

UNA DISCUSIÓN SOBRE LA CONFIABILIDAD DE LA TÉCNICA DE EXTRAPOLACIÓN DE LÍNEAS RECTAS DESDE EL PUNTO DE VISTA DE LA TEORÍA DE ERRORES

Wilton Pereira da Silva¹, Cleide M. D. P. S. e Silva¹, Jürgen w. Precker¹,
Diogo D. P. S. e Silva² e Cleiton D. P. S. e Silva³

¹Universidade Federal de Campina Grande, Depto. Física,
Campina Grande, Paraíba, Brasil

²Universidade Federal de Campina Grande, Depto. Matemática,
Campina Grande, Paraíba, Brasil

³Instituto Tecnológico de Aeronáutica, Depto. Sistemas e Controle,
São José dos Campos, São Paulo, Brasil

RESUMEN

El siguiente podría ser un experimento típico en el laboratorio de pregrado: la presión del aire es medida como función de la temperatura a volumen constante y en un rango definido de temperatura; luego, los resultados son representados gráficamente. Entonces, una línea recta es ajustada de forma visual (o por mínimos cuadrados) a los datos experimentales y ésta es extrapolada al valor de presión cero con la esperanza de encontrar el cero absoluto de temperatura. Pero, ¿se puede confiar en la información obtenida de tal extrapolación? Investigamos aquí este problema de una manera general utilizando métodos estadísticos para el ajuste con una línea recta, y aplicamos los resultados al problema del experimento descrito arriba.

1. INTRODUCCIÓN

Una técnica usada en física para obtener información a partir de un conjunto de datos experimentales es su representación gráfica. Una práctica común es extender la interpretación de los datos experimentales más allá del intervalo en el que fueron observados. De aquí el nombre “extrapolación”.

Como ejemplo se tiene la estimación experimental de la temperatura de cero absoluto midiendo la presión de un gas a volumen constante como función de la temperatura [1], y de aquí, la extrapolación para una presión $P=0$ del gas, una vez que la escala de temperatura absoluta ha sido definida para el gas ideal [2]. La validez de este procedimiento se da por sentada no solamente para $P=0$, sino para cualquier otro valor P de la presión.

Aunque esta técnica de extrapolación nació junto con las ciencias experimentales, recién algunas cuestiones sobre la precisión absoluta y relativa de los resultados se pueden contestar eficientemente en la actualidad. Junto a la expansión del uso de las computadoras, varios programas surgieron y permiten responder a estas cuestiones a través del análisis estadístico para un conjunto amplio de funciones (MATLAB, Origin, TableCurve, LAB Fit, etc.)

En este artículo investigamos cuán confidente es la técnica de extrapolación desde el punto de vista estadístico, siempre que el objeto de estudio se pueda representar por una línea recta. Por lo tanto, debemos revisar brevemente el ajuste de mínimos cuadrados de un

conjunto de datos (x_i, y_i) a una función $y = ax + b$, si las incertidumbres de y_i son desconocidas. Asimismo, necesitamos asociar las incertidumbres a la función de ajuste de tal forma que ésta se exprese finalmente de la forma $y(x) = y(x)_m \pm \sigma_{y(x)_m}$, donde $y(x)_m = ax + b$ es el valor medio de la función de ajuste y $\sigma_{y(x)_m}$ es su desviación estándar.

Restringiremos nuestra discusión a los errores estadísticos, e.g., supondremos que se puede despreciar errores sistemáticos de los datos (si no, vea la ref. [3]), y que la extrapolación no se saldrá del rango válido para el modelo físico en cuestión. Bajo estas restricciones, mostramos que la técnica de extrapolación produce resultados satisfactorios y confiables en el caso de un ajuste de línea recta y que estos resultados son comparables a aquellos obtenidos por interpolación.

2. ASPECTOS TEÓRICOS RELEVANTES

Antes de abordar los aspectos teóricos relevantes de este artículo, aceptaremos las siguientes hipótesis para la discusión a lo largo de todo el artículo:

- 1) los errores sistemáticos de las mediciones se pueden considerar despreciables;
- 2) la propagación de errores se puede calcular usando aproximaciones de primer orden;
- 3) la fluctuación de los puntos en torno a la función de ajuste se puede considerar gaussiana;
- 4) la ley física que es objeto de la extrapolación es válida en todo el intervalo de estudio;
- 5) la variable x no tiene errores.

