MODELOS ESTADISTICOS

TRADUCCION Y ADAPTACION DE "ANIMAL BREEDING THEORY" . Internordic licenciat course in quantitative genetics. E.P. Cunningham ed.

Todo análisis estadístico implica un modelo, aunque frecuentemente el modelo puede no estar explicitado. Los modelos utilizados en genética cuantitativa son casi universalmente lineales por lo que nos limitaremos a considerar este tipo.
Definiremos un modelo lineal como una función lineal de ciertos parámetros y variables. Un ejemplo simple es el siguiente:

$$Y_{ij} = \mu + t_i + e_{ij}$$

donde Y_{ij} es una variable observada, μ es una media poblacional fija, t_i el efecto de uno de tres tratamientos (i = 1,…, 3) y e_{ij} es un error aleatorio específico del jésimo individuo en el iésimo tratamiento. Nótese que el modelo carece de sentido hasta que se definen sus partes constituyentes; consideraremos esas definiciones como parte integral de cualquier modelo. Diferentes tipos de modelos son apropiados para diferentes conjuntos de datos y para diferentes usos pudiendo ser clasificados en varias formas. Para nuestros propósitos, adoptaremos la siguiente clasificación:

Modelos de regresión

Se utilizan para definir relaciones funcionales entre variables. En su forma más simple relacionan dos variables:

$$Y_i = \beta_o + \beta_1 X_i$$

donde β_0 y β_1 son parámetros fijos que definen la relación. Esto puede extenderse para tratar relaciones múltiples:

$$Y_i = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots$$

En este caso, una serie de constantes (β_0, β_1,β_2,...) define la relación de Y con las X's. En la mayoría de los datos biológicos en los que se utiliza regresión, se incluye una medida de incertidumbre que requiere un término adicional en el modelo:

$$Y_i = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + e_i$$

Todos los elementos han sido previamente definidos, con excepción de e_i que definiremos como una variable aleatoria específica del iésimo individuo y usualmente llamada término de error. Para la mayoría de nuestros propósitos, definiremos además al e_i como normalmente distribuido, con una varianza determinada y con media = 0, que notaremos:

$$N\left(0, \sigma_e^2\right)$$

Los modelos de regresión son útiles para analizar variación en un conjunto de datos multivariados y también pueden ser usados como ecuaciones predictivas.

Modelos fijos, a efectos fijos o de análisis de la varianza

Pueden considerarse un caso especial de modelos de regresión en el cual las X_i toman un valor de 0 o 1 dependiendo de la presencia o ausencia de determinada condición. La forma usual de un modelo fijo es:

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + e_{ijkl}$$

donde Y_{ijkl} es una variable observada, μ, los α_i, β_j, y γ_k son parámetros fijos y e_{ijkl} es una variable aleatoria, usualmente $N\left(0, \sigma_e^2\right)$. Este modelo describe el efecto de 3 clasificaciones, α, β, y γ sobre la variable observada.

Que la definición de una clasificación está compuesta de efectos fijos equivale a decir que los niveles de la clasificación que ocurren en el experimento son los únicos acerca de los cuales se desea extraer conclusiones. Sean, por ejemplo, 3 α_i representando 3 niveles de proteína en una ración. Entonces, si las α_i son definidas como fijas, esos 3 niveles constituyen la población total de niveles de proteína acerca de los cuales proveerá información el experimento. Uno de los propósitos de cualquier análisis basado en este modelo será el de estimar los parámetros individuales. En este ejemplo particular, las clasificaciones o factores involucrados están dispuestos en forma factorial o de clasificación cruzada. Frecuentemente, un factor puede representar subfactores dentro de otro y a esta clase de arreglos se la denomina anidada o jerárquica. Un ejemplo puede ser:

$$Y_{ijk} = \mu + \alpha_i + \beta_{ij} + e_{ijk}$$

donde Y_{ijk} es una variable observada (e.g. tasa de crecimiento de un cerdo), μ es una media poblacional, los α_i son los efectos de un factor (e.g. el padre del cerdo) y los β_{ij} son los efectos de un segundo factor (e.g. la madre del cerdo) anidado dentro del primero. Los e_{ijk} son términos aleatorios de error.

