Regresión y correlación

Oscar Centeno Mora

Introducción

- En los dos últimos capítulos se estudió la estimación y la prueba de hipótesis para hacer conjeturas sobre la población.

- El presente capítulo aplica técnicas para evaluar las relaciones entre dos variables.

- Se pretende estudiar la correlación y la regresión bivariada cómo técnicas para establecer relaciones entre dos variables cuantitativas. Para ambos casos, es estudia para dos variables

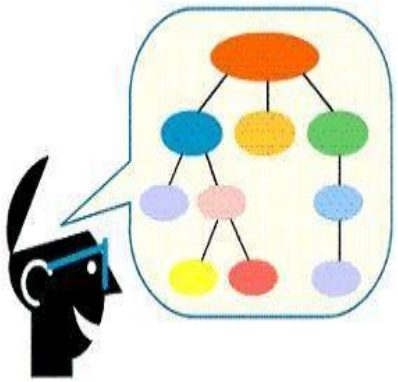
- A nivel de la regresión, se estudia la aplicación de intervalos de confianza y prueba de hipótesis para poder inferir los resultados a nivel población.

- Cualquier técnica siempre busca el poder inferir a nivel poblacional.

Índice

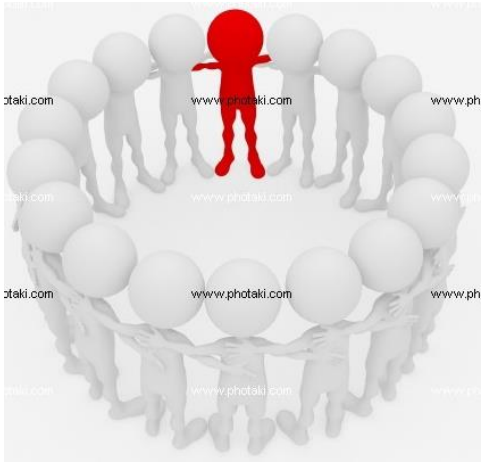
Relaciones: simétrica vs asimétrica

- Los análisis de la regresión y la correlación se deben de plantear ante todo como una cuestión conceptual.



- En el análisis de datos se quiere saber si existe o no relación entre dos variables: relación entre el peso y la estatura o relación entre la edad y las veces que se asiste al hospital.

- Por otra parte, a veces se quiere saber cuál es o son los factores que podrían explicar cierto acontecimiento: las horas de estudio son refleja de los altos o bajos puntajes en las calificaciones, las horas de ejercicio por semana sobre el porcentaje de grasa corporal, etc.

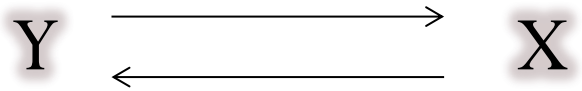


- De lo anterior, es importante saber qué es lo que se quiere en término de análisis de datos.

Relaciones: simétrica vs asimétrica

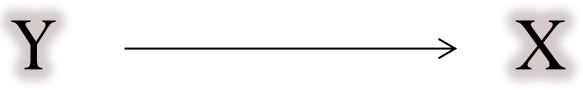
Relación simétrica: correlación

- Supone la relación entre dos variables de igual importancia.
- También conocido como correlación, y más generalmente como asociación.
- No se busca una explicación alguna. Solo conocer en que medida existe un asociación entre las variables.



Relación asimétrica: regresión

- Supone una relación de diferente importancia entre las variables
- Una variable trata de explicar a la otra, y esta explicación se suele expresar mediante una función matemática.
- La relación de asimetría se busca determinar una relación de causalidad.

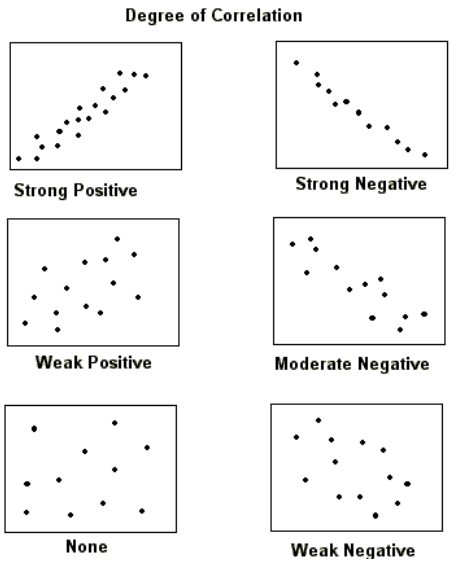


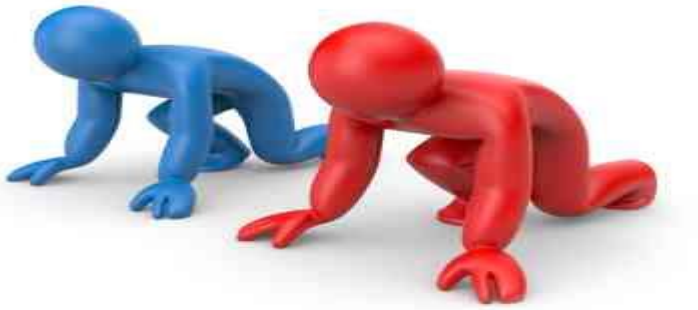
Relaciones: simétrica vs asimétrica

¿Cuál es la diferencia entre una asociación y una correlación?

Relaciones: simétrica vs asimétrica

- Se estudia primero la relación simétrica entre las variables, dígase también como la correlación.
- Seguido, se analiza la relación asimétrica entre las variables, denominada también cómo la regresión.
- Ningún de los dos anteriores busca competir a la hora de analizar datos. No son técnicas complementarias, dado que el análisis de correlación ayuda a la regresión en muchos análisis de soporte.

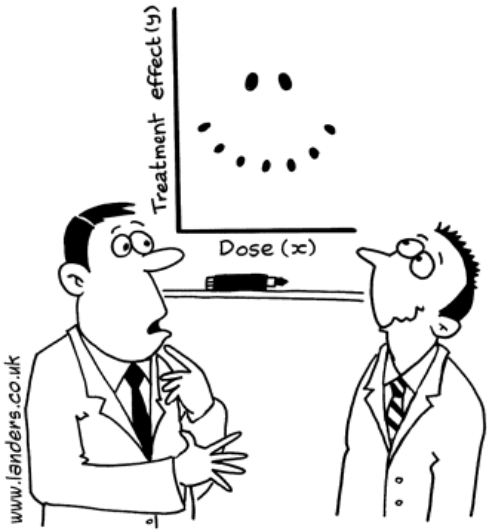




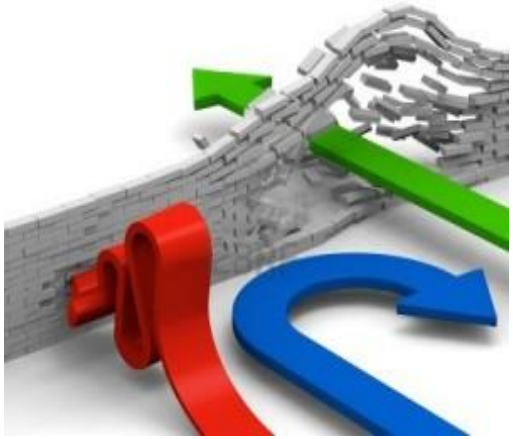
Índice

Análisis de correlación

- El análisis de correlación indica si dos variables están relacionadas o no.
- Por ejemplo la relación entre la edad y el aumento de salario, la edad y la producción de estrógenos, etc.
- Si el cambio en una variable está acompañado de un cambio en la otra, entonces se dice que las variables están correlacionadas.
- Al hablar de correlación, partimos de variables que son únicamente cuantitativas.



"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."



Análisis de correlación

- La correlación puede decir algo acerca de la relación entre las variables. Se utiliza para entender:

1. Relación es positiva o negativa
2. La fuerza de la relación

- En el análisis de datos, sobre todo en el análisis entre las variables cuantitativas, la correlación es una herramienta poderosa que brinda piezas vitales de información.

- Para medir la correlación, se suele utilizar el estadístico de rho, representado por la letra griega $\boldsymbol{\rho}$, siendo este el parámetro y $\boldsymbol{r}$ el estimador.

Análisis de correlación

- Para obtener la correlación estadística debemos de calcular el estadístico de rho, conocido también como la correlación de Pearson.

- La fórmula es la siguiente:

$$r = \frac{Cov(X,Y)}{s_X S_Y}$$

$$r = \frac{\sum_{i=1}^{n}(X_i - X) * (Y_i - Y)}{\sqrt{\sum_{i=1}^{n}(X_i - X)^2} * \sqrt{\sum_{i=1}^{n}(Y_i - Y)^2}}$$

$$r = \frac{n\sum_{i=1}^{n} XiYi - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{\left[n\sum_{i=1}^{n} {X_i}^2 - (\sum_{i=1}^{n} x_i)^2\right] * \left[n\sum_{i=1}^{n} {Y_i}^2 - (\sum_{i=1}^{n} Y_i)^2\right]}}$$

Análisis de correlación

- Si $r = 1$, existe una correlación positiva perfecta. El índice indica una dependencia total entre las dos variables denominada *relación directa*: cuando una de ellas aumenta, la otra también lo hace en proporción constante.

- Si $0 < r < 1$, existe una correlación positiva.

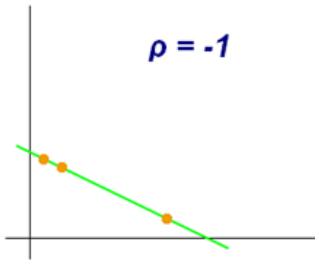
- Si $-1 < r < 0$, existe una correlación negativa.

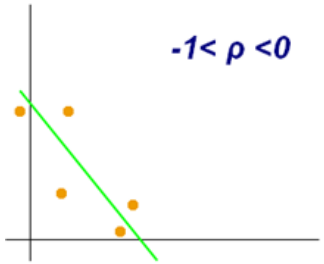
- Si $r = -1$, existe una correlación negativa perfecta. El índice indica una dependencia total entre las dos variables llamada*relación inversa*: cuando una de ellas aumenta, la otra disminuye en proporción constante.

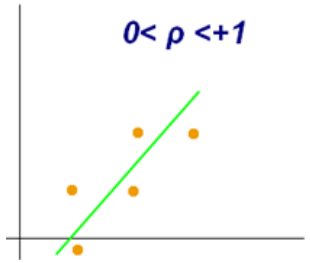
- Si $r = 0$, no existe relación lineal. Pero esto no necesariamente implica que las variables son independientes: pueden existir todavía relaciones no lineales entre las dos variables.

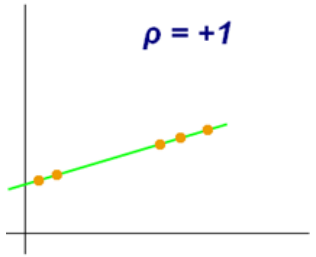
Análisis de correlación

Correlación negativa y positiva

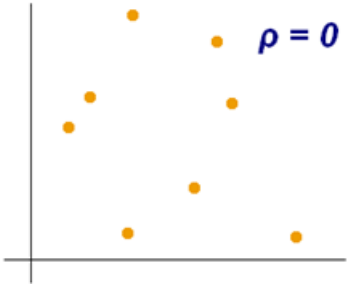








Correlación “nula”



Análisis de correlación

- En la aplicación de cierto medicamento, un doctor quiere ver si la aplicación de este tiene efecto según el peso y la edad.

- Para esto decide analizar antes la posible relación entre la edad y el peso de sus pacientes, y establecer una medida de relación para las características.

- ¿Cuál es el tipo de análisis que debe realizar?
  ¿Cuál sería el método que analiza lo anterior?
  ¿Cuál sería su expresión matemática?

