Tarea 09: Ajustes Lineales y Cálculo de Intervalos de Confianza

Eva Díaz

November 24, 2015

Curso:	Métodos Numéricos para la Ciencia e Ingeniería
Profesor:	Valentino González
Profesor Auxiliar:	Felipe Pesce

1 Introducción

La tarea consiste en realizar ajustes lineales de datos mediante la minimización de la función χ^2 (método de los cuadrados mínimos) y una simulación de Monte Carlo con el objetivo de encontrar los parámetros de la recta de ajuste. Además, mediante una simulación de Bootstrap, se calculan los intervalos de confianza de estos parámetros.

Estos procedimientos se aplican para resolver dos problemas: encontrar el valor numérico de la constante de Hubble utilizando dos conjuntos de datos distintos y para calcular la relación entre flujos luminosos provenientes de cuásares.

2 Constante de Hubble I

En 1929 Edwin Hubble comparó la velocidad de recesiń de las Nebulosas mediante corrimiento del espectro (la idea de galaxias lejanas era an reciente asi que se les llamaba nebulosas) con las distancias entre estas Nebulosas y la Tierra. Las distancias fueron medidas usando el método de las Cefeidas, que son estrellas de luminosidad variables cuyo período y luminosidad están fuertemente correlacionados. La llamada relación períodoluminosidad había sido recientemente calibrada.

Hubble no fue el primero en calcular esta relación pero su trabajo fue de los más influyentes de la época, convenciendo al mundo de la expansión del Universo. El modelo utilizado por Hubble es muy simple, dado por la siguiente ecuación.

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

El objetivo de esta parte es encontrar el valor de la constante de Hubble H_0 a partir de los datos originales de velocidad de recesión (en km/s) y distancia de las galaxias (en Mpc) utilizados por Edwin Hubble. Además se calcula el intervalo de confianza del 95% para el valor encotrado.

Para encontrar este valor se realiza un ajuste lineal de los datos de modo de encontrar la pendiente que los relaciona, esto encontrando la ecuación que minimiza la función χ^2.

Se define la función χ^2 como:

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

Esta función es una medida de la desviación de los valores observados de v respectos de los predichos por la expresión $H_0 * D$. El método de los cuadrados mínimos indica que el mejor valor de la constante H_0 es aquel que minimiza esta desviación. Este valor es fácil de encontar: se deriva χ^2 respecto del parámetro H_0 y a continuación se despeja éste.

$$\frac{d\chi^2}{dH_0} = \frac{d}{dH_0} \sum_i (v_i - H_0 D_i)^2 = 2 \sum_i v_i D_i - H_0 D_i^2 = 0 \tag{3}$$

Despejando H_0 de la expresión anterior se tiene que:

$$H_0 = \frac{\sum_i v_i D_i}{\sum_i D_i^2} \tag{4}$$

Por otro lado, siguiendo el mismo procedimiento anterior, se puede realizar el ajuste lineal con la distancia y la velocidad intercambiadas, de modo que al minimizar χ^2 se obtiene el mejor valor de $1/H_0$. Esto se muestra a continuación:

$$\chi^2 = \sum_i (D_i - (1/H_0) v_i)^2 \tag{5}$$

$$\frac{d\chi^2}{d(1/H_0)} = \frac{d}{d(1/H_0)} \sum_i (D_i - (1/H_0) v_i)^2 = 0 \tag{6}$$

