“Introducción a los Modelos de Pronósticos”

Dra. Fernanda Villarreal

Universidad Nacional del Sur- Departamento de Matemática

Septiembre 2016 - fvillarreal@uns.edu.ar

Introducción

- Planeación del futuro, un aspecto relevante en cualquier organización.

- El éxito a largo plazo depende de cuán bien la gerencia anticipa el futuro y elabora las estrategias apropiadas.

- El buen juicio, la intuición y tener conciencia del estado de la economía pueden dar a un gerente una idea aproximada o “intuición” de lo que es probable que suceda en el futuro.

- Sin embargo, es difícil convertir esta intuición en un número que pueda usarse, como el volumen de ventas del siguiente trimestre o el costo de la materia prima por unidad para el año próximo.

Pronóstico

“Es una estimación cuantitativa o cualitativa de uno o varios factores (variables) que conforman un evento futuro, con base en información actual o del pasado”.

- La estimación de pronósticos del volumen de ventas trimestrales para un producto en particular durante el año próximo afectará los programas de producción, los planes de compra de materias primas, las políticas de inventarios y las cuotas de ventas.

- En consecuencia, los malos pronósticos pueden dar como resultado un incremento en los costos de la empresa. ¿Cómo debemos proceder para proporcionar los pronósticos trimestrales del volumen de ventas?

- Revisar los datos históricos, con frecuencia ayuda a comprender mejor el patrón de las ventas pasadas, lo que conduce a mejores predicciones de las ventas futuras del producto.

- Los datos históricos de ventas forman una serie de tiempo.

- Una serie de tiempo es un conjunto de observaciones de una variable medida en puntos sucesivos en el tiempo o a lo largo de periodos sucesivos.

- En este curso se presentan varios procedimientos para analizar las series de tiempo.

- El objetivo de estos análisis es proporcionar buenos pronósticos o predicciones de los valores futuros de la serie de tiempo.

Métodos de elaboración de pronósticos

- Los métodos de elaboración de pronósticos se clasifican como cuantitativos o cualitativos.

- Los métodos cuantitativos se utilizan cuando:
  - se dispone de información pasada sobre la variable que se pronosticará
  - la información puede cuantificarse
  - es razonable suponer que el patrón del pasado seguirá ocurriendo en el futuro. En estos casos puede elaborarse un pronóstico con un método de series de tiempo o un método causal.

- Si los datos históricos se restringen a valores pasados de la variable que tratamos de pronosticar, el procedimiento de elaboración de pronósticos se llama método de serie de tiempo.

- El objetivo de los métodos de serie de tiempo es descubrir un patrón en los datos históricos y luego extrapolarlo hacia el futuro; el pronóstico se basa sólo en valores pasados de la variable que tratamos de pronosticar o en errores pasados.

- En este curso se explican tres métodos de series de tiempo: **suavización** (promedios móviles, promedios móviles ponderados y suavización exponencial), **proyección de tendencias y proyección de tendencias ajustada por influencia estacional.**

- Los métodos de elaboración de pronósticos causal se basan en el supuesto de que la variable que tratamos de pronosticar exhibe una relación de causa y efecto con una o más variables.

- En este curso se presenta el uso del **análisis de regresión** como un método de elaboración de pronósticos causal. Por ejemplo, los gastos de publicidad influyen en el volumen de ventas de muchos productos, de manera que el análisis de regresión puede utilizarse para desarrollar una ecuación que muestre cómo se relacionan estas dos variables.

- Utilizar un método de series de tiempo para elaborar el pronóstico en este ejemplo, implica que no se considerarían los gastos de publicidad; es decir, un método de serie de tiempo basaría el pronóstico sólo en las ventas pasadas.

- Los métodos cualitativos por lo general involucran el uso del juicio experto para elaborar pronósticos. Una ventaja de los procedimientos cualitativos es que pueden aplicarse cuando la información sobre la variable que se está pronosticando no puede cuantificarse o son escasos.

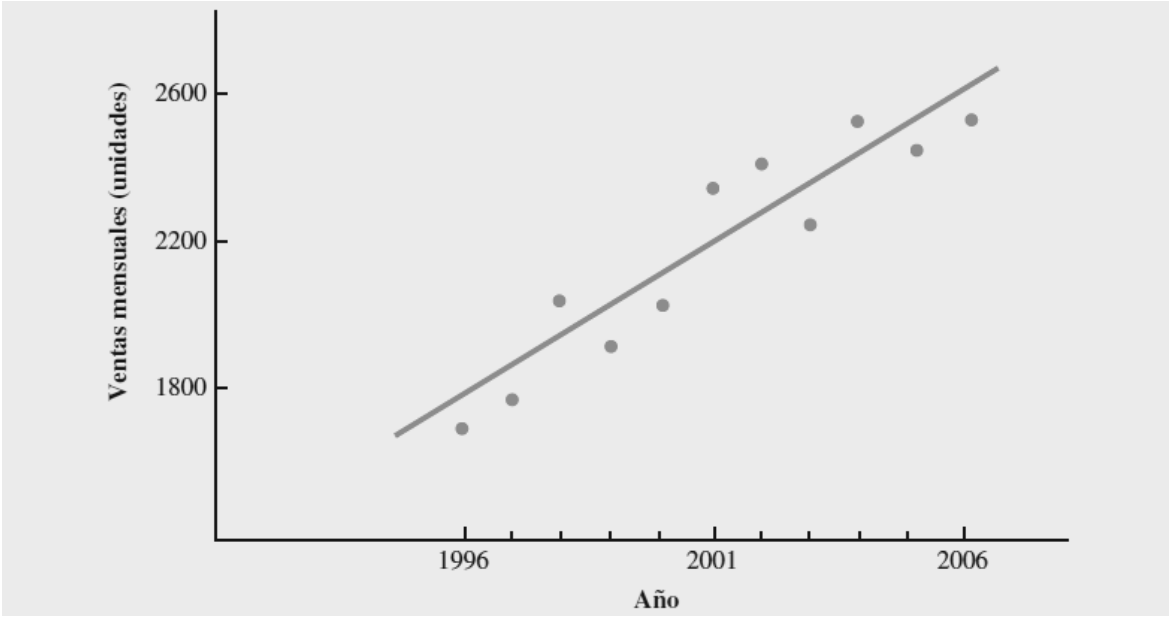
- Método Delphi
- Juicio experto
- Redacción de escenarios
- Enfoques intuitivos

PATRONES O COMPONENTES DE UNA SERIE DE TIEMPO

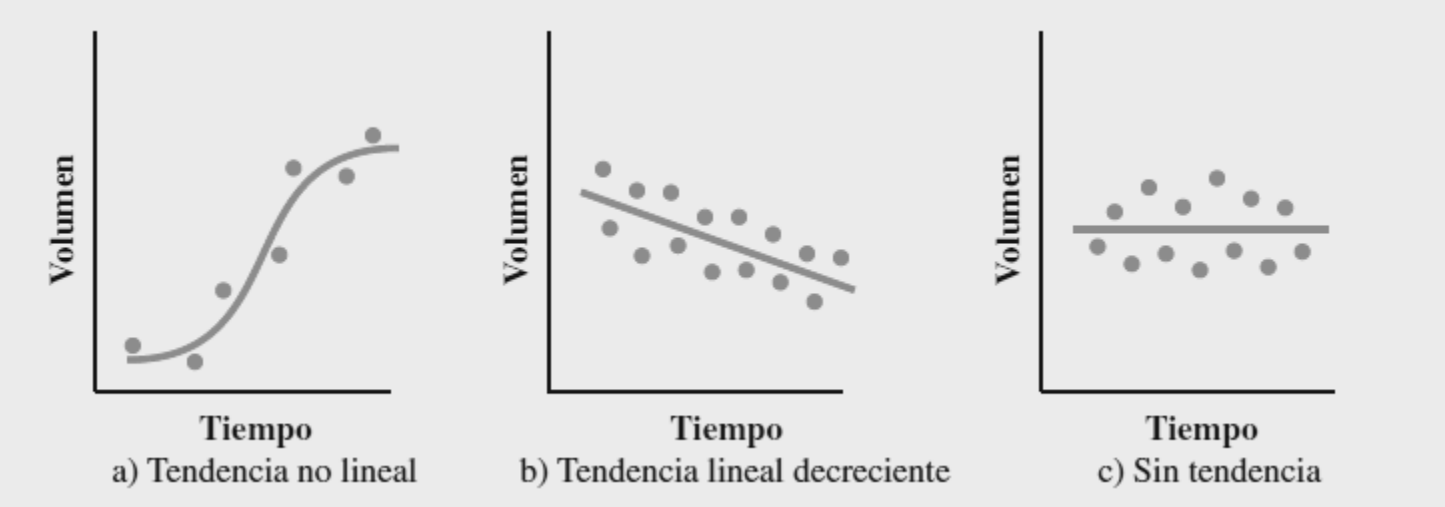
- El patrón o comportamiento de los datos en una serie de tiempo tiene varios componentes. El supuesto usual es que cuatro componentes separados: tendencia, cíclico, estacional e irregular, se combinen para proporcionar valores específicos de la serie de tiempo.

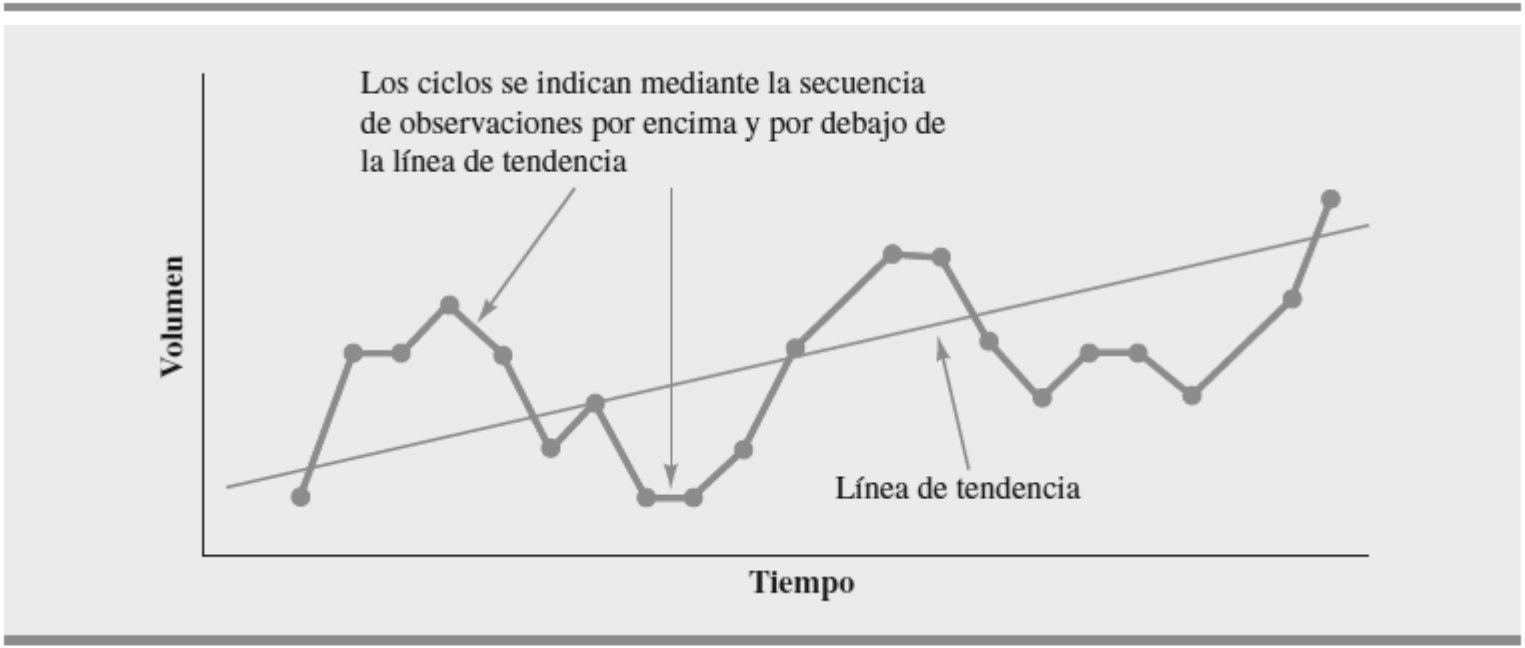
- TENDENCIA: componente de muy largo plazo
- CICLICO: componente de largo plazo
- ESTACIONAL: componente de corto plazo
- IRREGULAR: componente de muy corto plazo

- En el análisis de las series de tiempo, las mediciones pueden hacerse cada hora, diario, a la semana, cada mes, anualmente o en cualquier otro intervalo regular de tiempo. Aunque los datos de las series de tiempo suelen mostrar fluctuaciones aleatorias, las series de tiempo también muestran un desplazamiento o movimiento gradual hacia valores relativamente altos o bajos a través de un lapso largo. A este desplazamiento gradual de la serie de tiempo se le conoce como **la tendencia de la serie de tiempo.**
- Este desplazamiento o tendencia suele deberse a factores de largo plazo como variaciones en las características demográficas de la población, en la tecnología o en las preferencias del público.



Otros patrones de tendencia posibles



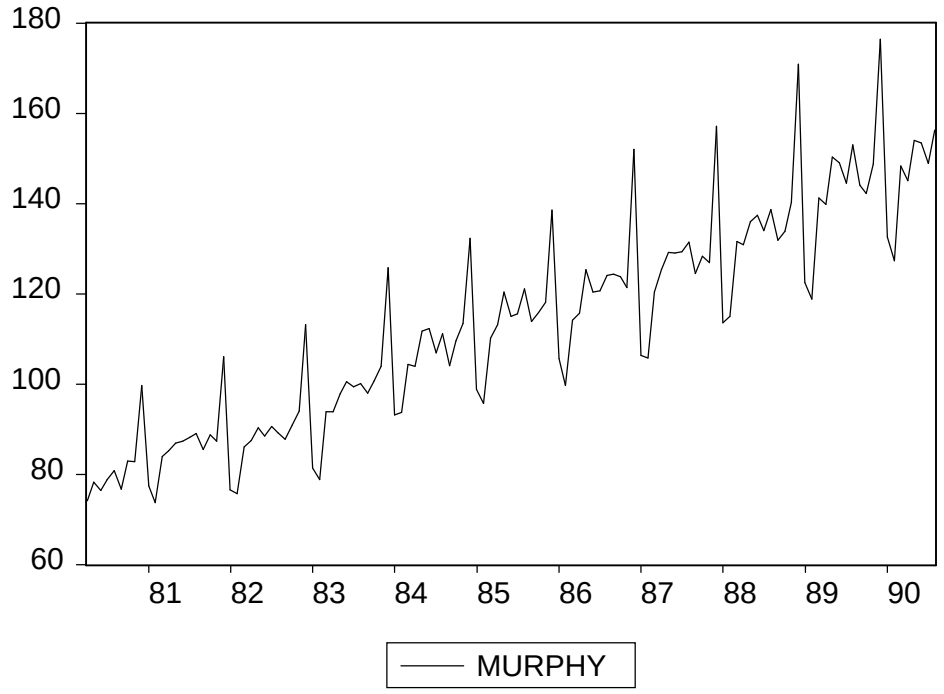


Aunque una serie de tiempo puede tener una tendencia a través de lapsos largos, no todos los valores futuros de la serie de tiempo caerán exactamente sobre la línea de tendencia. Las series de tiempo suelen mostrar secuencias de puntos que caen de manera alternante arriba y abajo de la línea de tendencia. Toda sucesión recurrente de puntos que caiga abajo y arriba de la línea de tendencia y que dure más de un año puede atribuirse al **componente cíclico de la serie de tiempo.** En la figura las observaciones son anuales.

- Patrón de cambio que se repite año con año en el mismo número de períodos.

FUERZAS QUE AFECTAN Y EXPLICAN ESTACIONALIDAD:

- períodos escolares
- períodos vacacionales
- productos de estación
- estaciones del año



Componente irregular

Mide la variabilidad de una serie cuando los demás componentes se han eliminado o no existen.

FUERZAS QUE AFECTAN Y EXPLICAN ALEATORIEDAD

- cambios climáticos
- desastres naturales
- huelgas
- hechos fortuitos

Este componente representa la variabilidad aleatoria en las series de tiempo y es resultado de factores a corto plazo, imprevistos y no recurrentes que afectan a la serie de tiempo. Como este componente representa la variabilidad aleatoria en las series de tiempo, es impredecible; no podemos intentar predecir su impacto en las series de tiempo.

Métodos de suavización

- En esta primera parte se presentan tres métodos de elaboración de pronósticos: promedios móviles, promedios móviles ponderados y suavización exponencial.

- El objetivo de cada uno de estos métodos es “suavizar” las fluctuaciones aleatorias causadas por el componente irregular de las series de tiempo, por lo que se conocen como métodos de suavización.

- Este tipo de métodos es apropiado para una serie de tiempo estable, es decir, una que no exhibe efectos significativos de tendencia, cíclicos o estacionales.

- Los métodos de suavización son fáciles de usar y por lo general proporcionan un alto nivel de precisión para pronósticos de corto alcance como un pronóstico para el siguiente periodo.

- Uno de los métodos, la suavización exponencial, tiene requisitos de datos mínimos y por tanto es un buen método para usar cuando se requieren pronósticos para cantidades grandes de artículos.

