JOURNAL OF SCIENCE EDUCATION (REVISTA DE EDUCACIÓN EN CIENCIAS) N 1, VOL 4, pp 8-14, 2003

USO DEL PAQUETE ESTADÍSTICO *SIMFIT* EN LA ENSEÑANZA DEL ANALISIS DE DATOS EN CIENCIAS EXPERIMENTALES

Using the *SIMFIT* Statistical Package to teach Data Analysis in Experimental Sciences

F. J. Burguillo[1] M. Holgado[1] y W. G. Bardsley[2]

[1]Dpto. de Química Física, Facultad de Farmacia, Universidad de Salamanca (España)
[2]School of Biological Sciences, Oxford Road, University of Manchester (U. K.)

Resumen

El propósito de este artículo es sugerir el uso de paquetes estadísticos en la enseñanza del análisis de datos como una alternativa a las hojas de cálculo y a los programas matemáticos. Para ello, se describen los contenidos teóricos y prácticos de un curso de análisis de datos que está basado íntegramente en el paquete estadístico *SIMFIT*. Este software ha sido desarrollado por uno de nosotros (W.G. Bardsley), es totalmente compatible con Windows 95/98/NT/2000/Me/XP y se encuentra disponible de modo gratuito en Internet tanto en inglés (http://www.simfit.man.ac.uk) como en español (http://simfit.usal.es). *SIMFIT* está compuesto de cuarenta programas de ajuste de curvas y estadística controlados desde un menú principal, de un programa de ayuda, de un manual de referencia en formato PDF y de varios archivos de prueba para practicar con todos los programas. Se exponen los contenidos de las sesiones teóricas del curso, que van desde el ajuste de curvas por regresión lineal y no lineal a la estadística aplicada. También se describen los talleres prácticos, que se desarrollan en un aula de informática y siguen la metodología del "caso práctico". Los casos contemplan las situaciones habituales que se dan en los laboratorios, como son el ajuste de curvas de calibrado por rectas o polinomios, la modelización de sistemas con exponenciales, la estadística descriptiva o las comparaciones tipo ANOVA.

Palabras clave: Análisis de datos, ajuste de curvas, tests estadísticos, simulación, gráficas.

Abstract

The aim of this paper is to suggest the use of statistical packages in the teaching of data analysis as an alternative to the spreadsheets and mathematical programs. In order to do so, the

theoretical and practical contents of a course on data analysis based entirely on the *SIMFIT* statistical package are described. This software was developed by one of us (W.G. Bardsley), it is fully Windows 95/98/NT/2000/Me/XP compatible, and is available free of charge through the Internet in both English (http://www.simfit.man.ac.uk) and Spanish (http://simfit.usal.es). *SIMFIT* consists of forty curve fitting and statistics programs drived from menus, a reference manual in PDF format, a help program and many test files to practice with each program. The contents of the theoretical lectures of the course are explained; they range from the linear and non-linear regression to the applied statistics. Practical workshops are also described; these are conducted in a computer room and follow the practical case method. The cases offer common situations found in the laboratories, such as curve fitting of calibration curves by straight lines or polynomials, the modelling of systems by exponentials, descriptive statistics, and ANOVA procedures.

Key words: Data analysis, Curve fitting, Statistical tests, Simulation, Graph plotting.

Introducción

En las licenciaturas de ciencias experimentales, los alumnos suelen cursar en distintas asignaturas los fundamentos del análisis de datos: incertidumbre de una medida, concepto de precisión y exactitud, réplicas, media y desviación estándar, ajuste de una recta por mínimos cuadrados...etc (Treptow (1998), Guare (1991), Malinowski (1995), Lieb (1997)). Para ello, en los primeros cursos, se utilizan calculadoras de bolsillo equipadas con funciones estadísticas, regresión lineal y gráficas. Sin embargo, en los últimos años de facultad y en todas las enseñanzas de postgradro, los alumnos dan un paso adelante y usan hojas de cálculo, programas matemáticos o paquetes estadísticos, adentrándose así en técnicas más avanzadas del análisis de datos, como son el ajuste de curvas por regresión no lineal, integración de ecuaciones diferenciales, pruebas estadísticas tipo ANOVA...etc.

Recientemente, se han publicado numerosos trabajos que han estudiado esta realidad, algunos analizando las aplicaciones gráficas de las calculadoras de bolsillo (Kim *et al.* (2000)), otros defendiendo las ventajas de las hojas de calculo tipo *Excel* (Harris (1998), Denton (2000)), muchos de ellos basados en las cualidades de programas matemáticos como *Mathcad* (Zielinski (2000 y 1995), Young *et al.* (1995)) o *Mathematica* (Ferreira *et al.* (1999), Bruce *et al.* (1993)) y solamente unos pocos apostando por el uso de paquetes estadísticos como *KaleidaGraph*, *Origin* o *SPSS* (Tellinghuisen (2000)). La pregunta acerca de cuál de estas estrategias es la más adecuada no es de fácil respuesta, ya que ésta depende de muchos factores: de la dificultad del problema, de la preparación matemática de los alumnos, del carácter amigable de los programas, así como de su popularidad entre alumnos y profesores. Sin duda, hacen falta estudios comparativos que clarifiquen las ventajas e inconvenientes de estas metodologías, sobre todo en cuanto a su rendimiento con alumnos que se enfrentan a estos temas por primera vez o cuya preparación matemática es escasa.

En este artículo se sugiere la conveniencia de enseñar las técnicas de análisis de datos mediante el uso de un paquete estadístico amigable, que ayude a los alumnos a centrarse más en los métodos en sí mismos que en el manejo informático de un programa y ofrecer a la vez la posibilidad de analizar sus datos en la misma forma que lo hacen los investigadores en la vida real. Para exponer estos aspectos, se analizan los contenidos teóricos y prácticos de un curso intensivo de análisis de datos que se imparte en la Universidad de Salamanca a postgraduados que se inician en la investigación. Este curso sería adaptable a las licenciaturas de Física, Matemáticas, Química, Biología, Farmacia o Medio Ambiente y podría formar parte de los estudios de postgrado en ciencias experimentales.

Metodología

Estructura del curso de análisis de datos

La duración del curso es de tres dias en jornada completa y su estructura es la siguiente:

- *Clases teóricas sobre aspectos fundamentales del análisis de datos.* Durante tres mañanas un profesor expone las bases teóricas de los distintos métodos del análisis de datos y hace numerosas demostraciones en pantalla grande mediante un ordenador y un cañón de proyección.

- *Talleres prácticos en el aula de informática:* Durante tres tardes cada alumno analiza distintos casos prácticos facilitados por el profesor, los cuales están pensados para cubrir los diferentes aspectos introducidos en las clases teóricas.

