# ANÁLISIS DE DATOS LONGITUDINALES

## Libro Fitzmaurice: Interpretación de covariables variables en el tiempo

A continuación, consideramos los aspectos de interpretación de las covariables variables en el tiempo. Utilizando una notación común para la variable de respuesta y las covariables en el estudio longitudinal. Específicamente, dejamos que denote la variable de respuesta para el i-ésimo sujeto en la j-ésima ocasión de medición (i = 1, ..., N; j = 1, ..., ni). Las variables de respuesta para el i-ésimo tema se pueden agrupar en un vector de respuesta nix1

i = 1, …, N

y asociado con cada respuesta, , hay un vector px1 de covariables

i = 1, …, N j = 1, …, ni

Tenga en cuenta que es el vector de covariables asociado con , la variable de respuesta para el i-ésimo sujeto en la ocasión j, y puede incluir dos tipos principales de covariables: covariables cuyos valores no cambian a lo largo de la duración del estudio y covariables cuyos valores cambian a través del tiempo. Las formas se conocen como covariables estacionarias o entre sujetos (p. Ej., Tratamientos experimentales fijos y de género), mientras que a las últimas se las conoce como variables variantes en el tiempo o covariables dentro del sujeto (p. Ej., Tiempo desde el inicio, estado actual de fumar, y exposiciones ambientales). En el primer caso, los mismos valores de las covariables se replican en las filas correspondientes de , para j = 1, ..., ni. En este último caso, los valores tomados por las covariables pueden variar con el tiempo (para al menos algunos individuos) y los valores en las filas correspondientes de pueden ser diferentes en cada ocasión de medición.

Al considerar las covariables variables en el tiempo, podemos distinguir las covariables que varían sistemáticamente a lo largo del tiempo, pero se fijan por diseño del estudio y las covariables que varían de forma aleatoria a lo largo del tiempo. Un ejemplo de una covariable variable en el tiempo que se fija por diseño es un indicador de grupo de tratamiento en una prueba cruzada. Otro ejemplo, y uno que se encuentra comúnmente en un estudio longitudinal, es el tiempo desde la línea de base (cuando las ocasiones de medición son fijadas por el diseño del estudio). Las covariables que varían aleatoriamente a lo largo del tiempo a menudo se denominan estocásticas, es decir, los valores de la covariable en cualquier ocasión no pueden predecirse con precisión, ya que están gobernados por un mecanismo aleatorio. Un ejemplo de covariable variable en el tiempo que es estocástica es el nivel actual de glucosa en sangre. En un estudio observacional de diabéticos, los niveles de azúcar en sangre de los participantes pueden variar aleatoriamente durante la duración del estudio. Los ejemplos adicionales incluyen el estado actual de fumar o los años acumulativos del paquete, la presión arterial, el nivel de colesterol, la ingesta de grasas y la exposición a contaminantes ambientales. Como veremos más adelante, cuando una covariable varía en el tiempo y es estocástica, surgen nuevos problemas con respecto a la interpretación y estimación de los parámetros de regresión en modelos para datos longitudinales.

Recuerde que muchos de los modelos para la respuesta promedio descritos en capítulos anteriores se pueden especificar como

(1)

para alguna función de enlace conocida g (·). Este uso de la notación vectorial y matricial implica que el modelo de la media en cada ocasión está dado por

j = 1, …, ni

Sin embargo, lo que a menudo se pasa por alto es la suposición implícita de que la media condicional de la respuesta j-ésima, dado , ..., , depende solo de .

