Analítica Predictiva

¿Qué es Analítica Predictiva?

- Proporciona herramientas para estimar aquellos datos de negocio que son desconocidos o inciertos, o que requieren de un proceso manual o costoso para su obtención.
- Más allá del puro análisis de la información histórica que realiza la analítica descriptiva, las **predicciones de datos** que realiza la analítica predictiva fortalecen las decisiones de negocio.
- Permite:
  - Anticipar comportamientos.
  - Detectar anomalías
  - Descubrir grupos afines

Herramientas: Segmentación

- Las **técnicas de segmentación** analizan grandes volúmenes de datos y detectan automáticamente grupos afines de elementos, o casos discordantes que no se ajusten a la dinámica estándar.
- Encuentra su fundamento en el **aprendizaje automático no supervisado**. Este transforma los datos en una representación numérica que permita calcular similitudes entre elementos, y así detectar grupos afines o elementos que sean muy diferentes de lo esperado.
- El resultado de un proceso de segmentación es un listado de los **grupos afines detectados** en los datos, para su posterior estudio por los analistas del cliente.
- La segmentación puede repetirse varias veces en el tiempo, a medida que se actualizan los datos bajo estudio, permitiendo así **descubrir cambios en los grupos identificados**, como puede ser la aparición de nuevos intereses en una base de clientes, o nuevas temáticas en un conjunto de documentos.

Algunas técnicas de segmentación

- Son muchas las técnicas que ayudan a segmentar y su uso está en función de lo que necesitemos analizar y conocer de los clientes.
- Se clasifican en 3 grupos por ser descriptivas, semidescriptivas y predictivas.
- **Técnicas descriptivas**
    - **Análisis de Frecuencia:** Miden la efectividad y evolución de la actividad de una actividad
    - **Tablas cruzadas**: Análisis de dos variables que describe relación y evolución de una variable en función de la otra.
    - **Clúster o agrupamiento**: Genera grupos de individuos lo más parecidos entre sí y al mismo tiempo lo más diferente posible de otros grupos. Una técnica multivariable cuyo objetivo es clasificar a una población amplia en un número de grupos pequeño, mutuamente excluyentes y exhaustivos, en base a las semejanzas y diferencias de los perfiles existentes entre los distintos elementos que componen la población.
        - No predice comportamiento ni diferencia entre “mejores y peores”
        - Se usa para determinar contenidos de comunicación sobre un grupo homogéneo.
- **Técnicas semidescriptivas**
    - **Chaid (CHi-square Automatic Interaction Detection):** Genera segmentos por variables más discriminantes y los diferencia de otros ayudando a identificar los mejores y lo peores.
- **Técnicas predictivas**
    - **Regresión múltiple:** Ayuda a predecir el comportamiento futuro de variables en función a la información acumulada.
    - Redes Neuronales, Inteligencia Articial.

Herramientas: Clasificación

- Se basa fuertemente el **aprendizaje automático supervisado**, que consiste en presentar diversos datos de clasificaciones pasadas, realizadas por expertos en el tema, a un sistema de clasificación.
- El resultado de todo este proceso de aprendizaje es un **sistema de clasificación** con la capacidad de procesar grandes volúmenes de datos, y dar tiempos de respuesta mucho más bajos que los necesarios para un tratamiento manual.
- La explotación de estos sistemas de clasificación se realiza mediante la aplicación de **reglas de negocio** sobre las clasificaciones generadas por los mismos.
- Para cada clasificación, el sistema genera un indicador de certidumbre o ***score*** que resume la confianza que el sistema tiene en su propia decisión.
- Las técnicas de clasificación automática permiten apoyar a los expertos y facilitar su labor para que puedan analizar un volumen mucho mayor de datos a un menor costo.

Algunas técnicas de clasificación

- La clasificación supervisada es una de las tareas que más frecuentemente son llevadas a cabo por los denominados Sistemas Inteligentes. Por lo tanto, un gran número de paradigmas desarrollados bien por la Estadística (Regresión Logística, Análisis Discriminante) o bien por la Inteligencia Artificial (Redes Neuronales, Inducción de Reglas, Árboles de Decisión, Redes Bayesianas) son capaces de realizar las tareas propias de la clasificación.
  - **Validación cruzada:** es una técnica utilizada para evaluar los resultados de un análisis estadístico cuando el conjunto de datos se ha segmentado en una muestra de entrenamiento y otra de prueba, la validación cruzada comprueba si los resultados del análisis son independientes de la partición. Aunque la validación cruzada es una técnica diseñada para modelos de regresión y predicción, su uso se ha extendido a muchos otros ejercicios de machine learning.
  - **Análisis Discriminante (AD):** introducido por Fisher (1936), se utiliza para predecir la pertenencia a un grupo (variable dependiente) a partir de un conjunto de predictores (variables independientes). El objetivo del AD es entender las diferencias de los grupos y predecir la verosimilitud de que una persona o un objeto pertenezca a una clase o grupo basándose en los valores que toma en los predictores.
  - **Algorítmo K-vecinos más cercanos: (**K nearest neighbors Fix y Hodges, 1951) es un método de clasificación. Es un método de clasificación no paramétrico, que estima el valor de la función de densidad de probabilidad o directamente la probabilidad a posteriori de que un elemento xx pertenezca a la clase CjCj a partir de la información proporcionada por el conjunto de prototipos o ejemplos. En el proceso de aprendizaje no se hace ninguna suposición acerca de la distribución de las variables predictoras.
  - **Arboles de clasificación:** Es un modelo no paramétrico, por tanto, no hay que estimar un modelo estadístico formal, son algoritmos para clasificar utilizando particiones sucesivas. Son apropiados cuando hay un número elevado de datos, siendo una de sus ventajas su carácter descriptivo que permite entender e interpretar fácilmente las decisiones tomadas por el modelo, revelando formas complejas en la estructura de datos que no se pueden detectar con los métodos convencionales de regresión.
  - **Máquinas de Soporte Vectorial:** (Support Vector Machines SVMs) son un conjunto de algoritmos de aprendizaje supervisados que desarrollan métodos relacionados con los problemas de clasificación y regresión.
  - **Red Neuronal Artificial** (RNA) es un modelo matemático inspirado en el comportamiento biológico de las neuronas y en cómo se organizan formando la estructura del cerebro. Las redes neuronales intentan aprender mediante ensayos repetidos como organizarse mejor a si mismas para conseguir maximizar la predicción.
  - **Naïve Bayes:** es uno de los clasificadores más utilizados por su simplicidad y rapidez. Se trata de una técnica de clasificación y predicción supervisada que construye modelos que predicen la probabilidad de posibles resultados, en base al Teorema de Bayes, también conocido como teorema de la probabilidad condicionada:

Herramientas: Predicción

- Es la capacidad de **anticiparse a futuros cambios** e incertidumbres en el negocio es fundamental para garantizar el éxito. Las técnicas de predicción tienen como objetivo acotar estas incertidumbres.
- Con procedimientos semejantes a los de la clasificación automática, las técnicas de predicción se basan en la disciplina científica conocida como **aprendizaje automático supervisado** y otras relacionadas, como el estudio de series temporales.
- Estas técnicas analizan históricos de las métricas que hay que predecir y, mediante técnicas estadísticas, las relacionan con factores externos que puedan influenciarlas. De este modo se construye un **modelo predictivo** capaz de estimar los valores más probables que tendrán estas métricas en el futuro próximo.
- De forma similar a los sistemas de clasificación, un **modelo de regresión** es una solución viva, que puede realimentarse de forma continua con nuevos datos a medida que se van obteniendo, reajustándose así de manera automática a cambios en tendencias o fenómenos inesperados.

Pronósticos

Predicción, Pronóstico y Prospectiva

- **Predicción:** estimación de un acontecimiento futuro que se basa en consideraciones subjetivas, en la habilidad, experiencia y buen juicio de las personas.
- **Pronóstico:** estimación de un acontecimiento futuro que se obtiene proyectando datos del pasado que se combinan sistemáticamente, aplicando técnicas estadísticas y de la ciencia administrativa.
- **Prospectiva:** conjunto de "tentativas sistemáticas para observar a largo plazo el futuro de la ciencia, la tecnología, la economía y la sociedad con el propósito de identificar las tecnologías emergentes que probablemente produzcan los mayores beneficios económicos y/o sociales". Es la ciencia que estudia el futuro para comprenderlo y poder influir en él

Uso de pronósticos

- Su éxito está basado en su aplicación efectiva en la planificación y toma de decisiones
- Los pronósticos son importantes para diferentes aspectos de la planeación, incluyendo aspectos tales como diseño del producto, diseño del proceso, inversión y reemplazo de equipo y planificación de la capacidad.
- Es además una herramienta para el control porque permite definir estándares para comparar.

Clasificación de enfoques

• **Intuitivo**: estimación de un evento futuro para una fecha posible. Implica hacer conjeturas, corazonadas y juicios subjetivos	Método Delphi, tormenta de ideas, grupo nominal, tanque de ideas, etc.
• **Formales:** estimación de cantidades basadas en técnicas estadísticas y datos anteriores.	Series de tiempo, métodos causales, simulación.

METODO	DESCRIPCION	HORIZONTE	COSTO
	Métodos Cualitativos		
Fuerza de ventas	Estimación del área de ventas como un todo	Corto y Mediano	Bajo, Medio
Opinión Ejecutiva	Gerentes de mercadotecnia, finanzas y producción preparan un pronóstico	Corto y Largo	Bajo, Medio
Venas y Gerentes	Los cálculos independientes de los vendedores son canalizados con proyecciones de los gerentes	Medio	Medio
Analogía histórica	Pronóstico proveniente de la comparación con un producto similar previamente introducido	Corto, Largo	Bajo, Medio
Delphi	Los expertos responden (anónimamente) una serie de preguntas, reciben retroalimentación y revisan sus cálculos.	Largo	Medio, Alto
Investigaciones de Mercado	Se usan cuestionarios y paneles para obtener datos que anticipen el comportamiento del consumidor.	Largo, Mediano y Corto	Medio, Alto
	Métodos Cuantitativos Series de Tiempo		
Promedio Simple	Se usa una regla simple que pronostica igual al último valor o igual más o menos algún porcentaje.	Corto	Bajo
Promedios móviles	El pronóstico es simplemente un promedio de los n más recientes	Corto	Bajo
Proyección de la tendencia	El pronóstico es una proyección lineal, exponencial u otra de la tendencia pasada.	Mediano, Largo	Bajo
Descomposición Estacional	Las series de tiempos se dividen en sus componentes de tendencia: estacional	Corto, Largo	Bajo
Suavización exponencial	Los pronósticos son promedios móviles ponderados exponencialmente, donde los últimos valores tienen mayor pesos	Corto	Bajo
	Métodos Cuantitativos Causales		
Regresión y correlación	Se usan una o más variables asociadas para pronosticar por medio de la ecuación de mínimos cuadrados (regresión) o de una asociación (correlación) con una variable explicativa	Corto, Mediano	Medio, Alto
Econométricos	Se usa una solución por ecuaciones simultáneas de regresión múltiple para una actividad económica	Corto, Largo	Alto

Los pronósticos en el ciclo de vida

DATOS DISPONIBLES: No hay
HORIZONTE DE TIEMPO Es necesario un horizonte largo
METODOS RECOMENDADOS Opinión y Juicio, Delphi.- Investigación de mercado.

DATOS Disponibles algunos
TIEMPO Sigue siendo necesario un horizonte largo; las tendencias y relaciones causa-efecto son importantes
METODOS Investigación de mercado: comparación histórica. Regresión y simulación computarizada

DATOS Considerable cantidad, de todo tipo.
TIEMPO Pronósticos a corto plazo
METODOS Series de tiempo, Regresión, suavización exponencial, econométricos

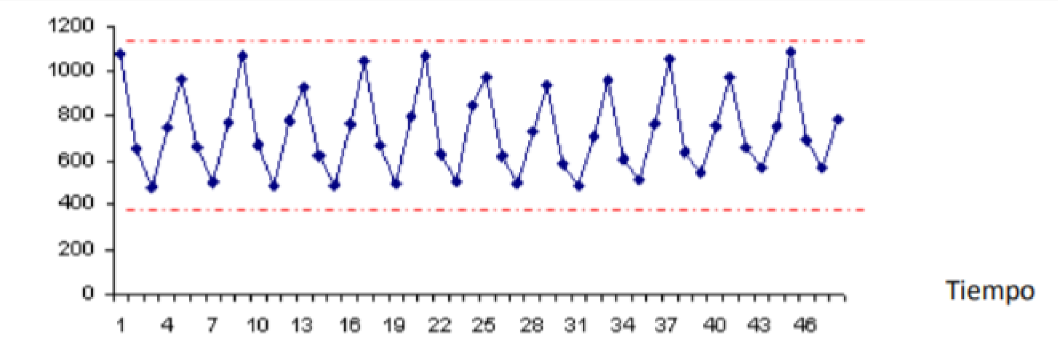
DATOS Abundantes
TIEMPO Horizonte reducido
METODOS Los mismos que para la madurez. Además el juicio, las analogías históricas y las investigaciones de mercado pueden señalar cambios.

Métodos formales

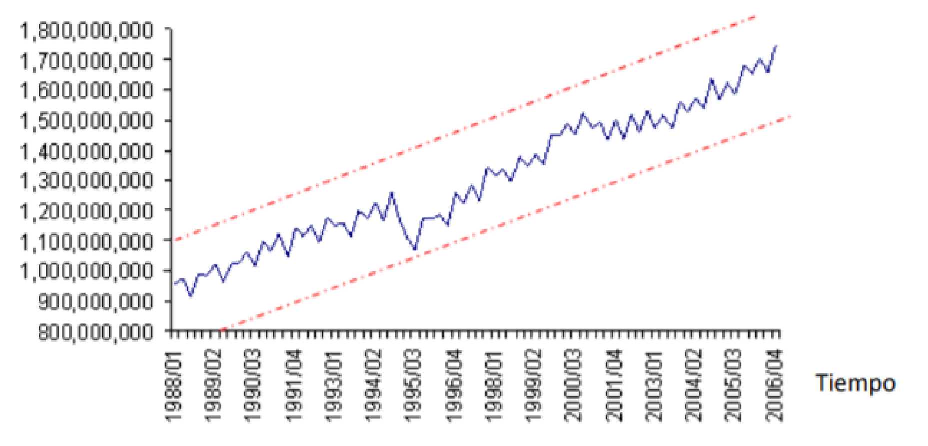
- Series de tiempo: es simplemente una lista cronológica de datos históricos, para la que la suposición esencial es que la historia predice el futuro de manera razonable
  - Promedio simple
  - Promedio móvil
  - Suavización exponencial
  - Regresión simple

Patrones de datos

- Patrón de horizontal (estacionario) Cuando los valores de una variable oscilan a través del tiempo alrededor de un nivel constante o medio, existe un patrón horizontal (estacionario)

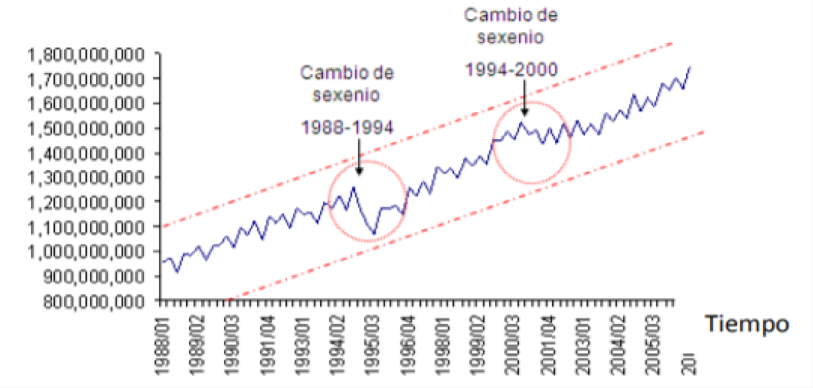


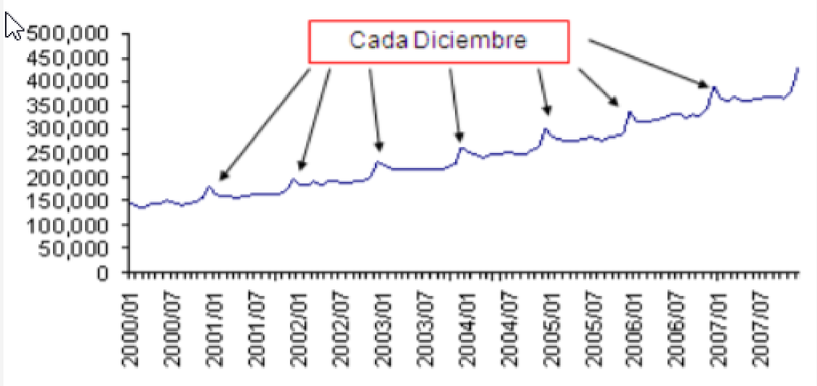
- Tendencias: existen cuando las series crecen o decrecen consistentemente sobre un largo periodo de tiempo.