$$2 \sum_i D_i v_i - (1/H_0) v_i^2 = 0 \tag{7}$$

$$\frac{1}{H_0} = \frac{\sum_i D_i v_i}{\sum_i v_i^2} \tag{8}$$

$$H_0 = \frac{\sum_i v_i^2}{\sum_i D_i v_i} \tag{9}$$

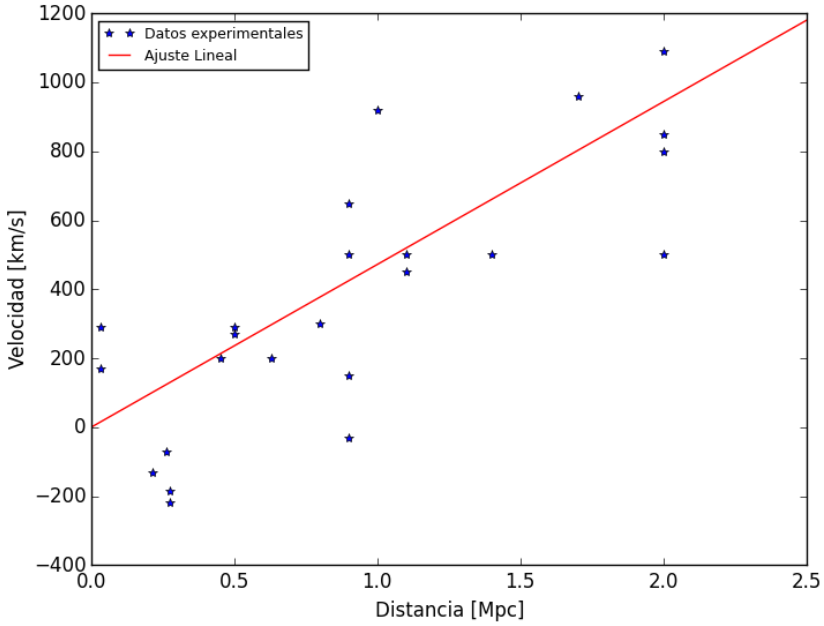

Figure 1: Comparación entre los datos de distancia y velocidad de las galaxias y su ajuste lineal.

Las ecuaciones 4 y 9 entregan dos formas de calcular la constante H_0, por lo que el valor final de la constante se tomará como el promedio de ambas. Se realizó el cálculo numérico de las dos expresiones, el cual resultó en los valores $H_0 = 493.94$ km/s/Mpc y $H_0 = 520.34$ km/s/Mpc, respectivamente. El promedio entre estos dos resultados entonces es de $H_0 = 472.14$ km/s/Mpc. La comparación entre los datos experimentales y el ajuste lineal realizado se muestra en la Figura 1.

Finalmente se calcula el intervalo de confianza del 95% de la constante de Hubble que se acaba de encontrar. Para hacer esto se utilizó el método de Bootstrap, debido a que los datos de que se disponen no incluyen errores de medición. Se aplicó el método sobre el promedio entre los dos valores de H_0 calculados, utilizando una semilla de 1943 y con 1000 simulaciones. Se encontró que el intervalo de confianza de la constante entontrada es [391.51 ; 555.57] km/s/Mpc. La Figura 2 ilustra los valores de H_0 obtenidos mediante este método y su frecuencia, así como los límites del intervalo de confianza.

