Métodos Numéricos para la Ciencia e Ingeniería
FI3104-1

Tarea 9

Maximiliano Dirk Vega Aguilera
18.451.231-9

1 Introducción

La tarea 8 consistió en utilizar algoritmos estadísticos para realizar un ajuste lineal a tres sets de datos y determinar un intervalo de confianza para los parámetros de cada ajuste. Para ello se utilizaron los métodos de simulación de Monte Carlo y de Bootstrap. Adicionalmente, se debió escribir el programa de forma que cumpliera las reglas de PEP8.

La primera parte consistió en derivar la constante de hubble (H_0) a partir de los datos originales utilizados por Hubble en 1929. En estos datos se relaciona la velocidad de recesión de las galaxias [km s^{-1}] (en ese tiempo llamadas nebulosas) y su distancia a la Tierra [Mpc], donde las distancias fueron obtenidas mediante la relación periodo-luminosidad que presentaban las estrellas cefeidas. El objetivo era encontrar el parámetro H_0 de la relación $v = H_0 * d$ y obtener un intervalo de confianza al 95% para este.

Para ello, a los datos se les hizo un ajuste lineal del tipo $y = a+bx$ considerando $a = 0$, ya que $a \neq 0$ no tiene sentido físico, pues una galaxia no puede tener una velocidad de recesión a distancia 0. Una vez obtenido el parámetro, y debido a que no se conocen los errores de las mediciones, se utilizó el método de Bootstrap para encontrar un intervalo de confianza al 95%.

Este método consiste en generar muestras sintéticas a partir de la muestra original y calcular los parámetros de ajuste lineal para cada muestra sintética. Las muestras sintéticas se obtienen tomando valores de la muestra original, al azar y con igual probabilidad, estos valores pueden repetirse. Una vez obtenido los parámetros de ajuste, se define el intervalo de confianza al 95% como el intervalo que contiene el 95% de los parámetros. Para que este modelo funcione correctamente, se acepta que con $N * log^2(N)$ muestras sintéticas se obtienen resultados aceptables, con N el tamaño de la muestra. Sin embargo, dado que los datos de esta primera parte son pocos, se consideró un $N = N^2$.

La segunda parte es análoga a la primera, pero utilizando datos de Super Novas tipo I para obtener las distancias. Dado que los datos son más precisos, se espera que el valor de H_0 obtenido fuese más cercano al real.

La tercera parte consistió en relacionar el flujo [nmaggies][1] de la banda i con el de la banda z a partir de los datos provenientes de un catalogo de cuasares, realizando un ajuste lineal y obteniendo un intervalo de confianza al 95% para los parámetros. Dado que se conocían los errores de las mediciones, se utilizó la simulación de Monte Carlo para encontrarlo.

Este método consiste en generar muestras sintéticas a partir de los datos y sus errores. Cada muestra sintética se obtiene de una distribución normal centrada en el dato con el error como desviación estándar, esto suponiendo que los errores son normales. Luego, se obtuvo el intervalo de confianza al 95% de forma análoga a lo anterior.

2 Desarrollo

Para el desarrollo de esta tarea se utilizó el lenguaje de programación python y sus paquetes numpy, matplotlib y scipy. Para las dos primeras partes se hizo el mismo procedimiento pero cambiando los archivos que contienen los datos.

Primero se obtuvieron los datos y se guardaron en arreglos separados, luego se realizó un ajuste lineal a partir de la minimización de la función χ^2 (considerando $y = a + bx$):

$$\chi^2(a,b) = \sum_{i=1}^{N} (\frac{y_i - a - bx_i}{\sigma_i})^2 \tag{1}$$

$$\frac{\partial \chi^2}{\partial a} = 0 = -2 \sum \frac{y_i - a - bx_i}{\sigma_i^2} \tag{2}$$

$$\frac{\partial \chi^2}{\partial b} = 0 = -2 \sum \frac{x_i(y_i - a - bx_i)}{\sigma_i^2} \tag{3}$$

Donde las soluciones son:

$$a = \frac{SS_{xx} - S_x S_{yy}}{\Delta} \tag{4}$$

$$b = \frac{SS_{xy} - S_x S_y}{\Delta} \tag{5}$$

Con:

$$\Delta = SS_{xx} - S_x^2 \tag{6}$$

$$S = \sum \frac{1}{\sigma_i^2} \;\; ; \;\; S_x = \sum \frac{x_i}{\sigma_i^2} \;\; ; \;\; S_y = \sum \frac{y_i}{\sigma_i^2} \;\; ; \;\; S_{xx} = \sum \frac{x_i^2}{\sigma_i^2} \;\; ; \;\; S_{xy} = \sum \frac{x_i y_i}{\sigma_i^2} \tag{7}$$

