UNAP
FINESI

IV Semestre

Modelos Discretos

Introducción a los Modelos de Regresión

Mtr. Alcides Ramos Calcina

Introducción a los Modelos de Regresión

Mtr. Alcides Ramos Calcina

Introducción

- Los **modelos de regresión** representa una técnica fundamental para la “construcción de modelos” que sean coherentes desde un punto de vista lógico.

- Se establece la **relación** de dependencia entre una variable (dependiente) y otra u otras (independientes o explicativas).

$$y_i = f\left(x_{1i}, x_{2i}, \cdots, x_{ki}\right)$$

DEFINICIÓN TÉCNICAS

Un modelo de regresión es un modelo matemático que busca determinar la relación entre una variable dependiente (Y), con respecto a otras variables, llamadas explicativas o independientes (X).

Introducción

OBJETIVOS

- El objetivo principal de construir un modelo de regresión puede ser:

Modelo con fines explicativos

- Evaluar cómo afecta el cambio en unas características determinadas (variables independientes) sobre otra característica en concreto (variable dependiente)

Modelo con fines predictivos

- intentar estimar o aproximar el valor de una característica (variable dependiente) en función de los valores que pueden tomar en conjunto otra serie de características (variables independientes).

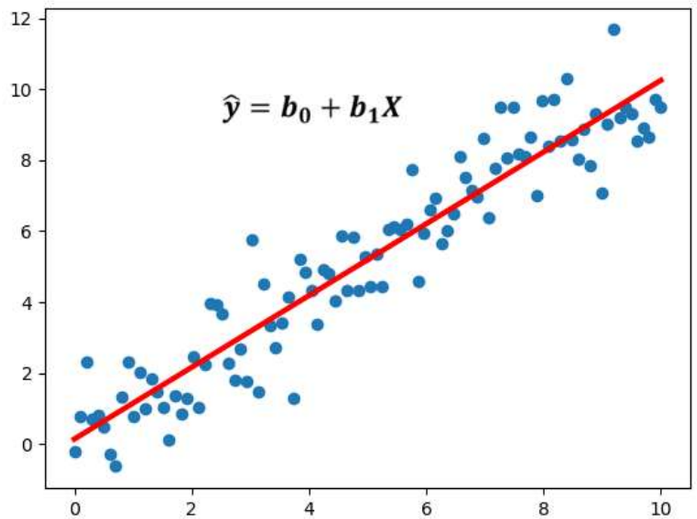
Clasificación

- Por el tipo de relación entre las variables

01

Regresión lineal

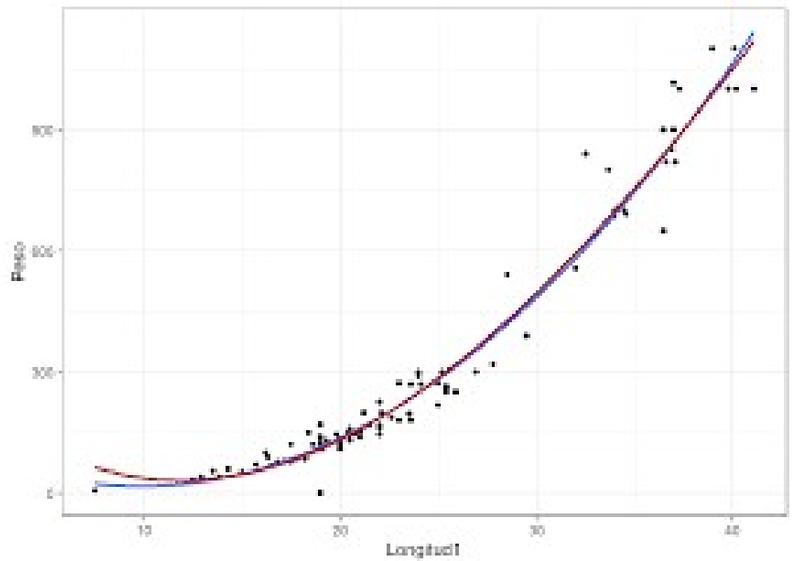
Función lineal formará una línea recta al ser grafica en un plano cartesiano



