Métodos matemáticos para la ciencia e Ingeniería: Ajustes de curvas e intervalos de confianza

Fernanda Pérez *(rut: 18.769.232-6)*

24 de Noviembre, 2015

1 Pregunta1

1.1 Introducción

En 1929 Edwin Hubble comparó la velocidad de recesión de las *Nebulosas* (la idea de galaxias lejanas era aún reciente así que se les llamaba nebulosas) con las distancias entre estas Nebulosas y la Tierra. Las distancias fueron medidas usando el método de las Cefeidas.
El archivo **hubble_original.dat** contiene las mediciones originales que utilizó Hubble. El modelo es el siguiente:

$$v = H_0 \cdot D \tag{1}$$

Donde H_0 es la constante de Hubble y generalmente se expresa en unidades de $\frac{km}{s \cdot Mpc}$. Se busca estimar la constante de Hubble utilizando los datos originales, incluyendo su intervalo de confianza al 95%.

1.2 Procedimiento

A partir de los datos en el archivo se crean los arrays de velocidades y distancias. El resultado es diferente si se modela $v = H_0 \cdot D$ (caso 1) ó $D = \frac{v}{H_0}$ (caso 2). Dado que no hay motivo para preferir uno sobre el otro se modelan ambos y se relacionan mediante bisección para buscar la solución final.

1.2.1 Caso 1

Se define una función chi cuadrado de la siguiente forma:

$$\chi^2 = \sum_i (v_i - H_0 \cdot D_i)^2 \tag{2}$$

Minimizando la función anterior con respecto a H_0, se llega a:

$$H_0 = \frac{\sum_i v_i \cdot D_i}{\sum_i {D_i}^2} = H_1 \tag{3}$$

1.2.2 Caso 2

Se define una función chi cuadrado de la siguiente forma:

$$\chi^2 = \sum_i (\frac{v_i}{H_0} - D_i)^2 \tag{4}$$

Minimizando la función anterior con respecto a $\frac{1}{H_0}$, se llega a:

$$H_0 = \frac{\sum_i {v_i}^2}{\sum_i v_i \cdot D_i} = H_2 \tag{5}$$

Utilizando la siguiente relación de bisección el posible calcular el H_{0final}:

$$H_{0final} = \frac{H_1 \cdot H_2 - 1 + \sqrt{(1 + H_1^2)(1 + H_2^2)}}{H_1 + H_2} \tag{6}$$

1.2.3 Simulación Bootstrap

Para calcular el intervalo de confianza se realiza una simulación *Bootstrap*. Para ello, en primer lugar se escoge una semilla (*seed=1234*) y un número de iteraciones definido ($Nboot = 500$). En cada iteración se genera un array de tamaño 24 (que es el tamaño de las muestra de velocidades y distancias), en que sus entradas son números random entre 0 y 24. Posteriormente se calcula H_1 y H_2 como se aprecia en las Ecuaciones 3 y 5, con v_i y D_i el valor de la entrada i-ésima de los arrays de velocidades y distancias respectivamente, donde i corresponde al valor en cada entrada del array generado con números random. Con H_1 y H_2 calculados, se procede a calcular H_{0final} (esto para cada iteración, creando un array al que llamamos *values_bis*).
Luego, ordenando el array obtenido de menor a mayor, es posible obtener los límites del intervalo de confianza de la siguiente forma:

$$limite\_bajo = values\_bis\_ordenado[int(Nboot * 0.025)] \tag{7}$$

$$limite\_alto = values\_bis\_ordenado[int(Nboot * 0.975)] \tag{8}$$

Donde int($\cdot$) es una función que aproxima su argumento al entero inferior.

1.3 Resultados

Realizando lo explicitado en la sección anterior se obtiene:

- $H_1 = 423.94 \frac{km}{s \cdot Mpc}$
- $H_2 = 520.34 \frac{km}{s \cdot Mpc}$
- $H_{0final} = 467.22 \frac{km}{s \cdot Mpc}$

Utilizando la simulación Bootstrap se obtiene un intervalo de confianza de $[388.66; 550.57]$. Se obtienen las Figuras 1 y 2.

