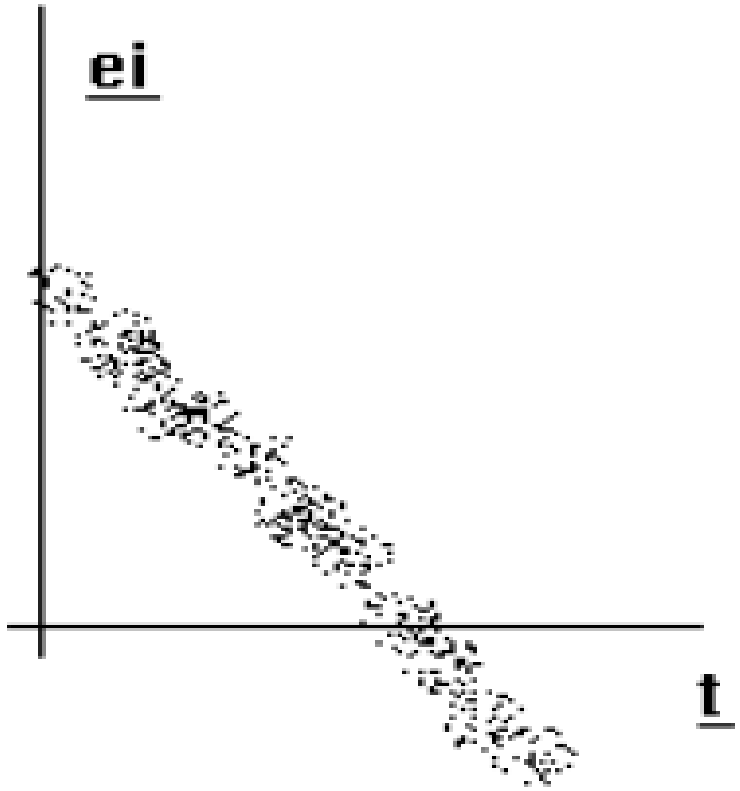
Método gráfico para detectar la autocorrelación a continuación podremos ver algunos posibles patrones de autocorrelación y no autocorrelación.

1.Se grafican los "e_i" en el tiempo (t_i, e_i)

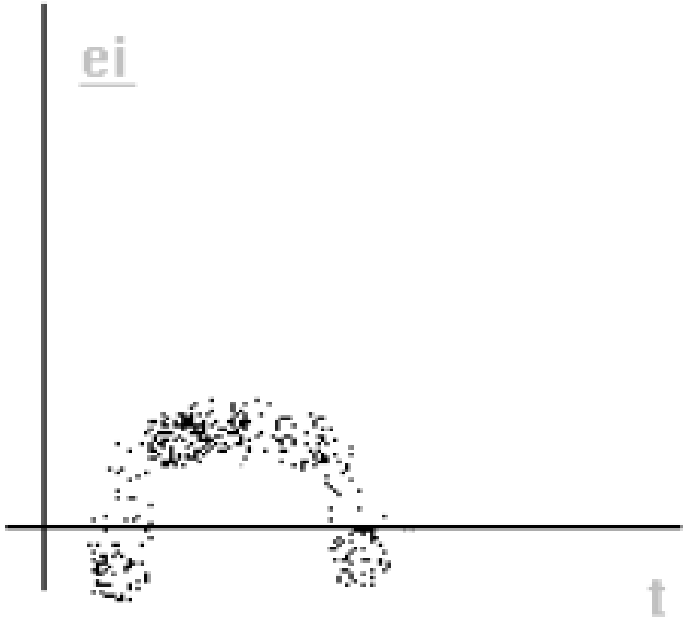


(a) que presenta un patrón cíclico

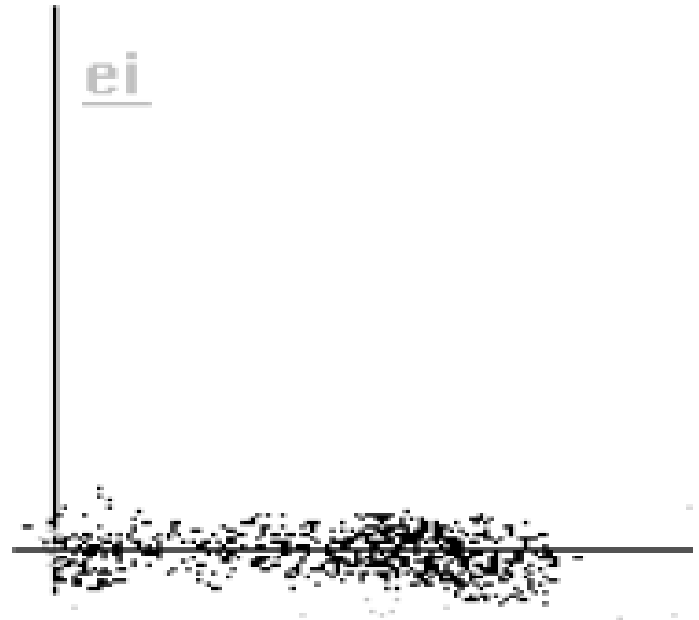




(b) y (c) se observa una tendencia lineal hacia arriba ó hacia abajo en las perturbaciones