¹Email: wiltonps@uol.com.br

Dado un conjunto de N puntos $(x_i, y_i \pm \sigma_{ymi})$, donde x_i y y_i son los valores medios de la abscisa y ordenada del i -ésimo punto respectivamente, y las σ_{ymi} son las incertidumbres de y_i , el ajuste de la función $y = ax + b$ se puede encontrar por las expresiones [4, 5, 6]:

$$a = \frac{1}{D} \left[\left(\sum_{i=1}^N \frac{x_i y_i}{\sigma_{ymi}^2} \right) \cdot \left(\sum_{i=1}^N \frac{1}{\sigma_{ymi}^2} \right) - \left(\sum_{i=1}^N \frac{x_i}{\sigma_{ymi}^2} \right) \cdot \left(\sum_{i=1}^N \frac{y_i}{\sigma_{ymi}^2} \right) \right] \quad (1)$$

y

$$b = \frac{1}{D} \left[\left(\sum_{i=1}^N \frac{x_i^2}{\sigma_{ymi}^2} \right) \cdot \left(\sum_{i=1}^N \frac{y_i}{\sigma_{ymi}^2} \right) - \left(\sum_{i=1}^N \frac{x_i}{\sigma_{ymi}^2} \right) \cdot \left(\sum_{i=1}^N \frac{x_i y_i}{\sigma_{ymi}^2} \right) \right], \quad (2)$$

donde

$$D = \left(\sum_{i=1}^N \frac{x_i^2}{\sigma_{ymi}^2} \right) \cdot \left(\sum_{i=1}^N \frac{1}{\sigma_{ymi}^2} \right) - \left(\sum_{i=1}^N \frac{x_i}{\sigma_{ymi}^2} \right)^2. \quad (3)$$

Las incertidumbres de los parámetros a y b , así como su covarianza están dados por

$$\sigma_{am} = \sqrt{\frac{1}{D} \sum_{i=1}^N \frac{1}{\sigma_{ymi}^2}}, \quad (4)$$

$$\sigma_{bm} = \sqrt{\frac{1}{D} \sum_{i=1}^N \frac{x_i^2}{\sigma_{ymi}^2}}, \quad (5)$$

$$\text{cov}(a, b) = -\frac{1}{D} \left(\sum_{i=1}^N \frac{x_i}{\sigma_{ymi}^2} \right), \quad (6)$$

donde D es la misma variable que en (3).

Para el caso en el que se desconozca las incertidumbres de los datos experimentales, e.g., tenemos puntos del tipo (x_i, y_i) , se puede asociar una incertidumbre común a cada uno de ellos de la siguiente forma: primero, sustituimos las incertidumbres desconocidas por el valor artificial 1, esto es, fijamos $\sigma_{ymi} = 1$ en (1) y (2), y determinamos los parámetros a y b por un pre-ajuste $y(x)$; luego, podemos determinar la varianza asociada a la función ajustada $y(x)$ por la siguiente expresión [7]:

$$\sigma_{y(x)}^2 = \frac{1}{N-2} \sum_{i=1}^N [y_i - y(x_i)]^2. \quad (7)$$

Ya que la raíz cuadrada de la varianza nos proporciona una indicación de la fluctuación de los datos experimentales en torno a la línea recta ajustada, es razonable admitir que las incertidumbres desconocidas de y_i son iguales a $\sigma_{y(x)}$. Así, las ecs. (4), (5) y (6) se reducen a:

$$\sigma_{am} = \sigma_{y(x)} \sqrt{\frac{N}{N \sum_{i=1}^N x_i^2 - \left(\sum_{i=1}^N x_i \right)^2}}, \quad (8)$$

$$\sigma_{bm} = \sigma_{y(x)} \sqrt{\frac{\sum_{i=1}^N x_i^2}{N \sum_{i=1}^N x_i^2 - \left(\sum_{i=1}^N x_i \right)^2}} \quad (9)$$

y

$$\text{cov}(a, b) = -\frac{\sigma_{y(x)}^2}{N \sum_{i=1}^N x_i^2 - \left(\sum_{i=1}^N x_i \right)^2} \sum_{i=1}^N x_i. \quad (10)$$