Eso en el caso general, pero acá consideramos $a = 0$ y $\sigma_i \bar{1}$ (debido a que no se conocen los errores), por lo que se obtiene la siguiente relación:

$$\frac{d\chi^2}{db} = 0 \tag{8}$$

$$b = \frac{\sum_i x_i y_i}{\sum_i x_i^2} = \frac{S_{xy}}{S_{xx}} \tag{9}$$

[1] 1 [nmaggie] $\sim 3.631\text{e}^{-6}$ [Jy]

Dado que no hay razón para preferir a la velocidad dependiente de la distancia o a la distancia dependiente de la velocidad, se realizó el análisis para ambos casos y se consideró la recta que pasa por entremedio de ambas, cuya pendiente viene dada por:

$$b_{biseccion} = \frac{b_1 b_2 - 1 + \sqrt{(1+b_1^2)(1+b_2^2)}}{b_1 + b_2} \tag{10}$$

Con esto, para los datos de las cefeidas y de las supernovas, se obtienen los gráficos de las figuras 1 y 2 respectivamente. Para los datos proveniente de las cefeidas se obtuvo un valor de $H_0 = 491.9[Km\ s^{-1}\ Mpc^{-1}]$ en un intervalo de confianza al 95% de (375.7, 550.0)$[Km\ s^{-1}\ Mpc^{-1}]$. Para los datos proveniente de las supernovas se obtuvo un valor de $H_0 = 68.8[Km\ s^{-1}\ Mpc^{-1}]$ en un intervalo de confianza al 95% de $(68.6, 73.2)[Km\ s^{-1}\ Mpc^{-1}]$.

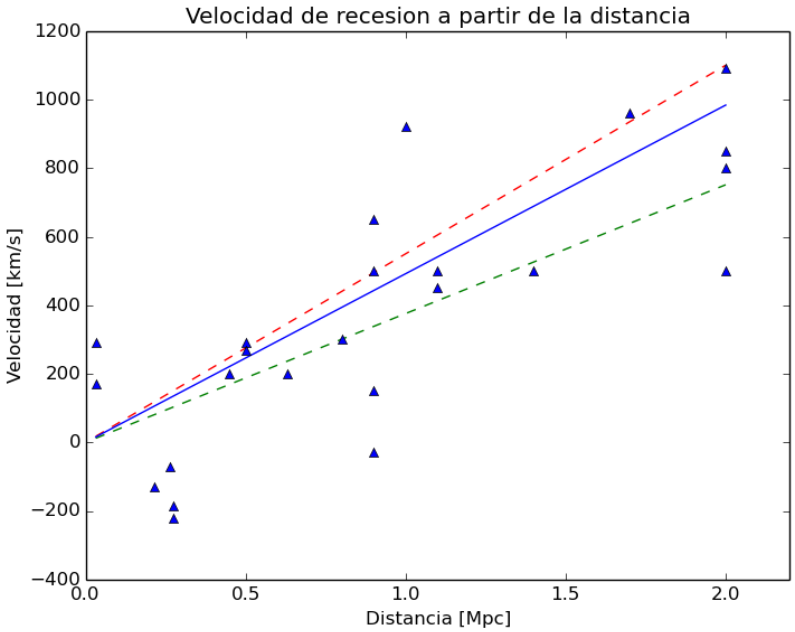

Figure 1: Distancias obtenidas con las estrellas cefeidas. La recta azul indica el ajuste lineal realizado a los datos. Las rectas roja y verde representan el límite superior e inferior del intervalo de confianza para el parámetro H_0.

La gran diferencia se debe en parte a que Hubble estimó mal las distancia utilizando las cefeidas, lo que naturalmente se traduce en un valor equivocado de H_0, además los datos presentan una dispersión mayor comparado con los datos de supernova. También vale destacar que los valores de H_0, para los dos casos, están dentro de sus intervalos de confianza, por lo que el método funciona bien.

Para la tercera parte se procedió de forma similar, pero se obtuvo el ajuste lineal a partir de la función polyfit de numpy y se cambió la forma en que se obtienen las muestras sintéticas. Para cada dato, se generó un dato sintético a partir un valor random obtenido de una distribución normal, centrada en ese dato y usando el error del dato como desviación estándar. Para cada muestra sintética generada se obtuvieron sus parámetros del ajuste

