Manejo de datos

José Ignacio Cabrera Martínez

20 de febrero de 2018

0.1. Los datos

Supongamos que después de realizar un experimento obtenemos la siguiente tabla de datos:

X ± 1 A	Y ± 0.8 B
10	2.1
15	2.5
20	6.1
25	5.7
30	7.7
35	7.1
40	9.8
45	11.9
50	12.3
55	14.6
60	14.0
65	17.1
70	18.6
75	19.9
80	20.0
85	24.0
90	24.3
95	25.0
100	26.7
105	28.8
110	28.1
115	30.2
120	32.9
125	33.4
130	33.9
135	36.1
140	38.1
145	37.4
150	40.0

Tabla 1: Datos obtenidos del experimento, X es la variable independiente y Y la variable dependiente, A es la unidad de la variable X y B es la unidad de la variable Y

Para comenzar a analizar nuestros datos, primero necesitamos poder visualizarlos de una manera más simple que como se pueden ver en la tabla 1, pues aunque podemos ver que crecen los valores de ambas variables, no es muy cómodo ver el comportamiento en forma de una tabla, es mas sencillo darse cuenta del comportamiento en una imagen. Por esta razón es mejor hacer una gráfica.

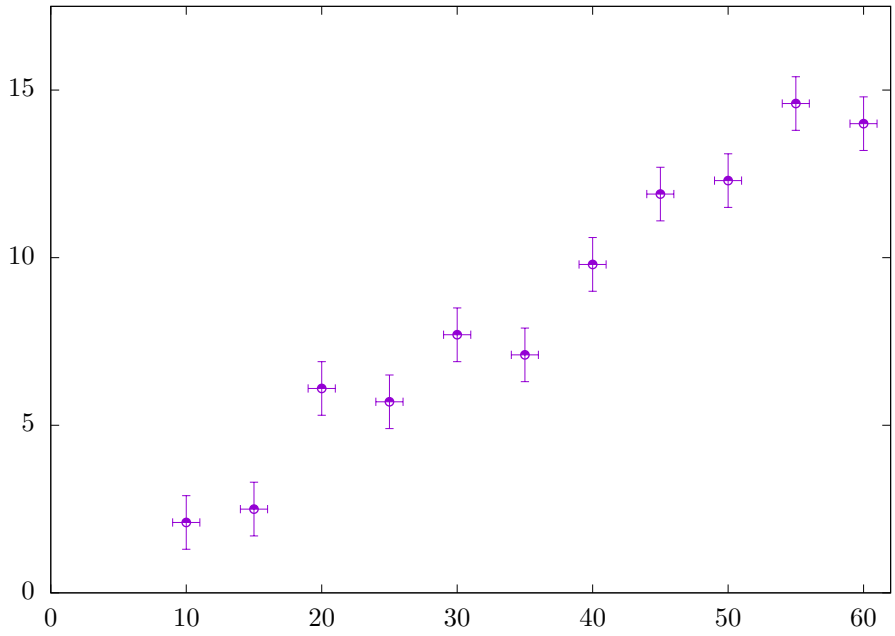

Figura 1: Gráfica con el intervalo de longitud de los ejes demasiado pequeño, por eso muchos puntos de nuestros datos quedaron fuera de la gráfica. Los ejes no están etiquetados, por esa razón no se sabe que hay en cada eje ni las unidades en que se medió, está es una mala gráfica.

0.2. Gráficas

Ahora la cuestión es como hacer una buena gráfica, lo primero que necesitamos definir es quienes serán los ejes de nuestra gráfica. Definiremos a nuestra variable independiente X como la abscisa y a la variable dependiente Y como la ordenada. Ahora necesitamos definir correctamente la longitud del intervalo de nuestros ejes, pues si la longitud del intervalo es muy pequeña parte de la información quedará fuera de la gráfica, como se puede notar en la figura 1, además en esta gráfica no sabemos que hay en cada eje, por lo cual no es muy útil, lo único bueno en esta gráfica es como está representada la incertidumbre de cada dato, como un segmento de línea que recorre todo el intervalo de X y de Y para cada dato.

En la figura 2 la longitud del intervalo de los ejes es demasiado grande y es complicado leer la información de nuestra gráfica, por otra parte en esta gráfica aunque los ejes están etiquetados con el nombre de cada variable en cada eje, no sabemos las unidades de cada variable, por esta razón la información que nos ofrece esta gráfica es incompleta.

En la figura 3 tenemos ya definida nuestra gráfica con una escala adecuada para ambos ejes, además los ejes están correctamente etiquetados y muestran las unidades de cada variable entre paréntesis, así es como se debe de hacer una gráfica.