Podemos determinar la incertidumbre de la función ajustada $y(x) = ax + b$ a partir de las ecs. (8), (9) y (10) por medio de propagación de errores. Para esto, recordemos que la propagación de la desviación estándar para una función $f(z_1, z_2, \dots, z_n)$ debido a las incertidumbres de los n parámetros z de f está dada por [8]:

$$\sigma_{fm} = \sqrt{\sum_{j=1}^n \sum_{k=1}^n \frac{\partial f}{\partial z_j} \frac{\partial f}{\partial z_k} \text{cov}(z_j, z_k)}. \quad (11)$$

Para una función f de dos parámetros, digamos a y b , (11) se puede escribir como

$$\sigma_{fm} = \sqrt{\left(\frac{\partial f}{\partial a} \sigma_{am} \right)^2 + \left(\frac{\partial f}{\partial b} \sigma_{bm} \right)^2 + 2 \frac{\partial f}{\partial a} \frac{\partial f}{\partial b} \text{cov}(a, b)}. \quad (12)$$

En el caso de una línea recta ajustada a puntos experimentales, y considerando los x_i libres de error, el error de la función $y(x) = ax + b$ ocurre debido a las incertidumbres de a y b , incluyendo la covarianza entre ellos. Así, de (12) se tiene la línea recta:

$$\sigma_{y(x)m} = \sqrt{(x \sigma_{am})^2 + \sigma_{bm}^2 + 2x \text{cov}(a, b)}. \quad (13)$$

Podemos por lo tanto escribir la función completa de ajuste como:

$$y(x) = (ax + b) \pm \sqrt{\sigma_{am}^2 x^2 + 2 \text{cov}(a, b)x + \sigma_{bm}^2}. \quad (14)$$

TABLA 1

Presión P del aire como una función de la temperatura t a volumen constante.

	1	2	3	4	5
P (cm Hg)	72,7	75,0	75,9	77,3	78,4
t ($^{\circ}C$)	30,0	38,0	42,3	48,0	54,0

De esta manera, la representación gráfica de la función de ajuste consiste de tres líneas: la línea interior representa el valor medio de la función dado por $y(x)_m = ax + b$, mientras que las dos líneas exteriores dan los límites inferior y superior de la banda de confiabilidad defida por $\pm\sigma_{y(x)_m}$ y calculada por medio de (13).

A continuación investigaremos si el resultado obtenido de (14) y extrapolado más allá del rango de los puntos experimentales es confiable, o si las incertidumbres crecen de tal forma que los límites aceptables de precisión se ven comprometidos. Para esto, apliquemos las ecuaciones referidas arriba a un conjunto de datos experimentales.

3. APLICACIÓN A UN CONJUNTO DE DATOS: DETERMINACIÓN DEL CERO ABSOLUTO Y UNA GENERALIZACIÓN DE LA TÉCNICA DE EXTRAPOLACIÓN

A fin de evaluar la técnica de extrapolación, apliquémosla a un experimento en el que se determine el cero absoluto de la temperatura. El experimento consiste de un bulbo con aire que contiene a un termómetro. El bulbo está inmerso en un recipiente con agua colocado sobre un calentador. Este sistema se conecta a un manómetro; el volumen de aire en el bulbo se mantiene constante durante las mediciones. Ya que este experimento es común en los laboratorios de pregrado, no entraremos en detalles sino más bien presentaremos los resultados experimentales. Además, nuestro interés principal comprende el análisis de datos desde el punto de vista de la teoría de errores, y no así el experimento en sí. La tabla 1 muestra los valores medidos de la presión absoluta del aire dentro del bulbo como una función de la temperatura.

Naturalmente, la temperatura t se tomaría como una variable independiente, pero ya que queremos hallar la temperatura a la cual la presión es cero, entonces ajustaremos la función $t(P) = aP + b$ a los datos de la tabla 1.