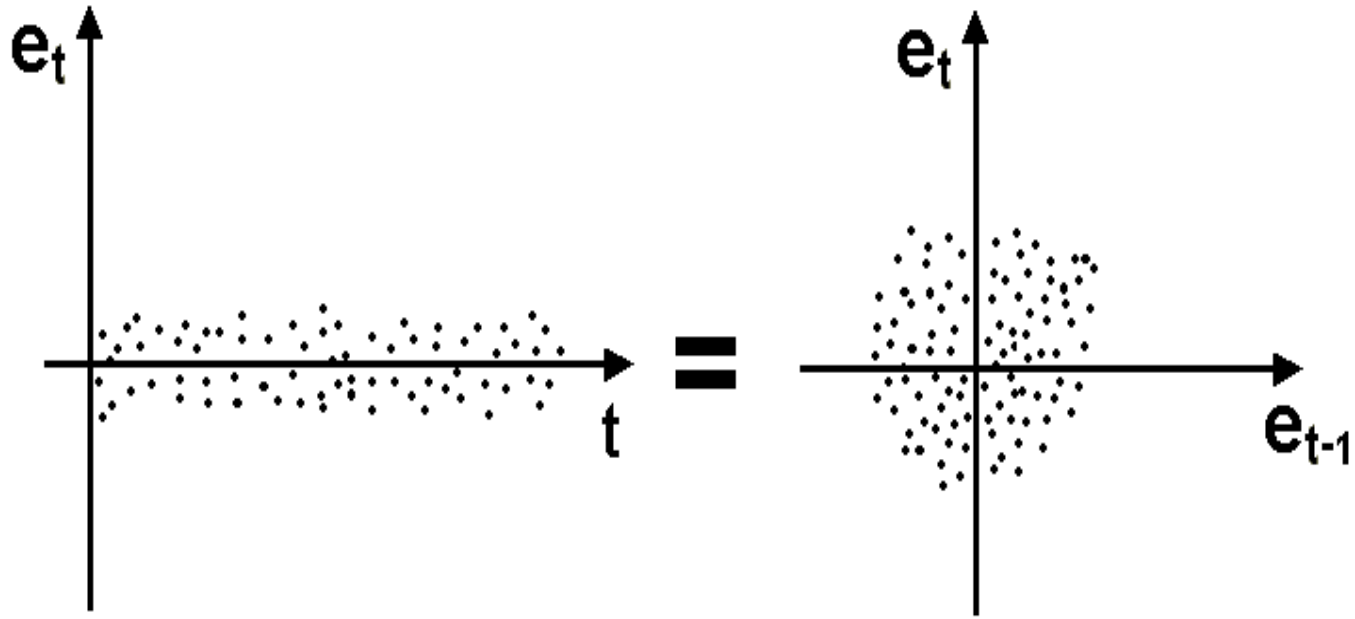


sugiere tanto una tendencia lineal como cuadrática en las perturbaciones

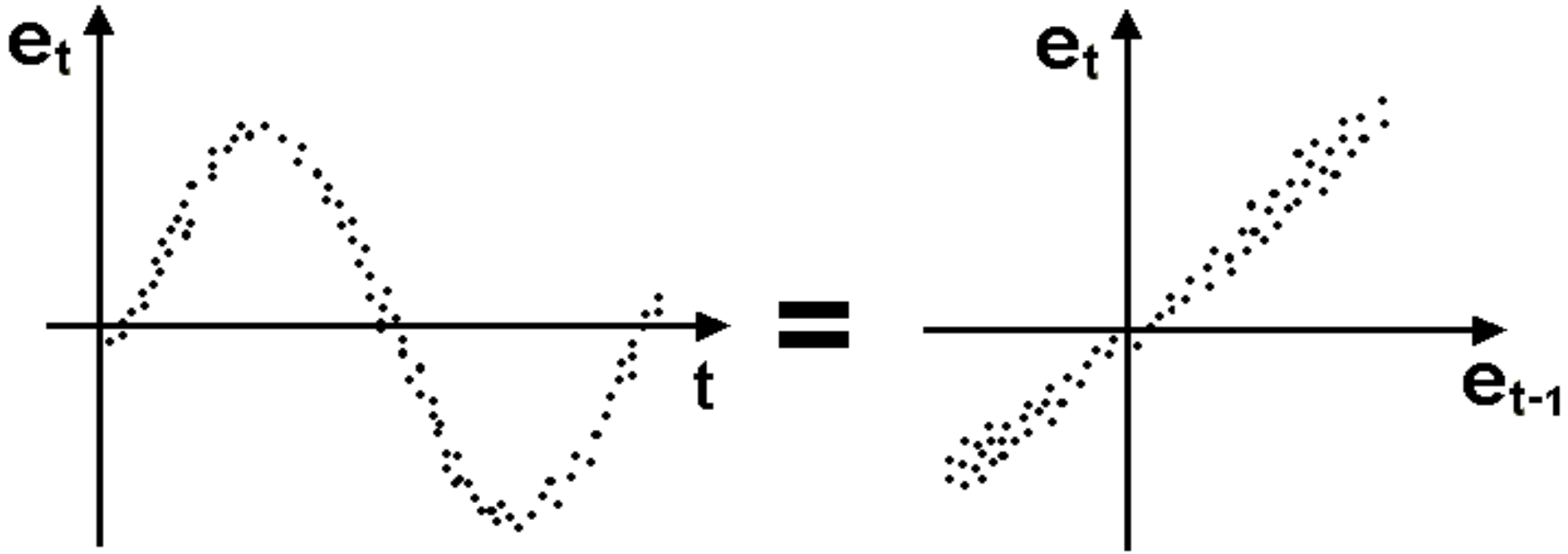


no presenta ningún patrón sistemático que apoye el supuesto de no autocorrelación del modelo de regresión lineal clásico

Patrones de No Autocorrelaión (Gráfico)



Patrones de Autocorrelación



Propiedades de los estimadores MCO del modelo de regresión en presencia de autocorrelación

- ✓ Siguen siendo insesgados, pues $E(b_i) = \beta_i$.
- ✓ Siguen siendo consistentes, o sea, a medida que crece el tamaño de la muestra se acercan a los verdaderos valores.
- ✓ Ya no son eficientes, pues su varianza se hace grande.

Consecuencias de la autocorrelación al aplicar MCO

- ✓ Los intervalos de confianza son más anchos de lo necesario.
- ✓ La varianza residual de la muestra (s_R^2) subestima el verdadero valor de la varianza poblacional (σ^2).
- ✓ Las pruebas “t” y “F” dejan de ser válidas, pudiendo llevar a conclusiones erróneas.
- ✓ El R^2 sobreestima la bondad del ajuste.
- ✓ Los estimadores MCO dejan de ser los mejores estimadores lineales insesgados (MELI)

Pruebas para detectar la autocorrelación

Existen diferentes pruebas para ello, estas son:

- Prueba de Durbin-Watson
- Prueba de Breusch-Godfrey
- Prueba χ^2 de independencia de los residuos
- Prueba de aleatoriedad o de corridas(Conocida también por Prueba de Geary)
- Prueba Ljung-Box Q-statistic
- Una prueba asintótica o de muestras grandes

Durbin-Watson: Hipótesis

Las hipótesis son:

H_0: $\rho = 0$ → *No hay autocorrelación*

H_1: $\rho \neq 0$ → *Hay autocorrelación*

Estadígrafo de prueba:

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \qquad 0 \leq d \leq 4$$

El estadístico “d” oscilará entre 0 y 4 cuando su valor es próximo a 0 nos indica que existe Autocorrelación positiva de 1er orden y cuando da valores cercanos a 4 indica que existe Autocorrelación negativa de 1er orden, sin embargo, cuando el término de error es independiente a lo largo del tiempo, el estadístico “d” estará próximo a 2 de ahí que la Prueba de Durbin’Watson se pueda plantear esquemáticamente de la siguiente forma

Zona critica



Determinación de d_L y d_U

Los valores teóricos de d se buscan en la tabla del nivel de significación escogido (teniendo en cuenta que si se planteó la prueba como de dos colas se usará $\alpha/2$ que representa en la tabla una significancia de 2.5%). En la tabla se selecciona la n correspondiente (número de observaciones) y k' número de variables explicatorias o independientes (x_2; x_3….)

Toma de Decisión



- Se Rechazo o no se Rechaza Ho en dependencia de la zona en que cae el valor de d que se obtuvo .

- Si el Stadist Durbin Watson < 1,4 Rechazo Ho (No se cumple el supuesto)

4. Supuesto de No multicolinealidad.

Se entiende por Multicolinealidad el efecto, como consecuencia de una fuerte interrelación entre dos variables independientes, ya que esto pudiera provocar que es difícil entonces conocer los efectos que cada variable tiene de forma individual sobre la variable dependiente.

Consecuencias de la Multicolinealidad.

- Grandes varianzas y covarianzas de los estimadores.
- Intervalos más amplios.
- Valores o pruebas “t” no significativas y R^2 altos.

La multicolinealidad no viola los supuestos básicos de la regresión, por eso es que se dice que es un problema de la muestra con que se trabaja

Algunas reglas para detectarla.

1. La matriz de correlación da una buena información, desde dos puntos de vistas, porque no sólo permite detectar si existe multicolinealidad entre las variables independientes, sino también cuales deben ser las variables independientes que deben estar en el modelo.
2. R^2 alto y pruebas "t" no significativas.

   La prueba del modelo general da un R^2 alto, pero las pruebas "t" de los coeficientes individuales, dan como resultado que ninguno ó pocos de ellos son significativamente diferentes de cero.

Es bueno recalcar que:

- Una variable independiente o predictora debe tener una correlación fuerte con la variable dependiente.

- Una variable independiente o predictora no debe tener una correlación demasiada alta con ninguna otra variable predictora.

Método para detectar la Multicolinealidad.

- Revisar la matriz correlación

	X_1	X_2	X_3
X_1	1	$r_{1,2}$	$r_{1,3}$
X_2	$r_{2,1}$	1	$r_{2,3}$
X_3	$r_{3,1}$	$r_{3,2}$	1

Si hay al menos un $r_{i,j} > 0,6$ las variables son linealmente dependientes, hay multicolinealidad no se cumple el supuesto de No Multicolinealidad

Método para detectar la Multicolinealidad.

- Revisar el VIF (Factor de Inflacción o incremento de varianza)

	VIF
Constante	12
X_1	5.6
X_2	8.7

Si algún VIF de las variables es > 10 No se cumple el supuesto de No Multicolinealidad

5.2. Especificaciones de un Modelo de Regresión Lineal General

5.2.1.Atributos de un buen Modelo de Regresión.

Parsimonia.

Un modelo nunca puede describir completamente y exactamente la realidad, porque sería muy complejo.

Debe ser tan simple como sea posible.

Según Milton Friedman “un modelo es importante si explica mucho en forma resumida”.

Atributos de un buen Modelo de Regresión.

- Bondad de Ajuste.

Un modelo con un R^2 elevado es un signo alentador. Aunque se necesita de otras precisiones.

Atributos de un buen Modelo de Regresión.

Identificabilidad.

Significa que para un conjunto de datos determinados los parámetros estimados deben tener valores únicos o lo que es lo mismo deben tener un único estimador.

Atributos de un buen Modelo de Regresión.

- Consitencia teórica.

Un modelo puede no ser bueno a pesar de tener un R^2 elevado ejemplo estimadores con signos equivocados (la demanda de pollo disminuye si los precios son menores)

Atributos de un buen Modelo de Regresión.

- Poder Predictivo.

Citando nuevamente a Friedman “… la única prueba de validez para un modelo es la comparación de sus predicciones con la experiencia”.

5.2.2. Posibles errores que se pueden cometer al formular un buen modelo.

- Omitir una variable relevante (significativa).
- Incluir una variable que no sea relevante (significativa).
- Errores al recoleccionar los datos.

CONCLUSIONES

- UNA VEZ QUE SE HAYA VERIFICADO EL BUEN AJUSTE DEL MODELO Y LA SIGNIFICACIÓN DE LAS VARIABLES, ES DE VITAL IMPORTANCIA VERIFICAR EL CUMPLIMIENTOS DE LOS SUPUESTOS DE LA REGRESIÓN.

  ELLOS SON: NORMALIDAD

  HOMOCEDASTICIDAD.

  NO AUTOCORRELACIÓN.

