ITEA (1999). Vol. 95A N.° 2, 131-142

AJUSTE Y COMPARACIÓN DE CURVAS DE CRECIMIENTO

A. Blasco*
L. Varona**

* Departamento de Ciencia Animal
Universidad Politécnica de Valencia
Apartado 22012
46071 Valencia
España

** UDL - IRTA
Rovira Roure, 177
Lleida
España

RESUMEN

En este artículo se delimitan las hipótesis que se utilizan en los modelos de crecimiento aplicándolo al caso de la función de Gompertz. Se proponen dos formas de ajuste para esta función, siguiendo los procedimientos de las dos principales escuelas de inferencia, la frecuentista y la bayesiana, y se discute la comparación de curvas. Aunque el artículo está centrado en la curva de Gompertz, las conclusiones, así como los procedimientos, son inmediatamente generalizables a cualquier curva de crecimiento.

Palabras clave: Curvas de crecimiento, Ajustes no lineales, Gompertz.

SUMMARY
FITTING AND COMPARISON OF GROWTH CURVES

In this paper, the underlying hypothesis of growth models are discussed, using the Gompertz growth curve. Two ways of fitting this curve are proposed, according to the procedures of the main inference schools: frequentist and bayesian. Curves comparison is also discussed. Although focused in the Gompertz growth curve, the procedures and results exposed in the paper can be applied to all growth curves.

Key words: Growth curves, Non linear adjustment, Gompertz.

Introducción

El ajuste de curvas de crecimiento es complicado cuando se utiliza la estadística clásica, y más complicado aún es la comparación entre dos curvas de crecimiento. Habitualmente hay que realizar hipótesis más o menos violentas para ajustar las curvas, hipótesis que en muchas ocasiones se admiten de forma implícita, y a veces sin ser conscientes de las consecuencias. La comparación de curvas es un problema complejo de estadística no lineal. Una revisión de los distintos tipos de curvas, la significación biológica de sus parámetros y

sus problemas de ajuste se encuentra en RICHARDS (1969). En este artículo se pretende delimitar con claridad cuáles son las hipótesis que se utilizan en los modelos de crecimiento y proponer dos formas de ajuste, siguiendo las dos principales escuelas de inferencia, la frecuentista y la bayesiana. Para no hablar en términos generales, difíciles en muchos casos de imaginar, nos ceñiremos, por su complejidad, a una de las curvas de crecimiento más usadas en ganadería, la de Gompertz, probando que los resultados pueden ser generalizados a cualquier curva.

Notación

Mayúsculas negrita, (**A**) con o sin subíndices ($\mathbf{G}_p$), son matrices.

Minúsculas negrita, (**u**) con o sin subíndices ($\mathbf{u}_a$), son vectores.

y: vector columna.

y': vector fila.

El resto son escalares.

El modelo

Tomemos la función de Gompertz, probablemente la más usada en la descripción del crecimiento de aves, cerdos y conejos. En el modelo, cada individuo i tiene una curva de crecimiento distinta. Cada individuo dispone de varios datos, tomados cada cierto tiempo j (por ejemplo, cada semana).

$$y_{ij} = a_i \bullet \exp\left[-b_i \bullet \exp\left(-k_i \bullet t_j\right)\right] + \varepsilon_{ij} \qquad (1)$$

Sobre los parámetros a_i, b_i, k_i, actúan efectos ambientales (estación, temperatura, tamaño de la camada en la que nació el individuo, sexo, etc.) que pueden estar correlacionados. Cada individuo tiene un valor genético para cada parámetro, que lógicamente está correlacionado con el valor genético de sus parientes. Si representamos por **a**, **b**, **k**, los vectores con los parámetros de todos los individuos,

$$\mathbf{a} = \mathbf{X}\beta_a + \mathbf{Z}\mathbf{u}_a + \mathbf{e}_a$$

$$\mathbf{b} = \mathbf{X}\beta_b + \mathbf{Z}\mathbf{u}_b + \mathbf{e}_b$$

$$\mathbf{k} = \mathbf{X}\beta_k + \mathbf{Z}\mathbf{u}_k + \mathbf{e}_k$$

Para evitar innecesarias complicaciones, asumimos que las matrices de diseño **X**, **Z** son las mismas (esto es, suponemos que si hay un efecto ambiental como el de estación, éste afecta a los tres parámetros, aunque sea de forma diferente), lo que permite simplificar la notación:

$$\mathbf{p} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

donde,

$\mathbf{p}' = [\ \mathbf{a}'\ ,\ \mathbf{b}'\ ,\ \mathbf{k}'\ ]$ es el vector de parámetros,

$\beta' = [\ \beta'_a\ ,\ \beta'_b\ ,\ \beta'_k\ ]$ es el vector de efectos ambientales sistemáticos,

$\mathbf{u}' = [\ \mathbf{u}'_a\ ,\ \mathbf{u}'_b\ ,\ \mathbf{u}'_k\ ]$ es el vector de efectos genéticos,

$\mathbf{e}' = [\ \mathbf{e}'_a\ ,\ \mathbf{e}'_b\ ,\ \mathbf{e}'_k\ ]$ es el vector de efectos aleatorios,

el vector **e** no incluye los errores de ajuste, que son los ε_{ij} de la ecuación (1), sino que pretende reflejar el hecho de que las curvas de crecimiento de los animales pueden estar afectadas por factores aleatorios no considerados en el modelo. Esto es distinto a que un dato concreto tomado sobre un animal en un momento dado se aleje de la curva de crecimiento de este animal debido a factores aleatorios relacionados con

el muestreo (por ejemplo, se toma el dato justo cuando acaba de comer el animal).

