Análisis Multivariado y Ajuste del Modelo

Ismael Solís Moreno

Análisis Multivariado

- Los estudios multivariados son similares a los univariados, a diferencia que tienen más de dos variables dependiente e independiente.

- Otra diferencia importante es que en un análisis de múltiples variables no hablamos de “correlación simple” ni de estadísticos descriptivos por sí solos, sino que apelamos a otras herramientas estadísticas llamadas “**multivariantes**”, tal es el caso de por ejemplo: *Análisis de varianza (ANOVA)*, *Estudio multifactorial* o *Regresiones Múltiples*.

Análisis Multivariado

Los investigadores emplean **estudios multivariantes** cuando requieren examinar la relación entre múltiples factores al mismo tiempo. Se diferencia claramente de los estudios univariados y bivariados en que **plantean más de una variable dependiente y varias independientes**.

Análisis Multivariado

Por ejemplo, si deseamos examinar la capacidad de tres nuevos productos químicos para limpiar un derrame de aceite, las **tres sustancias químicas** serían las **variables independientes**. En un análisis multivariante se podrían medir las propiedades de las sustancias químicas dispersantes, la desintoxicación del aceite, la toxicidad de la sustancia química y el efecto sobre el medio ambiente como **variables dependientes**.

Análisis Multivariado

El objetivo del análisis multivariado es variable en relación a lo que queremos conseguir con él. Estos son los diferentes escenarios que explican el objetivo del análisis multivariado:

- Optimizar los datos o simplificar la estructura: Esto ayuda a simplificar los datos en la mayor medida posible sin sacrificar información valiosa y sirve para facilitar la explicación de datos.

Análisis Multivariado

- Ordenar y agrupar: Cuando tengamos múltiples variables, se creará un conjunto de objetos o variables "similares" en función de las características medidas para ordenar y agrupar los datos.

- Investigar la relación de dependencia entre variables: La relación entre variables es algo que puede resultar preocupante para muchos. El análisis multivariado nos servirá para saber si todas las variables son independientes o dependientes entre sí.

Análisis Multivariado

- Relación predictiva entre variables: Deben determinarse para predecir el valor de una o más variables a partir de observaciones de otras variables.

- Construcción y prueba de hipótesis: Se prueban hipótesis estadísticas específicas expresadas en parámetros poblacionales multivariados. Esto se puede hacer para probar hipótesis o reafirmar hipótesis previas.