En la gráfica 3 podemos ver como según incrementa la variable independiente (X), la variable dependiente (Y) también se incrementa, eso hace pensar que ambas variables podrían estar correlacionadas, pero ¿qué tan correlacionadas pueden estar ambas variables? Para responder esta pregunta de una manera objetiva necesitamos definir un criterio que nos ayude a responder a esta pregunta. Para definir ese criterio recurriremos al concepto de varianza, desviación estándar y de

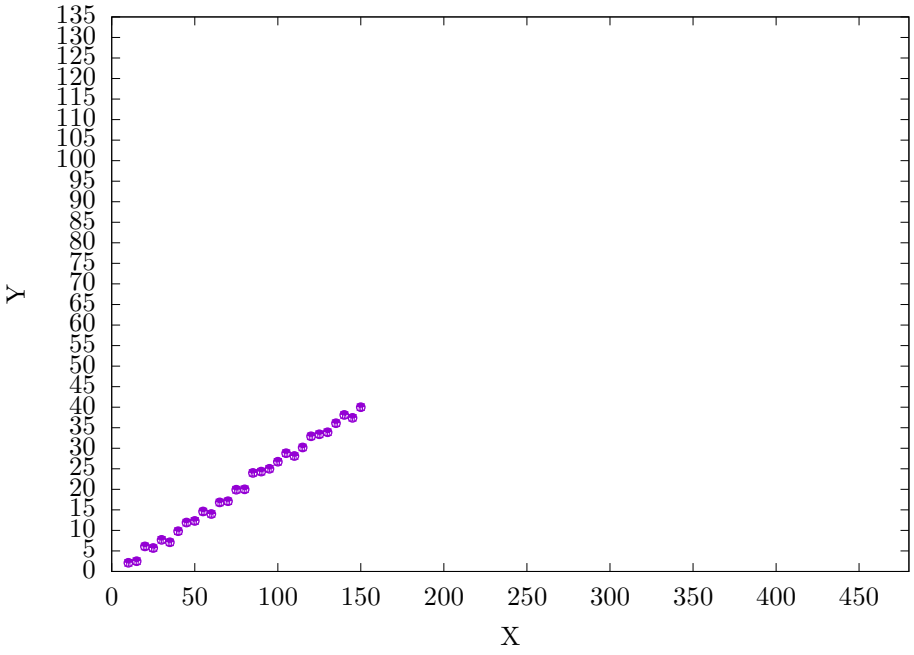

Figura 2: Gráfica con la escala de los ejes demasiado grande,por eso todos los datos están comprimidos en una pequeña porción de la gráfica, presentar de esta manera la información hace dificil la lectura de los datos.Los ejes están etiquetados pero no sabemos las unidades de cada eje, por tanto la información es incompleta, aunque esta gráfica es mejor que la gráfica 1 tampoco es una buena gráfica.

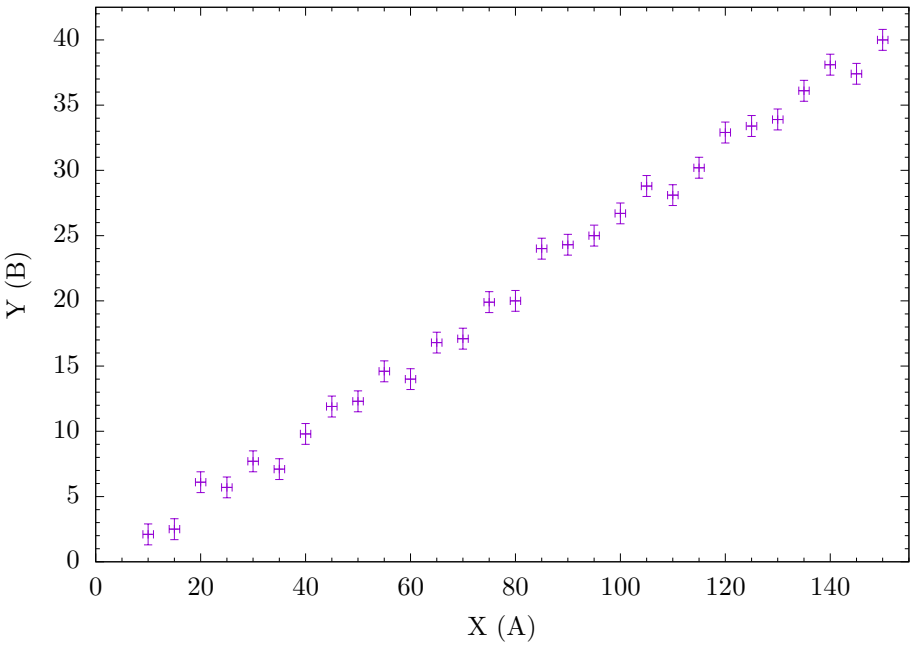

Figura 3: En esta gráfica cada eje tiene una escala adecuada para que se pueda leer la información sin problemas. Cada eje esta etiquetado con el nombre de cada variable y además se presneta entre parentesis las unidades en que se midieron las variables. Este es un buen ejemplo de como se debe de presentar una gráfica.

covarianza, pues estas dos cantidades nos ayudaran a definir un criterio de correlación que nos será útil.