De (1) y (2), y estableciendo $\sigma_{ymi} = 1$, obtenemos $a \cong 4,1809$ y $b \cong -274,70$. Con estos valores, se obtiene $\sigma_{y(x)} \cong 0,912416$ a partir de (7), y de (8), (9) y (10) obtenemos $\sigma_{am} \cong 0,207948$, $\sigma_{bm} \cong 15,7802$ y $cov(a,b) \cong -3,28037$. En consecuencia, $a = (4,18 \pm 0,21) ^{\circ}C/cm Hg$ y $b = (-275 \pm 16) ^{\circ}C$, así que finalmente podemos escribir

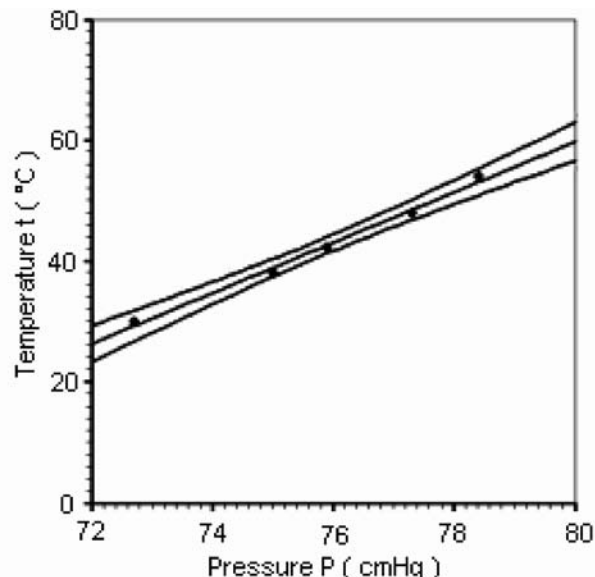


Figura 1. Función dada por (15) ajustada a los valores experimentales. La línea central representa el valor medio de la función, mientras las líneas inferior y superior limitan el 95 % de la banda de confiabilidad.

$$t(P) = (4.18P - 275) \pm$$

$$2.87 \sqrt{4,32424 \times 10^{-2}P^2 - 6,56074P + 249,015091}, \tag{15}$$

donde la incertidumbre fue calculada a partir de (13). El factor 2,87 que aparece delante de la raíz en (15) –que describe la incertidumbre– asegura que la banda de confiabilidad de la función que se ajustó solamente a 5 datos es del 95 %, habiéndose probado que los datos están distribuidos de manera normal en torno a la función de ajuste. Estableciendo $P=0$ cm Hg in (15), tenemos que $t = (-275 \pm 45) ^{\circ}C$. Es interesante observar que la precisión es mucho mayor que la correspondiente al cero absoluto. Naturalmente, esta determinación de la temperatura t para el valor específico de presión cero, no necesita tomar en cuenta la covarianza entre los parámetros ajustados a y b para el cálculo de la propagación de errores. Para este caso, que está ya discutido por Taylor [9], $t(0)$ es la constante b . Sin embargo, la covarianza debe considerarse para la determinación de t para cualquier otra presión P diferente de cero; es esta generalización la cuestión de nuestra discusión. Aún más: deseamos investigar si los resultados para $t(P)$ en la región extrapolada tienen una precisión compatible con los resultados dentro de la región apropiada. Si es así, podríamos admitir, desde un punto de vista estadístico, que los resultados de una extrapolación tienen la misma confiabilidad que aquellos obtenidos por interpolación, siempre que se respete nuestras hipótesis. Si, por otra parte, la incertidumbre relativa de $t(P)$ crece a medida que nos alejamos de la región de datos, los resultados obtenidos via extrapolación no serían acreedores a la misma confiabilidad que aquellos dentro de la región de datos.

