Pontificia Universidad JAVERIANA Bogotá

MAESTRÍA EN EPIDEMIOLOGÍA CLÍNICA

BIOESTADÍSTICA AVANZADA

MÓDULO I

Semana 5

Selección de variables - Interacción y confusión

Material de contenido y aplicación

Carlos Javier Rincón R.

Introducción

En esta semana se presentan dos temas cada uno relacionado al objetivo que se persigue al ajustar el modelo de regresión lineal. El primero es la selección de variables independientes cuando el objetivo es construir un modelo de predicción. El segundo es la evaluación de la modificación del efecto y la confusión cuando el objetivo es evaluar la asociación entre una exposición específica y un desenlace de interés. Se incluyen tres ejemplos de aplicación que los estudiantes deben replicar en sus equipos de forma local para afianzar los estos temas.

Procedimiento de selección de variables en un modelo de predicción

Uno de los objetivos que busca el ajuste de un modelo de regresión es el de construir un modelo de *predicción*. Es decir, se busca obtener un conjunto de variables $x´s$ que al evaluar su combinación lineal permite obtener valores cercanos al que se presentaría en un desenlace y de interés. Cuando se realiza un modelo de predicción, el conjunto de variables incluidas y la interpretación de sus respectivos coeficientes $\beta´s$ no son de interés, sino establecer si en conjunto logran acercarse al valor del desenlace que se puede presentar en una población objetivo.

Un paso inicial en la construcción de un modelo de predicción, es identificar el conjunto de p variables independientes para predecir el desenlace de interés. Esta identificación se basada en el conocimiento de expertos y en la evidencia reportada en la literatura. Es muy importante dedicar tiempo y esfuerzo en la identificación de este grupo de variables, ya que la mejor predicción posible que se podrá realizar estará limitada por este conjunto de variables. El modelo que incluye todo el conjunto de variables independientes se denomina como el *modelo completo*.

En el marco del modelo completo, una característica que se busca en un modelo de predicción, es que sea *parsimonioso*, es decir, un modelo que con el menor número posible de variables independientes, logre un valor tan cercano al desenlace como si se hubieran incluido todas las variables independientes disponibles del modelo completo.

En esta sección, se describen dos procedimiento para la selección de variables basado en pruebas F parciales (vistas en la semana 3) a fin de obtener modelos parsimoniosos a partir del modelo completo.

- Procedimiento de eliminación hacia-atrás (*backward*)

A continuación, se describen los pasos para realizar este procedimiento:

1. Ajustar el modelo.
2. Realizar una prueba F parcial a cada una de las variables independientes.
3. Seleccionar la variable con menor valor del estadística de prueba (o lo que es equivalente, con mayor valor p) y evaluar $H_0: \beta_j = 0$ (asumiendo un $\alpha = 0.1$). Si se rechaza la H_0, el proceso termina definiendo el modelo final. Si no se rechaza H_0, se retira la variable correspondiente y se inicia el proceso ejecutando nuevamente los pasos 1, 2 y 3.

- Procedimiento de selección hacia-adelante (*forward*)

A continuación, se describen los pasos para realizar este procedimiento:

1. Ajustar regresiones lineales simples tomando cada una de las variables independientes.
2. Realizar pruebas F parciales para cada variable independiente en cada regresión. Se elige la variable con mayor valor del estadístico de prueba (o lo que es equivalente, con menor valor p) y se evalúa $H_0: \beta_j = 0$ (asumiendo un $\alpha = 0.1$). Si se rechaza la H_0 esta variable ingresa al modelo.
3. Ajustar regresiones lineales adicionando al modelo definido en el paso 2 cada una de las variable independientes restantes.
4. Calcular pruebas F parciales para cada una de las variables adicionadas. Se selecciona la variable que tienen el mayor valor del estadístico de prueba (menor valor p) y se evalúa H_0. Si no se rechaza H_0 termina el proceso definiendo el modelo final como el modelo ajustado en el paso anterior; si se rechaza H_0 se incluye esta variable al modelo y se repite el paso 3 y 4.

Ejemplo de aplicación 1

A continuación, se presentan los dos proceso de selección de variables utilizando los datos previos de los 40 fetos y su variables:

- DBP: diámetro biparietal en milímetros (mm).
- LCC: longitud cráneo-caudal en milímetros (mm).
- DOF: diámetro occipitofrontal en milímetros (mm).
- edad: edad de la madre (menor o igual a 30 años, 31 a 35 años y mayor a 35 años).
- sexo: sexo del feto (H,M).

