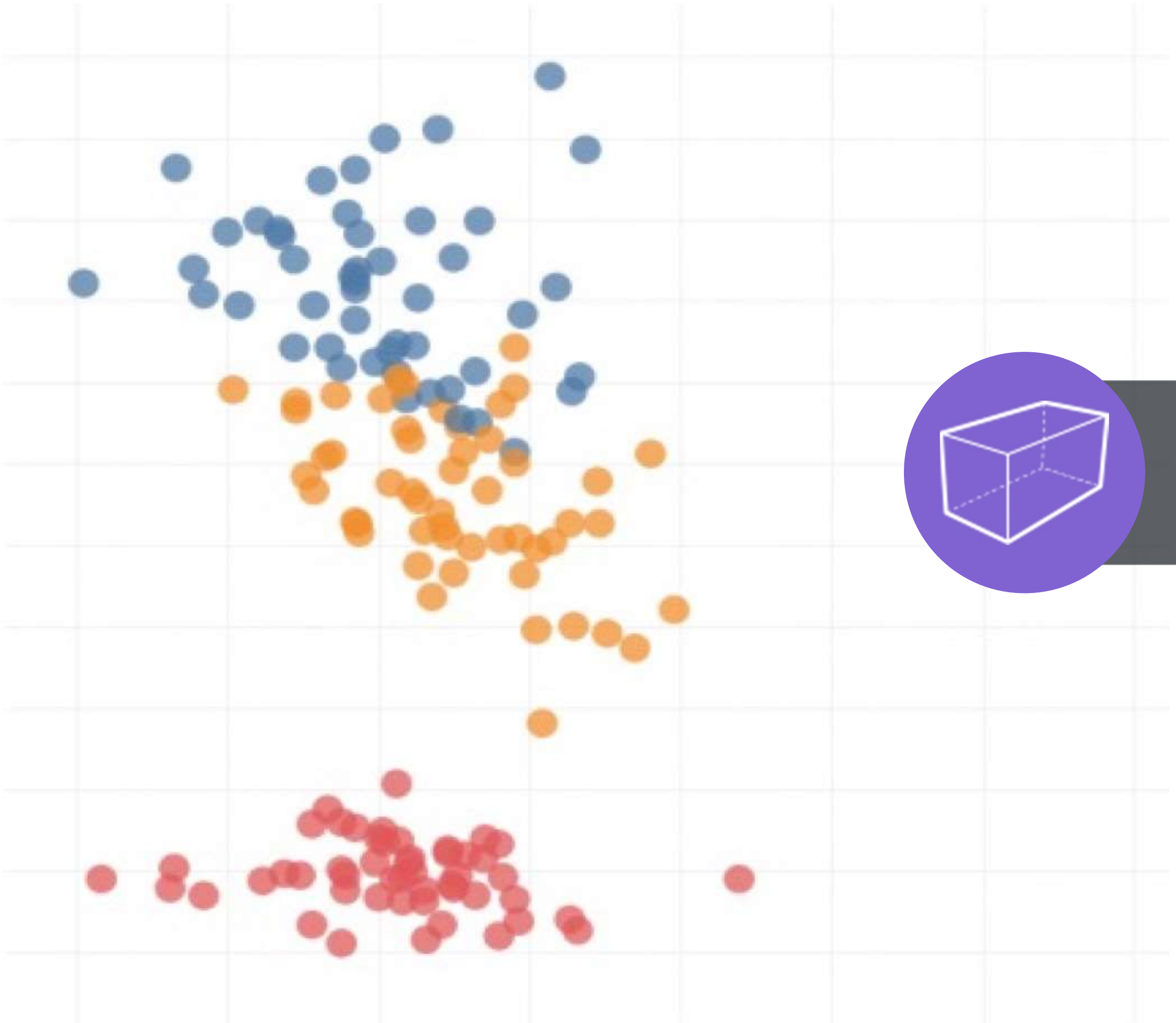
ALGES
ADVANCED LABORATORY FOR
GEOSTATISTICAL SUPERCOMPUTING

Mínimos cuadrados 1

Análisis Estadístico y Geoestadístico de Datos

Auxiliar: Fabián Soto F.
Profesor: Xavier Emery

Ingeniería de Minas
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE



Mínimos cuadrados

Resumen

Mínimos cuadrados

Se busca ajustar un modelo predictivo a los datos disponibles, minimizando el error

$$\chi^2 = \sum_{i=1}^{N} \varepsilon_i^2$$

N: número de datos

$\hat{y}_i$: valor estimado

y_i : valor del dato

$$\varepsilon_i = \hat{y}_i - y_i$$

Regresión lineal univariable

Se busca ajustar un modelo predictivo a los datos disponibles, minimizando el error