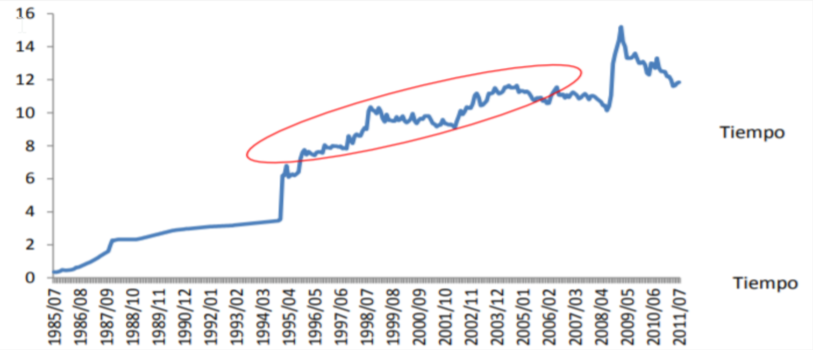


Patrones de datos

- **Patrones cíclicos**: Cuando una serie tiene tendencia, se puede observar un patrón adicional, un crecimiento o decrecimiento constante cada cierto tiempo (tres años o más); este patrón es el comportamiento cíclico.
- **Estacionalidad**: Cuando una serie se ve influenciada por factores que se repiten en la misma temporada del año, se dice que tiene un patrón estacional. Sólo puede existir en series que se miden con una frecuencia mensual, bimestral, trimestral, etc., pero no anual.
- **Aleatoriedad:** está en toda serie de tiempo; son cambios en la serie de corto plazo que por su aleatoriedad son difíciles modelarlos matemáticamente y por consecuencia no se proyectan al futuro para realizar pronósticos.







Promedio simple

- Todos los datos de los períodos anteriores tienen el mismo peso relativo. El promedio hace que los datos de mayor valor tiendan a ser equilibradas por los valores menores de otros períodos, reduciendo las posibilidades de error que se podrían cometer al dejarse llevar por fluctuaciones aleatorias que pueden ocurrir en un período.
- Se calcula en base a la expresión:

$$P = \frac{\sum_{i=1}^{k} d_i}{k}$$

- En donde,
  - d_i, i = 1, ... k, es la demanda de **todos** los períodos anteriores
  - k = número de períodos

Promedio móvil simple

- Combina los datos de demanda de la mayor parte de los períodos recientes, siendo su promedio el pronóstico para el período siguiente.
- El promedio se "mueve" en el tiempo, en el sentido de que, al transcurrir un período, la demanda del período más antiguo se descarta y se agrega, en su reemplazo, la demanda para el período más reciente, superando así la principal limitación del modelo del promedio simple.
- Se calcula como sigue:

$$MMS = \frac{\sum_{i=k-n+1}^{n} d_i}{n}$$

- Donde:
  - di es la demanda de cada uno de los n períodos anteriores. En este caso i va desde 1 hasta "n" períodos.
  - Si n =k, se tendrá el promedio simple.

Suavización exponencial de primer orden

- Se distingue porque da pesos de manera exponencial a cada una de las demandas anteriores a efectos de calcular el promedio.
- La demanda de los períodos más recientes recibe un peso mayor; los pesos de los períodos sucesivamente anteriores decaen de una manera no lineal (exponencial).
- El cálculo correspondiente requiere de 2 datos: el primero es la demanda real del período más reciente y el segundo es el pronóstico más reciente obtenido por cualquier otro método.
- A medida que termina cada período se realiza un nuevo pronóstico. Entonces:

$$\begin{array}{l} \text{Pronóstico de} \\ \text{la demanda del} \\ \text{período siguiente} \end{array} = \alpha \begin{bmatrix} \text{demanda} \\ \text{más} \\ \text{reciente} \end{bmatrix} + (1 - \alpha) \begin{bmatrix} \text{pronóstico} \\ \text{más} \\ \text{reciente} \end{bmatrix}$$

$$F_t = \alpha\, D_{t-1} + (1 - \alpha)\, F_{t-1}$$

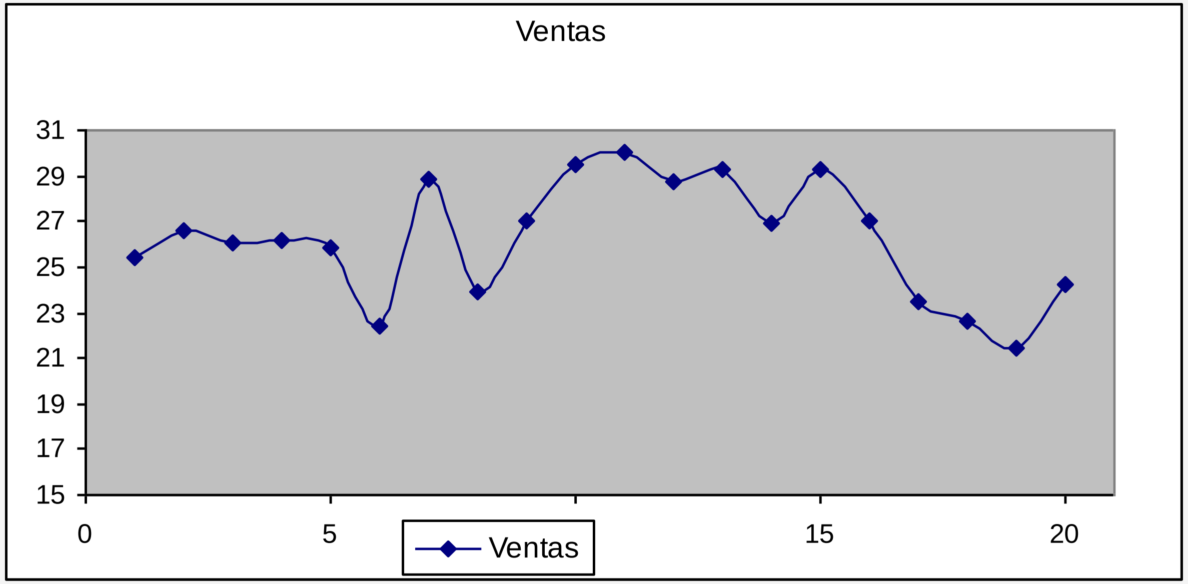
- Donde α es el coeficiente de suavización tal que $0 \leq \alpha \leq 1$

Selección de α

- Un elevado α sería más adecuado para los nuevos productos o para casos para los que la demanda subyacente está en proceso de cambio (esta es dinámica, o bien inestable). Un valor de α de 0.7, 0.8 o 0.9 puede resultar el más apropiado para estas condiciones, aun cuando el uso del suavizado exponencial es cuestionable si no se sabe si existen o no condiciones de inestabilidad.
- Si los datos son estables y se piensa que pueden ser representativos del futuro, el pronosticador podrá optar por un valor bajo de α para disminuir cualquier ruido que hubiera podido presentarse en forma súbita. Entonces, el procedimiento de pronóstico no reacciona de una manera drástica a las demandas más recientes. En estas condiciones de estabilidad, el coeficiente de suavización podría ser de 0.1, 0.2, o 0.3.
- Cuando la demanda es ligeramente inestable, coeficientes de suavización de 0.4, 0.5 o 0.6, pueden proporcionar los pronósticos más precisos.

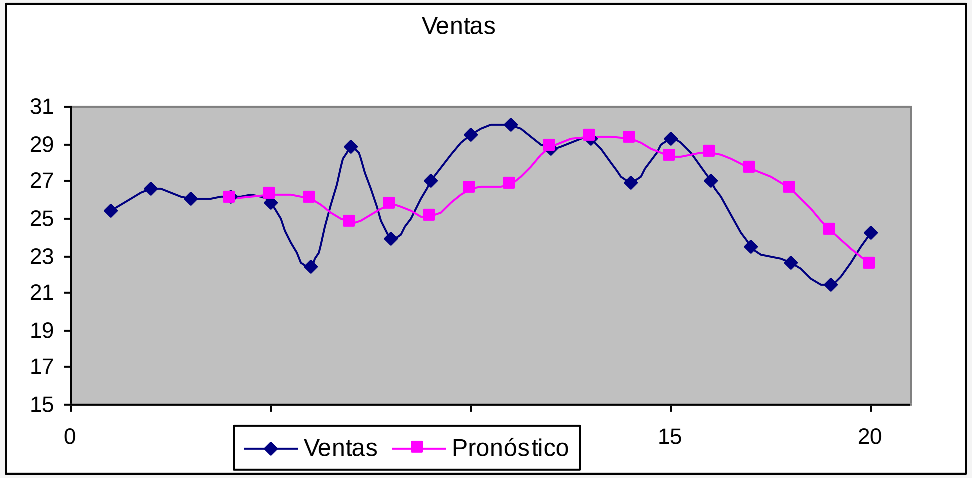
Ejemplo

Período	Ventas
1	25
2	27
3	26
4	26
5	26
6	22
7	29
8	24
9	27
10	29
11	30
12	29
13	29
14	27
15	29
16	27
17	23
18	23
19	21
20	24

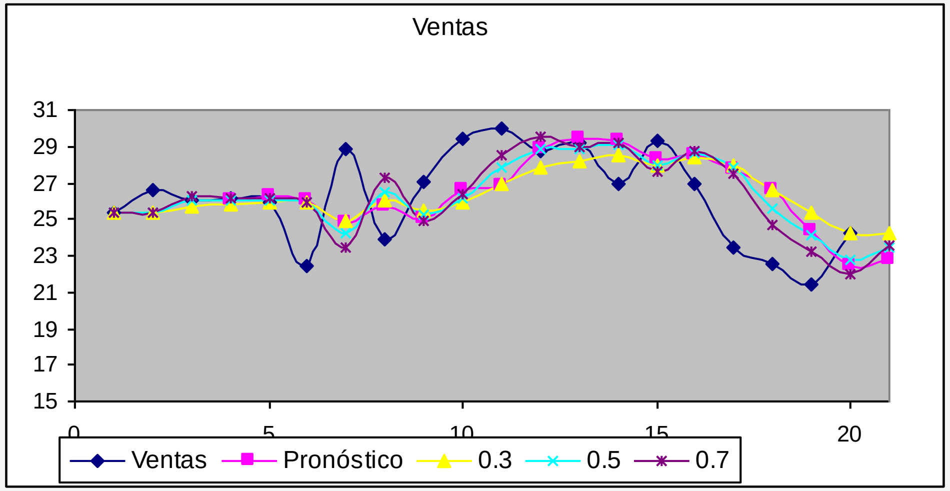


Pronóstico para el período 21
Promedio simple: 26.2 unidades
Promedio móvil con 3 períodos: 22.7 unidades

Período	Ventas	Pronóstico
1	25	
2	27	
3	26	
4	26	26.0
5	26	26.3
6	22	26.0
7	29	24.8
8	24	25.7
9	27	25.1
10	29	26.6
11	30	26.8
12	29	28.8
13	29	29.4
14	27	29.3
15	29	28.3
16	27	28.5
17	23	27.7
18	23	26.6
19	21	24.3
20	24	22.5
21		22.7



	Valor de α		
	0.3	0.5	0.7
1	25	25	25
2	25.4	25.4	25.4
3	25.8	26.0	26.3
4	25.8	26.0	26.1
5	25.9	26.1	26.1
6	25.9	26.0	26.0
7	24.9	24.2	23.5
8	26.1	26.5	27.2
9	25.4	25.2	24.9
10	25.9	26.1	26.4
11	27.0	27.8	28.5
12	27.9	28.9	29.5
13	28.1	28.8	29.0
14	28.5	29.0	29.2
15	28.0	28.0	27.6
16	28.4	28.6	28.8
17	28.0	27.8	27.5
18	26.6	25.6	24.7
19	25.4	24.1	23.2
20	24.2	22.8	22.0
21	24.2	23.5	23.5

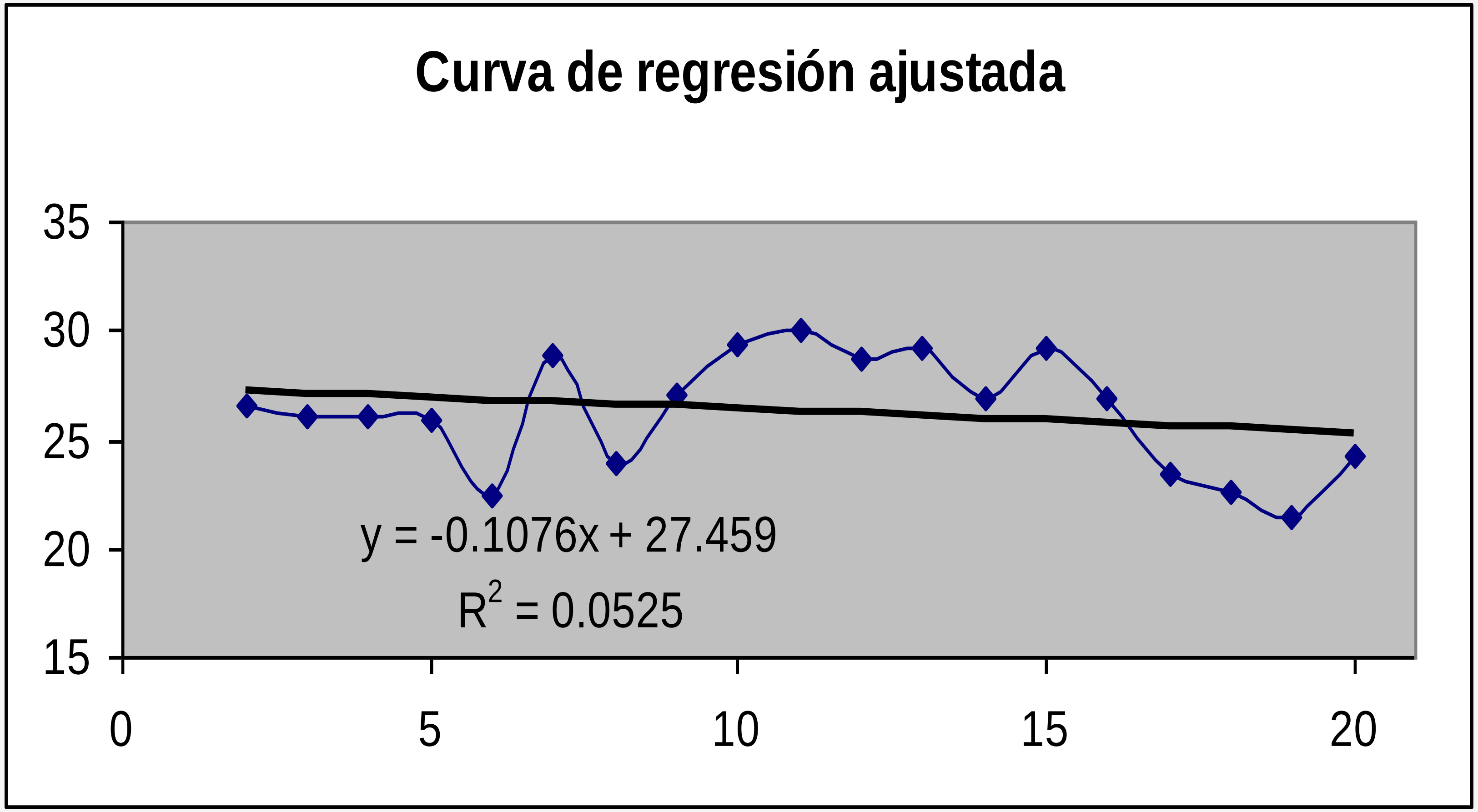


Regresión simple

- De la forma y = ax + b, donde x es función del período solamente.