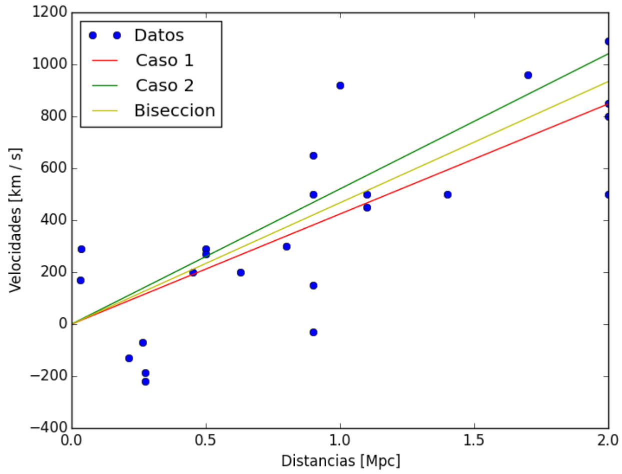

Figura 1: Gráfica de los datos del archivo **hubble_original.dat**, junto a las rectas obtenidas utilizando los modelo del caso 1 ($v = H_0 \cdot D$) y caso 2 ($D = \frac{v}{H_0}$), y la recta final utilizando el método de bisección de las dos anteriores.

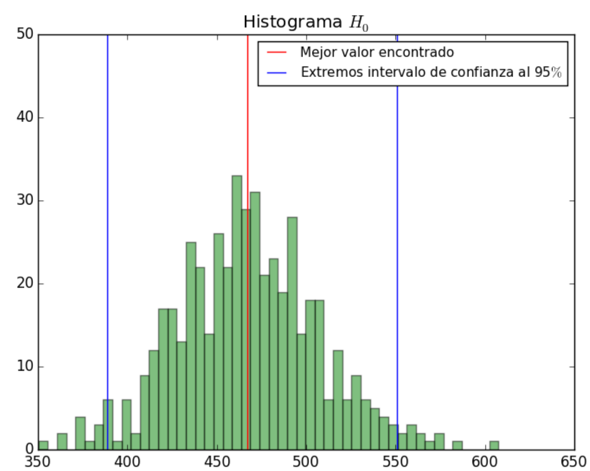

Figura 2: Histograma para H_0 obtenido al realizar simulación Bootstrap con los datos del archivo **hubble_original.dat**. Se indica con una recta roja el mejor valor encontrado (H_{0final}) y con dos azules los extremos encontrados del intervalo de confianza al 95%.

2 Pregunta 2

2.1 Introducción

Hubble cometió un error en su estimación de H_0. Una estimación más reciente de la constante de Hubble se obtiene con los datos ubicados en el archivo **SNIa.dat** (Freedman et al. 2000) que utiliza Super Novas tipo I para estimar distancias para una muestra de galaxias. Entre otras ventajas, el método permite estimar distancias muy superiores a las que se pueden medir con el método de las Cefeidas. Se busca volver a estimar la constante de Hubble, con estos datos, incluyendo su intervalo de confianza al 95%.

2.2 Procedimiento

Se utiliza el mismo procedimiento mencionado para la Pregunta 1, con la consideración de que en este caso los arrays de velocidades y distancias son de tamaño 36, y por lo tanto el array que se crea en la simulación *Bootstrap* será de este tamaño y tendrá en sus entradas número aleatorios entre 0 y 36.

2.3 Resultados

Realizando el procedimiento ya señalado, se obtiene:

- $H_1 = 70.67\frac{km}{s \cdot Mpc}$
- $H_2 = 70.02\frac{km}{s \cdot Mpc}$
- $H_{0final} = 70.84\frac{km}{s \cdot Mpc}$

Utilizando la simulación Bootstrap se obtiene un intervalo de confianza de $[68.70; 73.37]$. Se obtienen las Figuras 1 y 2.