- Los datos se presentan como siguen:

Edad	***Peso***
15	**60**
30	**75**
18	**67**
42	**80**
28	**60**
19	**65**
31	**92**

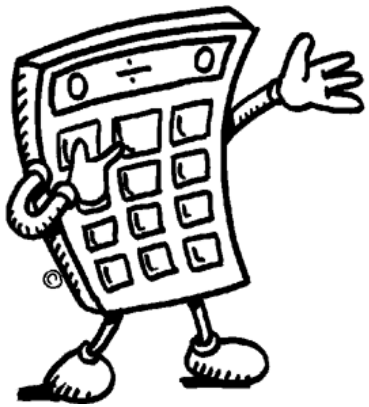
Análisis de correlación

- La fórmula:

$$r = \frac{n\sum_{i=1}^{n} XiYi - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{\left[n\sum_{i=1}^{n} {X_i}^2 - (\sum_{i=1}^{n} x_i)^2\right] * \left[n\sum_{i=1}^{n} {Y_i}^2 - (\sum_{i=1}^{n} Y_i)^2\right]}}$$

- Para obtener los resultados:

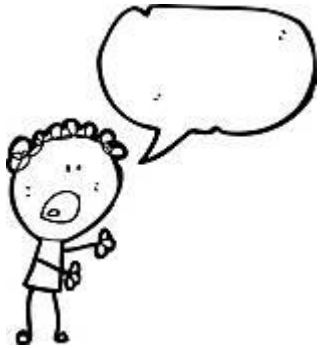
	Edad (X)	Peso (Y)	X^2	Y^2	XY
1	15	60	225	3600	900
2	30	75	900	5625	2250
3	18	67	324	4489	1206
4	42	80	1764	6400	3360
5	28	60	784	3600	1680
6	19	65	361	4225	1235
7	31	92	961	8464	2852
Total	183	499	5319	36403	13483



$$r = \frac{7 * 13483 - (183 * 499)}{\sqrt{[7 * 5319 - (183)^2] * [7 * 36403 - (499)^2]}} = 0{,}6564$$

Análisis de correlación

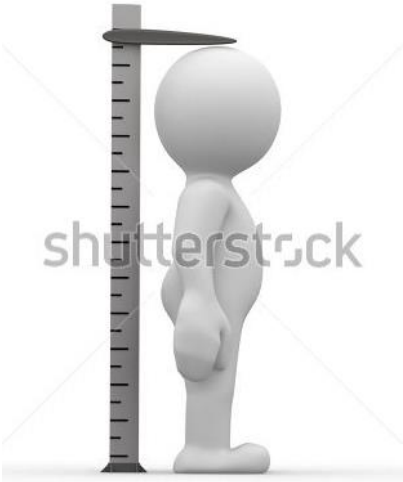
- Para su solución, se aplicó la fórmula de la correlación de Pearson.
- El resultado del coeficiente de correlación de Pearson resultó de 0,6564.
- Al ser el resultado del coeficiente de Pearson de 0.6564 esto indica una relación positiva.
- La relación entre el peso y la edad, o la edad y el peso, resultó ser positiva. Esto es, entre más años de la persona, esta suele poseer mayor peso, y viceversa.
- Nótese que no busca una explicación de una u otra variable, simplemente una relación entre ambas.



Índice

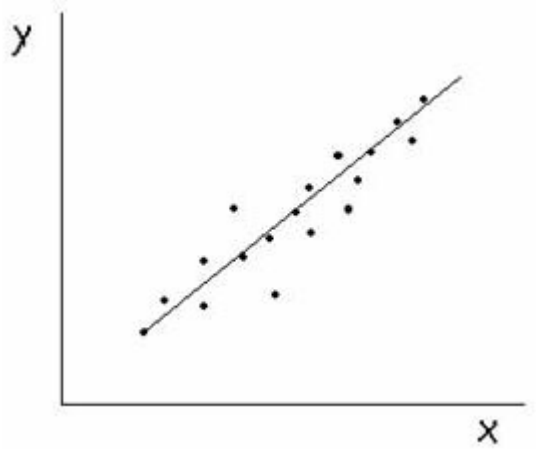
¿De dónde surge la regresión?

- Desarrollado por Sir Francis Galton a final del siglo XIX.
- Galton busco estudiar la relación entre las estaturas de los padres y de los hijos para predecir las edades.
- Galton desarrolló los modelos matemáticos para regresar la estatura del padre y así saber para una edad determinada la estatura del hijo.
- El descubrimiento causó un revolución analítica importante, por lo que el término de regresión ha perdurado en el tiempo.
- Siempre se mantuvo la idea de buscar una función matemática que explicase la relación causal entre dos variables.



Regresión bivariada y regresión multivariada

Regresión bivariada

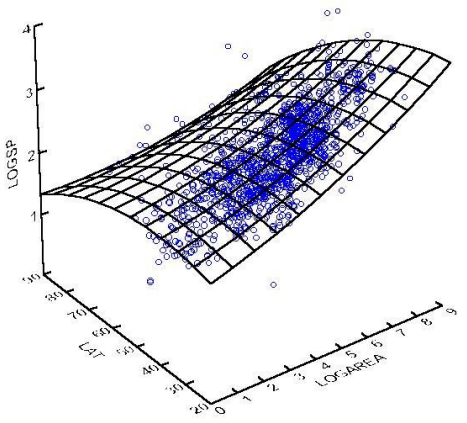


Una variable dependiente (Y)
Una variable independiente (X)

$$Y \sim X$$

$$\begin{bmatrix} Y1 \\ Y2 \\ \cdots \\ Y_n \end{bmatrix} \sim \begin{bmatrix} X1 \\ X2 \\ \cdots \\ X_n \end{bmatrix}$$

Regresión multivariada



Una variable dependiente (Y)
Dos o más variables independientes (X1, X2 , …, Xp)

$$Y \sim X$$

$$\begin{bmatrix} Y1 \\ Y2 \\ \cdots \\ Y_n \end{bmatrix} \sim \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

Índice

Etapas de un modelo de regresión bivariado

- Modelo de regresión cuenta con 6 etapas:

-Relación entre las variables

-Estimación de la recta de mejor ajuste (modelo de regresión)

-Diagnóstico del modelo de regresión

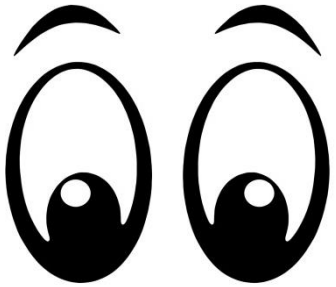
-Medidas remediales

-Estadísticas de bondad y ajuste

-Inferencia y prueba de hipótesis de los coeficientes del modelo

- En la estimación de un modelo de regresión, se debe de seguir este orden.

- Las etapa de “Diagnóstico del modelo de regresión” y “Medidas remediales” y “Estadísticas de bondad y ajuste” no se presentarán en el presente capítulo.



Relaciones entre las variables: e principio de la regresión

- Muchos estudios se basan en la creencia de identificar y cuantificar alguna relación asimétrica entre dos o más variables.

- Esto es, para una variable "Y", esta podría depender de cierta medida de otra variable "X".

- El análisis de regresión trata de **cuantificar o establecer una relación funcional** en la relación de estas variables.

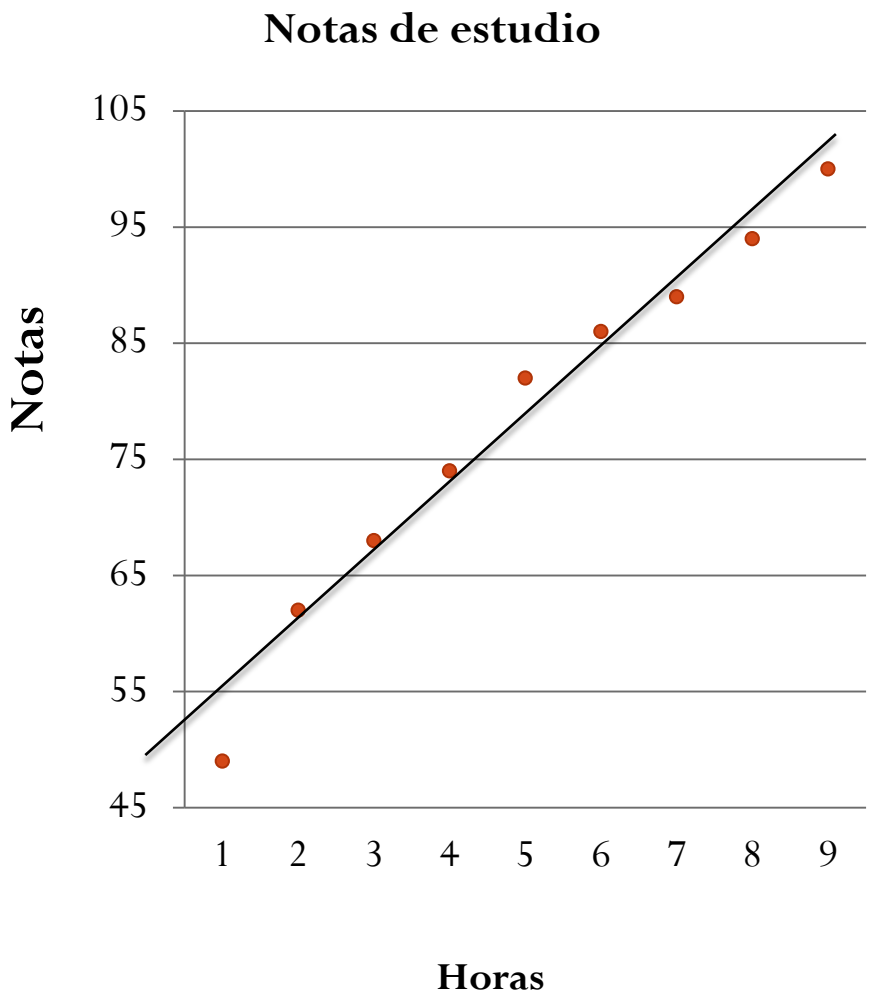
- Generalmente, se dice que "Y" depende de "X", donde ambas son variables cualesquiera con alguna relación.

Y es una función de X $\approx$ $Y = f(X)$

El principio de la regresión: ejemplo (1/2)

El director de una escuela desea analizar la relación entre las notas de los estudiantes y el tiempo que pasan estudiando.

- Se quiere saber si las notas están influencias por la cantidad de horas de estudio.
- La variable dependiente es "Notas" y la independiente "Horas" .
- El ejemplo muestra una relación entre las notas y las horas de estudio: entre más horas de estudio, mejores serán las calificaciones.