- *Consultoría sobre datos personales.* Los alumnos preguntan a los profesores el cómo analizar sus datos en concreto, lo que sirve para enriquecer el curso y motivar a los participantes.

Características del software utilizado en el curso

Se usa el paquete estadístico *SIMFIT* (Fig. 1), ya que se adapta perfectamente a nuestros objetivos y se encuentra disponible gratuitamente en Internet tanto en inglés (http://www.simfit.man.ac.uk) como en español (http://simfit.usal.es). Este software ha sido desarrollado por uno de nosotros (W.G. Bardsley) y es totalmente compatible con Windows 95/98/NT/2000/Me/XP. Entre sus características, cabe destacar:

- Consta de cuarenta programas controlados desde un menú principal que proporciona acceso a todos los módulos, a un programa de ayuda, al manual y a varios archivos de prueba para practicar con todos los programas.

- Tiene editores para introducir los datos desde el teclado, pero también se pueden importar desde *Excel* a través del portapapeles.

- Dispone de programas en formato amigable que ajustan tanto ecuaciones lineales (línea recta, polinomios) como no lineales (exponenciales, Michaelis-Menten, ecuaciones de crecimiento, unión de ligandos a macromoléculas...etc).

- Los propios programas escalan internamente los datos y calculan las estimas iniciales, ajustando luego en secuencia diferentes modelos dentro de la jerarquía elegida (por ejemplo: una exponencial, dos exponenciales, tres exponenciales...), calculando en cada caso la bondad del ajuste (residuales, r^2, test F, límites de confianza de los parámetros...). Con esta información, el usuario puede discriminar entre posibles modelos rivales y quedarse con el que cumpla el mayor número de requisitos estadísticos.

- Se superpone en una sola gráfica los datos y las curvas ajustadas, de forma que el usuario puede comparar visualmente los diferentes ajustes. Esta superposición puede hacerse también en otros tipos de gráfica (logarítmica, doble inversa, Scatchard, Hill...).

- Si la ecuación a ajustar no está disponible en la librería de *SIMFIT*, el usuario puede definir su propia ecuación mediante reglas sencillas y luego ajustarla a sus datos mediante un programa general (*QNFIT*) en el que todas las decisiones son tomadas interactivamente por el usuario.

- *SIMFIT* incluye todas las pruebas estadísticas habituales: estadística descriptiva de una muestra, comparación de medias, ANOVA, pruebas no paramétricas...etc.

- Se pueden hacer diferentes tipos de gráficas: curvas de frecuencia, histogramas, diagramas de barras y de sectores, gráficas 3D... En todas ellas es posible asignar distintos tamaños y colores a los diferenetes objetos.

Contenido de las sesiones teóricas del curso

Las clases teóricas constan de dos partes, una primera en la que se exponen los principios matemáticos base de cada método de ajuste de curvas o test estadístico y otra en la que se hace una demostración con ordenador, usando el programa adecuado de *SIMFIT* y un ejemplo que ilustre los aspectos matemáticos introducidos. Estas sesiones se recogen en la Tabla 1, donde puede apreciarse cómo los temas siguen un orden de dificultad creciente, con el fin de llevar al alumno de los conceptos básicos a los más avanzados. A continuación se comentan brevemente estas sesiones.

Ajuste de curvas por regresión lineal y no lineal. En primer lugar se abordan todos los conceptos relativos al ajuste de curvas, analizando las ecuaciones algebráicas habituales de una variable y varios parámetros, así como los conceptos de linealidad y no linealidad de una ecuación respecto a los parámetros. Se hace hincapié en que una cosa es que una ecuación sea "no lineal en las variables" (su representación y-x no sigue una linea recta) y otra que sea "no lineal en los parámetros" (considerada la "x" como constante la dependencia de "y" con los parámetros no se puede expresar como combinación de sumas y restas); por ejemplo, el polinomio cuadrático ($y=a+bx+cx^2$) sería no lineal en las variables pero lineal en los parámetros (a,b,c), mientras que una monoexponencial ($y=Ae^{-kx}$) sería no lineal tanto en las variables como en los parámetros

(*A,k*). Después de esto, se introducen los fundamentos de la regresión lineal y no lineal por mínimos cuadrados. Se explica que, en términos genéricos, el modelo de regresión de una variable dependiente "*y*" frente a una variable independiente "*x*" es de la forma:

$$y = f(x, \underline{\theta}) + \varepsilon \qquad [1]$$

donde "$\underline{\theta}$" es un vector de parámetros desconocidos (θ_1 , θ_2 , θ_3 ,..... θ_P) y "ε" es el error experimental. Según esto, dados "n" pares de observaciones (x_1,y_1), (x_2,y_2), (x_3,y_3),... (x_n,y_n), el método de los mínimos cuadrados consiste en calcular el valor de los parámetros $\underline{\theta}$ que minimiza la suma de los residuales al cuadrado. Pero, para que el ajuste tenga las propiedades estadísticas adecuadas, es necesario que los errores "ε_i" sigan una distribución normal de media cero (medidas no sesgadas) y varianza constante (medidas con la misma precisión). Si la varianza de los errores no es constante hay que utilizar la estrategia de los mínimos cuadrados ponderados (regresión lineal con pesos estadísticos), que consiste en calcular los valores de los parámetros "$\underline{\theta}$" que minimizan la siguiente función:

$$WSSQ = \sum w_i \left(y_i - f(xi, \underline{\theta}) \right)^2 \qquad [2]$$

donde "w_i" son los respectivos pesos estadísticos que se toman igual al inverso de la varianza ($1/s_i^2$), siendo "s_i" la desviación estándar de cada valor "y_i" obtenida normalmente a partir de réplicas, y donde *WSSQ* significa **"**suma ponderada de los residuales al cuadrado**"** (*weighted sum of squares*). Los métodos numéricos para calcular los parámetros **"**$\underline{\theta}$" que minimizan *WSSQ* dependen de la forma de la función "*f* " respecto a los parámetros "$\underline{\theta}$". Si **"***f* **"** es lineal en los parámetros, la solución es única, exacta y sus propiedades estadísticas están bien establecidas; en este caso el método se denomina "regresión lineal". Pero si **"** *f* **"** es no lineal en los parámetros,

entonces la solución es aproximada, ha de obtenerse por métodos iterativos y sus propiedades estadísticas son también aproximadas; denominándose a esta técnica "regresión no lineal". Los métodos iterativos de la regresión no lineal son básicamente de dos tipos: de búsqueda directa y de gradiente. Entre los de búsqueda directa destaca el "método de búsqueda al azar" (*random search*), que consiste en ir probando valores de los parámetros hasta encontrar aquellos valores que dan un menor WSSQ. Para el caso de 2 parámetros (θ_1 y θ_2), este método se podría esquematizar según el diagrama de la Figura 2-A. En esta figura se aprecia bien la importancia que tiene establecer correctamente los límites entre los que se pueden mover los parámetros, así como el disponer de una estima inicial de los parámetros para iniciar la búsqueda. Por su parte, los métodos de gradiente se basan en las derivadas de la función WSSQ respecto a los parámetros y los mas utilizados son: máximo descenso (*steepest descent*), Gauss Newton, Leverberg-Marquard y Quasi-Newton. La diferencia entre un método y otro se basa en la forma de calcular la dirección de búsqueda "u" y la longitud del paso "λ", como se esquematiza en la Figura 1-B para el caso particular de dos parámetros.