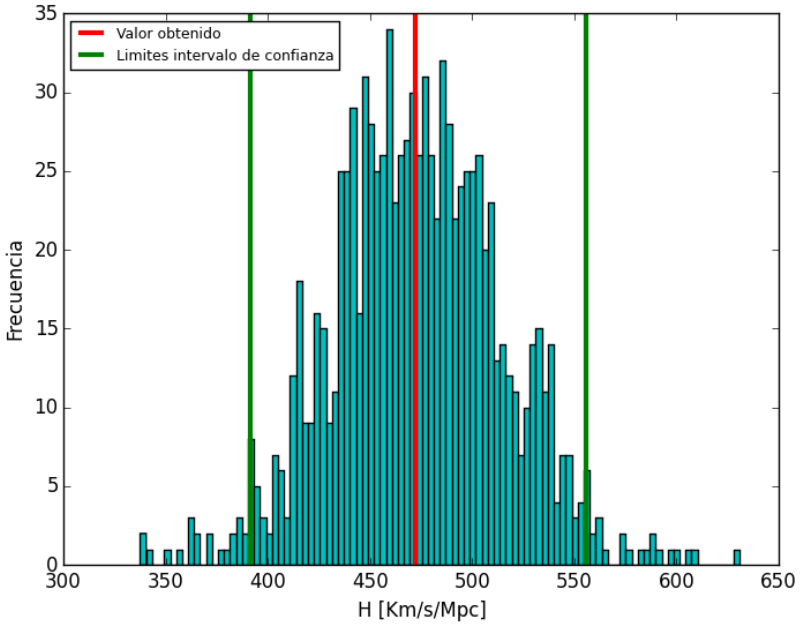

Figure 2: Histograma de valores de H_0. Número de simulaciones: 1000.

El valor encontrado de la constante de Hubble, $H_0 = 472.14$ km/s/Mpc, se condice con el valor calculado por Edwin Hubble en 1929 ($\sim$500 km/s/Mpc). Este primer valor era insuficiente ya en esa época debido a que indica que la edad del universo sería de unos 2000 millones de años, mientras que por estudios de los isótopos de las rocas terrestres se sabía que la edad de Tierra era de aproximadamente 4500 millones de años. Este error en los cálculos de Hubble proviene de, entre otras cosas, utilizar una calibración equivocada de la relación período-luminosidad de las Cefeidas, por lo cual la medición de las distancias es errónea.

El amplio intervalo de confianza encontrado responde a que la muestra examinada es muy pequeña, de tan sólo 24 galaxias. Se recomienda que el número de simulaciones sea de $N \log(N)^2$, es decir, 46 simulaciones para 24 galaxias. Debido al reducido tamaño de la muestra se considera razonable realizar un número mayor de simulaciones (de manera de obtener el valor más preciso posible del intervalo), razón por la cual se escogió realizar 1000 de éstas.

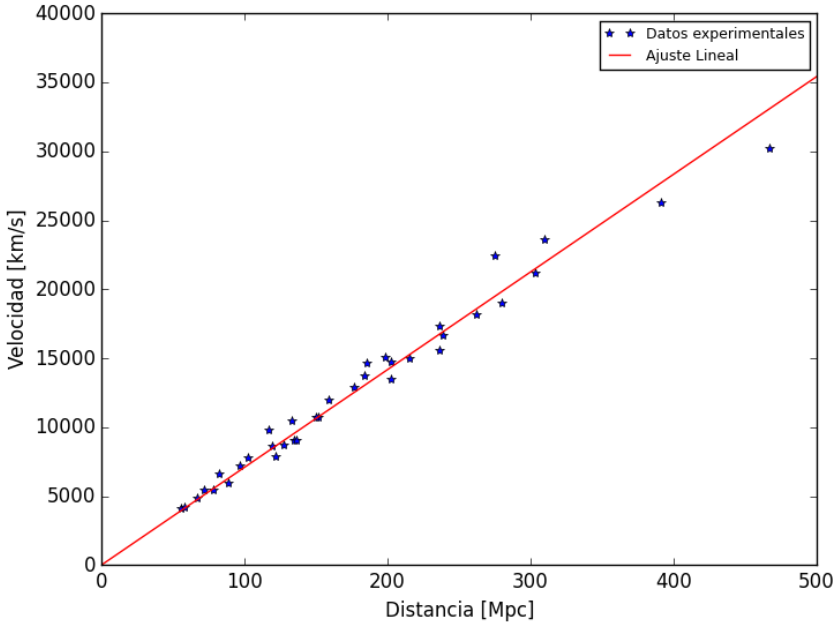

Figure 3: Datos experimentales y ajuste lineal de datos más recientes de distancia y velocidad de galaxias.

3 Constante de Hubble II

Una estimación más reciente de la constante de Hubble se puede realizar utilizando supernovas de tipo I para estimar las distancias de una muestra de galaxias. Esta forma de medición presenta ventajas que permiten estimar distancias muy superiores a las que se pueden medir con el método de las Cefeidas.