(2)

Con las covariables estacionarias, esta suposición se mantiene necesariamente desde = para todas las ocasiones k j. Además, con covariables variables en el tiempo que se fijan por diseño del estudio (por ejemplo, indicador de grupo de tratamiento en una prueba cruzada), la suposición también se cumple ya que los valores de las covariables en cualquier ocasión se determinan a priori por diseño del estudio y de manera completamente no relacionado con la respuesta longitudinal. Sin embargo, cuando una covariable es variable en el tiempo y estocástico (2) puede que no necesariamente se mantenga. Por ejemplo, la suposición será violada cuando el valor actual de , dado , predice el valor posterior de . En ese caso

y se dice que confunde la relación entre y . En general, cuando (2) no se cumple, los valores precedentes y / o posteriores de la covariable variable en el tiempo confunden la relación entre y ; esto puede llevar a estimaciones sesgadas de en (1)

Para corregir ideas, considere un estudio longitudinal diseñado para examinar los efectos del ejercicio físico en la reducción de los niveles de glucosa en sangre en pacientes con diabetes mellitus tipo 2. Dejamos que denote la cantidad acumulada de actividad física en la j-ésima ocasión y denota una medida de glucosa en sangre. El objetivo del estudio es determinar la relación entre y . Siguiente. Supongamos que los sujetos con niveles elevados de glucosa en la sangre aumentan a la vez su nivel de actividad física, mientras que los sujetos con la misma cantidad acumulada de actividad física en la ocasión j, pero con niveles normales de glucosa en sangre, continúan manteniendo su nivel habitual de actividad física actividad. Entonces, la suposición de que

no se sostiene y confunde la relación entre y . En particular, la fuerza de la relación entre y se subestimará si los sujetos con niveles elevados de glucosa en sangre posteriormente aumentan su cantidad de actividad física.

En general, cuando una covariable varía en el tiempo y es estocástica, se necesita mucho más cuidado al modelar su relación con la variable de respuesta. Es importante evaluar la suposición hecha en (2), es decir, que la media condicional de , dado todo el perfil covariable variable en el tiempo , ..., , depende solo del valor de la covariable en la ocasión j-ésima, . Sin embargo, tenga en cuenta que se puede definir en términos de funciones de las variables explicativas medidas en o precediendo a la ocasión j-ésima (por ejemplo, exposición acumulativa en la ocasión j). Cuando se viola (2) la relación entre la media de y , expresada en términos de , se verá confundida por los valores anteriores y / o posteriores de la covariable, y pueden producirse inferencias engañosas sobre .

Finalmente, incluso cuando (2) se mantiene, puede haber problemas con la interpretación de los parámetros de regresión que relacionan la respuesta media a las covariables estocásticamente variables en el tiempo. En particular, los parámetros de regresión en (1) pueden no tener la interpretación casual implícita. Por ejemplo, el modelo dado por (1) puede especificar correctamente la relación entre el nivel promedio de glucosa en sangre y la actividad física en la última ocasión de medición, ya que en la última ocasión , la cantidad acumulada de actividad física, es una función de todo el tiempo variando el perfil de las covariables Sin embargo, aunque (2) se mantenga, los parámetros de regresión Beta pueden no tener la interpretación causal implícita sin hacer suposiciones adicionales. Para ver por qué, consideremos una versión simplificada del ejemplo discutido anteriormente.

Supongamos que un grupo de diabéticos se mide en dos ocasiones. Deje que e indiquen los niveles de glucosa en sangre al inicio del estudio y el seguimiento, y y denotan medidas de actividad física en las dos ocasiones. Suponga que es de interés determinar la asociación entre la cantidad acumulada de actividad física, = + y el nivel de glucosa en sangre al finalizar el estudio, . Se supone el siguiente modelo:

,

donde, para el caso de la exposición, se supone una función de enlace de identidad. En este modelo, parece tener una interpretación como el efecto de un aumento de unidad en la cantidad acumulada de actividad física en el nivel medio de glucosa en sangre en el seguimiento, ya que

.

Sin embargo, debido a que varía en el tiempo y es estocástica, esta interpretación de se basa en la validez de cualquiera de los siguientes dos supuestos:

1. no es predicho por , dado y ;

o

1. no es pronosticado por , dado .