0.3. Correlación entre las variables

La desviación estándar está definida como:

$$\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{1}$$

para una distribución de valores de X, y $\overline{X}$ es el promedio de los valores de X; mientras que la varianza es simplemente σ^2. Por otra parte la covarianza para un conjunto de datos X y Y se puede calcular como:

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) \tag{2}$$

esta cantidad es una especie de varianza cruzada de los datos de X y Y, aprovechando esta propiedad podemos analizar que ocurre si por ejemplo la variable Y fuera una función lineal de X, es decir que:

$$Y = mX + b \tag{3}$$

entonces la covarianza sería:

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(mX_i + b - \overline{Y}) \tag{4}$$

con

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n}Y_i = \frac{1}{n}\sum_{i=1}^{n}(mX_i + b) \tag{5}$$

desarrollando lo anterior se tiene que:

$$\overline{Y} = \frac{1}{n}(\sum_{i=1}^{n}mX_i + \sum_{i=1}^{n}b) = m\frac{1}{n}\sum_{i=1}^{n}X_i + \frac{n}{n}b \tag{6}$$

$$\overline{Y} = m\overline{X} + b \tag{7}$$

Sustituyendo la ecuación 7 en la ecuación 4 entonces tenemos que:

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(mX_i + b - m\overline{X} - b) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(mX_i - m\overline{X}) \tag{8}$$

$$\sigma_{xy} = \frac{m}{n}\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X}) = m\sigma_x^2 \tag{9}$$

Es decir que en este caso la covarianza sería la varianza de X multiplicada por la constante m de la relación lineal . ahora desarrollando como seria la desviación estándar de Y bajo esta suposición tendríamos que:

$$\sigma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(mX_i + b - m\overline{X} - b)^2} = \sqrt{\frac{m^2}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{10}$$

resultando entonces

$$\sigma_y = \mid m \mid \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} = \mid m \mid \sigma_x \tag{11}$$

Con lo anterior podemos usar los resultados de la ecuación 9 y la ecuación 11 para definir un parámetro que nos ayude a medir la correlación entre las variables X y Y como

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{12}$$

Entonces si Y es una función lineal de X, utilizando las ecuaciones 9 y 11 en la ecuación 12 se obtiene:

$$\rho = \frac{m}{\mid m \mid} = \pm 1 \tag{13}$$

A ρ se le suele llamar el coeficiente de correlación de Pearson, y sus posibles valores están dentro del intervalo [-1,1]. $\rho = 1$ significa que las variables tienen una correlación lineal perfecta, $\rho \leqslant 1$ significa que se tiene una buena correlación entre las variables, hasta un valor ~ 0.7 es aceptable. Si ρ es 0.5 o menor, es una mala correlación, si $\rho \sim 0$ podemos decir que no hay correlación lineal entre las variables; pero que no haya correlación lineal no significa que no haya otros tipos de correlación entre las variables. Si $\rho \geq -1$ se tiene una anticorrelación lineal es decir que según crece una variable la otra decrece proporcionalmente, y finalmente si ρ está entre 0 y -5 es una mala correlación.

Ahora si aplicamos el coeficiente de correlación a nuestro conjunto de datos obtenidos de por el experimento (ver tabla 1), obtendremos que $\rho = 0.997$, esto significa que las variables X y Y están altamente correlacionadas linealmente.

Si queremos describir el comportamiento de los datos de nuestro experimento mediante un modelo matemático, el hecho de que ρ es tan cercano a 1 sugiere que el modelo matemático podría ser de la forma (aunque no necesariamente) $Y = mX + b$. Si el modelo es una línea recta, la pregunta es ¿Cómo se pueden encontrar los parámetros m y b de la recta? Para responder a esa pregunta, se esperaría que dichos parámetros se pudieran determinar a partir de los datos.

0.4. El método de mínimos cuadrados

Supongamos que la mejor línea recta que describe el comportamiento de los datos del experimento es una especie de promedio, los puntos mas alejados de la recta serían aquellos con mayor dispersión alrededor del promedio. Lo que desearíamos es que la dispersión de los puntos sea la más pequeña posible, para esto podemos definir una cantidad llamada *dif* que sea la suma de las diferencias entre los valores de las ordenadas y los valores teóricos que predice el modelo matemático, es decir:

$$dif = \sum_{i=1}^{n} (Y_i - f(X_i)) \tag{14}$$

Si tomamos esta definición, tendremos un problema importante y es que *dif* podría valer 0 sin que la dispersión de los datos sea pequeña; puede haber datos muy dispersos arriba y abajo del modelo matemático y la suma de diferencias podría cancelarse o dar valores muy cercanos a cero, por este motivo esta no es una buena cantidad que tratemos de minimizar, entonces se tomará mejor la cantidad *dif* definida como:

$$dif = \sum_{i=1}^{n} (Y_i - f(X_i))^2 \tag{15}$$

El modelo teórico *f(X)* puede ser tan simple o tan complicado como uno desee, pero en general siempre es mejor tratar de utilizar modelos lo mas sencillos posibles, como una línea recta, si el modelo mas simple no funciona, entonces se puede tratar con modelos mas complicados. Pero en general entre menos términos tenga un modelo, este se considera mejor, ya que es mas fácil interpretar dentro del experimento que representan pocos términos a tener que identificar que pueden representar muchos parámetros de un modelo.