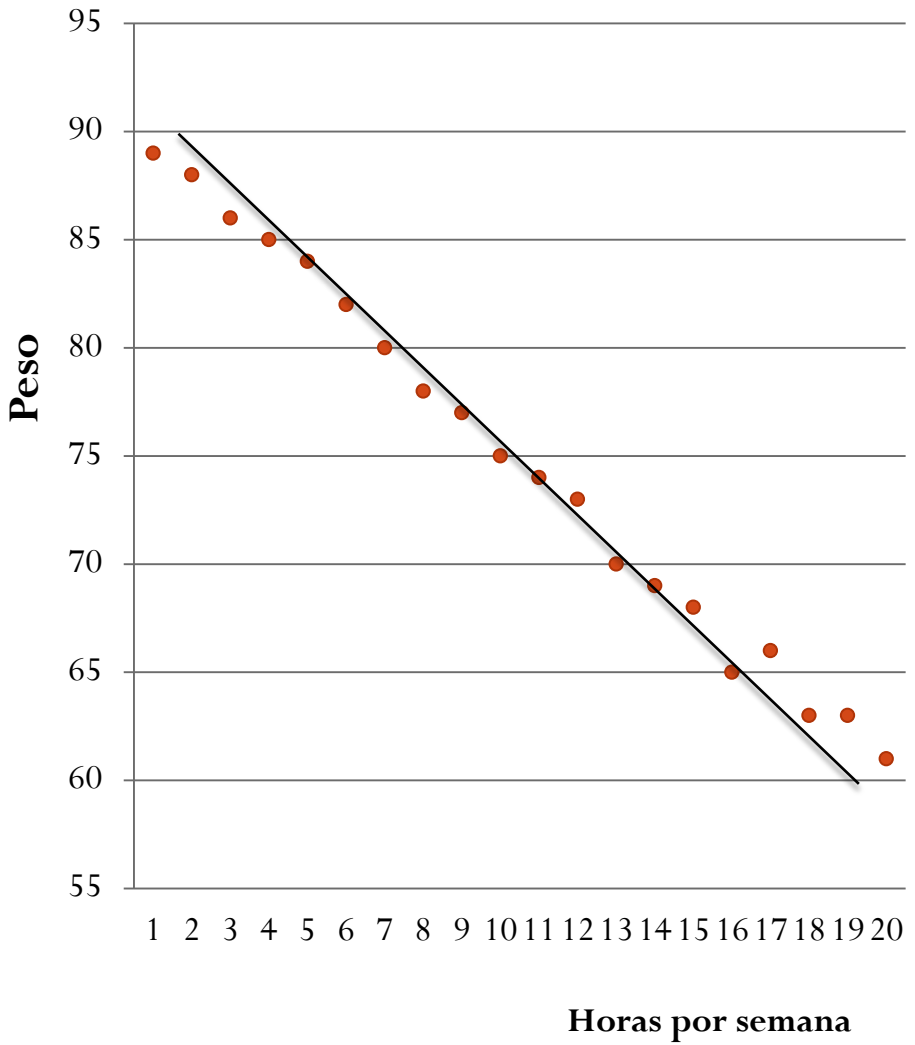
El principio de la regresión: ejemplo (2/2)

- El jefe de un gimnasio quiere investigar si las personas que bajan más de peso se debe a una rutina de más horas de ejercicio por semana.

→ Se desea averiguar si disminuir peso está en función de la cantidad de horas de ejercicio por semana.

→ La variable dependiente es el "peso", y la independiente "horas por semana".

→ Se observa que entre más horas de ejercicio en el gimnasio, mayor es la disminución de las personas (en kilogramos).



El principio de la regresión : dependencia e independencia

- Debido a que "Y" depende de "X", se dice que "Y" es la variable dependiente, y "X" es la variable independiente.
- Es importante **identificar cuál es la variable dependiente y cuál es la variable independiente** en el modelo de regresión.
- Esto depende de la lógica y del problema de investigación que se tenga .

Variable dependiente (Y): es la variable que se desea explicar o predecir; va a depender por lo tanto de otra variable para entender su cambio.

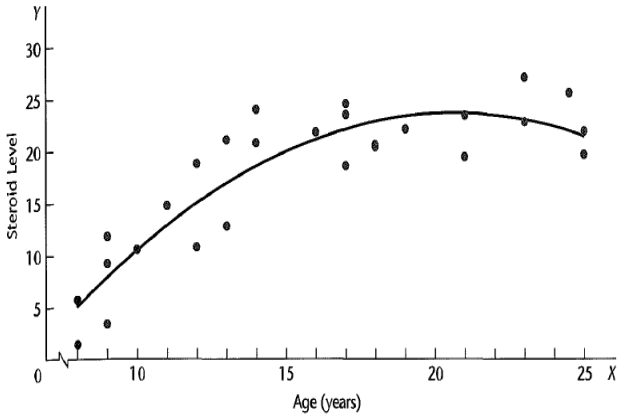
Variable independiente (X): es la variable independiente. Está es la que aporta el factor explicativo al movimiento de l variable dependiente.

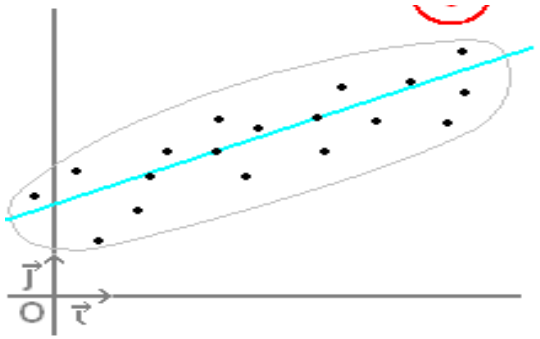
El principio de la regresión: tipo de relación entre “Y” y “X”

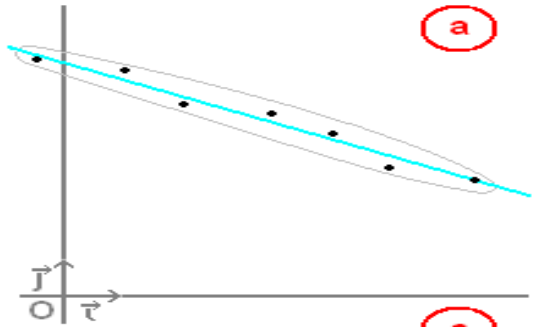
- En la relación entre las variables “X” y “Y” se pueden dar diversos tipos de asociaciones

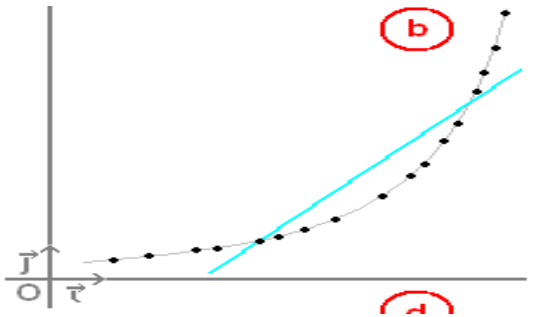
- Las relaciones pueden ser lineales (ejemplos 1 y 2), curvilíneas (nivel de esteroides en una persona según la edad), logarítmica (ventas de un empresa según cambios en la producción), etc.

- En la regresión lineal, se sostiene ***que a medida que X cambia,Y cambia en una cantidad constante.***









El principio de la regresión: tipo de relación entre “Y” y “X”

- Es importante hacer este comentario, ya que **NO todas las relaciones en la vida son lineales**. Sin embargo, se sigue bajo el paradigma de regresión lineal simple.

- Ejemplo:

-El nivel de estrógenos en las mujeres de 18-32 años. Es una función que aumenta, se estabiliza y vuelve a disminuir ($-x^2$)

-La producción inicial de una nueva fábrica ($\ln x$).

- Por lo tanto, se debe considerar de igual forma las relaciones lineales como las curvilíneas.

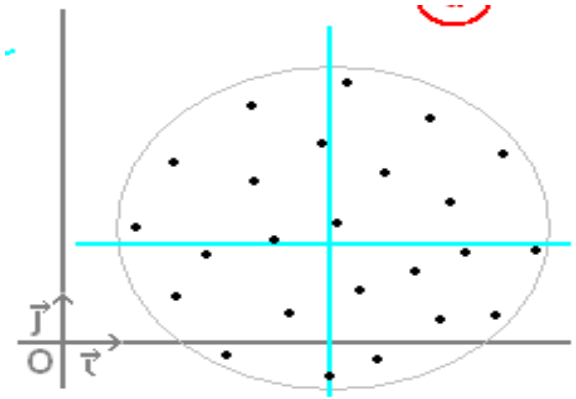
- Aunque sean relaciones curvilíneas, logarítmicas, u otras y no necesariamente lineales, estas seguirán siendo modelos linéales con transformaciones en las variables, para linearisar las respuesta.

El principio de la regresión: tipo de relación entre “Y” y “X”

- De igual forma, hay que contemplar el hecho de que del todo no haya relación.

- Aunque se aplique la regresión, es posible que no se obtenga resultados satisfactorios.

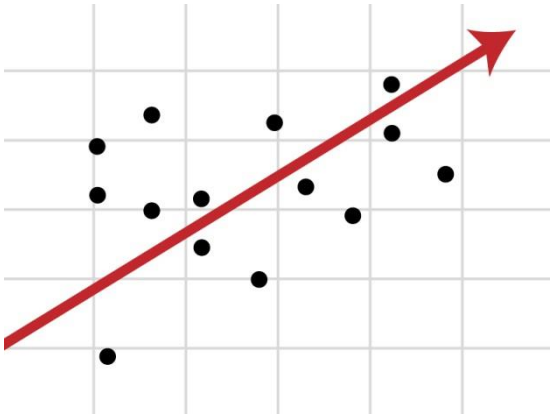
- Si se grafica los pares “Y” y “X”, y realmente no se observa alguna relación, simplemente esto quiere decir que no hay una relación entre las variables de interés, y por lo tanto “Y” no se puede ser explicada mediante “X”.



Índice

Determinación del modelo de regresión : concepto

- Para determinar un modelo de ***regresión simple lineal***, el objetivo se centra en determinar una recta capaz de modelar la relación de los datos.



- El modelo de la recta se expresa de la siguiente forma:

$$\hat{Y} = B_0 + B_1 X$$

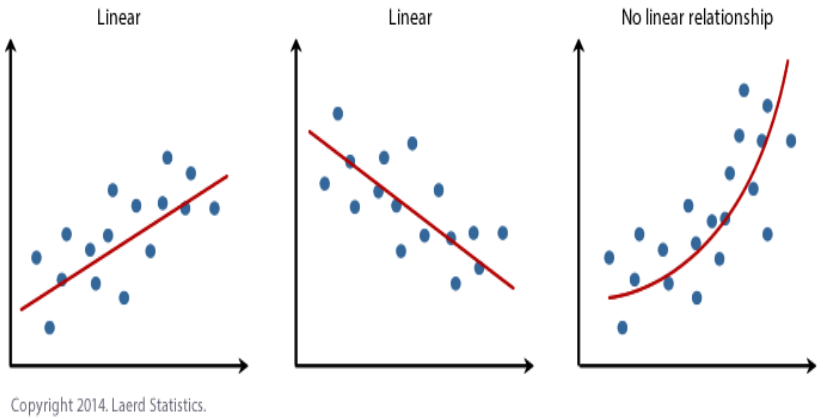
donde,

$\hat{Y}$ es la estimación de la recta para un valor X

X es el cambio en la variable independiente

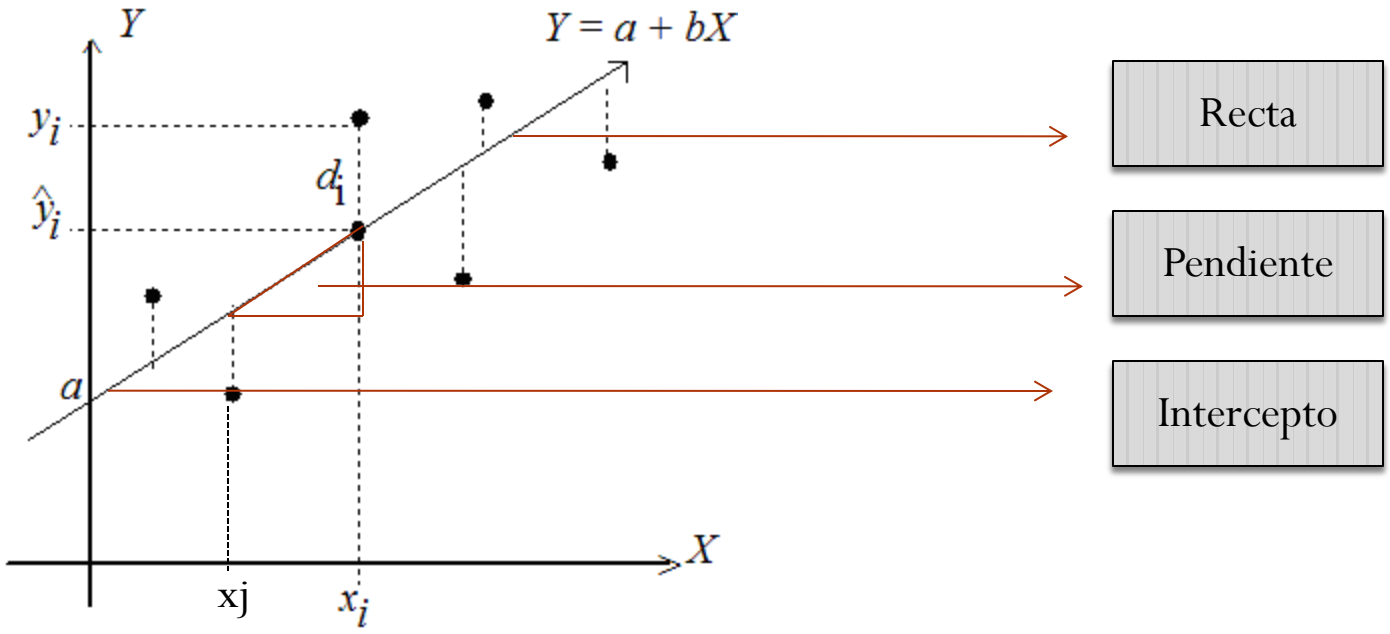
B_0 es el intercepto

B_1 es la pendiente de la recta



Determinación del modelo de regresión : concepto

- Un ejemplo, de la estimación de un modelo de regresión es el siguiente.



