Análisis de estudios longitudinales, datos agrupados y medidas repetidas

Preparado por Luis M. Molinero (Alce Ingeniería)
CorreoE: **bioestadistica@alceingenieria.net**

Artículo en formato PDF

Noviembre 2001

Los métodos estadísticos más utilizados en el análisis de variables numéricas continuas están en su mayoría diseñados para situaciones en las que se registra una única medida por cada unidad de observación (una medida por paciente), es el caso del análisis de la varianza y de las técnicas de regresión. Sin embargo, en la práctica nos encontramos con estudios en los que se toman varias medidas por paciente. Este tipo de estudios en los que para cada paciente una misma variable es registrada en diferentes momentos a lo largo del tiempo, se conocen como "**longitudinales**". Así, por ejemplo, se puede registrar diferentes medidas de presión arterial para cada paciente en diferentes días. Las diferentes presiones registradas para cada unidad de observación (paciente) están correlacionados −es razonable pensar que la variabilidad entre las medidas de cada sujeto sea menor que entre los diferentes sujetos− y no pueden por tanto considerarse como observaciones independientes, supuesto básico para estimar un modelo de regresión clásico.

Es verdad que mediante el análisis de la varianza es posible contemplar diseños en los que tenemos medidas repetidas sobre el mismo elemento de observación, pero tienen el inconveniente de que el número de observaciones por elemento debe ser idéntico (balanceado), requisito que salvo en los estudios experimentales es muy difícil de cumplir, y aún en éstos es posible que se den pérdidas que rompen el equilibrio del diseño.

Para ilustrar este tipo de situaciones, vamos a plantear un ejmeplo. Supongamos que queremos estudiar la influencia de la actividad física del sujeto en la variabilidad de la presión ambulatoria, así como las posibles diferencias entre hombres y mujeres y la influencia del índice de masa corporal. Para simplificar vamos a considerar sólo la PAS. Tenemos para cada sujeto diferentes lecturas de PAS, y para cada una de ellas un índice de actividad física (que suponemos medido en una escala de 0 a 100). Empezamos formulando un modelo de regresión lineal para cada paciente

$$\mathrm{PAS} = \beta_0 + \beta_1 \cdot \mathrm{Actividad} + e$$

Tenemos tantas ecuaciones de regresión (PAS en función de la Actividad) como pacientes, por lo que vamos a representar el modelo de una forma más general:

$$y_{ij} = \beta_{0i} + \beta_{1i} \cdot x_{ij} + e_{ij}$$

donde el subíndice *i* corresponde al paciente, *j* corresponde a cada observación para un mismo paciente, *y* es la variable dependiente (en el ejemplo PAS), *x* la variable independiente (de momento vamos a considerar una sola, que en el ejemplo planteado será el índice de actividad). *e* recoge la parte de variabilidad individual no explicada por la regresión (error o residuo).

Si estudiamos *N* pacientes, tenemos *N* ecuaciones de regresión, y por lo tanto N valores para los coeficientes β_0 y para β_1.

Los valores de esos coeficientes β_0 y β_1 pueden considerarse como una variable aleatoria y calcular su media y desviación típica. Si tenemos dos grupos de pacientes clasificados por ejemplo según el sexo, es razonable

calcular para cada grupo la media de β_1 (media de las pendientes de las regresiones individuales) y compararlas para ver si existen diferencias: para comprobar si influye de diferente manera el índice de actividad en la PAS en el grupo de hombres que en el de mujeres.

Si, de una forma más general, se contempla la posible influencia de una variable numérica continua (no cualitativa como el sexo), como por ejemplo el índice de masa corporar (eso sí hay un único valor de IMC para cada paciente, a lo largo de todas sus observaciones), nos interesa también conocer su posible influencia en esa relación entre la PAS ambulatoria y el índice de actividad. Al igual que hicimos antes con los valores de *PAS* y *Actividad* podemos ahora buscar ajustar una ecuación de regresión para los coeficientes β_0 y β_1 en función del IMC.