El objetivo de esta parte es repetir el procedimiento realizado en la primera parte (valor de la constante de Hubble e intervalo de confianza al 95%) pero con los datos proporcionados por Freedman et al (2000).

Mediante el procedimiento anterior de ajuste lineal se obtuvo que los valores de H_0 según las ecuaciones 4 y 9 es de $H_0 = 70.66$ km/s/Mpc y $H_0 = 71.01$ km/s/Mpc, respectivamente. El promedio entre estos dos valores da como resultado $H_0 = 70.84$ km/s/Mpc. La Figura 3 muetra la comparación entre los datos experimentales y el ajuste lineal que se hizo.

Como antes, se utilizó el método de Bootstrap para calcular el intervalo de

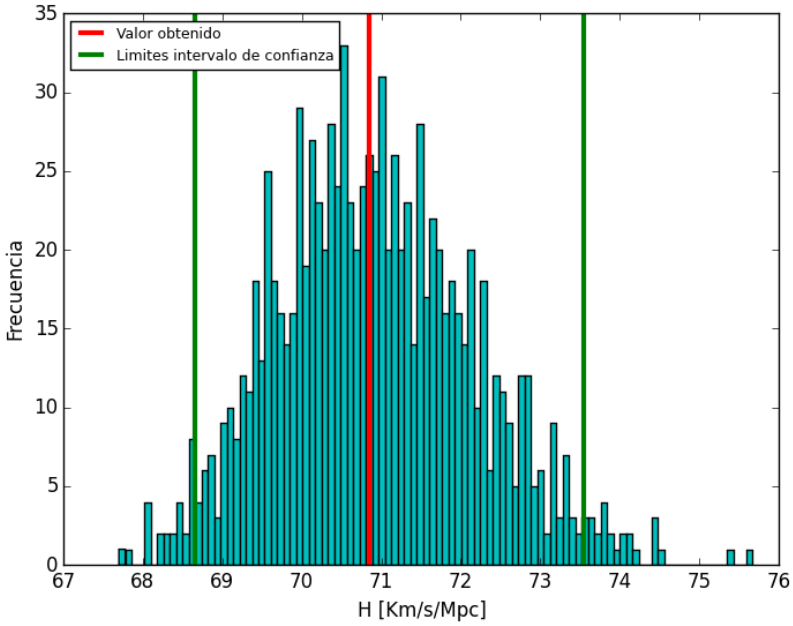

Figure 4: Histograma de H_0 hecho con datos recientes de distancia y velocidad de galaxias. Número de simulaciones: 1000.

confianza al 95% con una semilla de 1943 y 100 simulaciones se obtuvo que el intervalo es [68.65 ; 73.54] km/s/Mpc. La Figura 4 muestra los valores de H_0 obtenidos con Bootstrap y su frecuencia, así como los límites del intervalo de confianza.

El nuevo valor encontrado de la constante de Hubble, $H_0 = 70.84$ km/s/Mpc, es notoriamente menor que el encontrado en la parte anterior y obviamente más preciso y cercano al valor actual, que ronda entre 70 y 80 km/s/Mpc según varios estudios realizados desde el año 2000 en adelante. Según este valor de H_0 la edad del universo sería de 10000 millones de años, lo cual aún es insuficiente para dar cuenta de las estrellas más antiguas de los cúmulos globulares, con una edad de unos 14000 millones de años. Sin embargo, estudios posteriores de supernovas lejanas revelaron que existe otro factor que impulsa la expansión del universo: la denominada energía oscura. Ajustando el cálculo para considerar este último factor la edad del universo se acerca a los 14000 millones de años. Se tiene, por lo tanto, que el valor calculado de la constante de Hubble es preciso y se condice con lo observado.