Determinación del modelo de regresión: ejemplo

- Supóngase que los puntos del ejemplo anterior representan los casos, y cada caso posee un valor "X" (variable independiente) y un valor "Y" (variable dependiente).

- Supongamos que una estimación resultó como sigue:

$$\hat{Y} = 5 + 2X$$

- En este caso,

  5 sería el intercepto

  2 sería la pendiente

La pendiente expresa que para cada cambio de unidad en "X", se tiene un cambio de "2" en Y. De igual forma, la variable "Y", es una función de los valores de "X".

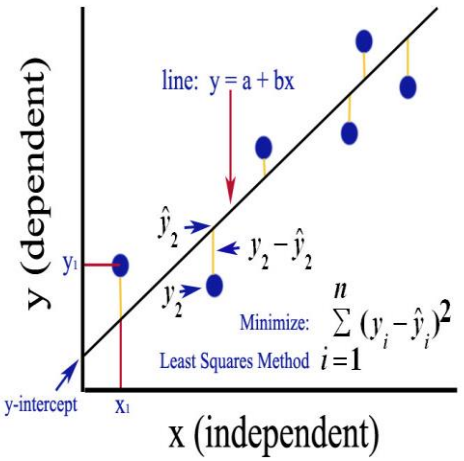
Determinación del modelo de regresión : concepto

- En una regresión lineal bivariada, la mayor importancia del modelo se centra en el valor de la pendiente:

  -Si esta es positiva, habrá una relación positiva.

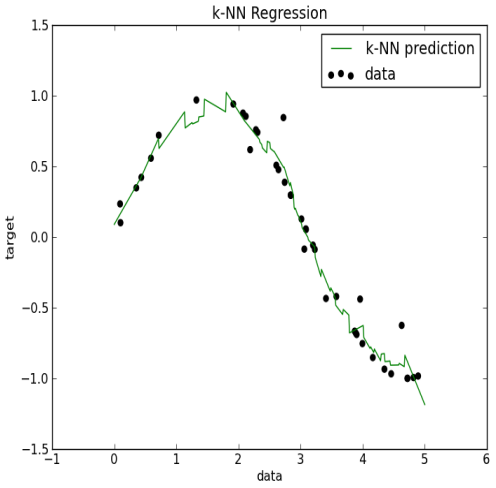
  -Si esta es negativa, habrá una relación negativa.

  -Si es nula (igual a "cero"), no habrá relación del todo

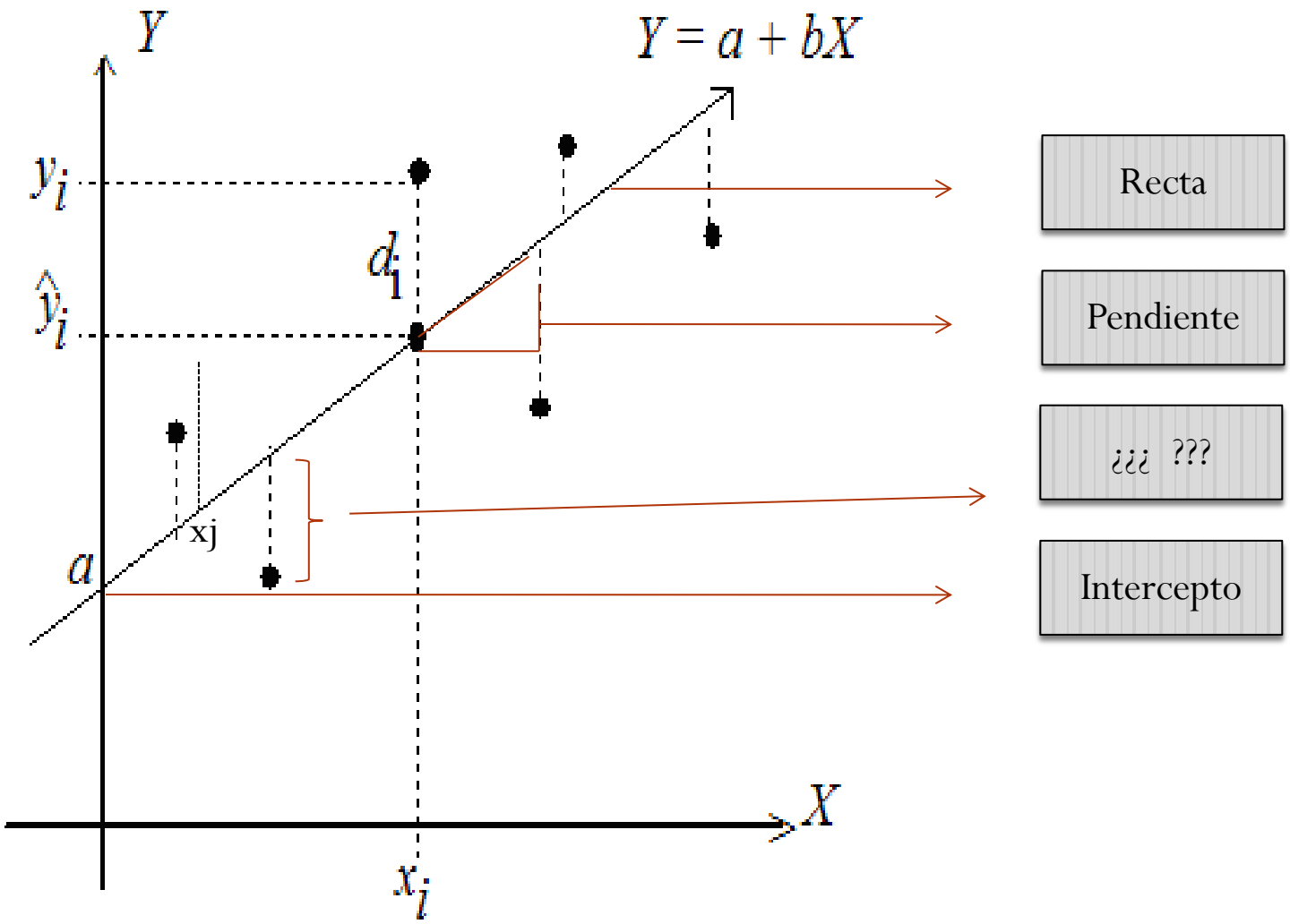


- Lo interesante es que si la pendiente es "nula" o igual a cero estadísticamente no existe ninguna relación entra las variables(en este caso no vale la pena estimar la función de regresión).

- Por tanto, "X" no puede utilizarse como variable explicativa de "Y", en caso de que la relación funcional brinda una relación prácticamente nula.



Determinación del modelo de regresión: matemática vs estadística



Modelo matemático o determinístico

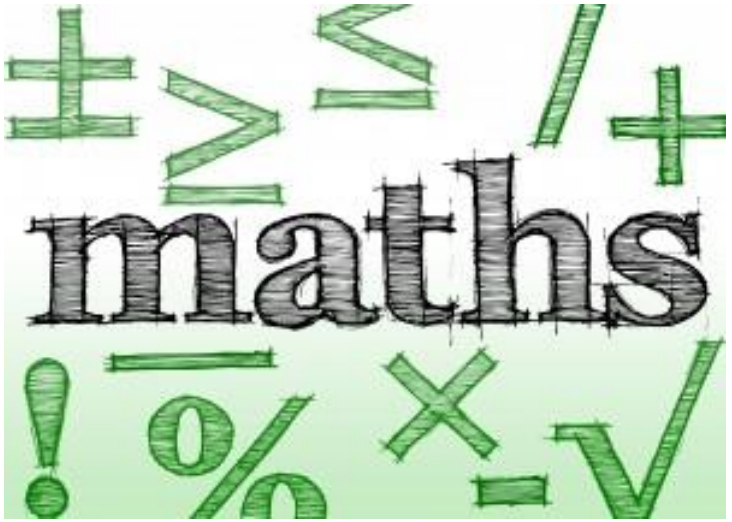
- Es importante diferencias entre un modelo determinístico (sin error), y un modelo de regresión (llamado estocástico, o con error).

- Una relación determinística puede expresarse mediante una fórmula matemática que siempre brindara un resultado exacto.

- Ejemplos : depósitos a plazo, velocidad, precio de una canasta básica, etc.

- Matemáticamente: $Y = a*X$

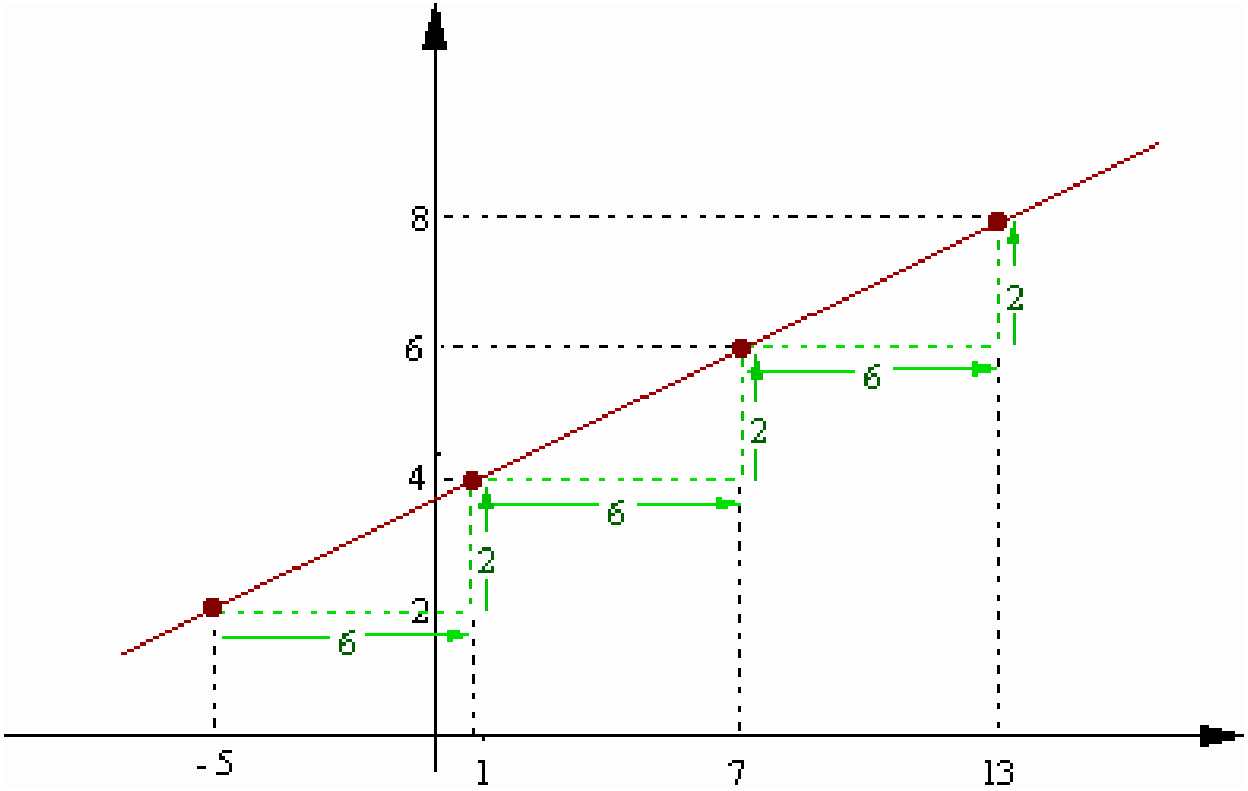
  En la relación anterior, mediante cambios constantes en X, se tendrá siempre el valor exacto de Y.



