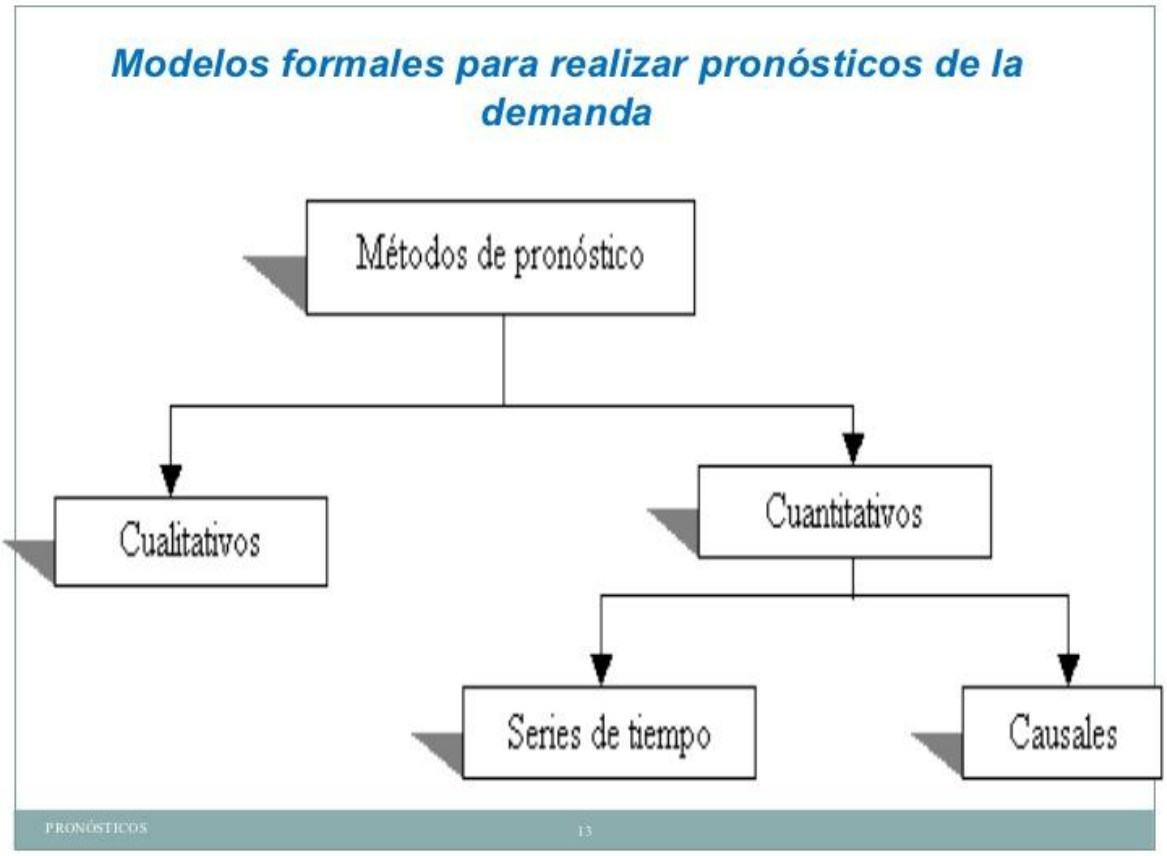
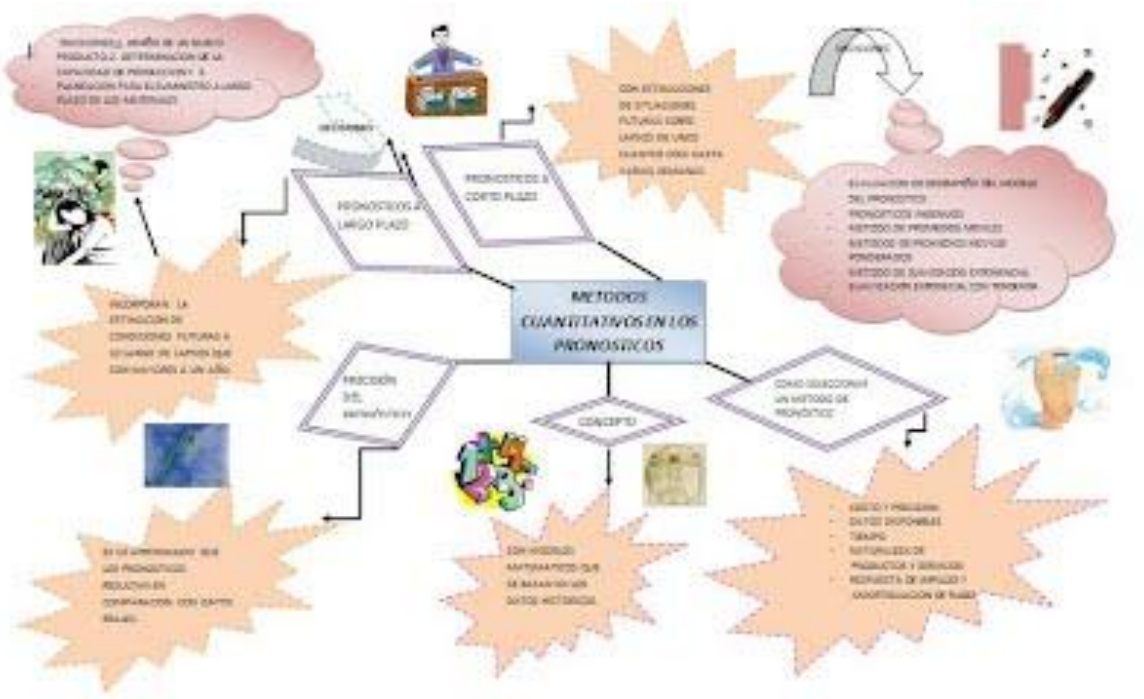
2.3.1 Métodos cuantitativos para los pronósticos.