Promedios móviles (simples de orden k)

$$F_{t+1} = \frac{Y_t + Y_{t-1} + \ldots\ldots + Y_{t-k+1}}{k}$$

El método de los promedios móviles utiliza el promedio de los k valores de datos más recientes en la serie de tiempo como el pronóstico para el siguiente periodo.

El término móvil indica que, mientras se dispone de una nueva observación para la serie de tiempo, reemplaza a la observación más antigua de la ecuación anterior y se calcula un promedio nuevo. Como resultado, el promedio cambiará, o se moverá, conforme surjan nuevas observaciones.

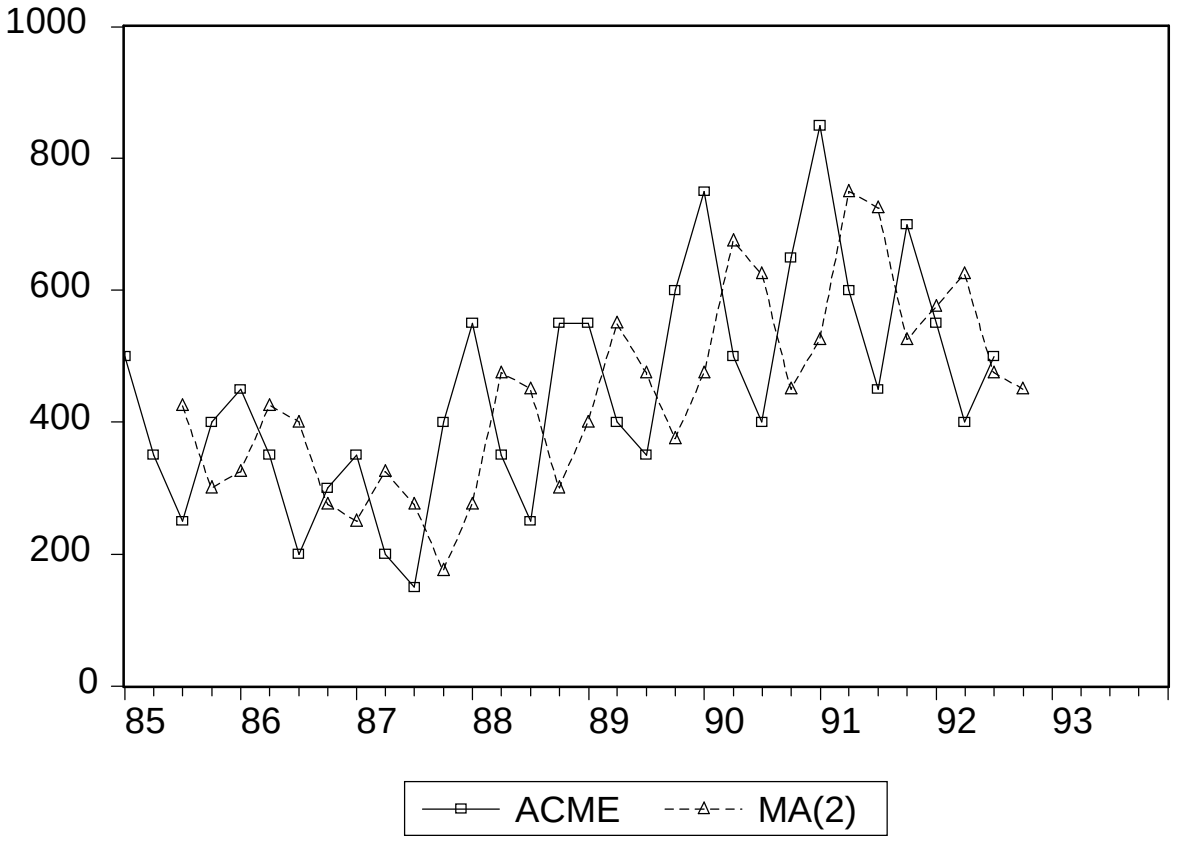
Y_t : observación en el período t F_t: pronóstico para el período t

Promedios móviles (simples de orden 3)

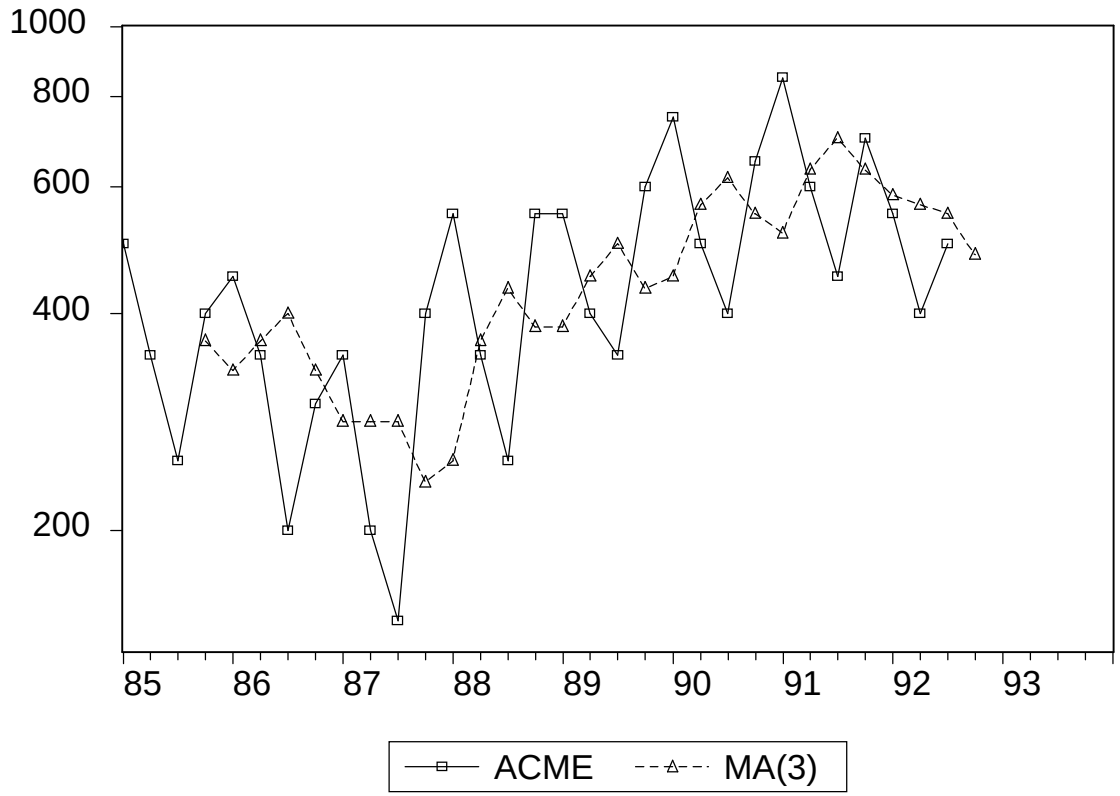
$$F_{t+1} = \frac{Y_t + Y_{t-1} + Y_{t-2}}{3}$$

- se promedian solo las últimas observaciones
- el orden se determina a priori
- un orden grande elimina los picos (suaviza)
- un orden pequeño permite seguir muy de cerca los cambios de corto plazo

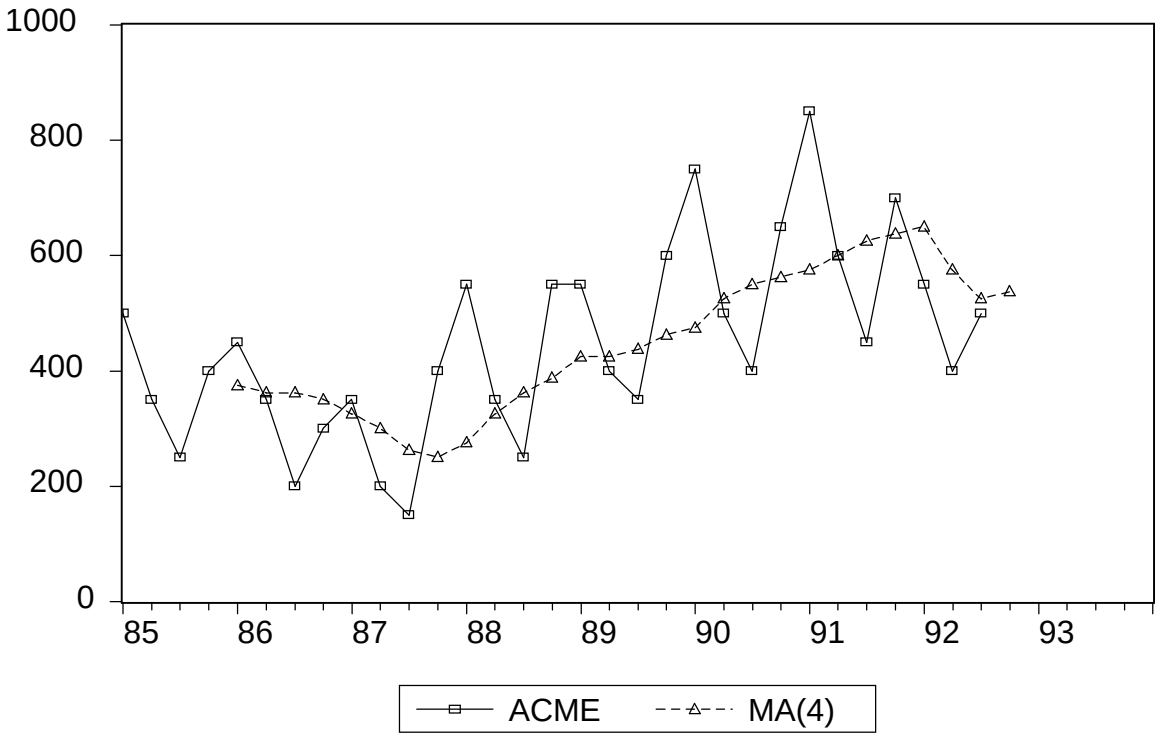
Promedios móviles (simples de orden 2)



PROMEDIO MÓVIL DE ORDEN 3

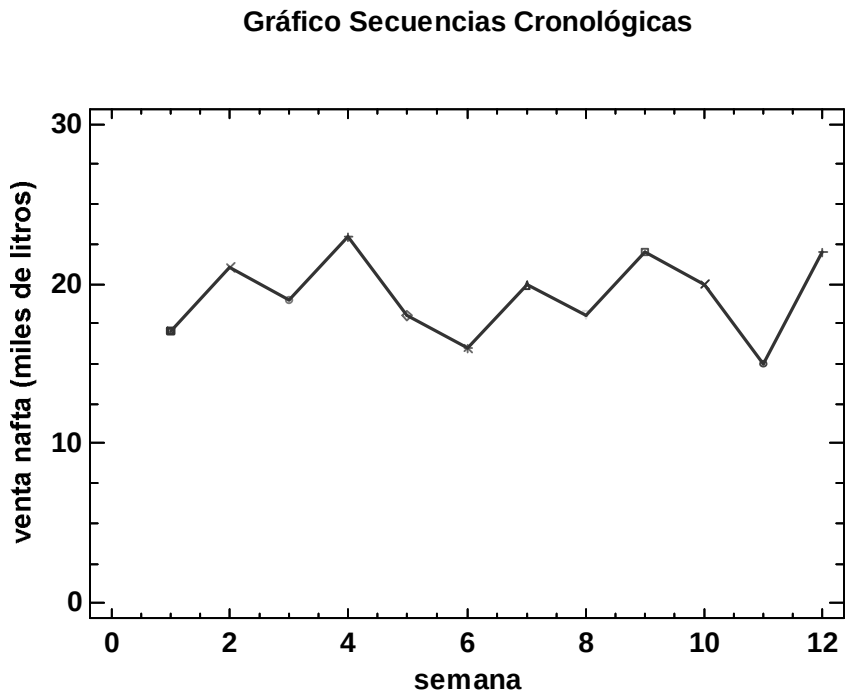


Promedios móviles (simples de orden 4)



Ejemplo

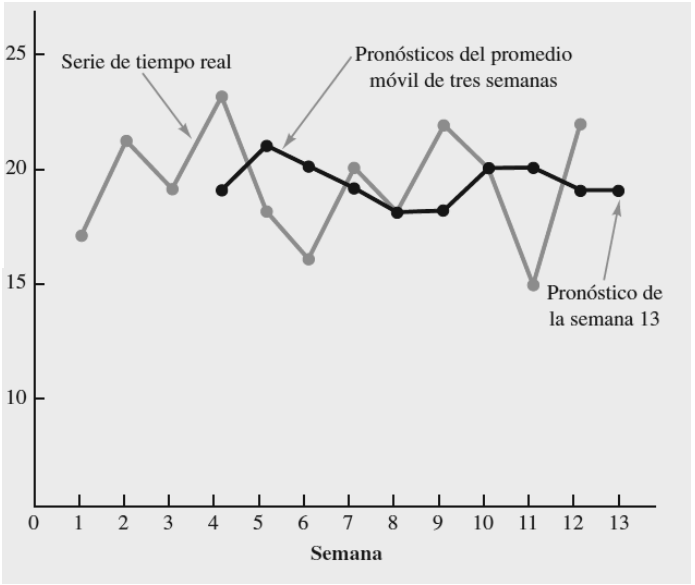
Litros de nafta vendidos por semana (en miles)



Ejemplo

Semana	Valor de la serie de tiempo	Pronóstico del promedio móvil	Error de pronóstico	Error de pronóstico al cuadrado
1	17			
2	21			
3	19			
4	23	19	4	16
5	18	21	−3	9
6	16	20	−4	16
7	20	19	1	1
8	18	18	0	0
9	22	18	4	16
10	20	20	0	0
11	15	20	−5	25
12	22	19	3	9
			Totales 0	92

pronóstico para la semana 13 es 19.



$e_t = Y_t - F_t$: residuo (error de pronóstico) en el período t

Precisión del pronóstico. Una consideración importante en la selección de un método de elaboración de pronósticos es la precisión del pronóstico. Desde luego, queremos pronosticar que los errores sean menores. Las últimas dos columnas de la tabla que contienen los errores de pronóstico y los errores de pronóstico al cuadrado, se pueden utilizar para desarrollar medidas de la precisión del pronóstico.

Medidas de error

- Error Medio (Me) : $ME = \frac{\Sigma e_i}{n}$ identifica sesgo
- Error Medio Absoluto: $MAD = \frac{\Sigma |e_i|}{n}$ distancia promedio
- Error Medio Cuadrático (Mse): penaliza errores grandes

$$MSE = \frac{\Sigma (e_i)^2}{n}$$

- Error Medio Absoluto Porcentual: proporción del error

$$MAPE = \frac{\Sigma |e_i / y|}{n}$$

MAPE proporciona una indicación de cuan grande son los errores de pronostico en comparación con los valores reales de la serie.

Promedios móviles ponderados

- En el método de promedios móviles, cada observación en el cálculo recibe el mismo peso.

- Una variación, conocida como promedios móviles ponderados, consiste en seleccionar diferentes pesos para cada valor de datos y luego calcular un promedio ponderado de los k valores de datos más recientes como el pronóstico.

- En la mayoría de los casos la observación más reciente recibe el mayor peso, y el peso disminuye para los valores de datos más antiguos. Por ejemplo, para la serie de tiempo de las venta de nafta semanal el cálculo de un promedio móvil ponderado de tres semanas, donde la observación más reciente recibe un peso del triple del peso dado a la observación más antigua y la siguiente observación más antigua recibe un peso del doble que la observación más antigua.
- Para la semana 4 el cálculo es:

$$3/6*19+2/6*21+1/6*17=19.33$$

En general, si creemos que el pasado reciente es un mejor pronosticador del futuro que el pasado distante, los pesos más grandes deben darse a las observaciones más recientes.

Suavización exponencial

$$F_{t+1} = \alpha Y_t + (1- \alpha) F_t \qquad 0 \leq \alpha \leq 1$$

La suavización exponencial utiliza un promedio ponderado de valores de series de tiempo pasadas como pronóstico.

La formula muestra que el pronóstico para el periodo t+1 es un promedio ponderado del valor real en el periodo t y el pronóstico para el periodo t.

Es un caso especial del método de promedios móviles ponderados en el cual seleccionamos sólo un peso, el peso para la observación más reciente.

Los pesos para los demás valores se calculan de forma automática y se vuelven cada vez más pequeños a medida que las observaciones se alejan en el pasado.
Podemos demostrar que el pronóstico de la suavización exponencial para cualquier periodo también es un promedio ponderado de todos los valores reales previos.

Por ejemplo para una serie de tiempo que consta de tres periodos de datos: Y_1, Y_2 y Y_3. Comenzamos $F_1=Y_1$

$F_2=\alpha Y_1 + (1-\alpha)F_1$

$= \alpha Y_1 + (1-\alpha)Y_1$

$= Y_1$

Por lo tanto, el pronóstico de suavización exponencial para el periodo dos es igual al valor real de la serie de tiempo en el periodo 1.

Para el periodo 3 el pronóstico es:

$$F_3=\alpha Y_2 + (1-\alpha)\, F_2= \alpha Y_2 + (1-\alpha)Y_1$$

Por ultimo al sustituir esta expresión para F_3 en la expresión para F_4, se obtiene:

$$F_4=\alpha Y_3 + (1-\alpha)F_3= \alpha Y_3 + (1-\alpha)\,[\alpha Y_2 + (1-\alpha)Y_1]$$

$$= \alpha Y_3 + \alpha\,(1-\alpha)\, Y_2 + (1-\alpha)^2 Y_1$$

Por consiguiente F_4 es un promedio ponderado de los primeros tres valores de la serie de tiempo.