Desde un punto de vista frecuentista se considera β como un conjunto de efectos fijos, pero desde un punto de vista bayesiano se debe intentar encontrar una función que represente la opinión que se tiene a priori en función de la información previa disponible. Una forma sencilla de hacer esto es suponer que

$$\beta \sim N(\mathbf{m}_\beta, \mathbf{V}_\beta)$$

donde $\mathbf{m}_\beta$, $\mathbf{V}_\beta$ son subjetivas e intentan dar lugar a una función Normal que represente el estado de creencias previas. Dada la dificultad de hacerlo en el caso multivariante, habitualmente $\mathbf{V}_\beta$ será una matriz diagonal (ver BLASCO, 1998 para una discusión sobre este punto).

Se considera que $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$

donde,

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_a & \mathbf{G}_{ab} & \mathbf{G}_{ak} \\ & \mathbf{G}_b & \mathbf{G}_{bk} \\ & & \mathbf{G}_k \end{bmatrix} = \begin{bmatrix} \mathbf{A}\sigma^2_{Ga} & \mathbf{A}\sigma_{Gab} & \mathbf{A}\sigma_{Gak} \\ & \mathbf{A}\sigma^2_{Gb} & \mathbf{A}\sigma_{Gbk} \\ & & \mathbf{A}\sigma^2_{Gk} \end{bmatrix} = \mathbf{A} \otimes \mathbf{G}_p$$

donde **A** es la matriz de parentesco de los individuos, y $\mathbf{G}_p$ la matriz de varianzas-covarianzas genéticas de los parámetros a, b, k,

$$\mathbf{G}_p = \begin{bmatrix} \sigma^2_{Ga} & \sigma_{Gab} & \sigma_{Gak} \\ & \sigma^2_{Gb} & \sigma_{Gbk} \\ & & \sigma^2_{Gk} \end{bmatrix}$$

Se considera que $\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})$

donde

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_a & \mathbf{R}_{ab} & \mathbf{R}_{ak} \\ & \mathbf{R}_b & \mathbf{R}_{bk} \\ & & \mathbf{R}_k \end{bmatrix} = \begin{bmatrix} \mathbf{I}\sigma^2_{Ra} & \mathbf{I}\sigma_{Rab} & \mathbf{I}\sigma_{Rak} \\ & \mathbf{I}\sigma^2_{Rb} & \mathbf{I}\sigma_{Rbk} \\ & & \mathbf{I}\sigma^2_{Rk} \end{bmatrix} = \mathbf{I} \otimes \mathbf{R}_p$$

donde **I** es la matriz de identidad, y $\mathbf{R}_p$ la matriz de varianzas-covarianzas de los efectos aleatorios no sistemáticos de los parámetros a, b, k.

$$\mathbf{R}_p = \begin{bmatrix} \sigma^2_{Ra} & \sigma_{Rab} & \sigma_{Rak} \\ & \sigma^2_{Rb} & \sigma_{Rbk} \\ & & \sigma^2_{Rk} \end{bmatrix}$$

Esto es, se considera que los efectos aleatorios no sistemáticos que actúan sobre a, b, k, pueden estar relacionados entre sí para cada individuo, pero que no hay correlaciones entre efectos de distintos individuos; es decir, que para un individuo i

$$r(e_{ai}, e_{bi}) \neq 0 \; ; \; r(e_{ai}, e_{ki}) \neq 0 \; ; \; r(e_{ki}, e_{bi}) \neq 0$$

mientras que para dos individuos i, j,

$$r(e_{ai}, e_{aj}) = 0 \; ; \; r(e_{ai}, e_{bj}) = 0 \; ; \text{ etc.}$$

Es razonable suponer que habrá una cierta correlación entre parámetros, por ejemplo, si uno representa la tasa de crecimiento (k) y otro el peso adulto (a) es difícil suponer que no estén relacionados. Si se quiere suponer que hay correlaciones entre individuos, es menester incluir un nuevo efecto aleatorio que las considere, aunque no hay motivos para pensar que deban existir.

Por tanto, se considera que,

$$(\mathbf{p} \mid \beta) \sim N(\mathbf{X}\beta \, , \, \mathbf{Z'GZ} + \mathbf{R})$$

$$(\mathbf{p} \mid \beta, \mathbf{u}) \sim N(\mathbf{X}\beta + \mathbf{Zu} \, , \, \mathbf{R}) \qquad (2)$$

Se considera, finalmente, que los errores de estimación ε_{ij} se distribuyen de forma Normal con media cero y varianza que va aumentando con el tiempo. La razón es que la varianza está ligada a la media del carácter: cuando los animales son jóvenes, todos los datos que se recogen se agrupan en torno a valores pequeños, y conforme los animales crecen la dispersión aumenta por un efecto de escala. Al llegar al estado

adulto, las medidas sucesivas que se toman tienen la misma dispersión. Si hemos considerado que la varianza aumenta con la media y ésta última lo hace con arreglo a una curva de crecimiento, podemos considerar que la varianza o que la desviación típica también aumentan con arreglo a una curva de crecimiento. De hecho el modelo es muy flexible, la mayor o menor linearidad de la pendiente que describe el crecimiento de la varianza estará determinada por los parámetros de la curva. Por tanto,

$$\varepsilon \sim N(\mathbf{0}, \mathbf{V})$$

donde,

$$\mathbf{V} = \begin{bmatrix} \mathbf{D} & \mathbf{0} & \dots & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{D} & & & . \\ . & & . & & . \\ . & & & . & . \\ \mathbf{0} & \mathbf{0} & \dots & \dots & \mathbf{D} \end{bmatrix} \qquad (3)$$

donde **D** es una matriz diagonal correspondiente a los datos de un individuo,

$$\mathbf{D} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & & . \\ . & & . & . \\ 0 & \dots & \dots & \sigma_j^2 \end{bmatrix}$$

donde,

$$\sigma_j = a_\varepsilon \bullet \exp\left[-b_\varepsilon \bullet \exp\left(-k_\varepsilon \bullet t_j\right)\right]$$