Para el caso de nuestros datos consideraremos un modelo de línea recta, es decir $f(X) = mX+b$, entonces la *dif* toma la forma:

$$dif = \sum_{i=1}^{n} (Y_i - mX_i - b)^2 \tag{16}$$

Ahora lo que se quiere hacer es minimizar *dif*, pero ¿respecto a que cantidades? Las cantidades que nos interesa determinar son m y b, entonces diremos que $dif = dif(m, b)$ y respecto a estos parámetros minimizaremos a dif. La manera de minimizar será derivando respecto a m y b e igualando a cero esas derivadas, *i. e.*

$$\frac{\partial dif}{\partial m} = \sum_{i=1}^{n} \frac{\partial (Y_i - mX_i - b)^2}{\partial m} = 0 \tag{17}$$

$$\frac{\partial dif}{\partial b} = \sum_{i=1}^{n} \frac{\partial (Y_i - mX_i - b)^2}{\partial b} = 0 \tag{18}$$

Desarrollando las ecuaciones 17 y 18 se tiene que:

$$\frac{\partial dif}{\partial m} = \sum_{i=1}^{n} 2(Y_i - mX_i - b)(-X_i) = -2\sum_{i=1}^{n}(Y_i - mX_i - b)(X_i) = 0 \tag{19}$$

$$\frac{\partial dif}{\partial b} = \sum_{i=1}^{n} 2(Y_i - mX_i - b)(-1) = -2\sum_{i=1}^{n}(Y_i - mX_i - b) = 0 \tag{20}$$

como las ecuaciones 19 y 20 son iguales a 0, se púede eliminar el factor -2. Si se desarrollan ambas ecuaciones entonces tendremos como resultado el sistema de ecuaciones:

$$\frac{\partial dif}{\partial m} = \sum_{i=1}^{n} X_i Y_i - m\sum_{i=1}^{n}(X_i)^2 - b\sum_{i=1}^{n} X_i = 0 \tag{21}$$

$$\frac{\partial dif}{\partial b} = \sum_{i=1}^{n} Y_i - m\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} b = 0 \tag{22}$$

resecribiendo las ecucaciones 21 y 22 tenemos:

$$m\sum_{i=1}^{n}(X_i)^2 + b\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} X_i Y_i \tag{23}$$

$$m\sum_{i=1}^{n} X_i + nb = \sum_{i=1}^{n} Y_i \tag{24}$$

Las ecuaciones 23 y 24 son dos ecuaciones lineales con dos incógnitas (m y b), en principio el sistema tiene una solución única, la cual podemos encontrar de varias maneras, si utilizamos el método de los determinantes, tendremos que el determinante principal es:

$$\triangle = \begin{vmatrix} \sum_{i=1}^{n}(X_i)^2 & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & n \end{vmatrix} = n\sum_{i=1}^{n}(X_i)^2 - \left(\sum_{i=1}^{n} X_i\right)^2 \tag{25}$$

es necesario recordar que para que exista una solución $\triangle \neq 0$ pues podría llegar a darse este caso $\triangle = 0$ aunque es poco probable. El valor de m es:

$$m = \frac{\begin{vmatrix} \sum_{i=1}^{n} X_i Y_i & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} Y_i & n \end{vmatrix}}{n\sum_{i=1}^{n}(X_i)^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n}(X_i)^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \tag{26}$$

y b es :

$$b = \frac{\begin{vmatrix} \sum_{i=1}^{n}(X_i)^2 & \sum_{i=1}^{n} X_i Y_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} Y_i \end{vmatrix}}{n\sum_{i=1}^{n}(X_i)^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = \frac{\sum_{i=1}^{n}(X_i)^2 \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} X_i Y_i \sum_{i=1}^{n} X_i}{n\sum_{i=1}^{n}(X_i)^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \tag{27}$$

De esta manera podemos determinar el valor de m y b a partir de nuestros datos. Utilizando las ecuaciones 26 y 27 con los datos de nuestra tabla 1, obtendremos que m= 0.273 B/A y b=-0.950 B (recordar que la unidad de la variable X en nuestro experimento es A y la unidad de la variable Y es B).

Hasta este punto cabe preguntarse que tan bien definidos están los parámetros m y b, pues finalmente estas cantidades deben de tener una cierta incertidumbre asociada, ahora bien la cuestion es ¿como determinar la incertidumbre de estas cantidades? Para responder a está pregunta es necesario recordar como propagar incertidumbres para una función de varias variables *f=f(x,y,...)*:

$$\sigma_f^2 = \left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + ... \tag{28}$$

en este caso nuestras funciones de varias variables son m y b, y las variables de las funciones son las Y_i, pues todo el desarrollo se hizo considerando únicamente la suma de las diferencias cuadráticas entre valores de las ordenadas (datos medidos) y el modelo de una línea recta.