Bondad de un ajuste. Se explican en el curso varios criterios para analizar la bondad de un ajuste. Unos se refieren a los residuales: valor del sumatorio de los residuales al cuadrado, varianza del ajuste, coeficiente de correlación, distribución gráfica de los residuales (al azar, con rachas), test estadístico de las rachas y test de los signos. Otros se refieren a los parámetros: varianza de los parámetros, coeficientes de variación, límites de confianza y test "t" de redundancia de un parámetro (su valor es tan próximo a cero que puede despreciarse). Todos estos criterios son exactos para la regresión lineal, mientras que en la regresión no lineal se aceptan como aproximados (Bardsley *et al.* (1995)). Por otra parte, cuando se analiza un sistema, lo normal es que

se dude entre modelos alternativos dentro de una secuencia jerárquica (1 exponencial, 2 exponenciales, 3 exponenciales...), por lo que se explican también algunos criterios para discriminar entre modelos, entre ellos el test estadístico "F", que valora si es o no significativa la mejora que experimenta habitualmente *WSSQ* al pasar de una ecuación de menos parámetros a otra con más parámetros. Otro criterio es la inspección visual de la superposición de los ajustes respectivos sobre los datos experimentales, analizando si estos datos se distribuyen al azar a ambos lados de la curva ajustada (buen ajuste) o si se agrupan en rachas (mal ajuste).

Vistos los aspectos fundamentales del ajuste de curvas, se pasa a explicar el interés que presentan ciertas ecuaciones en ciencias experimentales, así como el procedimiento para ajustarlas e interpretarlas.

Ajuste de una recta por regresión lineal y calibración. La ecuación de una recta tiene gran interés en ciencia, como es el caso de la ley de Beer en espectrofotometría. Habitualmente, se dispone de una serie de datos de la variable independiente (asumidos sin error) y de las correspondientes respuestas de la variable dependiente (cuyo error se suele determinar como la desviación estandar de 4 ó 5 réplicas). Normalmente, se hace un ajuste de regresión lineal con pesos estadísticos para estimar la pendiente y la ordenada en el origen, así como sus respectivos límites de confianza al 95%. En muchos casos, es necesario también hacer la predicción inversa (calibración), es decir, medido un valor de la variable "*y*" calcular el valor que le corresponde de la variable "*x*", así como sus límites de confianza.

Ajuste de polinomios. Los polinomios de distinto grado son una herramienta matemática muy utilizada en modelización empírica:

$$y = p_0 + p_1 x + p_2 x^2 + p_3 x^3 + ... + p_n x^n \qquad [3]$$

Debido a su sencillez, se suelen usar para ajustar curvas en general: de calibrado, curvas cinéticas para estimar velocidades iniciales...etc. Conviene hacer énfasis en que los polinomios son demasiado flexibles y pueden conducir a hiperajustes, por lo que conviene quedarse siempre con un ajuste de grado "n" que no sea muy elevado.

Ajuste de funciones exponenciales. Muchos sistemas experimentales en ciencias (Cinética Química, Farmacocinética...) se pueden modelizar con funciones exponenciales, que en términos genéricos adoptan la siguiente forma:

$$f(t) = A_1 e^{-k_1 t} + A_2 e^{-k_2 t} + ... + A_n e^{-k_n t} + C \qquad [4]$$

Dependiendo del valor de los parámetros A_i , k_i y C, esta ecuación puede interpretar curvas muy diferentes, bien de tipo creciente, decreciente e incluso con máximo o mínimo.

Ajuste de ecuaciones de Michaelis-Menten. Una mezcla de isoenzimas independientes que obedeciese cada una a una cinética de Michaelis-Menten se modelizaría con la siguiente ecuación:

$$v = \frac{V_{\max_1}[S]}{K_{m_1} + [S]} + \frac{V_{\max_2}[S]}{K_{m_2} + [S]} + ... + \frac{V_{\max_n}[S]}{K_{m_n} + [S]} \qquad [5]$$

que, para el caso de una única isoenzima, quedaría reducida a la conocida expresión:

$$v = \frac{V_{\max}[S]}{K_m + [S]} \qquad [6]$$

donde el parámetro V_{max} es la velocidad máxima y K_m es la constante de Michaelis. Tradicionalmente, los parámetros de esta ecuación se obtenían ajustando la ecuación de Lineweaver-Burk en dobles inversos:

$$\frac{1}{v} = \frac{1}{V_{\max}} + \frac{K_m}{V_{\max}} \cdot \frac{1}{[S]} \qquad [7]$$

ya que, al tratarse de una línea recta, los parámetros K_m y V_{max} se obtenían fácilmente por regresión lineal. Este procedimiento, sin embargo, ha sido aplicado incorrectamente en muchos casos, ya que el ajuste a la recta se hacía sin utilizar pesos estadísticos para los residuales, lo cual es erróneo como se verá mas adelante.

Ajuste de curvas de crecimiento y de supervivencia. En Biología y Medicina presenta un gran interés el interpretar cómo aumentan o disminuyen los individuos de una población (Dinámica de poblaciones, Epidemiología). Así, en los modelos de crecimiento se relaciona el número de individuos con el tiempo, entre ellos cabe destacar el modelo exponencial ([8]), el monomolecular ([9]) y el logístico ([10]):

$$N(t) = Ae^{kt} \quad [8] \qquad N(t) = A\left(1 - Be^{-kt}\right) \quad [9] \qquad N(t) = \frac{A}{1 + Be^{-kt}} \quad [10]$$

Análogamente, se han propuesto diferentes modelos de supervivencia que relacionan la proporción de individuos que sobrevive con el tiempo transcurrido, entre éstos cabría citar el modelo exponencial, de Weibull y de Gompertz.

Ajuste por tramos de cúbicas (cubic splines). A veces, la forma de la curva es lo suficientemente compleja como para no poder ser ajustada con un polinomio o por ecuaciones algebráicas sencillas, lo cual es frecuente en campos como la Biología o la Ecología, donde los sistemas suelen englobar varios procesos simultáneamente. En estos casos, se utiliza una técnica empírica denominada "ajuste por tramos de cúbicas empalmados por nudos". Esta técnica consiste en dividir el intervalo de los datos en varios subintervalos y en cada subintervalo ajustar una cúbica ($f(x)=a+bx+cx^2+dx^3$).