Es decir, se considera que los errores de estimación no están relacionados ni entre individuos ni entre dos medidas consecutivas de un individuo. Aunque es obvio que hay autocorrelación entre dos medidas sucesivas de un individuo, ésta ha sido tenida en cuenta al incluir efectos ambientales y genéticos en los parámetros, por lo que los errores no deben estar autocorrelacionados. Si hubiera otras fuentes de autocorrelación, deberían ser tenidas en cuenta en el modelo de los parámetros a, b, k. Si no es posible, porque se considera que hay efectos ambientales similares entre dos medidas sucesivas que pueden no estar recogidos en el modelo y que afectan al error de estimación, puede modificarse **D** para incluir covarianzas entre medidas sucesivas. Cómo hacerlo es un tanto arbitrario, pero suele suponerse que dos medidas sucesivas están correlacionadas con una correlación ρ, por lo que la primera y la tercera medida lo está con una correlación ρ^2, la primera y la cuarta con ρ^3, etc. Si las medidas no son tomadas de forma regular en el tiempo, hay que tenerlo en cuenta al definir la estructura de los errores.

Se considera también que las desviaciones típicas del error aumentan con el tiempo con arreglo a una curva de Gompertz de parámetros a_ε, b_ε, k_ε, y que son las mismas para todos los individuos. Para simplificar la notación, llamaremos

$$\mathbf{p'}_\varepsilon = [a_\varepsilon, b_\varepsilon, k_\varepsilon]$$

no consideraremos que sobre estos parámetros ejerzan su influencia factores genéticos ni ambientales sistemáticos.

Se desea estimar:

1) Los parámetros a_i, b_i, k_i, de cada individuo.

2) Los parámetros a, b, k, medios de la población.

3) Los parámetros genéticos de a, b, k (esto es, sus heredabilidades y correlaciones genéticas).

4) La diferencia entre parámetros a_i, b_i, k_i de dos individuos o entre los parámetros a, b, k medios de dos poblaciones.

5) Los efectos ambientales que se ejercen sobre los parámetros a, b, k.

6) Los parámetros a_ε, b_ε, k_ε que describen la evolución de la desviación típica del error de ajuste con el tiempo.

Solución frecuentista

1. Problemas de estimación

Ajustar esta curva presenta las siguientes dificultades:

1.1. *No existe una representación lineal de la curva*

Aunque se tomen logaritmos para hacer desaparecer un exponente, queda siempre una parte exponencial. Esto significa que o bien debe ajustarse mediante regresión no lineal o bien debe utilizarse alguna aproximación lineal, normalmente basada en una serie de Taylor. En cualquier caso, los ajustes no lineales se resuelven también por aproximaciones lineales usando habitualmente series de Taylor.

1.2. *En el caso de linealizar tomando logaritmos, el tratamiento de error de ajuste obliga a hipótesis forzadas*

Pongamos un ejemplo más simple, ajustar la curva exponencial

$$y = a \cdot \exp(-b \cdot t) + \varepsilon$$

pero esto impide linealizar la ecuación tomando logaritmos, por lo que se suele proponer el modelo,

$$y = a \cdot \exp(-b \cdot t) \cdot \exp(\varepsilon) = a \cdot \exp(-b \cdot t + \varepsilon)$$

que es fácilmente linealizable y se ajusta sin problemas por regresión simple,

$$\ln y = \ln a - b \cdot t + \varepsilon$$

pero que obliga a considerar que los errores son multiplicativos y exponenciales.

1.3. *Se desconocen las distribuciones de las estimas de los parámetros. No es posible comparar curvas de crecimiento*

Se sabe que los estimadores de los coeficientes de regresión se distribuyen como t-Student, pero en una regresión no lineal no se sabe qué distribución tienen los estimadores de los coeficientes a, b, k de la curva de crecimiento, lo que tiene la fatal consecuencia de *impedir la comparación de curvas*. Si se hacen transformaciones nos encontramos con el problema de que al deshacerlas, el antilogaritmo del error estándar de un coeficiente no es el error estándar del coeficiente. Por ejemplo, en la ecuación de Brody, una vez estimados b y [ln a], se desea saber el valor de a y para ello se toma el antilogaritmo de [ln a], pero el error estándar de a no se obtiene hallando el antilogaritmo del error estándar de [ln a].

1.4. *La corrección por efectos sistemáticos es oscura o imposible*

Se supone que el hecho de nacer en una camada numerosa, o de pertenecer a uno u otro sexo, afectan al peso adulto y a los parámetros b y k de forma distinta. Cómo hacer estas correcciones es oscuro, y habitualmente se ha preferido trabajar con datos precorregidos que no han tenido en cuenta que el efecto sistemático puede afectar de forma diferente a uno u otro parámetro de la curva, aparte de no considerar los errores cometidos en la precorrección en el resultado final.

1.5. *No se tiene en cuenta el parentesco, lo que altera la estructura de los errores*

Por métodos clásicos no está claro cómo incluir el efecto del parentesco en el ajuste a cada parámetro. Los modelos de regresión aleatoria, en la que cada parámetro es (o lleva asociada) una variable aleatoria

pretenden resolver parcialmente el problema, pero se encuentran la dificultad de que en ese caso los parámetros cambian de valor en función de la información disponible (i.e.: si hay pocos datos su valor se aproxima a cero). Es frecuente que en el ajuste de curvas de crecimiento no todos los individuos tengan datos hasta el final de la toma de medidas, por lo que hay parámetros mejor estimados que otros. Esto no es un problema para la estimación de una curva media pero sí para la estimación de curvas individuales.