$$b = \frac{n\left(\sum_i x_i y_i\right) - \left(\sum_i x_i\right)\left(\sum_i y_i\right)}{n\left(\sum_i x_i^2\right) - \left(\sum_i x_i\right)^2}$$

$$a = \frac{\sum_i y_i - b\sum_i x_i}{n}$$

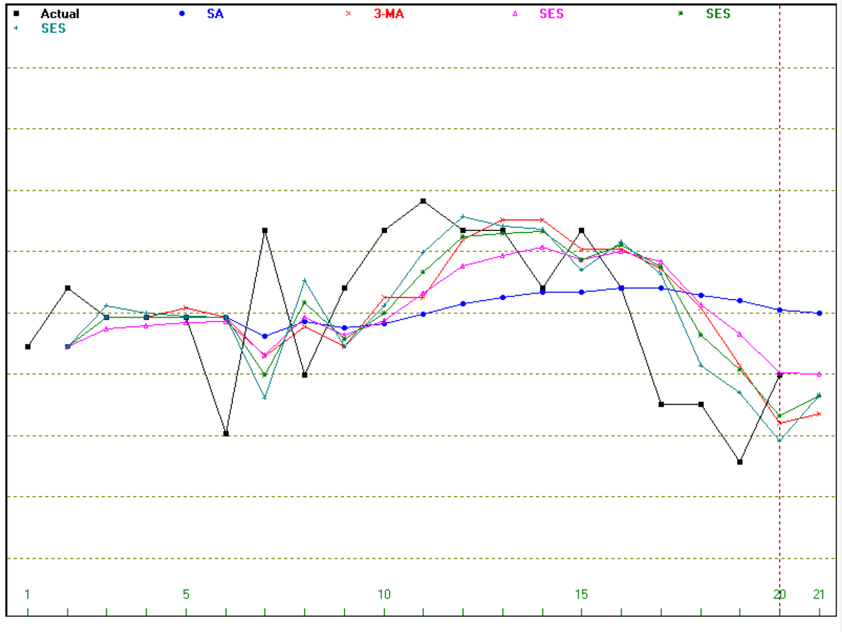

Curva de regresión ajustada
35
30
25
20
15
0
5
10
15
20
y = -0.1076x + 27.459
R2 = 0.0525

Resumen de valores obtenidos

Método	Pronóstico, período 21
Promedio simple	26.2
Promedio móvil 3 períodos	22.7
Suavización exponencial $\alpha = 0.3$	24.2
Suavización exponencial $\alpha = 0.5$	23.5
Suavización exponencial $\alpha = 0.7$	23.5
Regresión simple	25.2

. Forecast Result for Ejemplo

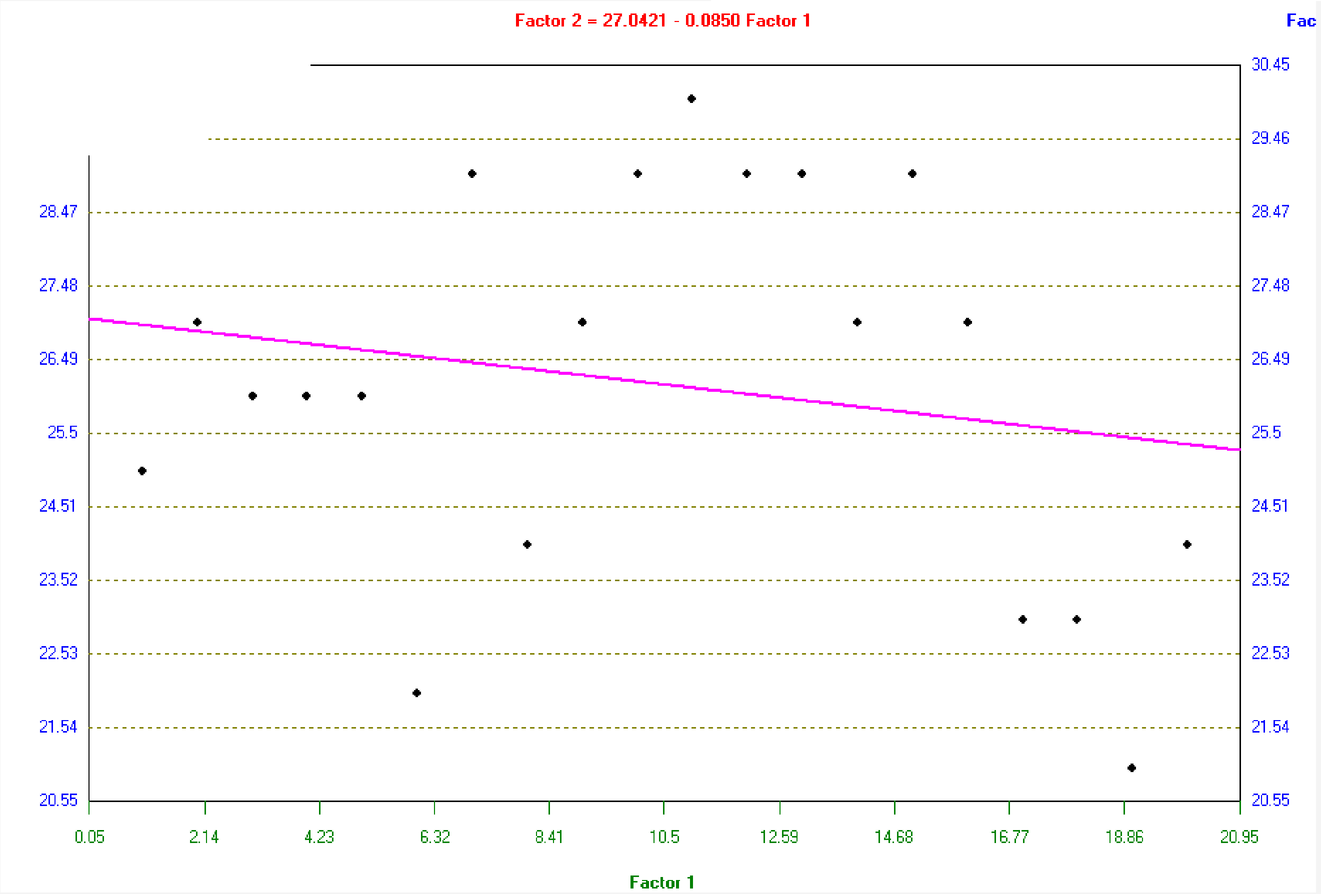
3-07-2009 Periodos	Actual Data	Forecast by SA	Forecast by 3-MA	Forecast by SES	Forecast by SES	Forecast by SES
1	25					
2	27	25		25	25	25
3	26	26		25.6	26	26.4
4	26	26	26	25.72	26	26.12
5	26	26	26.33333	25.804	26	26.036
6	22	26	26	25.8628	26	26.0108
7	29	25.33333	24.66667	24.70396	24	23.20324
8	24	25.85714	25.66667	25.99277	26.5	27.26097
9	27	25.625	25	25.39494	25.25	24.97829
10	29	25.77778	26.66667	25.87646	26.125	26.39349
11	30	26.1	26.66667	26.81352	27.5625	28.21805
12	29	26.45455	28.66667	27.76947	28.78125	29.46541
13	29	26.66667	29.33333	28.13863	28.89063	29.13962
14	27	26.84615	29.33333	28.39704	28.94531	29.04189
15	29	26.85714	28.33333	27.97793	27.97266	27.61257
16	27	27	28.33333	28.28455	28.48633	28.58377
17	23	27	27.66667	27.89918	27.74316	27.47513
18	23	26.76471	26.33333	26.42943	25.37158	24.34254
19	21	26.55556	24.33333	25.4006	24.18579	23.40276
20	24	26.26316	22.33333	24.08042	22.5929	21.72083
21		26.15	22.66667	24.05629	23.29645	23.31625
CFE		-0.1011848	-6.666668	-3.145695	-3.407104	-2.405363
MAD		2.251576	2.117647	2.081467	1.950382	2.008023
MSE		7.740797	6.640524	6.639411	6.098916	6.325839
MAPE		8.926395	8.452552	8.261073	7.743074	7.91053
rk.Signal		-4.493956E-02	-3.148149	-1.511287	-1.746891	-1.197876
R-square		4.456114E-02	0.4624788	0.2460335	0.4510994	0.6498051
			m=3	Alpha=0.3	Alpha=0.5	Alpha=0.7
				F(0)=25	F(0)=25	F(0)=25



Resultados utilizando WinQSB

03-07-2009 16:53:59	Dependent Variable	Independent Variable
Equation:	Factor 2 =	27.0421 - 8.496241E-02 Factor 1

03-07-2009	Variable/Item	Prediction and Values
1	Prediction for Factor 2	25.25789
2	Standard Deviation of Prediction	1.237538
3	Prediction Interval	[19.08318, 31.43261]
4	Confidence Interval of Prediction Mean	[22.65652, 27.85927]
5	Significance Level (alpha)	5%
6	Degree of Freedom	18
7	t Critical Value	2.102055
8	Factor 1	21



¿Cuál es el mejor modelo?

- El mejor indicador de un pronóstico es la precisión del método.
- Medidas de error
  - Error promedio
  - Error medio absoluto (MAD: mean absolute deviatio)
  - Promedio del error cuadrado (MSD: mean square deviation)
  - Error absoluto medio porcentual (MAPE: mean abslute percenage error)

Error promedio

- Se calcula como la diferencia entre los datos observados y el pronóstico. Debido al teorema del límite central, debe dar siempre un valor cercano a cero.

Desviación media absoluta

- A fin de evitar el problema del error promedio, se utiliza el promedio de la desviación media absoluta:

$$\frac{\sum \left| x_i - F_i \right|}{n}$$

Promedio de error cuadrado

- Penaliza más las desviaciones grandes

$$\frac{\sum \left| x_i - F_i \right|^2}{n}$$

Error absoluto medio porcentual

- También elimina el problema del signo. Otra ventaja es que permite comparación por ser un valor relativo, no absoluto.

$$PF_i = \frac{x_i - F_i}{x_i} \times 100$$

$$\frac{\sum \left| PF \right|}{n}$$

Métodos causales

- Muestran relación causa efecto
  - Regresión simple
  - Regresión múltiple
  - Box-Jenkins (ARIMA)

Análisis de Regresión

Introducción

- Tiene como objetivo modelar en forma matemática el comportamiento de una variable de respuesta en función de una o más variables independientes (factores).
- Para estimar los parámetros de un *modelo de regresión* son necesarios los datos.
- Estos pueden obtenerse de experimentos planeados, de observaciones de fenómenos no controlados o de registros históricos.

Relaciones entre variables aleatorias y regresión lineal.

- El término regresión fue introducido por Galton en su libro "*Natural inheritance*" (1889) refiriéndose a la "ley de la regresión universal":
  - "Cada peculiaridad en un hombre es compartida por sus descendientes, pero en media, en un grado menor."
    - Regresión a la media
  - Su trabajo se centraba en la descripción de los rasgos físicos de los descendientes (una variable) a partir de los de sus padres (otra variable).
  - Pearson (un amigo suyo) realizó un estudio con más de 1000 registros de grupos familiares observando una relación del tipo:
    - Altura del hijo = 85cm + **0,5** • altura del padre (aprox.)
    - Conclusión: los padres muy altos tienen tendencia a tener hijos que heredan parte de esta altura, aunque tienen tendencia a acercarse (*regresar*) a la media. Lo mismo puede decirse de los padres muy bajos.
- Hoy en día el sentido de regresión es el de predicción de una medida basándonos en el conocimiento de otra.

Regresión lineal simple

- Sean dos variables *X* y *Y*, suponga que se quiere explicar el comportamiento de *Y.*
- *Y* se le llama la *variable dependiente* o la *variable de respuesta* y a *X* se le conoce como *variable independiente* o *variable regresora*.
- La variable *X* no necesariamente es aleatoria, ya que en muchas ocasiones el investigador fija sus valores. *Y* siempre es una variable aleatoria.
- Una manera de estudiar el comportamiento de *Y* con respecto a *X* es mediante un modelo de regresión que consiste en ajustar un modelo matemático de la forma: $Y = f(X)$ a la pareja de puntos, con base en los valores que toma *X*.

El modelo de regresión:

- Suponga que las variables *X* y *Y* están relacionadas linealmente y que para cada valor de *X*, la variable dependiente, *Y*, es una variable aleatoria.
- Es decir, que cada observación de *Y* puede ser descrita por el modelo:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- donde ε es el error aleatorio normalmente distribuido con media cero y varianza σ^2. También suponga que los errores aleatorios no están correlacionados.
- El modelo se conoce como regresión lineal simple.