MÉTODOS CUANTITATIVOS

Los modelos cuantitativos de pronósticos son modelos matemáticos que se basan en datos históricos. Estos modelos suponen que los datos históricos son relevantes en el futuro. Casi siempre puede obtenerse información pertinente al respecto. Aquí, analizaremos varios modelos cuantitativos, la precisión del pronóstico, pronósticos a largo plazo y pronósticos a corto plazo.



Modelos cuantitativos de pronóstico:

1.- Regresión lineal. Modelo que utiliza el método de los mínimos cuadrados para identificar la relación entre una variable dependiente y una o más variables independientes, presentes en un conjunto de observaciones históricas. En la regresión simple, solo hay una variable independiente; en la regresión múltiple, hay más de una variable independiente, en por ejemplo, un pronóstico de ventas, son las ventas. Una modelo de regresión no necesariamente tiene que estar basado en una serie de tiempo, pues en estos casos el conocimiento de los valores futuros de la variable independiente (llamada también variable causal) se utiliza para predecir valores futuros de la variable dependiente. Por lo general, la regresión lineal se utiliza en pronósticos a largo plazo.

2.- Promedios móviles: Modelos de pronósticos del tipo de series de tiempo a corto plazo que pronostica las ventas para el siguiente periodo. En este modelo, el pronóstico aritmético de las ventas reales para un determinado número de los periodos pasados más recientes es el pronostico para el siguiente periodo.

3.- Promedio móvil ponderado: modelo parecido al modelo de promedio móvil arriba descrito, excepto que el pronóstico para el siguiente periodo es un promedio ponderado de las ventas pasadas, en lugar del promedio aritmético.

4.- Suavización exponencial: modelo también de pronóstico de series de tiempo a corto plazo que pronostica las ventas para el siguiente periodo. En este método, las ventas pronosticadas para el último periodo se modifican utilizando la información

correspondiente al error de pronóstico del último periodo. Esta modificación del pronóstico del último periodo se utiliza como pronostico para el siguiente periodo.

5.- Suavización exponencial con tenencia. El modelo de suavización exponencial arriba descrito, pero modificado para tomar en consideración datos con un patrón de tendencia. Estos patrones pueden estar presentes en datos a mediano plazo. También se conoce como suavización exponencial doble, ya que se suavizan tanto la estimación del promedio como la estimación de la tendencia utilizando dos constantes de suavización.

Regresión Lineal Simple

Nos centraremos en primer lugar, en el caso de que la función que relaciona las dos variables X e Y sea la más simple posible, es decir, una línea recta.

Por ello pasaremos a interpretar los coeficientes que determinan una línea recta.

Toda función de la forma Y=a+bX determina, al representarla en el plano una línea recta, donde X e Y son variables y a y b son constantes. Por ejemplo: Y=3+2X.

SIGNIFICADO DE a y b

a es la ordenada en el origen, es decir, es la altura a la que la recta corta al eje Y. Se denomina también ***término independiente.***

b, también denominada ***pendiente*** es la inclinación de la recta, es decir, es el incremento que se produce en la variable Y cuando la variable X aumenta una unidad.

Por ejemplo, en el caso anterior Y=3+**2**X, por cada unidad que incrementa la X, la Y presenta un incremento medio de 2 unidades.