$\boldsymbol{a}$: intercepto

$\boldsymbol{b}$: pendiente

$\hat{y}_i$: valor estimado de $\boldsymbol{y_i}$

$\boldsymbol{x_i}$: valor del dato de entrada

$$\hat{y}_i = \boldsymbol{a} + \boldsymbol{b}\boldsymbol{x_i}$$

Regresión lineal univariable

- Intervalos de confianza para $\boldsymbol{a}$ y $\boldsymbol{b}$

$$\hat{y}_i = \boldsymbol{a} + \boldsymbol{b}\boldsymbol{x}_i$$

- Análisis de significancia de los términos de la regresión

$$\boldsymbol{H_0}: b = \mathbf{0}$$
$$\boldsymbol{H_1}: \boldsymbol{b} \neq \mathbf{0}$$

- Coeficiente de determinación múltiple: mide cuánto explica la variable y al utilizar el modelo de regresión con las variables x (cercano a 1 explica mucho, cercano a 0 explica poco).

- Coeficiente de determinación múltiple ajustado: Permite comparar “justamente” modelos con diferente cantidad de parámetros.

Test de un modelo lineal sujeto a condiciones

H_0: Modelo reducido es aceptable

H_1: Modelo expandido es aceptable

$$\boldsymbol{F} = \frac{\frac{\boldsymbol{SE(H_0) - SE(H_1)}}{\boldsymbol{k}}}{\frac{\boldsymbol{SE(H_1)}}{\boldsymbol{N - M}}}$$

$\boldsymbol{k}$: condiciones lineales

$\boldsymbol{N}$: cantidad de datos

$\boldsymbol{M}$: coeficientes libres

$\boldsymbol{SE(H)}$: pendiente suma de cuadrados según la hipótesis $\boldsymbol{H}$

Se compara el estadístico con un Fisher (valor crítico) de $\boldsymbol{k}$ y $\boldsymbol{N}$-$\boldsymbol{M}$ grados de libertad