1.6. *No se tiene en cuenta la estructura de los errores debida a causas no genéticas*

No se suele tener en cuenta, por ejemplo, la autocorrelación ocasionada por medir a un individuo en periodos de tiempo consecutivos o el hecho de que la varianza del error aumenta con la edad hasta estabilizarse al llegar al estado adulto. Para abordar este último problema se puede utilizar una regresión dando un peso proporcional a la inversa de la varianza en cada momento de medida, pero no se incluye en el modelo el error cometido al estimar los pesos.

1.7. *La estimación de los parámetros genéticos de los coeficientes se realiza por métodos que no optimizan la información*

La mayor parte de autores estiman las curvas individuales y realizan un análisis genético de las estimaciones, con lo que pierden la información debida al parentesco y no incluyen el error de estimación de estos parámetros a, b, k en el error de los parámetros genéticos.

2. *Soluciones propuestas*

2.1. *Regresión lineal*

La solución más sencilla consiste en intentar encontrar una forma lineal de la curva y aplicar un programa de regresión lineal. En la mayor parte de las curvas esto es posible (véase el ejemplo del apartado anterior), pero no así en la función de Gompertz. Para poder linealizar esta función se ha sugerido (RICHARDS, 1969) aproximar, mediante una serie de Taylor,

$$\exp(-kt) \cong 1-kt$$

pero la aproximación exponencial deteniéndose en el primer término de la serie de Taylor es decididamente mala.

Incluso en el caso en el que las funciones se puedan linealizar, permanecen las dificultades expuestas en los apartados 2 y 3 del punto anterior.

2.2. *Regresión no lineal*

Hay programas de regresión no lineal (por ejemplo, en el paquete SAS), que permiten ajustar directamente la función de Gompertz. La mayor parte de soluciones se basan en linealizar la función aproximando mediante una serie de Taylor, por lo que los paquetes suelen requerir las derivadas respecto a los parámetros, aunque en ocasiones el paquete las aproxima mediante algún método de cálculo numérico.

En ocasiones los ajustes presentan problemas de no convergencia, pero estos problemas suelen ir ligados a la indefinición de alguno de los parámetros. Por ejemplo, en la función de Richards (RICHARDS, 1969) hay un parámetro que determina el punto de inflexión de la curva. Si en el periodo central del crecimiento los datos siguen una recta (como es usual en ganadería), ese parámetro está mal definido.

2.3. *Consideración de los efectos sistemáticos*

Respecto a los efectos sistemáticos, la solución más habitual es no considerarlos,

aunque a veces se realizan precorrecciones a los datos, por ejemplo por el tamaño de camada al nacimiento. El problema es que entonces no se trabaja con los datos sino con residuos de la corrección, con lo que se generan varios problemas: los parámetros de la curva están afectados de forma diferente por la corrección, lo que no se ha podido tener en cuenta; los errores de esta corrección no se tienen en cuenta al estimar los parámetros de la curva; finalmente, como se trabaja no con verdaderos residuos sino, lógicamente, con sus estimaciones, los valores que se obtienen al estimar los parámetros son distintos de los que se obtendrían de haber podido realizar la corrección de residuos y la estimación simultáneamente. Así las cosas, hay varias soluciones posibles:

a) *Cuando hay un efecto claro sobre los parámetros, por ejemplo el efecto sexo o el efecto raza*: Lo más sencillo es ajustar las curvas por separado haciendo desaparecer el problema. En el caso del efecto tamaño de camada al nacimiento, dado que no sería práctico separar las curvas por camadas, puede hacerse una precorrección a los datos, aunque muchas veces no será necesario, bien porque el objetivo es comparar el crecimiento de dos grupos de individuos cuyo tamaño de camada no es distinto, bien porque interesa conocer la situación real del crecimiento de esos grupos, incluyendo el hecho de que sus tamaños de camada sean diferentes.

b) *Cuando diversos niveles del efecto actúan sobre los parámetros, por ejemplo el efecto de estación*. En estos casos nos encontramos con datos de crecimiento tomados en varias estaciones. En general lo mejor es ignorar estos efectos por las complicaciones de corrección y de interpretación que traen. Como cada medida está tomada a una edad distinta, si se desea precorregir los datos habrá que hacerlo estimando el efecto de estación medida a medida (por ejemplo, semana de vida a semana de vida), puesto que el efecto de estación en las primeras semanas de vida es obvio que no es el mismo que en las últimas. Para poder hacer esta precorrección haría falta disponer de todas las medidas (todas las semanas, por ejemplo) en todas las estaciones. Ignorarlo tampoco debe conducir a problemas graves, ya que el crecimiento o el peso adulto no se van a modificar porque exista un efecto de estación en una semana concreta, puesto que son estimados con el conjunto de medidas. El problema sólo es serio si la mayor parte de los animales de un grupo crecieron en invierno y la mayor parte del otro en verano, pero esto es un problema de diseño que tiene mala solución en cualquier caso. Otro problema que puede ocurrir en especies de crecimiento rápido es que se tomen los datos de las primeras medidas en invierno y de las últimas en verano, lo que conduce a un cierto bandeo de la curva de crecimiento. De nuevo nos encontramos con un problema de diseño, aunque en especies de cría intensiva los efectos de estación se ven muy minimizados por las instalaciones.