  NO MULTICOLINEALIDAD

ESTUDIO INDEPENDIENTE PARA LA PROXIMA CLASE PRACTICA

1) En una industria del municipio se realizó un estudio sobre los costos de producción, procesando un número de observaciones de las variables:

Y = Costo de producción (CP) en miles de pesos

X_1 = Gasto de salario (GS) en miles de pesos

X_2 = Gasto de Materia Prima (GMP) en miles de pesos

Se procesó la información mediante el paquete Statgraphics correspondiente al modelo de regresión lineal, recogiéndose la siguiente información:

Parámetro	Estimador	Error Estand.	Statistic	P-Value	VIF
CONSTANT	0.540541	0.182307	2.965	0.0593	11.03
X1	0.202703	0.200893	1.00901	0.0373	0.58
X2	1.35135	0.275623	4.9029	0.0162	9.68

Análisis de varianza

Fuente	Sumade Cuadrados	GL	Cuadrado Medio	F-Ratio	P-Value
Model	6.30631	2	3.15315	17.95	0.0214
Residual	0.527027	12	0.175676		
Total (Corr.)	6.83333	14			

R^2 = 98.04% R^2(ajustado) = 97.36%

Durbin-Watson statistic = 1.82225 P valor Prueba K-S = 0.0004

Prueba de Bartlett: 1.58149 P. Valor de la Prueba de Bartlett = 0.000568

Matriz de Correlación

	X1	X2
X1	1	0,23
X2	0,23	1

Dado la información anterior:
Verifique el cumplimiento de todos los supuestos

Muchas gracias por su atención

Prof: Rodolfo Pérez León
Email: rperezleon@uart.edu.cu



UNIVERSIDAD DE ARTEMISA

ESTADÍSTICA II

MSc. Rodolfo Pérez Léon
Email: rperezleon@uart.edu.cu

UNIVERSIDAD DE ARTEMISA

Tema 6. SERIES DE TIEMPO.

6.1. Componentes de las Series de Tiempo
- 6.1.1. Tendencia.
- 6.1.2. Cíclico.
- 6.1.3. Estacionalidad
- 6.1.4. Irregular.

6.2. Clasificación de las Series de Tiempo.

6.3. Modelos Cásicos.

6.4. Cálculo de la Tendencia.
- 6.4.1. Media Móviles.
- 6.4.2. Mínimos Cuadrados Ordinarios.

UNIVERSIDAD DE ARTEMISA

Tema 6. SERIES DE TIEMPO.

6.5. Análisis de los residuos de la Series de Tiempo.
6.6. Análisis de la aleatoriedad de la Series de Tiempo.
6.7. Selección del mejor modelo.
6.8. Verificación de los residuos de un modelo.
6.9 Estimaciones con series de tiempo.

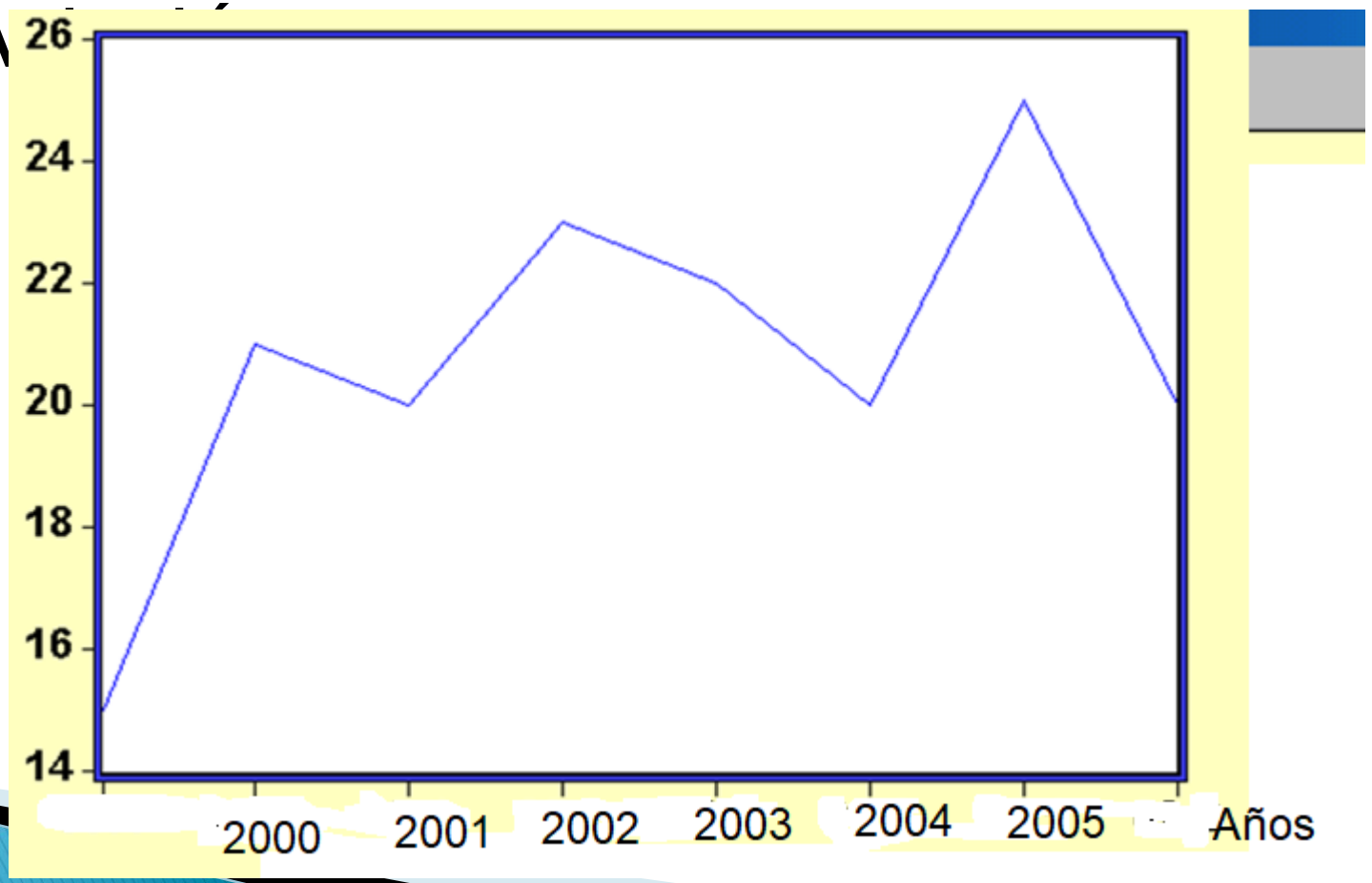
Objetivos: Que el alumno sea capaz de:

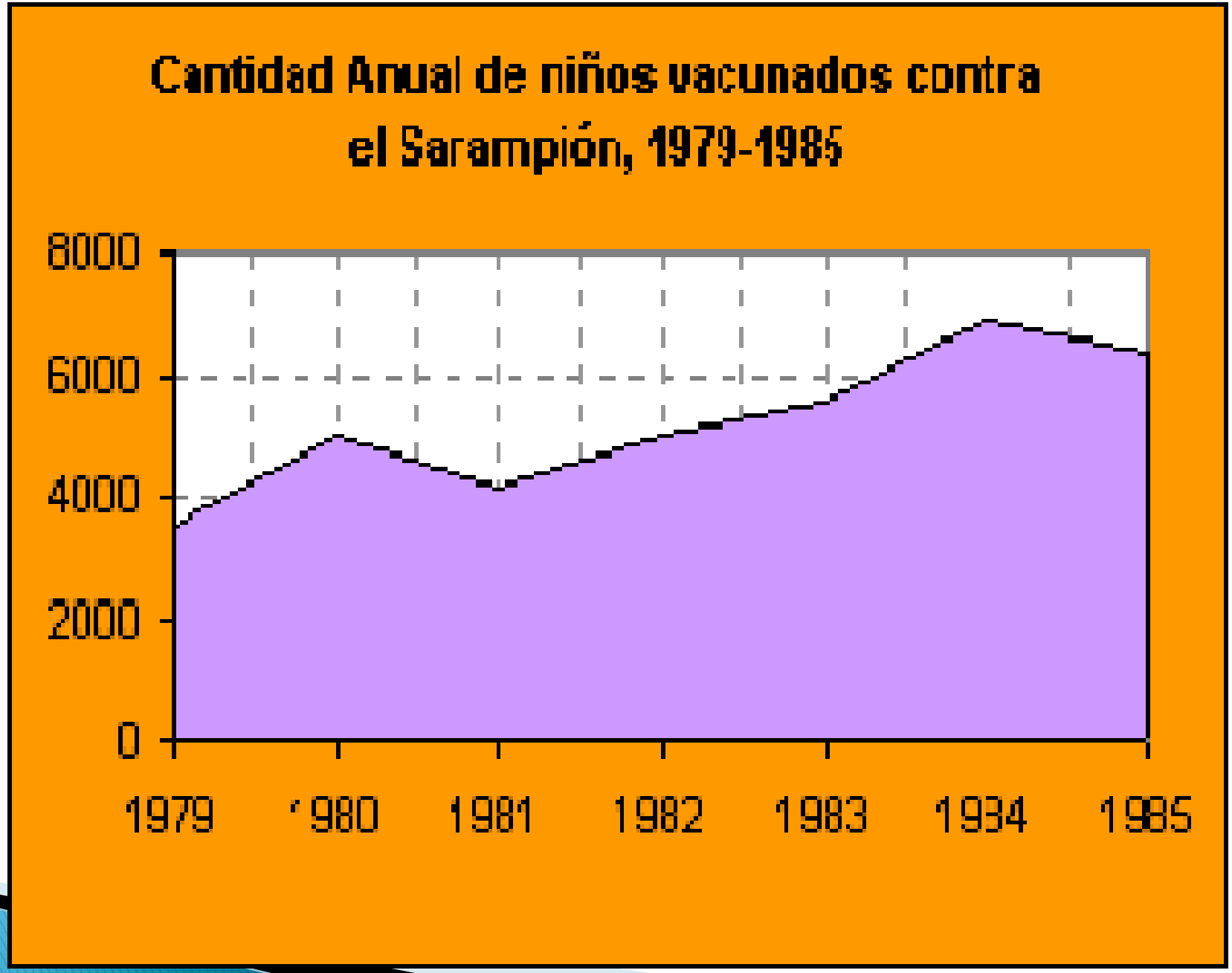
- Identificar los componentes de una serie de Tiempo.
- Calcular la tendencia en una Serie de Tiempo.
- Verificar la aleatoriedad de los residuos.
- Realizar estimaciones.