En un mismo modelo pueden ocurrir arreglos anidados y cruzados. Puede contener además una mezcla de elementos del tipo de un modelo de regresión. El modelo siguiente contiene factores cruzados, anidados y un factor de regresión:

$$Y_{ijkl} = \mu + \alpha_i + \beta_{ij} + \gamma_k + bx_{ijkl} + e_{ijkl}$$

Modelos aleatorios, a efectos aleatorios o de componentes de varianza

Los modelos aleatorios son indistinguibles de los modelos fijos; difieren solamente en la forma en que se definen sus constituyentes. El primer modelo fijo que se mostró, puede considerarse aleatorio como sigue:

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + e_{ijkl}$$

Y_{ijkl} es una variable observada, μ es la media de la población, las α_i son una muestra de efectos de una población de tales efectos con media cero y varianza σ_α^2; los β_j son una muestra de otra población con media cero y varianza σ_β^2; los γ_k son una tercera muestra de

una tercera población con media cero y varianza σ^2_γ y e_{ijkl} es, como antes, una variable aleatoria de una población con media cero y varianza σ^2_e.

Queda claro a partir de estas definiciones que la estimación de parámetros individuales para este tipo de modelo no aporta ningún tipo de información. Como cada clasificación es una muestra aleatoria, otro conjunto de datos puede aportar otra muestra diferente y dar resultados enteramente distintos. Lo que interesa en estos modelos son estadísticos que describan la población parental de la cual se originaron las muestras. La media de cada población es, por definición, igual a cero y así el único parámetro de utilidad es la varianza. El principal uso de los modelos aleatorios es aplicarlos a la estimación de esas varianzas. En el modelo anterior, los α_i pueden representar establecimientos, los β_i una muestra de toros de inseminación artificial y los γ_k una muestra de inseminadores. Si Y_{ijkl} es una medida de fertilidad de las vacas, entonces la varianza en fertilidad de cada una de las poblaciones que dan origen a esas tres muestras, pueden ser estimadas a partir de un análisis basado en ese modelo. Cada una de esas varianzas se conoce como 'componente de varianza'. Los componentes de varianza para poblaciones o grupos de individuos relacionados pueden ser interpretados en términos de variación genética y son el principal método para lograr estimaciones de los parámetros genéticos de una población. Los modelos aleatorios son, por ende, importantes en genética cuantitativa.

Modelos mixtos

Un modelo mixto es aquél compuesto por factores fijos y aleatorios. De hecho, todos los modelos que tratamos son mixtos, ya que cada uno contiene al menos un factor fijo (μ) y un factor aleatorio (e). Sin embargo, la diferencia tiene que ver con los factores que se encuentran entre la media y el término de error. El siguiente es un modelo mixto:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$$

donde Y_{ijk} es una variable observada, μ es la media poblacional, α_i es uno de un limitado número de efectos fijos, β_j es una variable aleatoria proveniente de una población con media cero y varianza σ^2_β y e_{ijk} es una variable aleatoria proveniente de una población con media cero y varianza σ^2_e

La presencia de efectos fijos puede complicar la estimación de los componentes de varianza σ^2_β y σ^2_e y la presencia de efectos aleatorios puede complicar la estimación de los efectos fijos α_i. Debido a ésto, el modelo mixto es, en general, menos deseable para propósitos analíticos que el fijo o el aleatorio. Frecuentemente, sin embargo, un modelo mixto es el único realista para una situación particular. El uso de modelos estadísticos en genética cuantitativa es la base para la estimación de componentes de varianza y efectos fijos de datos biológicos. Esas situaciones pueden, entonces, ser interpretadas en términos de la estructura genética de las poblaciones.