02

Regresión no lineal

Los valores de esta no formaran una línea recta al ser grafica

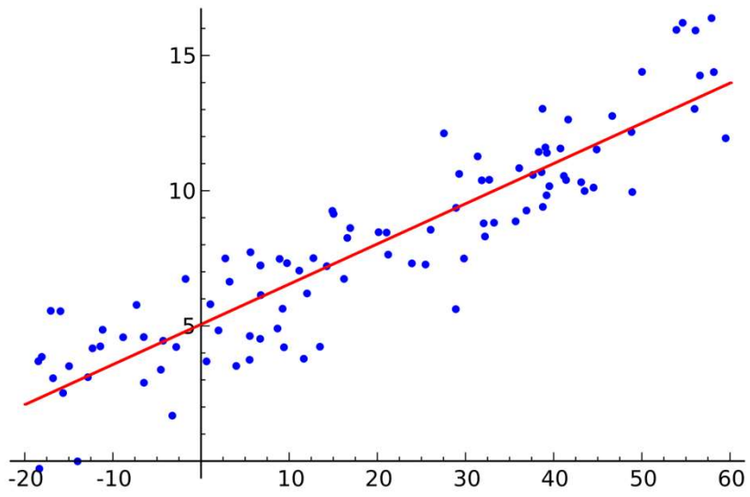


Clasificación

- Los modelos de Regresión Lineal de acuerdo a sus parámetros puede ser:

Lineal simple

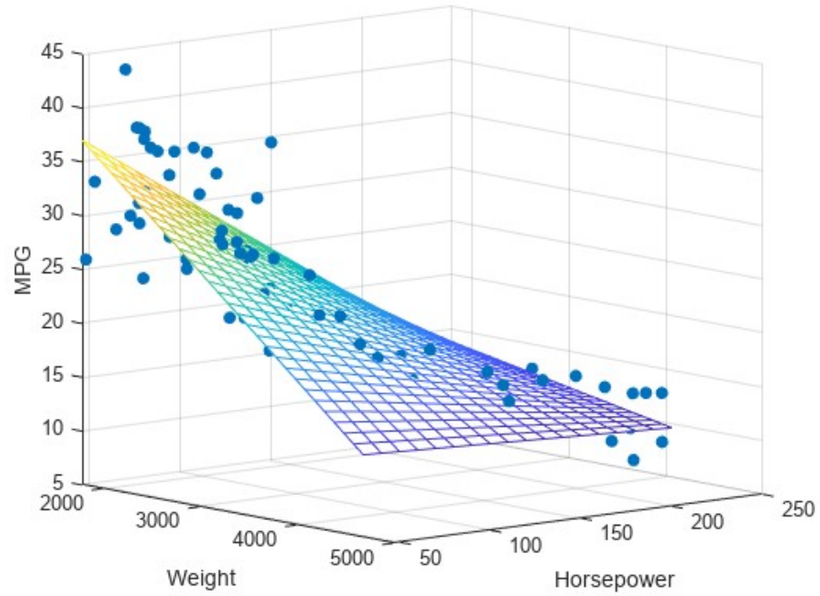
Una variable independiente.



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

Lineal múltiple

Dos o más variables indpendientes.



$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$

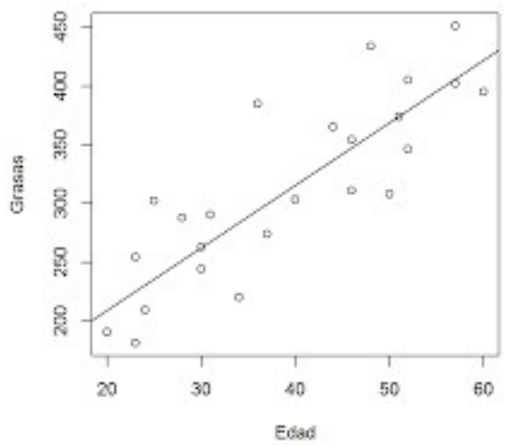
Clasificación

- Por la naturaleza de la variable dependiente (y):

CUANTITATIVA

Regresión lineal

Variable dependiente es una variable continua.