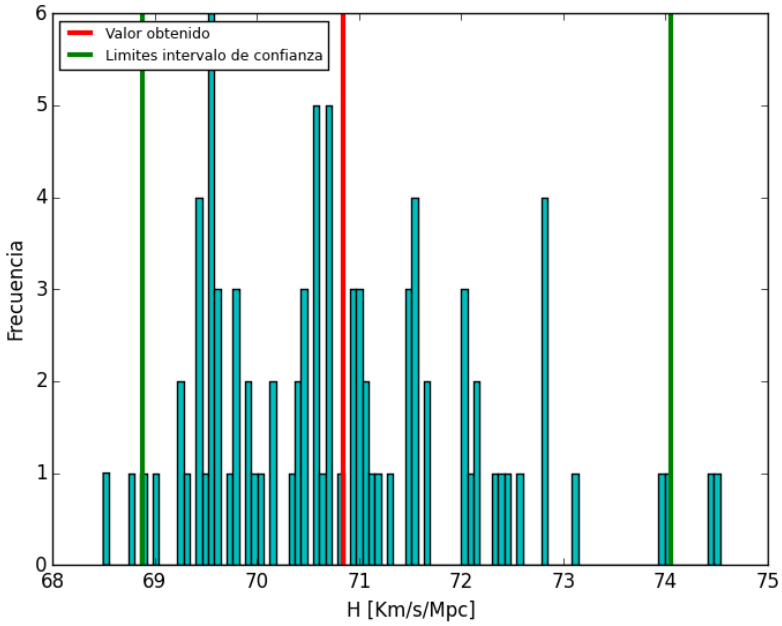

Figure 5: Histograma de H_0 hecho con datos recientes de distancia y velocidad de galaxias. Número de simulaciones: 88.

El intervalo de confianza disminuyó considerablemente respecto de la parte anterior y se puede observar en la Figura 4 que los valores H_0 se encuentran mucho menos dispersos que aquellos de la parte anterior (Figura 2). El número recomendado de simulaciones para el tamaño de esta muestra (36 galaxias) es de 88, sin embargo, al ser un número pequeño de muestras se decidió como en la parte anterior realizar 1000 simulaciones de Bootstrap de manera de obtener valores precisos para el intervalo.

Considerando, no obstante, que los datos utilizados ahora son más precisos que aquellos de la primera parte, se hace una opción viable aplicar el método de Bootstrap con menos simulaciones. Al aplicar Bootstrap para 88 simulaciones (Figura 5) se obtiene un intervalo de confianza de [68.87 ; 74.04] km/s/Mpc, el cual es bastante similar al obtenido con 1000 simulaciones y ratifica el cálculo acertado de H_0.xe

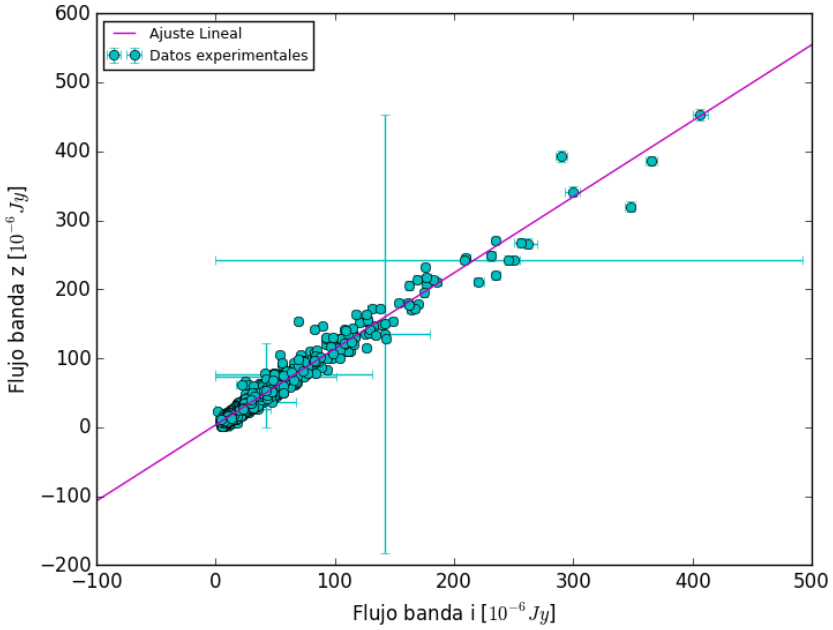

Figure 6: Datos de los flujos con sus errores y el ajuste lineal realizado mediante `np.polyfit`.