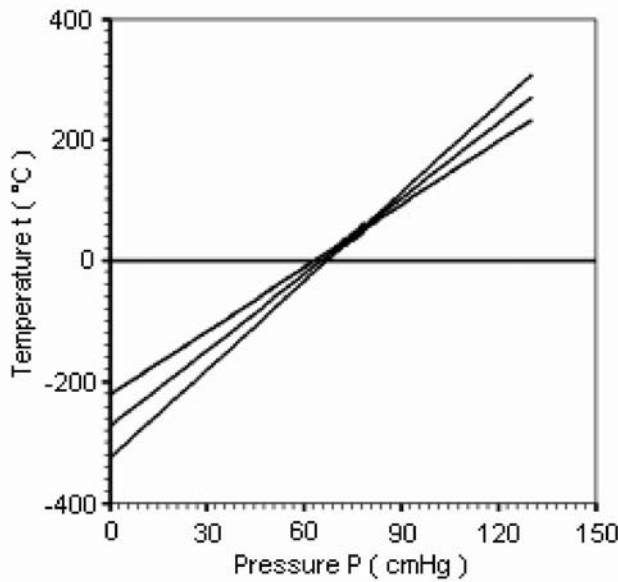


Figura 2. Función ajusta dada por (15), extrapolada desde $P = 72,7 \text{ cmHg}$ hasta $P = 0 \text{ cmHg}$, y de $P = 78,4 \text{ cmHg}$ hasta 130 cmHg . La banda muestra en 95 % de confiabilidad.

En la Fig. 1 se muestra la temperatura versus la presión dada por (15). Se puede ver que la banda de confiabilidad se dispersa a medida que se acerca a los extremos de la región de datos.

La Fig. 2 muestra el comportamiento de la incertidumbre de t entre 0 y 130 cmHg ; podemos ver que la incertidumbre crece fuertemente si nos alejamos de la región de datos en ambas direcciones.

La Fig. 3 muestra la incertidumbre de la temperatura t ,

$$\sigma_{t(P)_m} = 2,87 \sqrt{4,32424 \times 10^{-2} P^2 - 6,56074 P + 249,015091}, \quad (16)$$

dentro de la región de datos.

La incertidumbre es mínima en el valor promedio $P=75,86 \text{ cmHg}$, y crece en ambas direcciones a medida que nos alejamos de dicho valor. En nuestro caso, donde todas las mediciones tiene el mismo error común, el mínimo ocurre en la media aritmética. Si no, el mínimo aparecería en la media pesada estadísticamente.

La Fig. 4 muestra la incertidumbre de la temperatura t entre 0 y 130 cmHg ; podemos ver que la incertidumbre crece rápidamente más allá de la región de datos, lo que no resulta muy alentador.

Así, a fin de averiguar si la técnica de extrapolación es confiable, consideremos –a pesar de la creciente incertidumbre absoluta– la incertidumbre relativa σ_r de la temperatura t :

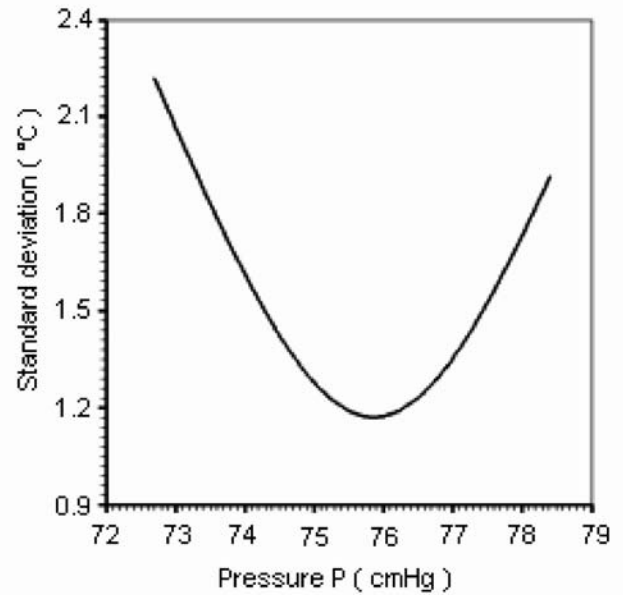


Figura 3. Desviación estándar de la función ajustada dada por (16), dentro de la región de datos. El mínimo ocurre en el valor promedio $P = 75,86 \text{ cmHg}$.

$$\sigma_r = 2,87 \sqrt{\frac{4,32424 \times 10^{-2} P^2 - 6,56074 P + 249,015091}{(4,18 P - 275)^2}}. \quad (17)$$

La Fig. 5 muestra que la incertidumbre relativa también crece mientras va del centro hacia los extremos de la región de datos. Hasta donde se puede apreciar, las cifras indican que la técnica de extrapolación no parece ser confiable.