En particular, si ninguno de estos supuestos se cumple, "confunde" la relación entre yy no tiene la interpretación causal deseada. Utilizamos vagamente el término "confusión" para enfatizar que oscurece o distorsiona la asociación de interés científico real entre y . Estrictamente hablando, puede considerarse tanto un "factor de confusión" como una llamada "variable intermedia" en el camino causal entre y . Cuando es tanto un factor de confusión como una variable intermedia, ya no se aplican los métodos estándar de ajuste para la confusión (ya que es predicho por , por lo que no debe ajustarse para , por lo que debe ajustarse en el análisis de la asociación entre y ). En cambio, se requieren métodos estadísticos avanzados para la inferencia causal cuando ni (i) ni (ii) se mantienen. Sin embargo, una discusión de métodos estadísticos para la inferencia causal está más allá del alcance de este capítulo; algunas referencias a la literatura estadística sobre este tema aparecen al final del capítulo.

Consideremos estas dos suposiciones en contexto. En un estudio longitudinal, es poco probable que (i) alguna vez se mantenga, ya que las respuestas repetidas generalmente están correlacionadas positivamente (dadas las covariables, y ). Por lo tanto, la interpretación causal de generalmente se basa en la validez de (ii). Por ejemplo, la suposición hecha en (ii) se violaría si los sujetos con niveles elevados de glucosa en sangre al inicio aumentan posteriormente su nivel de actividad física, mientras que los sujetos con la misma cantidad de actividad física al inicio, pero con niveles normales de glucosa en sangre, continúan para mantener su nivel habitual de actividad física. Cuando se cumple la suposición (ii), se dice que la covariable es externa con respecto a la variable de respuesta y tiene la interpretación causal deseada.

En resumen, cuando una covariable varía en el tiempo y es estocástica, debemos considerar la relación entre la respuesta en cualquier ocasión, digamos , y el valor posterior de la covariable, . Se dice que una covariable variable en el tiempo es externa cuando los valores actuales y anteriores de la respuesta en la ocasión j (, ..., ), dados los valores actuales y precedentes de la covariable que varía con el tiempo (, ... ., ), no predice el valor posterior de . Más formalmente, una covariable que varía en el tiempo es externa (o a veces se denomina exogena) cuando

, ... ., , ..., , ... ., (3)

de lo contrario, se dice que la covariable es interna (o endogena). Esto generaliza la suposición hecha en (ii). Tenga en cuenta que cuando una covariable es externa,

,

que es una suposición más débil que (2). Un ejemplo de una covariable externa es la contaminación del aire en estudios del crecimiento de las funciones pulmonares de los niños. Los niveles exteriores de contaminantes atmosféricos (p. Ej., Ozono, partículas finas suspendidas y dióxido de azufre) varían en el tiempo y son estocásticos, pero condicionados a valores pasados, los valores futuros no son pronosticados por las respuestas de la función pulmonar de los participantes y (3 ) sostiene. Sin embargo, debe tenerse en cuenta que la exposición personal de los niños a la contaminación atmosférica no se considerará una covariable externa si los niños con una función pulmonar deficiente cambian posteriormente su comportamiento diario (por ejemplo, pasar menos tiempo al aire libre) para evitar la exposición a altos niveles de contaminación del aire. En principio, es posible examinar la suposición de que una covariable variable en el tiempo es externa al considerar modelos de regresión para la dependencia de en , ..., (o alguna función (es) conocida (s) de , ..., ) y , ..., (o alguna función conocida de , ..., ). La ausencia de cualquier relación entre y , ..., , dado el perfil de covariable anterior, , ..., , proporciona soporte para la validez de la suposición de que el proceso de covariable es externo.

En conclusión, cuando las covariables son variables en el tiempo y estocásticas, los parámetros de regresión no necesariamente tienen la interpretación causal implícita incluso cuando (2) se cumple. A los parámetros de regresión se les puede dar una interpretación causal solo cuando se puede asumir además que las covariables variables en el tiempo son externas con respecto a la variable de respuesta (es decir, cuando (3) se cumple).