Constante suavización α=0.2

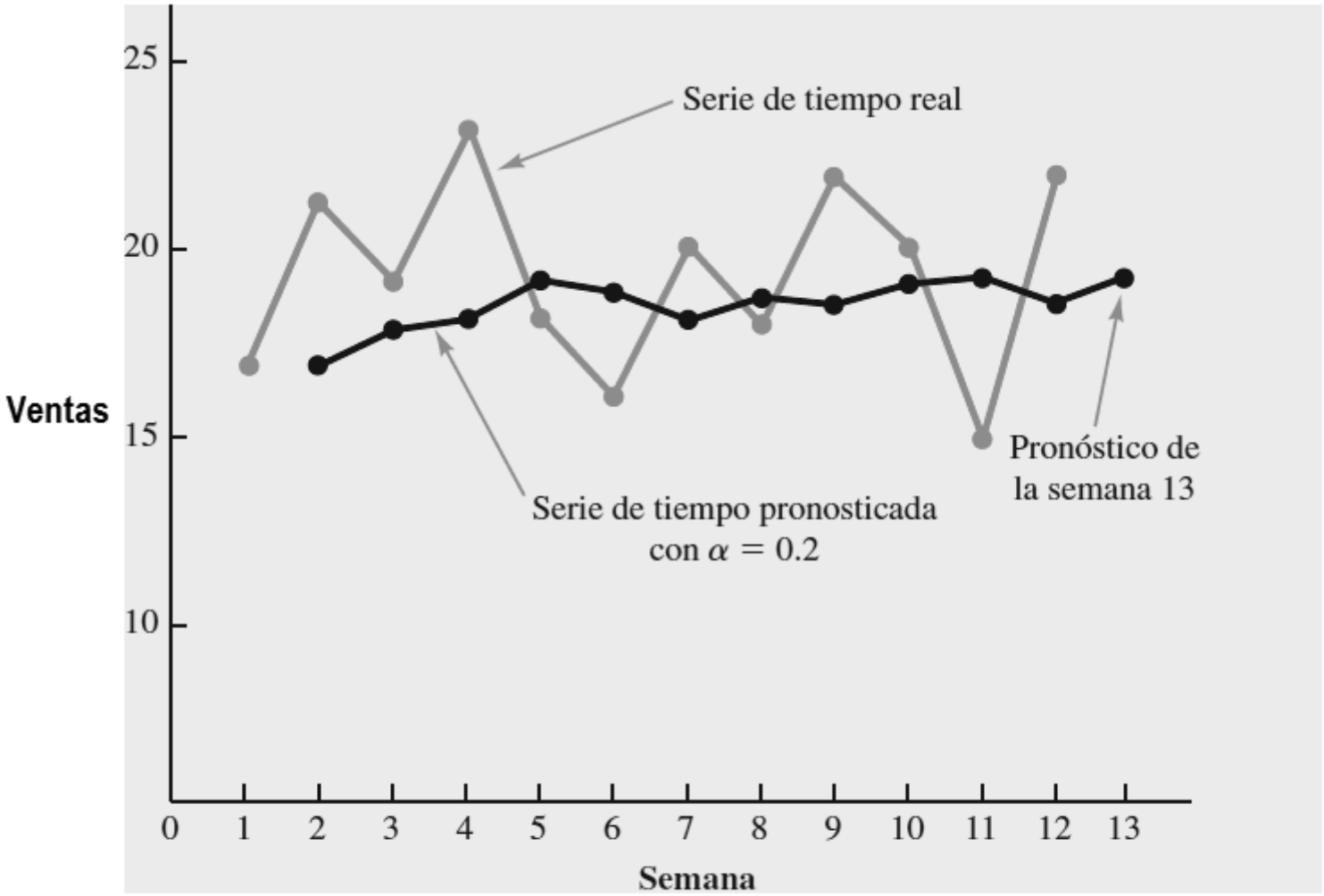
Semana (t)	Valor de la serie de tiempo (Y_t)	Pronóstico de suavización exponencial (F_t)	Error de pronóstico ($Y_t - F_t$)
1	17		
2	21	17.00	4.00
3	19	17.80	1.20
4	23	18.04	4.96
5	18	19.03	−1.03
6	16	18.83	−2.83
7	20	18.26	1.74
8	18	18.61	−0.61
9	22	18.49	3.51
10	20	19.19	0.81
11	15	19.35	−4.35
12	22	18.48	3.52

$$F_{13} = 0.2Y_{12} + 0.8F_{12} = 0.2(22) + 0.8(18.48) = 19.18$$

¿Qué valor de α ?

- Si la variabilidad aleatoria de la serie de tiempo es considerable, es preferible un valor pequeño para la constante de suavización. La razón de esta elección es que, dado que gran parte del error de pronóstico se debe a la variabilidad aleatoria, no queremos reaccionar de forma exagerada y ajustar los pronósticos demasiado rápido. Para una serie de tiempo con relativamente poca variabilidad, los valores más grandes de la constante de suavización tienen la ventaja de ajustar rápidamente los pronósticos cuando ocurren errores de pronóstico y por ende permiten que el pronóstico reaccione más rápido a las condiciones cambiantes.
- Elegimos el valor de α que minimiza el error de pronostico.

Observar como los pronósticos “suavizan” las fluctuaciones irregulares de la serie de tiempo.

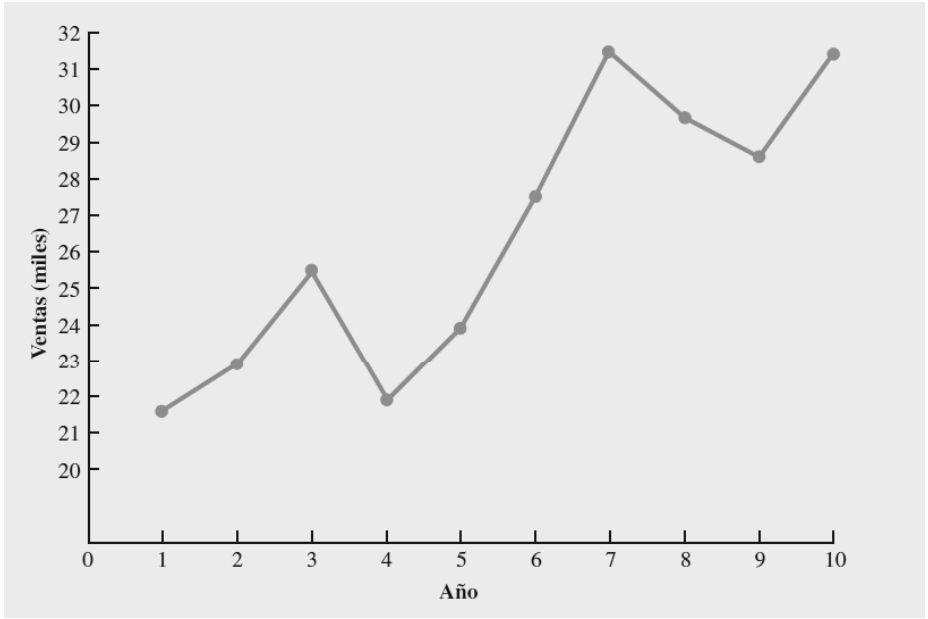


Proyección de la tendencia

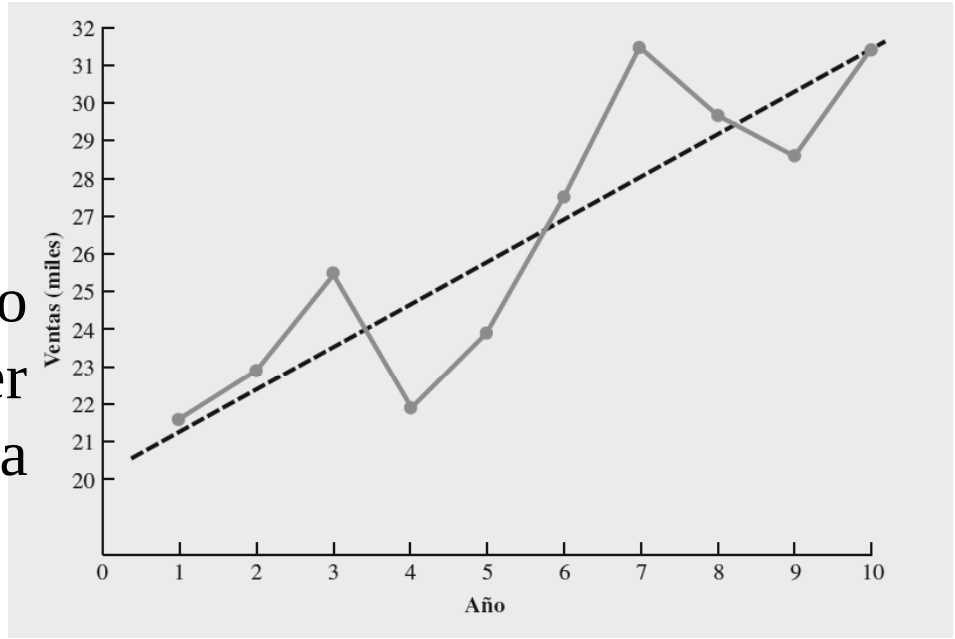
- En este punto se muestra cómo pronosticar los valores de una serie de tiempo que exhibe una tendencia lineal a largo plazo. El tipo de series de tiempo para las cuales el método de proyección de tendencias es aplicable, muestra un incremento o disminución constante en el tiempo. Debido a que este tipo de serie de tiempo no es estable, los métodos de suavización descritos en la sección anterior no son aplicables.

Ejemplo

Año (t)	Ventas (miles) (Y_t)
1	21.6
2	22.9
3	25.5
4	21.9
5	23.9
6	27.5
7	31.5
8	29.7
9	28.6
10	31.4



La serie de tiempo para el número de bicicletas vendidas parece tener un incremento general o una tendencia ascendente.



Para una tendencia lineal, el volumen de ventas estimado expresado como una función del tiempo.

$$T_t = b_0 + b_1 t$$

T_t = valor de tendencia para las ventas de bicicletas en el periodo t

Las ecuaciones para calcular b_1 y b_0 son

$$b_1 = \frac{\sum tY_t - (\sum t \sum Y_t)/n}{\sum t^2 - (\sum t)^2/n}$$

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

Ecuación para el componente de tendencia lineal para las series de tiempo de ventas de bicicletas.

$$T_t = 20.4 + 1.1t$$

La pendiente de 1.1 en la ecuación de tendencia indica que durante los 10 años pasados la empresa ha experimentado un crecimiento medio en las ventas de alrededor de 1100 unidades por año.

La proyección de tendencia del año siguiente,

$$T_{11} = 20.4 + 1.1 * 11 = 32.5$$

Componentes de tendencia y estacional

¿cómo pronosticar los valores de una serie de tiempo que tiene tanto un componente de tendencia como uno estacional?

La eliminación del efecto estacional de una serie de tiempo se conoce como desestacionalización de la serie de tiempo. Después de hacerlo, las comparaciones periodo a periodo son más significativas y pueden ayudar a identificar si existe una tendencia.

El enfoque que seguimos en este punto es apropiado en situaciones cuando sólo están presentes los efectos estacionales o en situaciones en que se dan tanto el componente estacional como el de tendencia.

El primer paso es calcular los índices estacionales y utilizarlos para desestacionalizar los datos.

Luego, si es evidente una tendencia en los datos desestacionalizados, utilizamos el análisis de regresión sobre los datos desestacionalizados para estimar la tendencia.

Modelo multiplicativo

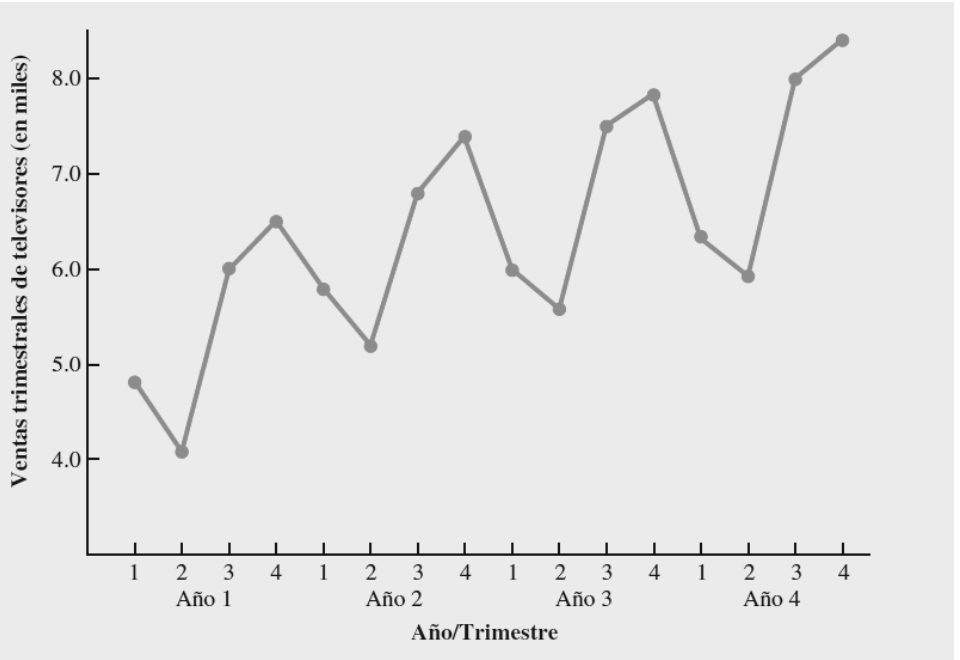
- Además de un componente de tendencia *T* y un componente estacional *S*, asumimos que la serie de tiempo también tiene un componente irregular *I*. El componente irregular representa los efectos aleatorios de la serie de tiempo que no pueden explicarse por medio de los componentes de tendencia y estacional.

- Con T_t , S_t e I_t para identificar los componentes de tendencia, estacional e irregular en el tiempo *t*, suponemos que el valor de la serie de tiempo real, denotado por Y_t, puede describirse por el **modelo multiplicativo de series de tiempo**.

$$Y_t = T_t \times S_t \times I_t$$

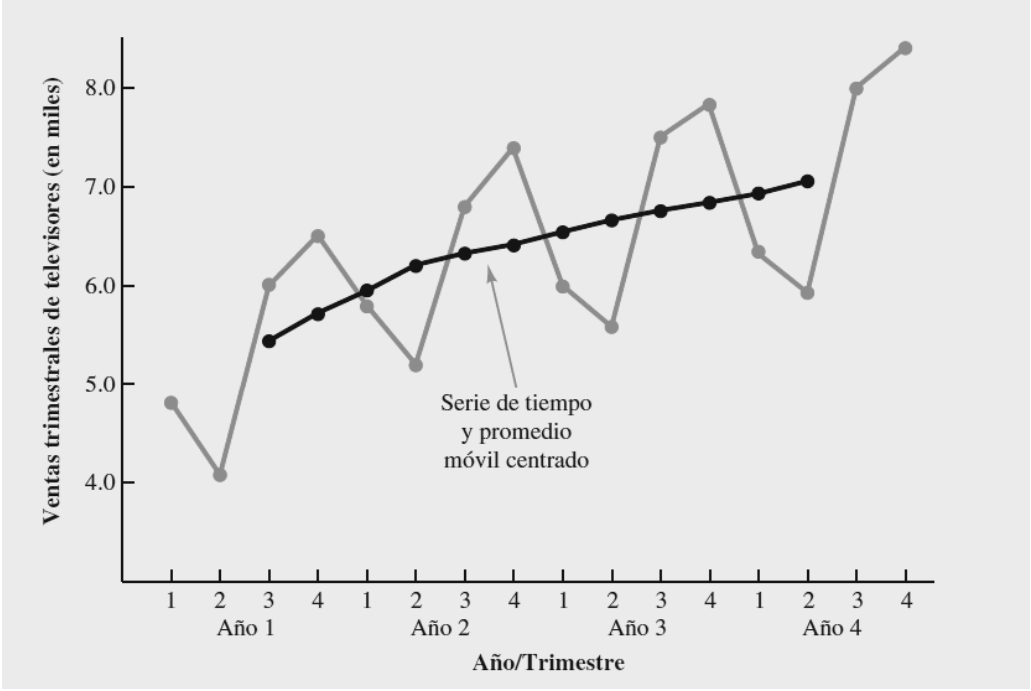
T_t es la tendencia medida en unidades del elemento que se pronostica. Sin embargo, los componentes S_t e I_t se miden en términos relativos, con valores por encima de 1.00, lo que indica efectos por encima de la tendencia, y valores por debajo de 1.00 que denotan efectos por debajo de la tendencia.

Año	Trimestre	Ventas (en miles)
1	1	4.8
	2	4.1
	3	6.0
	4	6.5
2	1	5.8
	2	5.2
	3	6.8
	4	7.4
3	1	6.0
	2	5.6
	3	7.5
	4	7.8
4	1	6.3
	2	5.9
	3	8.0
	4	8.4



Las ventas son menores en el segundo trimestre de cada año, seguidas por los niveles de ventas más altos en los trimestres 3 y 4. Por tanto, concluimos que existe un patrón estacional para las ventas de televisores.