Ya que tenemos una singularidad en $P=65,8 \text{ cmHg}$, la presión a la que la temperatura t toma el valor cero, dividimos el gráfico en dos regiones: desde $P = 0 \text{ cmHg}$ hasta $P = 60 \text{ cmHg}$, y desde $P = 70 \text{ cmHg}$ hasta $P = 130 \text{ cmHg}$. De manera interesante, vemos que la desviación estándar relativa se hace constante lejos de la región de datos. Primeramente, vemos que para $x(=P) > 78,4 \text{ cmHg}$, podemos entender este comportamiento a partir de (13): escribimos para la desviación estándar relativa

$$\frac{\sigma_{y(x)_m}}{y(x)_m} = 2,87 \sqrt{\frac{(x\sigma_{am})^2 + 2x\text{cov}(a, b) + \sigma_{bm}^2}{(ax + b)^2}}, \quad (18)$$

lo que nos da

$$\sigma_r = 2,87 \sigma_{ymi} / y(x)_m \rightarrow 2,87 \sigma_{am} / a \quad (19a)$$

en el límite $x \rightarrow \infty$. Sustituyendo $\sigma_{am} = 0,207948$ y $a = 4,1809$ en esta expresión obtenemos $\sigma_r = 0,143$, que se muestra como la línea segmentada en la Fig. 6. Para $x(=P) = 0$ tenemos

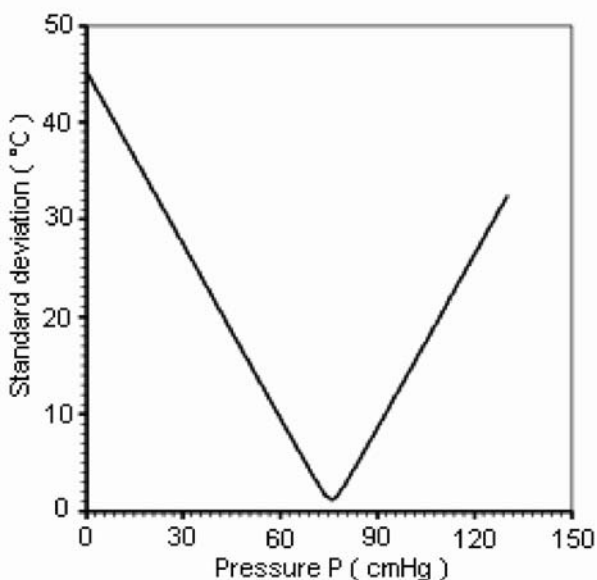


Figura 4. Desviación estándar de la función de ajuste dada por (16), extrapolada desde $P = 72,7$ hasta $P = 0$ cmHg , y desde $P = 78,4$ cmHg hasta 130 cmHg.

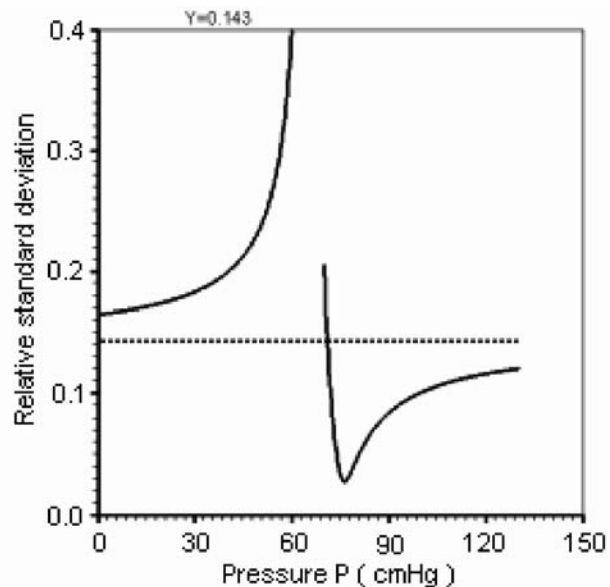


Figura 6. Desviación estándar relativa de la función ajustada de acuerdo a (17), Extrapolada desde $P = 60$ cmHg hasta $P = 0$ cmHg, y desde $P = 70$ cmHg hasta $P=130$ cmHg.