Test de un modelo lineal sujeto a condiciones

H_0: Modelo reducido es aceptable

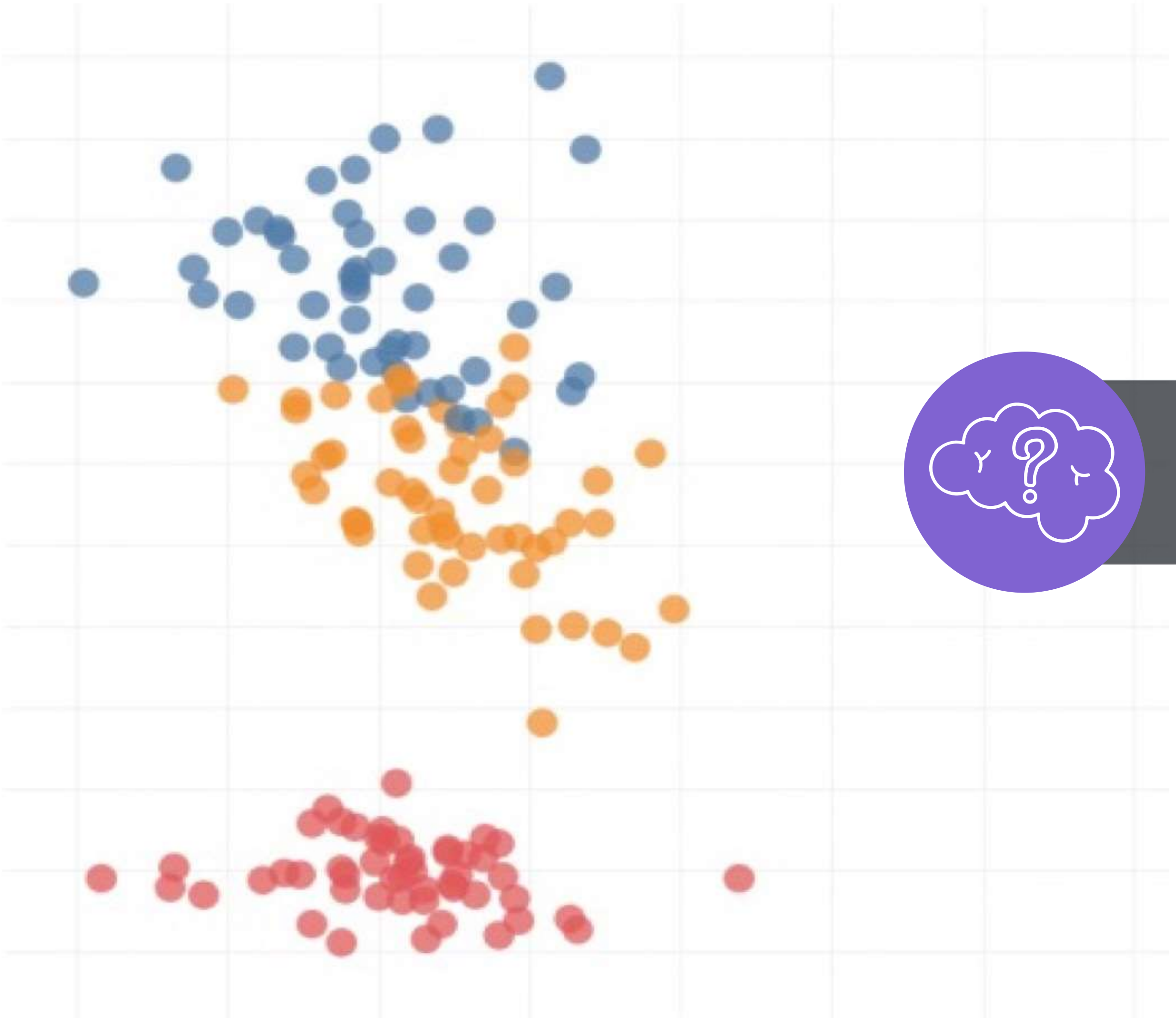
H_1: Modelo expandido es aceptable

$$F = \frac{\frac{SE(H_0) - SE(H_1)}{k}}{\frac{SE(H_1)}{N-M}}$$

- Para el cálculo de la suma de cuadrados (SE) se toma el valor predicho según el modelo de la hipótesis $\mathbf{k}$
- $\hat{y}_{k,\,i}$ es el valor predicho para el dato j según el modelo de la hipótesis $\mathbf{k}$
- ε_i representa el error cometido por el modelo en el dato i-ésimo

$$SE(H_k) = \sum_{i=1}^{N} \varepsilon_i^2$$

$$\varepsilon_i = \hat{y}_{k,\,i} - y_i$$



Problemas

Mínimos cuadrados

Problema 1:

- Modelar la relación entre los datos duplicados que fueron mandados a 2 laboratorios distintos (Archivo *Duplicados.xls*). ¿Es la regresión diferente a la identidad?

Problema 1:

- Modelar la relación entre los datos duplicados que fueron mandados a 2 laboratorios distintos (Archivo *Duplicados.xls*). ¿Es la regresión diferente a la identidad?

Problema 2:

- Se tienen valores para 2 variables x e y, medidas en 2 condiciones distintas (Pestaña *“Comparación Regresiones”* en el archivo *Regresiones.xls*).
- Se desea saber si las rectas de regresión bajo cada condición son iguales o no.

Problema 1

Realizaremos la predicción ($\hat{y}_i$) de los análisis del laboratorio B usando los datos obtenidos en el laboratorio A ($\boldsymbol{x_i}$)

- Debemos evaluar la significancia de la regresión.

$$\boldsymbol{H_0: b = 0}$$
$$\boldsymbol{H_1: b \neq 0}$$

$$\hat{y}_i = \boldsymbol{a + bx_i}$$

¿Qué significa que la regresión sea diferente de la identidad?

Problema 1

Para que la regresión sea igual a la identidad:

Se busca que la regresión $\hat{y}_i = \boldsymbol{a} + \boldsymbol{b}\boldsymbol{x}_i$ sea la identidad $\Longleftrightarrow$ $\boldsymbol{a} = \mathbf{0}$, $\boldsymbol{b} = \mathbf{1}$

- Este es el modelo reducido $\hat{y}_i = \boldsymbol{x}_i$
- Se comparará la significancia de este modelo versus el anterior

Problema 2

Tenemos datos en dos condiciones diferentes (A y B)

Condición A

x	y
0.536	0.54
0.543	0.45
0.377	0.279
0.94	1.176
0.629	0.582
0.826	1.098

Condición B

x	y
0.614	0.876
0.357	0.399
0.556	0.57
1.411	2.025
0.65	1.035
0.864	1.533
0.885	1.266
0.554	0.795

Problema 2

Tenemos dos modelos en dos condiciones diferentes (A y B)

$$\hat{y}_A = \boldsymbol{a} + \boldsymbol{b}\boldsymbol{x}_A$$
$$\hat{y}_B = \boldsymbol{a}' + \boldsymbol{b}'\boldsymbol{x}_B$$

¿Cómo combinar estos modelos en uno solo?

Variable indicador

- Es posible usar una variable indicador (i)
- Este vale 1 si estamos la condición A y vale 0 en caso contrario
- Podemos integrar ambas ecuaciones en una sola

$$\hat{y} = \mathrm{i}(\boldsymbol{a} + \boldsymbol{bx}) + (1 - \mathrm{i})(\boldsymbol{a'} + \boldsymbol{b'x})$$

Cuando i vale 1, se obtiene la ecuación para la condición A, y cuando vale 0 se obtiene la ecuación para la condición B.