Nuevamente los datos son los siguientes:

```
id<-c(1:40) # identificador del sujeto

DBP<-c(18,19,19.2,21,26,24.4,23,16.1,18.8,17,22,20.2,22.5,17.2,
20.3,19,22,20.3,19.5,19.5,20.8,23.8,22,22,21.8,19.6,18.6,
22.5,26.4,19.5,26.4,24.1,21.3,17.1,24.7,22.8,17.7,35.1,
19.3,23.3)

LCC<-c(50,58.9,55.5,64,82,79.1,72.3,49,56.9,61,63.4,54.5,74.8,
51,61.3,74,67,60.1,62.1,56.8,72.2,73.2,64,65.1,64.9,60.4,
58,73.7,82.7,59,82.5,68.1,55,49,76.9,72.6,62.2,90.8,64.9,
71.4)

DOF<-c(23,26.1,24.5,30,34,29.3,24.9,21.3,23,24,26.3,23.4,27.3,22.1,
26.1,29,29,23.6,23.1,25.5,27.4,29,24,28.5,26.3,27,25.3,22.8,
31.9,25.7,33.3,28.3,25.5,22.2,30.07,31,23.9,37.8,24.5,30.2)

edad<-c("mayor a 35 años","31 a 35 años","31 a 35 años","mayor a 35 años",
"31 a 35 años","mayor a 35 años","31 a 35 años","31 a 35 años",
"menor o igual a 30 años","mayor a 35 años","31 a 35 años",
```

```
"31 a 35 años","menor o igual a 30 años","mayor a 35 años",
"31 a 35 años","menor o igual a 30 años","mayor a 35 años",
"31 a 35 años","31 a 35 años","31 a 35 años","menor o igual a 30 años",
"31 a 35 años","31 a 35 años","31 a 35 años","31 a 35 años",
"menor o igual a 30 años","menor o igual a 30 años",
"menor o igual a 30 años","31 a 35 años","31 a 35 años","31 a 35 años",
"mayor a 35 años","menor o igual a 30 años","31 a 35 años",
"mayor a 35 años","menor o igual a 30 años","mayor a 35 años",
"mayor a 35 años","31 a 35 años","mayor a 35 años")

sexo<-c("M","H","H","H","H","H","M","H","M","M","M","H","H","M","M","M",
"M","H","M","M","H","H","H","M","H","M","H","M","M","H","M","M",
"M","M","M","M","H","H","H","M")

bd<-data.frame(id,DBP,LCC,DOF,edad,sexo) # Base de datos

bd$edad<-factor(bd$edad,levels = c("menor o igual a 30 años",
                                   "31 a 35 años","mayor a 35 años"))
bd$sexo<-factor(bd$sexo,levels = c("H","M"))
```

El objetivo del ejercicio es predecir DBP y el conjunto de variables independiente que definen el modelo completo son DOF, LCC, edad y sexo.

- método de selección hacia-atrás:

Iniciamos por ajustar el modelo completo y guardamos sus resultados en el objeto `m_comp` (paso 1).

```
m_comp<-lm(DBP~ DOF+LCC+edad+sexo,data=bd) # modelo completo
```

Evaluamos las pruebas parciales para cada una de las variables independientes (paso 2):

```
drop1(m_comp,test = "F")
```

	Df <dbl>	Sum of Sq <dbl>	RSS <dbl>	AIC <dbl>	F value <dbl>	Pr(>F) <dbl>
<none>	*NA*	*NA*	88.65511	43.83497	*NA*	*NA*
DOF	1	21.99797034	110.65308	50.70082	8.436411088	0.0064215976
LCC	1	37.05156488	125.70668	55.80287	14.209594247	0.0006234946
edad	2	6.54325805	95.19837	42.68334	1.254697909	0.2980318013
sexo	1	0.01290097	88.66802	41.84079	0.004947632	0.9443356202

5 rows

La variable con menor estadístico de prueba F parcial es la variable sexo. Se retira del modelo dado que no hay evidencia para rechaza H_0 y se ajusta un nuevo modelo con las variables restantes. Se calcular las pruebas F parciales de cada una de las variables independientes de este modelo:

```
drop1(update(m_comp, ~ . -sexo), test = "F")
```

	Df <dbl>	Sum of Sq <dbl>	RSS <dbl>	AIC <dbl>	F value <dbl>	Pr(>F) <dbl>
<none>	*NA*	*NA*	88.66802	41.84079	*NA*	*NA*
DOF	1	22.007193	110.67521	48.70882	8.686918	0.0056724503
LCC	1	37.063433	125.73145	53.81075	14.630080	0.0005164177
edad	2	6.630401	95.29842	40.72535	1.308612	0.2830888724