2.4. *Consideración de la estructura de los errores*

Hay diversas aproximaciones posibles:

a) *No considerar la estructura de los errores*. La estima de los parámetros sigue siendo insesgada, pero el error de estimación es mayor de lo que los cálculos ofrecen. Hay una tendencia entre los estadísticos a no dar relevancia a los problemas de heterocedasticidad de la varianza, lo que habitualmente es razonable, pero en los ajustes de curvas de crecimiento hay que notar que de las primeras medidas a las últimas el peso se puede multiplicar por cien,

por lo que la heterocedasticidad de las varianzas es realmente grande.

b) *Dar un peso a los errores proporcional a la inversa de la varianza en cada momento de medida (en cada semana, por ejemplo).* Los pesos se suelen calcular a partir de los propios datos. A veces se usan las inversas de las varianzas estimadas en cada momento de medición (en cada semana, por ejemplo), lo que debido a la escasez de datos suele conducir a incoherencias, como por ejemplo a que la varianza de los datos en una cierta edad sea menor que la de los datos tomados algunas semanas antes. A veces se intenta hilar más fino encontrando alguna ley subyacente, como hicimos nosotros al exponer el modelo. Como es cierto que la no consideración de que las varianzas son distintas no conduce a ninguna situación grave, salvo extremos como el no considerarla en absoluto, el no tratar de hallar una ley general y usar las varianzas de los datos sin más, no debe conducir a ningún problema.

c) *Consideración de la autocorrelación.* Hay software que permite introducir ciertas reglas en la construcción de la matriz de los errores de estimación ε_{ij}. El programa PROC MIXED del paquete estadístico SAS permite modificar la estructura de los errores con arreglo a un conjunto amplio de reglas del estilo de la expuesta anteriormente al hablar del modelo.

d) *Consideración de las relaciones genéticas.* Es posible ajustar un modelo mixto en el que los parámetros a, b, k tengan un componente aleatorio. El programa PROC MIXED, en unión con el macro NLMIX (de dominio público) permite además añadir la matriz de correlaciones entre efectos aleatorios; esto es, la de relaciones genéticas. El ajuste es de todas formas complejo, y no está garantizada la convergencia.

2.5. *Análisis de los parámetros genéticos de la curva*

Hasta la fecha el análisis se ha venido desarrollando en dos etapas, primero se han ajustado los parámetros de la curva para cada individuo y posteriormente se ha realizado un análisis genético de esos parámetros estimados (ver, p.ej., KACHMAN *et al.*, 1988), con lo que no se ha resuelto el problema expuesto en el punto 1.7. No es imposible encontrar una solución máximo verosímil basada en un modelo mixto como el descrito en el punto anterior (2.4), y así ha sido propuesta por (ZUCKER *et al.*, 1995), pero la cantidad de parámetros involucrados hace que esta propuesta sea extremadamente compleja de llevar a cabo.

Solución bayesiana

1. La solución analítica

1.1. *El problema*

Desde un punto de vista bayesiano, el problema consiste en encontrar la función de densidad posterior de los parámetros que se desea estimar, dados los datos. Esto es, encontrar

$$f(\mathbf{p}, \beta, \mathbf{u}, \mathbf{G}, \mathbf{R}, \mathbf{p}_\varepsilon \mid \mathbf{y})$$

La forma de hacerlo ha sido descrita por VARONA *et al.* (1997). Aplicando el teorema de Bayes,

$$f(\mathbf{p}, \beta, \mathbf{u}, \mathbf{G}, \mathbf{R}, \mathbf{p}_\varepsilon \mid \mathbf{y}) =$$
$$= f(\mathbf{y} \mid \mathbf{p}, \beta, \mathbf{u}, \mathbf{G}, \mathbf{R}, \mathbf{p}_\varepsilon) \cdot$$
$$\cdot f(\mathbf{p}, \beta, \mathbf{u}, \mathbf{G}, \mathbf{R}, \mathbf{p}_\varepsilon) / f(\mathbf{y})$$

1.2. *Cálculo de la verosimilitud*

En el modelo (ecuación 1), si se fija $\mathbf{p}$, β, $\mathbf{u}$, $\mathbf{G}$, $\mathbf{R}$, $\mathbf{p}_\varepsilon$ lo único que queda variable es el error de estimación ε_{ij}, por lo que

$$\mathbf{y} \sim N(\mathbf{m}, \mathbf{V})$$

donde **V** es la matriz de varianzas covarianzas de los errores, expuesta en (3), y en **m** se encuentran los valores m_{ij} que se obtendrían para cada momento de medida t_j en la ecuación (1) (sin el error ε_{ij}) al sustituir los valores de los parámetros a_i, b_i, k_i de cada animal, puesto que estos parámetros están fijados.

Como los errores ε_{ij} son independientes, y como la densidad conjunta de variables independientes es el producto de sus densidades, la expresión que queda es bastante simple:

$$f(\mathbf{y} \mid \mathbf{p}, \beta, \mathbf{u}, \mathbf{G}, \mathbf{R}, \mathbf{p}_\varepsilon) =$$

$$= \prod_i \prod_j f(y_{ij} \mid a_i, b_j, k_i, \sigma^2_{\varepsilon ij}) =$$

$$= \prod_i \prod_j \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp\left[\frac{(y_{ij} - m_{ij})^2}{\sigma^2_{ij}}\right]$$