- Comenzamos el procedimiento de cálculo utilizado para identificar la influencia estacional de cada trimestre.
- Con el fin de medir la variación estacional, es común usar el “método de razón de promedio móvil”. Esta técnica proporciona un índice que describe el grado de variación estacional.
- Los valores del promedio móvil centrado tienden a “suavizar” las fluctuaciones tanto estacional como irregular en la serie de tiempo. Los valores del promedio móvil calculados para cuatro trimestres de datos no incluyen las fluctuaciones debidas a influencias estacionales porque el efecto estacional se ha promediado. Cada punto en el promedio móvil centrado representa cuál sería el valor de la serie de tiempo sin influencias estacionales o irregulares.



Año	trimestre	Ventas(miles)	Total móvil (1)	promedio móvil	promedio móvil centrado	valores estacionales-irregulares	índice estacional	
1	1	4,8						
	2	4,1						
			21,4	5,35				
	3	6			5,475	1,096		
			22,4	5,6				
	4	6,5			5,7375	1,133		
			23,5	5,875				
	1	5,8			5,975	0,971	0,93	1 trimestre
2			24,3	6,075				
	2	5,2			6,1875	0,840	0,84	2 trimestre
			25,2	6,3				
	3	6,8			6,325	1,075	1,09	3 trimeste
			25,4	6,35				
	4	7,4			6,4	1,156	1,14	4 trimestre
			25,8	6,45				
3	1	6			6,5375	0,918		
			26,5	6,625				
	2	5,6			6,675	0,839		
			26,9	6,725				
	3	7,5			6,7625	1,109		
			27,2	6,8				
	4	7,8			6,8375	1,141		
			27,5	6,875				
4	1	6,3			6,9375	0,908		
			28	7				
	2	5,9			7,075	0,834		
			28,6	7,15				
	3	8						
	4	8,4						

(1)Un total móvil se asocia con el dato que ocupa el lugar del medio del conjunto de valores del cual fue calculado

- Al dividir cada observación de la serie de tiempo entre el valor del promedio móvil centrado correspondiente, podemos identificar el efecto estacional-irregular en la serie de tiempo. Por ejemplo, el tercer trimestre del año 1 muestra 6.0/5.475=1.096 como el componente estacional-irregular combinado. La tabla anterior resume los valores estacionales-irregulares resultantes para toda la serie de tiempo.
- Considere el tercer trimestre. Los resultados de los años 1, 2 y 3 muestran valores del tercer trimestre de 1.096, 1.075 y 1.109, respectivamente. Por tanto, en todos los casos el componente estacional-irregular parece tener una influencia por encima del promedio en el tercer trimestre. Las fluctuaciones durante los tres años pueden atribuirse al componente irregular, por lo que podemos promediar los valores calculados para eliminar la influencia irregular y obtener una estimación de la influencia estacional del tercer trimestre igual a 1,09.

Índice estacional	
0,93	1 trimestre
0,84	2 trimestre
1,09	3 trimestre
1,14	4 trimestre

Índice estacional	
0,93	1 trimestre
0,84	2 trimestre
1,09	3 trimestre
1,14	4 trimestre

- El trimestre de mejores ventas es el cuarto, con ventas que promedian 14% por encima del valor medio trimestral.
- El trimestre con peores ventas, o más lento, es el segundo, con un índice estacional de 0.84, que muestra que las ventas promediaron 16% por debajo de las ventas medias trimestrales.

Verificar: El modelo multiplicativo requiere que el índice estacional medio sea igual 1.00.

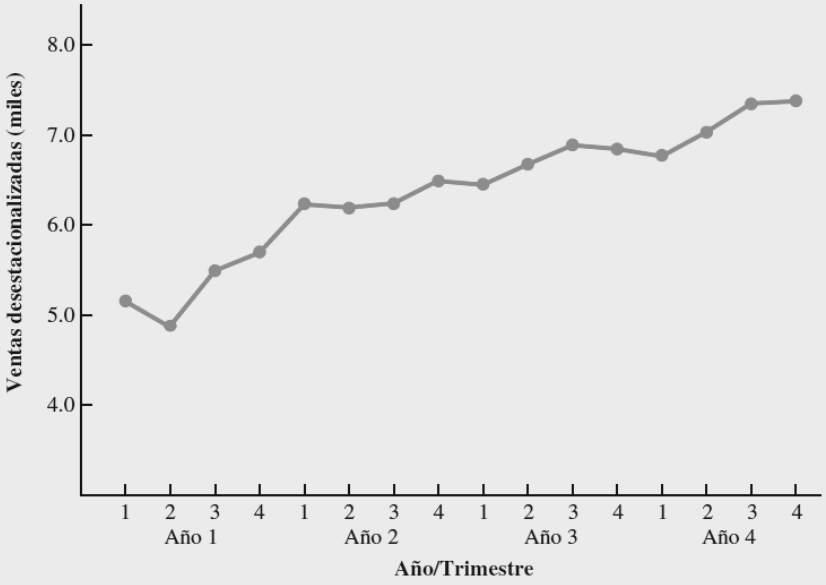
Desestacionalización de las series de tiempo

- El propósito de determinar índices estacionales es precisamente eliminar los efectos estacionales de una serie de tiempo. Este proceso se conoce como desestacionalización de las series de tiempo.

Año	Trimestre	Ventas (en miles) (Y_t)	Índice estacional (S_t)	Ventas desestacionalizadas ($Y_t/S_t = T_tI_t$)
1	1	4.8	0.93	5.16
	2	4.1	0.84	4.88
	3	6.0	1.09	5.50
	4	6.5	1.14	5.70
2	1	5.8	0.93	6.24
	2	5.2	0.84	6.19
	3	6.8	1.09	6.24
	4	7.4	1.14	6.49
3	1	6.0	0.93	6.45
	2	5.6	0.84	6.67
	3	7.5	1.09	6.88
	4	7.8	1.14	6.84
4	1	6.3	0.93	6.77
	2	5.9	0.84	7.02
	3	8.0	1.09	7.34
	4	8.4	1.14	7.37

La serie de tiempo parece tener una tendencia lineal ascendente. Para identificar esta tendencia, utilizamos el método de proyección de la tendencia; en este caso, los datos utilizados son los valores de las ventas trimestrales desestacionalizadas.

$$T_t = 5.101 + 0.148t$$



$T_t = 5.101 + 0.148\ t$

La pendiente de 0.148 indica que durante los 16 trimestres anteriores la empresa ha experimentado un crecimiento desestacionalizado medio en las ventas de aproximadamente 148 televisores por trimestre. Si suponemos que la tendencia de los 16 trimestres pasados en datos de ventas es un indicador razonablemente bueno del futuro, podemos utilizar esta ecuación para proyectar el componente de tendencia de la serie de tiempo para los 4 próximos trimestres del año 5.

t	proyección de tendencia. (en miles)
17	7,617
18	7,765
19	7,913
20	8,061

El paso final en el desarrollo del pronóstico, cuando tanto el componente de tendencia como el estacional están presentes, es utilizar el índice estacional para ajustar la proyección de tendencia.

Año	Trimestre	Pronóstico de tendencia	Pronóstico trimestral
5	1	7617	(7617)(0.93) = 7084
	2	7765	(7765)(0.84) = 6523
	3	7913	(7913)(1.09) = 8625
	4	8061	(8061)(1.14) = 9190

Observación importante

- En esta primera parte se utilizó la regresión lineal simple para ajustar una tendencia lineal a las series de tiempo de ventas de bicicletas y también para el caso de venta de televisores.

- Aquí obtuvimos una ecuación lineal que vinculaba dichas ventas con el periodo. Pero el número de bicicletas vendidas en realidad no se relaciona de manera causal con el tiempo, más bien el tiempo es un sustituto de las variables con que se relaciona en realidad el número de bicicletas vendidas, desconocidas o demasiado difíciles o costosas de medir.

- Por lo cual, el uso del análisis de regresión para la proyección de la tendencia no es un método de elaboración de pronósticos causal debido a que sólo se utilizaron los valores pasados de ventas, es decir, la variable que se pronostica.

- Cuando utilizamos el análisis de regresión para relacionar las variables que queremos pronosticar con otras variables que se supone influyen en la variable o la explican, se vuelve un método de elaboración de pronósticos causal.

Análisis de Regresión

- El Análisis de Regresión tiene como objetivo estudiar la relación entre variables. Permite expresar dicha relación en términos de una ecuación que conecta una variable de respuesta Y, con una o más variables explicativas $X_1, X_2, \ldots, X_k$.

Finalidad:

➢Determinación explícita del funcional que relaciona las variables. (Predicción)

➢Comprensión por parte del analista de las interrelaciones entre las variables que intervienen en el análisis.

Datos de corte transversal

- Una **base de datos de corte transversal** consiste en una muestra de individuos, hogares, empresas, ciudades, estados, países u otras unidades, tomada en algún punto dado en el tiempo. Algunas veces no todos los datos de estas unidades corresponden exactamente a un mismo momento.
- Por ejemplo, puede ser que, un conjunto de familias sea entrevistado durante diferentes semanas de un año. En un análisis de corte transversal puro, diferencias menores de tiempo en la recolección de los datos son ignoradas. Aun cuando un conjunto de familias haya sido entrevistado en semanas distintas de un mismo ano, se considerara como una base de datos de corte transversal.

- Se quiere estudiar la relación entre ROE (medida de desempeño de una empresa) y el pago que reciben los CEO.

- Relación entre salario y años de educación.

- Relación entre salario, años de educación y experiencia laboral

- Relación entre precio de una vivienda y metros cuadrados, cantidad de habitaciones, etc.

• A pesar de que el análisis de regresión tiene que ver con la dependencia de una variable respecto a otras variables, esto no implica causalidad necesariamente. La misma viene dada por consideraciones a priori o teóricas.

•A diferencia del análisis de correlación, en donde el principal objetivo es medir el grado de asociación lineal entre dos variables, aquí estamos interesados en estimar o predecir el valor promedio de una variable sobre la base de valores fijos de otras variables.

Análisis de regresión lineal simple

El análisis de regresión se relaciona en gran medida con la estimación y/o predicción de la media (de la población) o valor promedio de la variable dependiente, con base en los valores conocidos o fijos de las variables explicativas.

Población total de 60 familias de una comunidad hipotética.
Ingreso semanal (X) y gasto de consumo semanal (Y), en dólares.

X→ / Y↓	80	100	120	140	160	180	200	220	240	260
Consumo familiar semanal Y, $	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	–	88	–	113	125	140	–	160	189	185
	–	–	–	115	–	–	–	162	–	191
Total	325	462	445	707	678	750	685	1 043	966	1 211
Media condicional de Y, $E(Y\|X)$	65	77	89	101	113	125	137	149	161	173

Las 60 familias se dividen en 10 grupos de ingresos (de $80 a $260). Se tienen 10 valores fijos de X y los correspondientes valores de Y para cada uno de los valores X; así que hay 10 subpoblaciones Y

Se tienen 10 valores medios para las 10 subpoblaciones de Y.

⇦

A estos valores medios se les denomina valores esperados condicionales, en vista de que dependen de los valores dados a la variable condicional X. Se denota por E(Y/X)

Resulta importante distinguir dichos valores condicionales esperados del valor esperado incondicional del gasto de consumo semanal, E(Y).

E(Y)=7272/60=121,2

Es incondicional en el sentido de que para obtener esta cifra se omiten los niveles de ingresos de las diversas familias

¿Cuál es el valor esperado del gasto de consumo semanal de una familia?

La media incondicional: $121,20

¿Cuál es el valor esperado del gasto de consumo semanal de una familia cuyo ingreso semanal es $100 ,La media condicional: $77

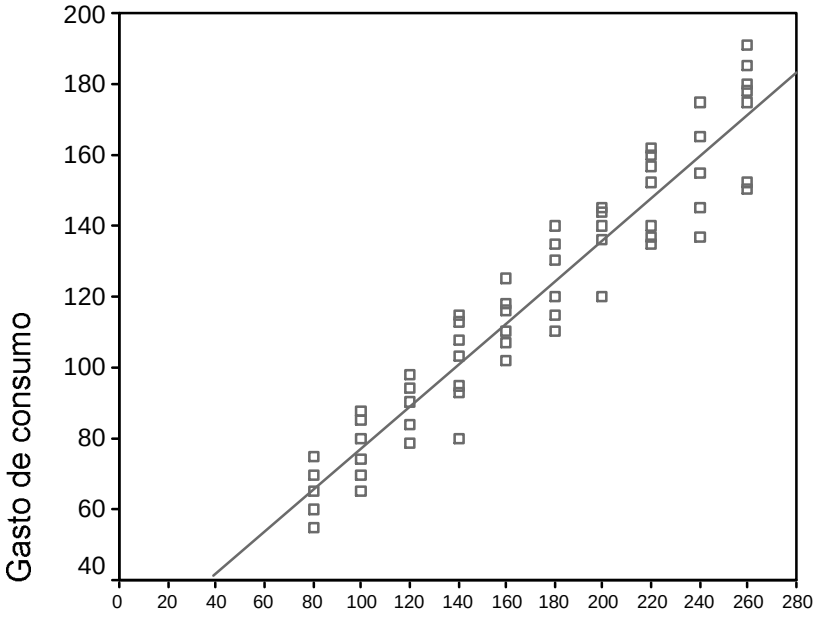
Saber el nivel de ingreso nos permite predecir mejor el valor medio del gasto de consumo.

Se puede observar en él gráfico de dispersión, al unir las medias condicionales la recta de regresión poblacional (RRP). (o regresión de Y sobre X).

El adjetivo “poblacional “ se debe al hecho de que en este ejemplo se consideró una población de 60 familias.

Gráfico de dispersión

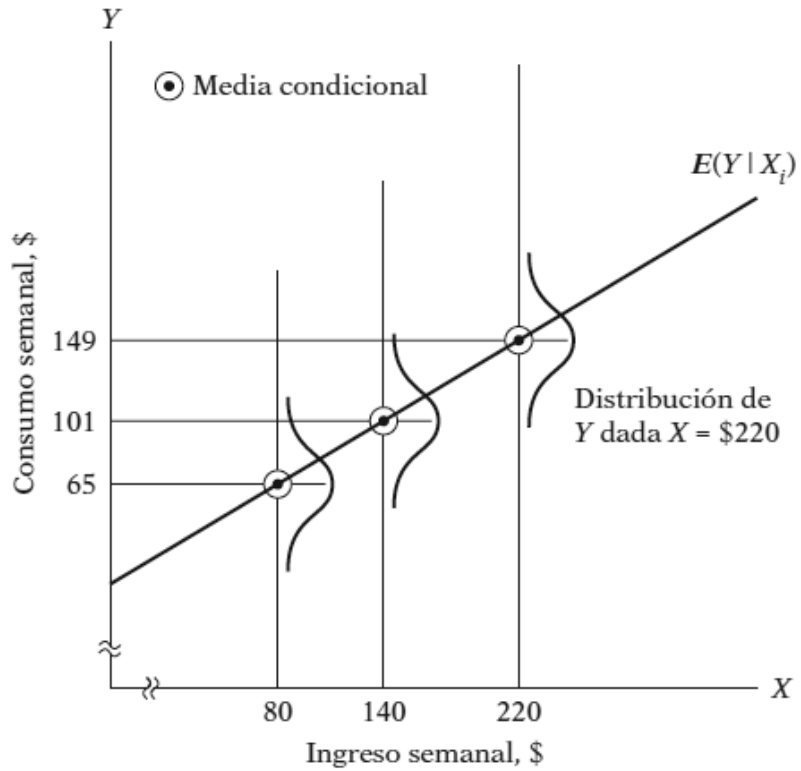
Gasto de consumo v/s Ingreso



A pesar de la variabilidad del gasto para cada ingreso, en promedio el consumo semanal se incrementa en la misma medida que el ingreso

Curva de regresión poblacional

Desde el punto de vista geométrico, **una curva de regresión poblacional** es simplemente el lugar geométrico de las medias condicionales de la variable dependiente para los valores fijos de la (s) variables explicativa(s).



Es la curva que conecta las medias de las subpoblaciones de Y que corresponden a los valores del regresor X.

Concepto de función de regresión poblacional (FRP)

Es claro que cada media condicional E(Y/Xi) es función de Xi, donde Xi es un valor dado de X.

E(Y/Xi)=f(Xi) y f(Xi) denota alguna función de la variable explicativa X.

¿Qué forma toma la función f(Xi)?