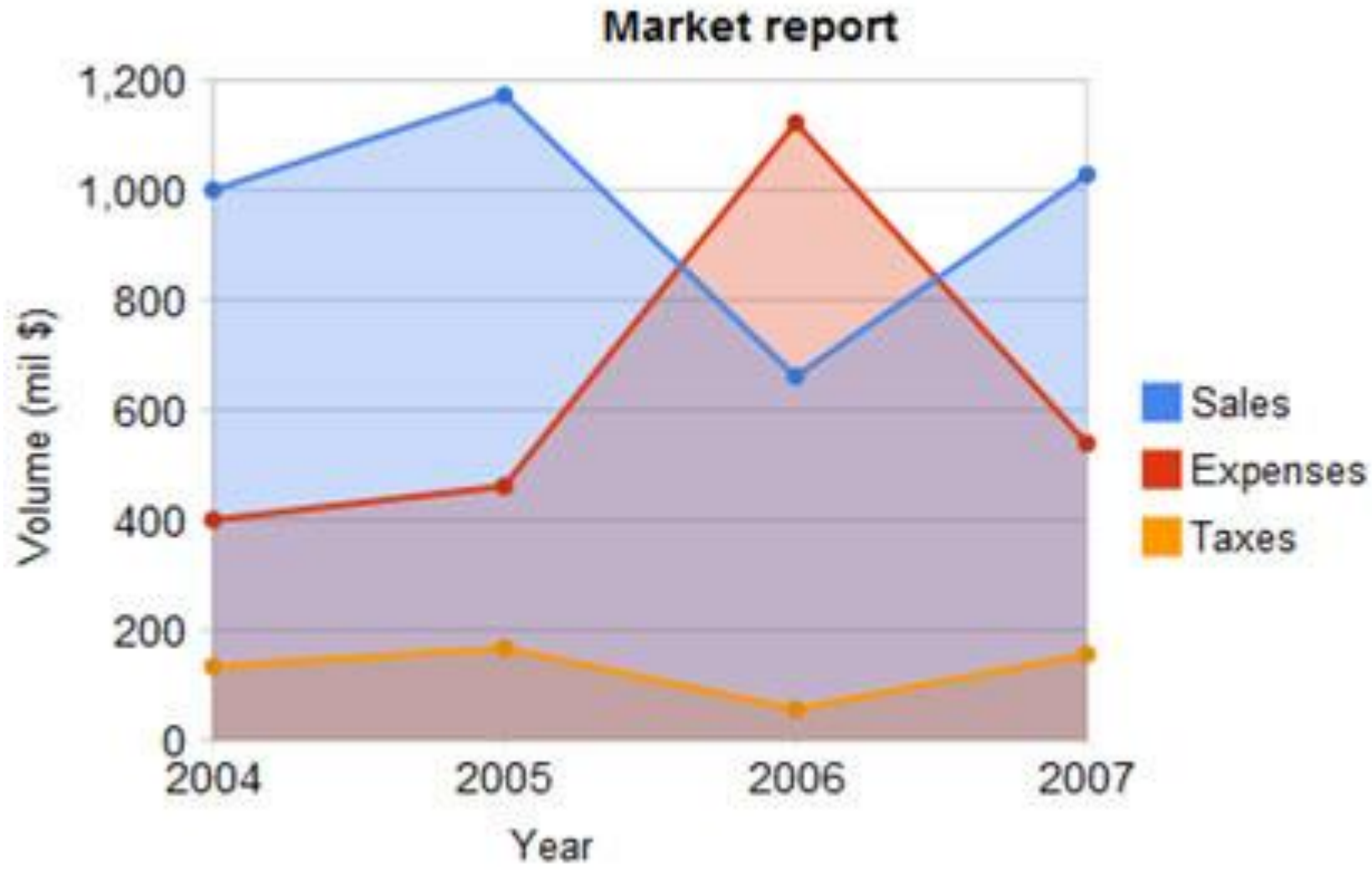
Una serie de tiempo, es un grupo de datos cuantitativos que se obtiene en períodos regulares, es decir en una sucesión de valores tomados en el tiempo, en el momento en que ocurrieron

Ejemplos:

- La producción azucarera anualmente
- Promedio trimestral de pasajeros transportados por Cubana de A




Cantidad Anual de niños vacunados contra el Sarampión, 1979-1985
8000
6000
4000
2000
0
1979
1980
1981
1982
1983
1934
1985


Market report
1,200
1,000
800
600
400
200
0
Volume (mil $)
2004
2005
2006
2007
Year
Sales
Expenses
Taxes

¿Cuáles son los Objetivos del Análisis de las Series de Tiempo?

- Aislar e identificar los factores que influyen en el comportamiento de la actividad que se analiza.
- Realizar pronósticos sobre la actividad con fines de planeación y control.

Componentes de una Serie de Tiempo.

- Componente Tendencia (T).
- Componente Cíclico (C) .
- Componente Estacional (S).
- Componente Irregular (I).