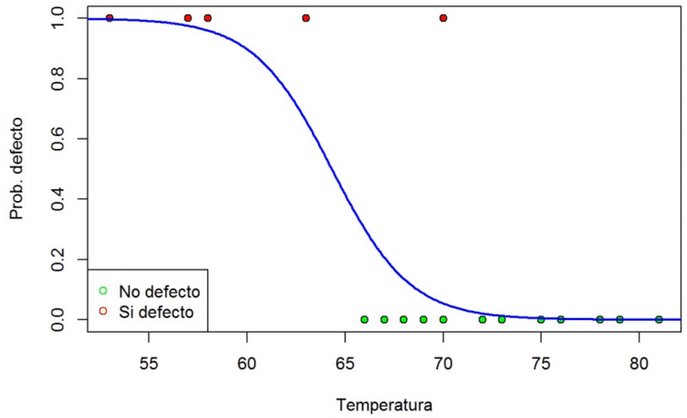


$$y_i = \beta_0 + \beta_1 x_i + e_i$$

CUALITATIVA

Regresión logística

Variable dependiente es dicotómica (categórica).



$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

Una breve Introducción a la Regresión Lineal

Mtr. Alcides Ramos Calcina

Regresión Lineal

El análisis de regresión esta relacionado, con el estudio de la dependencia estadística de una variable (variable dependiente) en función de una o más variables adicionales (variables independientes).