1.3. *Cálculo de las funciones de densidad a priori*

$$f(\mathbf{p}, \beta, \mathbf{u}, \mathbf{G}, \mathbf{R}, \mathbf{p}_\varepsilon) = f(\mathbf{p}, \beta, \mathbf{u}, \mathbf{G}, \mathbf{R})\, f(\mathbf{p}_\varepsilon)$$

ya que $\mathbf{p}_\varepsilon$ no depende del resto de parámetros. Utilizando reglas de probabilidad,

$$P(A,B) = P(A|B)\, P(B)$$

En nuestro caso,

$$f(\mathbf{p}, \beta, \mathbf{u}, \mathbf{G}, \mathbf{R}) =$$
$$= f(\mathbf{p} \mid \beta, \mathbf{u}, \mathbf{G}, \mathbf{R})\, f(\beta, \mathbf{u}, \mathbf{G}, \mathbf{R})$$

$$f(\beta, \mathbf{u}, \mathbf{G}, \mathbf{R}) = f(\beta, \mathbf{u} \mid \mathbf{G}, \mathbf{R})\, f(\mathbf{G}, \mathbf{R}) =$$
$$= f(\beta, \mathbf{u} \mid \mathbf{G}, \mathbf{R})\, f(\mathbf{G})\, f(\mathbf{R})$$

ya que las varianzas genéticas no están relacionadas con las ambientales. Es posible considerar dependencia entre ambos componentes de varianza en la distribucion a priori (WEISS *et al.*, 1997), aunque un modelo así nunca ha sido utilizado en mejora genética animal. En principio no hay razones por las que los dos tipos de componentes de varianza deban estar relacionados.

Supondremos también que los efectos sistemáticos son independientes de los aleatorios. En ocasiones se sabe con certeza que esto no es así (por ejemplo, cuando se evalúan toros lecheros se sabe que las mejores granjas, las que mejor ambiente procuran a las vacas, son las que más invierten en genética y traen semen de mejores animales), pero aquí no tenemos motivos para suponer que no puedan ser independientes, salvo errores en el diseño del experimento. Así pues,

$$f(\beta, \mathbf{u} \mid \mathbf{G}, \mathbf{R}) = f(\beta)\, f(\mathbf{u} \mid \mathbf{G})$$

Con lo que finalmente queda,

$$f(\mathbf{p}, \beta, \mathbf{u}, \mathbf{G}, \mathbf{R}, \mathbf{p}_e) = f(\mathbf{p} \mid \beta, \mathbf{u}, \mathbf{G}, \mathbf{R})\, f(\beta)$$
$$f(\mathbf{u} \mid \mathbf{G})\, f(\mathbf{G})\, f(\mathbf{R})\, f(\mathbf{p}_\varepsilon)$$

donde,

$$f(\mathbf{p} \mid \beta, \mathbf{u}, \mathbf{G}, \mathbf{R}) \sim N(\mathbf{X}\beta + \mathbf{Z}\mathbf{u}, \mathbf{R}) =$$
$$= N(\mathbf{X}\beta + \mathbf{Z}\mathbf{u}, \mathbf{I} \otimes \mathbf{R}_p)$$

$f(\beta) \sim N(\mathbf{m}_\beta, \mathbf{V}_\beta)$ según discutimos al hablar del modelo

$$f(\mathbf{u} \mid \mathbf{G}) \sim N(\mathbf{0}, \mathbf{G}) = N(\mathbf{0}, \mathbf{A} \otimes \mathbf{G}_p)$$

$f(\mathbf{G})$ y $f(\mathbf{R})$ deberían construirse en función de las creencias previas proporcionadas por la información previa disponible. Como no es posible establecer estrictamente estas creencias, suele recurrirse a alguna función que con pocos parámetros pueda cambiar de forma adaptándose al estado más o menos vago de creencias previas. Por conveniencias matemáticas suele usarse un función conjugada, en este caso una distribucion Whishart invertida, aunque hay

muchas otras soluciones posibles. Esta función depende de dos parámetros, una matriz de escala y otro parámetro al que equivocadamente se le denomina "grados de libertad" o "grados de credibilidad"; ambos parámetros modifican la forma de la función.

Es prácticamente obligado renunciar a describir el estado de creencias previo multivariante puesto que es difícil de definir, ya que habría que describir la opinión previa al experimento sobre cada conjunto de valores posible; esto es, en cada punto del espacio (ver BLASCO, 1998) para una discusión sobre este punto). Se puede, sin embargo, definir el estado de creencias para cada variable por separado utilizando una matriz de escala diagonal, puesto que en ese caso se obtienen chi-cuadrados invertidas que son fácilmente representables. En cualquier caso se debe intentar realizar los análisis variando los valores de estos parámetros para examinar hasta qué punto la opinión previa es importante en el resultado final.

$f(\mathbf{p}_\varepsilon) = \mathbf{c}$, un vector de constantes dentro de límites admisibles del espacio paramétrico para asegurar la propiedad de la distribución posterior conjunta. Es decir, los posibles valores de a_ε tienen todos la misma probabilidad a priori, con ciertos límites para evitar que la función sea impropia, y lo mismo podemos decir de b_ε y de k_ε. Con esto pretendemos reflejar un estado de incertidumbre sobre los valores de estos parámetros.

1.4. *Cálculo de* f(y)

El cálculo de $f(\mathbf{y})$ no es estrictamente necesario cuando lo que se desea es hallar la moda (el valor más probable) de la densidad posterior, puesto que como la densidad posterior no depende de **y** (está condicionada a **y**), $1/f(\mathbf{y})$ es una constante de proporcionalidad que se puede ignorar al buscar el máximo de la densidad posterior. Sin embargo si se desea hallar la media o usar el sistema de intervalos de confianza para la inferencia científica, hay que conocer exactamente la densidad posterior, por lo que hay que conocer el valor de $f(\mathbf{y})$. Este valor es difícil de computar porque

$$f(\mathbf{y}) = \int f(\mathbf{y}, \mathbf{p}, \beta, \mathbf{u}, \mathbf{G}, \mathbf{R}, \mathbf{p}_\varepsilon)\, f(\mathbf{p}, \beta, \mathbf{u}, \mathbf{G}, \mathbf{R}, \mathbf{p}_\varepsilon)\, d(\mathbf{p}, \beta, \mathbf{u}, \mathbf{G}, \mathbf{R}, \mathbf{p}_\varepsilon)$$

como **y** es un vector, esta integral es multidimensional. Aunque los elementos de **y** fueran independientes, esta integral pasaría a ser una integral múltiple de tantas dimensiones como datos, y en ambos casos el problema es irresoluble incluso por métodos numéricos. Si sólo se está interesado en las distribuciones marginales, la constante de integración es unidimensional y el problema es resoluble mediante integración numérica. De todas formas, hoy en día las técnicas de muestreo de Gibbs, que se mencionan en el apartado siguiente, han resuelto ambos problemas.