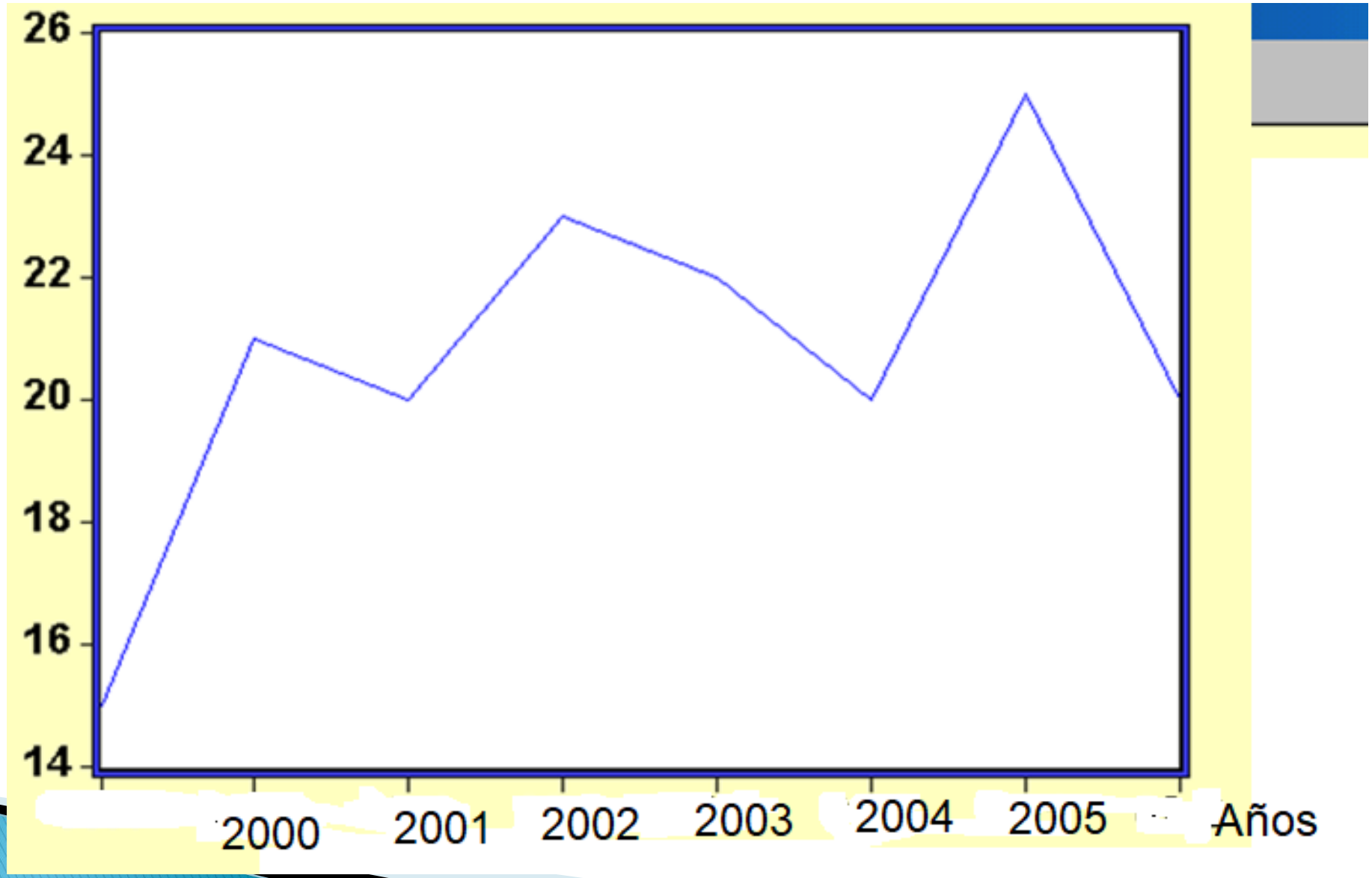
- **Componente Tendencia (T).**

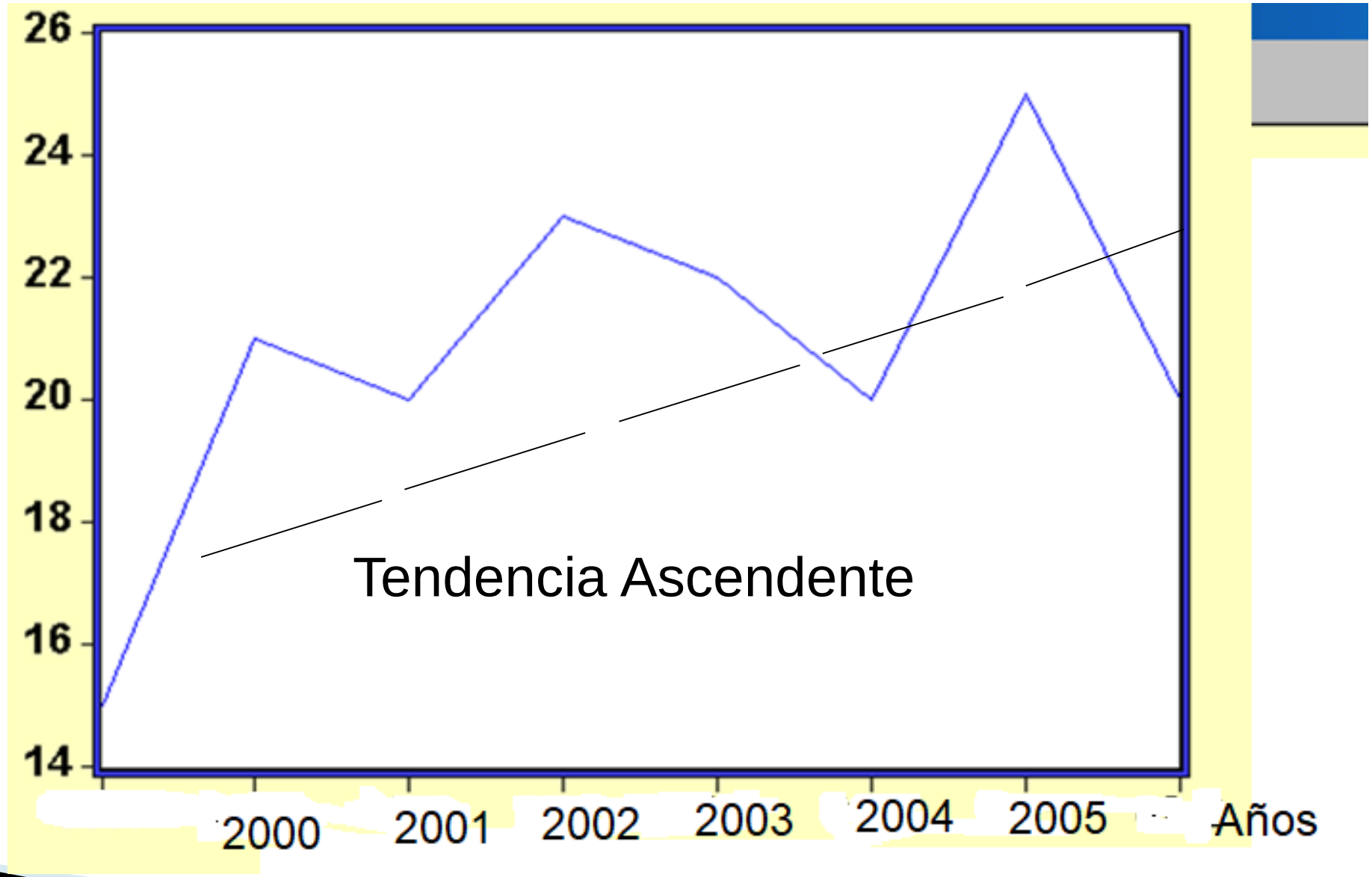
Es aquella parte de la serie que se caracteriza por un movimiento lento y suave, uniforme durante un período largo de tiempo.

La Tendencia puede describirse como una función dependiente del tiempo,

$$T = f(t)$$

Cuál es la tendencia en esta Serie ?




26
24
22
20
18
16
14
2000
2001
2002
2003
2004
2005
Años
Tendencia Ascendente

Utilidad de la tendencia.

- Permite la proyección temporal de la misma, constituyendo parte importante de cualquier estudio relacionado con la planificación a mediano y largo plazo.

- Permite aislar otros componentes del modelo.

Aspectos que inciden en la determinación de la tendencia.