Cálculo de los parámetros:

- Mediante mínimos cuadrados:

$$S=\sum_{i=1}^{n}(\varepsilon_i)^2=\sum_{i=1}^{n}\left(y_i-[\beta_0+\beta_1 x_i]\right)^2$$

- Derivando con respecto β_0 y β_1

$$\frac{\partial S}{\partial \beta_0}=-\sum_{i=1}^{n}2\left(y_i-[\beta_0+\beta_1 x_i]\right)$$

$$\frac{\partial S}{\partial \beta_1}=-\sum_{i=1}^{n}2x_i\left(y_i-[\beta_0+\beta_1 x_i]\right)$$

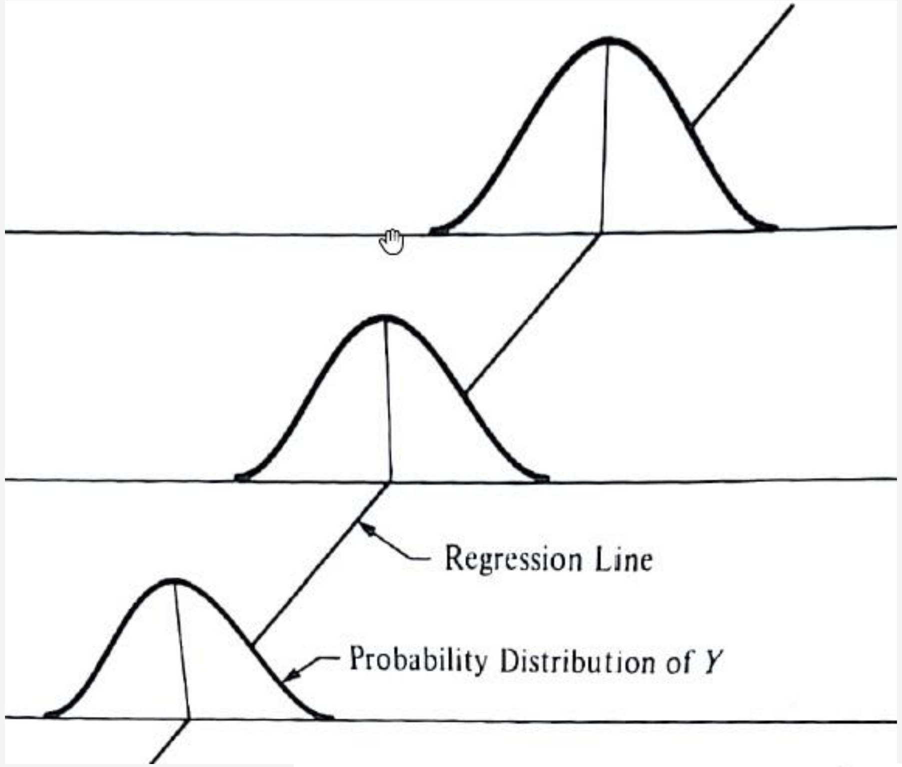
- Igualando a ce

$$\hat{\beta}_1=\frac{S_{xy}}{S_{xx}}$$

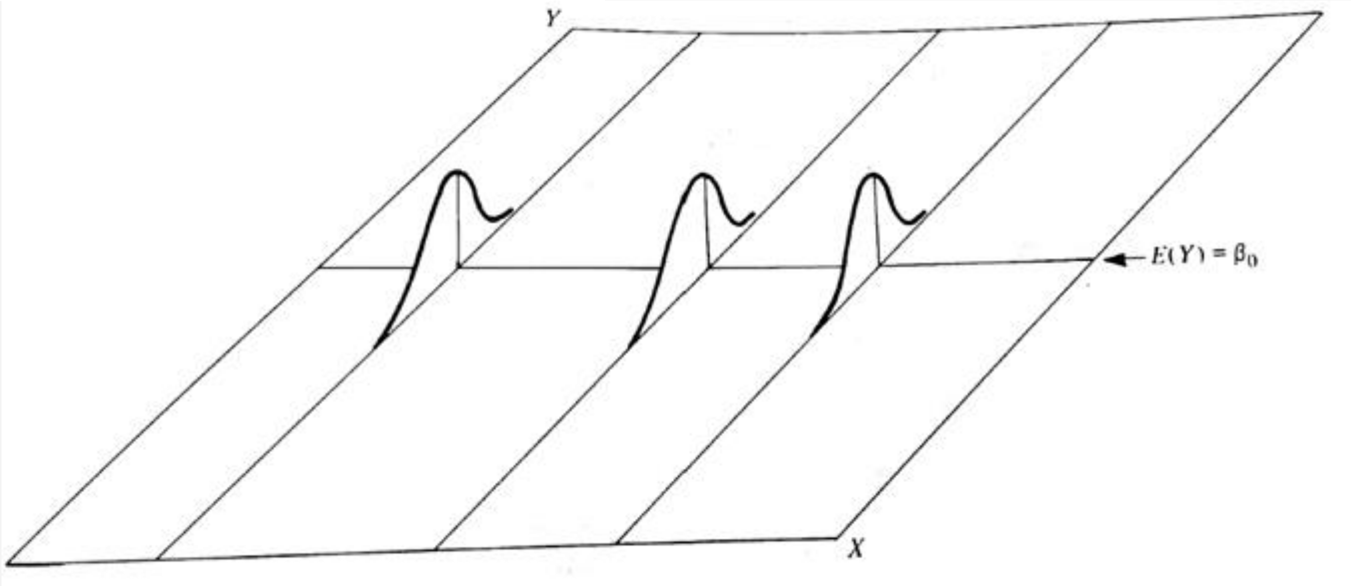
$$\hat{\beta}_0=\overline{y}-\hat{\beta}_1\overline{x}$$

$$S_{xy}=\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})=\sum_{i=1}^{n}x_i y_i-\frac{(\Sigma_{i=1}^{n}x_i)(\Sigma_{i=1}^{n}y_i)}{n}$$

$$S_{xx}=\sum_{i=1}^{n}(x_i-\overline{x})^2=\sum_{i=1}^{n}x_i^2-\frac{(\Sigma_{i=1}^{n}x_i)^2}{n}$$

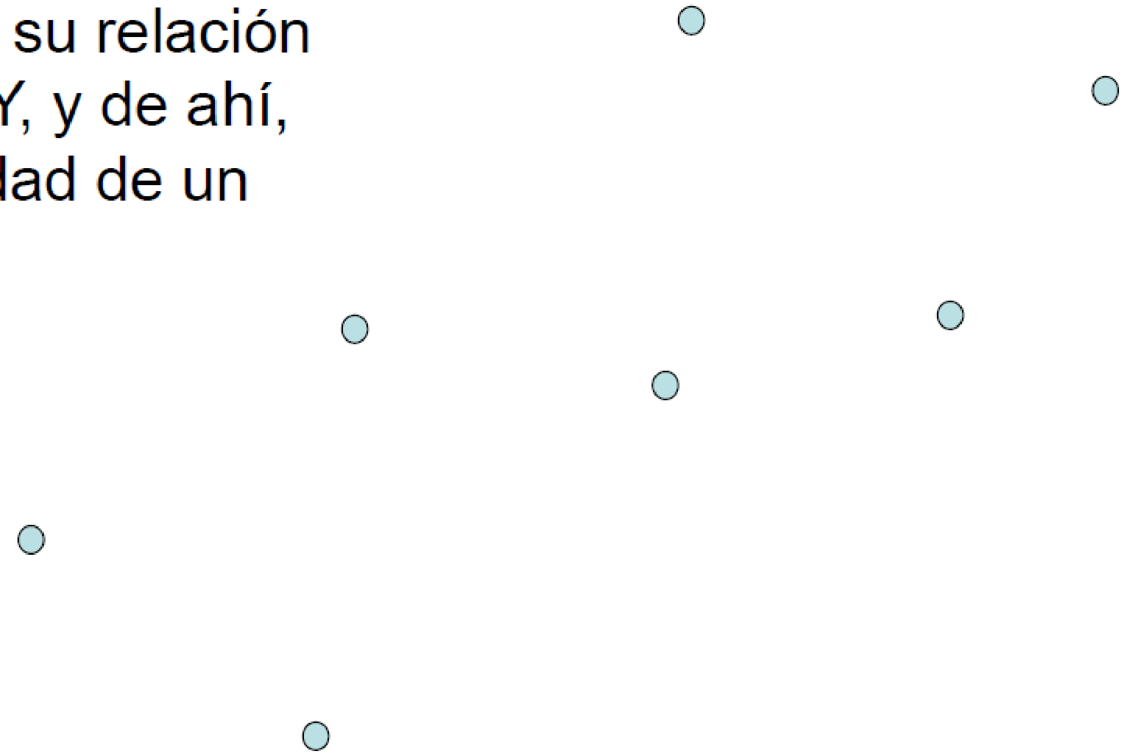


Comportamiento de la línea de regresión

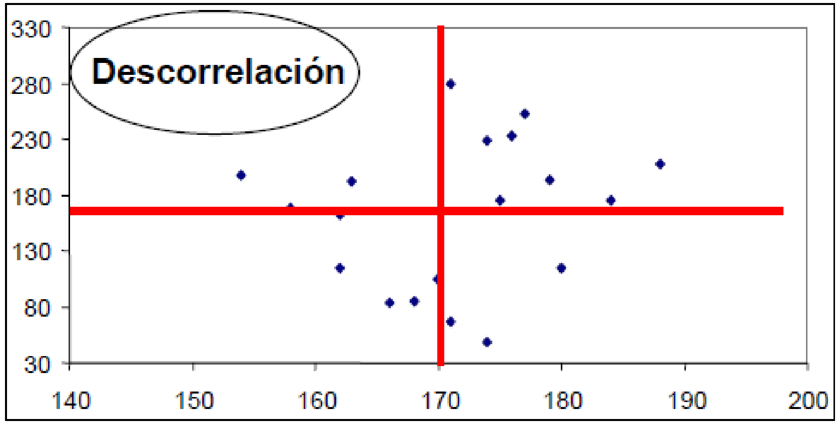


¿Qué tan buena es la regresión?

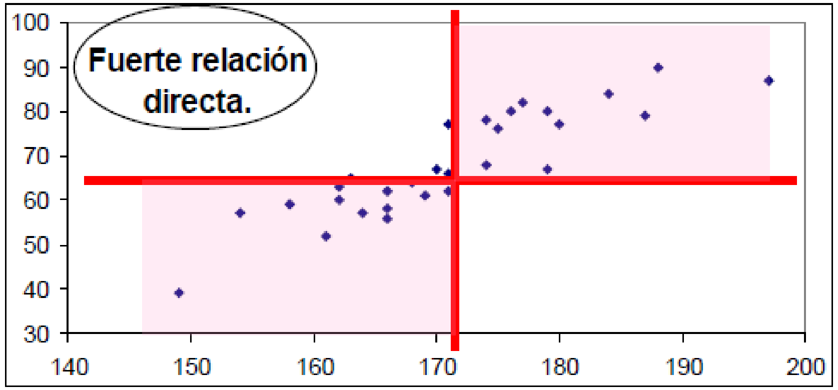
Imaginemos un diagrama de dispersión, y vamos a tratar de comprender en primer lugar qué es el error residual, su relación con la varianza de Y, y de ahí, cómo medir la bondad de un ajuste.



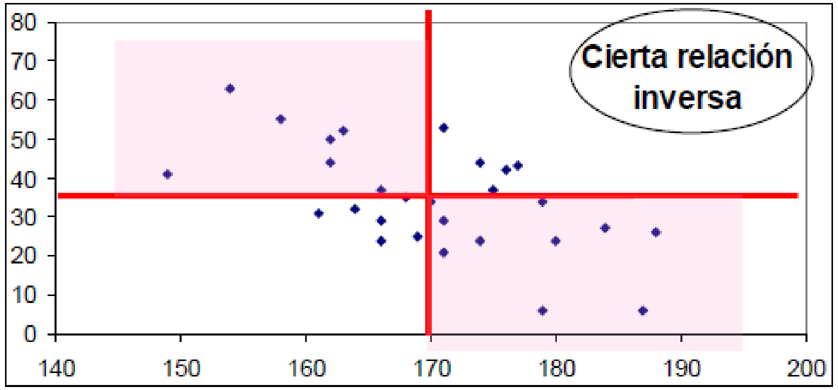
Reconociendo relaciones



Para valores de X por encima de la media tenemos valores de Y por encima y por debajo en proporciones similares: Descorrelación.

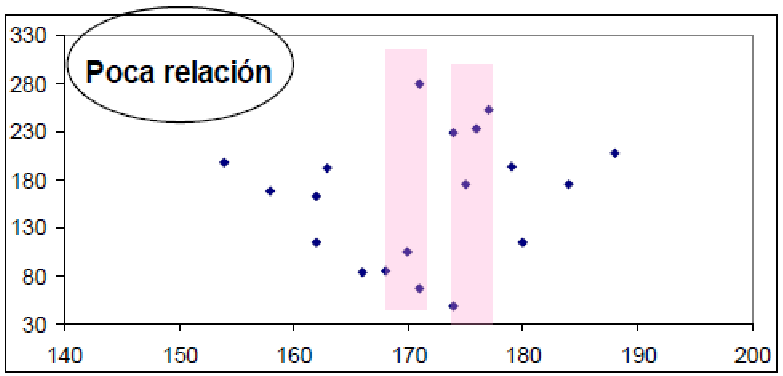


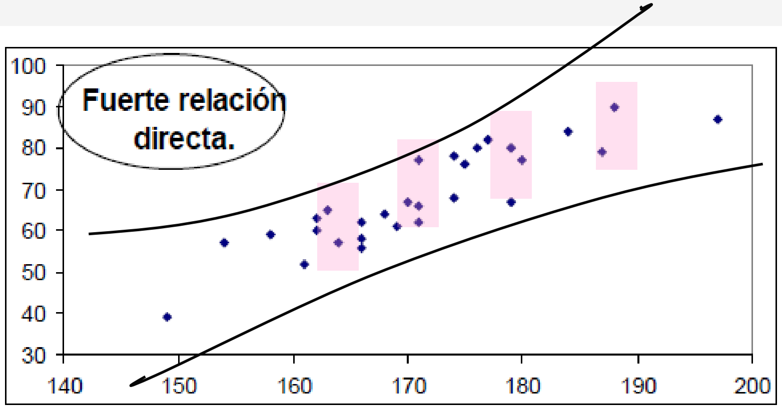
- Para los valores de X mayores que la media le corresponden valores de Y mayores también.
- Para los valores de X menores que la media le corresponden valores de Y menores también.
- Esto se llama relación directa o creciente entre X e Y.



Para los valores de X mayores que la media le corresponden valores de Y menores. Esto es relación inversa o decreciente.

Buena o mala relación

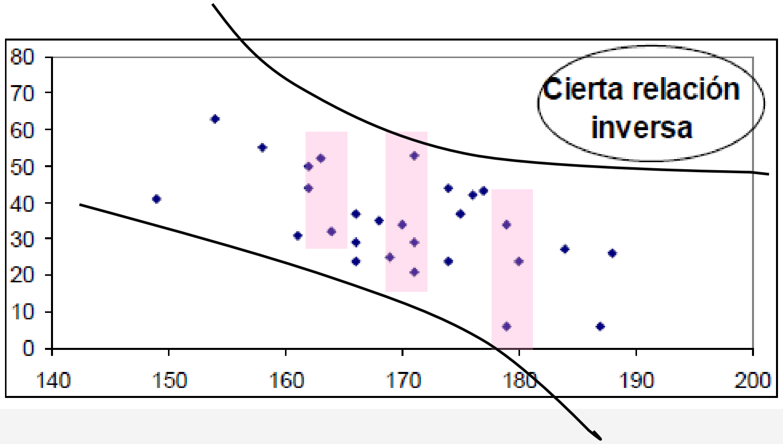




Dado un valor de X no podemos decir gran cosa sobre Y. Mala relación. Independencia.

- Conocido X sabemos que Y se mueve por una horquilla estrecha. Buena relación.

- Lo de “horquilla estrecha” hay que entenderlo con respecto a la dispersión que tiene la variable Y por si sola, cuando no se considera X.