Para β_1 en el paciente *i* planteamos la siguiente ecuación:

$$\beta_{i1} = a_{10} + a_{11} \cdot \mathrm{IMC}_i + r_i$$

y tendremos una ecuación similar para β_0

La combinación de los dos modelos de regresión nos permite considerar la influencia del grado de actividad en la variabilidad de la lectura de PAS ambulatoria, así como tener en cuenta las posibles diferencias debidas al IMC del sujeto.

Tanto en la primera ecuación de regresión como en la segunda pueden intervenir más variables independientes. Así en la primera ecuación para cada valor de PAS además del índice de actividad se podría haber incluido, por ejemplo, una variable dicotómica que indica si la lectura corresponde al día o a la noche. Y en la segunda ecuación podría intervenir también, además del IMC, el sexo, si fuma, tipo de medicación antihipertensiva...

Otra clase de estudios en los que nos encontramos con observaciones correlacionadas, no independientes, son los denominados **datos agrupados** (**clustered data**) , en los que existe un diseño jerárquico. Por ejemplo, pacientes agrupados en hospitales, en los que interesa conocer qué características del paciente afectan a la variable analizada o también qué características del grupo o bloque (en este caso el hospital) afectan asimismo a ese resultado. Un ejemplo podría ser un estudio para analizar qué factores de riesgo se asocian con hipertensión en pacientes diabéticos en atención primaria, en el que se incluyen centros con diferentes características. Cada centro aporta inicialmente el mismo número de pacientes al estudio, seleccionados de forma aleatoria. Es razonable pensar que con datos agrupados las observaciones pertenecientes al mismo grupo o bloque son en general más similares entre sí que con respecto a las de otros grupos, lo que violaría la condición de independencia entre las observaciones. Así en nuestro ejemplo puede ocurrir que los pacientes que atiende uno de los centros sean todos ancianos, otro centro fundamentalmente sujetos desempleados por encontrarse en una zona de población con alta tasa de paro...

Cuando las observaciones no son independientes las pruebas estadísticas habituales, que se basan en que sí existe esa independencia, tienden a producir errores estándar más pequeños, al considerar el tamaño de la muestra como el conjunto de todas las observaciones, con lo que se obtiene una sobreprecisión espuria, y más resultados "estadísticamente signficativos" de lo debido. Está claro que si vamos a extraer conclusiones respecto a los hospitales, nuestro tamaño de muestra no es el número de pacientes, sino el número de hospitales.

Esta estructura jerárquica de los datos: observaciones agrupadas en bloques, hace que este tipo de modelos se conozca con el nombre de **modelos multinivel** (**multilevel**), siendo los más utilizados los de 2 niveles. En el caso de los estudios longitudinales el nivel 2 lo constituyen los sujetos y el nivel 1 las observaciones sobre

cada sujeto. En los estudios de datos agrupados el nivel 2 correponde al bloque (por ejemplo hospital) y el nivel 1 a las unidades de observación (sujeto). Podríamos tener 3 niveles (o más): hospital, servicio, paciente.

Otros tipo de investigaciones con medidas correlacionadas son los **estudios de crecimiento**, en los que los sujetos se evalúan en diferentes edades o momentos, y también los de **curvas de dosis−respuesta** en los que se evalúa la respueta de cada sujeto para diferentes dosis del fármaco.

También en los **metanálisis** en los que se dispone de datos individuales de los pacientes, nos encontramos asimismo con un caso de análisis de datos agrupados; en el que los pacientes (nivel 1) se agrupan en bloques constituídos por cada estudio particular (nivel 2).

Podemos describir el **modelo jerárquico** de una forma general, que incluye cualquiera de los casos enunciados anteriormente. Para ello supongamos que tenemos *i* = 1 .. N unidades en el nivel 2 (bloques en el caso de datos agrupados, sujetos para los estudios longitudinales) y tenemos *j* = 1 .. n_i observaciones en el nivel 1 (sujetos en datos agrupados, observaciones repetidas en el contexto de estudios longitudinales), el modelo de regresión se puede escribir entonces de forma general como:

$$y_i = W_i \cdot \alpha + X_i \cdot b_i + e_i$$

donde *y* es el vector respuesta de dimensión n_i x 1 (obsérvese el subíndice *i* que indica que puede haber diferentes observaciones por sujeto o por bloque), **Wi** es una matriz diseño n_i x p para los **efectos fijos**, α es un vector p x 1 de coeficientes de regresión fijos a estimar (desconocidos), **Xi** es es una matriz diseño n_i x r para los **efectos aleatorios**, *bi* es un vector r x 1 de efectos individuales a estimar (desconocido), y *ei* es un vector n_i x 1 de residuos de regresión que corresponde a la variabilidad de las observaciones que aún queda sin explicar con nuestro modelo. He utilizado letras griegas para designar los efectos fijos y letras romanas para los efectos aleatorios.

