# Informe de Métodos Numéricos
# Tarea 9: Ajustes de Parámetros Experimentales

**Leonardo Leiva**

## 1. Resumen

## 2. Introducción: Marco Teórico

En este informe se hablará sobre algunos métodos que permiten aproximar parámetros relacionando dos o más variables variables (en la práctica experimentales). En otras palabras, buscar las constantes de un modelo matemático entre dos o más conjuntos de datos asociados. Dichas constantes pretenden ser las más cercanas al valor real, suponiendo que el modelo está correcto. Esto último se trabajará en parte en los intervalos de confianza y, probablemente, en la siguiente tarea.
Existen algunos métodos para buscar los parámetros de sets de datos. El más usual corresponde al algoritmo de Levenberg-Marquardt que corresponde a la minimización de la función $\chi^2$

### *2.1. Algoritmo de Levenberg-Marquardt (LMA)*

Corresponde a una minimización de los mínimos cuadrados de una función[1]. Se usa para optimizaciones lineales y no lineales. En general, teniendo los datos a aproximar y un modelo para aproximar, este algoritmo permite aproximar los parámetros por muy complejo que sea.
Corresponde a minimizar la siguiente función:

$$\chi^2(x_i, y_i, \beta) = \sum_{i=1}^{m} [y_i - f(x_i, \beta)]^2 \tag{1}$$

donde $y_i$ corresponde a los datos que se quieren aproximar a una función, $f$ es la función para aproximar y $\beta$ es un parámetro ajustable para lograr la aproximación. Notar que puede ser un valor o un set de valores a minimizar. La generalización a varias variables para $y_i$ y $x_i$ es simple.

El algoritmo resuelve por iteración el problema, variando cada vez el parámetro $\beta_{n+1} = \beta_n + d\beta$ hasta lograr que la suma $S(\beta)$ varíe muy poco con respecto a la iteración anterior, o que la variación $d\beta$ sea muy pequeña (según parámetros predeterminados que pueden ser fijados).

Para inicializar este algoritmo, se debe contar con el arreglo de $y_i$ a los que se quiere aproximar la función, los valores de $x_i$ asociados para evaluar en la función, y una adivinanza inicial para $\beta$. Debido a que este algoritmo encuentra el mínimo local más cercano según la adivinanza inicial que se le entregue, si se busca un mínimo global, la aproximación inicial debe ser lo suficientemente cercana.

El algoritmo retorna (entre otros resultados) los valores de los parámetros aproximados $\beta$ de la última iteración antes de cumplir las condiciones mencionadas (que la suma varíe poco o que la variación en $\beta$ llegue al límite definido).

### *2.2. Minimización de $\chi^2$*

Para el caso particular de que el parámetro relacione de forma lineal una variable con otra de la forma:

$$y = mx + n \tag{2}$$

Con m la pendiente, n el coeficiente de posición, $x$ e $y$ las variables. Por ahora no hay razón para creer que $x$ o $y$ sea la variable independiente. Para el caso particular de este problema, se tomará $n = 0$, de manera que hay que buscar las soluciones para la ecuación (2) y la siguiente:

$$x = y/m \tag{3}$$

Para solucionar se aplica la ecuación para $\chi^2$. con la nueva función $f(x_i, \beta) = y(x_i, \beta)$. Se usa $m = \beta$ y se deriva con respecto a $m$. Luego la derivada se impone igual a 0 para encontrar la minimización de la función $\chi^2$:

$$\frac{\partial(\chi^2)}{\partial\beta} = 2\sum_{i=1}^{m}[y_i - f(x_i, \beta)]\frac{\partial f(x_i, \beta)}{\partial\beta} \tag{4}$$

Usando que $y = f$, derivando $f$ con respecto a $\beta$ e imponiendo la derivada igual a cero:

$$\sum_{i=1}^{m}[y_i x_i - \beta x_i^2] = 0 \tag{5}$$

Se puede despejar $\beta$ para obtener una relación explicita, lo que corresponde a una regresión lineal, pero se prefirió usar una optimización de scipy para buscar los ceros de (5). Como el ajuste es lineal, la ecuación tiene un único 0, por lo tanto, un único mínimo. Se descarta que corresponda a un máximo, porque la función $\chi^2$ está hecha para que de valores positivos. Para este caso, además, se considera (como se dijo antes) que tanto $x$ como $y$ pueden ser la variable dependiente, por lo que se aproxima el valor de $\beta$ como:

$$\beta = \frac{\beta_1\beta_2 + 1 - \sqrt{(1+\beta_1^2)(1+\beta 2^2)}}{\beta_1 + \beta_2} \tag{6}$$