Análisis Multivariado - Ventajas

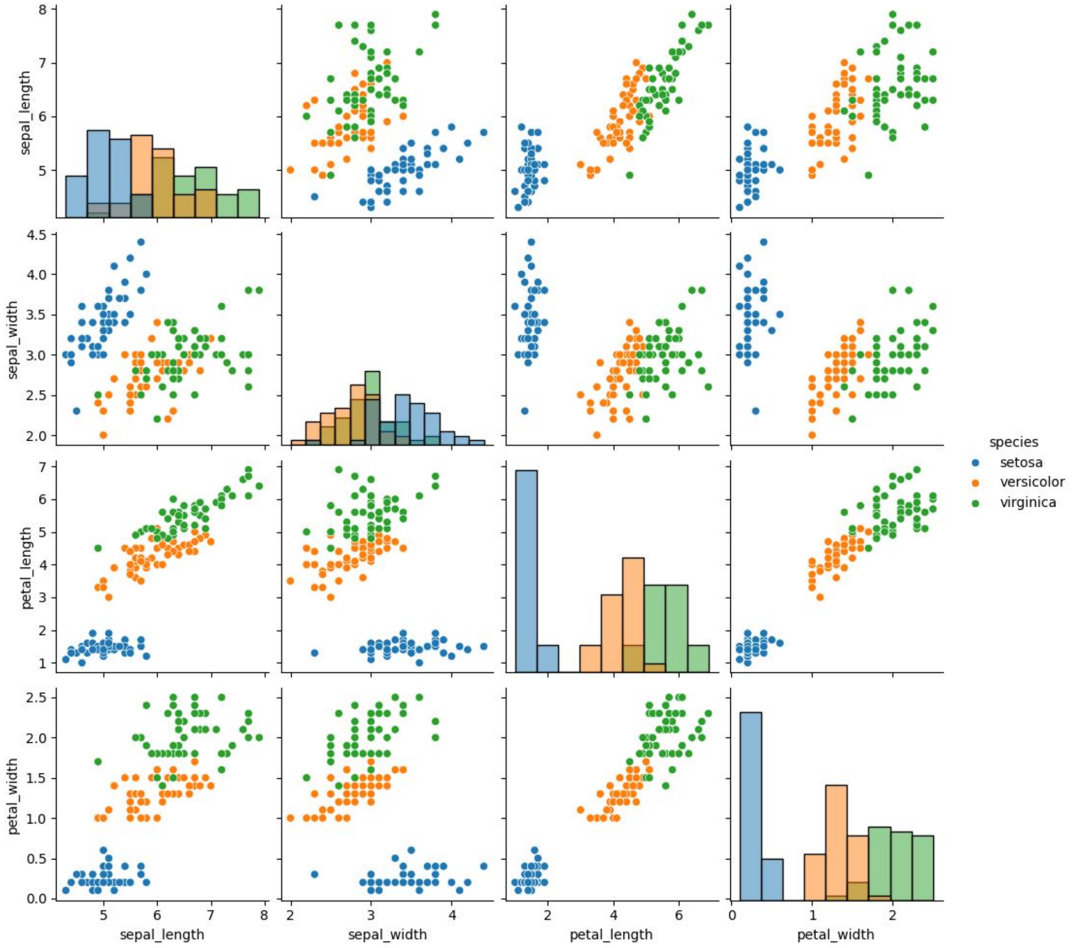
- Permite a los investigadores ver la *relación entre variables y cuantificar la relación entre ellas:* Se puede usar la **tabulación cruzada, correlación parcial y regresión múltiple para controlar la asociación entre variables**.
- Muestra *capacidad de obtener una visión general más realista y precisa* que cuando se analiza una sola variable.

Análisis Multivariado

- Sus técnicas son *complejas*, ***involucran matemáticas avanzadas y requieren procedimientos estadísticos*** para analizar datos.

- Los **resultados del modelado estadístico** ***no siempre son fáciles de entender***.