Con objeto de ilustar cómo se traduce este modelo a la hora de aplicarlo a datos reales, supongamos un ejemplo sencillo en el que estamos evaluando dos fármacos antihipertensivos en pacientes nuevos, divididos en dos grupos a los que se les asigna el tratamiento de forma aleatoria. Para simplificar vamos a fijarnos únicamente en la PAS que se registra antes del tratamiento (valor basal) y durante los cinco meses siguientes. Estamos interesados en evaluar si hay evidencia de mejora diferencial a lo largo del tiempo entre los pacientes tratados con uno u otro fármaco. Una representación matricial de este modelo para el paciente *i* es la siguiente:

$$\begin{bmatrix} PAS_{i0} \\ PAS_{i1} \\ PAS_{i2} \\ PAS_{i3} \\ PAS_{i4} \\ PAS_{i5} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} \cdot \begin{bmatrix} b_{i0} \\ b_{i1} \end{bmatrix} + \begin{bmatrix} Trat_i & Trat_i \times 0 \\ Trat_i & Trat_i \times 1 \\ Trat_i & Trat_i \times 2 \\ Trat_i & Trat_i \times 3 \\ Trat_i & Trat_i \times 4 \\ Trat_i & Trat_i \times 5 \end{bmatrix} \cdot \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} + \begin{bmatrix} e_{i0} \\ e_{i1} \\ e_{i2} \\ e_{i3} \\ e_{i4} \\ e_{i5} \end{bmatrix}$$

Tenemos pues un modelo que estima la PAS con dos factores aleatorios: una ordenada (primera columna de unos) y tendencia lineal a lo largo del tiempo (segunda columna 0..5, basal y meses siguientes); y dos factores fijos: tratamiento (primera columna de la segunda matriz. Asignamos el valor 0 para el primer tratamiento y 1 para el segundo tratamiento). El tratamiento de cada paciente no varía a lo largo de todo el estudio, por lo que para algunos pacientes tendremos una primera columna de 0 y para otros una columna de 1 en la segunda

matriz. Y, por último, en la segunda columna de la segunda matriz tenemos la interacción Tratamiento − semana (producto de ambas variables), que nos permitirá evaluar si los dos grupos de pacientes difieren en su evolución a lo largo del tiempo.

Este tipo de modelos aquí planteado se conoce también con el nombre de **modelos lineales mixtos** ("**linear mixed model**"), debido a que, como vemos, incorporan tanto factores fijos (tratamiento) como aleatorios (evolución del paciente a lo largo del tiempo).

En este ejemplo, la constante u ordenada ("intercept" en la linteratura anglosajona) determina el nivel basal medio de la PAS para los pacientes que reciben el primer tratamiento (TRAT=0). α_0 cuantifica cuánto más alta o baja (signo negativo) es la PAS basal en el segundo grupo de tratamiento (TRAT=1) respecto del primero. Esto es así por cómo se ha elegido la codificación (0 para el instante basal en la variable tiempo y 0 para el primer grupo de tratamiento), con otra codificación el significado sería diferente.

El coeficiente *b1* cuantifica el descenso global (si su signo es negativo) de la PAS de los pacientes a lo largo del estudio, y el coeficiente α_1 nos permite contrastar si hay diferencias en esa evolución entre los dos grupos de tratamiento.

El modelo nos permite también calcular la varianza y covarianza de los efectos aleatorios, es decir la variabilidad individual de la PAS basal y de la evolución.