Estas cúbicas se empalman suavemente en los nudos, para formar una suma de cúbicas que constituye finalmente la ecuación de ajuste. El número de nudos es elegido arbitrariamente por el usuario, por lo que éste debe tener la precaución de no elegir un número excesivo de nudos (no más de 2 ó 3), para no caer en un hiperajuste en el que la curva ajustada pasase por todos los puntos. Este tipo de ajustes se utiliza mucho en calibración y en el suavizado de curvas, ya que, una vez obtenida la suma de cúbicas, se puede utilizar esa función para hallar el área bajo la curva y las tangentes a la curva en ciertos puntos característicos.

Simulación de ecuaciones con datos exactos y con error. Dada una ecuación algebráica como la de la expresión [1], la simulación consiste en suponer unos valores para los parámetros $\underline{\theta}$ y para la variable "x" y generar a partir de ellos unos datos exactos. Posteriormente, y con el fin de simular mejor una experiencia real, se perturban esos datos exactos añadiéndoles un cierto error "ε", generado al azar mediante alguna distribución de probabilidad, normalmente la distribución Gaussiana. Las aplicaciones de estas simulaciones son enormes, ya que permiten explorar numerosas situaciones experimentales sin costo alguno de reactivos, por ejemplo decidir el margen de la variable independiente, el número de puntos a realizar, el espaciado entre los puntos (lineal, logarítmico...), número de réplicas por punto.... También sirven para la validación de modelos y para estudios estadísticos de Monte Carlo.

Ajuste de ecuaciones con 2 o más variables independientes. Este es el caso, por ejemplo, de las inhibiciones enzimáticas, en las que la velocidad de reacción depende tanto de la concentración de inhibidor como de la concentración de sustrato. Así, para una inhibición de tipo no competitivo la forma de la ecuación es la siguiente:

$$v = \frac{V_{\max} \cdot [S]}{K_m\left(1+\frac{[I]}{K_I}\right)+\left(1+\frac{[I]}{K_I}\right)[S]} \qquad [11]$$

Como puede verse, no se trata ahora del ajuste de una curva del tipo v=f([S]), sino del ajuste de una superficie del tipo v=f([S],[I]). Afortunadamente, la metodología es análoga a la ya comentada para ecuaciones de una variable y no suelen aparecer problemas especiales, por lo que es posible determinar los tres parámetros (V_{max}, K_m y K_I) a partir de un único ajuste.

Ajuste de modelos definidos por usuario. Siempre llega el dia en el que un investigador quiere proponer su propio modelo para explicar un determinado proceso. Normalmente son pequeñas variantes de otros modelos ya establecidos, pero en cualquier caso se impone que el paquete estadístico permita escribir la ecuación del modelo con una determinada sintaxis y que disponga de una rutina general que sea capaz de ajustar la ecuación a nuestros datos experimentales. En el curso, se explica el método desarrollado por *SIMFIT* para suministrar estas ecuaciones definidas por el usuario. Consiste en escribir un archivo de texto ASCII, en el que se va declarando el número de variables, el número de parámetros y una serie de instrucciones secuenciales (*add, substract, multiply,divide, log ...*) con las operaciones matemáticas que figuren en la ecuación. Una vez que se ha escrito el archivo, se llama al programa *QNFIT* de *SIMFIT*, que lee por un lado la ecuación del usuario y por otro los datos experimentales, luego pide al usuario que elija las estimas iniciales y los límites para los parámetros y por último realiza al ajuste y nos proporciona los parámetros de la ecuación ajustada.

Integración y ajuste de ecuaciones diferenciales. Cuando se trata de una ecuación diferencial simple, lo habitual es que sea de una variable independiente, de una variable

dependiente y de varios parámetros; en este caso el problema se suele reducir a integrar analíticamente la ecuación diferencial y a realizar el ajuste con base en la ecuación integrada correspondiente. Así ocurre, por ejemplo, con la cinética de orden uno:

$$-\frac{d[A]}{dt} = k[A] \quad \Rightarrow \quad [A] = [A]_0 \cdot e^{-kt} \qquad [12]$$

Si se trata de varias ecuaciones diferenciales simultáneas, el caso más frecuente es el de un "sistema de ecuaciones diferenciales ordinarias", en el que sólo hay una variable independiente (normalmente el tiempo), varias variables dependientes (concentraciones, número de animales...) y diferentes parámetros (Toby and Toby (1999)). Un ejemplo de estos sistemas sería la cinética de las reacciones consecutivas en Química o la dinámica de poblaciones depredador-presa en Biología. En muchos de estos sistemas ya no es fácil encontrar la solución analítica exacta, por lo que una alternativa sería la siguiente: a) integrar numéricamente las ecuaciones diferenciales (métodos de Runge-Kutta, Gear, Adams) utilizando diferentes parámetros y condiciones iniciales, b) superponer las curvas integradas a los datos experimentales para ver si coinciden entre sí, c) terminar de ajustar las ecuaciones a los datos experimentales mediante un algoritmo que combine simultáneamente la integración y el ajuste.

Una vez terminados los aspectos teóricos relativos al ajuste de curvas y la modelización matemática, se pasa a estudiar en el curso lo que se suele conocer como estadística aplicada. Algunos de los temas tratados se citan a continuación.

Estadística descriptiva, curvas de frecuencia y funciones de probabilidad. Cuando se dispone de los datos de un muestreo, siempre es aconsejable hacer un análisis descriptivo. Lo habitual es determinar la media, la desviación estándar, mediana,

cuartiles...etc, también es costumbre representar los datos en forma de histograma o como distribución acumulativa en escalones, por último se suele hacer un test de Shapiro-Wilks para probar si los datos siguen una distribución normal o Gaussiana.

Comparación de 2 medias con pruebas paramétricas y no paramétricas. Si se dispone de 2 muestras, lo primero sería comprobar si ambas siguen una distribución normal (test de Shapiro-Wilks) y si tienen la misma varianza (test F). Si se cumplen estas dos condiciones, lo recomendable es aplicar el test "t de Student", con el fin de probar si las medias son estadísticamente iguales o diferentes. Sin embargo, cuando alguna de las muestras tiene pocos datos y además éstos no siguen una distribución normal, conviene utilizar alguna prueba no paramétrica como el test "U de Mann-Whitney".

Comparación de n medias por análisis de la varianza (ANOVA). Esta técnica se utiliza cuando hay que comparar más de 2 muestras. Se suele usar cuando las muestras siguen distribuciones normales con la misma varianza y se quiere comprobar si todas las muestras tienen la misma media. Cuando el ANOVA sugiere que al menos un par de muestras (entre las "n") difieren significativamente, entonces se debe aplicar el test "Q de Tukey" para detectar cuáles son esas muestras. El método ANOVA no se debe utilizar cuando las muestras no presentan la misma varianza, en este caso hay dos alternativas: a) hacer alguna transformación de los datos (raiz cuadrada, log, arcsen...) para estabilizar las varianzas, b) aplicar el test no paramétrico de Kruskal-Wallis.