4 Flujo luminoso de cuásares

Se dispone de una sección recortada del catálogo de cuásares del Data Release 9 del Sloan Digital Sky Survey (SDSS). El objetivo de esta parte es ajustar linealmente la relación entre los flujos de las bandas i y z, datos incluidos en este catálogo. Se dispone además de los errores de medición de estos datos, razón por la cual se realizó una simulación de Montecarlo en lugar de un Bootstrapping para los intervalos de confianza.

Para realizar el ajuste lineal simplemente se utilizó la función `np.polyfit` con los datos entregados, obteniéndose como resultado los valores $m = 1.10$ y $n = 3.14$ para la pendiente y el coeficiente de posición, respectivamente. La Figura 6 ilustra los datos de que se dispone para el flujo de las bandas (con sus errores respectivos) y el ajuste lineal mediante `np.polyfit`.

Para calcular el intervalo de confianza al 95%, tanto de la pendiente como del coeficiente de posición, se realizó una simulación de Montecarlo en donde se asumió que los errores de los datos proporcionados son de naturaleza gaussiana,

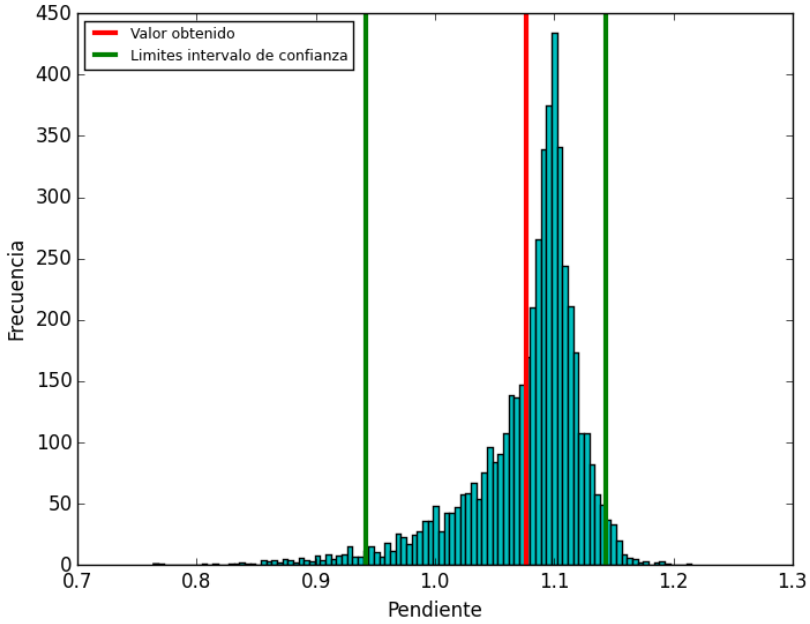

Figure 7: Histograma de valores de la pendiente con el promedio obtenido con Montecarlo marcado en rojo.

de modo que los valores calculados de i y z quedan determinados por:

$$\text{flujo calculado}_i = \text{dato}_i + \text{error}_i * r \tag{10}$$

$$\text{flujo calculado}_z = \text{dato}_z + \text{error}_z * r \tag{11}$$

Donde r, por tratarse de un error gaussiano, es una variable aleatoria de distribución normal. Después de 10000 iteraciónes de este proceso se determinó que los valores promedio de la pendiente y el coeficiente de posición son $\bar{m} = 1.07$ y $\bar{n} = 3.97$, respectivamente.