La cantidad σ_{Y_i} se puede factorizar, pues es la misma para todas las variables Y_i, ya que es una medida de dispersión alrededor del modelo matemático, entonces:

$$\sigma_{Y_i} = \sqrt{\frac{1}{n-l}\sum_{i=1}^{n}\left(Y_i - mX_i - b\right)^2} = \sigma_Y \tag{29}$$

donde l es el número de parámetros que tienen nuestro modelo, para nuestro caso *l=2* y n es como siempre nuestro número total de datos.

Para calcular la incertidumbre de m es conveniente desarrollar primero esta cantidad, desarrollando un poco la ecuación 26 se :

$$m = \frac{nX_1Y_1 - Y_1\sum_{i=1}^{n} X_i + nX_2Y_2 - Y_2\sum_{i=1}^{n} X_i + ...}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \tag{30}$$

las derivadas son sobre las Y_i, así que si se toma la derivada del k esimo elemento, esta derivada será:

$$\frac{\partial m}{\partial Y_k} = \frac{nX_k - \sum_{i=1}^{n} X_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \tag{31}$$

entonces

$$\left(\frac{\partial m}{\partial Y_k}\right)^2 = \frac{n^2 X_k^2 + \left(\sum_{i=1}^{n} X_i\right)^2 - 2nX_k \sum_{i=1}^{n} X_i}{\left(n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right)^2} \tag{32}$$

el término anterior es solo una de las n derivadas elevadas cuadrado que se tienen que sumar para poder calcular la incertidumbre asociada a m, entonces sumando los n términos tenemos que:

$$\sum_{k=1}^{n}\left(\frac{\partial m}{\partial Y_k}\right)^2 = \sum_{k=1}^{n} \frac{n^2 X_k^2 + \left(\sum_{i=1}^{n} X_i\right)^2 - 2nX_k \sum_{i=1}^{n} X_i}{\left(n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right)^2} \tag{33}$$

$$= \frac{n^2 \sum_{k=1}^{n} X_k^2 + \sum_{k=1}^{n}\left(\sum_{i=1}^{n} X_i\right)^2 - 2n\sum_{k=1}^{n} X_k \sum_{i=1}^{n} X_i}{\left(n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right)^2} \tag{34}$$

los índices i y k se pueden intercambiar entre si ya que solo representan la forma de numerar los datos, entonces si cambiamos al índice k por i entonces la ecuación 34 queda:

$$\sum_{i=1}^{n}\left(\frac{\partial m}{\partial Y_i}\right)^2 = \frac{n^2 \sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n}\left(\sum_{i=1}^{n} X_i\right)^2 - 2n\sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i}{\left(n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right)^2} \tag{35}$$

$$= \frac{n^2 \sum_{i=1}^{n} X_i^2 + n\left(\sum_{i=1}^{n} X_i\right)^2 - 2n\left(\sum_{i=1}^{n} X_i\right)^2}{\left(n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right)^2} \tag{36}$$

$$= \frac{n^2 \sum_{i=1}^{n} X_i^2 - n\left(\sum_{i=1}^{n} X_i\right)^2}{\left(n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right)^2} \tag{37}$$

$$= n\frac{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}{\left(n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right)^2} \tag{38}$$

con lo cual la ecuación 38 se reduce a:

$$\sum_{i=1}^{n}\left(\frac{\partial m}{\partial Y_i}\right)^2 = \frac{n}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \tag{39}$$

con lo anterior se tiene finalmente que la incertidumbre de m (σ_m) es:

$$\sigma_m = \sigma_Y \sqrt{\frac{n}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}} \tag{40}$$

$$\sigma_m = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - mX_i - b\right)^2}\sqrt{\frac{n}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}} \tag{41}$$

Para encontrar la incertidumbre del parámetro b es necesario hacer el mismo desarrollo, así que desarrollando la ecuación 27:

$$b = \frac{Y_1\sum_{i=1}^{n} X_i^2 - Y_1X_1\sum_{i=1}^{n} X_i + Y_2\sum_{i=1}^{n} X_i^2 - Y_2X_2\sum_{i=1}^{n} X_i + ...}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \tag{42}$$

tomando la derivada respecto al la variable Y_k se obtiene :

$$\frac{\partial b}{\partial Y_k} = \frac{\sum_{i=1}^{n} X_i^2 - X_k\sum_{i=1}^{n} X_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \tag{43}$$

elevando al cuadrado:

$$\left(\frac{\partial b}{\partial Y_k}\right)^2 = \frac{\left(\sum_{i=1}^{n} X_i^2\right)^2 + X_k^2\left(\sum_{i=1}^{n} X_i\right)^2 - 2X_k\sum_{i=1}^{n} X_i^2\sum_{i=1}^{n} X_i}{\left(n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right)^2} \tag{44}$$