Contenido de las sesiones prácticas

Los casos prácticos que se analizan en el curso aparecen recogidos en la Tabla 2. y están disponibles en http://simfit.usal.es. Cada caso consta de unos datos

experimentales, del correpondiente planteamiento teórico que orienta el tipo de análisis a realizar y de un procedimiento "paso a paso" que describe todos los detalles del camino a seguir. A continuación, se describen los objetivos de estos casos y se presentan, a modo de ejemplo, algunas de las pantallas que aparecen en el ordenador.

Se comienza con unos casos sencillos en los que, tomando como base un calibrado colorimétrico de fosfato, se le proporcionan al alumno 6 datos puntuales de absorbancia a distintas concentraciones de fosfato patrón (caso 1), con el fin de que ajuste una recta a esos puntos por regresión lineal sin pesos estadísticos. En el caso2 se le pide que repita el ajuste anterior, pero ahora con unos pesos estadísticos correspondientes a un error relativo constante del 5 % ($W_i=1/(0.05y_i)^2$). Por último (caso3) se le proporcionan unos nuevos datos de calibrado pero ahora con 4 réplicas a cada concentración de fosfato, de modo que pueda realizar un ajuste de regresión lineal con pesos estadísticos calculados a partir de las réplicas ($W_i=1/s_i^2$), y obtenga así la correspondiente recta de calibrado y las curvas de confianza al 95% (Fig. 3-sup. izq.). Por otra parte, el alumno calcula la concentración de fosfato en dos muestras en las que se ha medido la absorbancia (predicción inversa o calibración), así como las horquillas de confianza al 95% para dichas concentraciones.

A continuación, se aborda el análisis de la cinética de desintegración del Radon (caso 4), con el fin de que el alumno ajuste una monoexponencial a los datos (ecuación [12]) mediante regresión no lineal, asumiendo para los pesos un error relativo constante del 5% . El resultado es el característico de una cinética de orden 1 (Fig. 3-sup. dcha.) y el valor obtenido para la constante de velocidad es de $7.45\ 10^{-3} \pm 0.28\ 10^{-3}\ h^{-1}$, donde el ± se refiere a los límites de confianza al 95% calculados según la expresión habitual:

$$p_i = \pm t_{(n-m,\alpha/2)}\sqrt{\mathrm{var}(p_i)} \qquad [13]$$

donde "*t*" es la "t de Student" con "*n*-*m*" grados de libertad (*n*=nº de puntos, *m*=parámetros) y riesgo "α" (normalmente 0.05) y donde var(p_i) es la varianza del parámetro.

En el caso 5 se analiza la cinética de hidrólisis del p-nitrofenil fosfato por fosfatasa alcalina de E. Coli. El objetivo aquí es ajustar la ecuación de Michaelis-Menten ([6]) a unos datos v-[S] con 5 réplicas de velocidad a cada concentración de sustrato, de manera que el alumno pueda seguir practicando con la técnica de regresión no lineal con pesos estadísticos. En el caso siguiente (caso 6), se ajusta la ecuación de Lineweaver-Burk en dobles inversos ([7]) a los mismos datos que los del caso 5, usando ahora una regresión lineal sin pesos estadísticos. El alumno observa enseguida que este ajuste no proporciona los valores de V_{max} y K_m correctos, los que se habían obtenido por regresión no lineal a la ecuación directa en el caso 5. Sin embargo, cuando se ajusta esa misma ecuación doble inversa, pero utilizando los siguientes pesos estadísticos obtenidos en base a las leyes de propagación del error:

$$s_i(1/v_i) = \frac{s_i(v_i)}{v_i^2} \quad \Rightarrow \quad w_i = \frac{1}{(s_i(1/v_i))^2} = \frac{v_i^4}{(s_i(v_i))^2} \qquad [14]$$

entonces el resultado obtenido sí que es el mismo. Es decir, se trata de demostrar al alumno que lo más aconsejable es ajustar siempre las ecuaciones directas por regresión no lineal y no sus transformaciones lineales por regresión lineal. Sin embargo, a efectos gráficos, se pueden representar los datos y la curva ajustada (por regresión no lineal) tanto en el espacio directo como en la linealización de Lineweaver-Burk (Fig. 3-inf. izda.).

El caso 7 supone un salto cualitativo importante, ya que se trata de discernir si en una mezcla de 2 isoenzimas tienen ambas el mismo comportamiento cinético (misma V_{max} y K_m) o tienen comportamiento distinto (existen unos parámetros V_{max1} y K_{m1} para la isoenzima 1 y V_{max2} y K_{m2} para la isoenzima 2). Para ello, se ajustan en secuencia la ecuación genérica [5] con grado 1 y con grado 2, utilizando regresión no lineal con pesos estadísticos procedentes de réplicas. El alumno va comparando la bondad de los dos ajustes para discriminar cuál es el mejor de ellos. Al final, comprueba que hay evidencias acerca que el grado 1 es insuficiente para explicar los datos y que el mayor número de parámetros que supone el grado 2 estaría justificado estadísticamente. Entre esas evidencias, cabe destacar la inspección visual de cómo se superponen las curvas ajustadas a los puntos experimentales (Fig. 3-inf. dcha.), observándose que la curva de grado 2 se superpone mejor a los puntos que la de grado 1, tanto en el espacio directo (v-[S]) como en la representación de Eadie-Hofstee (v-(v/[S])).

El caso 8 consiste en encontrar la ecuación que mejor se ajuste a los datos de crecimiento de la levadura *Yarrowia lipolytica* en n-hexadecano. Concretamente, se dispone de la absorbancia del cultivo a 600 nm a distintos tiempos de incubación, existiendo en cada caso 5 réplicas de cada medida. El alumno va probando, mediante regresión no lineal con pesos estadísticos, el ajuste de diferentes ecuaciones usadas habitualmente para las curvas de crecimiento ([8],[9],[10]), llegando a la conclusión que el mejor ajuste es el que se logra con la ecuación logística ([10]), como se aprecia en la figura 4-sup. izda.