4 rows

La variable con menor estadístico de prueba F parcial es edad. Se retira del modelo dado que no hay evidencia para rechazar H_0. Se ajusta un nuevo modelo con las variables restantes y se evalúan nuevamente las pruebas F parciales:

```
drop1(update(m_comp, ~ . -sexo - edad), test = "F")
```

	Df <dbl>	Sum of Sq <dbl>	RSS <dbl>	AIC <dbl>	F value <dbl>	Pr(>F) <dbl>
<none>	*NA*	*NA*	95.29842	40.72535	*NA*	*NA*
DOF	1	25.93313	121.23155	48.35292	10.06864	0.0030315389
LCC	1	33.91096	129.20938	50.90219	13.16607	0.0008557403

3 rows

La estadística de prueba F parcial menor es el de la variable DOF. Se rechaza su H_0 por lo tanto el proceso termina y las variables seleccionadas son DOF y LCC. El modelo final es:

```
m_final<-lm(DBP~DOF+LCC,data=bd)
summary(m_final)
```

```
##
## Call:
## lm(formula = DBP ~ DOF + LCC, data = bd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7036 -0.8989  0.0193  0.7632  4.9369
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.73049    1.90803  -0.383 0.704023
```

```
## DOF          0.41164    0.12973   3.173 0.003032 **
## LCC          0.16887    0.04654   3.629 0.000856 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.605 on 37 degrees of freedom
## Multiple R-squared:  0.7957, Adjusted R-squared:  0.7847
## F-statistic: 72.06 on 2 and 37 DF,  p-value: 1.737e-13
```

- Método de selección hacia adelante.

Se ajusta un modelo únicamente con intercepto (modelo nulo), y se adicionan cada una de las variables independientes obteniendo las regresiones lineales simple (paso 1)

```
m_int<-lm(DBP~1,data=bd)
add1(m_int, scope = ~ DOF+LCC+edad+sexo, test = "F")
```

	Df <dbl>	Sum of Sq <dbl>	RSS <dbl>	AIC <dbl>	F value <dbl>	Pr(>F) <dbl>
<none>	*NA*	*NA*	466.4790	100.25334	*NA*	*NA*
DOF	1	337.2696216	129.2094	50.90219	99.18974756	3.822314e-12
LCC	1	345.2474509	121.2315	48.35292	108.21773074	1.126402e-12
edad	2	13.4529596	453.0260	103.08281	0.54937185	5.819488e-01
sexo	1	0.9644545	465.5145	102.17056	0.07872852	7.805499e-01

5 rows

La variable adicional que presenta el mayor estadístico de prueba F parcial es LCC y se rechaza su H_0 por lo tanto se agrega al modelo nulo. Se ajusta el modelo con LCC y adicionando cada una de las variables restantes:

```
add1(update(m_int,~ .+LCC),scope = ~ DOF+LCC+edad+sexo, test = "F")
```

	Df <dbl>	Sum of Sq <dbl>	RSS <dbl>	AIC <dbl>	F value <dbl>	Pr(>F) <dbl>
<none>	*NA*	*NA*	121.23155	48.35292	*NA*	*NA*
DOF	1	25.93313223	95.29842	40.72535	10.06864462	0.003031539
edad	2	10.55634087	110.67521	48.70882	1.71686269	0.194010161
sexo	1	0.05805827	121.17349	50.33375	0.01772794	0.894799204

4 rows

La variable con mayor estadístico de prueba F parcial es DOF, y se rechaza su H_0, por lo tanto se agrega esta variable al modelo. Se ajusta el modelo con LCC y DOF, y se adiciona cada una de las variables restantes:

```
add1(update(m_int,~.+LCC+DOF), scope = ~ DOF+LCC+edad+sexo, test = "F")
```

	Df <dbl>	Sum of Sq <dbl>	RSS <dbl>	AIC <dbl>	F value <dbl>	Pr(>F) <dbl>
<none>	*NA*	*NA*	95.29842	40.72535	*NA*	*NA*
edad	2	6.6304013	88.66802	41.84079	1.3086119	0.2830889
sexo	1	0.1000442	95.19837	42.68334	0.0378325	0.8468734

3 rows

La variable con mayor estadística de prueba F parcial es edad, y no se rechaza su H_0, por lo tanto no se adiciona esta variable y termina el proceso. El modelo final toma como variables independientes LCC y DOF.