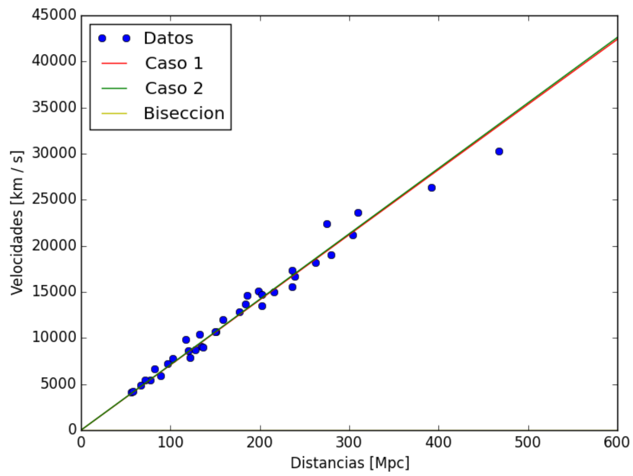

Figura 3: Gráfica de los datos del archivo **SNIa.dat**, junto a las rectas obtenidas utilizando los modelo del caso 1 ($v = H_0 \cdot D$) y caso 2 ($D = \frac{v}{H_0}$), y la recta final utilizando el método de bisección de las dos anteriores.

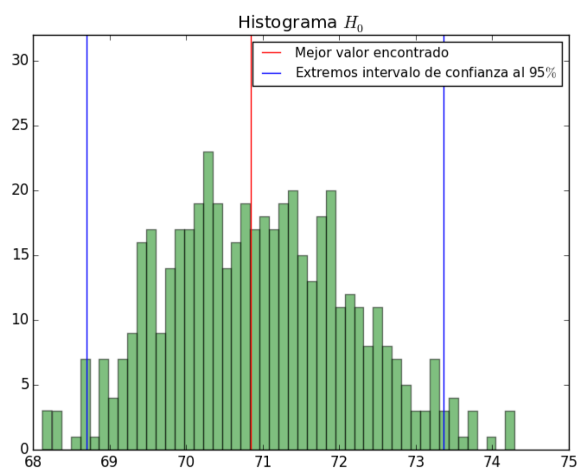

Figura 4: Histograma para H_0 obtenido al realizar simulación Bootstrap con los datos del archivo **SNIa.dat**. Se indica con una recta roja el mejor valor encontrado (${H_0}_{final}$) y con dos azules los extremos encontrados del intervalo de confianza al 95%.

3 Conclusiones Preguntas 1 y 2

Dada la gran diferencia entre los valores de H_{0final} obtenidos en cada pregunta ($467.22\frac{km}{s \cdot Mpc}$ vs $70.84\frac{km}{s \cdot Mpc}$), se infiere que los datos originales de Hubble estaban bastante equivocados. Esto nos da la noción de la importancia de la relación período-luminosidad.

Las rectas obtenidas que mejor modelan los datos (rectas amarillas en Figuras 1 y 3) hacen sentido con la gráfica misma de los datos.

Los largos relativos de los intervalos obtenidos en cada pregunta ($[388.66; 550.57]$ vs $[68.70; 73.37]$), hacen sentido dado que los datos del Problema 2 son una mejor estimación que los datos del problema 1.

4 Pregunta 3

4.1 Introducción

El archivo **DR9Q.dat** es una sección recortada del catálogo de cuasares del Data Release 9 del Sloan Digital Sky Survey (SDSS). Se busca encontrar la línea recta que mejor modela la relación entre el flujo en la banda i (columna 81°) y la banda z (columna 83°), incluyendo los intervalos de confianza al 95% para los parámetros de la línea recta. Los errores para el flujo en la banda i y z se encuentran en las columnas número 82 y 84 respectivamente.