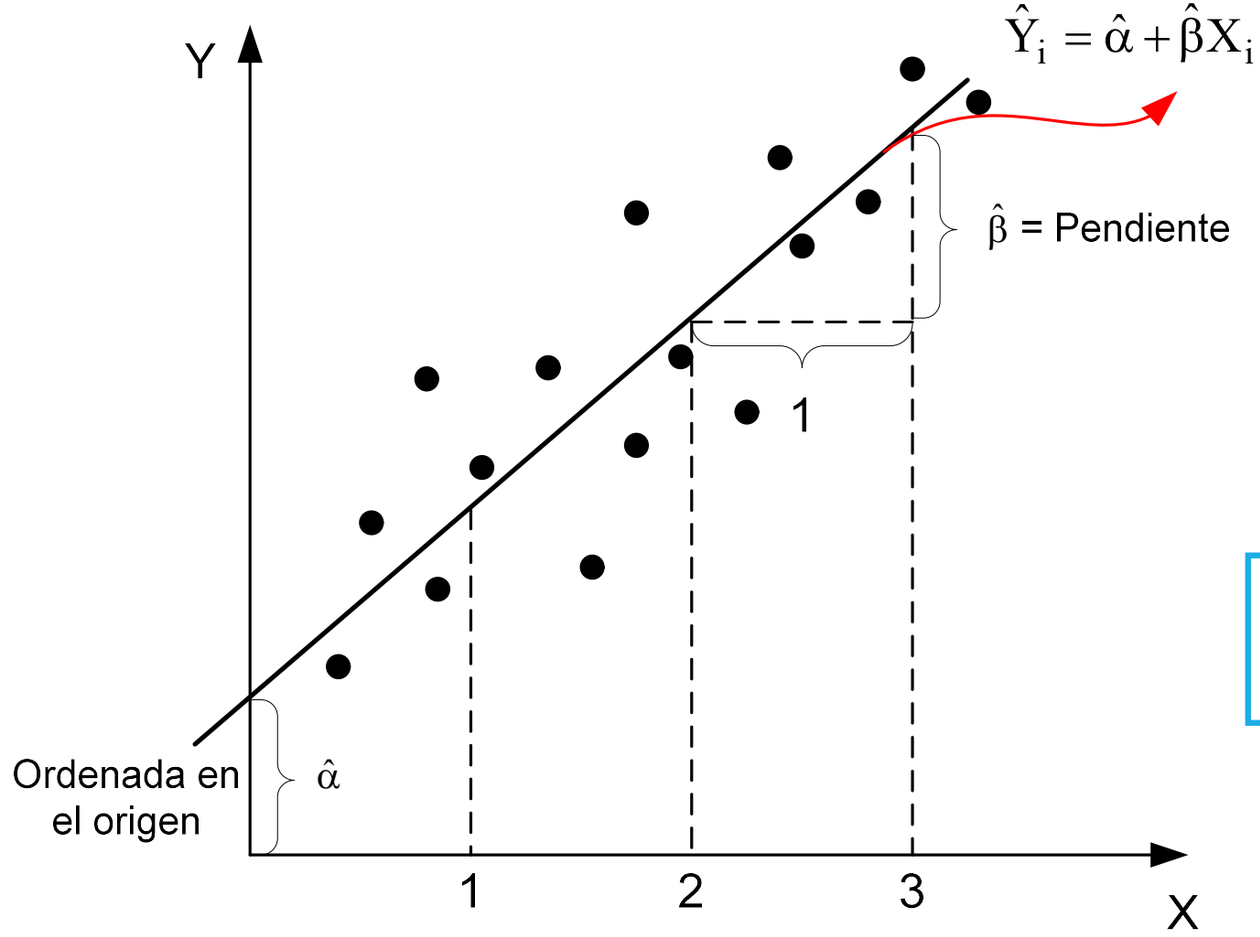
SIGNIFICADO DE LA REGRESIÓN

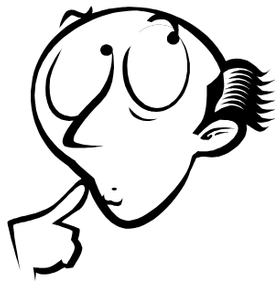
De manera básica la regresión tiene dos significados:

- Una surge de la distribución conjunta de probabilidad de dos variables aleatorias.
- Y el otro es empírico y nace de la necesidad de ajustar alguna función a un conjunto de datos.

Modelo de Regresión Lineal Simple

Modelo: $y = f(x) \Rightarrow$ $y = \alpha + \beta x + \varepsilon$





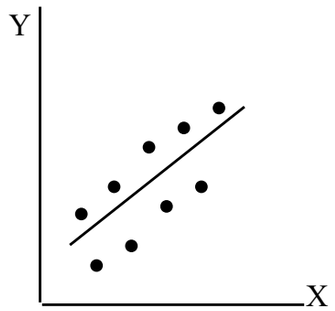
Finalidad

Estimar los valores de *y* (variable dependiente) a partir de los valores de *x* (variable independiente)

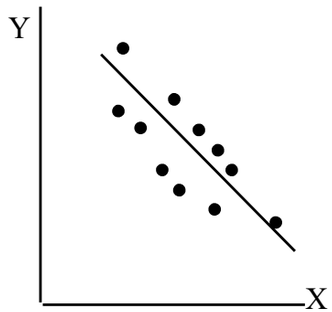
Ajuste de curvas

Una línea recta puede ser obtenida por el método de los mínimos cuadrados. Graficamos la pareja ordenada (X,Y) en un plano cartesiano, podemos obtener diversos diagramas de dispersión.

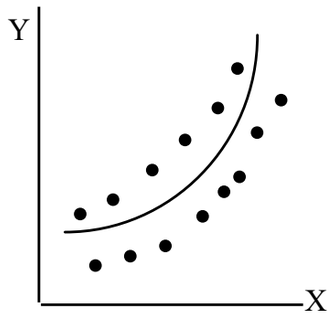
(a) Lineal directa

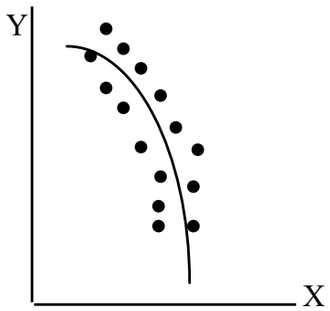


(b) Lineal inversa

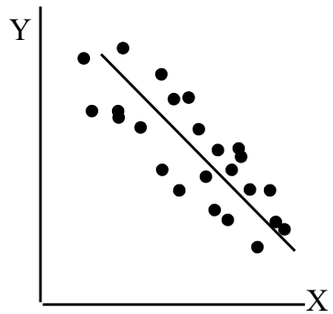


(c) Curvilínea directa

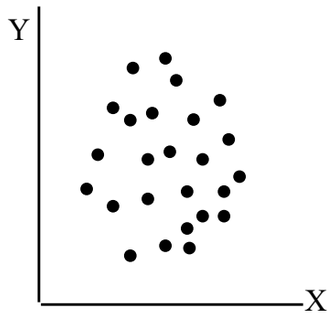




(d) Curvilinea inversa



(e) Lineal inversa con más dispersión



(d) Ninguna relación

Ajuste de curvas

Ahora el problema consiste en estimar los parámetros α y β por el método de Mínimos Cuadrados.