2. Modus operandi

Finalmente tenemos que

$$f(\mathbf{p}, \beta, \mathbf{u}, \mathbf{G}, \mathbf{R}, \mathbf{p}_\varepsilon \mid \mathbf{y}) =$$

$$= f(\mathbf{y} \mid \mathbf{p}, \beta, \mathbf{u}, \mathbf{G}, \mathbf{R}, \mathbf{p}_\varepsilon)\, f(\mathbf{p} \mid \beta, \mathbf{u}, \mathbf{G}, \mathbf{R})\, f(\beta)$$

$$f(\mathbf{u}|\mathbf{G})\, f(\mathbf{G})\, f(\mathbf{R})\, f(\mathbf{p}_\varepsilon) / f(\mathbf{y}) \qquad (4)$$

que es un producto de funciones Normales o Whishart invertida. Se puede intentar hallar el máximo de esa función, lo que dará la solución más probable, o hallar la media, lo que dará la solución que minimiza el riesgo cuadrático. En el primer caso nos encontramos con un problema de la misma envergadura que el de hallar soluciones máximo verosímiles: el número de

parámetros y la complejidad de las funciones hacen de ésta una tarea muy compleja. En el segundo caso habría que integrar la función y hallar la constante de proporcionalidad, lo que tampoco parece viable.

Una solución propuesta recientemente es la de extraer muestras al azar de la función de densidad posterior para representar aproximadamente esta función de densidad. Las inferencias se hacen a partir de los puntos muestreados de la densidad posterior multivariante. Creando histogramas o dibujando las función de densidad marginales a partir de esos puntos, se puede obtener una estima de la moda; la media de esos valores es una estima de la media de la densidad posterior; y finalmente, ordenándolos, se puede obtener la mediana con facilidad, así como cualquier intervalo de confianza.

El problema se centra ahora en cómo obtener esos valores *tomados al azar* de la función de densidad posterior. En el caso multivariante no es posible computacionalmente –al menos de momento– muestrear directamente de la función de densidad posterior, y hay que transformar el problema en univariante, o en problemas de menos dimensiones, mediante técnicas de muestreo de Gibbs. Las técnicas de muestreo de Gibbs, basadas en el muestreo de las funciones condicionales de la densidad posterior, permiten, además, eludir el cálculo de f(**y**) (ver SORENSEN, 1997) para una amplia revisión). Se trata, pues, de extraer muestras al azar de las funciones

$$f(\mathbf{p}|\,\beta, \mathbf{u}, \mathbf{G}, \mathbf{R}, \mathbf{p}_\varepsilon, \mathbf{y}),\ f(\beta|\,\mathbf{p}, \mathbf{u}, \mathbf{G}, \mathbf{R}, \mathbf{p}_\varepsilon, \mathbf{y}),\ f(\mathbf{u}|\,\beta, \mathbf{p}, \mathbf{G}, \mathbf{R}, \mathbf{p}_\varepsilon, \mathbf{y}),$$

$$f(\mathbf{G}|\,\mathbf{p}, \beta, \mathbf{u}, \mathbf{R}, \mathbf{p}_\varepsilon, \mathbf{y}),\ f(\mathbf{R}|\,\mathbf{p}, \beta, \mathbf{u}, \mathbf{G}, \mathbf{p}_\varepsilon, \mathbf{y}),\ f(\mathbf{p}_\varepsilon|\,\mathbf{p}, \beta, \mathbf{u}, \mathbf{G}, \mathbf{R}, \mathbf{y})$$

Para iniciar el proceso se toman valores arbitrarios de β, **u**, **G**, **R**, $\mathbf{p}_\varepsilon$, con ellos:

1) se muestrea al azar un valor de **P** en la primera función $f(\mathbf{p}|\,\beta, \mathbf{u}, \mathbf{G}, \mathbf{R}, \mathbf{p}_\varepsilon, \mathbf{y})$,

2) con ese valor y los **u**, **G**, **R**, $\mathbf{p}_e$ arbitrarios de antes se muestrea al azar un valor de β en la siguiente función $f(\beta|\,\mathbf{p}, \mathbf{u}, \mathbf{G}, \mathbf{R}, \mathbf{p}_\varepsilon, \mathbf{y})$,

3) con esos dos valores de **p** y β se muestrea un valor al azar de **u** en la tercera función $f(\mathbf{u}|\,\beta, \mathbf{p}, \mathbf{G}, \mathbf{R}, \mathbf{p}_\varepsilon, \mathbf{y})$, y así sucesivamente hasta que se tienen valores al azar de β, **u**, **G**, **R**, $\mathbf{p}_\varepsilon$, momento en el que se reinicia el ciclo muestreando un valor al azar de **p** en la primera función. Al cabo de varios ciclos, los valores muestreados pertenecen a la densidad posterior $f(\mathbf{p}, \beta, \mathbf{u}, \mathbf{G}, \mathbf{R}, \mathbf{p}_\varepsilon | \mathbf{y})$.