En esta práctica tanto el método de selección hacia adelante y hacia atrás llegan al mismo modelo, pero esto no siempre ocurre.

Interacción y confusión

Un objetivo distinto al de la predicción, cuando se ajusta un modelo de regresión, es estimar la magnitud de la relación entre una variable de interés (que denominaremos ahora *exposición*) y un desenlace y. El interés ahora será obtener una estimación válida y precisa del coeficiente β asociado a la exposición. Cuando se ajusta un modelo de regresión con este propósito, se debe considerar un conjunto de variables adicionales que al incluirlas y retirarlas como variables independientes del modelo, se obtienen estimaciones muy diferentes sobre este coeficiente. Este conjunto de variables se denominan variables de *confusión*. También se deben considerar otro conjunto de variables que al estimar la relación de la exposición con el desenlace por cada una de las combinaciones de los distintos valores de estas variables, se obtienen magnitudes de la relación diferentes. Este conjunto de variables se denominan *modificadoras del efecto* y se incorporan en el modelo como *interacciones*. A continuación, revisaremos la modificación del efecto y la confusión desde el abordaje de los modelos de regresión lineal iniciando por la modificación del efecto.

Modificación del efecto

Dado una desenlace y, una variable de exposición x_e y una variable x_2 de naturaleza continua como modificadora del efecto, bajo esta identificación de variables e incorporando un término de interacción (identificado por ×), obtenemos la siguiente expresión del modelo:

$$y = \beta_0 + \beta_1 x_e + \beta_2 x_2 + \beta_3 x_e \times x_2$$

Como nuestro objetivo es evaluar la relación entre x_e y y, vamos a comparar dos modelos, variando en una unidad la variable x_e, así:

- Modelo 1:

$$(y|x_e = a+1) = \beta_0 + \beta_1(a+1) + \beta_2 x_2 + \beta_3(a+1) \times x_2$$

simplificando, tenemos

$$(y|x_e = a+1) = \beta_0 + \beta_1 a + \beta_1 + \beta_2 x_2 + \beta_3 a x_2 + \beta_3 x_2$$

- Modelo 2:

$$(y|x_e = a) = \beta_0 + \beta_1 a + \beta_2 x_2 + \beta_3 a \times x_2$$

Comparar los dos modelos anteriores quiere decir, hacer la diferencia entre los dos modelos, así:

$$(y|x_e = a+1) - (y|x_e = a+1) = \beta_1 + \beta_3 x_2$$

el resultado anterior indica que la diferencia en el desenlace por una unidad de cambio en la variable de exposición se expresa en *función de la variable* x_2, es decir, esta diferencia cambia (aumenta o disminuye dependiendo el signo de β_3) $\beta_3$veces el valor de la variable x_2. En consecuencia, si evaluamos la hipótesis nula H_0: $\beta_3 = 0$ utilizando una prueba F parcial, si hay evidencia para rechazarla, concluimos que x_2 es una variable modificadora del efecto; caso contrario x_2 no es una variable modificadora del efecto y la relación entre x_e y y es igual a β_1.

Ahora, repliquemos el proceso de comparación anterior, pero cuando x_2 es una variable categórica con tres posibles valores, $x_2 = [referencia, A, B]$. Es decir, que debemos incorporar esta variable en el modelo a partir de dos variables indicadoras $x_{2A} = \{1\ si\ x_2 = A, 0\ en\ otros\ caso\}$ y $x_{2B} = \{1\ si\ x_2 = B, 0\ en\ otros\ caso\}$. En este caso, la expresión del modelo, incorporando las interacciones, es la siguiente:

$$y = \beta_0 + \beta_1 x_e + \beta_2 x_{2A} + \beta_3 x_{2B} + \beta_4 x_e \times x_{2A} + \beta_5 x_e \times x_{2B}$$

Bajo este modelo, la comparación del desenlace cuando hay un cambio en una unidad en la variable x_e se construye a partir de los dos modelos, así:

- Modelo 1:

$$(y|x_e = a+1) = \beta_0 + \beta_1(a+1) + \beta_2 x_{2A} + \beta_3 x_{2B} + \beta_4(a+1)x_{2A} + \beta_5(a+1)x_{2B}$$