Donde $\beta_1$ es $\beta$ calculado para (2) y $\beta_2$ es $\beta$ calculado para (3).
Para el caso de que los coeficientes de posición no sean $0$, vale decir, tener la ecuación (2) como está y la (3) de la forma $y = x/m - n/m$. En este caso se usará la pendiente de la

fórmula (6) y como para definir una recta se necesita la pendiente y un punto cualquiera, ese punto será el que se obtiene de intersectar las rectas que se forman con $\beta_1$ y $\beta_2$. Con un poco de álgebra se puede mostrar que la recta que se obtiene es:

$$y = \beta x + y_0 - a x_0 \tag{7}$$

Donde $(x_0, y_0)$ corresponde a la intersección de las rectas anteriormente descritas.

### *2.3. Intervalos de Confianza*

En estadísitca, corresponde a un par de valores para una muestra entre los cuales un valor desconocido tiene una determinada probabilidad de acierto, en otra. Se tiene que para un intervalo más grande la probabilidad de acierto es mayor, mientras que al disminuir el intervalo se tiene una estimación más precisa, pero aumenta la posibilidad de fallar. Usualmente la distribución de estos valores desconocidos en el intervalo de confianza es Normal, pero también se tiene una gran cantidad de casos en los que esto no es así.

### *2.4. Método de Bootstrap*

Este método permite obtener una aproximación en donde se puede encontrar el intervalo de confianza al $\alpha\,\%$, es decir, que el $\alpha\,\%$ de los errores estén en el intervalo. Corresponde a un método de la estadística Bayesiana ya que se basa en usar los datos que se tienen para obtener información sobre los errores.
Si se tiene una muestra de $N$ datos experimentales y se busca obtener un parámetro de esta muestra, se escogen, al azar y con repetición $N$ valores de la muestra y se calcula el parámetro buscado para esta nueva muestra. Cabe destacar que es necesaria la repetición de valores, porque si no se obtendría siempre la muestra original. Este proceso se repite suficientes veces para lograr convergencia. Se puede demostrar que para este algoritmo basta con usar $N_{iteración} = N log_{10}(N)^2$, aunque para muestras pequeñas resulta necesario usar un $N_{iteración}$ más grande.
Los $N_{iteración}$ valores de parámetro buscados obtenidos se ordenan de menor a mayor y el intervalo de confianza se obtiene de este arreglo ordenado: el valor mínimo será el que se encuentre en la posición que represente el $(100 - \alpha)/2\,\%$ de la muestra y el máximo será $100 - \alpha/2\,\%$.

### *2.5. Método de Monte Carlo*

La idea es la misma que para el método anterior pero la forma de proceder es diferente. Para este caso se usarán los errores asociados a cada dato experimental (que pueden o no, ser los mismos para todos). Se usa un $N_{iteración}$ suficientemente grande para lograr convergencia y se realiza $N_{iteración}$ iteraciones, donde se obtiene una muestra nueva a partir de la muestra original de la forma:

$$x_{muestranueva} = x_{muestraoriginal} + dx * r \tag{8}$$

Donde $x_{muestra}$ son todos los valores de la muestra, $dx$ es el error para cada valor y $r$ es una variable aleatoria que distribuye como una normal $N(0,1)$ (centrada en 0 con varianza 1). Notar que es una ecuación vectorial, vale decir, a priori se tendrían valores diferentes a la muestra original para cada caso. Con esta nueva muestra se calcula el parámetro que se quiere determinar.
El intervalo de confianza se obtiene de la misma manera que el el método de bootstrap en base a un porcentaje de confianza.

## 3. Problemas Presentados y Método de Resolución

Con las herramientas destritas anteriormente se resuelven tres problemas: Aproximación de la Constante de Hubble con los datos originales medidos por Edwin Hubble; luego se aproximará la misma constante con datos más actualizados y que permiten una aproximación más realista; por último, se realizará una aproximación lineal de dos variables de flujo a partir de datos experimentales del catálogo de quásares.