En una situación real no tenemos la totalidad de la población para efectuar el análisis.
La forma funcional de la FRP es, una pregunta empírica, aunque en casos específicos la teoría puede tener algo que decir. Por ejemplo, un economista podría plantear que el gasto de consumo está relacionado linealmente con el ingreso.
Por tanto, como una primera aproximación podemos suponer que la FRP es una función lineal de Xi

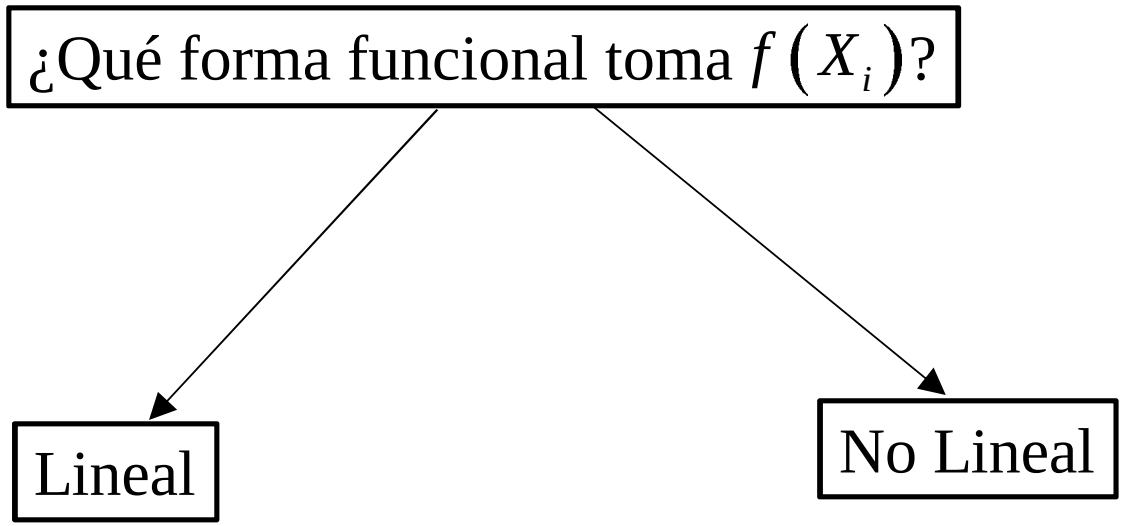
$$E(Y / X_i) = \beta_1 + \beta_2 X_i$$

β_1 y β_2 son parámetros no conocidos pero fijos que se denominan coeficientes de regresión

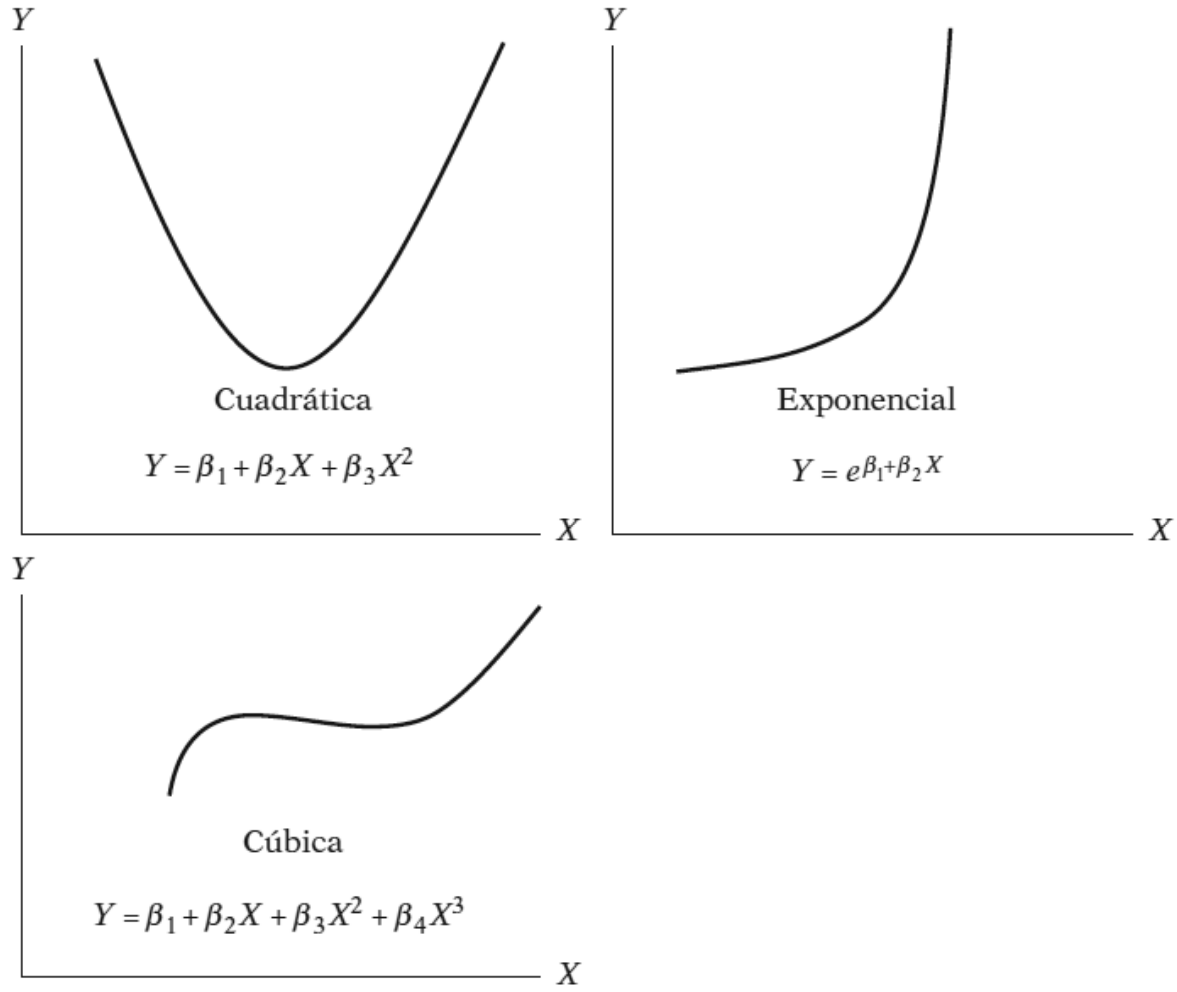
- ***Función de Regresión Poblacional***

$$E\left(Y\middle|X_i\right) = f\left(X_i\right)$$

El valor esperado de la distribución de *Y* esta funcionalmente relacionado con X_i, pero...



El término regresión “lineal” siempre significará una regresión lineal en los parámetros.



•Entre otras las formas funcionales *lineales* se destacan:

$$\boxed{Y = \alpha + \beta . X}$$

$$Y = \alpha . X^{\beta}$$

$$Y = \exp(\alpha + \beta . X)$$

• La primer ecuación es lineal

•La segunda ecuación se puede trasformar en:

$$\log Y = \log \alpha + \beta \log X$$

La tercer ecuación se puede transformar en

$$\log Y = \alpha + \beta X$$

• Veamos la interpretación de cada coeficiente β

Modelo	Variable dependiente	Variable independiente	Interpretación de β_1
Nivel-nivel	y	x	$\Delta y = \beta_1 \Delta x$
Log-nivel	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

$$\widehat{\log(wage)} = 0.584 + 0.083\, educ$$

$$n = 526,\ R^2 = 0.186.$$

El coeficiente de *educ* tiene una interpretación porcentual multiplicándolo por 100: $\widehat{wage}$ aumenta 8.3% por cada año más de educación. Esto es a lo que los economistas se refieren cuando hablan de "rendimiento de un año más de educación".

$$\widehat{\log(salary)} = 4.822 + 0.257 \log(sales)$$

$$n = 209,\ R^2 = 0.211.$$

El coeficiente de log(*sales*) es la elasticidad estimada de *salary* (sueldo) respecto a *sales* (ventas). Esto implica que por cada aumento de 1% en las ventas de la empresa hay un aumento de aproximadamente 0.257% en el sueldo de los CEO —la interpretación usual de una elasticidad.

Ecuación de regresión poblacional FRP

$$E(Y / X_i) = \beta_1 + \beta_2 X_i \quad \Leftarrow$$

Ecuación de regresión poblacional FRP

Donde β_1 y β_2 son parámetros no conocidos pero fijos que se denominan coeficientes de regresión.

En el análisis de regresión el interés es estimar la FRP, es decir estimar los valores de β_1 y β_2 no conocidos con base en las observaciones de Y y X

Especificación estocástica de la FRP

¿Qué podemos decir sobre la relación entre el gasto de consumo de una familia individual y un nivel dado de ingresos?

Se observa en la figura , que dado el nivel de ingresos de Xi, el gasto de consumo de una familia individual está agrupado alrededor del consumo promedio de todas las familias en ese nivel de Xi, esto es, alrededor de su esperanza condicional. Por consiguiente, podemos expresar la desviación de un Yi individual alrededor de su valor esperado de la siguiente manera:

$$Y_i = E(Y / X_i) + u_i \quad \text{o} \quad u_i = Y_i - E(Y / X_i)$$

Donde la desviación u_i es una variable aleatoria no observable que toma valores positivos o negativos. Técnicamente **, u_i es conocida como perturbación estocástica o término de error estocástico.**

Especificación estocástica de la FRP

Se puede decir que el gasto de una familia individual, dado su nivel de ingresos, puede ser expresado como la suma de dos componentes

$$Y_i = E(Y / X_i) + u_i$$

La media del gasto de consumo de todas las familias con el mismo nivel de ingresos.

Componente aleatorio .
Es un sustituto para todas aquellas variables que son omitidas del modelo pero que colectivamente afectan a Y

Especificación estocástica de la FRP

$$Y_i = E(Y / X_i) + u_i = \beta_1 + \beta_2 X_i + u_i$$

La ecuación plantea que el gasto de consumo de una familia está relacionado linealmente con su ingreso, más el término de perturbación. Así los gastos de consumo individual, dado X=$80, pueden ser expresados como

$$Y_1 = 55 = \beta_1 + \beta_2(80) + u_2$$
$$Y_2 = 60 = \beta_1 + \beta_2(80) + u_2$$
$$Y_3 = 65 = \beta_1 + \beta_2(80) + u_3$$
$$Y_4 = 70 = \beta_1 + \beta_2(80) + u_4$$
$$Y_5 = 75 = \beta_1 + \beta_2(80) + u_5$$

Así, el supuesto de que la recta de regresión pasa a través de las medias condicionales de Y implica que los valores de la media condicional de u_i son cero.

Especificación estocástica de la FRP

La especificación estocástica

$$Y_i = E(Y / X_i) + u_i = \beta_1 + \beta_2 X_i + u_i$$

Tiene la ventaja que muestra claramente otras variables además del ingreso, que afectan el gasto de consumo y que un gasto de consumo de familias individuales no puede ser explicado en su totalidad solamente por la(s) variable(s) incluida(s) en el modelo de regresión.

Función de regresión muestral (FRM)

En la práctica lo que se tiene al alcance no es más que una muestra de valores de Y que corresponden a algunos valores fijos de X. Por consiguiente la labor ahora es estimar la FRP con base en información muestral.
Supóngase que no se conocía la población de la tabla 1 y que la única información que se tenía era una muestra de valores de Y seleccionada aleatoriamente para valores dados de X tal como se presenta en la tabla 2

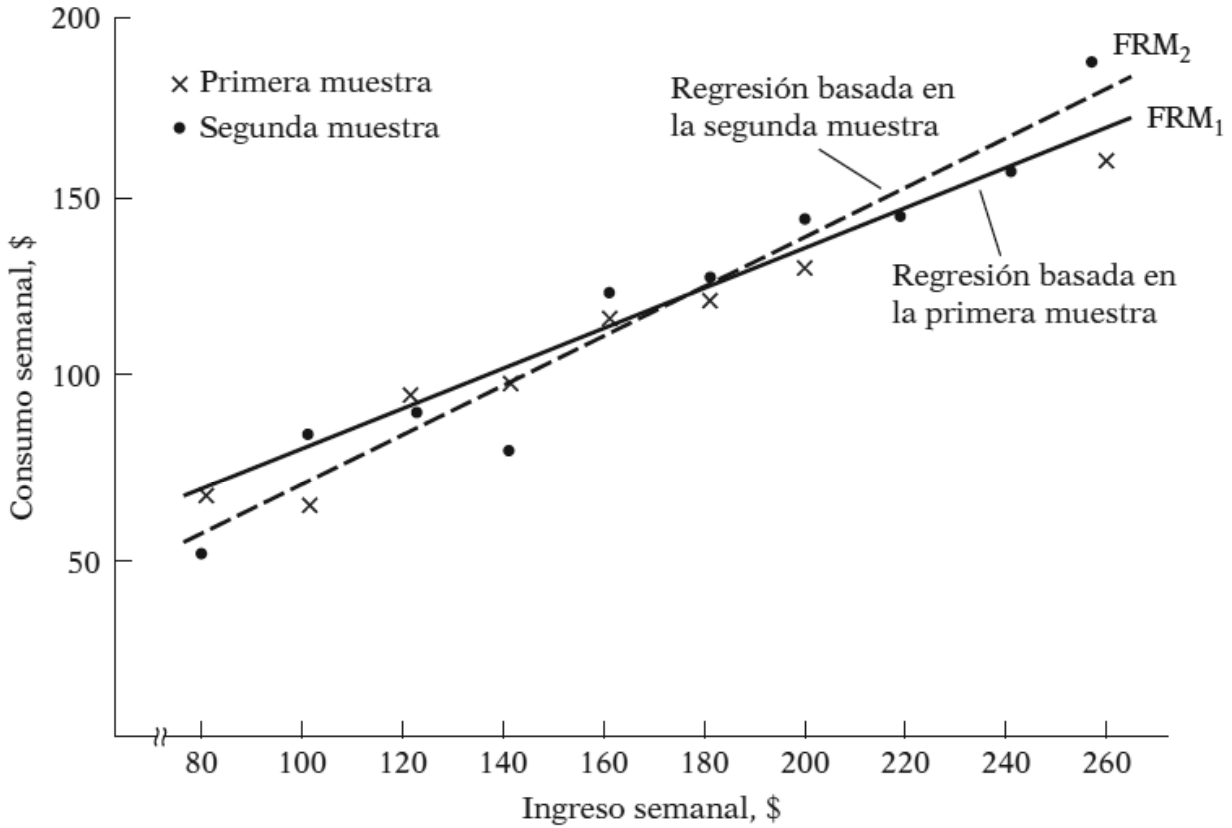
De la muestra de la tabla 2,
¿se puede predecir el gasto de consumo semanal promedio Y para la población correspondiente a los valores de X seleccionados?

¿Se puede estimar la forma FRP a partir de la información muestral?

Y	X
70	80
65	100
90	120
95	140
110	160
115	180
120	200
140	220
155	240
150	260

Función de regresión muestral (FRM)

Consideremos otra muestra tomada de la población de la tabla.
Las rectas de la figura se conocen como rectas de regresión muestral. En general, se podrían obtener N FRM diferentes para N muestras diferentes y estas FRM no necesariamente son iguales



Y	X
55	80
88	100
90	120
80	140
118	160
120	180
145	200
135	220
145	240
175	260

Ahora, en forma análoga a la FRP en la cual se basa la recta de regresión poblacional, se puede desarrollar el concepto de función de regresión muestral. La contraparte muestral puede escribirse como:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

Es la contraparte de

$$E(Y / X_i) = \beta_1 + \beta_2 X_i$$

Donde

$\hat{Y}_i = e\text{stimador de } \mathrm{E(Y/X)}$

$\hat{\beta}_1 = \text{estimador de } \beta_1$

$\hat{\beta}_2 = \text{estimador de } \beta_2$

Atención!: que un estimador, conocido también como estadístico (muestral), no es más que una regla, fórmula o método para estimar el parámetro poblacional a partir de la información suministrada por la muestra disponible. Un valor numérico particular obtenido por el estimador en un análisis se conoce como estimación. Cabe señalar que un estimador es aleatorio, pero una estimación no.