1. Período de tiempo escogido

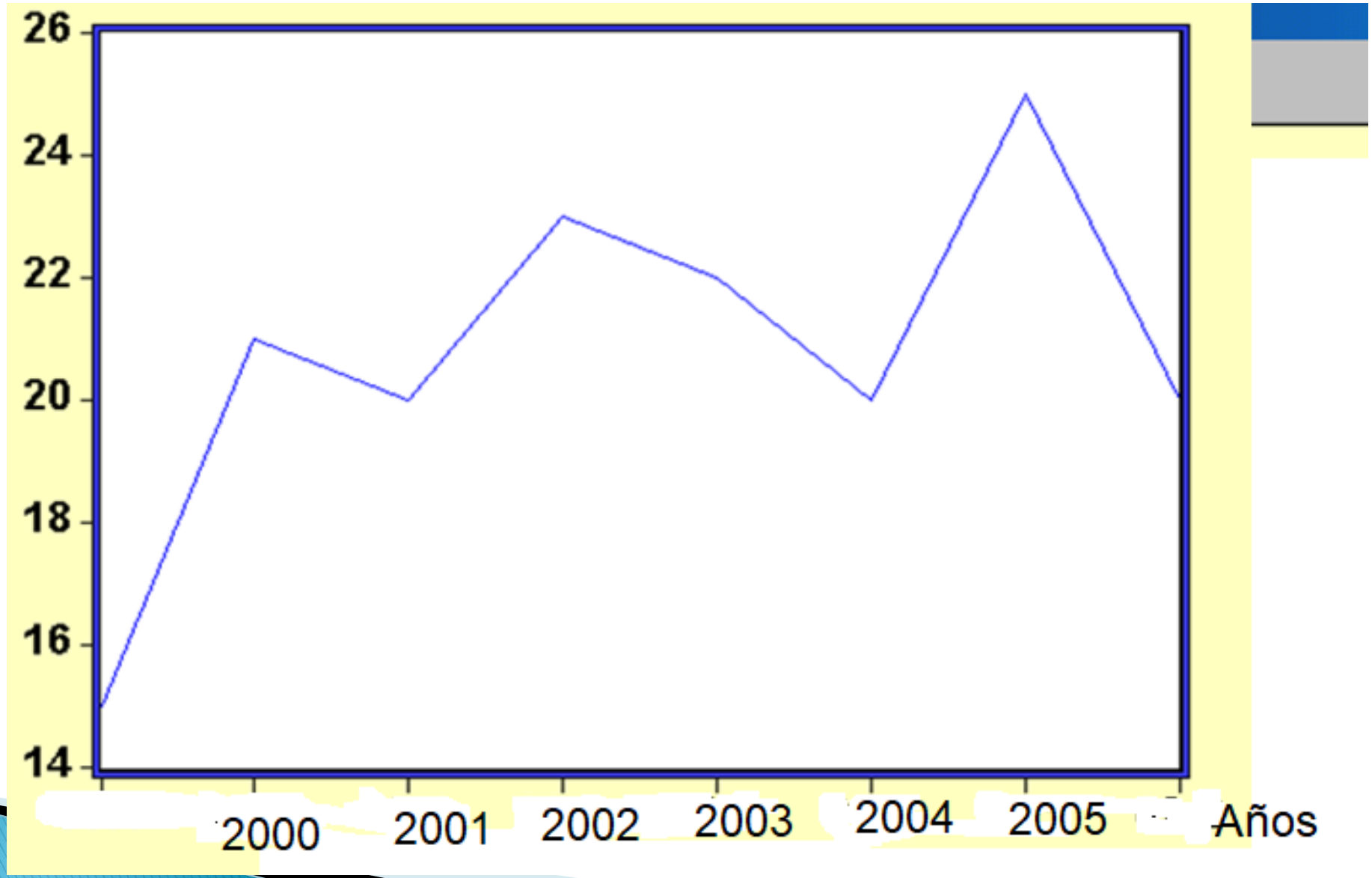
2. Tipo de ecuación propuesta

3. Método de estimación utilizado.

- **Componente Cíclico(C).**

Muestra los desplazamientos o movimientos ascendentes y descendentes de la serie.
Son movimientos oscilatorios alrededor de la tendencia con una configuración “no” regular, ni en amplitud ni en períodos.

Componente Cíclico



Se plantea que en países de economía planificada, en las series relacionadas con la actividad económica, el componente cíclico es irrelevante por lo que no se considera.

- **Componente Estacional(S)**

Resulta el componente, producto de las fluctuaciones periódicas que se presentan regularmente en la serie, durante subintervalos de un año generalmente.

- **Componente Estacional(S)**

Aunque su denominación procede de las estaciones del año, la idea de la estacionalidad puede referirse por ejemplo, a meses, trimestres, u otros subperíodos de una año, donde se manifiestan movimientos oscilatorios con periodicidad regular.

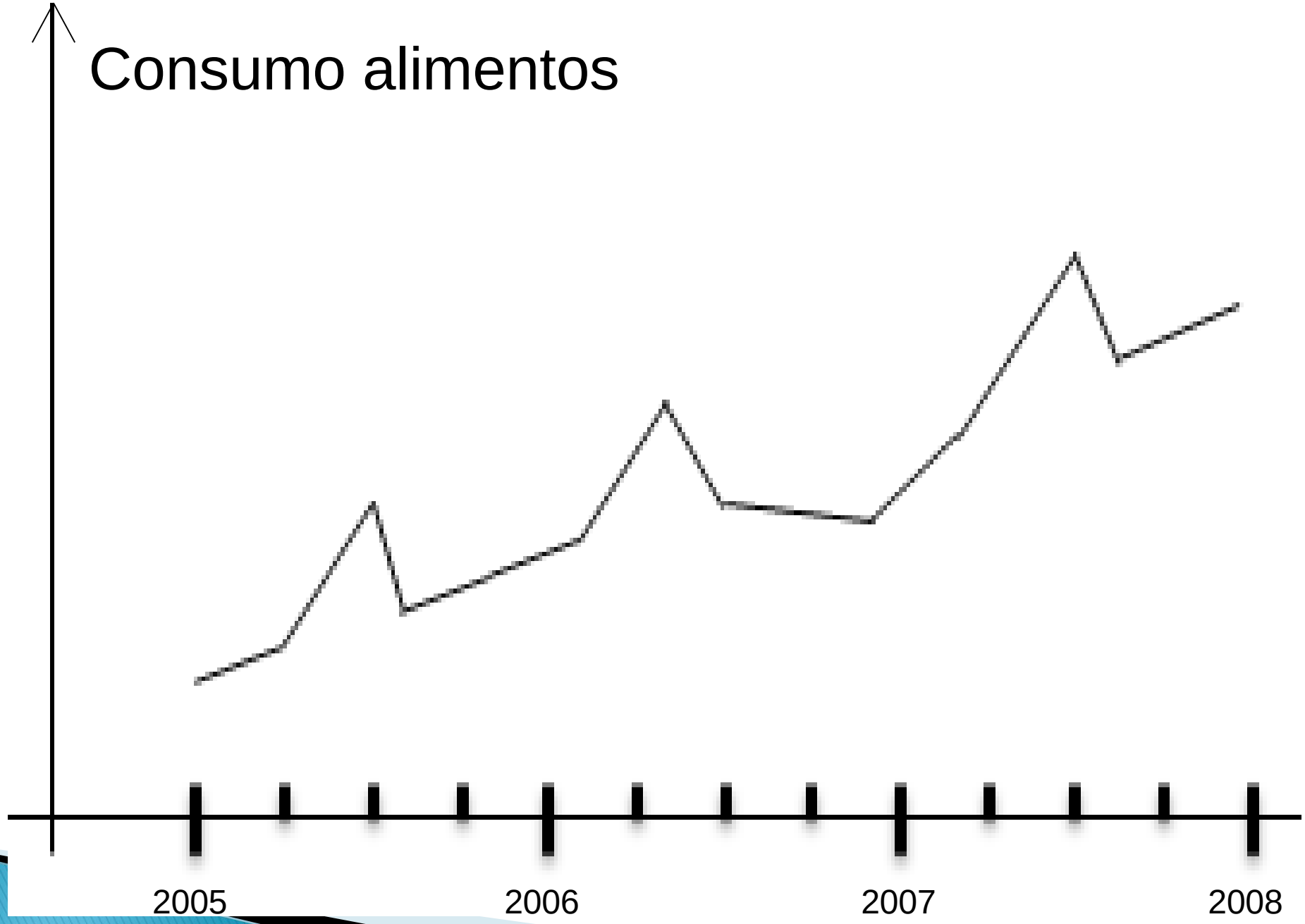
Principales factores que motivan la existencia de Estacionalidad

- Condiciones Climatológicas.
- Problemas de Calendario.
- Decisiones institucionales

Componente Estacional(S).

Ejemplo.

- Aumento del turismo nacional en los meses de julio y agosto.
- Aumento del número de pasajeros que se transportan en los meses de vacaciones.
- Disminución de lluvias caídas en los meses de enero hasta abril.
- Disminución de ventas de viandas en el ultimo semestre del año.