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$\hat{\beta} = \frac{S_{xy}}{S_x^2} = \frac{Cov(x,y)}{S_x^2}$$

- Para el caso de muestras pequeñas, el denominador de la varianza es “n-1”.
- Muestras grades, se divide entre “n”

La varianza residual es:

$$S_R^2 = \frac{nS_y^2 - \hat{\beta} nCov(x,y)}{n-2}$$

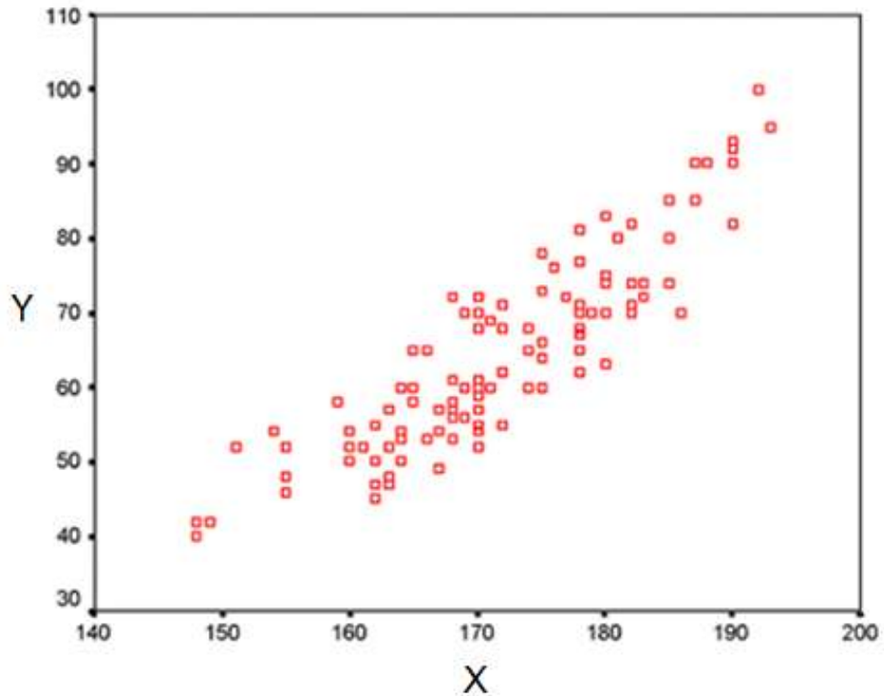
con,

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} \qquad \bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}$$

$$S_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} \qquad S_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}$$

Diagrama de dispersión

- Los diagramas de dispersión no sólo muestran la relación existente entre variables, sino también resaltan las observaciones individuales que se desvían de la relación general. Estas observaciones son conocidas como outliers o valores atípicos, que son puntos de los datos que aparecen separados del resto.



Ejemplo 1

El departamento de personal de una empresa informática dedicada a la introducción de datos ha llevado a cabo un programa de formación inicial del personal. La tabla siguiente indica el progreso en pulsaciones por minuto (p.p.m.) obtenido en mecanografía de ocho estudiantes que siguieron el programa y el número de semanas que hace que lo siguen:

Número de semanas	Ganancia de velocidad (p.p.m)
3	87
5	119
2	47
8	195
6	162
9	234
3	72
4	110

a) Represente el diagrama de dispersión. ¿Cree Ud. que es razonable suponer que existe una relación lineal entre el número de semanas y la ganancia de velocidad?