### *3.1. Pregunta 1: Aproximación Inicial de Hubble*

Se realiza una aproximación de la constante de Hubble en base a los datos experimentales medidos en 1929 por Edwin Hubble. Dichos datos se encuentran en el repositorio "data/hubble_original.dat". Se encuentra ahí la comparación entre la velocidad de recesión de las Nebulosas (aún no estaba muy clara la consepcion de "galaxias") con la distancia entre estas con respecto a la Tierra. Se midió mediante el método de las Cefeidas, que corresponde a la medición luminosidad variable de estrellas. Dicha variación se relaciona fuertemente con el periodo. Dicha relación estaba recién calibrada.
Se presume que la relación entre velocidad de recesión y distancia es lineal de la forma:

$$v = H_0 * d \tag{9}$$

Donde $v$ es la velocidad de recesión, $d$ es distancia y $H_0$ es la constante de Hubble. El valor teórico $H_0 = 70$ [(km/s)/Mpc] $+2{,}4/-3{,}2$.
Con los datos extraidos desde el archivo mencionado se realiza, inicialmente, una aproximación de la constante de Hubble por medio del Algoritmo de Levenberg-Marquardt usando la rutina 'scipy.leastsq'. Luego se realiza minimizando manualmente la función $\chi^2$ de la ecuación (5). Se calcula el valor de $H_0$ mediante la bisección de la recta de considerar (9) y $d = v/H_0$.
Por último, los intervalos de confianza se calcularon mediante bootstrap debido a que los errores no eran dados. Se buscan intervalos de confianza al $95\,\%$. El número de simulaciones que se usa es $N^2$ dada la pequeña cantidad de muestras. El algoritmo usado esta en "parte1.py".

### *3.2. Pregunta 2: Aproximación Actualizada de la Constante de Hubble*

Para este caso se utiliza el mismo modelo para relacionar la velocidad de recesión con la distancia entre las nebulosas y la Tierra, pero para los datos utilizados en este caso se tiene una correción de la calibración de la relación periodo-luminosidad (entre otros errores corregidos). Los datos en este caso están en "data/SNIa.dat". El método de medición utiliza supernovas tipo 1 que, además, permite medir distancias mucho mayores.
Se realiza el mismo procedimiento que para la parte 1 con los datos nuevos: se usa la minimización de $\chi^2$, la bisección entre las rectas y los intervalos de confianza por bootstrap. El algoritmo está en "parte2.py".

### *3.3. Pregunta 3*

Para esta parte se busca la línea recta que mejor modela la relación entre el flujo de la banda i con la banda z de la sección recortada del catálogo de quásares del Data Release 9 del Sloan Digital Sky Survey (SDSS). Se buscan los intervalos de confianza al $95\,\%$. Se busca una relación de la forma:

$$i = mz + n \tag{10}$$

Con $m$ la pendiente y $n$ el coeficiente de posición Considerando que tampoco hay razón para asumir que $i$ depende de $z$ o visceversa (igual que en las partes anteriores), se realiza tanto el ajuste para $i$ dependiendo de $z$ como al revés. Para ello se usó el algoritmo de "polyfit"de numpy, el cual entrega la pendiente y el coeficiente de la recta que se aproxima mejor a los datos experimentales que se le entregan. Con estos coeficientes se usó la ecuación (6) para la pendiente y el coeficiente de posición de la ecuación (7). Se hizo una modificación al método que calcula la bisección entre las dos rectas que se usó para la parte 1 y 2.
Dado que los errores para el flujo de ambas bandas son datos, se usa el algoritmo de Monte Carlo para obtener los intervalos de confianza. El algoritmo está en "parte3.py".

## 4. Resultados y Análisis

### *4.1. La constante de Hubble: Primer Intento*

Para una estimación inicial del valor de la constante de Hubble se obtuvo $H_0 = 467{,}219$ [Km/s /Mpc] vía bisección de la relación directa y la inversa. La recta de color verde representa dicha pendiente. Los valores de la relación directa (en rojo) y la inversa (en celeste corresponde a $H_0 = 424{,}973$ [Km/s /Mpc] y $H_0 = 520{,}343$ [Km/s /Mpc] respectivamente. Los datos experimentales a través de los que se obtuvo dichas rectas corresponde a los puntos azules.
Se observa que la aproximación no es muy buena dada la gran dispersión de los datos respecto a la recta del modelo (ver igura 1). En este intervalo se puede dudar, incluso, de que el modelo adoptado esté bien, pero es necesario notar que la distancia está en $[Mpc]$