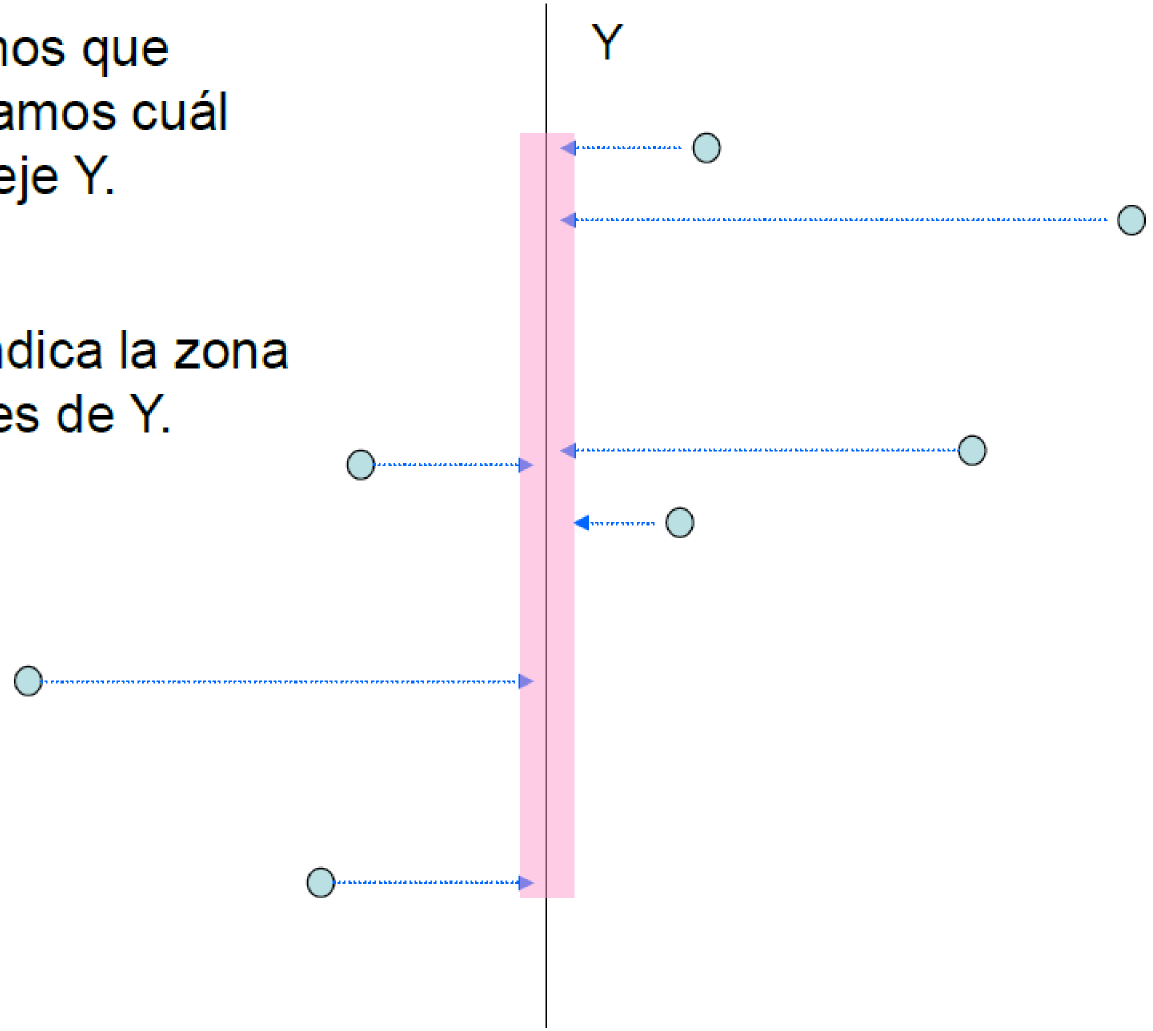


Interpretación de la variabilidad de Y

En primer lugar olvidemos que existe la variable X. Veamos cuál es la variabilidad en el eje Y.

La franja sombreada indica la zona donde varían los valores de Y.

Proyección sobre el eje Y = olvidar X.

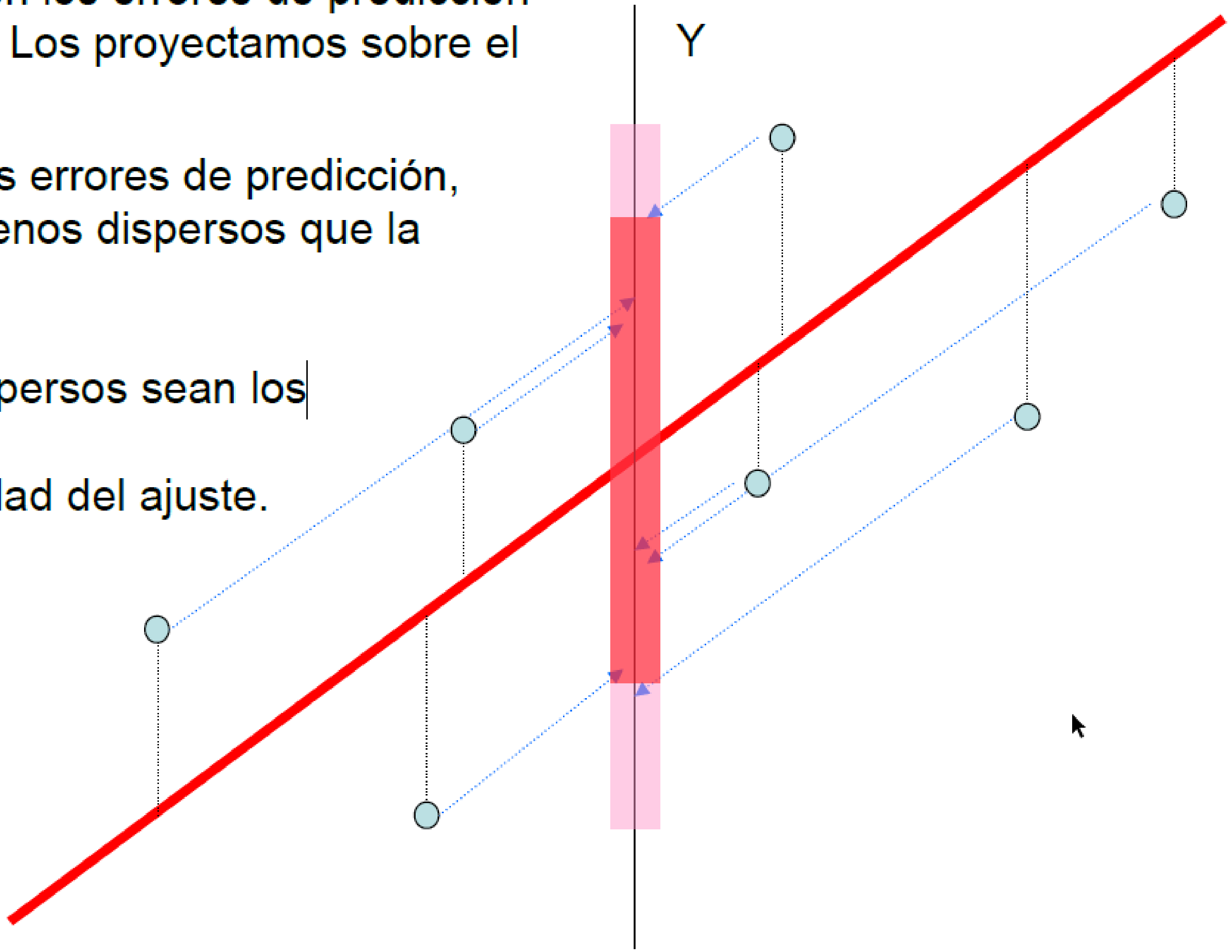


El residuo

Fijémonos ahora en los errores de predicción (líneas verticales). Los proyectamos sobre el eje Y.

Se observa que los errores de predicción, residuos, están menos dispersos que la variable Y original.

Cuanto menos dispersos sean los residuos,
mejor será la bondad del ajuste.



Covarianza

- La covarianza entre dos variables, $\mathbf{S_{xy}}$, indica si la posible relación entre dos variables es directa o inversa:
- Directa: $S_{xy} > 0$
- Inversa: $S_{xy} < 0$
- Descorreladas: $S_{xy} = 0$
- El signo de la covarianza da una idea de la nube de datos, pero no dice nada del grado de relación.

$$\boxed{S_{xy} = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})}$$

Coeficiente de correlación de Pearson (r)

$$r = \frac{S_{xy}}{S_x S_y}$$

- Indica si los datos tienen una tendencia a disponerse alineadamente (excluyendo rectas horizontales y verticales).
- Tiene el mismo signo que S_{xy} . Por tanto de su signo obtenemos el que la posible relación sea directa o inversa.
- Es útil para determinar si hay relación lineal entre dos variables, pero no servirá para otro tipo de relaciones (cuadrática, logarítmica,...)
- Sólo toma valores en [-1,1].
- Cuanto más cerca esté r de +1 o -1 mejor será el grado de relación lineal.

Coeficiente de determinación R^2

- El **coeficiente de determinación** es la proporción de la varianza total de la variable explicada por la regresión y sirve para reflejar la bondad del ajuste de un modelo a la variable que se pretende explicar.
- El valor de R^2 oscila entre 0 y 1. Cuanto más cerca de 1 se sitúe su valor, mayor será el ajuste del modelo a la variable que estamos intentando explicar. De forma inversa, cuanto más cerca de cero, menos ajustado estará el modelo y, por tanto, menos fiable será.
- Es muy importante tener clara la **diferencia** entre el **coeficiente de correlación** y el **coeficiente** de **determinación**:
  - R^2: mide la proporción de variación de la variable dependiente explicada por la variable independiente.
  - r: mide el grado de asociación entre las dos variables.
- **Coeficiente de determinación ajustado**: Al añadir variables al modelo, R^2 podría aumentar y se puede concluir, de manera errónea, que el conjunto de variables elegido es capaz de explicar una mayor parte de la variación de la variable independiente. R^2 ajustado se utiliza para ver el grado de intensidad o efectividad que tienen las variables independientes en explicar la variable dependiente. Dice qué porcentaje de variación de la variable dependiente es explicado colectivamente por todas las variables independientes.

Bondad de ajuste o R^2

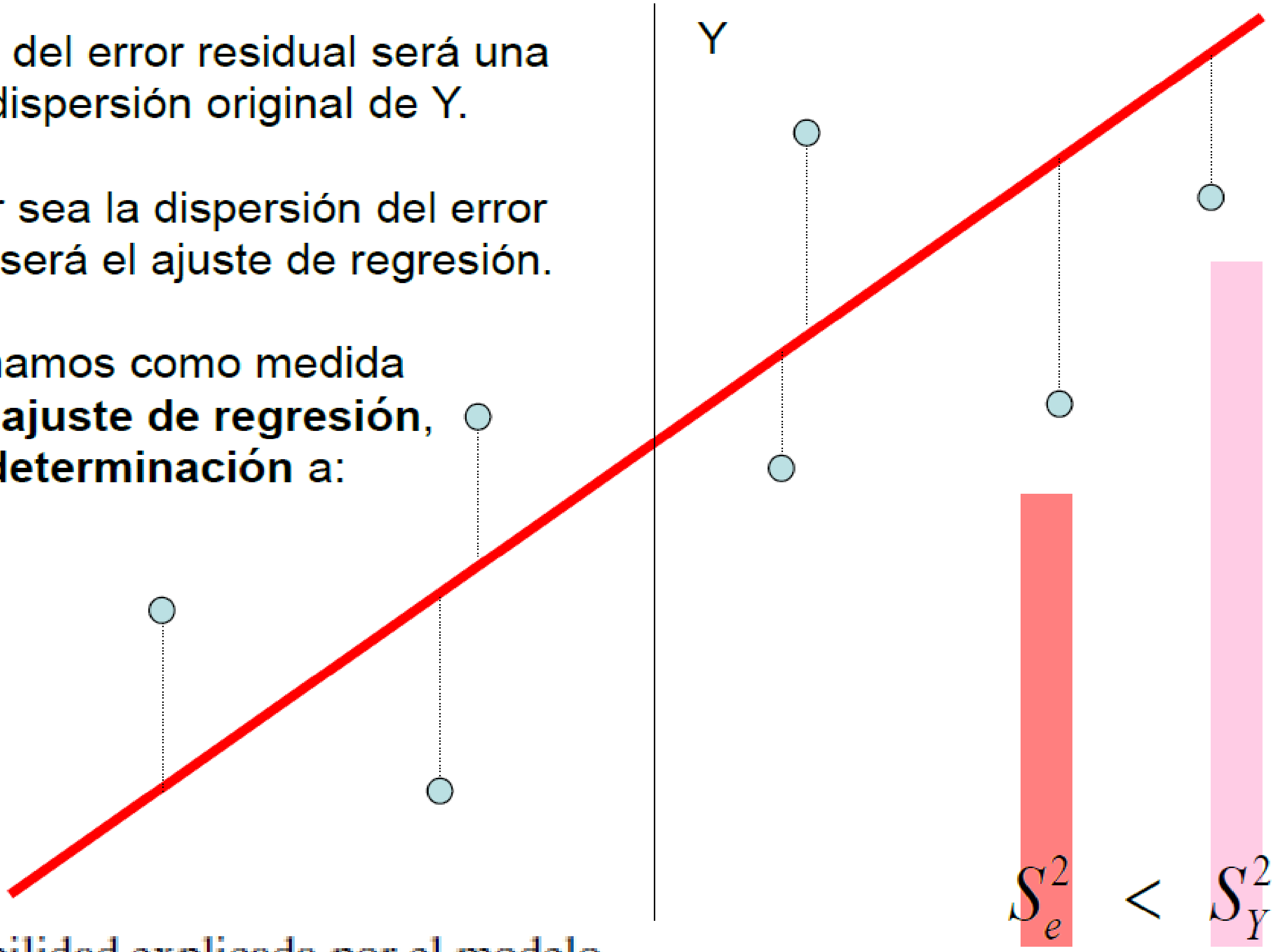
Resumiendo:

- La dispersión del error residual será una fracción de la dispersión original de Y.

- Cuanto menor sea la dispersión del error residual mejor será el ajuste de regresión.

Eso hace que definamos como medida de **bondad de un ajuste de regresión**, o **coeficiente de determinación** a:

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



$$R^2 = \frac{\text{Variabilidad explicada por el modelo}}{\text{Variabilidad total}}$$

Prueba de hipótesis del modelo de regresión

- Es otra forma de probar que tan bien se explica la relación entre *X* y *Y.*
- Una forma de hacer esto es probar una serie de hipótesis sobre el modelo.
- Para ello se supone una distribución normal para el término de error, ε_i, con media cero y varianza σ^2.
- La hipótesis de mayor interés plantea que la pendiente es significativamente diferente de cero.
- Esto se logra al probar la siguiente hipótesis:

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_A &: \beta_1 \neq 0 \end{aligned}$$

Análisis de regresión para el modelo

Parámetro	Estimación	Error estándar	Estadístico	Valor-*p*
Intercepción	$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$	$\sqrt{CM_E\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$	$\dfrac{\hat{\beta}_0}{\sqrt{CM_E\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}}$	$\Pr(T > \lvert t_0 \rvert)$
Pendiente	$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$	$\sqrt{CM_E / S_{xx}}$	$\dfrac{\hat{\beta}_1}{\sqrt{CM_E / S_{xx}}}$	$\Pr(T > \lvert t_0 \rvert)$

Intervalo de confianza para los parámetros de la regresión

$$\hat{\beta}_1 \pm t_{(\alpha/2, n-2)} \sqrt{CM_E / S_{xx}}$$

$$\hat{\beta}_0 \pm t_{(\alpha/2, n-2)} \sqrt{CM_E \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$$

ANOVA del modelo de regresión

- Se descompone la variabilidad observada, y a partir de ello se prueba la hipótesis.