4.2 Procedimiento

En primer lugar, a partir de los datos del archivo, se crean los arrays *flujo_i*, *flujo_z*, *error_i* y *error_z* donde los valores son multiplicados por *3.631* con el fin de dejar todo en unidades de $10^{-6} Jy$.
Para encontrar la recta que mejor modela la relación pedida, se hace un ajuste lineal utilizando *np.polyfit(flujo_i, flujo_z, 1)*.

Para calcular los intervalos de confianza se utiliza una simulación de *Monte Carlo*. Para ello se escoge una semilla (*seed=1234*) y una cantidad de iteraciones definida ($Nmc = 10000$).

4.2.1 Simulación de Monte Carlo

En cada iteración se crea un array (r) de tamaño 36 (que es el tamaño de las muestra de flujos y errores), en que sus entradas son números random entre 0 y 1. Luego se crean los siguientes arrays:

$$muestra\_i = flujo\_i + error\_i \cdot r \tag{9}$$

$$muestra\_z = flujo\_z + error\_z \cdot r \tag{10}$$

Dado que no se sabe cuál es la coordenada dependiente y cuál la independiente, deben considerarse ambos casos tal cual como se hizo en las Preguntas 1 y 2. Se realizan dos ajustes lineale, *np.polyfit(muestra_i, muestra_z, 1)* y *np.polyfit(muestra_z, muestra_i, 1)*, se obtiene las pendientes y los coeficientes de posición, que se *mezclan* de alguna manera que sea simétrico para ambos casos (esto para cada iteración, creando dos arrays: *pendientes* y *coefs_posicion*).

Se ordenan de menor a mayor ambos arrays obtenidos con la simulación y luego los límites de los intervalos de confianza para cada parámetro son calculados de la siguiente manera:

$$limite\_bajo\_pendiente = pendientes\_ordenado[int(Nmc * 0.025)] \tag{11}$$

$$limite\_alto\_pendiente = pendientes\_ordenado[int(Nmc * 0.975)] \tag{12}$$

$$limite\_bajo\_coefs = coefs\_posicion\_ordenado[int(Nmc * 0.025)] \tag{13}$$

$$limite\_alto\_coefs = coefs\_posicion\_ordenado[int(Nmc * 0.975)] \tag{14}$$

Una posible forma [1]de *mezclar* los resultados de las pendientes y coeficientes de posición es la siguiente:

$$pendiente\_final = tan(\frac{arctan(pendiente\_1) + arctan(pendiente\_2)}{2}) \tag{15}$$

$$coef\_final = y - pendiente\_final \cdot x \tag{16}$$

Donde:

$$x = \frac{coef\_1 \cdot pendiente\_2 + coef\_2}{1 - pendiente\_1 \cdot pendiente\_2} \tag{17}$$

$$y = pendiente\_1 \cdot x + coef\_1 \tag{18}$$

[1] Esta forma fue sacada del informe de Bruno Scheihing.

4.3 Resultados

Realizando el ajuste lineal se obtiene la Figura 3.

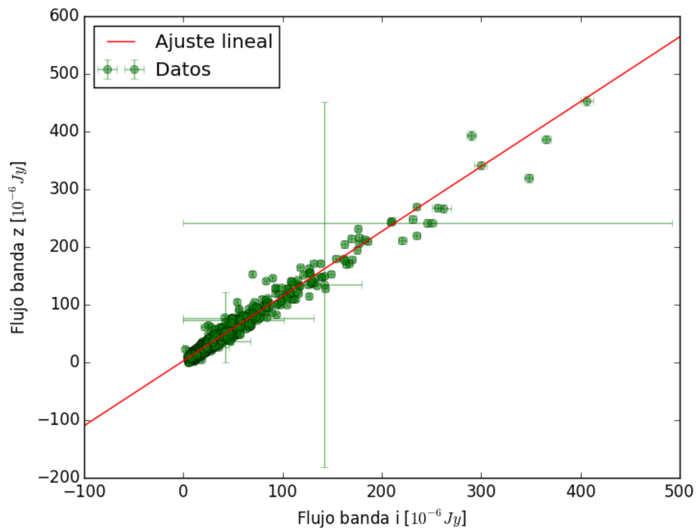

Figura 5: Gráfica de los flujos en las bandas i y z junto a sus respectivas barras de errores. La recta roja es obtenida realizando el ajuste lineal (pendiente = 1.10, coeficiente de posición = 3.15).