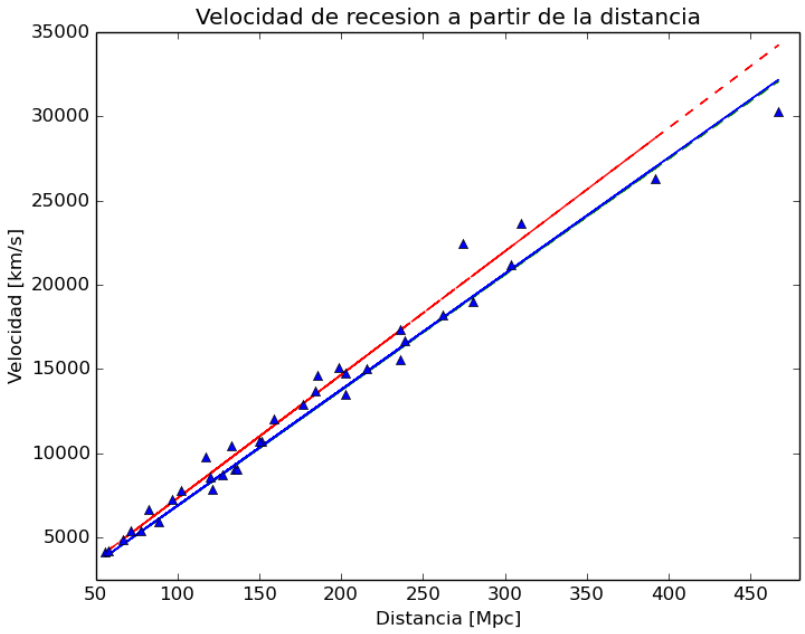

Figure 2: Distancias obtenidas con las supernovas de tipo I. La recta azul indica el ajuste lineal realizado a los datos. Las rectas roja y verde representan el límite superior e inferior del intervalo de confianza para el parámetro H_0. Notar que la recta azul casi calza con la recta verde.

lineal de la misma forma que los obtenido para los datos reales. Estos parámetros fueron ordenados y se calculó el intervalo que contiene el 95% de ellos, de modo que el menor y mayor valor del arreglo restante define el intervalo de confianza al 95%. Además, al obtener los datos se dividieron por 3.631 de modo que estuviesen en las unidades de [1e^{-6} Jy].

Se obtuvo el gráfico de la figura 3, con los parámetros b = 1.10 en un intervalo de confianza al 95% de (0.93, 1.13) y a = 0.23 en un intervalo de confianza al 95% de (0.19, 0.62).

A pesar de que hay 2 datos con errores grandes, estos no afectaron al método de simulación de Monte Carlo, debido principalmente a la gran cantidad de datos existentes.

3 Conclusión

Los métodos estadísticos nos permiten realizar una aproximación a la función real que modela una situación a partir de los datos obtenidos de forma experimental. Mientras más datos conozcamos de lo que se quiere modelar, mejores resultados se obtendrán y los errores de las mediciones perderán importancia (dentro de cierto margen de precisión).

Finalmente, es importante realizar una correcta determinación de los datos para que se obtengan conclusiones lo más cercanas a la realidad posible, esto quedó en evidencia en el cálculo de la constante de hubble.

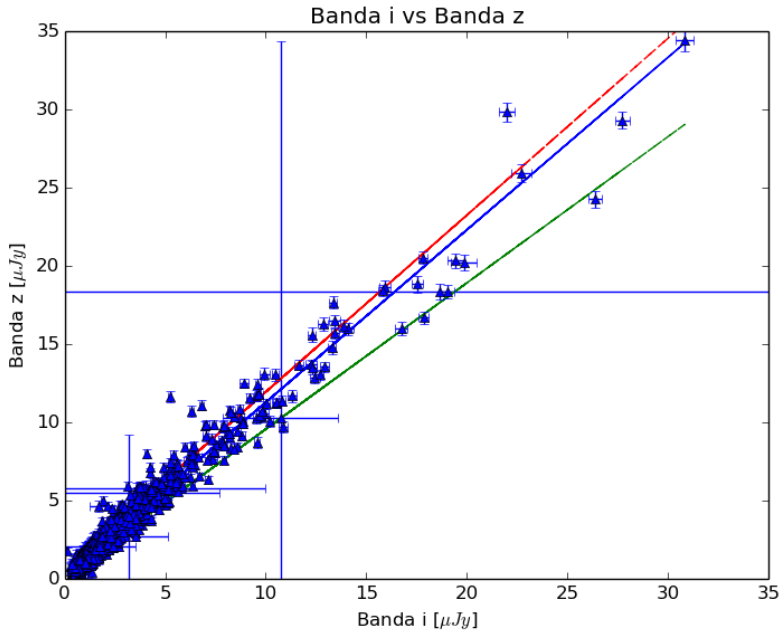

Figure 3: La recta azul indica el ajuste lineal realizado a los datos. Las rectas roja y verde representan el límite superior e inferior del intervalo de confianza para los parámetros a y b del ajuste lineal.