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_0	Valor-*p*
Regresión	$SC_R = \hat{\beta}_1 S_{xy}$	1	CM_R	CM_R/CM_E	$\Pr(F > F_0)$
Error o residual	$SC_E = S_{yy} - \hat{\beta}_1 S_{xy}$	$n - 2$	CM_E		
Total	S_{yy}	$n - 1$			

Ejemplo:

- Se quiere investigar la forma en que se relaciona la cantidad de fibra (madera) en la pulpa con la resistencia del producto (papel).
- Los datos obtenidos en un estudio experimental se muestran en la tabla

Porcentaje de fibra	Resistencia
4	134
6	145
8	142
10	149
12	144
14	160
16	156
18	157
20	168
22	166
24	167
26	171
28	174
30	183

Análisis de regresión lineal

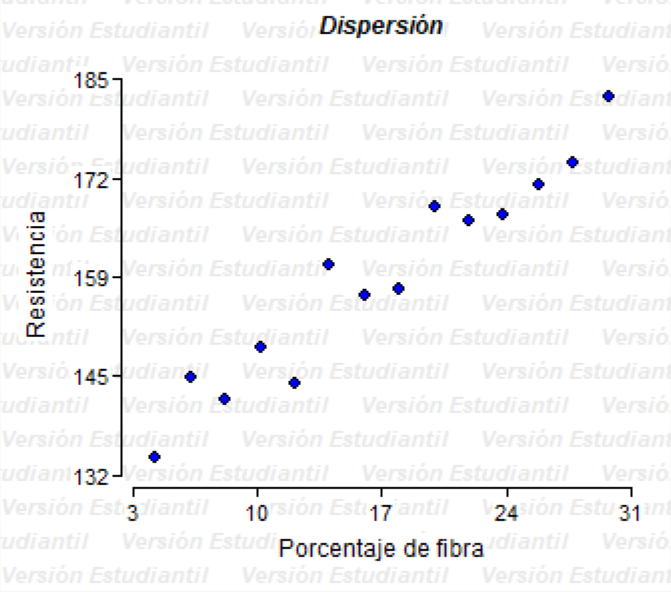
Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
Resistencia	14	0.93	0.92	20.22	81.51	83.43

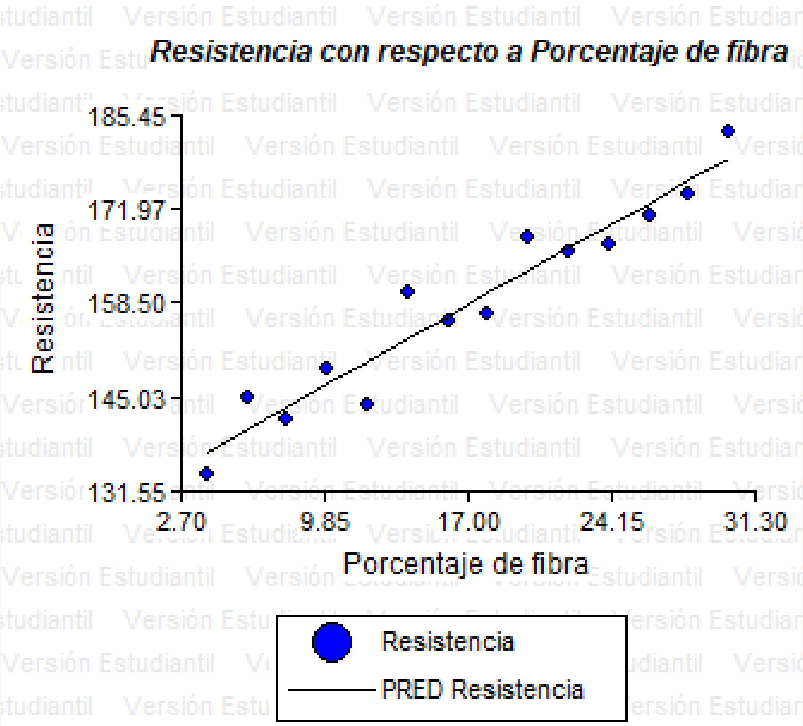
Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	130.67	2.42	125.41	135.94	54.05	<0.0001		
Porcentaje de fibra	1.62	0.13	1.34	1.90	12.64	<0.0001	159.75	1.00

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo.	2400.53	1	2400.53	159.75	<0.0001
Porcentaje de fibra	2400.53	1	2400.53	159.75	<0.0001
Error	180.33	12	15.03		
Total	2580.86	13			





SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.964432318
R Square	0.930129695
Adjusted R Square	0.92430717
Standard Error	3.876481166
Observations	14

ANOVA

	df	SS	MS	F	gnificance F
Regression	1	2400.53	2400.53	159.747	2.7E-08
Residual	12	180.325	15.0271		
Total	13	2580.86			

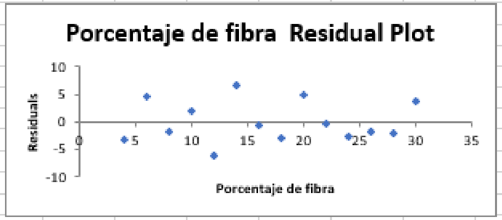
	Coefficients	andard Err	t Stat	P-value	.ower 95%	Jpper 95%	ower 95.0%	pper 95.0%
Intercept	130.6747253	2.41779	54.0472	1.1E-15	125.407	135.943	125.407	135.943
Porcentaje de fibra	1.624175824	0.1285	12.6391	2.7E-08	1.34419	1.90416	1.34419	1.90416

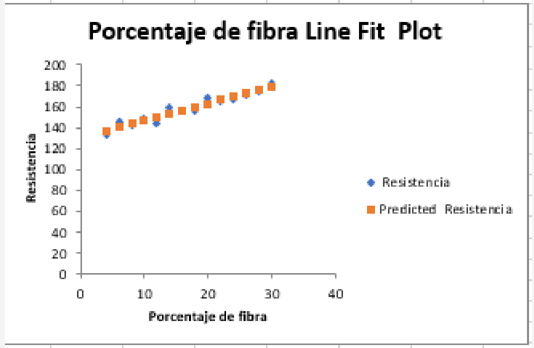
RESIDUAL OUTPUT

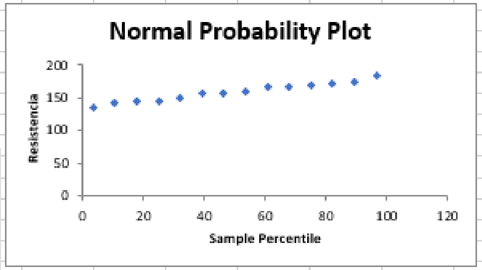
Observation	edicted Resistenc	Residuals
1	137.1714286	-3.17143
2	140.4197802	4.58022
3	143.6681319	-1.66813
4	146.9164835	2.08352
5	150.1648352	-6.16484
6	153.4131868	6.58681
7	156.6615385	-0.66154
8	159.9098901	-2.90989
9	163.1582418	4.84176
10	166.4065934	-0.40659
11	169.6549451	-2.65495
12	172.9032967	-1.9033
13	176.1516484	-2.15165
14	179.4	3.6

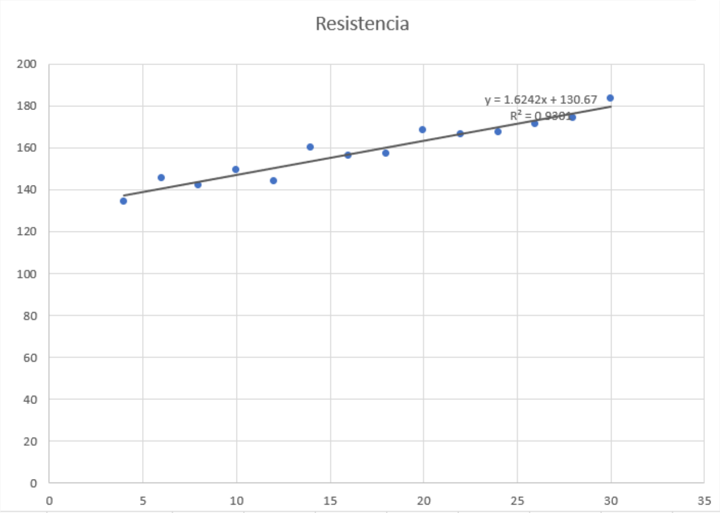
PROBABILITY OUTPUT

Percentile	Resistencia
3.57143	134
10.7143	142
17.8571	144
25	145
32.1429	149
39.2857	156
46.4286	157
53.5714	160
60.7143	166
67.8571	167
75	168
82.1429	171
89.2857	174
96.4286	183









Regresión lineal múltiple

- En muchas situaciones prácticas existen varias variables independientes que se cree que influyen o están relacionadas con una variable de respuesta *Y*, y por lo tanto será necesario tomar en cuenta si se quiere predecir o entender mejor el comportamiento de *Y*.
- Sea $X_1, X_2, \ldots, X_k$ variables independientes o regresoras, y sea *Y* una variable de respuesta, entonces el *modelo de regresión lineal múltiple* con *k* variables independientes es el polinomio de primer orden:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

- Donde los β_j son los parámetros del modelo que se conocen como *coeficientes de regresión* y ε es el error aleatorio, con media cero, $E(\varepsilon) = 0$ y $V(\varepsilon) = \sigma^2$.
- Así:
  - β_0 es el término independiente. Es el valor esperado de Y cuando $X_1, \ldots, X_k$ son cero.
  - $\beta_1, \beta_2, \ldots \beta_k$ son los coeficientes parciales de la regresión tal que β_1 mide el cambio en Y por cada cambio unitario en X_1, manteniendo $X_2, X_3, \ldots, X_k$ constantes y así sucesivamente.
  - ε es el error de observación debido a variables no controladas.

Estructura de datos

Y	X_1	X_2	...	X_k
y_1	x_{11}	x_{21}	...	x_{k1}
y_2	x_{12}	x_{22}	...	x_{k2}
	$\vdots$	$\vdots$		$\vdots$
y_n	y_{1n}	x_{2n}	...	x_{kn}

Análisis de regresión

Parámetro	Estimación	Error estándar	Estadístico	Valor-*p*
Intercepción	$\hat{\beta}_0$	$\sqrt{CM_E C_{11}}$	$\frac{\hat{\beta}_0}{\sqrt{CM_E C_{11}}}$	$\Pr(T > \|t_0\|)$
β_1	$\hat{\beta}_1$	$\sqrt{CM_E C_{22}}$	$\frac{\hat{\beta}_1}{\sqrt{CM_E C_{22}}}$	$\Pr(T > \|t_0\|)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
β_k	$\hat{\beta}_k$	$\sqrt{CM_E C_{k+1,k+1}}$	$\frac{\hat{\beta}_k}{\sqrt{CM_E C_{k+1,k+1}}}$	$\Pr(T > \|t_0\|)$

Prueba de hipótesis

$$H_0 : \beta_1 = \beta_2 = \dots \beta_k = 0$$
$$H_A : \beta_j \neq 0 \qquad \text{para al menos un } j = 1, 2, \dots, k$$

ANOVA

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_0	Valor-*p*
Regresión	$SC_R = \hat{\beta}'\mathbf{X}'\mathbf{y} - \frac{(\Sigma_{i=1}^n y_i)^2}{n}$	k	CM_R	CM_R/CM_E	$\Pr(F > F_0)$
Error o residuo	$SC_E = \mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{y}$	$n-k-1$	CM_E		
Total	$S_{yy} = \mathbf{y}'\mathbf{y} - \frac{(\Sigma_{i=1}^n y_i)^2}{n}$	$n-1$			

Ejemplo

- Se tiene una muestra de 8 mujeres, donde se quiere predecir el peso en función a una serie de variables. La información se muestra a continuación:

Análisis de regresión lineal

Variable	N	R²	R² Aj	ECMP	AIC	BIC
Peso	8	0.50	0.00	1056.81	49.10	49.66

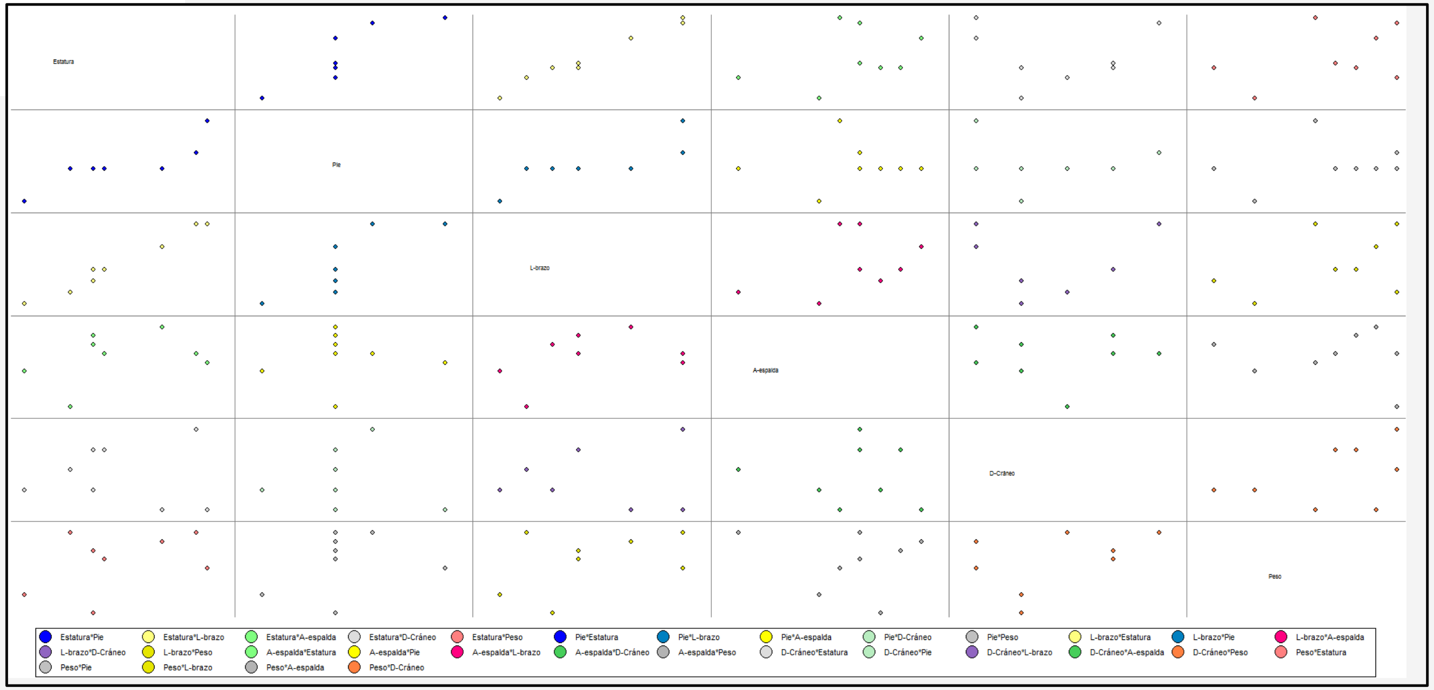
Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	-32.60	111.52	-512.44	447.24	-0.29	0.7976		
Estatura	0.83	2.61	-10.41	12.08	0.32	0.7803	4.10	78.80
Pie	-1.48	2.72	-13.17	10.20	-0.55	0.6394	4.30	5.28
L-brazo	-0.38	5.43	-23.72	22.97	-0.07	0.9510	4.00	74.95
A-espalda	-0.52	0.79	-3.94	2.90	-0.66	0.5787	4.43	1.80
D-Cráneo	0.89	1.36	-4.97	6.75	0.65	0.5802	4.43	1.53

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo.	37.80	5	7.56	0.40	0.8226
Estatura	1.91	1	1.91	0.10	0.7803
Pie	5.63	1	5.63	0.30	0.6394
L-brazo	0.09	1	0.09	4.8E-03	0.9510
A-espalda	8.14	1	8.14	0.43	0.5787
D-Cráneo	8.06	1	8.06	0.43	0.5802
Error	37.70	2	18.85		
Total	75.50	7			

Registro	estatura	pie	l_brazo	a_espalda	d_cráneo	peso
1	158	36	68	43	55	43
2	152	34	66	40	55	45
3	168	39	73	41	54	48
4	159	36	69	42	57	49
5	158	36	69	44	57	50
6	164	36	71	45	54	51
7	156	36	67	36	56	52
8	167	37	73	42	58	52



Ejemplo:

- Se presenta un experimento secuencial para optimizar la producción de un colorante natural. En la etapa final se delimitó una zona de experimentación donde se sospecha que se encuentran las condiciones óptimas para la producción de este colorante en función de la concentración de carbono ($X1$) y temperatura ($X2$).