Pairplot



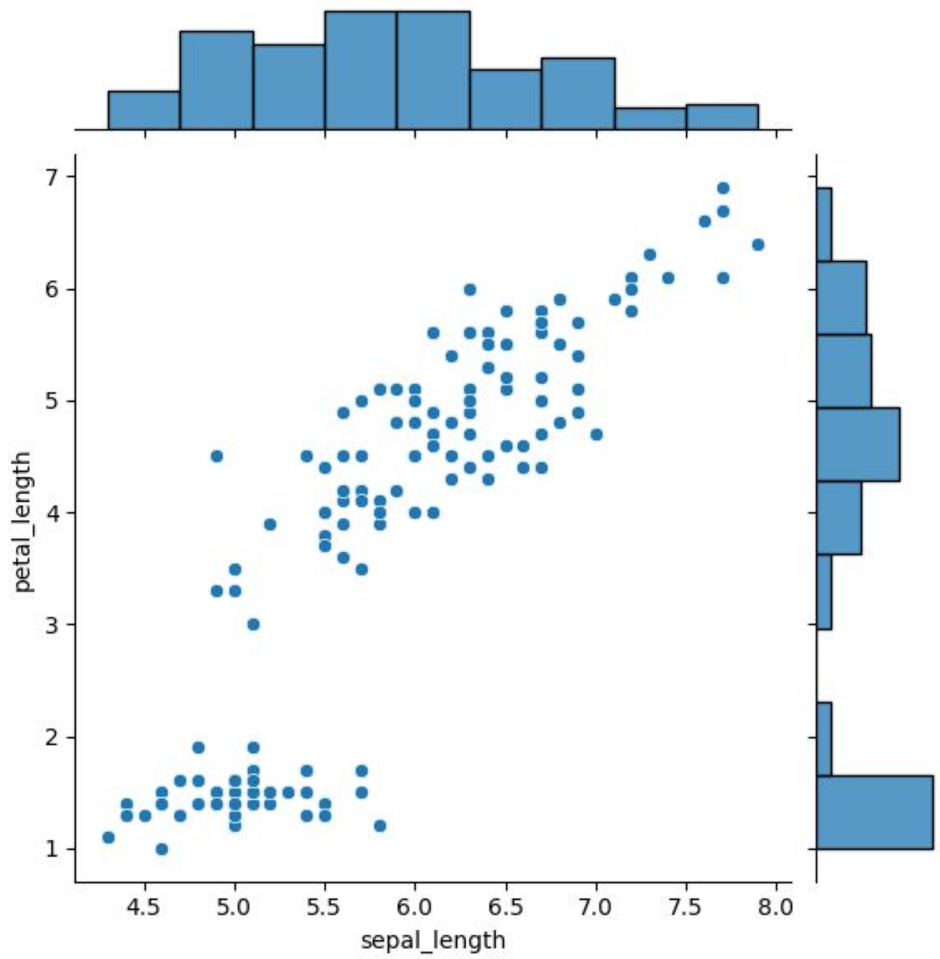

```python
import seaborn as sns
import matplotlib.pyplot as plt

# Cargar el conjunto de datos 'iris' de Seaborn
iris = sns.load_dataset('iris')

# Crear un pairplot
sns.pairplot(iris, hue='species', diag_kind='hist')

# Mostrar el gráfico
plt.show()
```

Joinplot




```python
import seaborn as sns
import matplotlib.pyplot as plt

# Cargar el conjunto de datos 'iris'
iris = sns.load_dataset('iris')

# Crear un jointplot
sns.jointplot(x='sepal_length', y='petal_length', data=iris, kind='scatter')

# Mostrar el gráfico
plt.show()
```

```python
import seaborn as sns
import matplotlib.pyplot as plt
import itertools

# Cargar el conjunto de datos 'iris'
iris = sns.load_dataset('iris')

# Listar las columnas numéricas del dataset (excluyendo 'species')
numeric_columns = iris.select_dtypes(include=['float64']).columns

# Generar todas las combinaciones posibles de dos columnas
combinations = list(itertools.combinations(numeric_columns, 2))

# Crear un jointplot para cada combinación de columnas
for x, y in combinations:
    sns.jointplot(x=x, y=y, data=iris, kind='scatter')
    plt.show()
```

Análisis Multivariado

Ejercicio en Clase

- Utilizando el dataset de mpg de la clase pasada crea lo siguiente:
  - Pairplot de las variables numéricas del dataset
  - Una matriz de calor sintetizada
  - Joinplot de la variable mpg contra las demás variables numéricas en el dataset.