- Modelo 2:

$$(y|x_e = a) = \beta_0 + \beta_1 a + \beta_2 x_{2A} + \beta_3 x_{2B} + \beta_4 a x_{2A} + \beta_5 a x_{2B}$$

entonces, al realizar la diferencia entre estos dos modelo tenemos que:

$$(y|x_e = a+1) - (y|x_e = a) = \beta_1 + \beta_4 x_{2A} + \beta_5 x_{2B}$$

La expresión anterior indica, que la relación entre x_e y y en el grupo de referencia $x_2 = referecia$ es igual a β_1, en el grupo $x_2 = A$ es igual a β_1+β_4 y en el grupo $x_2 = B$ es igual a $\beta_1 + \beta_5$. Si evaluamos la H_0: $\beta_4 = \beta_5 = 0$ a través de una prueba F parcial múltiple, de rechazarla, x_2 es una variable modificadora del efecto, caso contrario x_2 no es una variable modificadora del efecto y la relación en las tres categorías sería igual a β_1.

Del proceso de comparación anterior, se concluye que al incluir una interacción, también se deben incluir los términos de cada una de las variables que la componen (efecto marginales).

Ejemplo de aplicación modificación del efecto

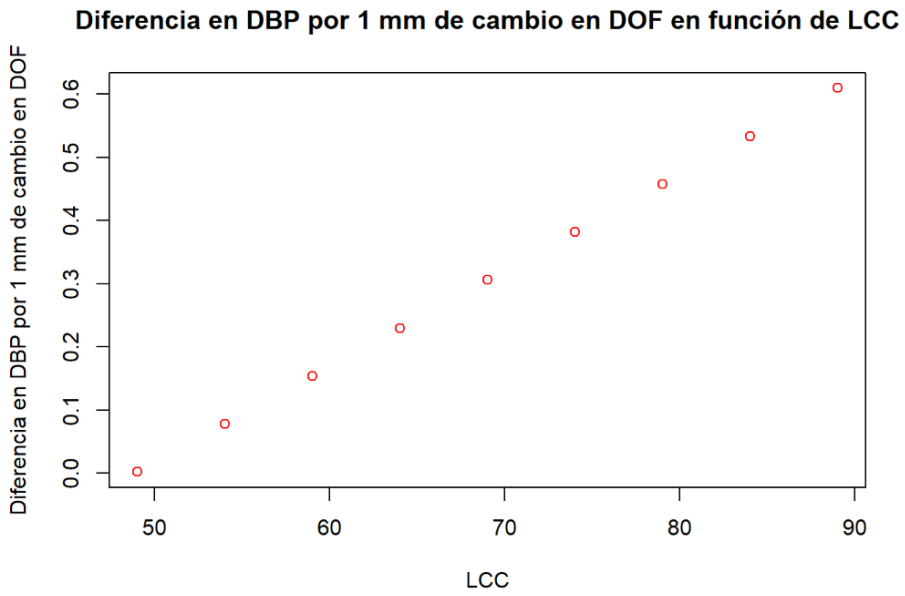
Utilizaremos nuevamente la información de las medidas registradas de DBP, DOF, LCC, edad y sexo en los 40 fetos. Con el objetivo de evaluar la relación entre $x_e = DOF$ y $y = DBP$ considerando la variable LCC como posible modificadora del efecto, ajustamos el siguiente modelo:

```
# interacción con una variable continua
m.lcc<-lm(DBP~DOF*LCC,data=bd)
summary(m.lcc)
```

```
##
## Call:
## lm(formula = DBP ~ DOF * LCC, data = bd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3397 -0.7039 -0.0027  0.8139  2.5092
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.411406  10.343681   2.747  0.00934 **
## DOF         -0.742058   0.420718  -1.764  0.08625 .
## LCC         -0.218076   0.141916  -1.537  0.13312
## DOF:LCC      0.015184   0.005312   2.858  0.00704 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.469 on 36 degrees of freedom
## Multiple R-squared:  0.8335, Adjusted R-squared:  0.8196
## F-statistic: 60.07 on 3 and 36 DF,  p-value: 4.338e-14
```

Del resultado anterior, tenemos evidencia para rechaza la hipótesis nula H_0: $\beta_{DOF \times LCC} = 0$, indicando que la relación entre DBP y DOF se encuentra en función de LCC. Específicamente tenemos que: $(DBP|DOF = a + 1) - (DBP|DOF = a) = -0.742058 + 0.015184 \times LCC$. Dado que el rango de valor de LCC va de 49 mm hasta 90.8 mm, podemos representar esta relación con el siguiente gráfico:

```
LCC<-seq(49,90,5)
diff<--0.742058+0.015184*LCC
plot(LCC,diff,
     main = "Diferencia en DBP por 1 mm de cambio en DOF en función de LCC",
     xlab = "LCC",ylab = "Diferencia en DBP por 1 mm de cambio en DOF",
     col="red")
```



Del gráfico anterior, concluimos que la diferencia en el DBP por 1 mm de cambio en el DOP puede estar entre 0.001958 ($LCC = 49$) hasta 0.6366492 ($LCC = 90$).