Y	***X1***	***X2***
5707	9	17
5940	13	17
3015	9	25
2673	13	25
5804	8	21
6700	13	21
5310	11	15
7250	11	26
7521	11	21
7642	11	21
7500	11	21
7545	11	21

Análisis de regresión lineal

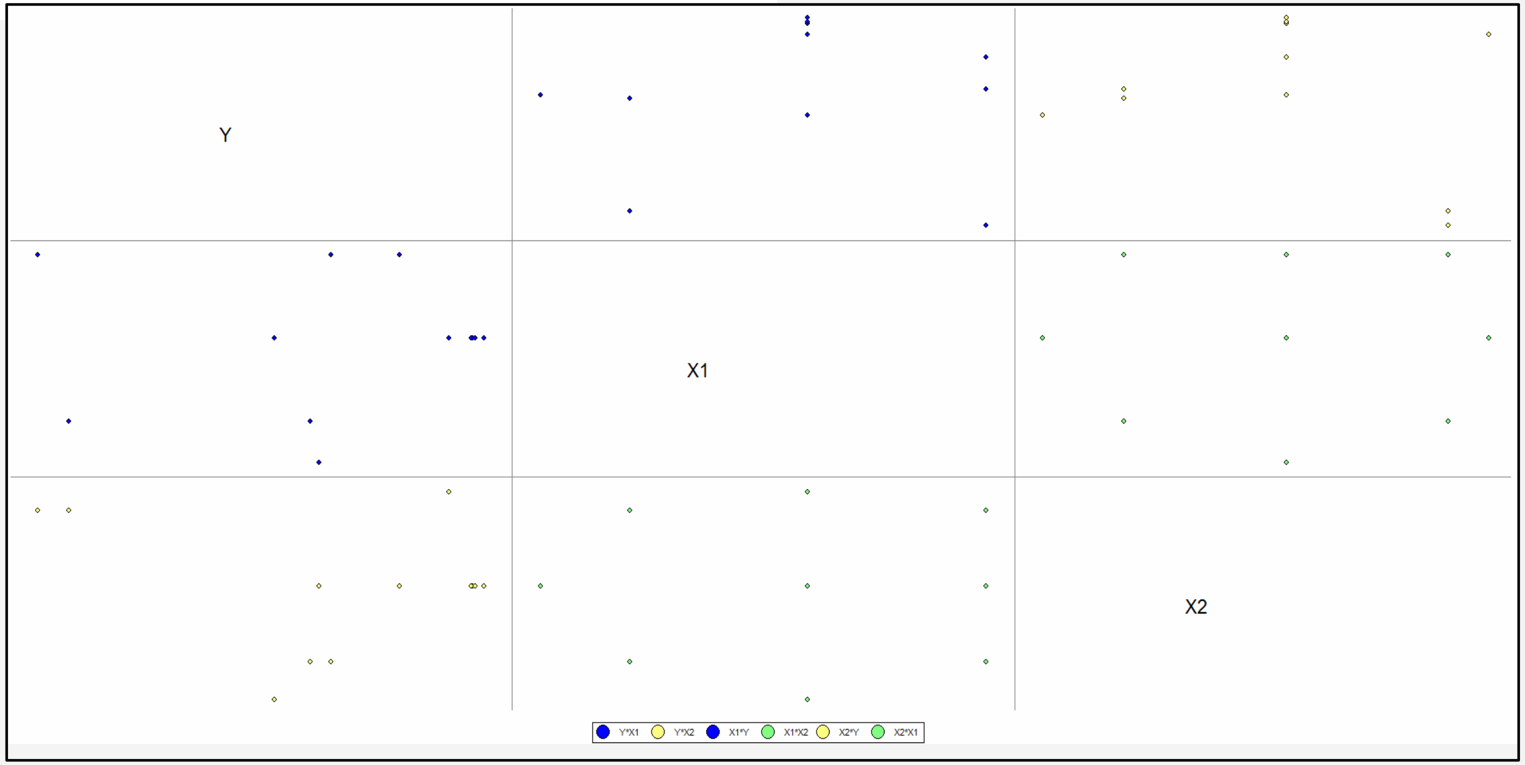
Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
Y	12	0.05	0.00	6864357.82	219.11	221.05

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	7608.77	5130.96	-3998.28	19215.81	1.48	0.1722		
X1	62.65	343.47	-714.33	839.63	0.18	0.8593	1.03	1.00
X2	-107.19	165.25	-481.02	266.64	-0.65	0.5328	1.42	1.00

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo.	1549956.17	2	774978.09	0.23	0.8012
X1	113501.03	1	113501.03	0.03	0.8593
X2	1435332.87	1	1435332.87	0.42	0.5328
Error	30701588.75	9	3411287.64		
Total	32251544.92	11			

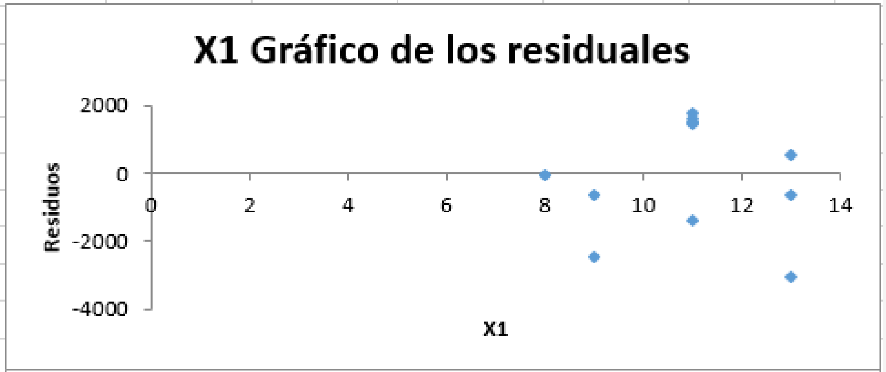


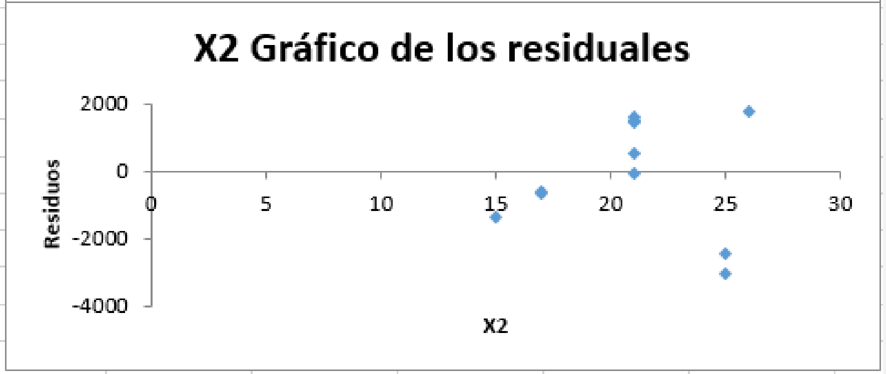
Estadísticas de la regresión	
Coeficiente de corr	0.219222157
Coeficiente de dete	0.048058354
R^2 ajustado	-0.163484234
Error típico	1846.967146
Observaciones	12

ANÁLISIS DE VARIANZA

	Grados de libertad	*a de cuadra*	*dio de los cua*	*F*	*Valor crítico de F*
Regresión	2	1549956.17	774978.085	0.22718052	0.801210202
Residuos	9	30701588.7	3411287.64		
Total	11	32251544.9			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95.0%*	*Superior 95.0%*
Intercepción	7608.766922	5130.96375	1.48291184	0.17224599	-3998.279483	19215.81333	-3998.279483	19215.81333
X1	62.65074055	343.467361	0.18240668	0.85930614	-714.3264105	839.6278916	-714.3264105	839.6278916
X2	-107.1930282	165.253016	-0.64866004	0.53275799	-481.0213219	266.6352655	-481.0213219	266.6352655





Regresión no lineal o de orden superior

- Los parámetros son lineales cuando cada término del modelo es aditivo y contiene solo un parámetro que multiplica el término.
- Es un problema de inferencia para un modelo del tipo

$$y = f(x, \theta) + \varepsilon$$

- Está fundamentado en datos multidimensionales *x*, *y*, donde f es una función no lineal de algunos parámetros desconocidos θ.
- Se pretende obtener los valores de los parámetros de tal manera que se obtenga la mejor curva de ajuste.
- Se utiliza cuando no pueda modelarse adecuadamente la relación con parámetros lineales.

- Se quiere probar un modelo de segundo orden:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} + \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 + \varepsilon_i$$

y	X1	x2	x1x2	$x1^2$	$x2^2$
5707	9	17	153	81	289
5940	13	17	221	169	289
3015	9	25	225	81	625
2673	13	25	325	169	625
5804	8	21	168	64	441
6700	13	21	273	169	441
5310	11	15	165	121	225
7250	11	26	286	121	676
7521	11	21	231	121	441
7642	11	21	231	121	441
7500	11	21	231	121	441
7545	11	21	231	121	441

Resumen

Estadísticas de la regresión	
Coeficiente (	0.79266847
Coeficiente (	0.62832331
R^2 ajustadc	0.31859273
Error típico	1413.45602
Observacion	12

ANÁLISIS DE VARIANZA

	Grados de libert	*na de cuadrad*	*dio de los cua*	*F*	*alor crítico de F*
Regresión	5	20264397.4	4052879.47	2.02861245	0.20715042
Residuos	6	11987147.6	1997857.93		
Total	11	32251544.9			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*nferior 95.0%*	*uperior 95.0%*
Intercepción	-71007.2838	32101.5037	-2.2119613	0.06895682	-149556.834	7542.26602	-149556.834	7542.26602
X1	8091.25305	3863.2781	2.09440088	0.08110317	-1361.84793	17544.354	-1361.84793	17544.354
x2	3473.11924	1771.74197	1.96028502	0.09765656	-862.17719	7808.41567	-862.17719	7808.41567
x1x2	-17.96875	88.3410014	-0.20340215	0.84554363	-234.131393	198.193893	-234.131393	198.193893
X1sq	-355.726778	157.612486	-2.2569708	0.06481733	-741.390638	29.937082	-741.390638	29.937082
X2sq	-81.6321728	35.6846521	-2.28759895	0.06214839	-168.949371	5.68502528	-168.949371	5.68502528

Análisis de regresión lineal

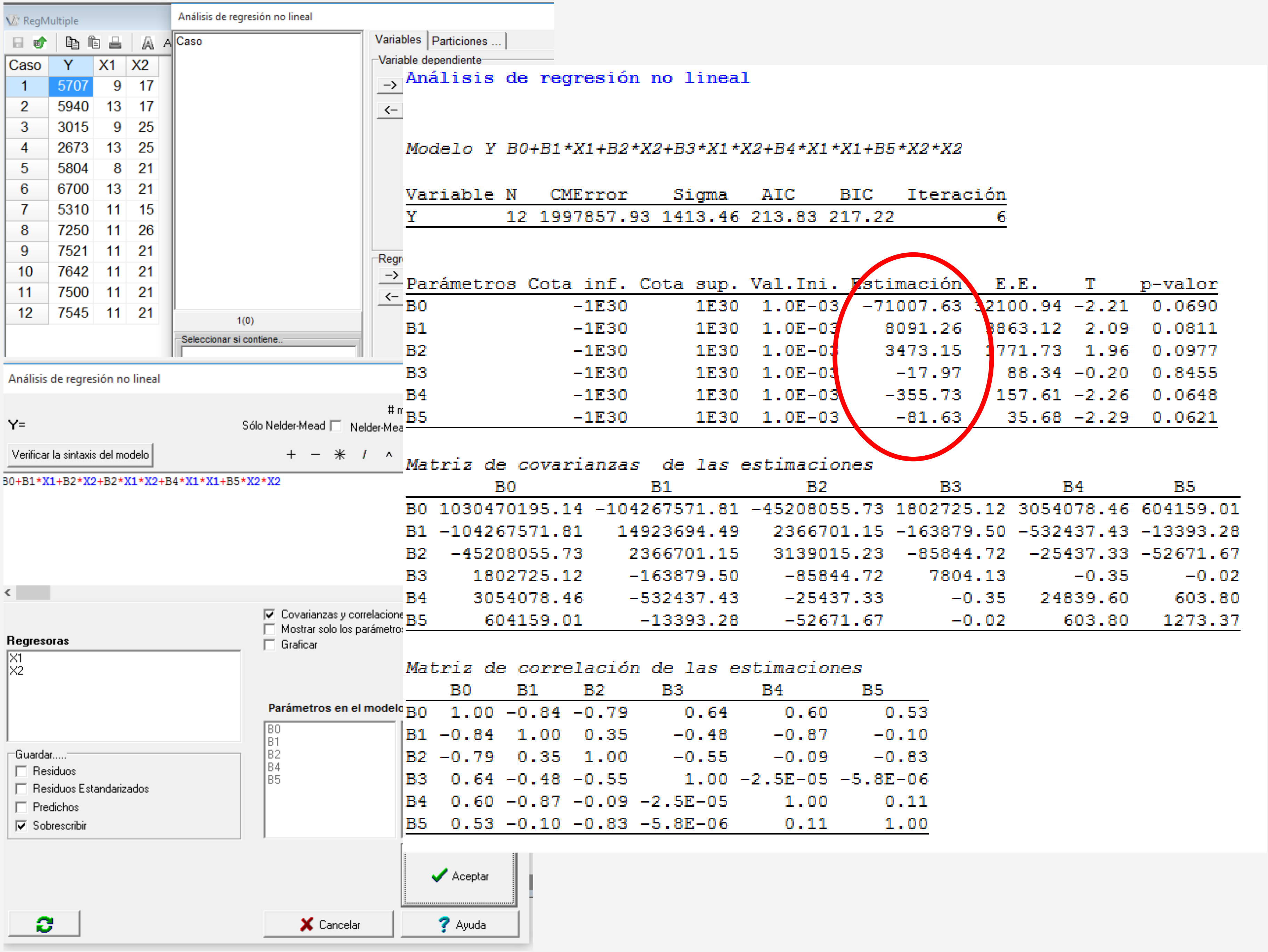
Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
Y	12	0.63	0.32	16032059.28	213.83	217.22