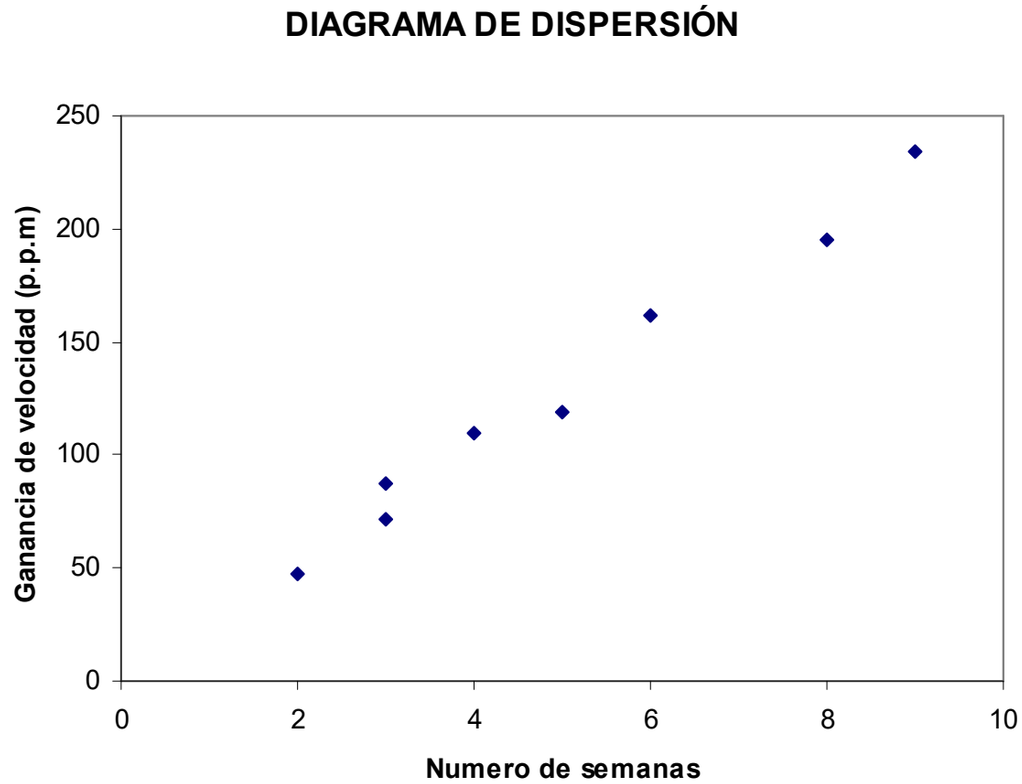
b) Encuentre la recta de regresión. Interprete los parámetros obtenidos.

c) ¿Qué ganancia de velocidad podemos esperar de un estudiante que hace siete semanas que va a clase?

d) Calcule la varianza residual

Ejemplo 1

a) Diagrama de dispersión



El diagrama de dispersión nos muestra que la relación entre las dos variables es lineal con pendiente positiva, de manera que cuantas más semanas pasan, mayor es la ganancia de velocidad. Por tanto, tiene sentido buscar la recta de regresión.

Ejemplo 1

b) A partir de la tabla siguiente, realizamos los cálculos de los parámetros de la recta.

i	Xi	Yi	$X_i - \overline{X}$	$Y_i - \overline{Y}$	$(X_i - \overline{X})^2$	$(Y_i - \overline{Y})^2$	$(X_i - \overline{X})(Y_i - \overline{Y})$
1	3	87	-2	-41.25	4	1701.56	82.50
2	5	119	0	-9.25	0	85.56	0
3	2	47	-3	-81.25	9	6601.56	243.75
4	8	195	3	66.75	9	4455.56	200.25
5	6	162	1	33.75	1	1139.06	33.75
6	9	234	4	105.75	16	11183.06	423.00
7	3	72	-2	-56.25	4	3164.06	112.50
8	4	110	-1	-18.25	1	333.06	18.25
Σ	**40**	**1026**			**44**	**28663.50**	**1114.00**

Medias muestrales:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{40}{8} = 5$$

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{1026}{8} = 128.25$$

Varianzas muestrales:

$$S_x^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} = \frac{44}{7} = 6.286$$

$$S_y^2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n-1} = \frac{28663.50}{7} = 4094.786$$

Ejemplo 1

Covarianza muestral:

$$Cov(x,y) = S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1} = \frac{1114}{7} = 159.143$$

Ya podemos calcular los coeficientes de la recta de regresión:

$$\hat{\beta} = \frac{S_{xy}}{S_x^2} = \frac{Cov(x,y)}{S_x^2} = \frac{159.143}{6.286} = 25.318 \qquad \hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} = 128.250 - 25.318(5) = 1.659$$

La recta de regresión obtenida es:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i = 1.659 + 25.318x$$

- α: La ordenada en el origen no tiene ninguna interpretación con sentido, ya que correspondería a la ganancia de velocidad por cero semanas de clases.
- β: La pendiente nos indica, por cada semana de clase se tiene una ganancia de velocidad de aproximadamente 25.3 p.p.m

Ejemplo 1

c) Para una persona que hace siete semanas que va a clase, podemos calcular la ganancia de velocidad a partir de la recta de regresión, considerando x = 7:

$$\hat{Y}_i = 1.659 + 25.318(7) = 178.885$$

Es decir, aproximadamente una ganancia de 179 pulsaciones por minuto.

d) La varianza residual es:

$$S_R^2 = \frac{nS_y^2 - \hat{\beta} nCov(x, y)}{n - 2} = \frac{8(4094.786) - 25.318(8)(159.143)}{8 - 2}$$

$$S_R^2 = 87.471$$

Entonces, la desviación estándar residual $S_R = 9.353$

Prueba de hipótesis para β

Una gran parte del interés se centra en los procedimientos de inferencia respecto a β. La razón de esto es el hecho de que β dice mucho acerca de la relación entre X e Y.