En la recta de regresión -como ya veremos- ***b*** recibe el nombre de ***Coeficiente de regresión.***

Si b>0, entonces cuando X aumenta Y también lo hace (relación directa).

Si b<0, entonces, cuando X aumenta Y disminuye (relación inversa).

Ver figura 6.4a y b respectivamente.

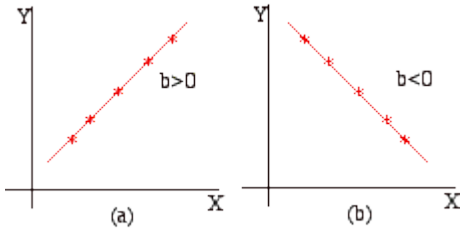


Figura 6.4: Signo de la pendiente en una recta de regresión

ESTIMACIÓN DE LA RECTA DE REGRESIÓN POR EL MÉTODO DE LOS MÍNIMOS CUADRADOS

Sean X e Y dos variables aleatorias medidas sobre los mismos individuos, y sean (x_i,y_i) los pares de observaciones sobre dichos individuos.

En primer lugar procederemos a representar el diagrama de dispersión, o nube de puntos. Supongamos que es la obtenida en la figura 6.5. Aunque la nube revele una gran dispersión, podemos observar una cierta tendencia lineal al aumentar X e Y (tendencia que no es del todo exacta; por ejemplo si suponemos que X es la edad e Y es la talla, obviamente, la talla no sólo depende de la edad, además también puede haber errores de medida).

Por esa nube de puntos podemos hacer pasar infinitas rectas. De todas ellas debemos elegir una ¿cual?... Obviamente elegiremos la mejor de todas en algún sentido.

La recta de regresión debe tener carácter de línea media, debe ajustarse bien a la mayoría de los datos, es decir, pasar lo más cerca posible de todos y cada uno de los puntos.

Llamaremos a la mejor de todas Y*=a+bX (Y* para distinguir los valores de la tabla de los que se habrían producido con la recta si la relación fuese funcional).

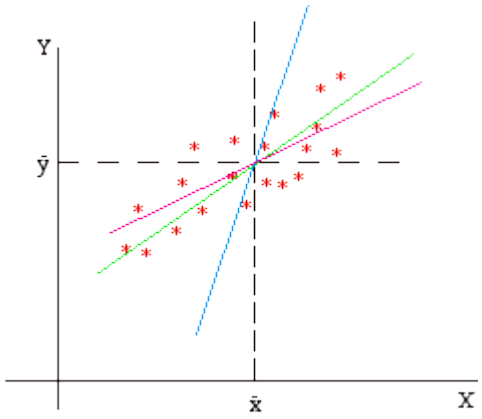


Figura 6.5: Nube de puntos y posibles rectas que pueden pasar por ella.

Que pase lo más cerca posible de todos los puntos, es decir que diste poco de todos y cada uno de ellos significa que hemos de adoptar un criterio particular que en general se conoce como MÍNIMOS CUADRADOS. Este criterio significa que la suma de los cuadrados de las distancias verticales de los puntos a la recta debe ser lo más pequeña posible (ver figura 6.6). (Obviamente, este es uno de los posibles criterios a adoptar, pero es el más utilizado).

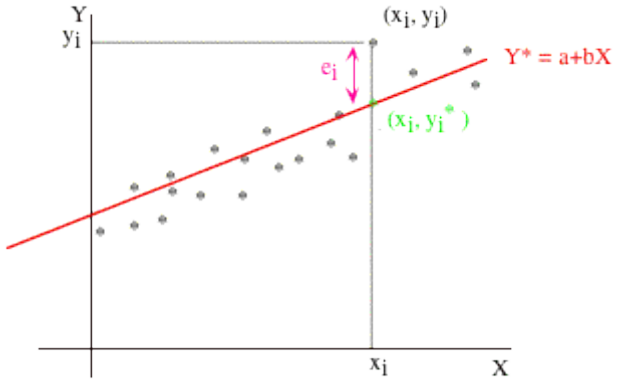


Figura 6.6: Recta de regresión mostrando los residuos o errores que se minimizan

en el procedimiento de ajuste de los Mínimos cuadrados.

Estas distancias verticales se denominan errores o residuos.