Esta relación es determinística, la relación es exacta y no hay error

Modelo matemático o determinístico

- Relación determinística: relación perfecta todos los puntos caen sobre la curva
- Ejemplos: estimación de los Km/hora de un vehículo.



Recta pasa por todos los puntos

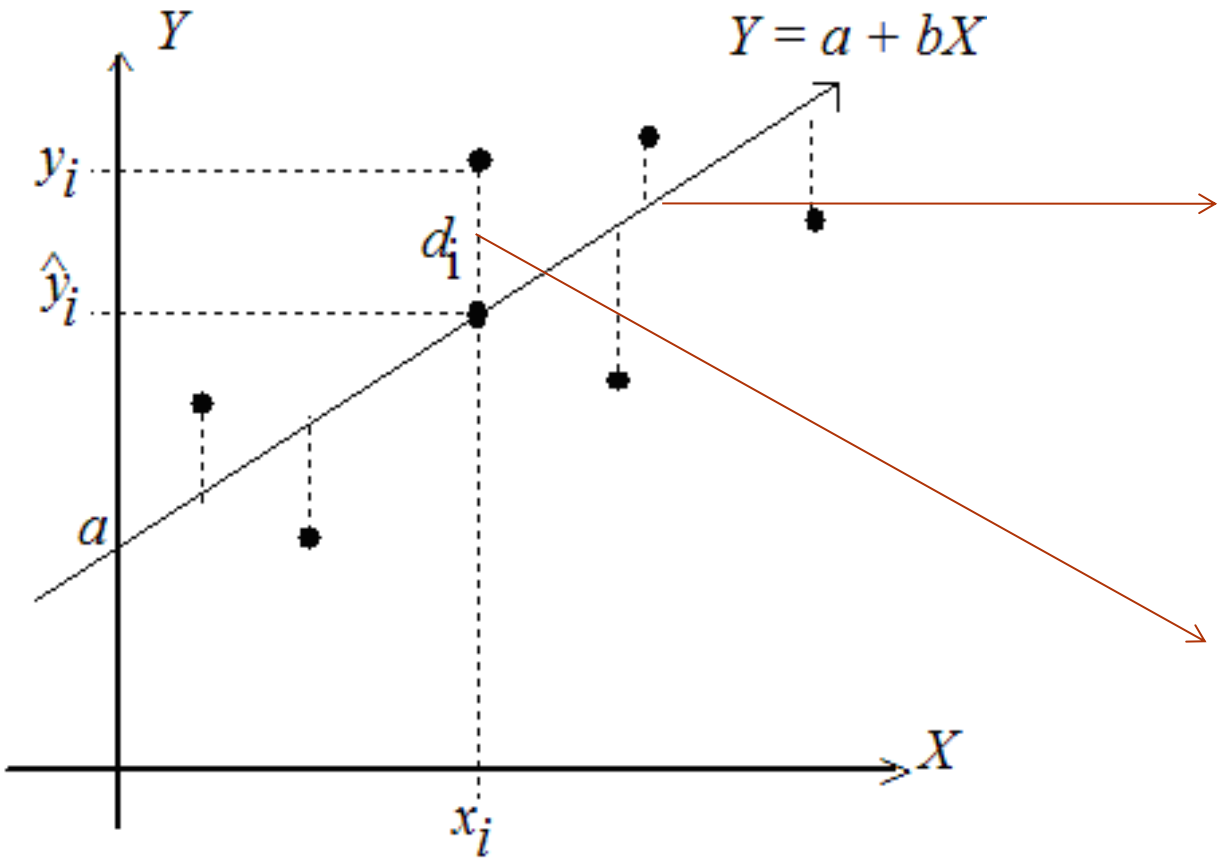
Ajuste perfecto

Modelos estocástico o estadístico

- Infortunadamente, la mayoría de los hechos fuera de la matemática y la física no son determinísticos, y contienen error.
- En casi el 100% de veces se encuentra que al utilizar una variable para explicar otra existe un cierto nivel de variabilidad en la relación establecida.
- Esto se debe a que entre la variable dependiente e independiente existen relaciones con cierta aleatoriedad.
- Por lo tanto, habrá una diferencia o, mejor dicho, un grado de **error** en el intento por explicar o predecir la variable dependiente.

Modelos estocástico o estadístico

En resumen: las relaciones estocásticas contienen un componente denominado **error**, el cuál es la diferencia entre el valor de la observación y el valor predicho por la recta de regresión.



Recta

Error: diferencia entre el valor estimado y la observación Yi

Modelos estocástico o estadístico

- Un modelo de regresión se constituye de dos partes.

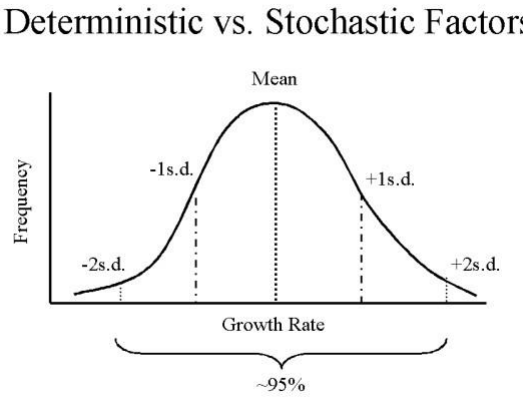
- ***La parte determinística***: es la estimación de la recta de mejor ajuste. Esta se expresa como:

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

- ***La parte estocástica***: es la diferencia entre la recta de estimación y el valor de la observación. Esta se expresa como:

$$\varepsilon = Y - \hat{Y}$$

- La unión entre estos dos elementos constituyen un modelo de regresión, el cual intenta brindar la mejor aproximación en la predicción de las observaciones.



Deterministic	Probabilistic
▪ Each activity has a planned value. ▪ For the schedule each task has a predecessor and a successor. ▪ The longest path through the network is the critical path. ▪ The total duration of the project is a fixed value - it is deterministic. ▪ The total cost is the sum of all the activity costs. ▪ Risks are defined and handled as static entities.	▪ The program elements are not random, but they are random variables drawn from a probability distribution. ▪ Three point estimates "can" be used to describe task duration random variables. ▪ The total duration of the project is a random number. ▪ The total cost is a random number. ▪ Risks are stochastic processes that have probabilistic outcomes for cost, schedule and technical performance.

Modelos estocástico o estadístico

- Finalmente, cuál es la diferencia entre los siguientes dos modelos:

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

$$Y = b_0 + b_1 X + \varepsilon$$

$$Y_i = b_0 + b_1 X + \varepsilon_i$$

- Y cuál es la diferencia entre los siguientes dos modelos:

$$\hat{Y} = \beta_0 + \beta_1 X$$

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

Determinación del modelo de regresión

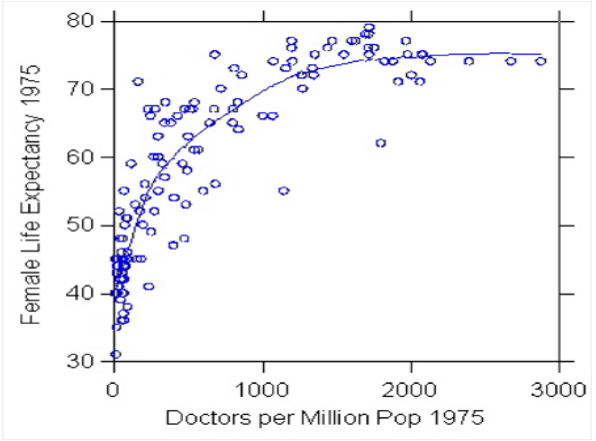
- Finalmente, comprendido el concepto de la recta de mejor ajuste, modelo determinístico, modelo estocástico y error, entre otros, solo falta saber como se estima de recta de mejor ajuste.

- Se recuerda que la recta de mejor ajuste tiene por fórmula:

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

- De igual forma, se debe preguntar como interactúa la parte estocástica para así determinar la recta de mejor ajuste, la cual es determinística.

- Todo esto se explica a continuación.



Estimación: recta de mejor ajuste.

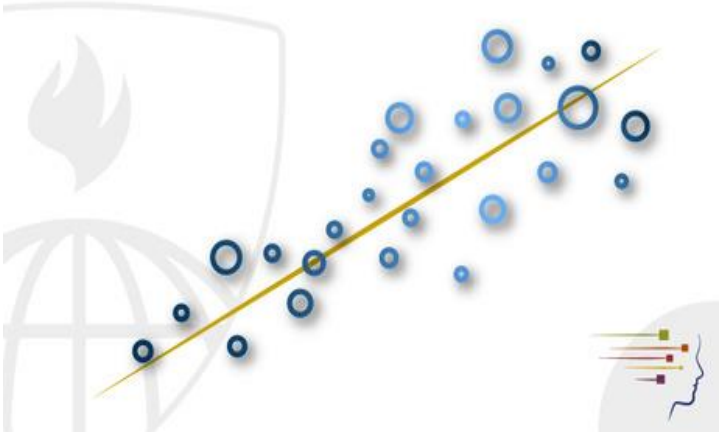
- El propósito del análisis de regresión es determinar la mejor recta que se ajuste a los datos.

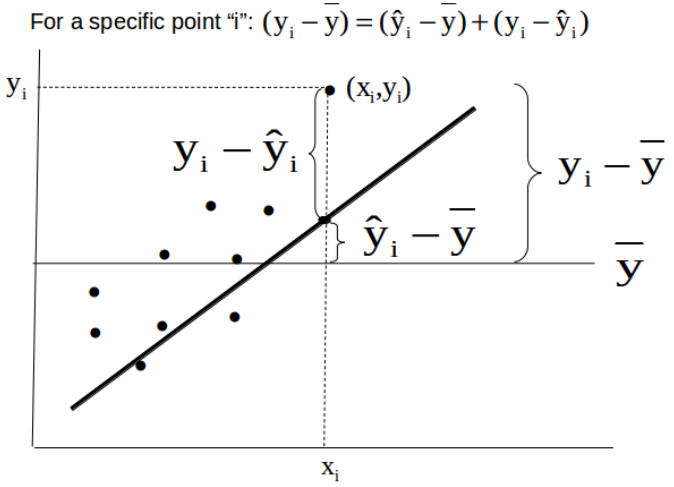
- Esto es, para la ecuación: $\hat{Y} = b_0 + b_1 X$

  -Se quiere estimar los valores de **b0** y **b1** que mejor se adecuen a los datos (la recta que se aproxime más a la totalidad de los datos).

- El método que se utiliza para obtenerla se denomina **Mínimos Cuadrados Ordinarios** *(*MCO).

- Los MCO producirá la mejor de todas las ecuaciones posibles para aproximar los datos en las relaciones lineales.

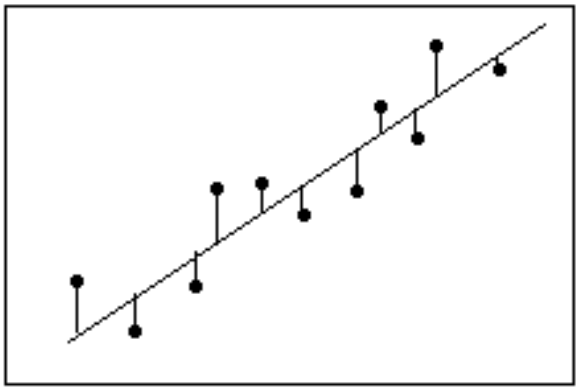




Estimación: recta de mejor ajuste.