Cuando X e Y están linealmente relacionados:

$$\beta = \begin{cases} (+) \text{ positivo: Y aumenta a medida que X aumenta } \rightarrow \text{ RL Directa} \\ 0\text{: No existe relación lineal entre X e Y} \\ (-) \text{ Negativo: Y disminuye a medida que X aumenta } \rightarrow \text{ RL Inversa} \end{cases}$$

Observemos que si en el modelo de regresión lineal la pendiente es cero, entonces la variable X no tiene ningún efecto sobre la variable Y. En este caso diremos que X no es una variable explicativa del modelo.

Prueba de hipótesis para β

Como en todos los contrastes de hipótesis, daremos los pasos siguientes:

i) Planteamiento de hipótesis

$$H_0: \beta = \beta_0 \quad \text{vs} \quad H_a: \begin{cases} \beta > \beta_0 \\ \beta < \beta_0 \\ \beta \neq \beta_0 \end{cases}$$

ii) Nivel de significancia

$$\alpha = 0.05 = 5\%$$

iii) Estadístico de prueba: t-Student

$$t_c = \frac{\hat{\beta} - \beta_0}{S}\sqrt{nS_x^2}$$

con $t_{\frac{\alpha}{2},(n-2)}$ Grados de Libertad

Donde:

$$S = \sqrt{\frac{\left[\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right] - \hat{\beta}\left[\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right]}{n-2}}$$

iv) Decisión

Si $t_c > t_t$, se rechaza H_0.

También se puede utilizar el valor de probabilidad asociado al estadístico t:

Si $p \leq \alpha$, se rechaza H_0.

v) Conclusión

Ejemplo 2

Continuando con el ejemplo 1 de la empresa informática dedicada a la introducción de datos, queremos contrastar la hipótesis nula de que la variable X no es explicativa de la variable Y, es decir, que la pendiente de la recta de regresión es cero.

Solución:

- Planteamiento de hipótesis

$$H_0 : \beta = \beta_0$$

$$H_1 : \beta \neq \beta_0$$

- Nivel de significancia:

$$\alpha = 0.05 \ (5\%)$$

- Estadístico de prueba: (t-Student)

$$t_c = \frac{\hat{\beta} - \beta_0}{S}\sqrt{nS_x^2} = \frac{25.318 - 0}{9.353}\sqrt{8(6.286)} = 19.196$$

El valor tabulado de t (tabla distribución t-Student) es:

$$t_t = t_{(n-2);\alpha/2} = t_{6;0.025} = 2.447$$

- Decisión

Como $t_c = 19.197 > t_t = 2.447$, e rechaza H_0.

- Conclusión

Podemos concluir que la variable ganancia de velocidad es explicativa por el número de semanas con un 95% de confianza

Intervalo de confianza para β

IC: $P\left[\hat{\beta} \pm t_{\frac{\alpha}{2}} \sqrt{S_{\beta}^2}\right] = 1 - \alpha$ o IC:

Donde:

$\mathrm{EE}(\hat{\beta}) = \sqrt{\mathrm{S}_{\beta}^2}$, es el error estándar del estimador.

Modelo de Regresión Lineal General

- La formulación matemática de estos modelos es la siguiente:

$$Y = f(X_1, X_2, X_3, ..., X_k) + e_i$$

- Donde e_i es el error de observación debido a variables no controladas.

- En el **modelo de Regresión Lineal General es:**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + e_i$$

Un primer objetivo es estimar los parámetros del mismo $\beta_0, \beta_1, \beta_2 ..... \beta_k$, y la función de distribución del error a partir de una muestra de *n* observaciones.

Modelo de Regresión Lineal General

- De la expresión matemática del modelo de regresión lineal general se deduce que para $i = 1,2,...,n$ se verifica la siguiente igualdad**:**

$$Y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + ... + \beta_k X_{ik} + e_i$$

Donde e_i es el error aleatorio o perturbación de la observación i-ésima.