El caso 9 aborda el problema de cómo ajustar curvas cinéticas concentración-tiempo con el fin de estimar las velocidades iniciales y las asíntotas finales (metodología análoga a la del caso 8). Por su parte, el caso 10, introduce al alumno en

una técnica de ajuste totalmente diferente, es el método de los tramos cúbicos empalmados (*cubic splines*). Se trata de hacer una de curva de calibrado del peso de unos animales, siguiendo el patrón en función de su edad. Como la forma de los datos es de tipo sigmoideo y el calibrado sólo tiene fines empíricos, se opta por un ajuste de tramos cúbicos que da buenos resultados. El caso 11 ofrece al alumno la posibilidad de seguir investigando estos ajustes por tramos cúbicos, pero ahora sobre un problema más complejo; se trata de comparar cómo varía el número de células inflamatorias con el tiempo de cicatrización tanto en ratas tratadas como en ratas control (datos supuestos, no reales). El problema se aborda ajustando ambas curvas por tramos cúbicos (Fig. 4-sup. dcha.) y luego comparando el área bajo cada curva (*ABC*) mediante el llamado porcentaje de diferencia entre curvas:

$$\%Diferencia = \frac{Integral|curva1 - curva2|}{ABC1 + ABC2} \qquad [15]$$

El caso 12 introduce al alumno en las técnicas de simulación de experimentos, primero generando datos exactos a partir de una ecuación y luego perturbando dichos datos por adición de errores aleatorios basados en una distribución de probabilidad elegida por el usuario (normal, exponencial, uniforme , de Cauchi).

El caso 13 consiste en encontrar el mejor modelo para interpretar la cinética plamática de digoxina administrada por vía intravenosa. El alumno sabe que los modelos mas frecuentes en Farmacocinética son: el bicompartimental, formado por un comportamiento central (plasma) y un comportamiento periférico (órganos) y el tricompartimental, que es análogo al anterior pero con un segundo comportamiento periférico adicional (tejidos profundos). Ambos modelos siguen ecuaciones poliexponenciales decrecientes del tipo de la expresión [4], por lo que el alumno

compara el ajuste a 2 exponenciales (modelo bicompartimental) con el de 3 exponenciales (tricompartimental) y decide cuál de los dos es el más adecuado para sus datos de digoxina.

El caso 14 supone un avance muy importante, ya que aborda el análisis conjunto de unos datos de velocidad frente a concentración de sustrato a distintas concentraciones de inhibidor. Habitualmente, esta situación se realiza analizando separadamente los datos v-[S] a cada concentración fija de inhibidor ([I]), pero en este caso se le pide al alumno que analice todos los datos conjuntamente. Para ello, ha de realizar sucesivos ajustes con el programa *QNFIT* a las funciones de dos variables v([S],[I]) que describen la inhibición competitiva, no competitiva o acompetitiva. Al comparar los tres ajustes, llega a la conclusión que el mejor de ellos es el que proporciona la inhibición no competitiva (ecuación [11]), cuya representación es la que aparece en la figura 4 inf. izda..

El caso 15 es un ejercicio en el que los alumnos escriben funciones sencillas definidas por usuario, comprueban que cumplen las reglas de sintaxis de *SIMFIT* y las ajustan a datos simulados. El caso 16 supone otro paso adelante, ya que aborda la integración y ajuste de ecuaciones diferenciales. Como ejemplo se utilizan las ecuaciones de Lotka-Volterra, que se integran primero con diferentes parámetros y condiciones iniciales, con el fin de investigar su comportamiento, para ajustarlas finalmente a unos datos supuestos de un depredador y su presa (Fig. 4-inf. izda).

Los últimos casos (17, 18 y 19) se refieren a diferentes aspectos de estadística aplicada (ver Tabla 2), que al ser más conocidos no parece interesante describirlos aquí.

Resultados y conclusiones

La presión por celebrar los cursos nos ha impedido evaluar los resultados a través de encuestas realizadas a los alumnos. A modo de apreciación subjetiva, podríamos resumir nuestra experiencia en las siguientes conclusiones:

1. El análisis de datos es una materia cada día más necesaria en las carreras de ciencias experimentales. No obstante, las enseñanzas de esta disciplina suelen estar dispersas en diferentes asignaturas, por lo que los contenidos y la nomenclatura no suelen estar unificados.

2. Una manera de armonizar estas enseñanzas podría ser el utilizar un paquete estadístico como alternativa a las hojas de cálculo y a los programas matemáticos, lo que ayudaría a que los alumnos se centraran más en los métodos estadísticos en sí que en los aspectos informáticos de los programas.

3. *SIMFIT* es un paquete estadístico amigable, aunque avanzado, que incluye la mayoría de las técnicas habituales de ajuste de curvas y de estadística. Su uso suele persuadir a los alumnos que, para hacer un buen análisis de datos, no basta con "apretar un botón" sino que hace falta entender muy bien lo que se está haciendo.

4. *SIMFIT* es un software académico gratuito, copias que los profesores pueden distribuir libremente entre sus alumnos, y así asignarles trabajos para realizar en casa o en pequeños grupos.

Agradecimientos:

A los profesores José Martínez Lanao y Ramón Ardanuy Albajar, por su inestimable ayuda en la celebración de los cursos de análisis de datos en la Universidad de Salamanca. We also acknowledge the support given to the SIMFIT project by the University of Manchester (U. K.) and by Salford Software, whose FTN95 compiler has been used to compile the Windows version of *SIMFIT*.

Bibliografía

Bardsley W. G., Bukhari N. A. J., Ferguson M. W. J., Cachaza J. A. and Burguillo F. J. (1995) "Evaluation of model discrimination, parameter estimation and goodness of fit in nonlinear regression problems by test statistics distributions". *Computers in Chemistry* **19**, 75-84.

Bruce, J.J.; Anderson, B.D. and Bruce D. (1993) "Investigating the harmonic oscillator using Mathematica". *J. Chem. Educ.* **70**, A122.

Denton, P. (2000) "Analysis of First Order Kinetics Using Microsoft Excel Solver". *J. Chem. Educ.* **77**, 1524-1525.

Ferreira, M.M.C.; Ferreira Jr., W.C., Lino, C.S. and Porto, M.E.G. (1999) "Uncovering Oscillations, Complexity and Chaos in Chemical Kinetics Using Mathematica". *J. Chem. Educ.* **76**, 861.

Guare, Ch. J. (1991) "Error, Precision and Uncertainty". *J. Chem. Educ.* **68**, 649-652.

Harris, D.C. (1998) “Nonlinear Least-Squares Curve Fitting with Microsoft Excel Solver”. *J. Chem. Educ.* **75**, 119-121.

Kim M-H., Ly, S-Y. and Hong, T-K. (2000) “Comparisons and Demostrations of Scientific Calulators”. *J. Chem. Educ.* **77**, 1367-1370.

Lieb, S.G. (1997) “Simplex Method of Nonlinear Least-Squares. A logical Complementary Method to Linear Least-Squares Analysis of Data”. *J. Chem. Educ.* **74**, 1008.

Malinowski, E.R. (1995) “A computer program for Calculating Standard Deviations from Standard Devaitions. *J. Chem. Educ.* **72**, 1079-1082.