Entonces el criterio puede expresarse:

$$D = \sum_{i=1}^{n} e_i \qquad \text{mínima}$$

Dado que la recta de regresión deberá tener carácter de línea media, esa suma de distancias deberá anularse (lo mismo que sucedía, como veíamos en la primera unidad didáctica al tratar de hallar la suma de las diferencias con respecto a la media aritmética). Por las mismas razones que entonces, para evaluar la dispersión, trabajaremos con esas distancias, pero al cuadrado, de modo que la función que deberemos minimizar será:

$$D = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - y_i^*\right)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

donde y_i^* son los valores estimados según el modelo Y=a+bX

En la anterior expresión lo conocemos todo, excepto **a** y **b**. Para encontrar dichos valores, con la condición de que D sea mínima, deberemos hallar las derivadas parciales de D con respecto a **a** y a **b**, y resolver el sistema resultante, al igualar las ecuaciones obtenidas a 0. Es decir, el problema se reduce a un problema de mínimos.

Así, obtendremos:

$$\frac{\partial D}{\partial a} = 2\sum_{i=1}^{n} (y_i - a - bx_i)(-1) = 0$$
$$\frac{\partial D}{\partial b} = 2\sum_{i=1}^{n} (y_i - a - bx_i)(-x_i) = 0$$

Adecuando convenientemente las ecuaciones anteriores, obtenemos:

$$\sum_{i=1}^{n} (y_i - a - bx_i) = 0$$
$$\sum_{i=1}^{n} (y_i - a - bx_i)(x_i) = 0$$

Operando y reorganizando términos, obtenemos las denominadas ***Ecuaciones Normales de Gauss***:

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$
$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Resolviendo el sistema, obtenemos las expresiones para a y b:

$$a = \bar{y} - b\bar{x}$$
$$b = \frac{S_{XY}}{s_X^2}$$

La interpretación de **a** y **b**, es análoga a la que comentábamos en el apartado 6.1.3.2, sólo que como ya dijimos entonces, **b** recibe el nombre de ***Coeficiente de Regresión*****.**

Como podemos observar, en el numerador de **b**, aparece la covarianza, y en el denominador la varianza de la variable independiente. Esto hace que el signo de **b** sea el mismo signo que el de la covarianza, por lo que si b>0, entonces, existe una relación directa entre las variables, y si b<0 entonces la relación es inversa.

En nuestro ejemplo de talla y edad, **b** sería el incremento medio que se produce en la talla, por cada incremento unitario de edad; si la edad está en años, por cada año aumente la edad.

Si queremos predecir un valor y_i a partir de un valor concreto de x_i, utilizaremos la expresión de la ecuación donde ahora ya, a y b son conocidos. No olvidemos que ese era uno de los objetivos del análisis, tratar de conocer valores de Y a partir de los de X:

$$y^*_i = a+bx_i$$

REPRESENTATIVIDAD DE LA RECTA DE REGRESIÓN.

- **Poder explicativo del modelo**

La recta de regresión, tiene carácter de línea media, como ya se ha señalado con anterioridad, tratando por lo tanto de resumir o sintetizar la información suministrada por los datos.

Si tiene carácter de linea media (de promedio, en definitiva), deberá ir acompañada *siempre* de una medida que nos hable de su representatividad, es decir, de lo buena que es la recta, ya que el haber obtenido la mejor de todas no da garantías de que sea buena.

Necesitamos, por tanto, una medida de dispersión, que tenga en cuenta la dispersión de cada observación con respecto a la recta, es decir, lo alejado que se encuentra cada punto de la recta.

Es decir, deberemos evaluar esas distancias verticales a la recta, es decir, los errores o residuales.

Si las dispersiones son pequeñas, la recta será un buen representante de la nube de puntos, o lo que es lo mismo, la ***bondad de ajuste del modelo será alta***. Si la dispersión es grande, la bondad de ajuste será baja.

Una forma de medir dicha bondad de ajuste es precisamente evaluando la suma de los cuadrados de los errores. Por tanto, llamaremos ***Varianza residual*** a la expresión:

$$S_e^2 = \frac{\sum_{i=1}^{n} \left(y_i - y_i^*\right)^2}{n}$$

Si la varianza residual es grande, el modelo será malo, es decir, la recta no explicará el comportamiento general de la nube.