Interpretación y estadísticos de la regresión

- El coeficiente β_1 mide el cambio experimentado por Y asociado con una variación de X_{i1} en una unidad. De forma similar β_2 mide la variación experimentada por Y como consecuencia de un incremento de X_{i2} en una unidad.
- En ambos casos, el supuesto de que las demás variables explicativas permanecen constantes es crucial para esta interpretación de los coeficientes.
- En base a estos supuestos, los parámetros desconocidos $\beta_{1\ y}$ β_2 se denominan ***Coeficientes de regresión parciales***, puesto que corresponden a los valores de las derivadas parciales de Y con respecto a X_{i1} y X_{i2}, respectivamente.

Modelo Lineal Múltiple Matricial

Para plantear el modelo lineal múltiple, por la vía matricial, será:

Modelo ..... $\underline{Y} = \underline{x}\,\underline{\beta} + \varepsilon_{\underline{i}}$

Ecuación de Regresión..... $\underline{Y} = \underline{x}\,\underline{\beta}$

Ecuación de Estimación.... $\hat{Y} = \underline{x}\,\underline{b}$

Quedando en la ecuación de estimación, x como un vector fila

Es decir:

$$x = \begin{vmatrix} 1 & x_1 & x_2 \end{vmatrix} \quad \text{y} \quad b = \begin{vmatrix} b_o \\ b_1 \\ b_2 \end{vmatrix}$$

que se ha obtenido a partir de las Ecuaciones Normales, representada por la vía Matricial:

$$b = (x'y)(x'x)^{-1}$$

Modelo Lineal Múltiple Matricial

Las dimensiones de la matriz “x” y “y”:

$$(x'x) = \begin{vmatrix} n & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i}x_{2i} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{2i}^2 \end{vmatrix}$$

$$(x'y) = \begin{vmatrix} \sum Y_i \\ \sum x_{1i}Y_i \\ \sum x_{2i}Y_i \end{vmatrix}$$

(x’x) ....Siempre será una matriz cuadrada y simétrica.

Ya aquí el invertir la matriz se hace un poco más complejo, ya que tenemos una matriz 3 x 3, y tendríamos que aplicar algún método de invertir matriz.

Ejemplo 3

Supóngase que el precio de un producto se desea expresar en función de la cantidad de unidades del mismo por cada núcleo familiar y el precio de su materia prima fundamental. Si se tienen 10 núcleos familiares con la siguiente información:

Y = precio de un producto en pesos;
X_1 = cantidad de unidades del producto;
X_2 = precio de la materia prima fundamental en soles.

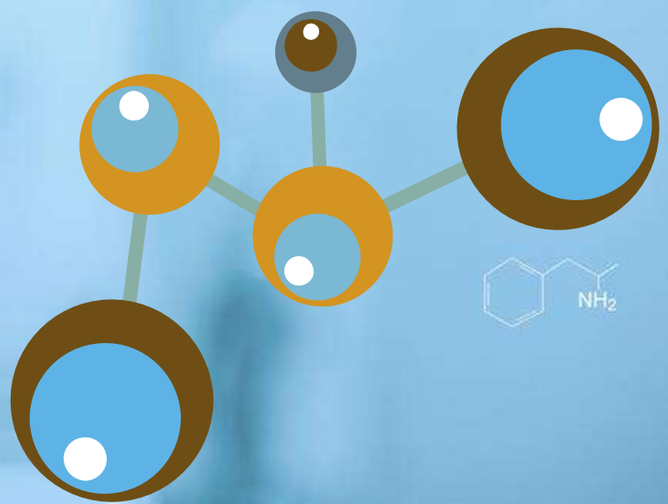
Familia	Y	X_1	X_2
1	1.6	4	0.4
2	2	2	0.6
3	1.5	4	0.4
4	1.7	2	0.4
5	2.2	1	0.7
6	2.5	1	0.8
7	3	1	1
8	1.9	2	0.5
9	1.8	2	0.4
10	2.8	1	0.8

a) Grafique los diagramas de dispersión
b) Determine la ecuación de estimación
c) Interprete los coeficientes b_1 y b_2
d) Determine la precisión de la ecuación estimada, y diga si la misma puede ser utilizada para predecir el precio del producto.

FINESI

Modelos Discretos

IV Semestre



GRACIAS

https://aulavirtual2.unap.edu.pe/