Tellinghuisen, J. (2000) “Nonlinear Least-Squares Using Microcomputer Data Analysis Programs: KaleidaGraph in the Physical Chemistry Teaching Laboratory”. *J. Chem. Educ.* **77**, 1233-1239.

Toby, S. and Toby, F.S. (1999) “The Simulation of Dinamic Systems”. *J. Chem. Educ.* **76,** 1584-1590.

Treptow, R.S. (1998) “Precision and Accuracy in Measurements”. *J. Chem. Educ.* **75**, 992-995.

Young, S. H.; Madura, J.D.; Wierzbicki, A. (1995) “Integration of Numerical Methods into the Undergraduate Physical Chemistry Curriculum Using Mathcad”. *J. Chem. Educ.* **72**, 606.

Zielinski, T.J. (2000) “Symbolic Software in the Chemistry Curriculum”. *J. Chem. Educ.* **77**, 668-670.

Zielinski, T.J. (1995) "Promoting Higher-Order Thinking Skills: Uses of Mathcad and Classical Chemical Kinetics to Foster Student Development". *J. Chem. Educ.* **72**, 631

Received 9.04.2002, accepted 7.08.2002

Tabla 1. Contenido de las sesiones teóricas del curso

- ***Ajuste de curvas por regresión lineal y no lineal***
- Bondad de un ajuste y discriminación entre modelos
- Ajuste de rectas y polinomios, curvas de calibrado
- Ajuste de funciones exponenciales
- Ajuste de ecuaciones de Michaelis Menten
- Curvas de crecimiento y de supervivencia
- Ajuste por tramos de cúbicas (cubic splines)
- Simulación de ecuaciones con datos exactos y con error
- Ajuste de ecuaciones con 2 o más variables
- Ajuste de modelos definidos por usuario
- Integración y ajuste de ecuaciones diferenciales
- Estadística descriptiva, histogramas y curvas de frecuencia
- Comparación de 2 medias con pruebas paramétricas y no paramétricas
- Comparación de n medias por análisis de la varianza (ANOVA)

Tabla 2. Contenido de las sesiones prácticas del curso

- ***Recta de calibrado para análisis de fosfato (casos 1-3)***
- Cinética de desintegración del Radon (caso 4)
- V_{max} y K_m de la fosfatasa alcalina de *E. coli*. (casos 5 y 6).
- V_{max} y K_m en una mezcla de 2 isoenzimas (caso 7)
- Ajuste de la curva de crecimiento de la levadura *Y. lipolytica* (caso 8)
- Cálculo de velocidades iniciales en Cinética Química
- Calibrado del peso de un animal con la edad (Caso 10)
- Comparación de curvas, areas bajo curvas y tangentes a curvas (Caso 11)
- Simulación de experimentos con datos exactos y con error (caso 12)
- Modelización de la farmacocinética de digoxina (caso 13)
- Análisis de la inhibición de una enzima (caso 14)
- Escribiendo y ajustando funciones definidas por usuario (caso 15)
- Ecuaciones diferenciales de Lotka-Volterra para depredador-presa(caso 16)
- Estadística descriptiva de la altura de 25 jovenes (caso 17)
- Comparación de 2 medias: altura de hombres y mujeres (Caso 18)
- ANOVA: crecimiento de semillas en 3 medios nutricionales (caso 19)

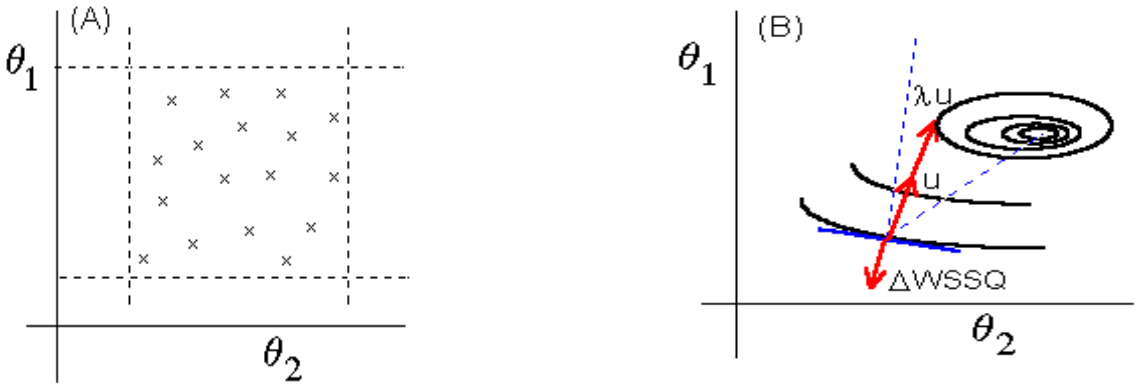
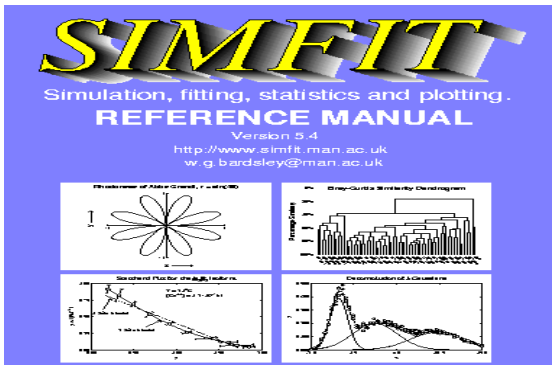
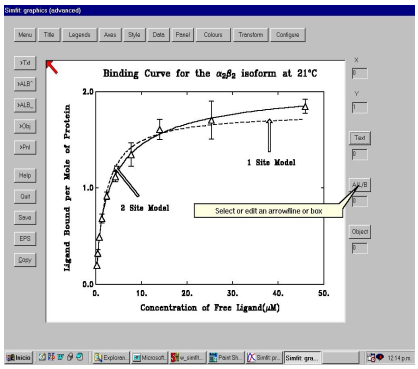


Figura 1. Métodos iterativos de optimización en regresión no lineal para el caso de 2 parámetros θ_1 y θ_2:
(A) de búsqueda al azar, (B) de gradiente, donde "ΔWSSQ" es el gradiente de la suma ponderada de los residuales al cuadrado, "u" es la dirección de búsqueda y "λ" es la longitud del paso.