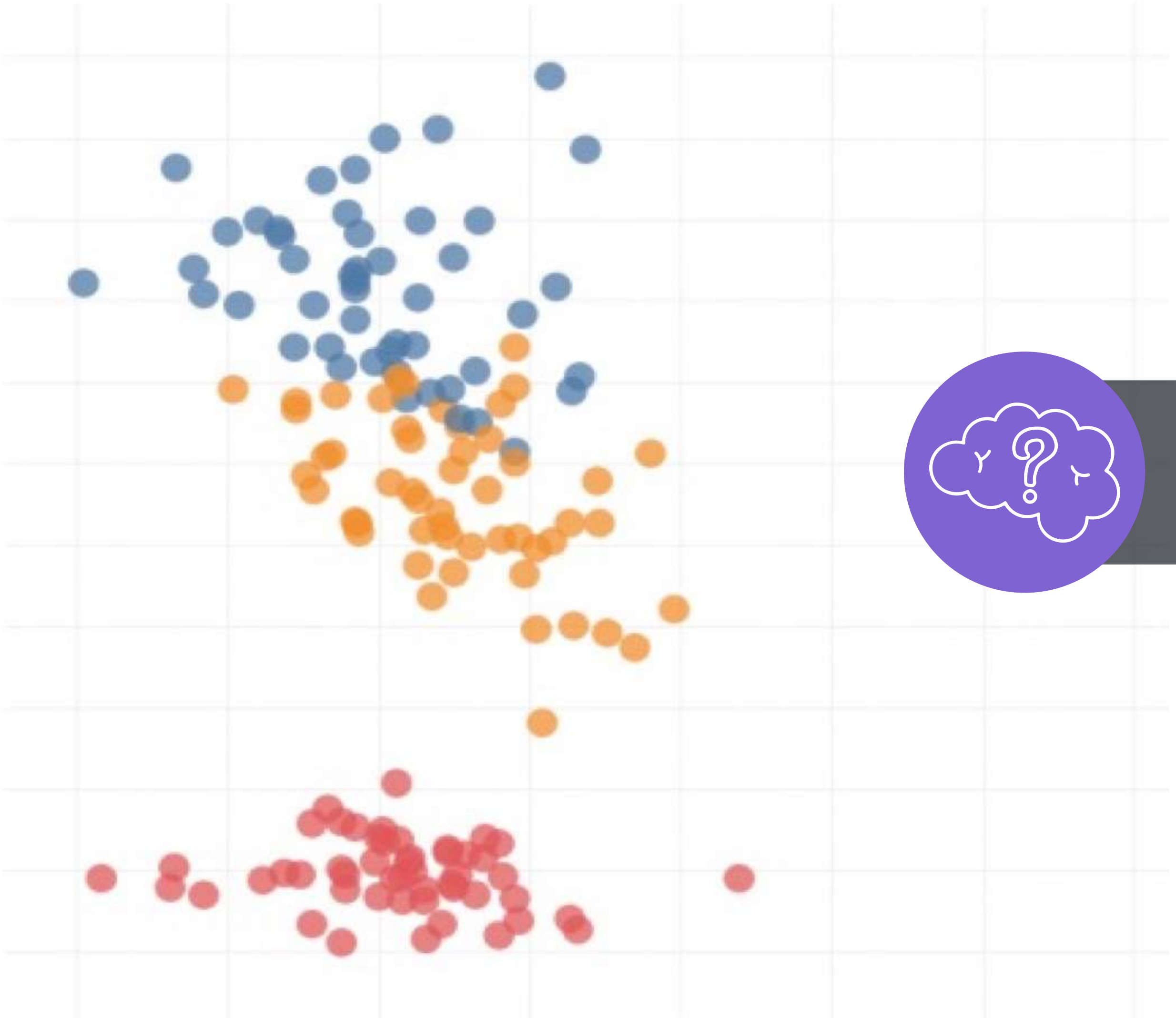
Ecuación final

Reordenando la ecuación final obtenemos:

$$\hat{y} = \boldsymbol{a}' + \boldsymbol{b}'\boldsymbol{x} + (\boldsymbol{a} - \boldsymbol{a}')\boldsymbol{i} + (\boldsymbol{b} - \boldsymbol{b}')\boldsymbol{x}\boldsymbol{i} \iff \hat{y} = \boldsymbol{a}' + \boldsymbol{b}'\boldsymbol{x} + \boldsymbol{c}\boldsymbol{i} + \boldsymbol{d}\boldsymbol{x}\boldsymbol{i}$$

Con:

$$\boldsymbol{c} = (\boldsymbol{a} - \boldsymbol{a}')$$
$$\mathbf{d} = (\boldsymbol{b} - \boldsymbol{b}')$$



¿Dudas?