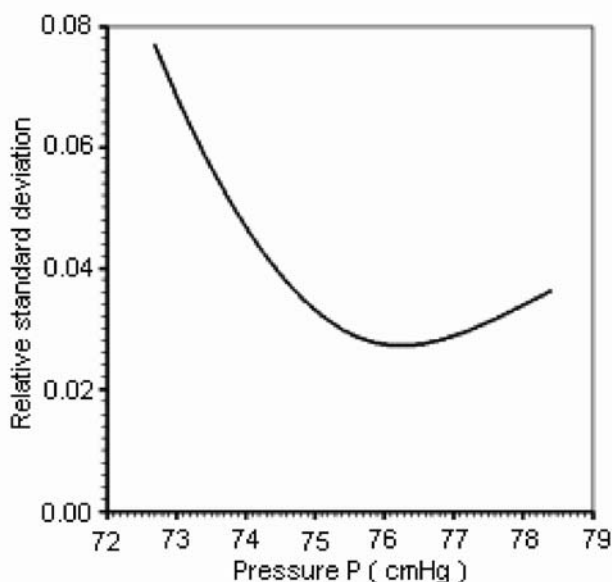


Figura 5. Desviación estándar relativa de la función ajustada de acuerdo a (17), dentro de la región de datos. Si se aplica (17) desde 0 cmHg hasta 130 cmHg, obtenemos el resultado que se muestra en la Fig. 6.

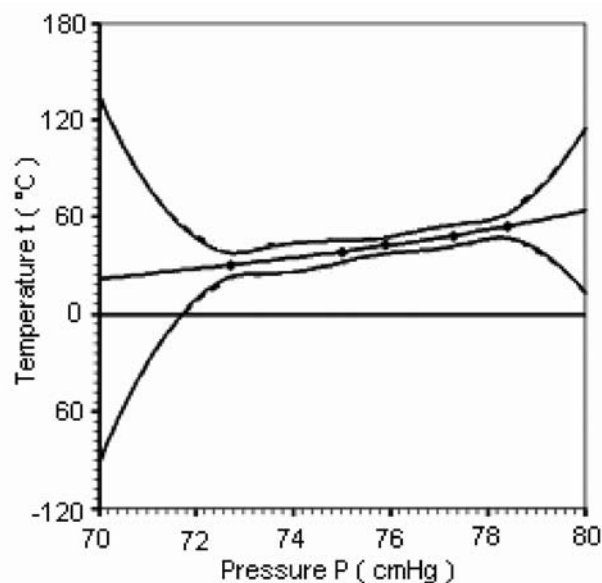


Figura 7. Función de tercer grado ajustada a los mismos datos experimentales dados en la Tabla 1. Una pequeña extrapolación ya crea incertidumbres tan grandes que ninguna información se puede extraer desde el exterior de la región de datos.

$$\sigma_r = 2,87\sigma_{y_{mi}}/y(x)_m \rightarrow 2,87\sigma_{b_m}/b, \quad (19b)$$

cuyo resultado es $\sigma_r = 0,165$.

Esto significa que si estamos lejos de la región de datos hacia la derecha (o bien si P puede tomar un valor mucho menor que cero), la pendiente de la línea recta es el parámetro dominante y, en consecuencia, la covarianza entre los parámetros se vuelve despreciable. Este comportamiento está de acuerdo con la referencia [10],

donde se discute sobre la covarianza de un par de variables correlacionadas.

Para una presión $P = 68,5$ cmHg, la temperatura toma el valor $t=0$ °C y su incertidumbre relativa es infinita. Pero, tal como muestra la Fig. 4, la incertidumbre relativa de la temperatura es finita y completamente aceptable al compararse con los valores involucrados. Esta singularidad es pues una consecuencia de la escala de temperatura elegida.

4. DISCUSIONES Y CONCLUSIONES

De las Figs. 3 y 4, que muestran las desviaciones estándar según (16), podemos concluir que las incertidumbres más pequeñas para una función de primer orden con dos parámetros ajustados, están dentro de la región de datos, con un mínimo en la presión promedio $P = 75,86 \text{ cm Hg}$. Al alejarse de la región de datos, la incertidumbre absoluta crece en ambas direcciones.