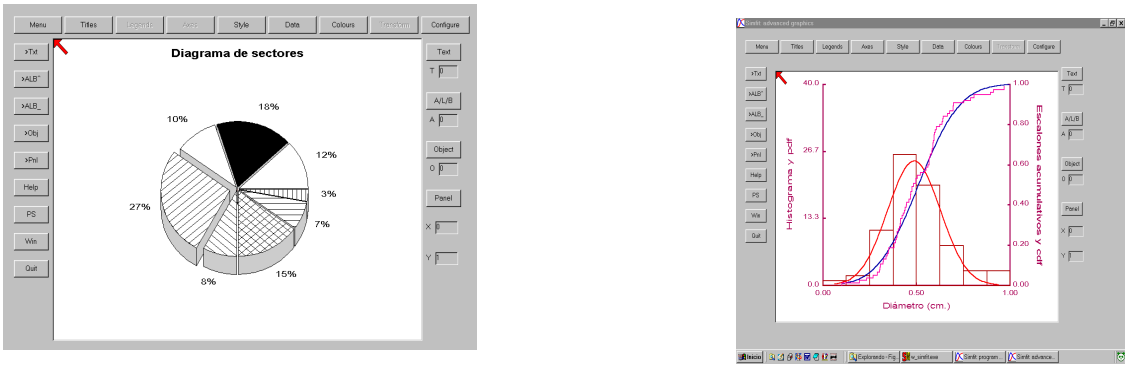
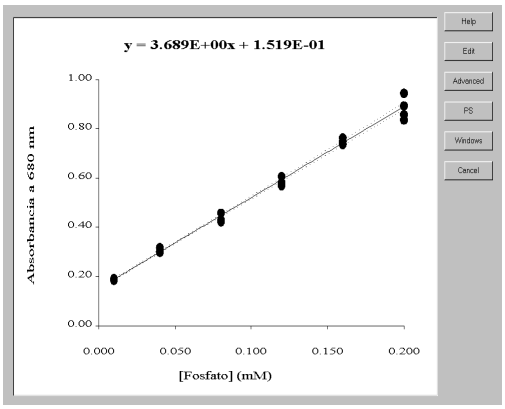
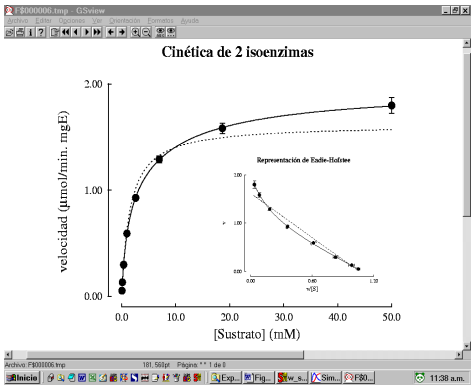


Figura 2. Algunas pantallas ilustrativas del paquete estadístico *SIMFIT* . De izquierda a derecha y de arriba a abajo: manual de referencia en formato PDF; un ajuste por regresión no lineal; diagrama de sectores; histograma y curvas de frecuencia en una estadística descriptiva.





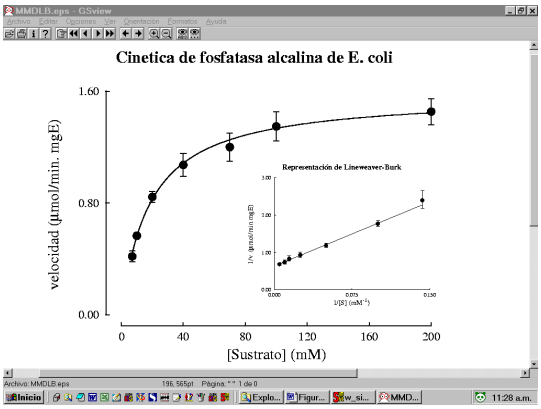
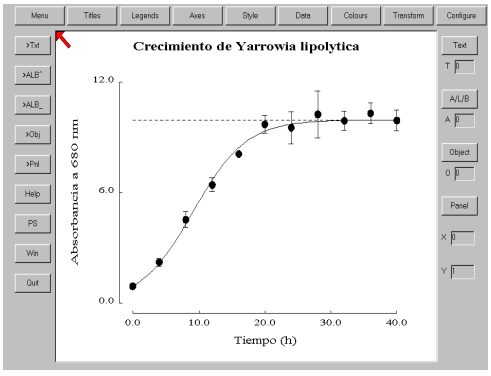
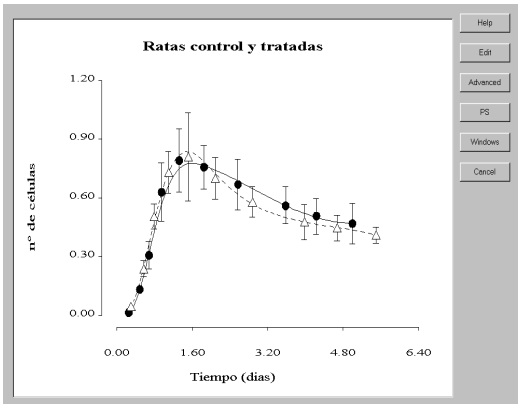
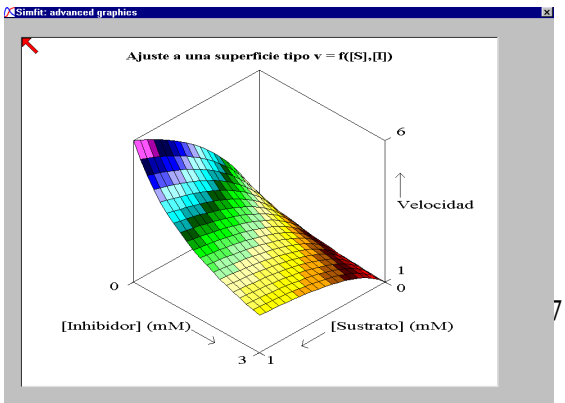


Figura 3. De izquierda a derecha y de arriba a abajo: recta de calibrado para fosfato; cinética de desintegración del Radon ajustada a una monoexponencial; cinética de la fosfatasa alcalina de *E. coli* ajustada a una ecuación de Michaelis-Menten; cinética de 2 isoenzimas donde se compara el ajuste a 1 y 2 términos de Michaelis-Menten.







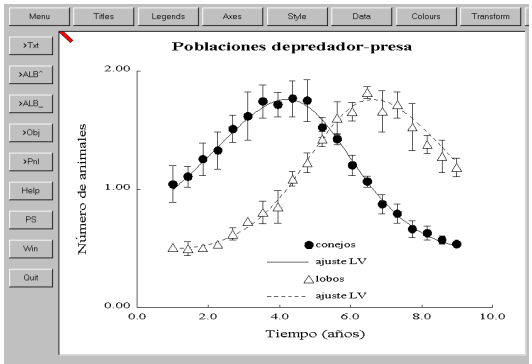


Figura 4. De izquierda a derecha y de arriba a abajo: ajuste logístico a la curva de crecimiento de *Yarrowia lipolytica*; ajuste y comparación de curvas por "cubic splines"; ajuste de una superficie v([S[,[I]) a datos de una inhibición enzimática; ajuste de las ecuaciones diferenciales de Lotka-Volterra a datos depredador-presa.