- El fundamento de que con los MCO se obtenga la mejor recta de estimación de cualquier otra, se debe a que los errores (o residual), se minimiza con este método:

- De ahí el nombre de "Mínimos Cuadrados Ordinales": produce una recta tal que la suma de los errores al cuadrado es menor de lo que sería con cualquier otra recta.

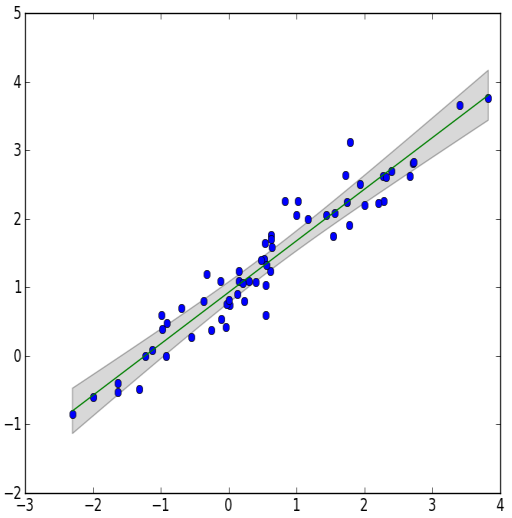


La suma de errores al cuadrado se minimiza con $\sum (Y_i - \hat{Y}_i)^2 = min$

Estimación de la recta de mejor ajuste

- Entonces, el objetivo es calcular la recta de mejor ajuste:

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



- Para esto básicamente se debe llevar a cabo 2 etapas:

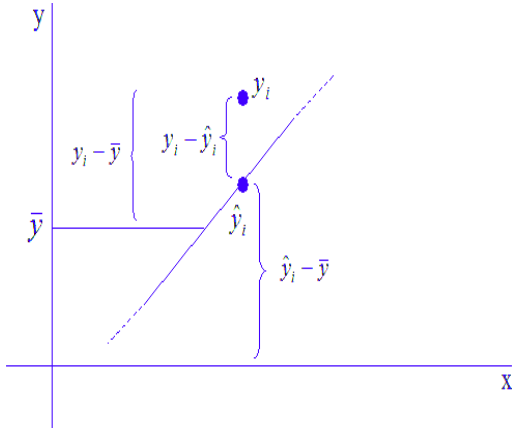
1. Cálculo de la suma de cuadros y productos cruzados de las variables en cuestión, "X" y "Y":

a. Suma de cuadrados de "X" (SCx).

b. Suma de cuadrados de "Y" (SCy).

c. Suma de cuadrados cruzada de "X" y "X" (SCxy).

2. Luego, calcular la pendiente b1, y el intercepto b0.



Estimación de la recta de mejor ajuste

- 1. Obtención de las sumas de cuadrado y productos cruzados

Suma de los cuadros de "X" $$SCx = \sum (X_i - \bar{X})^2$$

Suma de los cuadros de "Y" $$SCy = \sum (Y_i - \bar{Y})^2$$

Suma de los cuadros de "X" e "Y" $$SCxy = \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

Estimación de la recta de mejor ajuste

- 2. Estimación de la pendiente e intercepto en la recta de regresión:

Pendiente: $$b_1 = \frac{SCxy}{SCx}$$

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

Con esta información tenemos la recta de regresión: $$\hat{Y} = b_0 + b_1 X$$

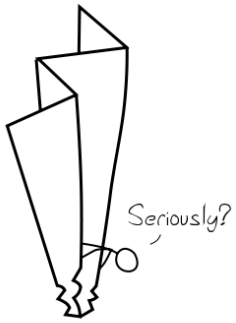
Estimación de la recta de mejor ajuste: ejemplo

- Ejemplo:

La gerencia de Hop Scoth Airlines, la aerolínea más pequeña del mundo, considera que existe una relación directa entre los gastos publicitarios y el número de pasajeros que escogen viajar por Hop Scoth. Para determinar si esta relación existe, y si es así cuál podría ser la naturaleza exacta, los estadísticos empleados por Hop Scotch decidieron utilizar los procedimientos MCO para determinar el modelo de regresión.

- Se recolectaron los valores mensuales por gastos de publicidad y número de pasajeros para los n=15 mese más recientes. Los datos se presentan a continuación (se expresan en unidades de 1=1000).

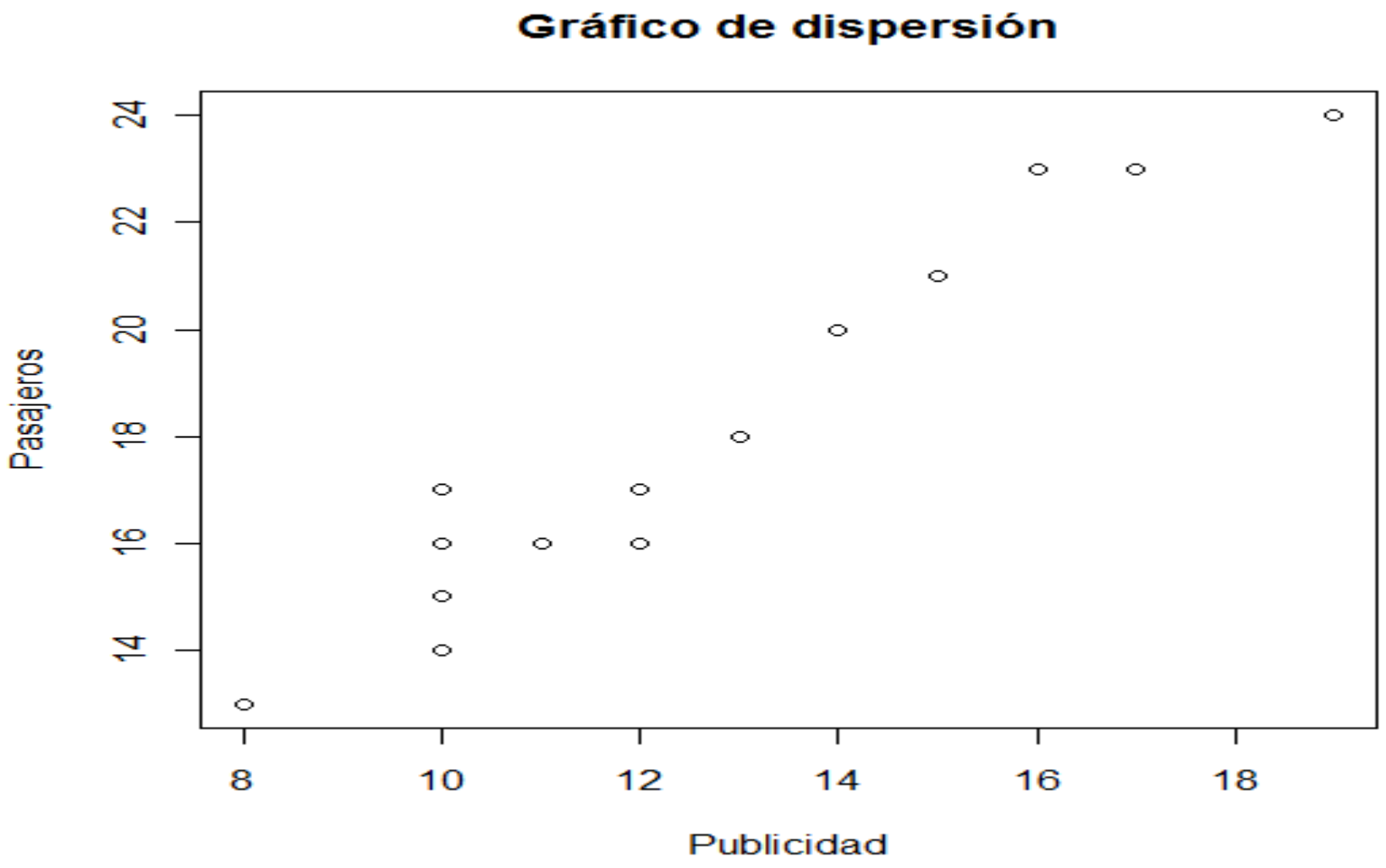
- Antes de estimar el modelo de regresión, qué se podría decir sobre el tipo de datos empleados.



Estimación de la recta de mejor ajuste: ejemplo

Observaciones	Publicidad (X) en miles $	Pasajeros (Y) en miles $
1	10	15
2	12	17
3	8	13
4	17	23
5	10	16
6	15	21
7	10	14
8	14	20
9	19	24
10	10	17
11	11	16
12	13	18
13	16	23
14	10	15
15	12	16

Estimación de la recta de mejor ajuste: gráfico



Estimación de la recta de mejor ajuste

- Antes de realizar cualquier tipo de análisis, se debe preguntar:

¿Cuál es el tipo de análisis a llevar a cabo?

- En este caso se trata de un análisis de regresión, ya que se considera que existe una relación directa entre los gastos publicitarios y el número de pasajeros que escogen viajar por Hop Scoth.

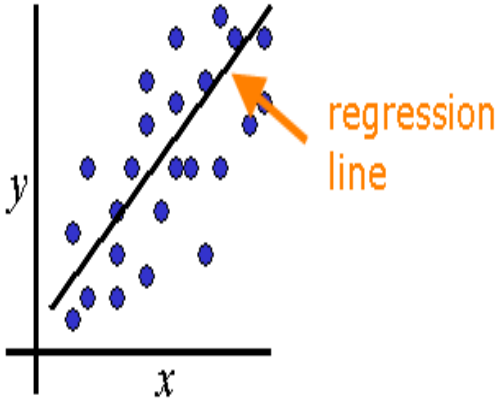
- Se quiere conocer el grado de relación de una variable en función de otra…

- En términos prácticos, se desea saber si los gastos en publicidad tienen relación causal con el número de pasajeros que viajan por Hop Scoth.

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

Estimación de la recta de mejor ajuste: ejemplo

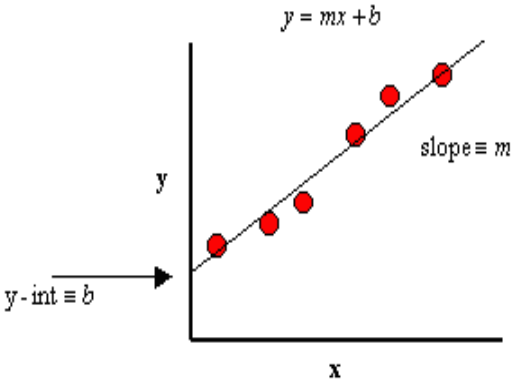
- Formalmente, cada variable debe ser analizada con estadísticas descriptivas, a lo cual denominamos como inspección de los datos (histogramas, valores faltantes, valores extremos, etc. en cada una de las variables). Esto lo omitimos por ahora.



- Sabiendo que se debe estimar una regresión, y que además se cuenta con una relación de dos variables (bivariada), se debe estimar un modelo de regresión simple o bivariada.

