En busca de la causalidad

En Medicina es frecuente que tratemos de buscar relaciones de causa efecto. Si queremos demostrar que el fármaco X produce un efecto, no tenemos más que tomar dos grupos de personas, a un grupo le damos el fármaco, al otro grupo no se lo damos y vemos si hay diferencias.

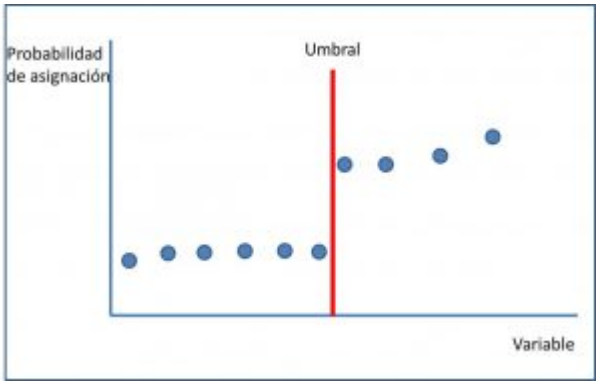
Pero la cosa no es tan sencilla, porque nunca podemos estar seguros de que las diferencias en efecto entre los dos grupos se deban en realidad a otros factores distintos al tratamiento que hemos empleado. Estos factores son los llamados factores de confusión, que pueden ser conocidos o desconocidos y que nos pueden sesgar los resultados de la comparación.

Para resolver este problema se inventó el elemento clave de un ensayo clínico, la aleatorización. Si repartimos los participantes en el ensayo entre las dos ramas de forma aleatoria conseguiremos que estas variables de confusión se repartan de forma homogénea entre las dos ramas del ensayo, con lo que cualquier diferencia entre las dos tendrá que ser debida a la intervención. Solo así podremos establecer relaciones de causa-efecto entre nuestra exposición o tratamiento y la variable de resultado que midamos.

El problema de los estudios cuasi-experimentales y de los observacionales es que carecen de aleatorización. Por este motivo, nunca podremos estar seguros de que las diferencias se deban a la exposición y no a cualquier variable confusora, por lo que no podemos establecer con seguridad relaciones causales.

Este es un inconveniente molesto, ya que muchas veces será imposible realizar ensayos aleatorizados ya sea por motivos éticos, económicos, de la naturaleza de la intervención o de lo que sea. Por eso se han inventado algunas argucias para poder establecer relaciones causales en ausencia de aleatorización. Una de estas técnicas es la de los *propensity score* que vimos en una entrada anterior. Otra es la que vamos a desarrollar hoy, que tiene el bonito nombre de regresión discontinua.

La regresión discontinua es un diseño cuasi-experimental que permite realizar inferencia causal en ausencia de aleatorización. Se puede aplicar cuando la exposición de interés se asigna, al menos parcialmente, según el valor de una variable aleatoria continua si esta variable cae por encima o



por debajo de un determinado valor umbral. Pensemos, por ejemplo, en un fármaco hipocolesterolemiante que pautaremos cuando el colesterol LDL aumente por encima de un valor determinado, o de una terapia antirretroviral en un enfermo de sida que indicaremos cuando su contaje de CD4 disminuya por debajo de determinado valor. Existe una discontinuidad en el valor umbral de la variable que produce un cambio brusco en la probabilidad de asignación al grupo de intervención, tal como os muestro en la figura adjunta.

En estos casos en los que la asignación del tratamiento depende, al menos en parte, del valor de una variable continua, la asignación en las proximidades del umbral es casi como si fuese aleatoria. ¿Por qué? Porque las determinaciones están sujetas a una variabilidad aleatoria por error de muestreo (además de la propia variabilidad de las variables biológicas), lo que hace que los individuos que están muy cerca del umbral, por encima o por debajo, sean muy similares en cuanto a las variables que puedan actuar como confusoras (el estar por encima o por debajo del umbral puede depender de la variabilidad aleatoria del resultado de la medición de la variable), de manera similar a como ocurre en un ensayo clínico. A fin de cuentas, podemos pensar que un ensayo clínico no es más que un diseño de discontinuidad en el que el umbral es un número aleatorio.

La matemática de la regresión discontinua es solo para iniciados y no es mi intención explicarla aquí (primero tendría que entenderla yo), así que nos vamos a conformar con conocer algunos términos que nos servirán para entender los trabajos que empleen esta metodología.

La regresión discontinua puede ser nítida o difusa. En la nítida, la probabilidad de asignación cambia de cero a uno en el umbral (la asignación del tratamiento sigue una regla determinista). Por ejemplo, se inicia el tratamiento cuando se cruza el umbral, con independencia de otros factores. Por otra parte, en la difusa hay otros factores en juego que hacen que en el umbral la probabilidad de asignación cambie, pero no de cero a uno, sino que puede depender de esos otros factores añadidos.

Así, el resultado del modelo de regresión varía un poco según se trate de una regresión discontinua nítida o difusa. En el caso de la regresión nítida se calcula el llamado efecto causal medio, según el cual los participantes son asignados a la intervención con seguridad si traspasan el umbral. En el caso de la regresión difusa, la asignación ya no se realiza según un modelo determinista, sino según uno probabilístico (según el valor respecto al umbral y el de otros factores que el investigador puede

considerar importantes). En estos casos hay que hacer un análisis por intención de tratamiento según la diferencia de la probabilidad de asignación cerca del punto de corte (algunos pueden no traspasar el umbral pero ser asignados a la intervención porque así lo considere el investigador según los otros factores).

Así, en el modelo probabilístico habrá que medir el efecto en los cumplidores (los asignados a la intervención), por lo que el modelo de regresión nos dará el efecto causal medio de los cumplidores, que es la medida típica de la regresión discontinua difusa.

Y creo que aquí lo vamos a dejar por hoy. No hemos hablado nada sobre la ecuación de regresión, pero baste decir que tiene en cuenta las pendientes de la función de probabilidad de asignación antes y después del umbral y una variable de interacción para la posibilidad de que los efectos del tratamiento sean heterogéneos a ambos lados del umbral. Como veis, todo bastante complicado, pero para eso están los paquetes estadísticos como R o Stata que implementan estos modelos sin apenas esfuerzo.

Para terminar, decir solo que lo habitual es ver modelos que utilizan regresión lineal para variables de resultado cuantitativas, pero existen extensiones del modelo que utilizan variables dicotómicas y técnicas de regresión logística, e incluso modelos con estudios de supervivencia y variables de tiempo a suceso. Pero esa es otra historia…