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	-71007.28	32101.50	-149556.82	7542.25	-2.21	0.0690		
X1	8091.25	3863.28	-1361.85	17544.35	2.09	0.0811	8.39	216.02
X2	3473.12	1771.74	-862.18	7808.41	1.96	0.0977	7.84	196.27
X1X2	-17.97	88.34	-234.13	198.19	-0.20	0.8455	4.04	109.71
X1sq	-355.73	157.61	-741.39	29.94	-2.26	0.0648	9.09	166.08
X2sq	-81.63	35.68	-168.95	5.69	-2.29	0.0621	9.23	137.18

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo.	20264397.36	5	4052879.47	2.03	0.2072
X1	8763633.84	1	8763633.84	4.39	0.0811
X2	7677203.30	1	7677203.30	3.84	0.0977
X1X2	82656.25	1	82656.25	0.04	0.8455
X1sq	10176922.80	1	10176922.80	5.09	0.0648
X2sq	10455008.25	1	10455008.25	5.23	0.0621
Error	11987147.56	6	1997857.93		
Total	32251544.92	11			



Caso	Y	X1	X2
1	5707	9	17
2	5940	13	17
3	3015	9	25
4	2673	13	25
5	5804	8	21
6	6700	13	21
7	5310	11	15
8	7250	11	26
9	7521	11	21
10	7642	11	21
11	7500	11	21
12	7545	11	21

Análisis de regresión no lineal

Y=

B0+B1*X1+B2*X2+B2*X1*X2+B4*X1*X1+B5*X2*X2

Regresoras: X1, X2

Parámetros en el modelo: B0, B1, B2, B4, B5

Análisis de regresión no lineal

*Modelo Y B0+B1*X1+B2*X2+B3*X1*X2+B4*X1*X1+B5*X2*X2*

Variable	N	CMError	Sigma	AIC	BIC	Iteración
Y	12	1997857.93	1413.46	213.83	217.22	6

Parámetros	Cota inf.	Cota sup.	Val.Ini.	Estimación	E.E.	T	p-valor
B0	-1E30	1E30	1.0E-03	-71007.63	32100.94	-2.21	0.0690
B1	-1E30	1E30	1.0E-03	8091.26	3863.12	2.09	0.0811
B2	-1E30	1E30	1.0E-03	3473.15	1771.73	1.96	0.0977
B3	-1E30	1E30	1.0E-03	-17.97	88.34	-0.20	0.8455
B4	-1E30	1E30	1.0E-03	-355.73	157.61	-2.26	0.0648
B5	-1E30	1E30	1.0E-03	-81.63	35.68	-2.29	0.0621

Matriz de covarianzas de las estimaciones

	B0	B1	B2	B3	B4	B5
B0	1030470195.14	-104267571.81	-45208055.73	1802725.12	3054078.46	604159.01
B1	-104267571.81	14923694.49	2366701.15	-163879.50	-532437.43	-13393.28
B2	-45208055.73	2366701.15	3139015.23	-85844.72	-25437.33	-52671.67
B3	1802725.12	-163879.50	-85844.72	7804.13	-0.35	-0.02
B4	3054078.46	-532437.43	-25437.33	-0.35	24839.60	603.80
B5	604159.01	-13393.28	-52671.67	-0.02	603.80	1273.37

Matriz de correlación de las estimaciones

	B0	B1	B2	B3	B4	B5
B0	1.00	-0.84	-0.79	0.64	0.60	0.53
B1	-0.84	1.00	0.35	-0.48	-0.87	-0.10
B2	-0.79	0.35	1.00	-0.55	-0.09	-0.83
B3	0.64	-0.48	-0.55	1.00	-2.5E-05	-5.8E-06
B4	0.60	-0.87	-0.09	-2.5E-05	1.00	0.11
B5	0.53	-0.10	-0.83	-5.8E-06	0.11	1.00

Estrategias de modelado

- En un análisis predictivo el mejor modelo es el que produce predicciones más *fiables* para una nueva observación, mientras que en un análisis estimativo el mejor modelo es el que produce estimaciones más *precisas* para el coeficiente de la variable de interés.
- En ambos casos se prefiere el modelo más sencillo posible (a este modo de seleccionar modelos se le denomina *parsimonia*).
- Sin embargo, hay una serie de pasos que deben realizarse siempre:
  - Especificación del modelo máximo.
  - Especificación de un criterio de comparación de modelos y definición de una estrategia para realizarla.
  - Evaluación de la fiabilidad del modelo.

Especificando el modelo máximo

- El criterio para decidir qué variables forman el modelo máximo se establece en función de sus objetivos y del conocimiento teórico que se tenga sobre el problema, evidentemente cuanto menor sea el conocimiento previo mayor tenderá a ser el modelo máximo.
- Un modelo máximo grande minimiza la probabilidad de error tipo II o *infra-ajuste*, que en un análisis de regresión consiste en no considerar una variable que realmente tiene un coeficiente de regresión distinto de cero.
- Un modelo máximo pequeño minimiza la probabilidad de error tipo I o *sobreajuste*(incluir en el modelo una variable independiente cuyo coeficiente de regresión realmente sea cero).
- Aunque el sobreajuste no introduce sesgos en la estimación de los coeficientes, un infra-ajuste puede producirlos, pero que un modelo máximo grande aumenta la probabilidad de problemas de colinealidad.

Comparando los modelos

- Debe establecerse cómo y con qué se comparan los modelos.
- Si bien hay varios estadísticos sugeridos para comparar modelos, el más frecuentemente usado es la F parcial, recordando que cuando los dos modelos sólo difieren en una variable, el contraste sobre la F parcial es exactamente el mismo que el realizado con la t sobre el coeficiente de regresión
- Por otro lado a veces interesa contrastar varias variables conjuntamente mejor que una a una (por ejemplo todos los términos no lineales) o, incluso, es necesario hacerlo (por ejemplo para variables indicadoras).
- En un análisis estimativo el criterio para incluir o excluir variables distintas a las de interés, es sobre todo los cambios en los coeficientes y no los cambios en la significación del modelo.

Selección de las variables regresoras

- La selección de variables regresoras es un procedimiento estadístico importante por que no todas las variables regresoras tienen igual importancia.
- Algunas variables regresoras pueden perjudicar la confiabilidad del modelo, en especial si están correlacionadas entre ellas.
- Computacionalmente es más fácil trabajar con un conjunto pequeño de variables regresoras.
- Es más económico recolectar información para un modelo con pocas variables.
- Si se reduce el número de variables entonces el modelo cumple con el principio de la parsimonia.

Algunos métodos de selección de las variables

- La idea de estos métodos es elegir el mejor modelo en forma secuencial pero incluyendo o excluyendo una sola variable regresora en cada paso de acuerdo a ciertos criterios.
- El proceso secuencial termina cuando se satisface una regla de parada establecida.
- Hay tres algoritmos usados: Backward Elimination, Forward Selection y Stepwise Selección.

Backward elimination

- Se comienza con el modelo completo y en cada paso se va eliminando una variable. Toda variable que sale, no puede entrar.
- Si todas las variables regresoras son importantes, es decir tienen p-value pequeños para la prueba t, entonces el mejor modelo es el que tiene todas las variables regresoras disponibles.
- En caso contrario, en cada paso la variable que se elimina del modelo es aquella que satisface cualquiera de los siguientes requisitos equivalentes entre sí:
  - Aquella variable que tiene el estadístico de t, en valor absoluto, más pequeño entre las variables incluidas aún en el modelo.
  - Aquella variable que produce la menor disminución en el R^2 al ser eliminada del modelo.
  - Aquella variable que tiene la correlación parcial (en valor absoluto) más pequeña con la variable de respuesta, tomando en cuenta las variables aún presentes en el modelo.
- El proceso termina cuando se llega a un modelo con un número prefijado $p*$ de variables regresoras.

Forward elimination

- Se empieza con aquella variable regresora que tiene la más alta correlación con la variable respuesta.
- En este caso, toda variable que entra no puede salir.
- En el siguiente paso se añade al modelo la variable que reúne cualquiera de estos requisitos equivalentes entre sí.
  - Aquella variable que produce el mayor incremento en el R^2 al ser añadida al modelo.
  - El proceso termina cuando se llega a un modelo con un número prejado $p*$ de variables predictoras.

Stepwise selection

- Se empieza con un modelo de regresión simple y en cada paso se puede añadir una variable en forma similar al método forward
- Se coteja si alguna de las variables que ya están presentes en el modelo puede ser eliminada.
- El proceso termina cuando ninguna de las variables fuera del modelo tiene importancia suficiente como para ingresar al modelo.

Evaluando la fiabilidad del modelo

- Una vez encontrado el mejor modelo hay que evaluar su fiabilidad, es decir, evaluar si se comporta igual en otras muestras extraídas de la misma población.
- Evidentemente, el modo más completo de evaluarlo será repetir el estudio con otra muestra y comprobar que se obtienen los mismos resultados, aunque generalmente esta aproximación resulta excesivamente costosa.
- Otra aproximación alternativa consiste en partir aleatoriamente la muestra en dos grupos y ajustar el modelo con cada uno de ellos y si se obtienen los mismos resultados se considera que el modelo es fiable.
- Esta aproximación es demasiado estricta ya que, en la práctica, casi nunca se obtienen los mismos resultados.
- Una validación menos estricta consiste en ajustar el modelo sobre uno de los grupos (grupo de trabajo) y calcular su R^2, que se puede interpretar como el cuadrado del coeficiente de correlación simple entre la variable dependiente y las estimaciones obtenidas en la regresión.

Se tiene un análisis de la eficiencia en peso, de un proceso de producción. Se escogieron seis variables que se cree pueden regir el proceso. Determine el mejor modelo de regresión

N	Y	X1	X2	X3	X4	X5	X6
1	138.00	8.00	534.00	107.39	138.08	690.00	138.08
2	58.00	12.00	182.00	43.80	72.40	290.00	72.40
3	30.00	10.00	546.00	0.43	35.55	150.00	35.55
4	30.00	13.00	367.00	12.06	30.19	150.00	30.19
5	69.00	5.00	478.00	64.41	63.01	345.00	-63.01
6	49.00	6.00	330.00	25.89	46.63	245.00	-46.63
7	30.00	7.00	165.00	18.29	32.20	150.00	32.20
8	136.00	12.00	184.00	86.63	139.33	680.00	139.33
9	119.00	11.00	355.00	104.37	97.54	595.00	-97.54
10	93.00	5.00	469.00	62.52	98.09	465.00	-98.09
11	19.00	8.00	246.00	16.65	21.30	95.00	-21.30
12	59.00	8.00	323.00	53.97	45.33	295.00	-45.33
13	143.00	12.00	395.00	115.52	142.41	715.00	-142.41
14	107.00	12.00	423.00	84.25	121.20	535.00	-121.20
15	81.00	8.00	525.00	56.16	69.19	405.00	-69.19
16	71.00	15.00	184.00	58.55	61.44	355.00	-61.44
17	103.00	5.00	160.00	3.66	100.49	515.00	-100.49
18	112.00	8.00	259.00	98.80	113.76	560.00	113.76
19	76.00	8.00	395.00	4.63	80.45	380.00	80.45
20	149.00	5.00	476.00	79.87	148.40	745.00	-148.40
21	126.00	11.00	173.00	111.57	122.65	630.00	122.65
22	47.00	7.00	175.00	19.65	46.43	235.00	-46.43
23	140.00	8.00	209.00	113.07	142.50	700.00	142.50
24	59.00	10.00	292.00	39.36	50.62	295.00	-50.62
25	85.00	12.00	220.00	34.39	85.51	425.00	-85.51
26	19.00	10.00	251.00	6.47	3.57	95.00	-3.57
27	6.00	8.00	354.00	5.91	0.10	30.00	0.10
28	78.00	14.00	222.00	2.56	59.08	390.00	59.08
29	12.00	14.00	560.00	4.39	7.52	60.00	7.52
30	26.00	11.00	501.00	4.70	19.68	130.00	19.68

Análisis de regresión lineal

Variable	N	R^2	R^2 Aj	ECMP	AIC	BIC
y	30	0.23	0.03	3041.60	318.61	329.82

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	72.37	40.73	-11.88	156.62	1.78	0.0888		
x1	-2.38	2.85	-8.27	3.52	-0.83	0.4128	5.70	1.08
x2	0.01	0.06	-0.12	0.13	0.09	0.9293	5.01	1.08
x3	-0.42	0.37	-1.19	0.35	-1.13	0.2702	6.28	3.55
x4	-2.07	0.99	-4.11	-0.03	-2.10	0.0470	9.41	32.05
x5	0.50	0.21	0.06	0.94	2.34	0.0280	10.50	34.76
x6	-0.02	0.10	-0.22	0.18	-0.22	0.8257	5.05	1.13

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo.	12861.59	6	2143.60	1.17	0.3570
x1	1276.67	1	1276.67	0.70	0.4128
x2	14.75	1	14.75	0.01	0.9293
x3	2342.26	1	2342.26	1.28	0.2702
x4	8085.20	1	8085.20	4.41	0.0470
x5	10090.01	1	10090.01	5.50	0.0280
x6	91.00	1	91.00	0.05	0.8257
Error	42204.58	23	1834.98		
Total	55066.17	29			

Análisis de regresión lineal

Variable	N	R²	R² Aj	ECMP	AIC	BIC
y	30	0.15	0.09	2078.14	313.68	319.28

Eliminación backward. Máximo p-valor para retener: 0.15
Número original de regresoras: 6,regresoras retenidas en el modelo 2

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	55.59	15.82	23.12	88.06	3.51	0.0016		
x4	-2.00	0.93	-3.91	-0.09	-2.15	0.0408	5.62	30.32
x5	0.42	0.19	0.03	0.82	2.19	0.0373	5.80	30.32

Error cuadrático medio: 1731.518962

Variable	N	R²	R² Aj	ECMP	AIC	BIC
y	30	0.00	0.00	2032.05	314.59	317.39

Selección Forward. Máximo p-valor para entrar: 0.15
Número original de regresoras: 6,regresoras retenidas en el modelo 0

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	69.83	7.96	53.56	86.10	8.78	<0.0001		

Error cuadrático medio: 1898.833333

Variable	N	R²	R² Aj	ECMP	AIC	BIC
y	30	0.00	0.00	2032.05	314.59	317.39

Selección Stepwise.
Máximo p-valor para entrar: 0.15
Máximo p-valor para retener: 0.15
Número original de regresoras: 6,regresoras retenidas en el modelo 0

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	69.83	7.96	53.56	86.10	8.78	<0.0001		

Error cuadrático medio: 1898.833333

Riesgos de la regresión

- La relación entre las variables puede ser espuria y el usuario es quien debe investigar si tal relación es de tipo causa-efecto. Se debe tomar en cuenta que algunas de las razones por las que las variables X y Y aparecen relacionadas de manera significativa son:
  - X influye sobre Y.
  - Y influye sobre X.
  - X y Y interactúan entre sí, una tercera variable Z influye sobre ambas y es lcausante de tal relación.
  - X y Y actúan en forma similar debido al azar.
  - X y Y aparecen relacionados debido a que la muestra no es representativa.
- Hacer extrapolaciones indiscriminadas con base en el modelo. Para no incurrir en esto se debe tener cuidado en cuanto a extrapolar más allá de la región que contienen las observaciones originales.

Ejemplo de relaciones espurias:

- http://www.tylervigen.com/spurious-correlations
- https://rodas5.us.es/items/ea5754ec-0941-2a2a-32e1-b34998971302/1/viewcontent/5742ce5f-de69-479e-8251-f507e4488223?_sl.t=true