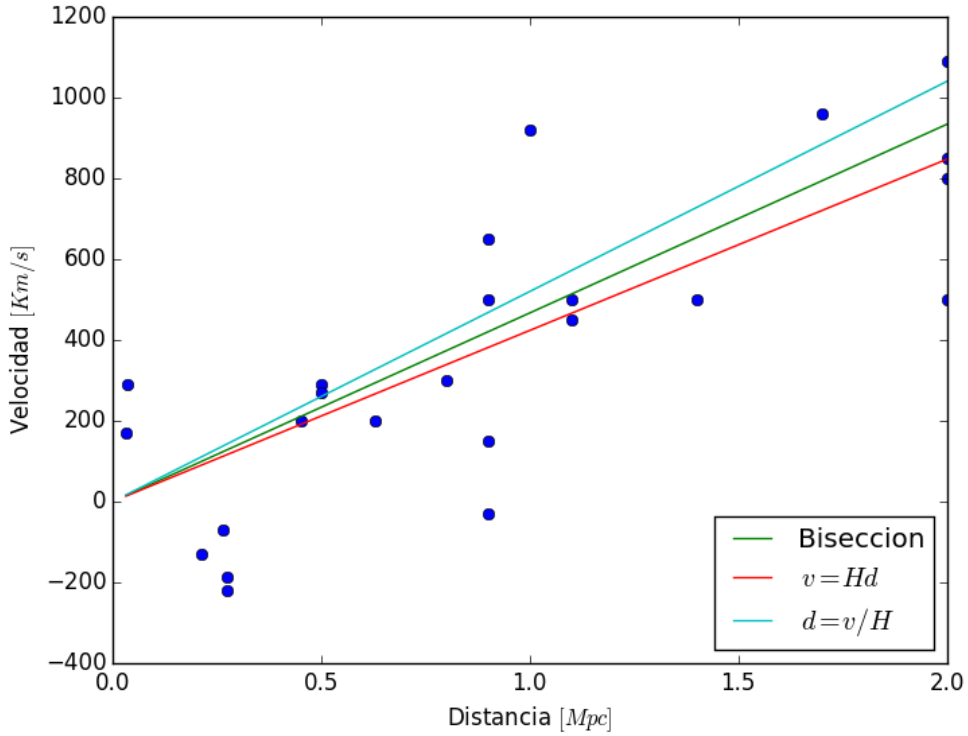


Figura 1: Relación de la Velocidad de Recesión y la distancia de las estrellas a la Tierra para los datos obtenidos por Hubble mediante la medición de Ceferidas. Los puntos azules son los datos experimentales. Las rectas en rojo y celeste son las aproximaciones iniciales de la ecuación (9) y la relación inversa. La recta en verde es una mejor aproximación considerando las dos anteriores

($1Mpc = 3,0857 * 10^{16}$). Considerando esto, se puede observar que un modelo lineal es bastante bueno y que las posibles variaciones son debido a errores aleatorios y de medición, más que error en el modelo adoptado. A pesar de esto, al usar esta escala se puede notar que hay bastante dispersión de los datos con respecto a la recta que los modela.
El gráfico de los intervalos de confianza representa también el gran margen de error de los datos tomados: los $H_0$ obtenidos por bootstrap distribuyen parecido a una Gaussiana y se observa que el ancho (varianza) es más de $50[Km/s/Mpc]$. Se requerirían más datos para obtener un resultado con menos ruido. El intervalo de confianza al $95\,\%$ es $[423{,}937, 520{,}343]$, muy parecido a las pendientes obtenidas para los métodos sin usar bisección. Es un rango bastante amplio: Podemos estar bastante seguros de que una medición se encontrará en ese intervalo, pero la precisión es bastante mala.
Además, y no menos importante, se observa discordancia con el valor teórico de $H_0$. Es, casi 7 veces más. Se puede comprobar que, efectivamente habían errores en las mediciones que no se pueden atribuir a errores aleatorios. Como se mencionó anteriormente, la calibración período–luminosidad estaba equivocada, aunque hay otros errores que también ayudaron, fue el que más influyó.