La fórmula práctica para el cálculo de la varianza residual, si el procedimiento de ajuste es el de los mínimos cuadrados es la siguiente:

$$S_e^2 = \frac{\sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i y_i}{n}$$

La cota máxima de la varianza residual es la varianza que tratamos de explicar mediante el modelo de regresión, es decir, la varianza de la variable dependiente. Por tanto, sin más que hacer relativa la varianza residual respecto de su máximo valor, y multiplicando por 100, obtendremos el porcentaje de variaciones no explicado por el modelo:

$$\%\ \text{de variaciones sin explicar} = \frac{S_e^2}{s_y^2} 100$$

Ahora, ya es fácil obtener una media que nos indique el porcentaje de variaciones controladas o explicadas mediante el modelo, que se conoce como ***Coeficiente de Determinación***, que denotaremos con R^2. Su expresión en tantos por 1, será:

$$R^2 = 1 - \frac{S_e^2}{s_y^2}$$

Como puede observarse, a partir de la expresión anterior: $0 < R^2 < 1$. Por tanto:

Si $R^2=1$, entonces no hay residuos, habrá una dependencia funcional. Cuanto más se acerque dicho valor a la unidad, mayor ***poder explicativo*** tendrá el modelo de regresión.

Si R^2=0, X no explica en absoluto ninguna de las variaciones de la variable Y, de modo que o bien el modelo es inadecuado, o bien las variables son independientes. Cuanto más cercano a 0 esté dicho valor, menor poder explicativo.

- **Poder explicativo vs poder predictivo**

Un modelo de regresión con un alto porcentaje de variaciones explicado, puede no ser bueno para predecir, ya que el que la mayoría de los puntos se encuentren cercanos a la recta de regresión, no implica que todos lo estén, y puede ocurrir, que justamente para aquel rango de valores en el que el investigador está interesado, se alejen de la recta, y por tanto, el valor predecido puede alejarse mucho de la realidad.

La única forma de poder evaluar el poder predictivo del modelo es tras la observación y el análisis de los gráficos de residuales, es decir, de diagramas de dispersión, en los que en el eje de ordenadas se colocan los residuales, y en el eje de abscisas se colocan o bien X, Y, o Y*.

Sólo si la banda de residuales es homogénea, y se encuentran todos los puntos no demasiado alejados del 0 (aunque depende de la escala de medida), diremos, que un modelo con un alto poder explicativo, también es bueno para predecir.

Un análisis detallado de los residuales se realizará en la sección 6.2.

CAUSALIDAD

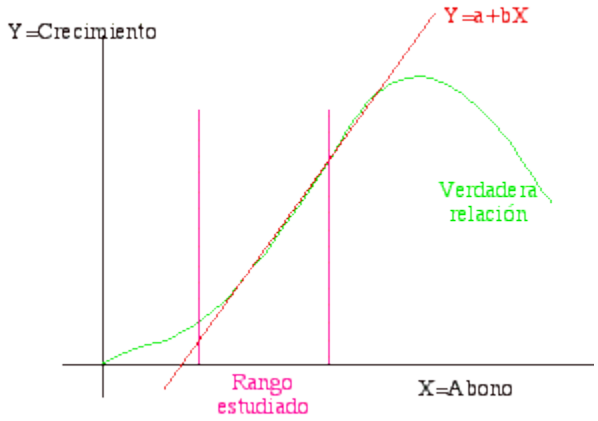
Es muy importante resaltar el hecho, de que un modelo sea capaz de explicar de manera adecuada las variaciones de la variable dependiente en función de la independiente, no implica que la primera sea causa de la segunda.

Es un error muy común confundir causalidad con *casualidad*. El hecho de que las variables estén relacionadas no implica que una sea causa de la otra, ya que puede ocurrir el hecho de que se esté dando una variación concomitante, por el simple hecho de que las dos son causa de una tercera. Por ejemplo, si realizamos un estudio en el que se analice el número de canas (X) y la presión arterial (Y), podríamos encontrar una relación lineal casi perfecta. Eso no significa que el tener canas aumente la presión arterial, lo que verdaderamente está ocurriendo es que es la edad, la causante, de que se tengan más canas y una tendencia a tener más alta la presión arterial.

EXTRAPOLACIÓN

Es importante, resaltar el hecho de que a la hora de hacer predicciones, no deben extrapolarse los resultados más allá del rango de la variable X utilizado para ajustar el modelo, ya que más allá de ese rango no sabemos qué puede estar ocurriendo.

Por todos es conocido que las plantas necesitan abono para poder crecer. Desde pequeños hemos aprendido que hay que abonarlas, de modo que en principio, cuanto más abono se les suministre más crecerán. Pero... ¿qué ocurriría si abonásemos demasiado el suelo?. Obviamente la planta moriría. Bien, esto se traduce, en que conforme aumenta la cantidad de abono, el crecimiento es más notable, pero a partir de un punto, la planta deja de crecer, y es más se muere. Esto queda reflejado en la figura 6.7. De ahí el peligro de extrapolar los resultados.


Y=Crecimiento
Y=a+bX
Verdadera relación
Rango estudiado
X=Abono