Función de regresión muestral (FRM) en su forma estocástica

La FRM en su forma estocástica se puede expresar como

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{\mu}_i$$

Donde $\hat{\mu}_i$ denota el término residual (muestral)

Conceptualmente es análogo a μ_i y puede ser considerado como un estimación de μ_i

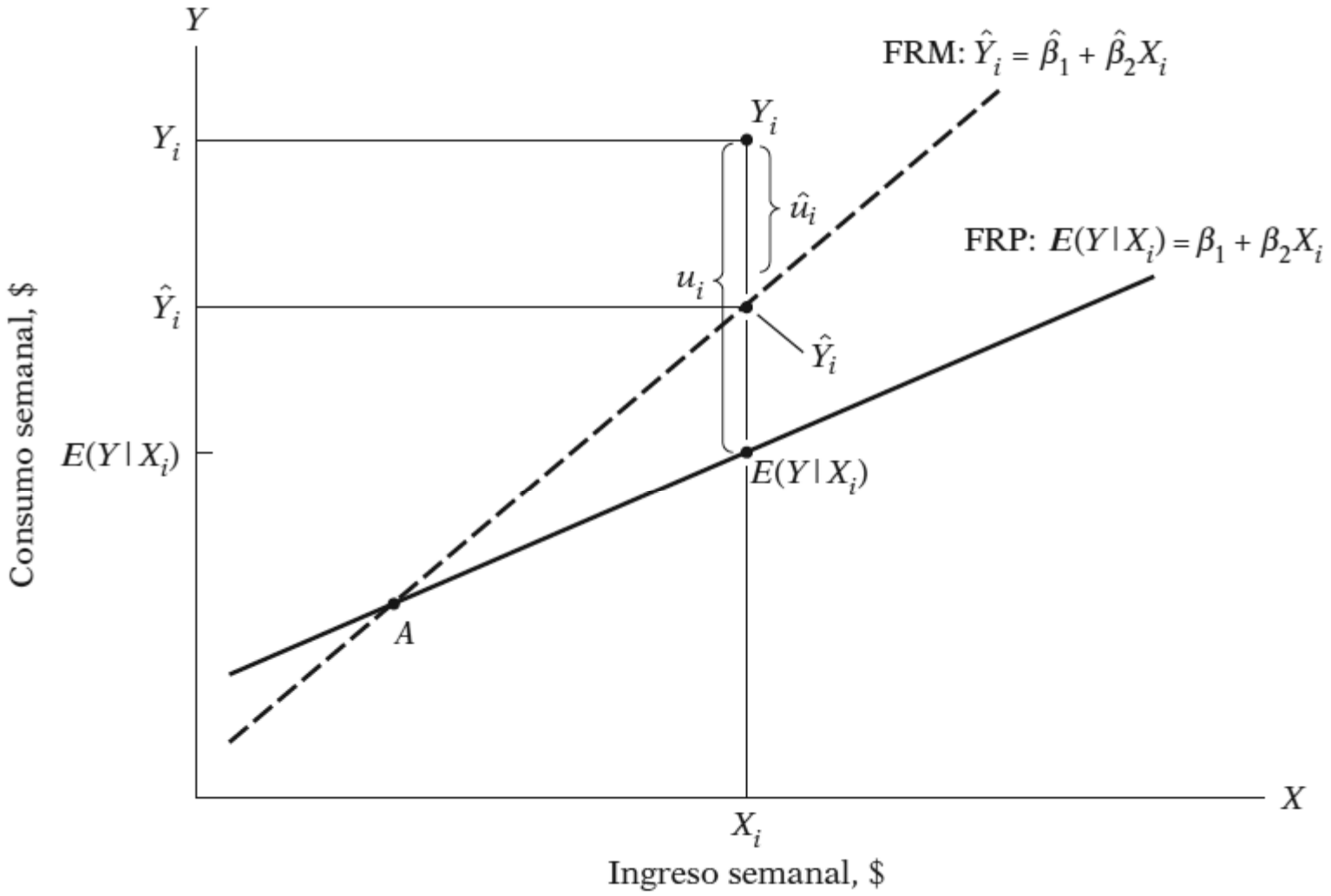
El objetivo principal en el análisis de regresión es estimar la FRP

$$Y_i = \beta_1 + \beta_2 X_i + \mu_i$$

Con base en la FRM

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{\mu}_i$$

Rectas de regresión muestral y poblacional



Debido a fluctuaciones muestrales, la estimación de la FRP basada en la FRM es, en el mejor de los casos, una aproximación.

Rectas de regresión muestral y poblacional

Para $X=X_i$, se tiene una observación muestral $Y=Y_i$. En términos de la FRM, la Y_i observada puede ser expresada como

$$Y_i = \hat{Y}_i + \hat{\mu}_i$$

Y en términos de la FRP, puede ser expresada como

$$Y_i = E(Y / X_i) + \mu_i$$

Dado que la FRM es apenas una aproximación de la FRP, ¿se puede diseñar un método que haga que esta aproximación sea lo más ajustada posible?

Función de regresión simple: problema de estimación

La tarea consiste en estimar la función de regresión poblacional (FRP) con base en la función de regresión muestral (FRM) en la forma más precisa posible.
Los dos métodos de estimación que suelen utilizarse son:

1) Los mínimos cuadrados ordinarios (MCO)
2) La máxima verosimilitud (MV).

El método de MCO es el que más se emplea en el análisis de regresión.

Método de mínimos cuadrados ordinarios (MCO)

El método MCO se atribuye a Carl Friedrich Gauss un matemático alemán. Bajo ciertos supuestos el método tiene algunas propiedades estadísticas muy atractivas que lo han convertido en uno de los más eficaces y populares del análisis de regresión.

Primero se estima $\hat{u}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$

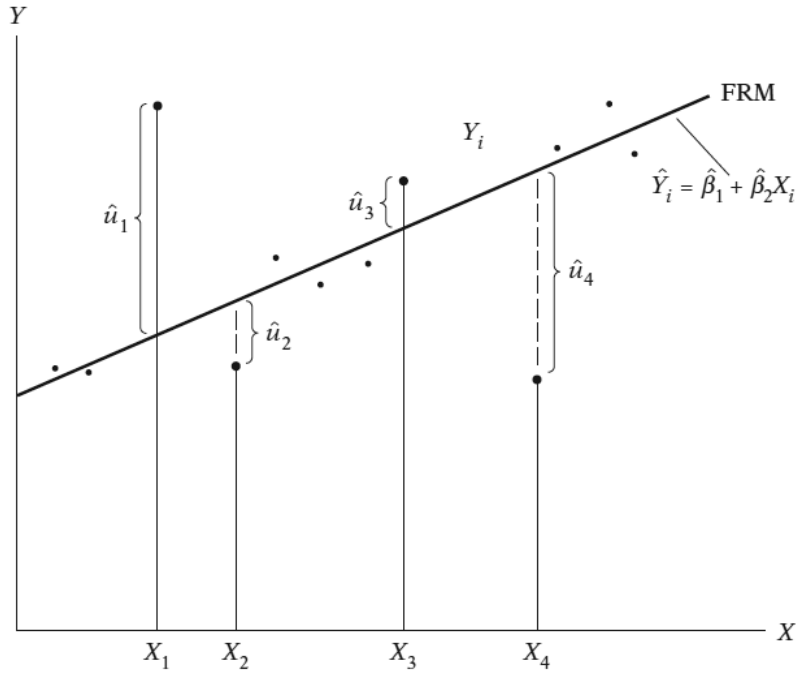
que muestra que los residuos son simplemente las diferencias entre los valores observados y los estimados de Y.

Ahora, dados n pares de observaciones de Y y X, se está interesado en determinar la FRM de tal manera que esté lo más cerca posible a la Y observada.

Método de mínimos cuadrados ordinarios (MCO)

Con este fin se puede adoptar el siguiente criterio: seleccionar la FRM de tal manera que la suma de los residuos : $\sum \hat{u}_i = \sum \left(Y_i - \hat{Y}_i \right)$ sea la menor posible.

Este criterio, no es muy bueno porque a todos los residuos se les da la misma importancia sin considerar qué tan cerca o qué tan dispersas estén las observaciones individuales de la FRM. Debido a lo anterior, es muy posible que la suma algebraica de los residuos sea pequeña (aun cero) a pesar de que las $\hat{u}_i$ están bastante dispersas alrededor de FRM.



Método de mínimos cuadrados ordinarios (MCO)

Se puede evitar este problema si se adopta el criterio de mínimos cuadrados, el cual establece que la FRM puede determinarse en forma tal que

$$\sum \hat{u}_i^{\,2} = \sum \left(Y_i - \hat{Y}_i \right)^2 = \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i \right)^2$$

sea la menor posible. Este método da más peso a los residuos tales como $\hat{u}_1$ y $\hat{u}_4$ que a los residuos $\hat{u}_2$ y $\hat{u}_3$

El procedimiento de MCO genera las siguientes ecuaciones para estimar β_1 y β_2 donde n es el tamaño de la muestra

Método de mínimos cuadrados ordinarios (MCO)

$$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i$$

$$\sum Y_i X_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2$$

Ecuaciones normales

Resolviendo las ecuaciones normales simultáneamente se obtiene

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} \qquad \hat{\beta}_1 = \overline{\mathrm{Y}} - \hat{\beta}_2 \overline{\mathrm{X}}$$

Estimadores de mínimos cuadrados

- Los estimadores obtenidos se conocen como estimadores de mínimos cuadrados, pues se derivan del principio de mínimos cuadrados. Estos estimadores tienen propiedades numéricas por haber sido obtenidos con el método de MCO: “Propiedades numéricas son las que se mantienen como consecuencia del uso de mínimos cuadrados ordinarios, sin considerar la forma como se generaron los datos”.

- Existen también las propiedades estadísticas de los estimadores MCO, es decir, propiedades “que se mantienen sólo con ciertos supuestos sobre la forma como se generaron los datos”.

- Si deseamos estimar sólo $\beta 1$ y $\beta 2$, basta el método MCO presentado de la sección anterior.

- Por consiguiente, mientras no se especifique la forma como se crean o se generan X_i y u_i no hay manera de hacer alguna inferencia estadística sobre Y_i, ni tampoco, sobre $\beta 1$ y $\beta 2$.

- Así, los supuestos sobre la(s) variable(s) X_i y el término de error son relevantes para lograr una interpretación válida de los valores estimados de la regresión.

Modelo clásico de regresión lineal supuestos detrás del método MCO

El modelo de Gauss, modelo clásico o estándar de regresión lineal (MCRL) el cual es el cimiento de la mayor parte de la teoría econométrica, plantea 9 supuestos.

Supuesto 1: Modelo de regresión lineal
El modelo de regresión es lineal en los parámetros

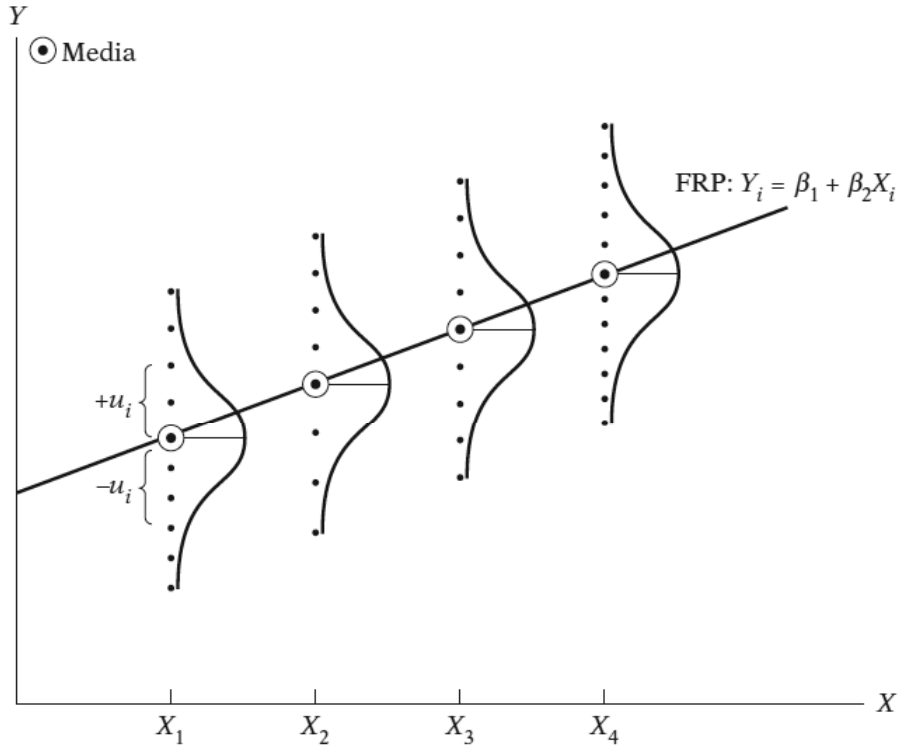
$$Y_i = \beta_1 + \beta_2 X_i + \mu_i$$ modelo simple

Supuesto 2: Los valores de X son fijos en muestreo repetido.
Significa que el análisis de regresión es un análisis de regresión condicional, esto es, condicionado a los valores dados del (los) regresor X.

Supuesto 3: El valor medio de la perturbación ui es igual a cero.

Dado el valor de X, el valor esperado del término aleatorio de perturbación ui es cero.

$$E(u_i / X_i) = 0$$

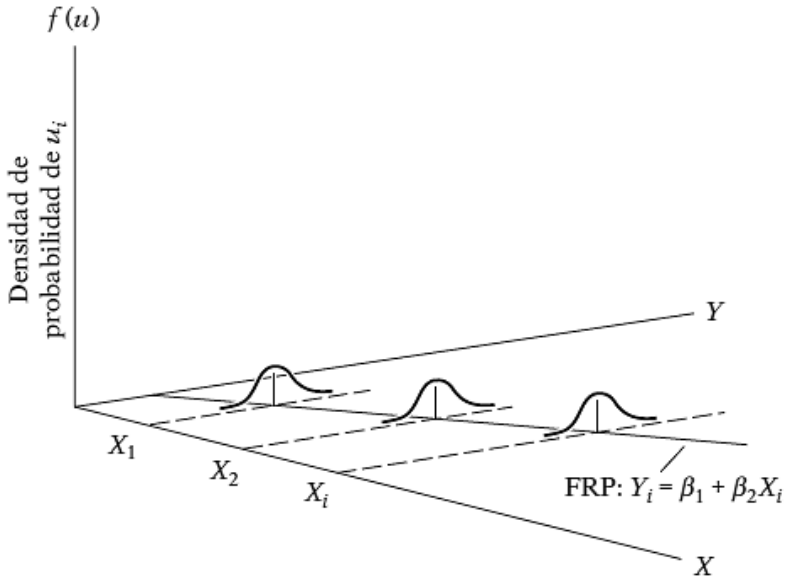


Nótese que el supuesto E(ui/Xi)=0 implica que

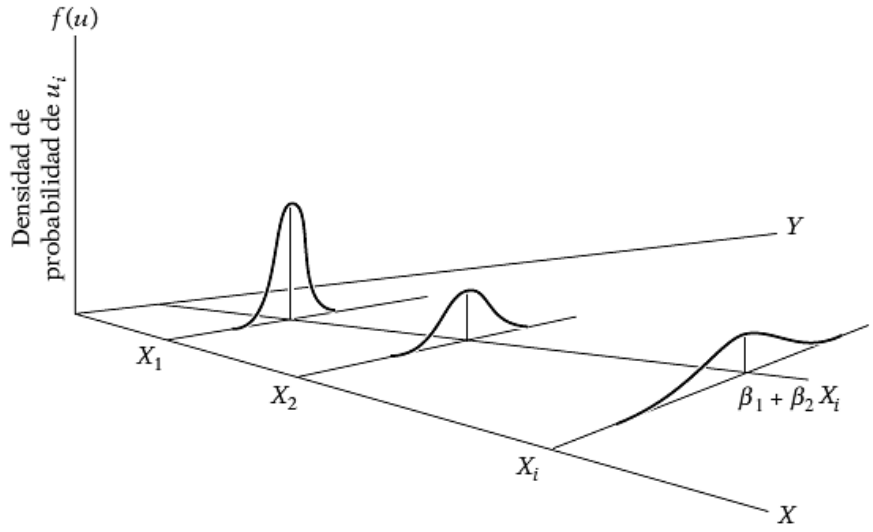
$$E(Y / X_i) = \beta_1 + \beta_2 X_i$$

Supuesto 4: Homocedasticidad o igual varianza de ui.
Dado el valor de X, la varianza de ui es la misma para todas las observaciones, es decir, las varianzas condicionales de ui son idénticas.

$$\text{var}(ui / Xi) = \sigma^2$$



Homocedasticidad



Heterocedasticidad

Homocedasticidad

Caso donde se cumple el supuesto:

Figura: Gráfico de y vs x

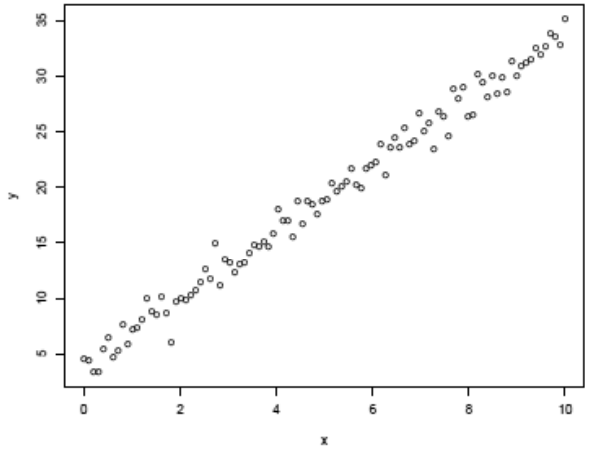
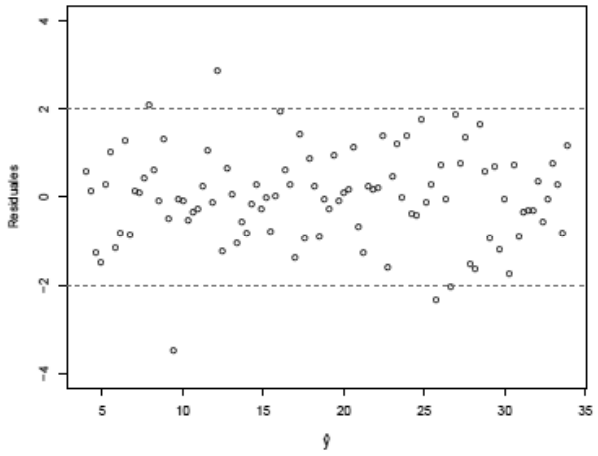


Figura: Gráfico de residuales vs y ajustados



Homocedasticidad

Caso donde no se cumple el supuesto:

Figura: Gráfico de y vs x

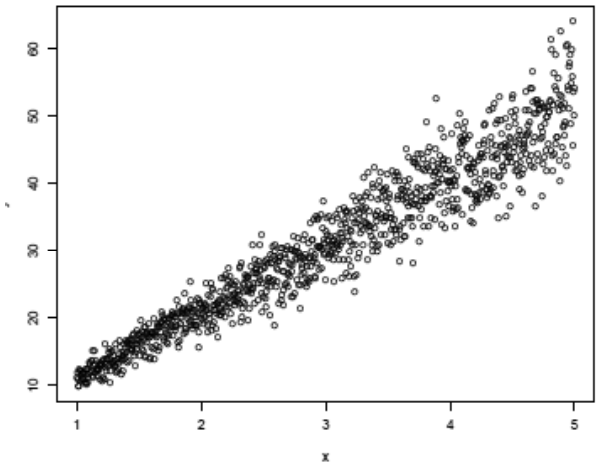
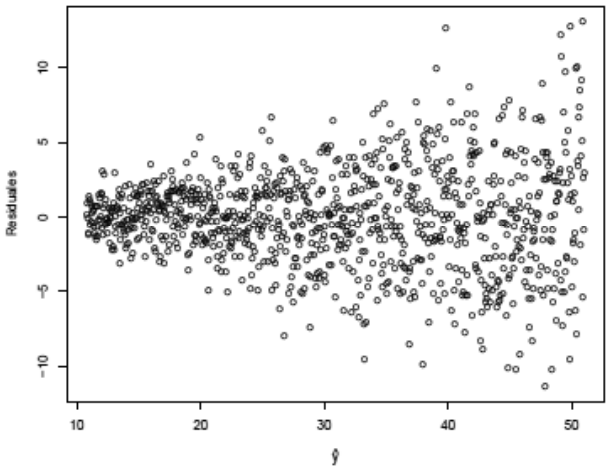


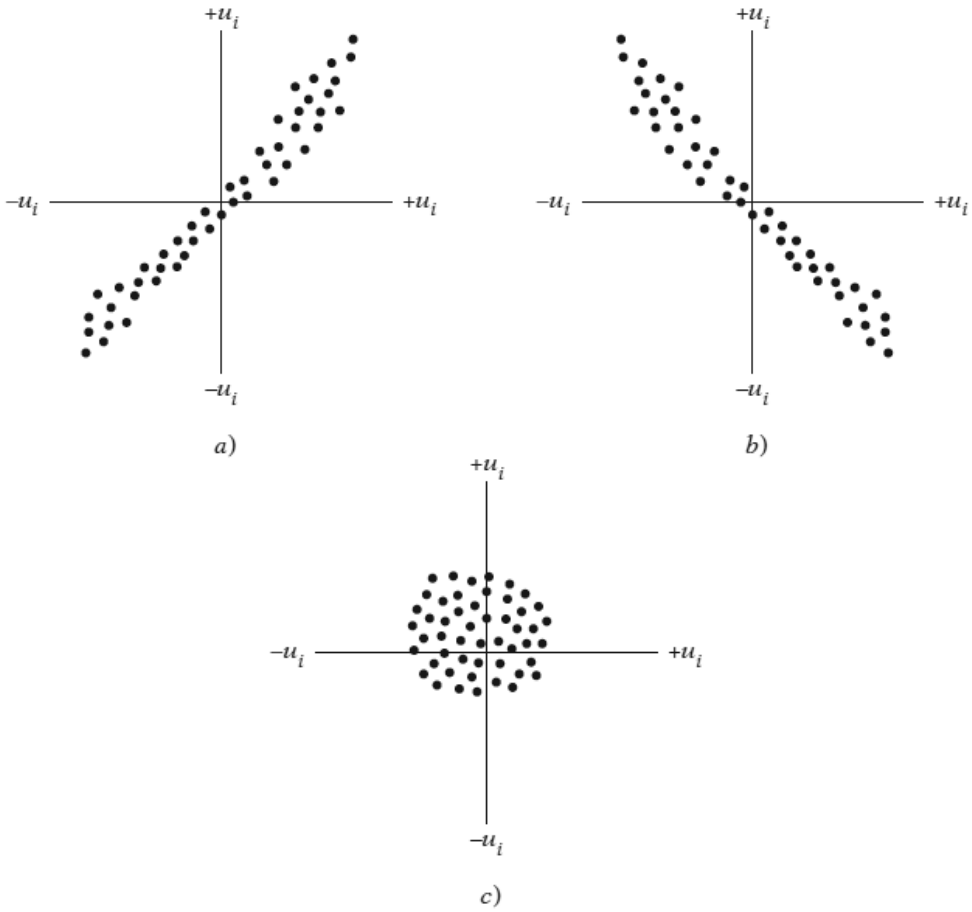
Figura: Gráfico de residuales vs y ajustados



Supuesto 5: No existe auto correlación entre las perturbaciones.

Dados dos valores cualquiera de X, Xi y Xj , la correlación entre dos ui y uj es cero.

$$\text{cov}(ui, uj / Xi, X_j) = 0$$



Supuesto 6: La covarianza entre ui y Xi es cero o E(uiXi)=0

$$\text{cov}(ui, X_i) = 0$$

Supuesto 7: El número de observaciones n debe ser mayor que el número de parámetros por estimar.

Supuesto 8: Variabilidad en los valores de X.

No todos los valores de X en una muestra dada deben ser iguales.

$$\text{var}(X) > 0$$

Supuesto 9: No hay sesgo de especificación

El supuesto de normalidad: El modelo clásico de regresión lineal normal

Recordemos que con los supuestos vistos anteriormente los estimadores de MCO $\hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2$ satisfacían diferentes propiedades estadísticas muy deseables, tales como insesgamiento y varianza mínima . Si nuestro objetivo es únicamente la estimación puntual el método de MCO será suficiente, sin embargo la estimación puntual es sólo la formulación de un aspecto de la inferencia estadística.

Nuestro interés no consiste solamente en estimar la función muestral de regresión (FRM), sino también en utilizarla para obtener inferencias respecto a la función de regresión poblacional (FRP).

El supuesto de normalidad: El modelo clásico de regresión lineal normal

La regresión lineal normal clásica supone que cada ui, está normalmente distribuida con

$$\textit{Media}: \qquad E(u_i) = 0$$

$$\textit{Varianza}: \qquad E\left[u_i - E(u_i)\right]^2 = E(u_i^2) = \sigma^2$$

$$Cov(u_i, u_j): \qquad E\left(\left[u_i - E(u_i)\right]\left[u_j - E(u_j)\right]\right) = E(u_i u_j) = 0 \quad i \neq j$$

Estos supuestos pueden expresarse en forma más compacta como

$$u_i \sim \mathrm{N}(0, \sigma^2)$$

Normalidad

Caso donde se cumple el supuesto:

Figura: Histograma de los residuales

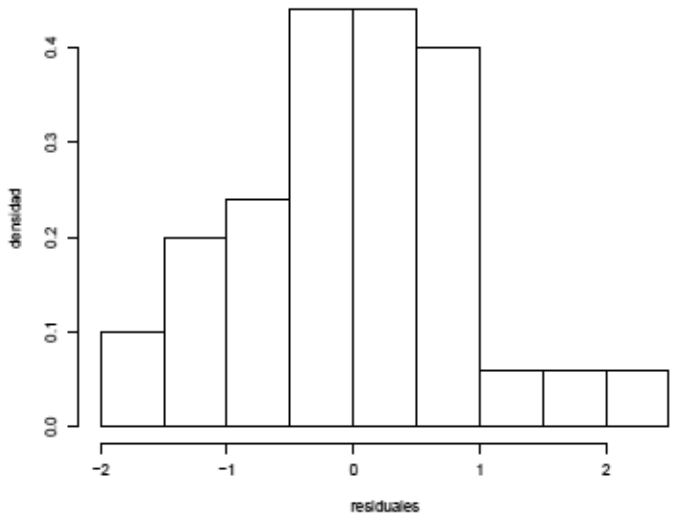
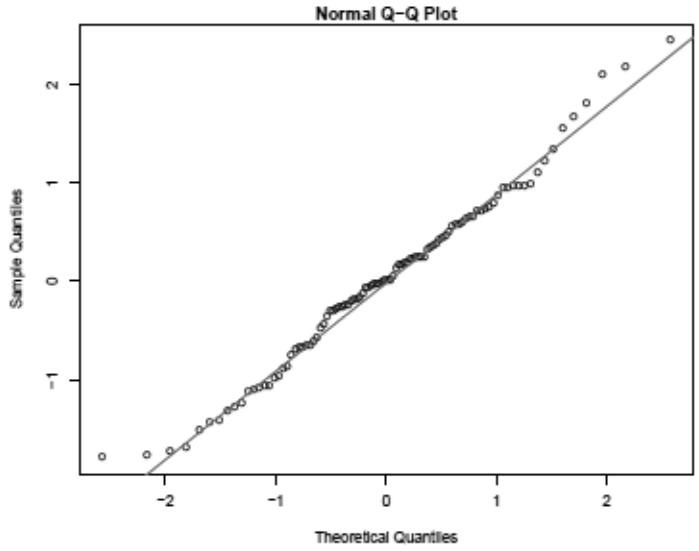


Figura: qq-plot de los residuales



Prueba de hipótesis para β_2

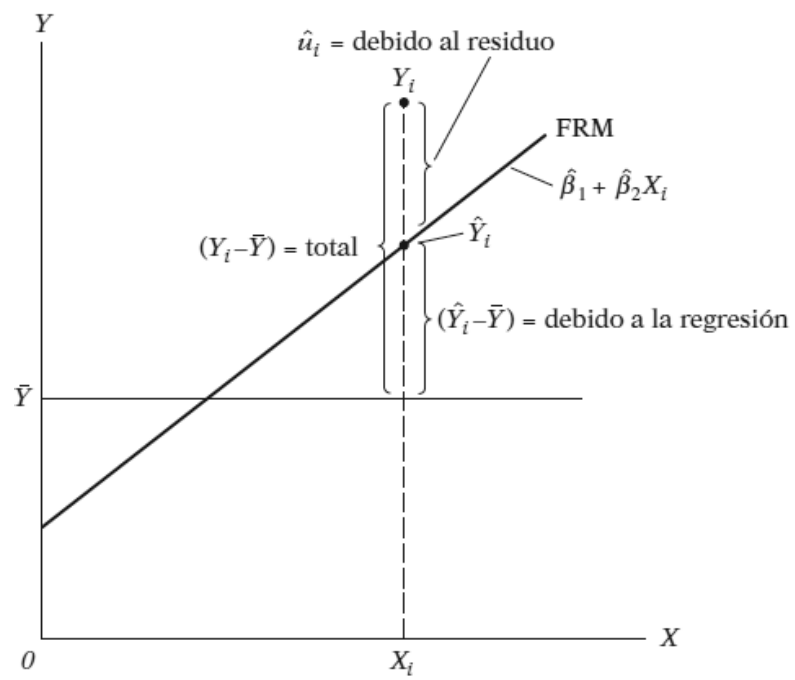
Tipo de hipótesis	H_0: hipótesis nula	H_1: hipótesis alternativa	Regla de decisión: rechazar H_0 si
Dos colas	$\beta_2 = \beta_2^*$	$\beta_2 \neq \beta_2^*$	$\lvert t \rvert > t_{\alpha/2,\text{gl}}$
Cola derecha	$\beta_2 \leq \beta_2^*$	$\beta_2 > \beta_2^*$	$t > t_{\alpha,\text{gl}}$
Cola izquierda	$\beta_2 \geq \beta_2^*$	$\beta_2 < \beta_2^*$	$t < -t_{\alpha,\text{gl}}$

$$t = \frac{\hat{\beta}_2 - \beta_2}{\text{ee}\,(\hat{\beta}_2)}$$ sigue la distribución t con n − 2 gl.

• La bondad de ajuste de la recta de regresión es equivalente a determinar cuán bien se ajusta la recta de regresión a los datos muestrales. Como medida de esto surge el **coeficiente de determinación muestral** (ó r^2):

$$r^2 = \hat{\beta}_2^2 \left(\frac{S_x^2}{S_y^2} \right)$$

•Verbalmente, r^2 mide la proporción o el porcentaje de la variación total en Y explicada por el modelo de regresión.



Dos propiedades de r^2:

1. Es una cantidad no negativa.

2. Sus límites son $0 \leq r^2 \leq 1$. Un r^2 de 1 significa un ajuste perfecto. Por otra parte, un r^2 de cero significa que no hay relación alguna entre la variable regresada y la variable regresora, es decir, la mejor predicción de cualquier valor de Y es simplemente el valor de su media. En esta situación, por consiguiente, la línea de regresión será horizontal al eje X.

REGRESION LINEAL MULTIPLE

Generalizando la función de regresión poblacional (FRP) de dos variables se puede escribir la FRP de tres variables así:

$$Y_i = \beta_1 + \beta_2 \mathrm{X}_{1\mathrm{i}} + \beta_3 \mathrm{X}_{2\mathrm{i}} + \mu_{\mathrm{i}}$$

donde *Y es* la variable dependiente, X_1 y X_2 las variables explicativas (o regresoras). u_i es el término de perturbación estocástica, e i la iésima observación.
Los coeficientes se denominan coeficientes de regresión parcial

Se continúa operando dentro del marco del modelo clásico de regresión lineal (MCRL).

Modelo de tres variables

Supuestos

Específicamente. se supone lo siguiente

- Valor medio de ui, igual a cero

$$E(u_i / X_{1i}, X_{2i}) = 0 \quad \text{para} \quad \text{cada} \quad \text{i}$$

- No correlación serial

$$\text{cov}(u_i, u_j) = 0 \quad \text{i} \neq \text{j}$$

- Homocedasticidad

$$\text{var}(u_i) = \sigma^2$$

Supuestos

- Covarianza entre ui y cada variable X igual a cero

$$\text{cov}(\ u_i\ ,\ X_{1i}\ ) = \text{cov}(\ u_i\ ,\ X_{2i}\ ) = 0$$

- No hay sesgo de especificación

  El modelo está especificado correctamente

- No hay colinealidad exacta entre las variables X

  No hay relación lineal exacta entre X_1 y X_2

Adicionalmente, se supone que el modelo de regresión múltiple es lineal en los parámetros, que los valores de las regresoras son fijos en muestreos repetido y que hay suficiente variabilidad en dichos valores.

Interpretación de la ecuación de regresión múltiple

Dados los supuestos del modelo de regresión clásico, se cumple que, al tomar la esperanza condicional de *Y* *a* ambos lados de

se obtiene $$Y_i = \beta_1 + \beta_2 \mathrm{X}_{1\mathrm{i}} + \beta_3 \mathrm{X}_{2\mathrm{i}} + \mu_{\mathrm{i}}$$

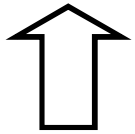
$$E(Y_i / X_{1i}, X_{2i}) = \beta_1 + \beta_2 \mathrm{X}_{1\mathrm{i}} + \beta_3 \mathrm{X}_{2\mathrm{i}}$$

Expresado en palabras, de la expresión anterior se obtiene la media condicional o el valor esperado de *Y* condicionado a los valores dados o fijos de las variables X_1 y X_2. Por consiguiente, igual que en el caso de dos variables, el análisis de regresión múltiple es el análisis de regresión condicional, sobre los valores fijos de las variables explicativas, y lo que obtenemos es el valor promedio o la media de *Y,* o la respuesta media de *Y* a valores dados de las regresoras X.

Nota: Las propiedades de los estimadores MCO del modelo de regresión múltiples son similares a aquellas del modelo con dos variables

Significado de los coeficientes de regresión parcial

Los coeficientes de regresión β_2 y β_3 se denominan coeficientes de regresión parcial.



β_2 mide el cambio en el valor de la media de Y, E(Y) por unidad de cambio en X_1 permaneciendo X_2 constante.

β_3 mide el cambio en el valor medio de Y, E(Y) por unidad de cambio en X_2 cuando el valor de X_1 se conserva constante.