Este clase de modelos se puede extender a relaciones no lineales entre las observaciones y los términos fijos y aleatorios, como puede ser por ejemplo una regresión logística, cuando la variable resultado es un suceso dicotómico o politómico (más de dos respuestas nominales). En este caso se habla de **modelos lineales mixtos generalizados** ("**generalized linera mixed model**"), y el concepto e interpretación es similar al aquí descrito aunque la matemática es todavía más compleja.

Los modelos lineales generalizados permiten también manejar **observaciones multivariantes**, es decir situaciones en las que se considera más de una variable dependiente (variables objetivo) para los sujetos. En nuestro ejemplo podría interesar el análisis de forma conjunta de la PAS y PAD. La formulación planteada por los modelos lineales mixtos es más flexible que la regresión multivariante tradicional, ya que por ejemplo permite usar covariantes diferentes y comunes para cada variable dependiente, y además no se excluyen los pacientes con ausencias en alguna de las variables dependientes.

En el modelo planteado hasta ahora se ha supuesto que los residuos e_{ij} son independientes, pero en los estudios longitudinales, en los que las observaciones siguen una secuencia temporal, es razonable pensar que las observaciones contiguas se parezcan más entre sí que las observaciones separadas en el tiempo. En estos casos se puede considerar incluir esa relación para los residuos, usando técnicas del área de la estadística matemática conocida como "**análisis de series temporales**" y que no vamos a comentar de momento para no complicar aún más las cosas.

La aplicación de estas técnicas está poco difundida en la literatura biomédica, probablemente debido a su complejidad, y a que hasta hace poco no existía el software adecuado para efectuar los cálculos, y de hecho he encontrado pocas referencias en la literatura médica para ilustrar este artículo. No obstante es indudable su gran utilidad en bastantes situaciones, aunque también en muchas otras, en las que no se da la complejidad estructural aquí planteada, será suficiente con las técnicas de regresión clásicas.

En cualquier caso los modelos matemáticos y estadísticos no son sino una herramienta más para ayudar en la investigación de teorías que deben estar bien argumentadas y nunca a la inversa, y el objetivo ha de ser siempre simplificar y clarificar la interpretación de los datos y no añadir complejidad adicional mediante artificios matemáticos.

Enlaces de interés

- Multilevel Statistical Models
  Harvey Goldstein
  Kendall's Library of Statistics 3 Internet Edition
  Libro on−line que se puede descargar en formato MS−Word. Orientado hacia las ciencias sociales. Cada capítulo se puede descargar de forma independiente. El capítulo 1 constituye una buena introducción, muy bien escrita y sin fórmulas. A partir del capítulo 2 la complejidad aumenta.
- Applied multilevel analysis
  J.J.Hox
  Libro on−line en formato PDF. También orientado hacia las ciencias sociales. Los capítulos 1 y 2 son descriptivos de los modelos de regresión multinivel y no excesivamente complicados de entender.
- Mortality in England and Wales, 1979−1992. An introduction to Multilevel Modelling using MLwiN
  Leyland, A. H., and McLeod
  MRC Social and Public Health Sciences Unit, Glasgow: 2000; Occasional paper no. 1.
- Multilevel Models
  Ian Plewis
- Referencias sobre modelos multinivel

Bibliografía seleccionada

- Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data
  Cnaan A, Laird NM, Slasor P.
  Statistics in Medicine 1997; 16 (20): 2349−2380
- The Influence of Physical Activity on the Variability of Ambulatory Blood Pressure
  Andrew C. Leary, Peter T. Donnan, Thomas M. MacDonald, and Michael B. Murphy
  Am J Hypertens 2000; 13:1067–1073
- Inflammation and dietary protein intake exert competing effects on serum albumin and creatinine in hemodialysis patiens
  George A. Kaysen, Glenn M. Chertow, Rohini Adhikarla, Belinda Yount, Claudio Ronco, and Nathan W. Levin
  Kidney International, 2001; 60:333−340
- What can go wrong when you assume that correlated data are independent: an illustration from the evaluation of a childhood health intervention in Brazil Michael J. Cannon, Lee Warner, J. Augusto Taddei , David G. Kleinbaum
  Statistics in Medicine 2001; 20 (9−10): 1461−1467

Indice de artículos

Principio de la página