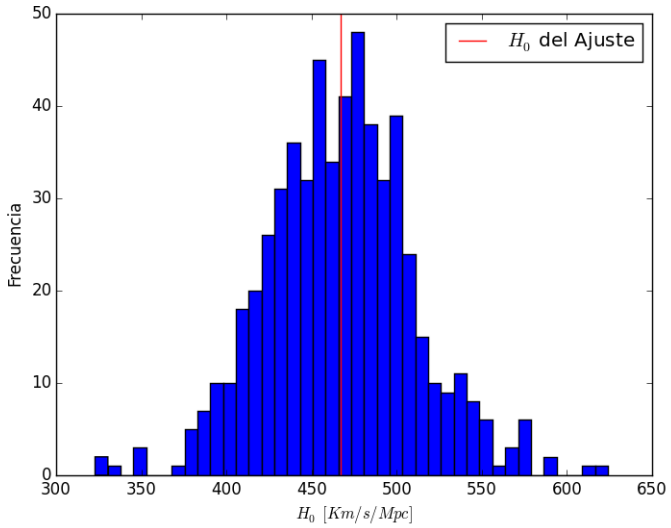


Figura 2: Histograma que representa los intervalos de confianza de $H_0$ para la primera estimación (datos originales). Se usó método de Bootstrap.

### *4.2. La constante de Hubble: Intento Actualizado*

Para esta aproximación se observa una considerable mejora en la cercanía entre los valores experimentales y el modelo lineal adoptado. De esta forma, incluso las rectas para antes de la bisección modela muy bien el comportamiento (ver figura 3). Casi no se aprecia la diferencia. En la figura (4) se observa un acercamiento que permite ver la diferencia para la primera sección de la recta. El modelo se muestra mucho más concordante con los datos obtenidos. Aún así se observa que los datos más grandes se alejan un poco del modelo, pero también hay que considerar que no hay muchos datos en esas secciones.
Cabe destacar también la mejora en los rangos de distancias obtenidas: Para la parte anterior se tenían, a lo mas, datos para $2Mpc$, mientras que para esta se tienen hasta casi $500Mpc$. Esto también mejora y afianza el modelo obtenido.
Los intervalos de confianza refuerzan eso. Para la figura 5 se observa que la forma es similar a la figura 2, pero el ancho de la varianza es mucho menor: aproximadamente $2Mpc$. El valor obtenido para este modelo es de $H_0 = 70{,}84[Km/s/Mpc]$, que está muy cerca del valor teórico. También el intervalo de confianza se ajusta al intervalo de confianza obtenido. Al $95\,\%$ se obtiene $[68{,}67, 73{,}68]$.
Los valores de las pendientes para la recta que representa la solución directa a la ecuación (9) es $H_0 = 70{,}66[Km/s/Mpc]$, mientras que para la relación inversa $H_0 = 71{,}01[Km/s/Mpc]$. Esto indica que los datos eran mucho más simétricos que para la parte anterior, sumado con que los errores son más pequeños.