- Los siguientes son los cálculos para estimar el siguiente modelo, la recta de mejor ajuste:

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



Estimación de la recta de mejor ajuste: ejemplo

1. Cálculo de la suma de cuadros y productos cruzados de:

Suma de los cuadros de "X"

$$SCx = \sum (X_i - \bar{X})^2$$

$$SCx = 137.7333$$

Suma de los cuadros de "Y"

$$SCy = \sum (Y_i - \bar{Y})^2$$

$$SCy = 171.7333$$

Suma de los cuadros de "X" "Y"

$$SCxy = \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

$$SCxy = 148.9333$$

Estimación de la recta de mejor ajuste: ejemplo

- Cálculo de los promedios de las variables:

$$\bar{X} = \frac{187}{15} = 12.4667 \qquad \bar{Y} = \frac{268}{15} = 17.8667$$

- 2. Estimación de la pendiente e intercepto en la recta de regresión (el resultado se debe multiplicar por 1000) :

Pendiente: $b_1 = \frac{SCxy}{SCx}$ $b_1 = 1.08$

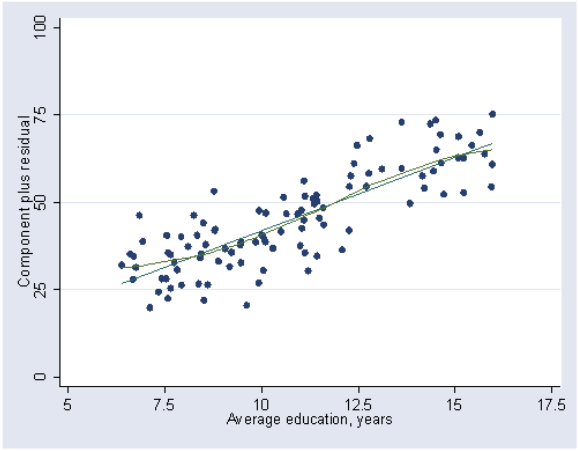
Intercepto: $b_0 = \bar{Y} - b_1\bar{X}$ $b_0 = 4.3865$

Estimación de la recta de mejor ajuste: ejemplo

- Entonces, de acuerdo a los valores de b1 y b0, el modelo de regresión sería el siguiente:

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

$$\hat{Y} = 4.3865 + 1.08 X_i$$



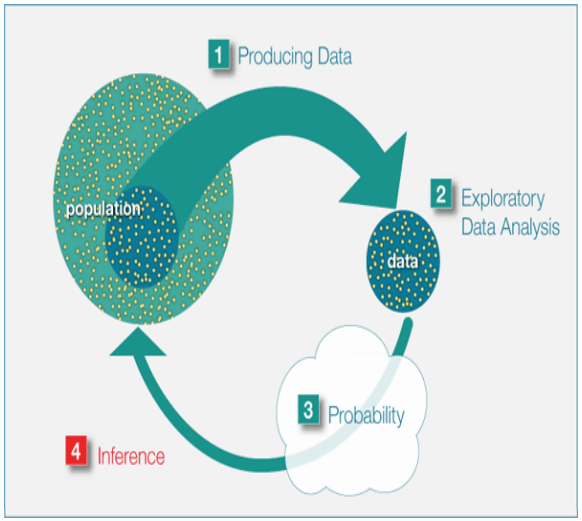
- Ahora: ¿Interpretación de este modelo?

→ Pista: la interpretación se hace a partir de la estimación del valor **b1** (1.08*1000=1080). ¿Qué se podría decir en términos del problema en cuestión?

Índice

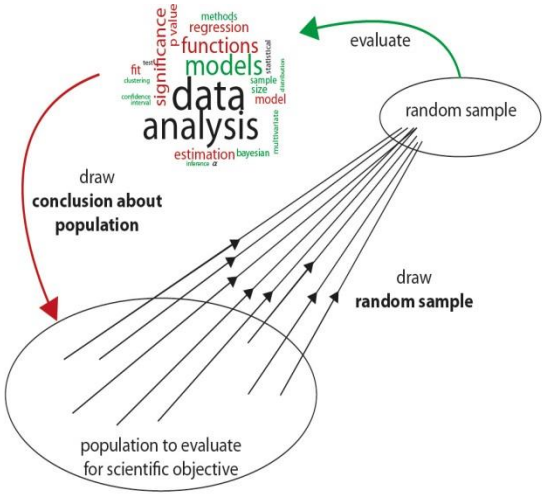
Estimación y prueba de hipótesis

- Para el caso de Hop Scoth Airlines, los resultados sugieren una relación entre las ventas y la publicidad. El coeficiente de regresión b1=1.08 (pendiente) y el coeficiente de determinación r=0.9683 (r^2=0.9378), indican que, a medida que hay un aumento en los gastos publicitarios, se produce un aumento en el número de pasajeros.



- Sin embargo, estos resultados se basan en una muestra de sólo n=15 observaciones. Se debe hacer la siguiente pregunta: ¿los resultados anteriores pueden ser inferidos a nivel poblacional?

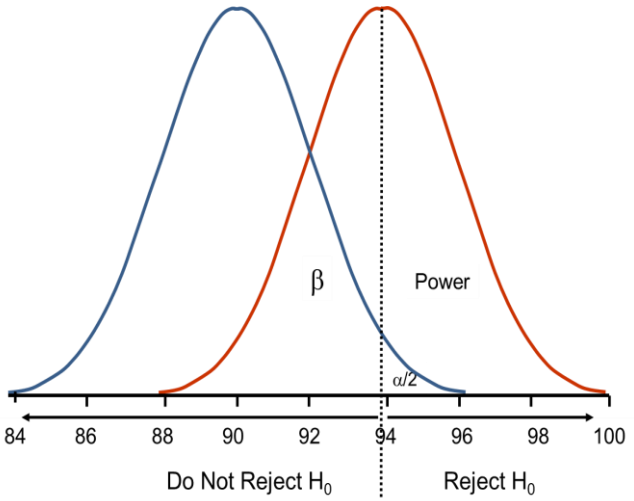
- Existe la posibilidad que al haber error de muestreo, los parámetros poblacionales no son distintos de cero, y por ende, los resultados antes obtenidos no son predictores del crecimiento del número de pasajeros.



- Se debe verificar que los resultados de los coeficientes pueden ser inferidos a nivel población.

Estimación y prueba de hipótesis

- Para probar lo anterior, debemos hacer prueba de hipótesis a nivel de los coeficientes. Estas pruebas se plantean como sigue:



1. Prueba a nivel del coeficiente B_1

$$H_0: B_1 = 0$$

$$H_A: B_1 \neq 0$$

2. Prueba a nivel del coeficiente ρ

$$H_0: \rho = 0$$

$$H_A: \rho \neq 0$$

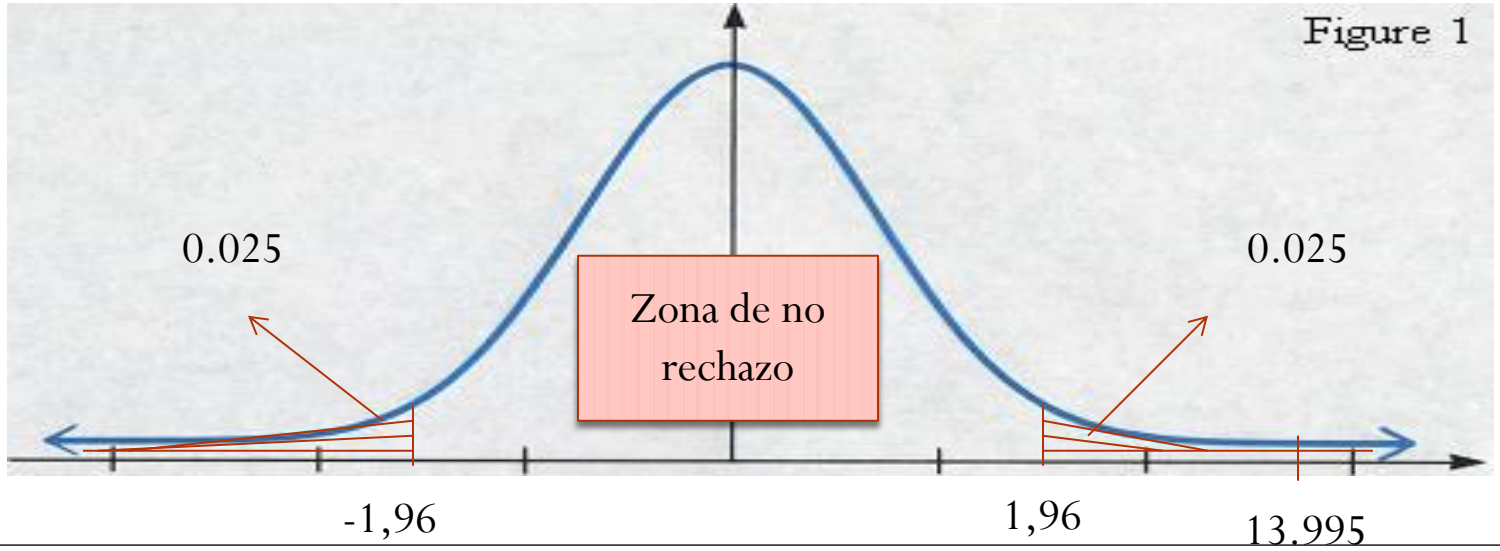
Veremos primero la prueba del coeficiente Rho (ρ), seguido de la prueba del parámetro *Beta* (β).

Estimación y prueba de hipótesis : recordatorio

- Para la construcción de una prueba de hipótesis, se requiere proceder en dos etapas:

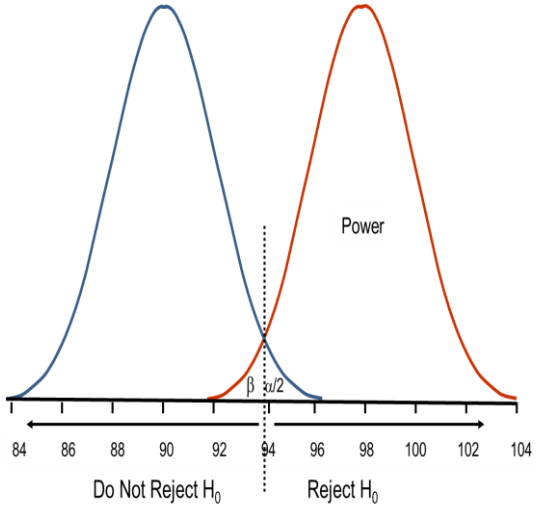
1. Calcular el estadístico asociado a la prueba de hipótesis.

2. Comparar el estadístico calculado o de prueba con el estadístico procedente de una función de distribución teórica (acá la Z de la normal estándar). A partir de la comparación, determinar si la hipótesis se rechaza o no se rechaza.



Estimación y prueba de hipótesis : ρ

- La primera prueba de hipótesis que se debe de llevar a cabo es el probar si existe o no una relación lineal entre las variables puestas en causa. Para esto, que mejor prueba de hipótesis que el probar si el coeficiente ρ es diferente de *0*.

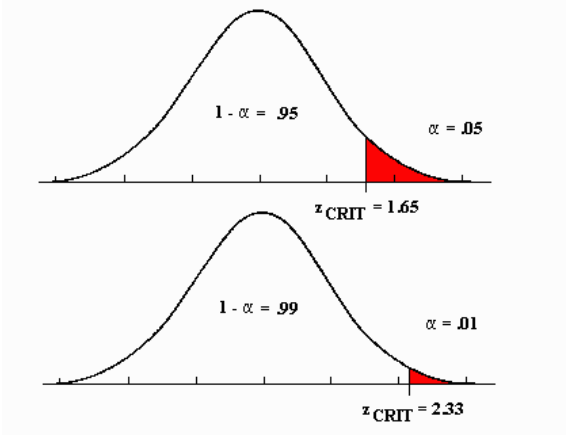


- Como el análisis respecto a la correlación entre pasajeros y publicidad se basa en los datos muestrales, el error de muestreo podría llevarnos a conclusiones no apropiados. Se debe probar si la relación general entre las variables es fidedigna.