Consumo alimentos
2005
2006
2007
2008

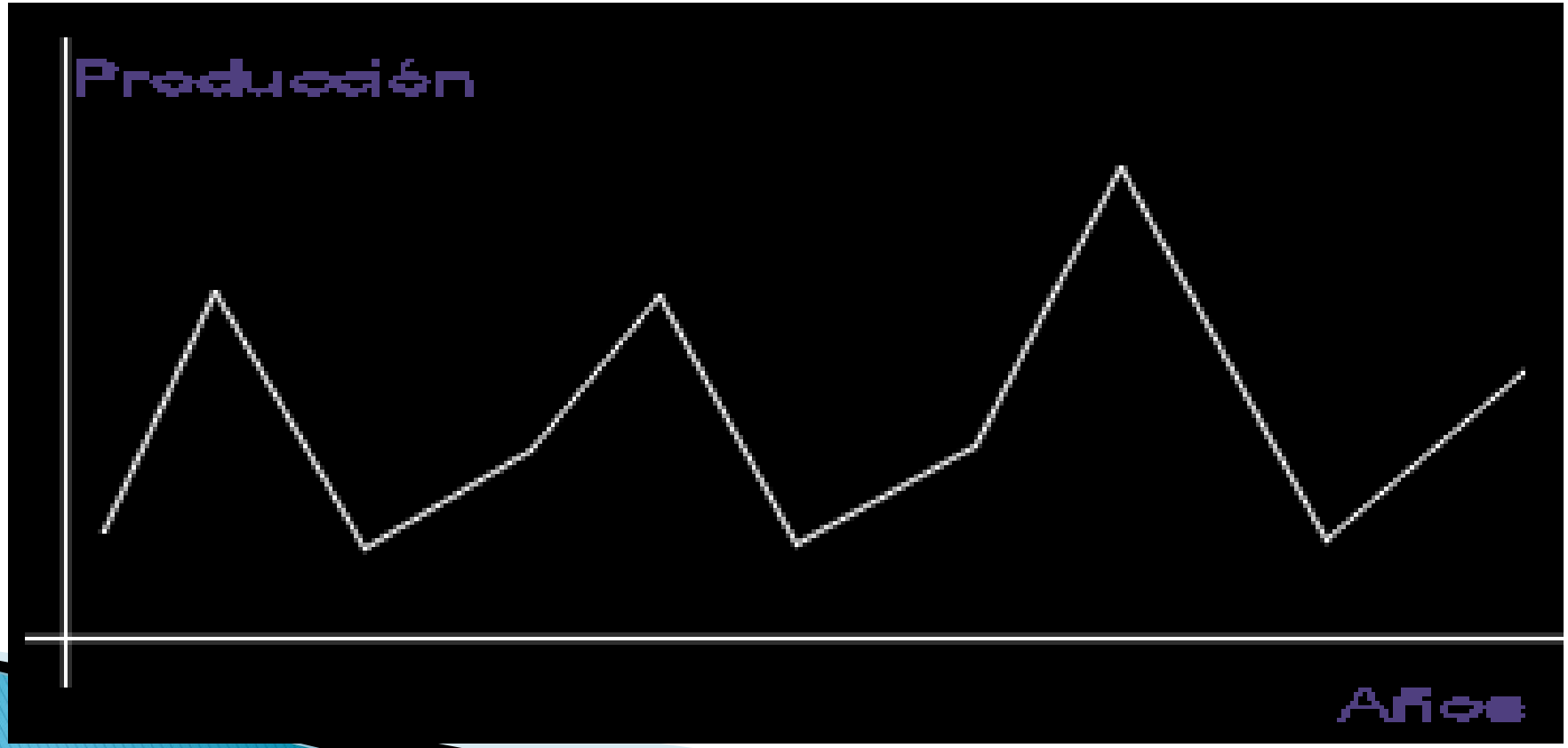
- **Componente Irregular (I).**

Refleja aquellos movimientos esporádicos que ocurren en la serie y que no presentan regularidad alguna, ni en intensidad, ni periodicidad.

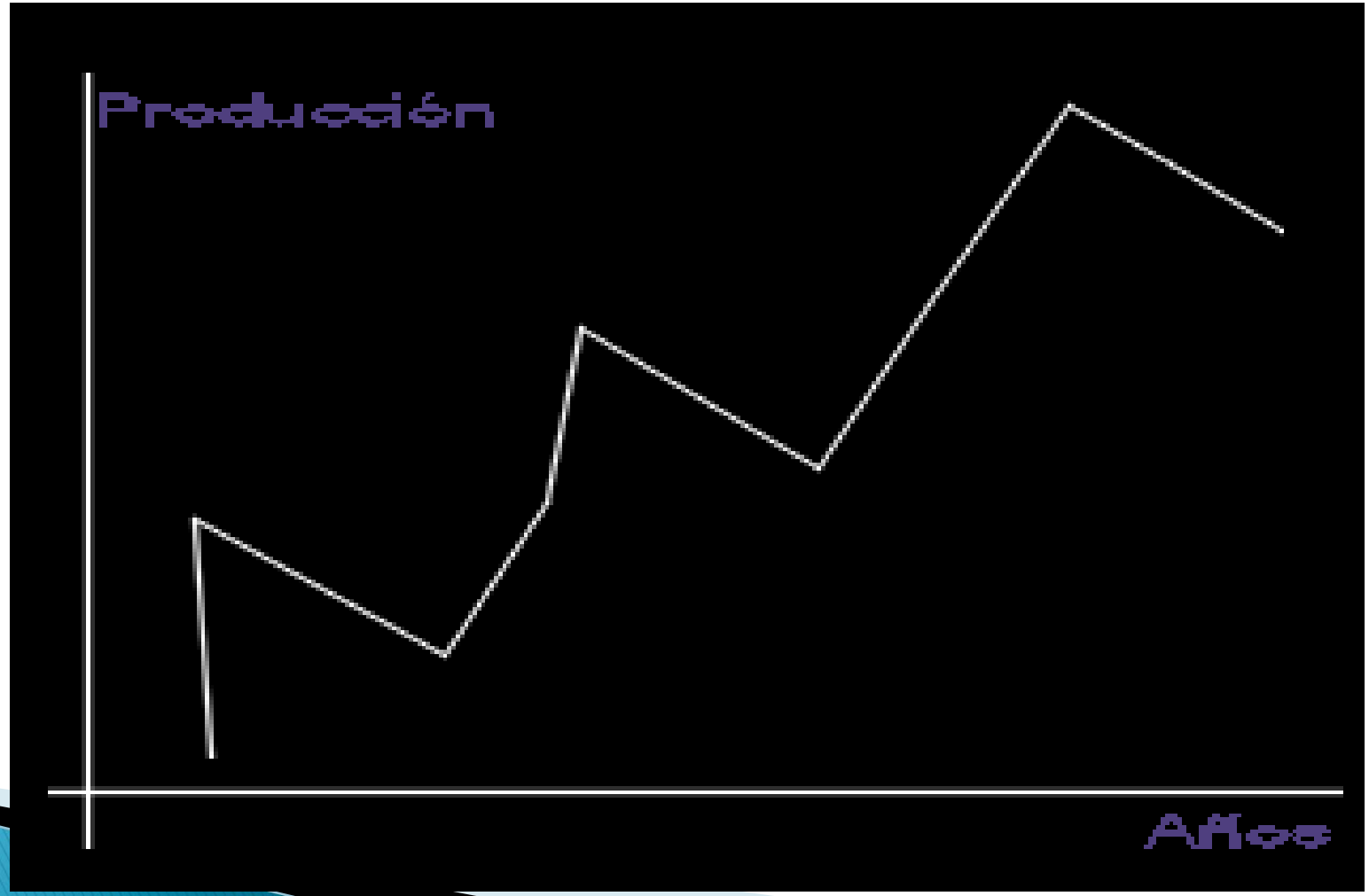
Puede considerarse que son motivados por acontecimientos fortuitos, como catástrofes naturales, incidentes sociales

Clasificación de las Series de Tiempo.

Estacionarias: Los Movimientos oscilan alrededor de una media constante.



Dinámicas: Los Movimientos no oscilan alrededor de una media constante



Tipos de Modelos

Aditivo

$$Y_t = T_t + C_t + S_t + I_t$$

Multiplicativo

$$Y_t = T_t C_t S_t I_t$$

Ajuste de Tendencia
Métodos de Estimación

De los métodos de estimación los más usados son:

1. Medias Móviles
2. Mínimos Cuadrados.

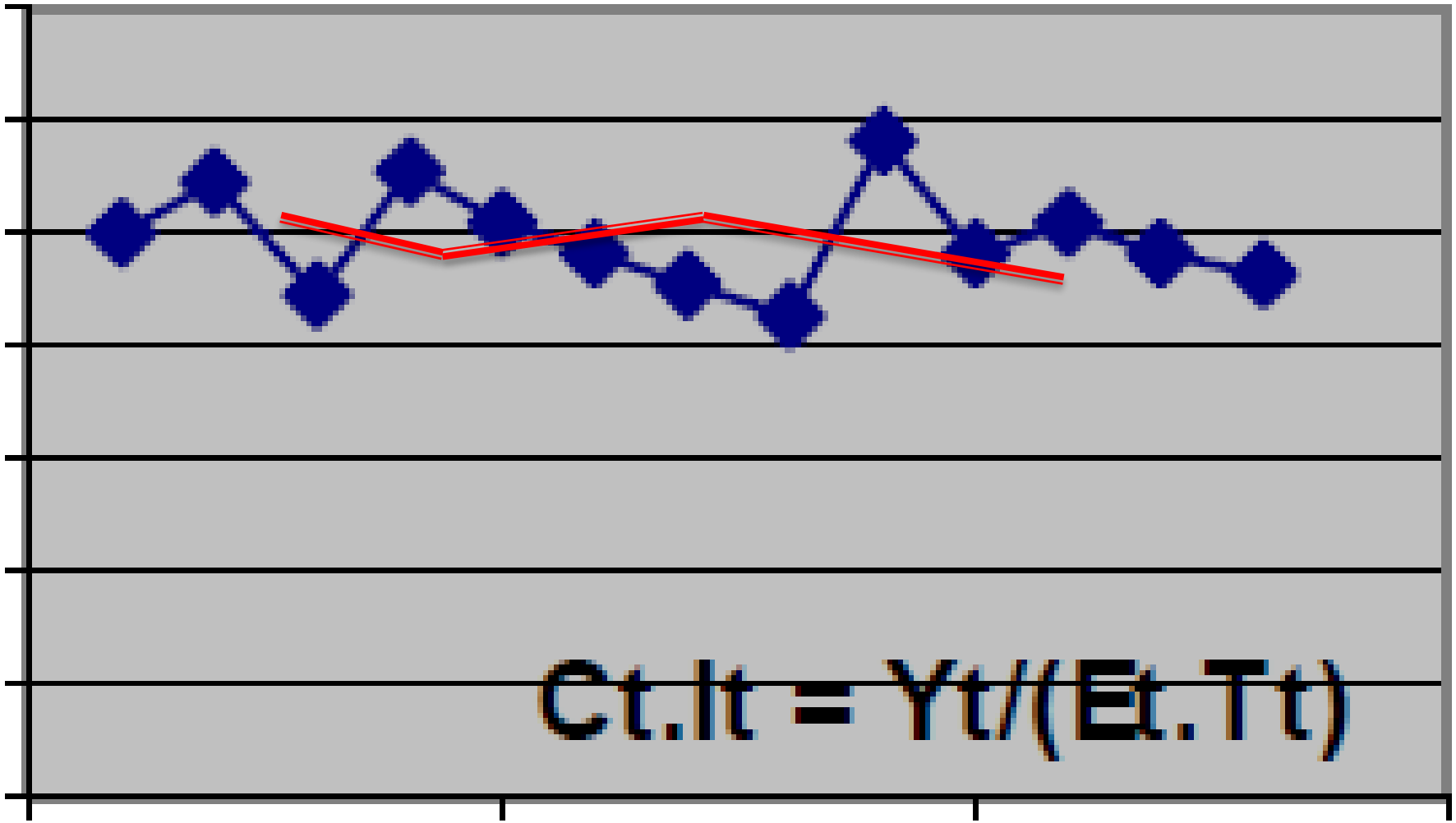
Utilizar uno u otro se basa en el propósito con que se estima la tendencia y la distribución de los datos que componen la serie

¿Cuándo usamos uno u otro?