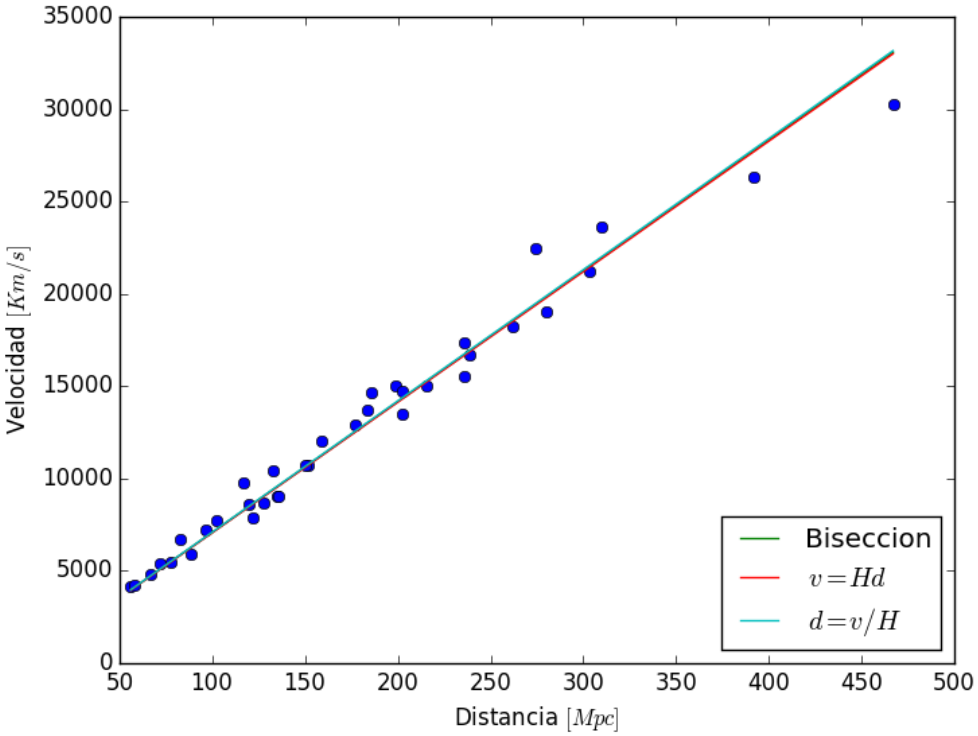


Figura 3: Relación de la Velocidad de Recesión y la distancia de las estrellas a la Tierra para mediciones modernas. Los puntos azules son los datos experimentales. Las rectas en rojo y celeste son las aproximaciones iniciales de la ecuación (9) y la relación inversa. La recta en verde es una mejor aproximación considerando las dos anteriores

### *4.3. Los Flujos de Banda*

Modelando la relación de flujo de forma lineal, se obtuvo que la pendiente para la relación (10) es $m = 1{,}123$ y el coeficiente de posición $n = 2{,}501$. La variación entre el cálculo directo de la relación (10) con respecto a la relación inversa es pequeña y se observa en la figura (6). Se puede observar también que la relación lineal es bastante razonable según los datos obtenidos.

Los intervalos de confianza obtenidos por Monte Carlo muestran que la toma de datos es bastante sólida en el sentido que alejarse de los valores obtenidos para la pendiente y el coeficiente de posición es muy poco probable. De hecho, una pequeña variación hace que la probabilidad disminuya considerablemente. La desviación estándar que se puede interpretar de los histogramas en la figura (7) y (8) es bastante pequeña. De esto se obtiene que, obtener un intervalo de confianza para porcentajes altos varía bastante entre ellos.

Para el 95 % se tiene $[1{,}033, 1{,}226]$ para la pendiente y $[-0{,}55, 4{,}989]$ para el coeficiente de posición. se observa mayor dispersión para el coeficiente de posición, pero no tiene demasiada importancia en el modelo: la pendiente domina más la relación de dependencia.

La gran cantidad de datos permite eliminar el ruido en las simulaciones para los intervalos

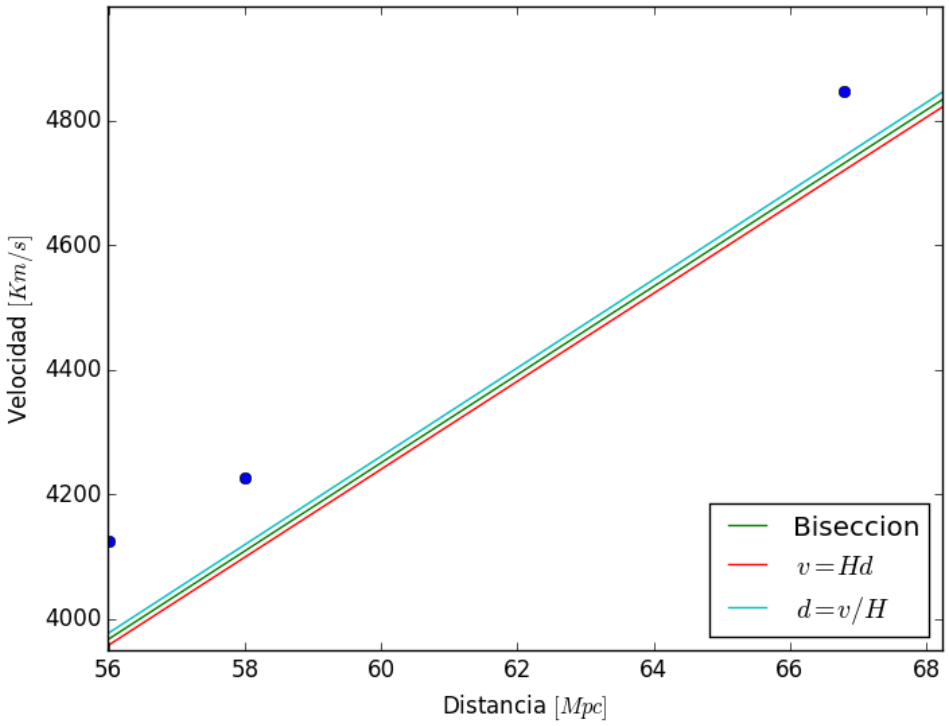


Figura 4: Acercamiento de la figura 3 para observar la diferencia entre las rectas modelo obtenidas.