- Puede ser que la correlación a nivel poblacional sea cero y que una muestra engañosa hizo que se asumiera equivocadamente una relación. Por consiguiente se debe probar la siguiente hipótesis:

$$H_0: \rho = 0$$

$$H_A: \rho \neq 0$$



Estimación y prueba de hipótesis : ρ

- Se calcula el estadístico de prueba para la hipótesis de presencia de relación lineal:

Prueba Z para el coef. de regre. poblacional $$Z = \frac{r - \rho}{S_r}$$

- En donde tenemos que S_r es el error estándar del coeficiente de correlación

Error estándar del coef. de correlación $$S_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

Estimación y prueba de hipótesis : ρ

- Para el caso de Hop Scotch Airlines:

$$S_r = \sqrt{\frac{1 - 0.93776}{15 - 2}} = 0.069 \qquad Z = \frac{0.9683 - 0}{0.069} = 13.995$$

- Si $\alpha = 0.05$ y $Z_{0.05} = \pm 1.96$:

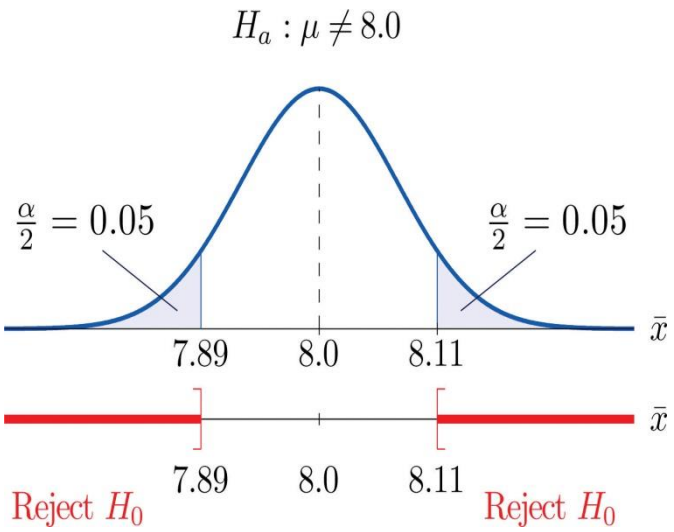
La regla de decisión es: "No rechazar si t está entre ± 1,96. De otro modo rechazar".

- Debido a que $Z = 13.995 > 1.96$, se rechaza la hipótesis nula.

- Esto indica que, estadísticamente se comprueba la presencia de la relación entre los gastos en publicidad y el aumento en el número de pasajeros que viajan por Hop Scotch Airlines.

Estimación y prueba de hipótesis : β

- Interesa también saber si la relación encontrada es positiva o negativa, y además determinar la magnitud de la relación. Para esto se plantea la hipótesis sobre el coeficiente de la pendiente para poder comprobar lo anterior.

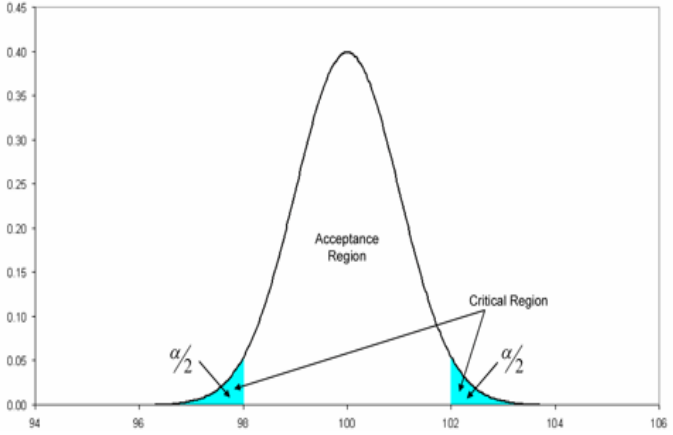


- La prueba de hipótesis se plantea de la siguiente forma:

$$H_0: B_1 = 0$$

$$H_A: B_1 \neq 0$$

- Dado que no conocemos el valor poblacional, debemos estimar $\beta 1$ con los datos de una muestra. El estimador vendría a ser b1.



- Entonces, ahora tendríamos que hacer la prueba de hipótesis para demostrar la relación específica del estimador.

Estimación y prueba de hipótesis : β

- Se calcula el estadístico de prueba para la hipótesis sobre la pendiente:

Prueba *z* para el coef. de regre. poblacional $$Z = \frac{b_1 - \beta_1}{S_{b_1}}$$

- En donde tenemos que S_{b1} es el error estándar del coeficiente de regresión

Error estándar del coef. de regresión $$S_{b_1} = \frac{Se}{\sqrt{SCx}}$$

Estimación y prueba de hipótesis : β

- Para el caso de Hop Scotch Airlines:

$$S_{b_1} = \frac{0.907}{\sqrt{137.733}} = 0.07726 \qquad Z = \frac{1.0813-0}{0.07726} = 13.995$$

- Si se selecciona un valor de α=0.05, entonces para la comparación de prueba de hipótesis $Z_{0.05} = \pm 1{,}96$. La regla de decisión sería:

Regla de decisión: "No rechazar si *t* está entre $\pm$ 1.96, de lo contrario rechazar".

- De acuerdo, al valor del estadístico t calculado, se rechaza la hipótesis nula.

Estimación y prueba de hipótesis : β

- Debido a que el estadístico calculado es z=13.995, la hipótesis nula de $\beta_1=0$ se rechazar. Al nivel del 5%, estadísticamente existe una relación entre pasajeros y publicidad.

- Si la hipótesis nula no hubiera sido rechazada, se concluiría que la publicidad y los pasajeros no están relacionados. Descartando el modelo , se utilizaría una variable explicativa diferente.

- Ahora se debe verificar cuál es la magnitud de la relación.

Estimación y prueba de hipótesis : β

- Debido a que se ha rechazado la hipótesis nula de que B_1=0, la pregunta es, "¿Cuál es la magnitud de la relación?". Esta pregunta puede responderse calculando un intervalo de confianza (IC) para B_1,

Intervalos de confianza : $$\beta_1 = b_1 \pm t(S_{b_1})$$

- Si se utiliza un nivel de confianza del 95%,

IC para B_1=1.08 $\pm$ (1,96)(0.07726)

0.9285 $\leq$ B_1 $\leq$ 1.2314

Esto significa que se puede estar al 95% seguro que los intervalos de 9285 y 12315 contienen el valor poblacional de β. En términos prácticos, estadísticamente se comprueba la relación positiva entre los gastos de publicidad y el aumento en el número de pasajeros que viajan por Hop Scotch Airlines

Estimación y prueba de hipótesis

- A un nivel de confianza del 95%, se concluye que el coeficiente de correlación poblacional es estadísticamente diferente de 0, y que por lo tanto, hay una relación entre los gasto de publicidad y el número de pasajeros.

- Al igual que con la prueba para β_1, la hipótesis nula no se rechaza y se concluye que hay una relación positiva entre los gasto de publicidad y el número de pasajeros.

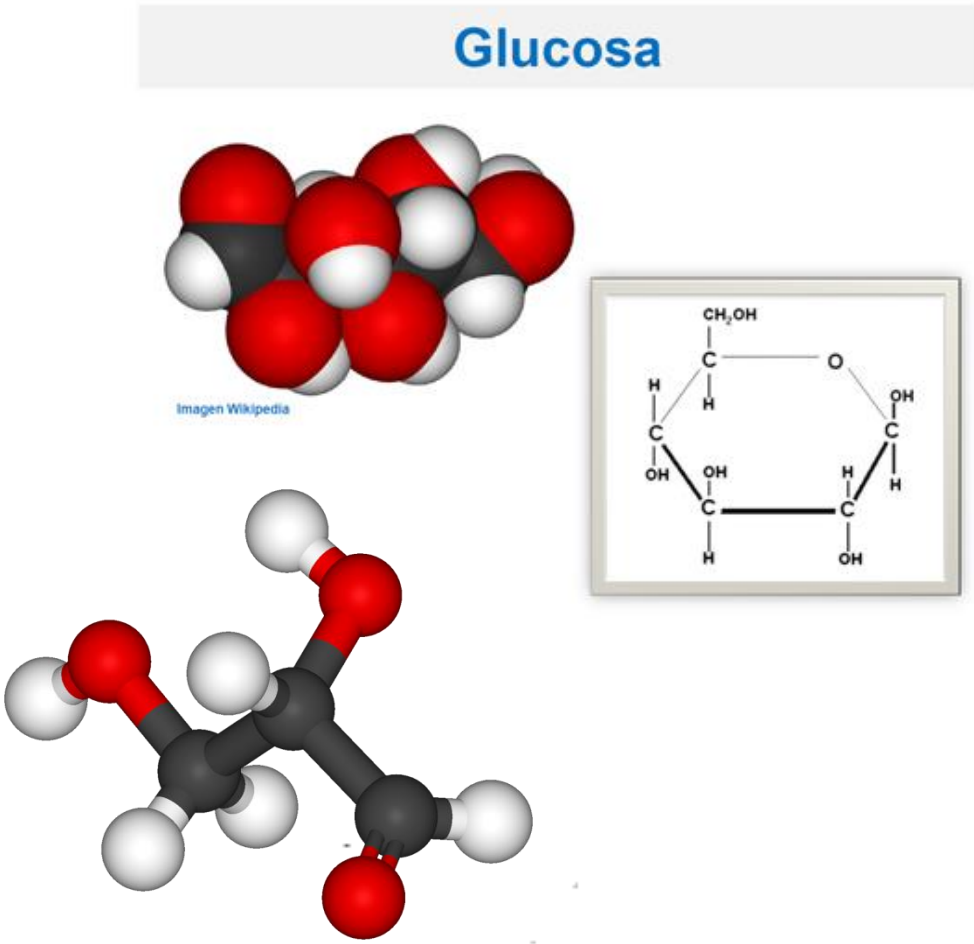
- El hecho de tener el mismo estadística calculado (13.995) sea para los estimadores de β_1 y ρ no es coincidencia. Siempre se obtendrá los mismos resultados en estas 2 pruebas de hipótesis en un modelo de regresión simple.

- Sin embargo se debe realizar ambas pruebas ya que esta igualdad no se mantiene en un modelo de regresión múltiple como se verá en el próximo capítulo.

Ejercicio de práctica

Los siguientes datos corresponden a los pesos (kg) y niveles de glucosa en la sangre (mg/100 ml) de 16 varones adultos aparentemente sanos:

Peso (X)	Glucosa (Y)
64.0	108
75.3	109
73.0	104
82.1	102
76.2	105
95.7	121
59.4	79
93.4	107
82.1	101
78.9	85
76.7	99
82.1	100
83.9	108
73.0	104
64.4	102
77.6	87

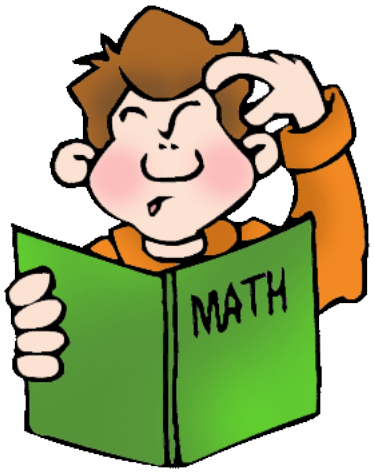


Ejercicio de práctica

Para el ejercicio anterior determine:

1. La correlación entre el peso y la glucosa.
2. Interprete el resultado de la correlación
3. La ecuación de regresión lineal simple.
4. Interprete el valor del coeficiente B1 de peso
5. Realiza la prueba de hipótesis para el coeficiente de peso, bajo el supuesto de que el coeficiente debería ser diferente de 0.
6. Realice la prueba para el coeficiente Rho, bajo el supuesto de que el coeficiente debería ser diferente de 0
7. Determine el intervalo de confianza del 95% para el coeficiente de peso.

   Para todas las pruebas se parte de $\alpha=0.05$



Conclusión

- El presente capítulo abordó los principios de la regresión bivariada y el análisis de correlación.

- Dentro de los temas vistos en la metodología de un modelo de regresión bivariada se expusieron:

-La relación lógica entre las variables

-La estimación de la recta de mejor ajuste

-Prueba de hipótesis y la estimación de los estimadores de regresión.

- El utilizar la regresión o la correlación debe de pasar ante todo sobre un marco conceptual de lo que se desea analizar: una relación simétrica o asimétrica de las variables.


arte