1. *Medias móviles.* Si los datos de la serie no muestran un comportamiento uniforme, que pudiera ser descrito por una ecuación matemática, o si nuestro interés está referido a la obtención del componente estacional.
2. *Mínimos Cuadrados.* Si estamos interesados en hacer extrapolaciones de la tendencia o pronósticos de la serie y esta sigue un crecimiento uniforme(de tipo lineal o exponencial)

Método de la Medias Móviles.

Este método consiste en suavizar la serie, por lo tanto se utiliza cuando la impresión visual de la tendencia general a largo plazo o los movimientos de la serie, se oscurece por las amplias fluctuaciones en sus componentes cíclicos e irregulares y no se aprecia si hay un efecto descendente o ascendente.


Ct.It = Yt/(Et.Tt)

A diferencia del método de los mínimos cuadrados, el método de promedios móviles para estudiar la tendencia es demasiado subjetivo y depende de la duración del período seleccionado para elaborar los promedios

¿ PROMEDIOS MÓVILES, COMO SE CALCULAN?

PROMEDIOS MÓVILES PARA UN PERÍODO DE DURACIÓN SELECCIONADO: K

- **SERIE DE MEDIAS ARITMÉTICAS CALCULADAS A LO LARGO DEL TIEMPO, CADA MEDIA SE CALCULA PARA UNA SUCESIÓN DE VALORES OBSERVADOS QUE TENGAN ESA DURACIÓN EN PARTICULAR : K**

- **PROMEDIOS MÓVILES DE 5 AÑOS, UNA SERIE DE MEDIAS OBTENIDAS AL PROMEDIAR SUCESIONES CONSECUTIVAS QUE CONTIENEN K: 5 VALORES.**

Ejemplo. Los datos muestran las ganancias en MP de la producción mensual promedio de carbón en el periodo de 1948 hasta 1956

Año	Dato
1948	50,0
1949	36,5
1950	43,0
1951	44,5
1952	38,9
1953	38,1
1954	32,6
1955	38,7
1956	41,7

- **Suponga que se desea calcular promedios móviles de 5 años, para una serie de n = 9 años,**

K = 5, mientras que i = 3, 4, 5, 6, 7

$$\overline{Y}_3^5 = \frac{1}{5}\left(Y_1 + Y_2 + Y_3 + Y_4 + Y_5\right) = 212.9/5$$

$$\overline{Y}_4^5 = \frac{1}{5}\left(Y_2 + Y_3 + Y_4 + Y_5 + Y_6\right) = 201.0/5$$

$$\overline{Y}_5^5 = \frac{1}{5}\left(Y_3 + Y_4 + Y_5 + Y_6 + Y_7\right) = 197.1/5$$

.

.

Año	Dato	Total Móvil 5 años	Promedio Móvil 5 años
1948	50,0		
1949	36,5		
1950	43,0	212,9	42,6
1951	44,5	201,0	40,2
1952	38,9	197,1	39,4
1953	38,1	192,8	39,6
1954	32,6	190,0	38,0
1955	38,7		
1956	41,7		

Método de los Mínimos Cuadrados.

Este procedimiento es el más utilizado en la estimación de la tendencia y se emplea cuando se quiere que se refleje el movimiento gradual a través de una ecuación matemática.

Este método como se vio ya en el tema anterior permite ajustar una línea a una serie de datos, en el caso de la línea recta, donde la fórmula general viene dada por:

$Y_i = b_0 + b_1X + U_i$

Se obtienen los valores calculados para los dos coeficientes donde b_0 es la ordenada en el origen y b_1 la pendiente

$$b_i = \frac{n\sum X_i Y_j - \sum X \sum Y}{n\sum x_i^2 - \left(\sum X\right)^2}$$

$$b_0 = \overline{Y} - b_1 \overline{X}$$

Ejemplo.

Los datos representan las ganancias en MCUC de una empresa de cítrico.

Año	Ganancia I semestre	Ganancia II semestre
2001	15	21
2002	20	23
2003	22	20
2004	20	25

Año	Sem.	x	y	xy	x^2	y^2
01	I	1	15	15	1	225
	II	2	21	42	4	441
02	I	3	20	60	9	400
	II	4	23	92	16	529
03	I	5	22	110	25	484
	II	6	20	120	36	400
04	I	7	20	140	49	400
	II	8	25	150	64	625
	Σ	36	166	779	204	3505

$$b_i = \frac{n\sum X_i Y_j - \sum X \sum Y}{n\sum x_i^2 - \left(\sum X\right)^2}$$

$$b_i = \frac{8\sum(779) - (36)(166)}{8(204) - (36)(36)}$$

$$b_1 = 0{,}76$$

$$b_0 = \overline{Y} - b_1 \overline{X}$$

$$b_0 = \overline{Y} - b_1 \overline{X}$$

$$b_o = (166 / 8) - 0{,}76(36/8)$$

$$b_o = 17.33$$

$$T = 17.33 + 0{,}36\ t$$

$$T = 17.33 + 0{,}36\ t$$

Interpretación de b_1

Cada semestre que pasa las ganancias aumentan en 0.76 MCUC

Análisis de los residuos

- La bondad de ajuste de una Serie de Tiempo se puede realizar a través del R2 pero no es suficiente, se debe verificar el cumplimiento de los supuestos sobre los residuos que son los ya estudiado en el Tema II.

En la próxima clase estudiaremos

- Verificación de si la serie es completamente aleatoria.
- Verificación de si los residuos de un modelo son completamente aleatorios.
- Selección del mejor modelo.
- Realizar estimaciones.

CONCLUSIONES

- Componentes de la Serie de Tiempo.
- Métodos para el Cálculo de la Tendencia

Clase Práctica No. 6

En un estudio realizado para analizar el impacto del cambio climático sobre el incremento de la temperatura en la provincia de Artemisa, desde 2000 hasta 2020, utilizando la opción de series temporales por el STATGRAF arrojó las siguientes salidas:

Estimated Autocorrelations for CP

	Autocorrelatión	Stand. Error	Lower 95% Prob. Límit	Upper 95% Prob. Límit
1	0,59163	0.22941	-0.44964	0.44964
2	0.32846	0.29912	-0.58628	0.58628
3	0.20771	0.31754	-0:62237	0:62237
4	0,10443	0.32461	-0.63623	0.63623
5	0.06678	0.32637	-0.63963	0.63963
6	0.08724	0.32709	-0.64110	0.64110

- Test. for Randomnes of CP
- Box-Price Test.
- Test basaed on first 6 autocorrelations.
- Test Statistic = 9.95687.
- P. Value = 0,018648

Comparation of models:
Data CP

- Modelos:
- (A) Randow walk
- (B) Constant mean = 566.5
- (C) Linear Trend = 1.03 + 1.0901(t)
- (D) Simple moving average of 5 terms.
- (E) Simple exponetial with $\alpha = 0{,}9999$

Modelo	RMSE	RUNS	RUNM	AUTO	MEAM
(A)	122,851	***	OK	OK	OK
(B)	244.398	***	***	OK	***
(C)	120.855	OK	OK	OK	OK
(D)	212.926	OK	OK	OK	***
(E)	122.012	OK	OK	OK	OK

RESPONDA LAS SIGUIENTES PREGUNTAS

Atendiendo a estos resultados responda:

a) ¿Puede la serie considerarse completamente aleatoria?

b) ¿Cuál de los modelos resultó ser el mejor? Justifique su respuesta.

c) Estime la temperatura para el año 2028.

Atendiendo a estos resultados responda:

d) ¿Pueden los residuos del mejor modelo considerarse completamente aleatorios?

e) Interprete el coeficiente de regresión en el contexto dado.

f) Cuales son los componentes de una serie de tiempo, describir cada uno de ellos.

FIN DEL CURSO DE ESTADÍSTICA II

Prof: M.S. Ing. Rodolfo Pérez León
Email: rperezleon@uart.edu.cu