Estudio de casos y controles sobre factores de riesgo en el cáncer colorrectal

2009-11-14

1. Objetivos del estudio

Estimar el riesgo de padecer cáncer colorrectal asociado a una serie de variables

Población

Casos:

436 pacientes con cáncer de colon o recto

Controles:

416 pacientes hospitalizados por otras enfermedades

Variables de interés

Las variables se encuentran en el archivo datos.txt en este orden y con estos códigos, separadas por un tabulador.

Los valores perdidos tienen el código NA.

La primera fila contiene los nombres.

`Grupo`

1:control, 2:caso

`Sexo`

1: hombre, 2: mujer

`Edad`

(C) edad actual en años

`IMC`

(C) índice de masa corporal ($peso/talla^2$) actual

`Calorias`

(C) ingesta calórica diaria promedio del último año

`Tumor`

localización del tumor: 1: control, 2: recto, 3: colon izquierdo, 4: colon derecho

`Extension`

extensión del tumor: 1: control, 2: A/B (local), 3: C (regional), 4: D (metástasis)

`Alcohol`

1: abstemio, 2: bebedor

`Alc.dur`

duración del consumo de alcohol en años

`Tabaco`

1: nunca, 2: ex -fumador, 3: fumador

`Tab.dur`

duración del consumo de tabaco en años

`Familiares`

familiares con cáncer: 1: no, 2: sí colon, 3: sí otros

1.1. Análisis preliminar

1.1.1. Leer los datos

Importar la matriz de datos en R mediante la función `read.table`. Ello crea una `data.frame` que tiene como columnas las variables del estudio y como filas los individuos. Asegurarse de que los datos se importaron correctamente. La función `dim` proporciona el número de filas y de columnas. La función `names` muestra los nombres de las columnas.

1.1.2. Etiquetar las variables categóricas

Definir las variables categóricas como factores con la función `factor`. Asignar etiquetas tal como se muestra en la tabla de definición de variables. Para comprobar que se han creado

correctamente se puede usar la función `table` o `summary`, que muestra las frecuencias de cada categoría. ¿Observas alguna diferencia entre el resultado de las dos funciones?

1.1.3. Evaluar la consistencia interna de los datos

Algunas variables deben mostrar consistencia en los valores. Por ejemplo, las variables que sólo corresponden a los casos como la localización del tumor o la extensión. Con la función `table` se puede comprobar que no hay errores de clasificación. Pensar otras pruebas de consistencia.

1.1.4. Valores perdidos

Detectar la presencia de valores perdidos en las variables mediante la función `summary`. La función `is.na` los identifica. Diferenciar valores perdidos de valores que no procede (duración del consumo si no se consume). Identificar patrones en valores perdidos, es decir, se acumulan en los mismos individuos o están repartidos al azar. Si una variable tiene valores perdidos crear una copia con nombre variable.i y en la copia reemplazar los `NA` según estos criterios:

a) Si es cuantitativa y tiene valores perdidos, reemplazarlos por un valor aleatorio que siga una distribución normal con la misma media y desviación estándar que la variable. Podéis usar la función `rnorm`.

b) Si es cuantitativa y tiene valores que no procede, reemplazarlos por cero.

c) Si es categórica, reemplazarlos por un valor aleatorio con igual probabilidad que las proporciones observadas para cada categoría. Como R no tiene una función para generar valores aleatorios según la distribución multinomial, se puede emplear la uniforme entre 0 y 1 con `runif` y asignar la categoría según una partición proporcional a las frecuencias relativas.

1.1.5. Nuevas variables

1. Agrupar categorías Crear nueva variables:
   a) tabaco2 con categorías: no-fumador, fumador
   b) tumor2 con categorías: control, recto, colon
   c) familiares2 con categorías: no-cáncer cáncer
2. Categorizar variables continuas
   a) Crear nueva variable edad4 con 4 categorías edad con cortes en 60, 70 y 75. Emplear la función cut.
   b) Crear nuevas variables imc4 y calorias4 con 4 categorías con cortes en los percentiles 25, 50 y 75.
   c) Crear nuevas variables alc.dur3 y tab.dur3 con los que no han consumido, los que están entre 1 y 40 años y los que han consumido más de 40 años.
   d) Una vez creadas todas las nuevas variables, crear una nueva `data.frame` llamada `datos.tot` con las antiguas y las nuevas y borrar las variables aisladas para que no molesten.

1.2. Describir las características de la población

Las funciones `summary` y `table` proporcionan descripciones básicas. Para obtener las proporciones de las categorías podemos emplear la función Table, que hay que descargar de la web, pues no está en R-base. Para obtener estadísticos descriptivos de variables

cuantitativas estratificados por categorías de otras variables se puede emplear la función tapply. Funciones resumen útiles son `mean, sd, min, max, median, quantile, mad, IQR`. También puede ser útil emplear gráficas para ver la forma de la distribución de las variables. Probar las funciones `plot, hist, boxplot, pie, pairs`.

1.3. Analizar la asociación de cada variable con la enfermedad

1. Tests de hipótesis: Para cada variable hay que plantear la hipótesis nula y la alternativa y realizar el test de ji- cuadrado mediante la función `chisq.test`.

2. Estimar el riesgo y realizar el test de razón de verosimilitudes: Ajustar un modelo logístico a cada variable con la función `glm`. Calcular el odds-ratio y su intervalo de confianza al 95 % mediante la función `intervals`. Interpretar los resultados en relación al test de hipótesis del apartado anterior. Realizar el test de hipótesis basado en la razón de verosimilitudes con la función `anova` y comparar con el resultado de `chisq.test`. ¿Hay discrepancias entre los tests y los IC95 % en alguna variable? ¿En ese caso, qué resultado es más fiable?

3. Análisis detallado de las variables cuantitativas: Para las variables cuantitativas, analizar también las variables categorizadas, de las que se debe plantear el test de asociación y el de tendencia. Comparar los resultados de los diferentes tests de hipótesis y según la variable esté como continua, categorizada y categorizada tratada como continua.

1.4. Detectar potenciales factores de confusión

En cualquier estudio se suele contar con unas variables de interés, que forman el objetivo del análisis y otras variables que, sin tener un interés per se, deben tenerse en cuenta pues pueden afectar la estimación de los efectos. Entre estas variables, denominadas confusoras, casi siempre están la edad y el sexo, y puede haber otras. Una variable, para ser confusora, debe estar asociada simultáneamente a la enfermedad (ser factor de riesgo) y a la exposición que puede confundir.

Analizar si edad, sexo, imc, calorías, tabaco y familiares son variables potencialmente confusoras para el alcohol.

1.5. Análisis ajustado

Analizar el alcohol ajustando por las variables potencialmente confusoras. Valorar si se da un cambio en la estimación de los beta superior al 20cambio en relación a la amplitud del intervalo de confianza. Hacer los tests de hipótesis ajustados por las variables potencialmente confusoras.

¿Hay algún caso en el que el ajuste cambie las conclusiones?

1.6. Detectar interacciones

Analizar si el efecto estimado para el alcohol se modifica por otras variables. Evaluar como posibles interacciones de interés: edad, sexo, tabaco, imc, calorias y familiares.

Hacer los tests de hipótesis de interacción. Si el p-valor de la interacción es inferior a 0.15 interpretar por qué. Para ello estimar el efecto de alcohol o tabaco en cada categoría de la variable que interactúa.

1.7. Resumen de los resultados

Diseñar tablas resumen con los valores más importantes de los análisis realizados. Redactar un pequeño informe comentando los resultados e interpretando los análisis.