Ahora haciendo la suma sobre todos los elementos k

$$\sum_{k=1}^{n}\left(\frac{\partial b}{\partial Y_k}\right)^2=\frac{\sum_{k=1}^{n}\left(\sum_{i=1}^{n}X_i^2\right)^2+\sum_{k=1}^{n}X_k^2\left(\sum_{i=1}^{n}X_i\right)^2-2\sum_{k=1}^{n}X_k\sum_{i=1}^{n}X_i^2\sum_{i=1}^{n}X_i}{\left(n\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2\right)^2} \tag{45}$$

$$=\frac{n\left(\sum_{i=1}^{n}X_i^2\right)^2+\sum_{k=i}^{n}X_k^2\left(\sum_{i=1}^{n}X_i\right)^2-2\sum_{k=1}^{n}X_k\sum_{i=1}^{n}X_i^2\sum_{i=1}^{n}X_i}{\left(n\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2\right)^2} \tag{46}$$

cambiando el índice k por i:

$$\sum_{i=1}^{n}\left(\frac{\partial b}{\partial Y_i}\right)^2=\frac{n\left(\sum_{i=1}^{n}X_i^2\right)^2+\sum_{i=1}^{n}X_i^2\left(\sum_{i=1}^{n}X_i\right)^2-2\sum_{i=1}^{n}X_i\sum_{i=1}^{n}X_i^2\sum_{i=1}^{n}X_i}{\left(n\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2\right)^2} \tag{47}$$

$$=\frac{n\left(\sum_{i=1}^{n}X_i^2\right)^2+\sum_{i=1}^{n}X_i^2\left(\sum_{i=1}^{n}X_i\right)^2-2\sum_{i=1}^{n}X_i^2\left(\sum_{i=1}^{n}X_i\right)^2}{\left(n\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2\right)^2} \tag{48}$$

$$=\frac{n\left(\sum_{i=1}^{n}X_i^2\right)^2-\sum_{i=1}^{n}X_i^2\left(\sum_{i=1}^{n}X_i\right)^2}{\left(n\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2\right)^2} \tag{49}$$

$$=\sum_{i=1}^{n}X_i^2\frac{n\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2}{\left(n\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2\right)^2} \tag{50}$$

simplificando la última expresión, se tiene como resultado:

$$\sum_{i=1}^{n}\left(\frac{\partial b}{\partial Y_i}\right)^2=\frac{\sum_{i=1}^{n}X_i^2}{n\sum_{i=1}^{n}X_i^2-\left(\sum_{i=1}^{n}X_i\right)^2} \tag{51}$$

Con las ecuaciones 29 y 51 se puede escribir la incertidumbre del parámetro b (σ_b)como:

$$\sigma_b = \sigma_Y \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}} \tag{52}$$

$$\sigma_b = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(Y_i - mX_i - b)^2}\sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}} \tag{53}$$

finalmente si se utilizan las ecuaciones 41 y 53 para nuestro conjunto de datos se tiene que $\sigma_m = 0.004$ y $\sigma_b = 0.346$. Con esta información podemos reportar al modelo matemático para nuestros datos como:

$$Y = (0.273 \pm 0.004)X - (0.950 \pm 0.346) \tag{54}$$

Al procedimiento utilizado se le conoce como el método de mínimos cuadrados. Aquí se utilizó para el caso de una línea recta, pero de la misma forma se puede utilizar para cualquier función *f(x)*, claro que la función que se escoja dependerá de como se como se comportan los datos.

0.4.1. Cuando hay errores diferentes en las ordenadas

Todo el desarrollo anterior es útil si los errores de todas las ordenadas son iguales, pero ¿qué ocurre si las ordenadas tienen errores diferentes? En ese caso se necesita considerar de alguna manera como influye el error de cada dato. Sería bueno que los datos con errores mas pequeños fueran mas importantes que los datos con errores mas grandes, una forma de lograr esto es usar una función ligeramente diferente a la función *dif* (ecuación 15). la función que se minimizará es :

$$\chi^2 = \sum_{i=1}^{n} \frac{(Y_i - f(X_i))^2}{\sigma_i^2} \tag{55}$$

Un poco mas adelante se hablara con mas detalle de esta función. A la función anterior se le aplica el mismo proceso de minimización descritos anteriormente para encontrar los parámetros m y b del ajuste a una linea recta, con lo cual se obtiene que:

$$b = \frac{\sum_{i=1}^{n}\frac{X_i^2}{\sigma_i^2}\sum_{i=1}^{n}\frac{Y_i}{\sigma_i^2} - \sum_{i=1}^{n}\frac{X_i}{\sigma_i^2}\sum_{i=1}^{n}\frac{X_iY_i}{\sigma_i^2}}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\sum_{i=1}^{n}\frac{X_i^2}{\sigma_i^2} - \left(\sum_{i=1}^{n}\frac{X_i}{\sigma_i^2}\right)^2} \tag{56}$$

$$m = \frac{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\sum_{i=1}^{n}\frac{X_iY_i}{\sigma_i^2} - \sum_{i=1}^{n}\frac{X_i}{\sigma_i^2}\sum_{i=1}^{n}\frac{Y_i}{\sigma_i^2}}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}\sum_{i=1}^{n}\frac{X_i^2}{\sigma_i^2} - \left(\sum_{i=1}^{n}\frac{X_i}{\sigma_i^2}\right)^2} \tag{57}$$