Ajuste del Modelo

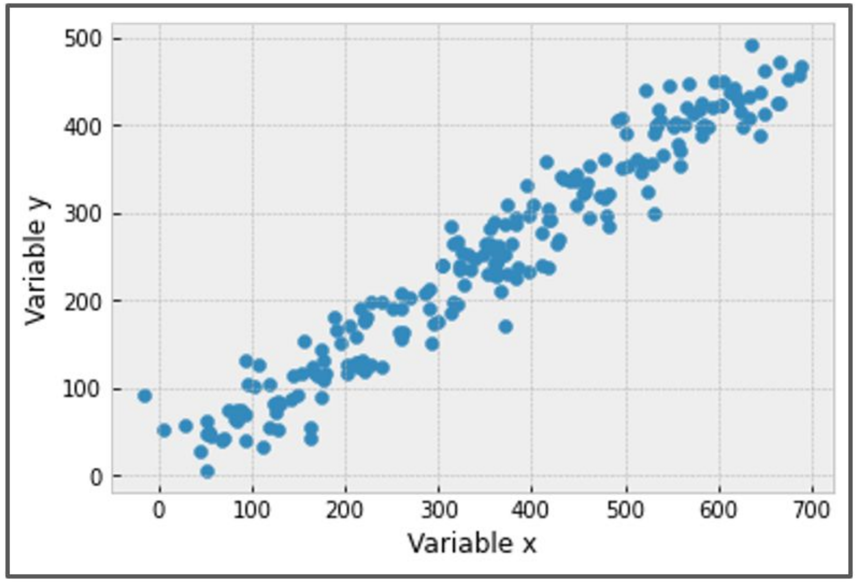
Punto de Partida

🔎 Planteamos la hipótesis de que ***podría*** **existir algún tipo de dependencia** de una variable con respecto a la otra.

🧐 Si este tipo de dependencia existe, **queremos ver de qué forma se da esa relación**.

Ajuste del Modelo

veamos el siguiente gráfico:



Pareciera que **las variables tienen una fuerte correlación positiva**, y si lo pensamos en términos de dependencia, quiere decir que cuando la variable **x** aumenta, entonces también lo hace la variable **y**, y viceversa.

<u>Atención:</u> *cuando planteamos que ante un cambio en la variable x se produce un cambio en la variable y. A esto lo llamaremos* ***dependencia de la variable y*** *hacia la variable* ***x.***

Ajuste del Modelo

Esto podemos verlo como una función matemática estándar

- $y = f(x)$ > donde la variable *y* es una función de *x*, o sea que **en definitiva y depende del cambio de *x***.
- Otra forma de decir lo mismo es que ***x*** es una variable independiente, o sea que **su cambio no depende de nuestro modelo.**

Ajuste del Modelo

Función Lineal

- $y = a + bx$ 👉 donde **a** y **b** son números reales.
- Esta función genera una recta en el plano.
- El valor de *a* (ordenada al origen) **muestra cuál es el valor de *y* cuando *x* vale 0.**
- El valor de *b* (pendiente), por su parte, **indica el grado de inclinación de la recta.**