Para poder aplicar estas técnicas es menester saber cómo muestrear al azar de las funciones condicionales. Es sencillo escribirlas, puesto que de (4) se trata de tomar como constante todo aquello que está condicionado. Por ejemplo, para el caso de $f(\mathbf{p}|\,\beta, \mathbf{u}, \mathbf{G}, \mathbf{R}, \mathbf{p}_\varepsilon, \mathbf{y})$ se escribiría de forma explícita (4), se tomaría como constantes β, **u**, **G**, **R**, $\mathbf{p}_\varepsilon$ y como variable **p**, y se intentaría ver si esa función es de alguna forma conocida (Normal, Wishart u otra) de la que hayan algoritmos para extraer valores al azar. Haciendo esto se obtiene que

$f(\beta|\,\mathbf{p}, \mathbf{u}, \mathbf{G}, \mathbf{R}, \mathbf{p}_\varepsilon, \mathbf{y})$, $f(\mathbf{u}|\,\beta, \mathbf{p}, \mathbf{G}, \mathbf{R}, \mathbf{p}_\varepsilon, \mathbf{y})$ son Normales de parámetros conocidos,

$f(\mathbf{G}|\,\mathbf{p}, \beta, \mathbf{u}, \mathbf{R}, \mathbf{p}_\varepsilon, \mathbf{y})$ y $f(\mathbf{R}|\,\mathbf{p}, \beta, \mathbf{u}, \mathbf{G}, \mathbf{p}_\varepsilon, \mathbf{y})$ son Wishart invertidas,

$f(\mathbf{p}_\varepsilon|\,\mathbf{p}, \beta, \mathbf{u}, \mathbf{G}, \mathbf{R}, \mathbf{y})$ y $f(\mathbf{p}|\,\beta, \mathbf{u}, \mathbf{G}, \mathbf{R}, \mathbf{p}_\varepsilon, \mathbf{y})$ no pertenecen a ninguna familia conocida, por lo que para extraer valores al azar de ellas es necesario utilizar técnicas de aceptacion-rechazo (RIPLEY, 1987) o introducir muestreos mediante un algoritmo Metropolis-Hastings de estas variables (ver TANNER, 1993, para una revisión).

Interpretación de resultados

Una vez se dispone de una muestra aleatoria de la densidad posterior, cada punto de la densidad conjunta f(**p**, β, **u**, **G**, **R**, $\mathbf{p}_\varepsilon$ | **y**) pertenece a su vez a cada una de las densidades marginales f(**p**|**y**), f(β|**y**), f(**u**|**y**), f(**G**|**y**), f(**R**|**y**), f($\mathbf{p}_\varepsilon$|**y**) y en cada una de ellas se puede calcular un estimador del parámetro, que suele ser la media de la muestra o la moda (normalmente se usan muestras de 5.000 a 50.000 puntos para minimizar estos errores de estimación, llamados aquí de Monte-Carlo). La precisión viene dada por las regiones de confianza (el equivalente a los intervalos de confianza), calculadas simplemente observando entre qué números alrededor del estimador se hallan el 95% de los puntos de la muestra.

Comparar curvas de crecimiento es sencillo. Supongamos que se quieren comparar las curvas de crecimiento medias de dos grupos de animales. Para ello basta con realizar el análisis de los datos de los dos grupos simultáneamente introduciendo como efecto fijo el efecto de grupo, con lo que tendremos dos niveles para cada parámetro a, b, k. Luego se calcula en cada iteración del proceso Gibbs las diferencias entre los dos niveles de este efecto fijo, y con ello obtenemos puntos de la función de densidad posterior de las diferencias entre grupos para a, b, k. A partir de ahí se actúa como antes: se calcula un estimador o los intervalos de confianza para esa diferencia. También se pueden calcular las medias de los valores aditivos de cada parámetro, y en cada iteración hallar la media de los de cada grupo y restar esas dos medias. Con ello se obtendrían puntos de la función de densidad posterior de las diferencias genéticas entre grupos para a, b, k.

La programación de todas estas técnicas no es compleja, y es de esperar que en un futuro inmediato vaya apareciendo software que resuelva estos problemas con cierta facilidad.

Bibliografía

BLASCO A., 1998. La controversia bayesiana en mejora animal. *ITEA* (94A: 5-42).

KACHMAN S.D., BAKER R.L., GIANOLA D., 1988. Phenotypic and genetic variability of estimated growth curve parameters in mice. *Teor. Appl. Genet.* 76, 148-156.

RICHARDS F.J., 1969. The quantitative analysis of growth. *Plant physiology*. Steward, F.C. (Ed.). Academic Press, 1-76.

RIPLEY B.D., 1987. Stochastic simulation. *Wiley*. New York.

SORENSEN D., 1997. Gibbs sampling in quantitative genetics. *National Institute of Animal Sciences. Internal rapport* N.º 82. Tjele, Dinamarca, 188.

TANNER M.A., 1993. Tools for Statistical Inference. Springer-Verlag.

VARONA L., MORENO C., GARCÍA L.A., ALTARRIBA J., 1997. Multiple Trait genetic analysis of underlying biological variables of production functions. *Lives. Prod. Sci.* 47, 201-209.

WEISS R.E., WANG Y., IBRAHIM J. G. 1997., Predictive model selection for repeated measures random effects models using Bayes Factors. Biometrics 53, 592-602.

ZUCKER D.M., ZERBE G.O., WU M.C., 1995. Inference of the Association between coefficient in a multivariate growth curve model. Biometrics 51, 413-424.

(Aceptado para publicación el 22 de diciembre de 1998)