y para los errores de los parámetros b y m se tiene:

$$\sigma_b^2 = \frac{\sum_{i=1}^{n} \frac{X_i^2}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2} \sum_{i=1}^{n} \frac{X_i^2}{\sigma_i^2} - \left(\sum_{i=1}^{n} \frac{X_i}{\sigma_i^2} \right)^2} \tag{58}$$

y

$$\sigma_m^2 = \frac{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2} \sum_{i=1}^{n} \frac{X_i^2}{\sigma_i^2} - \left(\sum_{i=1}^{n} \frac{X_i}{\sigma_i^2} \right)^2} \tag{59}$$

Un caso particularmente interesante es si los datos pueden ser modelados por lo que comúnmente se llama una ley de potencias es decir

$$Y = f(X) = CX^{\alpha} \tag{60}$$

en este caso lo correcto sería hacer la minimización de la función *dif* (ó χ^2 si los errores no son iguales para las ordenadas) como se hizo para el caso de una línea recta, pero para este nivel puede ser suficiente hacer un pequeño truco. Este truco consiste en sacar logaritmos a las variables X y Y, pues en ese caso lo que tenemos es que:

$$\log(Y) = \log(C) + \alpha \log(X) \tag{61}$$

es decir en el plano logarítmico nuestros datos tienen un comportamiento lineal y podemos utilizar los resultados de ajustar nuestros datos a un modelo lineal. Una vez encontrados los parámetros de la línea recta en el plano logarítmico, podemos aplicar la función inversa del logaritmo para recuperar el comportamiento de nuestros datos en el plano X Y.

0.5. Evaluación del modelo matemático

Una vez que determinamos el valor der los parámetros del modelo matemático, que describe el comportamiento del fenómeno que estudiamos en nuestro experimento, sería bueno saber que tan bien nuestro modelo describe el comportamiento de nuestros datos. Para eso lo primero que podemos hacer, es una gráfica de nuestros datos y el modelo, como se puede ver en la figura 4. En dicha figura se puede observar que el modelo lineal *f(X)* describe bastante bien el comportamiento de nuestros datos, pero esto no es un criterio suficiente ni objetivo para decir que nuestro modelo es bueno o malo. Una buena costumbre es hacer una gráfica de nuestra variable dependiente Y menos nuestro modelo f(X) como las ordenadas y nuestra variable independiente (X) como las abscisas, tal como se muestra en la figura 5.

Si nuestro modelo describe bien el comportamiento de nuestro experimento, se esperaría que en este tipo de gráfica de diferencias, los datos estén distribuidos aleatoriamente arriba y abajo del valor 0. Si los datos presentan un comportamiento sistemático, por ejemplo que en la gráfica a

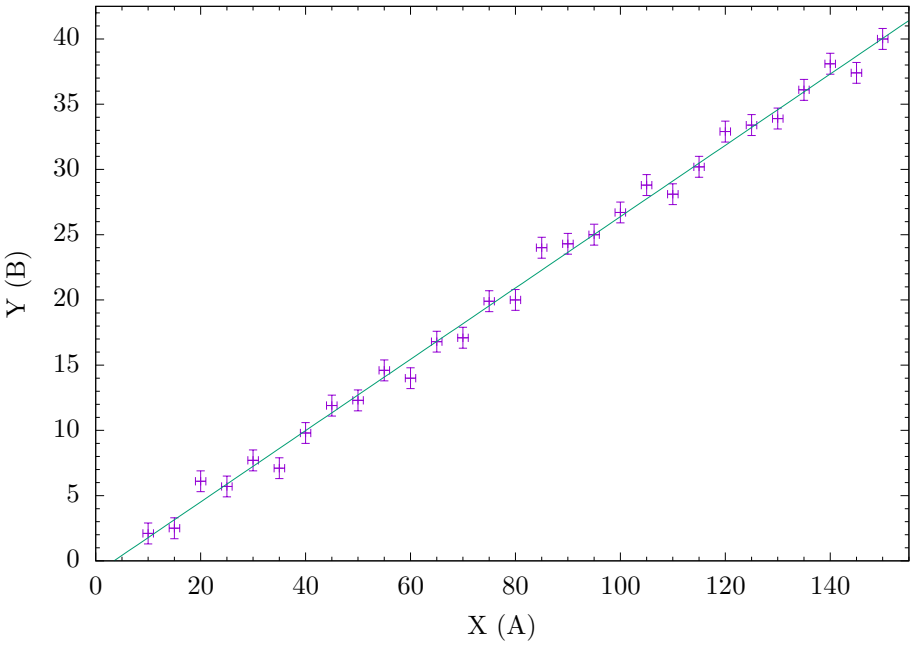

Figura 4: Gráfica de los datos y el modelo lineal encontrado

la izquierda los datos estén por debajo de cero, al centro estén arriba de cero y a la derecha estén otra vez por debajo de cero, esto podría indicar que nuestro modelo no es del todo correcto. Pero de todas maneras este tampoco es un criterio para evaluar a nuestro modelo.