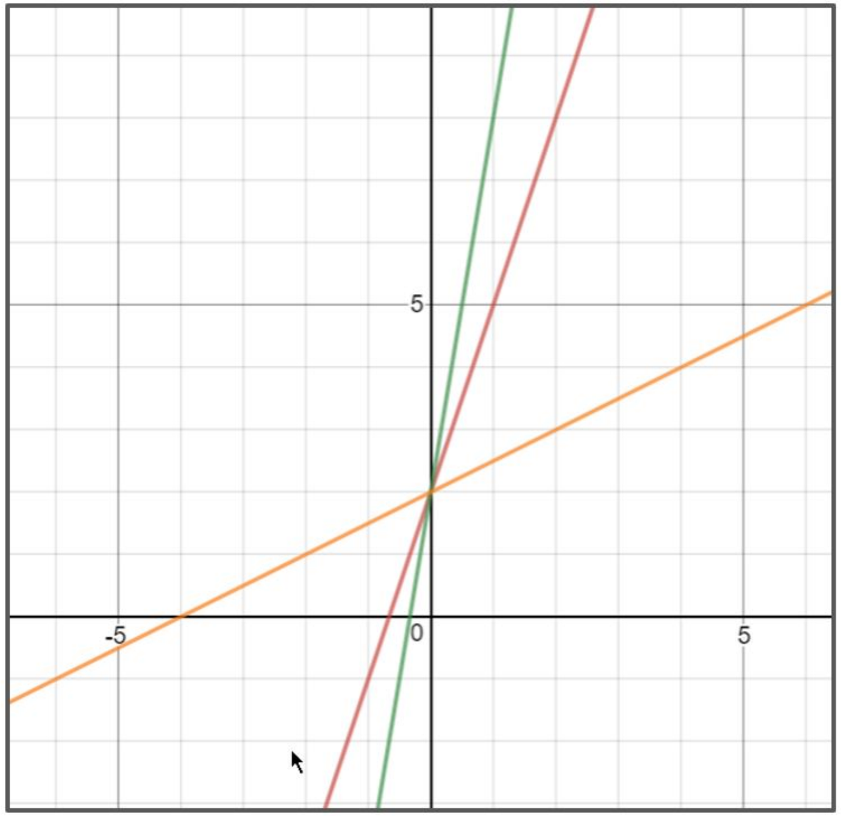
Ajuste del Modelo

Consideraciones

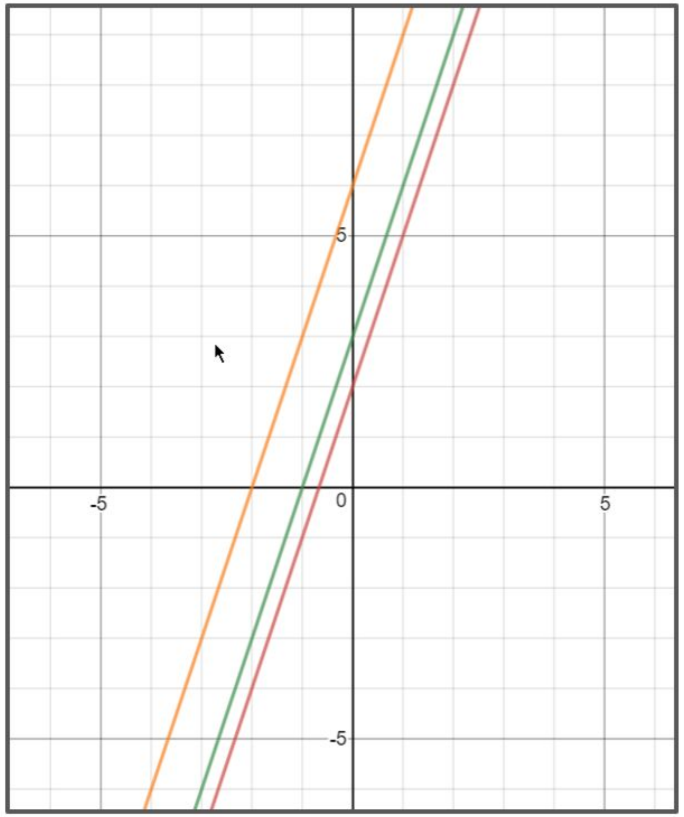
- Una **recta totalmente horizontal** tiene una **pendiente igual a cero.**
- Una **recta inclinada** en el sentido de la correlación positiva tiene una **pendiente positiva**.
- Una **recta inclinada** en el sentido de la correlación negativa tiene una **pendiente negativa**.
- Una **recta vertical** tiene pendiente **infinita**.

Ajuste del Modelo

👉 El **mismo valor de *a* con distintos valores de *b*,** aquí cambia la pendiente o inclinación



👉 Un **valor fijo de *b* para distintos valores de *a***, aquí cambia la posición de la recta pero su inclinación permanece igual.