Finalmente, la simulación entrega que el intervalo de confianza de la pendiente es de [0.94 ; 1.14], mientras que para el coeficiente de posición es [2.34 ; 7.90]. Las Figuras 7 y 8 muestran los valores de m y n y su frecuencia de aparición en la simulación de Montecarlo, así como también los límites de los intervalos de confianza.

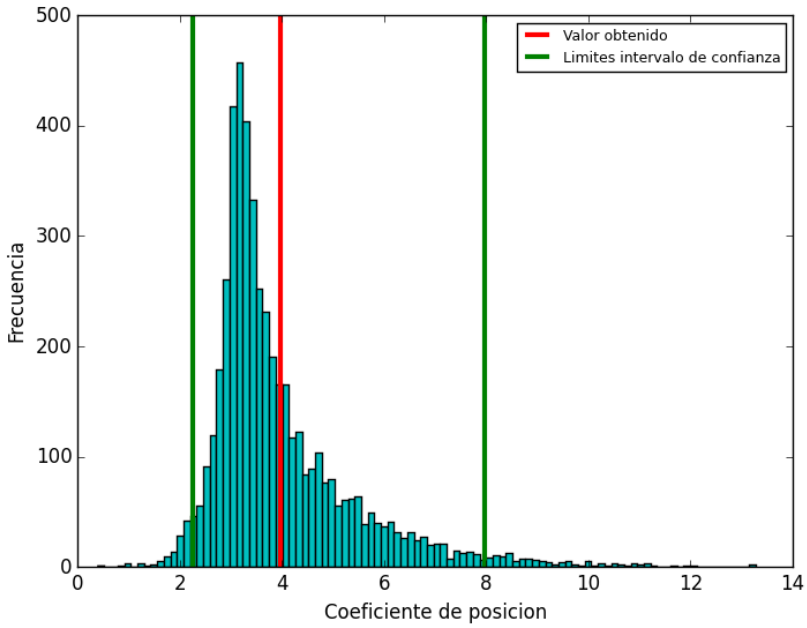

Figure 8: Histograma de valores del coeficiente de posición con el promedio obtenido con Montecarlo marcado en rojo.

Se puede apreciar que el ajuste lineal realizado es suficientemente acertado (a pesar de tratarse de un simple ajuste polinomial), lo cual quiere decir que los datos proporcionados junto a sus errores son confiables (con contadas excepciones que son fácilmente apreciables en la Figura 6) y que el modelo funciona.

Los intervalos de confianza entregan un buen rango de valores para m y n, considerando la escala en que se mueven los valores para los flujos i y z, como se puede observar en la Figura 6.

5 Conclusiones

Ambos cálculos de la constante de Hubble se condicen con lo calculado en su tiempo por Hubble y por Freedman y compañia, lo cual indica que los algoritmos aquí aplicados para el ajuste fueron implementados exitosamente y que además el modelo propuesto por Hubble es bueno. Los intervalos de confianza encontrados tienen relación directa con la precisión en el cálculo: mientras que el primero es un intervalo grande debido al error cometido por Hubble, el segundo es considerablemente más pequeño y deja a la constante en un rango aceptable hoy en día.

El ajuste obtenido para la relación entre el flujo i y z es excelente, como se puede apreciar visualmente en los gráficos anteriores. Esto indica que el modelo propuesto para esta relación es muy bueno y tiene una buena probabilidad de ser efectivamente lineal afín. El coeficiente de posición resultó muy cercano al origen, por lo que existe la posibilidad de que con una cantidad mayor de datos y con mejores mediciones este parámetro resulte ser nulo y el modelo sea completamente lineal. Los intervalos de confianza, por otra parte, son muy pequeños comparados con los valores en que se mueven los datos y entregan buenos valores para los parámetros de la recta, lo cual confirma la validez del modelo.