Prueba de la significación global de la regresión

La significación global de la regresión se puede probar con la relación de la varianza explicada a la varianza no explicada: Esta sigue una distribución F con k-1 y n-k grados de libertad, donde n es el número de observaciones y k es el número de parámetros estimados.

$$F_{k-1,n-k} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

Si la relación F calculada excede el valor tabulado de F al nivel especificado de significación y grados de libertad, se acepta la hipótesis de que los parámetros de la regresión no son todos iguales a cero y que R cuadrado es significativamente diferente de cero.

$$H_0 : \beta_1 = \beta_2 = \beta_3 = ... \beta_n = 0$$

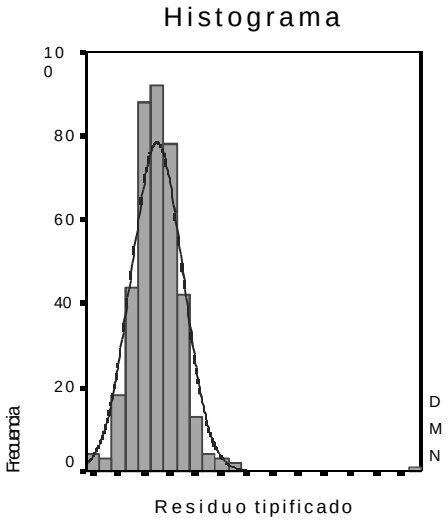
$$H_1 : \text{No todas las } \beta \text{ son cero}$$

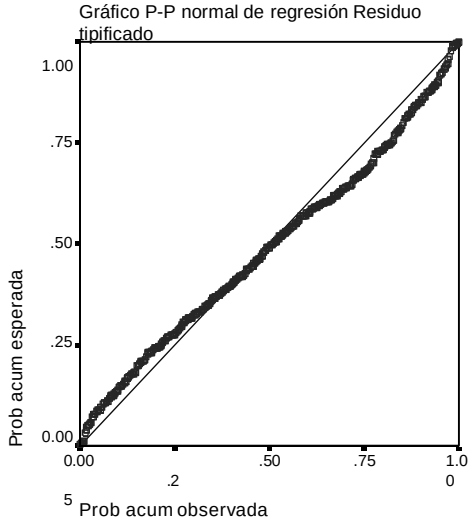
Chequear Supuestos

1. Normalidad de los residuos
2. No autocorrelación
3. Homocedasticidad
4. Linealidad
5. No multicolinealidad

Por ejemplo Normalidad de los residuos

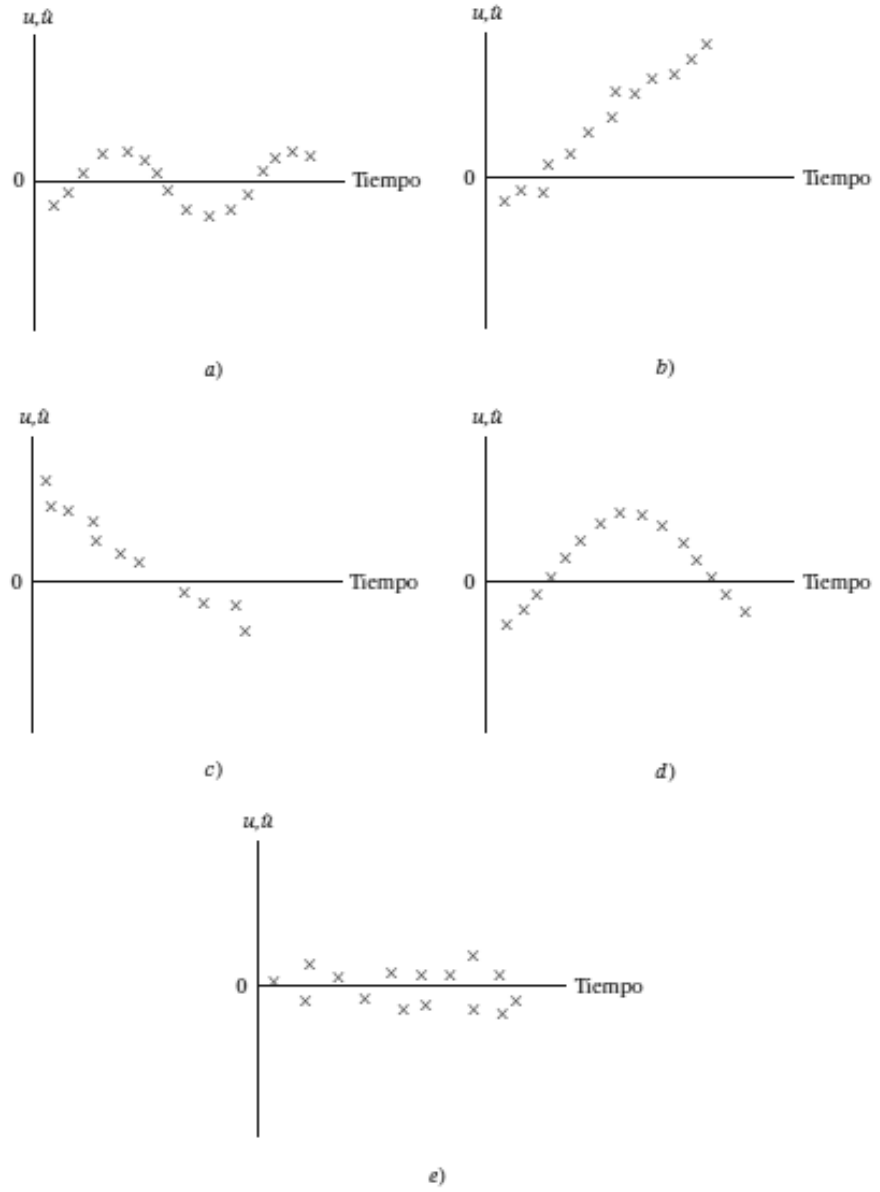
Gráficos: Histograma, gráfico probabilístico normal



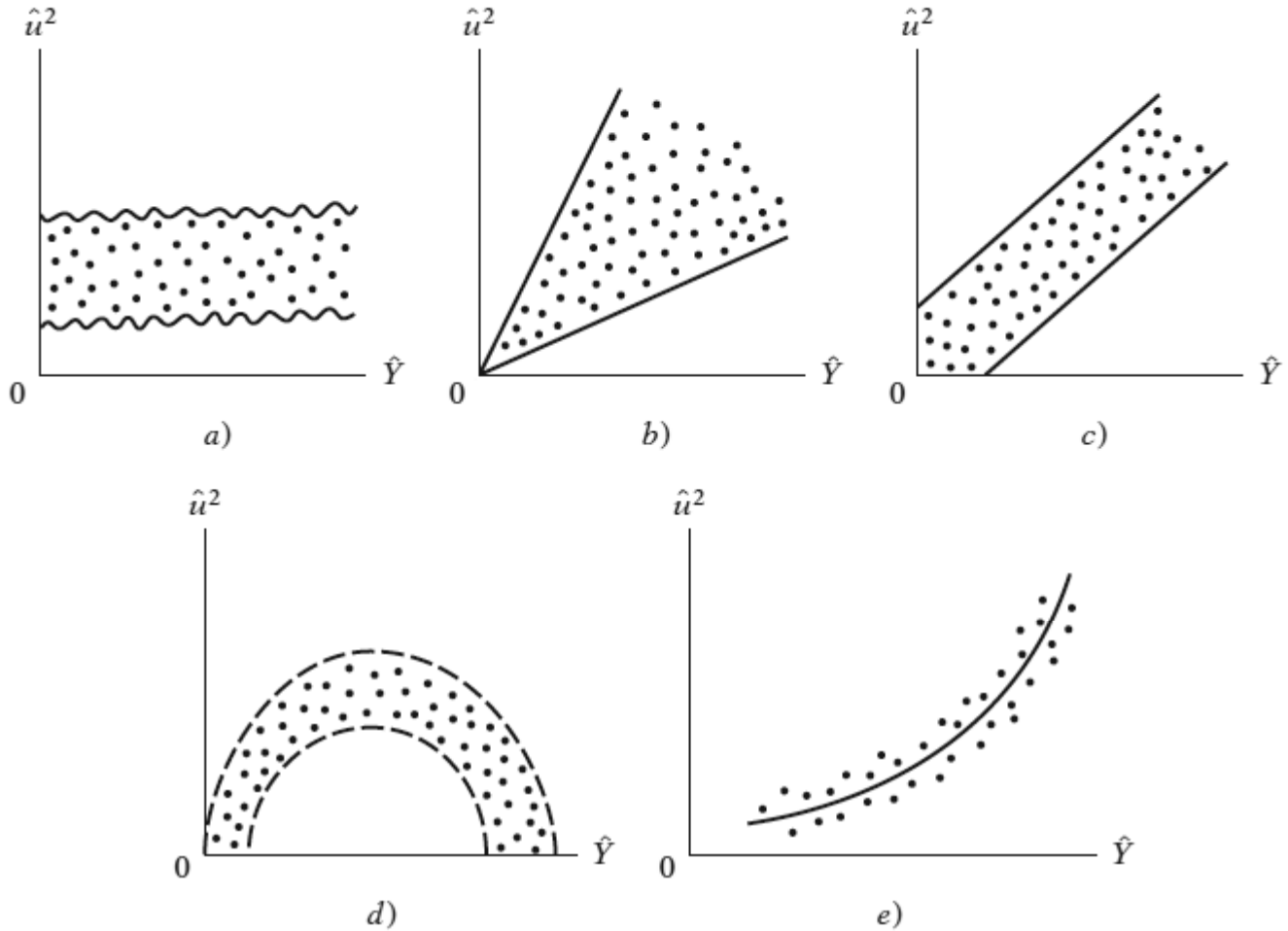


No autocorrelación

El caso e) indica que no hay un patrón sistemático, apoyando el supuesto de no autocorrelación de los residuos.



El primer gráfico (arriba-izq) nos estaría diciendo que no habría heteroscedasticidad. Sin embargo los otros gráficos muestran patrones definidos.



Linealidad

- Se observan los gráficos de regresión parcial. Para examinar la relación entre la variable dependiente y cada una de las independientes por separado.

No Multicolinealidad

- FIV>10 se dice que la variable es altamente colineal.
- IC(índice de condición) si esta entre 10 y 30 existe multicolinealidad entre moderada y fuerte y si excede 30, existe multicolinealidad severa.
- Al enfrentar el problema de multicolinealidad severa, una de las soluciones mas simples consiste en omitir del modelo una de las variables colineales.

Pruebas formales

Homocedasticidad: Prueba de Goldfeld-Quant, prueba de White.

Incorrelación de los errores: Prueba de Durbin-Watson, prueba de rachas.

Normalidad de los errores: Pruebas de Shapiro-Wilks, prueba de Anderson Darling

Ejemplo

Analizar la relación existente entre el grado de estrés de los trabajadores Y a partir del tamaño de la empresa en que trabajan X_1, el número de año que llevan en el puesto de trabajo actual X_2 , salario anual percibido X_3 y la edad del trabajador X_4.

obs	X1 tamaño	X2 Años puesto	X3 Salario anual	X4 edad	Y gradodee
1	812	15	30	38	101
2	334	8	20	52	60
3	377	5	20	27	10
4	303	10	54	36	27
5	505	13	52	34	89
6	401	4	27	45	60
7	177	6	26	50	16
8	598	9	52	60	184
9	412	16	34	44	34
10	127	2	28	39	17
11	601	8	42	41	78
12	297	11	84	58	141
13	205	4	31	51	11
14	603	5	38	63	104
15	484	8	41	30	76

Una vez hallada la relación pedida entre las variables, evaluar la capacidad predictiva del modelo y hallar predicciones del grado de estrés de los trabajadores para los valores siguientes de las variables independientes:

x1	x2	x3	x4
302	9	44	42
351	8	65	62
381	9	52	53

Análisis de regresión con Datos de Series de tiempo

En términos formales, a una secuencia de variables aleatorias indexadas en el tiempo se le llama proceso estocástico o proceso de series de tiempo (“estocástico” es sinónimo de aleatorio). Cuando se conforma una base de datos de series de tiempo, se obtiene un resultado posible, o realización, del proceso estocástico. Únicamente se puede ver una sola realización, ya que no es posible retroceder en el tiempo y empezar de nuevo el proceso. (Esto es análogo al análisis de corte transversal en el que únicamente se puede reunir una sola muestra aleatoria.) No obstante, si ciertas condiciones históricas fueran distintas, por lo general se obtendría una realización diferente para el proceso estocástico y es por ello que los datos de series de tiempo se consideran como el resultado de variables aleatorias.

Datos para regresión múltiple con series de tiempo

Periodo	Serie de tiempo (Y_t)	Valor de las variables independientes						
		x_{1t}	x_{2t}	x_{3t}	.	.	.	x_{kt}
1	Y_1	x_{11}	x_{21}	x_{31}	.	.	.	x_{k1}
2	Y_2	x_{12}	x_{22}	x_{32}	.	.	.	x_{k2}
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
n	Y_n	x_{1n}	x_{2n}	x_{3n}	.	.	.	x_{kn}

Y_t valor de la serie de tiempo en el periodo t

x_{1t} = valor de la variable independiente 1 en el periodo t

x_{2t} = valor de la variable independiente 2 en el periodo t

$$\hat{Y}_t = \hat{\beta}_1 + \hat{\beta}_2 X_{1t} + \hat{\beta}_3 X_{2t}$$

Ecuación regresión estimada con 2 variables independientes.

Condiciones para utilizar análisis estadístico inferencial con datos de series de tiempo

- Un proceso estocástico es estacionario en sentido estricto o fuerte cuando la distribución de probabilidad conjunta de la serie es invariante con respecto al tiempo.
- Un proceso estocástico es estacionario en el sentido débil si su media y su varianza son constantes en el tiempo y si el valor de la covarianza entre dos periodos depende sólo de la distancia o rezago entre estos dos periodos, y no del tiempo en el cual se calculó la covarianza.
- Ergodicidad: Las observaciones muy lejanas en el tiempo no están correlacionadas. Es necesaria para poder contar con suficientes observaciones independientes para estimar los parámetros del modelo.

- Una serie no estacionaria tendrá media y/o varianza que cambian en el tiempo
- Si una serie es no estacionaria se puede estudiar su comportamiento sólo durante el período de observación.
- Cada conjunto de datos pertenecerá a un episodio particular
- No puede generalizarse
- Tienen poco valor práctico

Por tanto, las series de tiempo estacionarias y débilmente dependientes son ideales para el análisis de regresión múltiple.

COINTEGRACION

- La regresión de una variable de serie de tiempo sobre una o mas variables de serie de tiempo, frecuentemente puede dar resultados sin sentido o espurios. Este fenómeno se conoce como regresión espuria.
- Una forma de protegerse de esta es establecer si las series de tiempo están cointegradas.
- Cointegración significa que a pesar de no ser estacionarias a nivel individual, una combinación lineal de dos o mas series de tiempo puede ser estacionaria.

Ejemplo

Gasto en consumo personal contra Ingreso disponible

$CP_t = \beta_1 + \beta_2 ID_t + u_t$

Se puede expresar

$u_t = CP_t - \beta_1 - \beta_2 ID_t$

- Se somete el u_t estimado a un test de raíz unitaria (para probar estacionariedad)
- Si es estacionaria la regresión de consumo contra ingreso sería cointegrada. Existe una relación de equilibrio o largo plazo

Ejemplo final

- La novak corporation desea desarrollar un modelo de pronostico para la proyección de la ventas futura. Ya que la corporación tiene tiendas a lo largo de una extensa región se eligen los ingresos personales disponibles (x_1) como variable explicativa posible. A continuación se presentan los siguientes datos anuales desde 2000 a 2016:

fila	año	ventas(millones)	ingreso personal (millones)	tasa desempleo(%)
1	2000	8	336,1	5,5
2	2001	8,2	349,4	5,5
3	2002	8,5	362,9	6,7
4	2003	9,2	383,9	5,5
5	2004	10,2	402,8	5,7
6	2005	11,4	437	5,2
7	2006	12,8	472,2	4,5
8	2007	13,6	510,4	3,8
9	2008	14,6	544,5	3,8
10	2009	16,4	588,1	3,6
11	2010	17,8	630,4	3,5
12	2011	18,6	685,9	4,9
13	2012	20	742,8	5,9
14	2013	21,9	801,3	5,6
15	2014	24,9	903,1	4,9
16	2015	27,3	983,6	5,6
17	2016	29,1	1076,7	8,5

$$\hat{Y}_t = -0.014 + 0.03X_{1t} + 0.35X_{2t}$$

El modelo final estimado incorpora la tasa de desempleo.

La función $\hat{Y}_t = -0.014 + 0.03X_{1t} + 0.35X_{2t}$ puede utilizarse para predecir las ventas ya que se cumplen todos los supuestos. Datos de expertos se utilizan para estimar el ingreso personal y la tasa de desempleo para la región para generar un pronostico de las ventas de Novak para 2017.

Si x_1 (1185) y x_2 (7,8) El pronostico de ventas esperado para 2017 es 32,8 millones.

$$\hat{Y}_t = -0.014 + 0.03X_{1t} + 0.35X_{2t}$$

- En otro tipo de modelo de pronóstico basado en la regresión, las variables independientes son todos los valores anteriores de la misma serie de tiempo. Por ejemplo, si los valores de la serie de tiempo se denotan Y_1, Y_2, ..., Y_n, y la variable independiente es Y_t, se trata de hallar una ecuación de regresión estimada que relacione Y_t con los valores más recientes de la serie de tiempo Y_{t-1}, Y_{t-2}, etc. Si se emplean como variables independientes los tres periodos más recientes, la ecuación estimada de regresión será

$$\hat{Y}_t = \hat{\beta}_1 + \hat{\beta}_2 Y_{t-1} + \hat{\beta}_3 Y_{t-2} + \hat{\beta}_4 Y_{t-3}$$

- A los modelos de regresión que tienen variables independientes con los valores anteriores de la serie de tiempo se les conoce como modelos autorregresivos.

Bibliografía de referencia

- Anderson R., Sweeney D., Williams T., Camm J. y Cochran J. (2015), "Quantitative Methods for Business" Cengage Learning. USA.
- Anderson D., Sweeney D. y Williams T. (2008), "Estadística para Administración y Economía". 10º edición. Ed. Thomson. México.
- Canavos, G. (2003), "Probabilidad y Estadística. Teoría y aplicaciones". Mc Graw Hill. Interamericana de México.
- Enders, W. (2008), "Applied Econommetric Time Series". Editorial Wiley. Inglaterra.
- Guisande Gonzalez, C., Vaamonde Liste, A. y Barreiro Felpeto, A. (2011), "Tratamiento de datos con R, Statistica y SPSS". Ed. Díaz de Santos. España.
- Gujarati, D. y Porter D. (2010), "Econometría" 5° Edición. Mc Graw Hill. México.
- Hanke, J. and Wichern, D. (2006). "Pronósticos en los negocios". Editoral Pearson. México.
- Levine D., Stephan D, Krehbiel T., and Berenson M. (2008). "Statistics for Managers". Pearson New Jersey.

Software: Statgraphics - QM for Windows (Pearson)