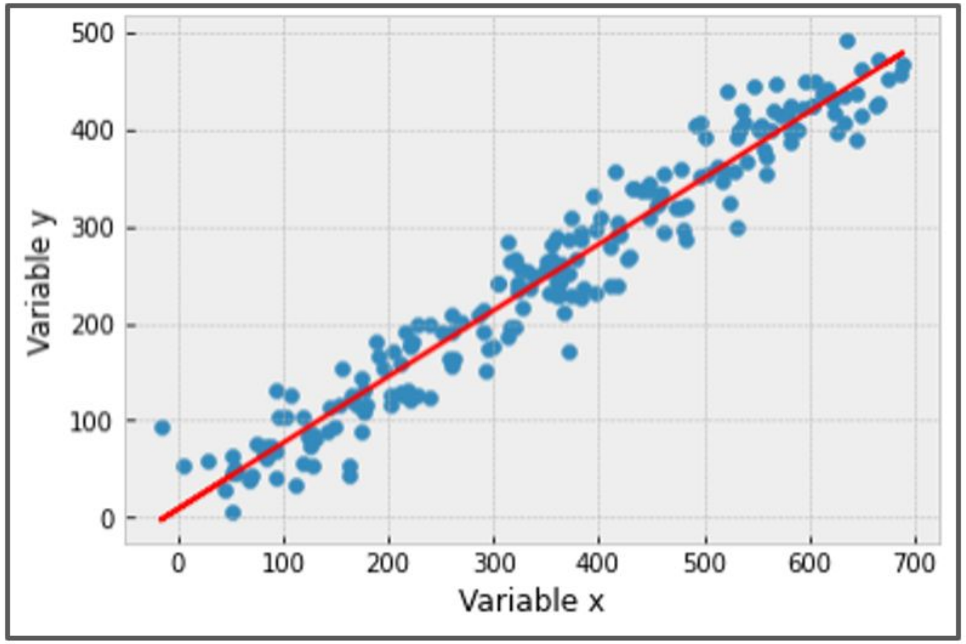


Ajuste del Modelo

Si tenemos un conjunto de puntos en las variables *x* e *y*, y de alguna forma *y* depende de *x*, una forma es trazar una recta que de alguna manera puede representar a esos puntos, tomando un criterio para la representación y trazar una recta que cumpla con él.