Utilizando la simulación de Monte Carlo se obtienen los siguientes intervalos de confianza:

- Pendiente: $[1.04; 1.23]$
- Coeficiente de posición: $[-0.49; 4.95]$

A partir de la misma simulación se generan las Figuras 6 y 7.

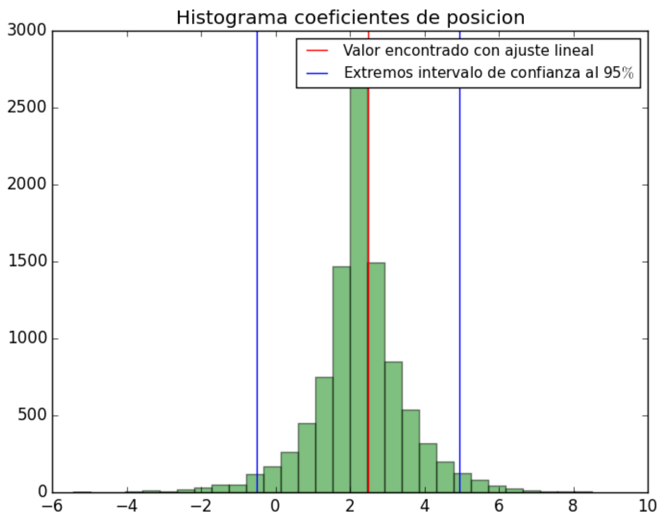

Figura 6: Histograma de los coeficientes de posición, obtenido al realizar simulación de Monte Carlo. Se indica con una recta roja el valor encontrado con el ajuste lineal y con dos azules los extremos encontrados del intervalo de confianza al 95%.

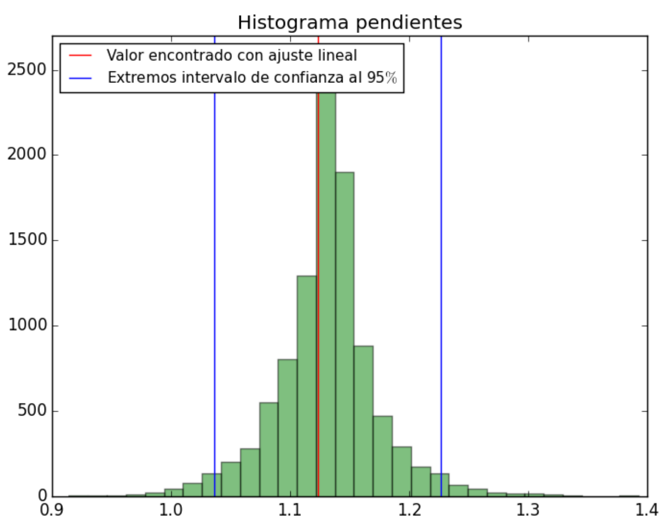

Figura 7: Histograma de las pendientes, obtenido al realizar simulación de Monte Carlo. Se indica con una recta roja el valor encontrado con el ajuste lineal y con dos azules los extremos encontrados del intervalo de confianza al 95%.

4.4 Conclusiones

La recta generada mediante ajuste lineal de la Figura 5 hace sentido con la gráfica misma de los datos.
El intervalo de confianza para la pendiente resulta ser bastante pequeo, $[1.04; 1.23]$, en cambio el del coeficiente de posición no, $[-0.49; 4.95]$.
En las Figuras 6 y 7, los valores encontrados con ajuste lineal no están exactamente centrados en la curva que genera el histograma, esto puede deberse a la forma en que se *mezclaron* las variables en la simulación de Monte Carlo con el fin de buscar simetría.
Se podrían intentar otras formas de *mezclar* variables para buscar una mejor simetría.