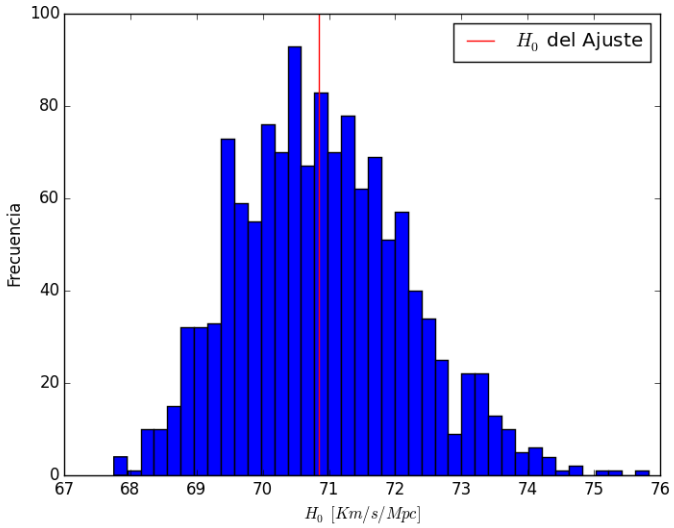


Figura 5: Histograma que representa los intervalos de confianza de $H_0$ para la estimaciones modernas. Se usó método de Bootstrap.

de confianza siempre que se considere un número razonable de iteraciones.

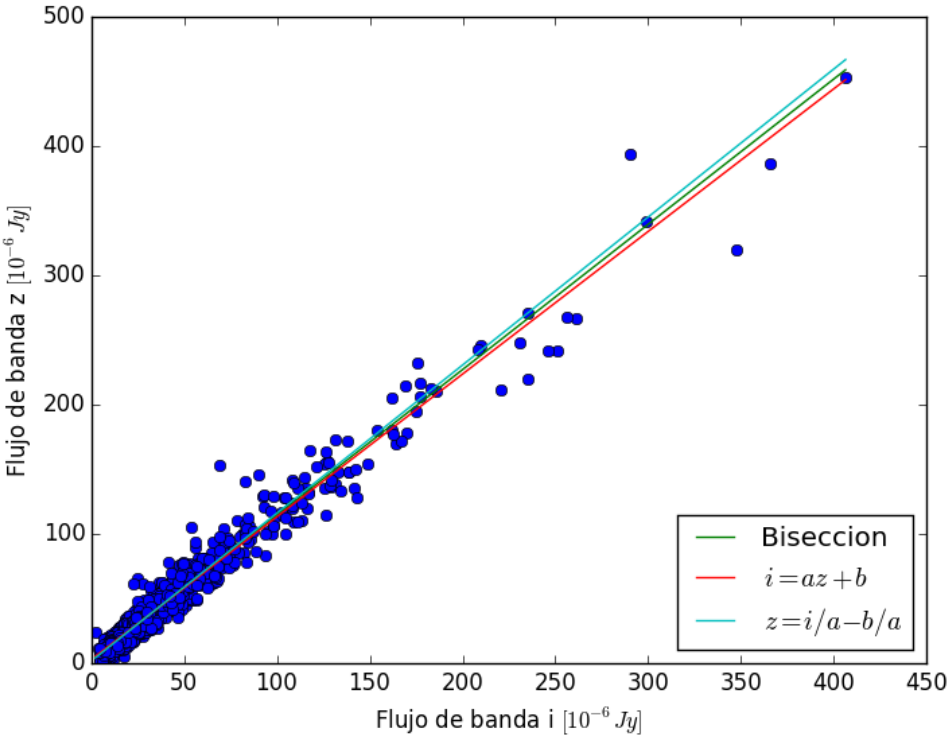


Figura 6: Acercamiento de la figura 3 para observar la diferencia entre las rectas modelo obtenidas.

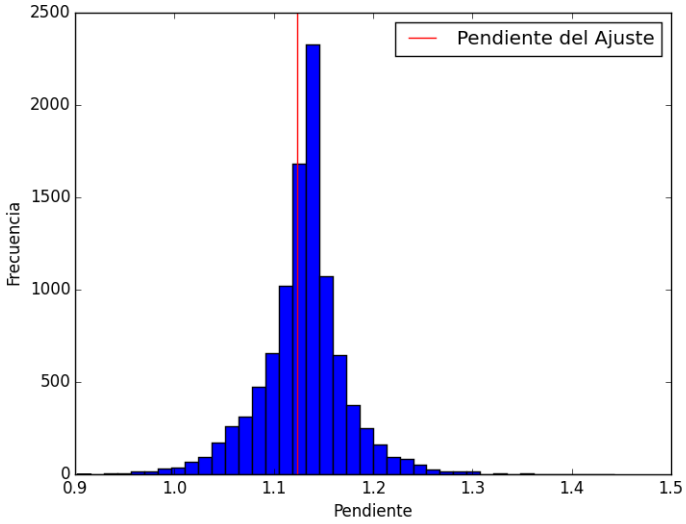


Figura 7: Histograma que representa los intervalos de confianza de $H_0$ para la estimaciones modernas. Se usó método de Bootstrap.