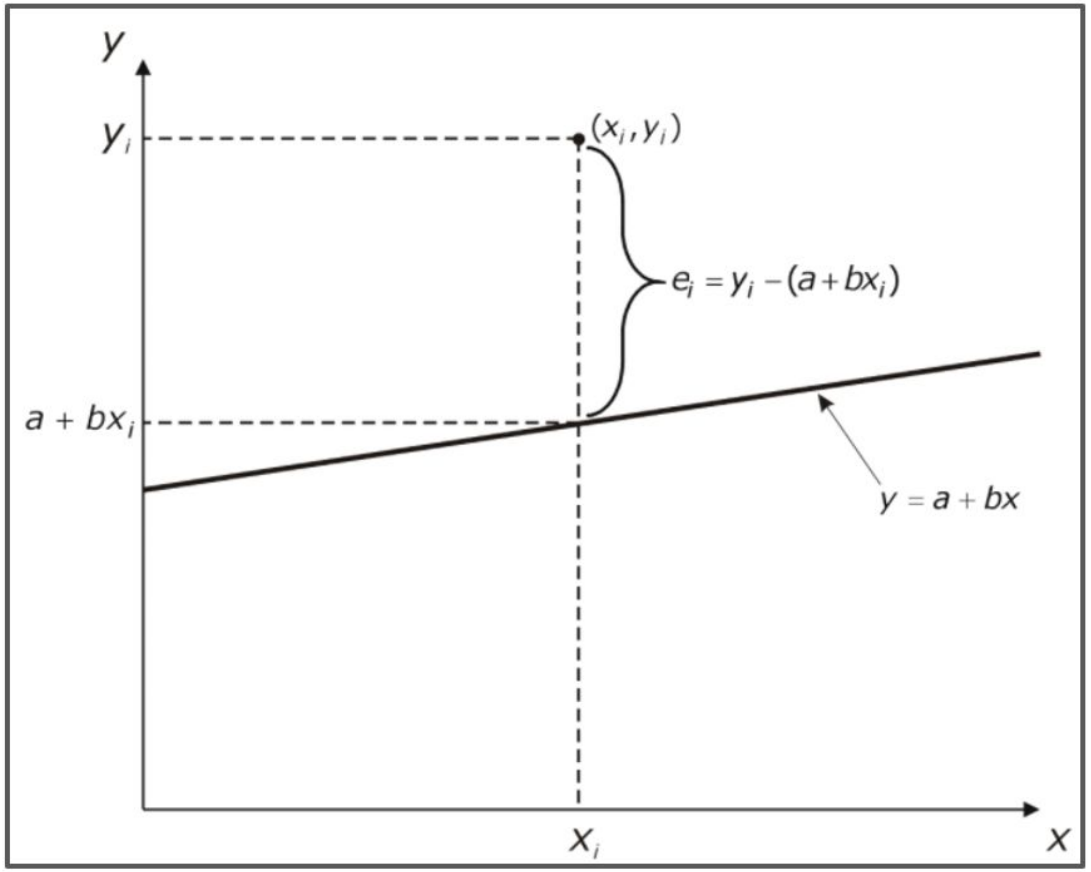
Ajuste del Modelo

Por ejemplo, una recta que pase “lo más al centro posible” del conjunto de puntos...



Aquí realizamos **un ajuste de la recta a los datos**. A la técnica que utilizamos para realizar este ajuste a un conjunto de puntos por parte de una recta la llamaremos **“método de mínimos cuadrados”.**

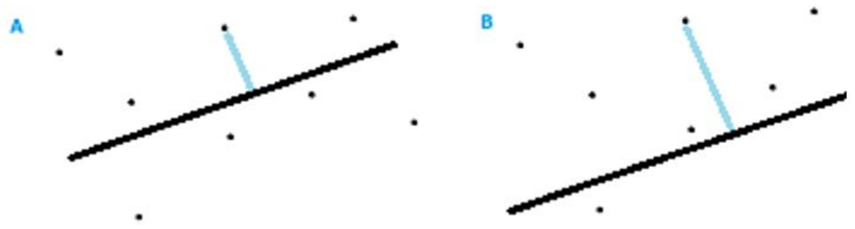
Ajuste del Modelo



- Existe una fórmula (¡que no veremos aquí!) para encontrar precisamente la recta que cumple con la condición de la fórmula de mínimos cuadrados.
- **El método de mínimos cuadrados es el método por defecto que utiliza el modelo de regresión lineal.**

Ajuste del Modelo

- Se toma cada **punto individual y se calcula su distancia vertical a la recta** (denominada error y simbolizada con la letra e).
- Se realiza entonces **la suma de todas las distancias verticales elevadas al cuadrado**. En fórmula $\sum_i [y - (a + bx)]^2$



Ajuste del Modelo

Ejemplo. Encontrar la recta que mejor se ajusta a los siguientes datos:

x	y	x·y	x^2
7	2	14	49
1	9	9	1
10	2	20	100
5	5	25	25
4	7	28	16
3	11	33	9
13	2	26	169
10	5	50	100
2	14	28	4
Σ 55	57	233	473

$$y = \underbrace{m}x + \underbrace{b}$$

$$m = \frac{\Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}}{\Sigma x^2 - \frac{(\Sigma x)^2}{n}} = \frac{233 - \frac{55 \cdot 57}{9}}{473 - \frac{(55)^2}{9}} \simeq -0,84$$

$$b = \bar{y} - m\bar{x} = \frac{\Sigma y}{n} - (-0,84)\frac{\Sigma x}{n} = \frac{57}{9} + 0,84$$

Ajuste del Modelo

Actividad en Clase

Revisar los siguientes videos para entender conceptualmente como funciona el ajuste del modelo lineal usando el método de mínimos cuadrados.

https://www.youtube.com/watch?v=k964_uNn3l0

https://www.youtube.com/watch?v=gUdU6BgnJ2c