Podemos construir un criterio para evaluar la calidad de nuestro modelo utilizando la ecuación 15 y la dispersión o la incertidumbre de nuestra variable dependiente, esto se conoce como el criterio de χ^2, definiremos a χ^2 como:

$$\chi^2 = \sum_{i=1}^{n} \frac{(Y_i - f(X_i))^2}{\sigma_i^2} \tag{62}$$

donde *f(X)* es el modelo matemático o hipótesis, σ_i es la desviación estándar de nuestra variable Y si esta la determinamos de manera estadística (si no es el caso podemos usar el incertidumbre de Y). En el caso de que el modelo fuera perfecto $\chi^2 = 0$, pero este caso no ocurre, en general el modelo es bueno cuando χ^2 es pequeña, y cuando esta cantidad es grande podemos decir que el modelo es malo. Aún con esto, este criterio es muy ambiguo, ya que es muy arbitrario decir que un número es grande o pequeño si no se tiene una referencia. Para eliminar esta ambigüedad se define la cantidad χ^2_{red}, que será una especie de normalización de la χ^2. Definiremos a χ^2_{red} como:

$$\chi^2_{red} = \frac{\chi^2}{NGL} \tag{63}$$

donde *NGL* es el número de grados de libertad que tenemos. El *NGL* lo definiremos como el número de datos que tenemos (n) menos el número de parámetros que tienen nuestro modelo, *i. e.*:

$$ngl = n - \text{ número de parámetros del modelo} \tag{64}$$

en el caso de nuestro modelo lineal tenemos dos parámetros (m y b) y nuestro número de datos es 29, por lo tanto el número de grados de libertad es $NGL = 29 - 2 = 27$, entonces para nuestro

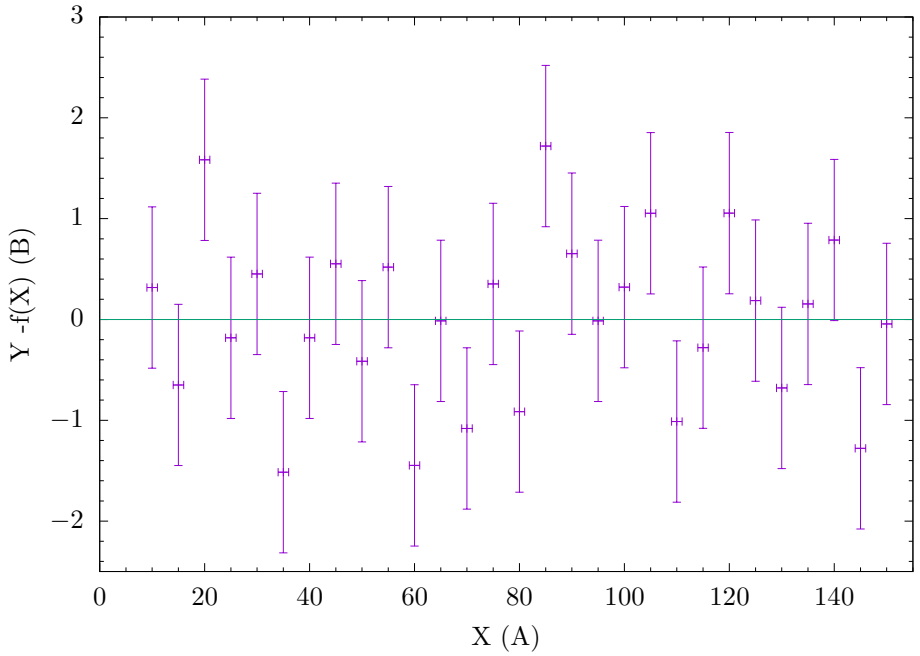

Figura 5: Gráfica de la diferencia entre la variable dependiente y el modelo (Y-f(X))

caso de modelo lineal con 29 datos tendremos que :

$$\chi^2_{red} = \frac{\sum_{i=1}^{n} \frac{(Y_i - mX_i - b)^2}{\sigma_i^2}}{27} \tag{65}$$

Calculando los valores para nuestros datos se obtiene $\chi^2 = 31.495$ y $\chi^2_{red} = 1.166$. El criterio de la χ^2_{red} nos dice que si esta cantidad es cercana a cero nuestro modelo sería perfecto, pero algunos criterios estadísticos nos dicen que es muy raro que esto ocurra (por no decir que se está haciendo trampa). Si χ^2_{red} es menor que 1 podemos asegurar que nuestro modelo es bastante bueno, si χ^2_{red} está entre 1 y 2 nuestro modelo es entre bueno y aceptable, y si χ^2_{red} es mayor que 2 podemos decir que nuestro modelo es malo. En el caso particular de nuestro conjunto de datos tenemos que $\chi^2_{red} = 1.166$, esto indica que nuestro modelo lineal es bastante bueno para describir el comportamiento de nuestros datos.