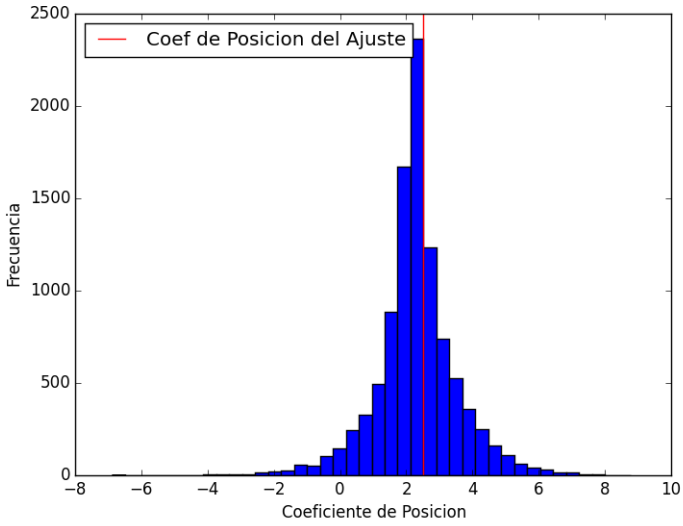


Figura 8: Histograma que representa los intervalos de confianza de $H_0$ para la estimaciones modernas. Se usó método de Bootstrap.

## 5. Conclusiones

Se destaca que los métodos de ajuste de parámetros son muy útiles en el sentido de que permiten hacerse una idea de como se relacionan dos variables. Inicialmente se deben observar los datos y pensar que modelo podría tener la relación. Luego se puede usar la minimización de $\chi^2$ para tener una idea del parámetro que mejor se ajusta y, finalmente, ver que relación teórica se obtiene. Al comparar con los datos experimentales se puede concluir si el modelo está bien o no. También se puede tener que el modelo se parezca, pero este incompleto por falta de datos o porque a grandes escalas se comporta de formas complejas. Esto solo se puede resolver por medio de la toma de datos a rangos más grandes.
Esto último ayuda mucho a la hora de analizar muestras con mucho ruido como las partes 1 y 2: Inicialmente se tenían datos que parecen no tener mucha relación a la escala que se está observando, pero al obtener un muestreo más amplio se concluye claramente que el comportamiento era como se suponía.
Los intervalos de confianza pueden ser una buena herramienta para determinar la factibilidad del modelo adoptado. En el primer caso se podía ver que, a pesar de que el modelo lineal parecía aplicable, debía haber un error en el procedimiento por la gran diferencia de los posibles valores obtenidos a partir de Bootstrap, reflejado en un ancho intervalo de confianza. Para el caso de la parte 2, los intervalos eran mucho más acotados y permitían concluir que el modelo era bueno. Esto significa una mejora considerable entre ambos.
También se destaca la importancia de que, en caso de no tener claro que variable depende de cual, es necesario establecer una relación que establezca una simetría entre ambas variables, dado que los métodos de optimización no lo son. Las diferencias no son despreciables, por lo que, adoptar un modelo no simétrico podría llevar a errores importantes en el resultado.

En cuanto a los problemas planteados, se sugiere realizar algún tipo de modelamiento diferente y más desafiante (no lineal, por ejemplo), sobre todo en el caso de ser tres problemas. El problema 3 no tenía mucho contexto y no resultaba muy desafiante al ser muy similar a los anteriores. A pesar de que existía diferencia para el método utilizado en el cálculo de los intervalos de confianza, no era muy diferente a lo que ya se había hecho.

## Referencias

[1] Algoritmo de Levenberg-Marquardt https://en.wikipedia.org/wiki/Levenberg-Marquardt_algorithm
[2] Documentación de scipy del Algoritmo de Levenberg-Marquardt http://docs.scipy.org/doc/scipy/reference/tutorial/optimize.html