Por otra parte, de las Figs. 5 y 6, que muestran las desviaciones estándar según (17), podemos ver que las incertidumbres relativas se comportan razonablemente tanto dentro como fuera de la región de datos, excluyendo el intervalo en torno a la singularidad para $t = 0 \text{ }^\circ\text{C}$, lo que es una peculiaridad de la escala de temperaturas usada.

Como se puede ver de (18), (19) y la Fig. 6, a medida que nos alejamos de la región de datos, la precisión relativa de la función ajustada depende menos del parámetro b y de la covarianza entre a y b . Por lo tanto, la confiabilidad de un resultado extrapolado obtenido por $y(x)_m$, para valores de x suficientemente alejados de la región de datos, puede indicarse por la precisión relativa del coeficiente angular a de la línea recta.

Tomando asimismo en cuenta que la incertidumbre absoluta de la recta ajustada es finita en toda la región analizada, tal como se muestra en las Figs. 3 y 4, podemos concluir que la aplicación de la técnica de extrapolación produce resultados tan satisfactorios como los de la interpolación. Naturalmente, esta conclusión tiene exclusivamente por base los errores estadísticos y no así los errores sistemáticos. Aunque esta conclusión pueda parecer algo obvia, no es válida para cualquier función ajustada, como se muestra, por ejemplo, en la Fig. 7 para el caso de un polinomio de tercer grado: una interpolación sería aceptable pero ya no una extrapolación.

Desde luego, no podemos probar que la información extraída de la línea recta extrapolada es verdadera. Pero al menos, si uno está dispuesto a seguir nuestra argumentación, podemos concluir que desde el punto de vista

de la estadística, no hay objeción para extrapolar líneas rectas y explorarlas en busca de información. Además, podemos cuantificar las incertidumbres involucradas en la extrapolación de una línea recta. Aunque hemos restringido nuestra investigación a los puntos sujetos a una línea recta, sería interesante estudiar la confiabilidad de extrapolaciones de otros tipos de funciones que se saben sujetas a un conjunto dado de datos experimentales.

REFERENCIAS

- [1] Frederick J. Keller, W. Edward Gettys and Malcolm J. Skove, *Physics*, (McGraw-Hill, Inc., 1993), 2da. ed., vol. 1, capítulo 16.
- [2] David Halliday, Robert Resnick and John Merrill, *Fundamentals of Physics*, (John Wiley & Sons, Inc., 1988), 3ra. ed., vol. 2, capítulo 19.
- [3] John Bechhoefer, Curve fits in the Presence of Random and Systematic Error, *Am. J. Phys.* **68** (5) 424-429 (2000).
- [4] P. R. Bevington and D. K. Robinson, *Data Reduction and Error Analysis for the Physical Sciences*, (McGraw-Hill, New York, 1992) 2da. ed., pp 96-114.
- [5] W. H. Press, S. A. Teukolski, W. T. Vetterling and B. P. Flannery, *Numerical Recipes in Fortran: The Art of Scientific Computing*, (Cambridge U. P., Cambridge, 1992) 2da. ed., pp 650-659.
- [6] Silva, Wilton P. and Silva, Cleide M. D. P. S., *Tratamento de Dados Experimentais*, (UFPB/Editora Universitária, João Pessoa, PB, 1998), 2da. ed., pp 152-163.
- [7] Silva, Wilton P. et al., Geração de Incertezas de Funções Redutíveis ao Primeiro Grau Ajustadas pelo Método dos Mínimos Quadrados, *Rev. Bras. Ensino de Física*, **21**, nº 3, setembro, 1999
- [8] Helene, Otaviano A. M. e Vanin, Vitor R., *Tratamento Estatístico de Dados em Física Experimental*, (Ed. Edgard Blücher LTDA, São Paulo, 1981), pp. 49-54.
- [9] John R. Taylor, *An Introduction to Error Analysis* 2nd Edition, p.192, University Science Books, Sausalito, California (1997)
- [10] John R. Taylor, Simple Examples of Correlations in Error Propagation, *Am. J. Phys.* **53** (7), 663-667 (1985)