Ahora, consideremos la variable edad como posible modificadora del efecto de la relación entre DBP y DOF. Ajustamos el siguiente modelo:

```
# interacción con una variable categórica
m.edad<-lm(DBP~DOF+edad+DOF*edad,data=bd)
summary(m.edad)
```

```
##
## Call:
## lm(formula = DBP ~ DOF + edad + DOF * edad, data = bd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4562 -1.1560  0.0487  1.0771  2.7807
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               16.8735     5.4886   3.074 0.004143 **
## DOF                        0.1428     0.2064   0.692 0.493547
## edad31 a 35 años         -14.7245     6.1085  -2.410 0.021487 *
## edadmayor a 35 años      -25.3577     6.2896  -4.032 0.000296 ***
## DOF:edad31 a 35 años       0.5829     0.2300   2.535 0.016022 *
## DOF:edadmayor a 35 años    0.9552     0.2332   4.097 0.000245 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.552 on 34 degrees of freedom
## Multiple R-squared:  0.8244, Adjusted R-squared:  0.7986
## F-statistic: 31.92 on 5 and 34 DF,  p-value: 6.422e-12
```

En la salida anterior encontramos que los dos términos de interacción son significativos.

Evaluamos ahora una prueba F parcial múltiple comparado el modelo anterior contra un modelo reducido $DBP = DOF + edad$ para evaluar la hipótesis nula
$H_0: \beta_{DOF \times \text{edad31a35}} = \beta_{DOF \times edad>35} = 0.$

```
m2<-lm(DBP~DOF+edad,data=bd)
anova(m.edad,m2) # Prueba F parcial multiple
```

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	34	81.92169	*NA*	*NA*	*NA*	*NA*
2	36	125.73145	-2	-43.80976	9.091193	0.0006874398

2 rows

Se rechaza la hipótesis nula indicado que edad es una modificadora del efecto. Específicamente, tenemos que en el grupo de referencia menores de 31 años, la diferencia en DBP por 1 mm de cambio en DOF es igual a $\beta_1 = 0.1428$, en el grupo de edad de 31 a 35 es de $\beta_1 + \beta\beta_{DOF \times \text{edad31a35}} = 0.1428 + 0.5829 = 0.7257$ y en el grupo de edad de mayores a 35 es de $\beta 1 + \beta_{DOF \times edad>35} = 0.1428 + 0.9552 = 1.098$

Confusión

Ahora revisemos el escenario considerando las variables de confusión. Primero, la confusión se evalúa asumiendo que estas variables no son modificadoras del efecto; si alguna de estas lo es, se debe incorporar como una interacción en el modelo y evaluarlas como se describió en la sección anterior. Bajo este supuesto, iniciemos asumiendo que tenemos dos variables x_2 y x_3 que se consideran posibles variables confusoras de la relación entre x_e y y. Bajo esta premisa, podemos estimar la magnitud de la relación entre X_e y y ajustando un modelo de regresión lineal simple $y = \beta_0 + \beta_1 x_e$ donde β_1 se denomina como la medida del efecto *cruda*. También, podemos estimar la misma relación ajustando un modelo de regresión lineal múltiple $y = \beta_0 + \beta_1 x_e + \beta_2 x_2 + \beta_3 x_3$ donde β_1 se denomina como la medida de efecto ajustada o controlando por x_1 y x_2. Si la estimación cruda se considera diferente a la estimación ajustada, concluimos que x_1 y x_2 son variables confusoras y se deben incluir en el modelo ajustado; caso contrario se pueden retirar la dos variables del modelo. Establecer que hay diferencia entre las dos estimaciones no se basa en la evaluación de una prueba de hipótesis estadística, es un juicio basado en expertos que establecen que la diferencia en la relación se puede considerar *clínicamente significativa*. En ocasiones se utiliza un cambio relativo del 10% como un cambio clínicamente significativo.

Ejemplo de aplicación variables de confusión

Retomando las mediciones de los fetos, vamos a explorar la relación entre xe=DOF y y=DBP, donde las variables LCC, edad y sexo se consideran potenciales confusoras (asumimos que ninguna de estas variables son modificadoras del efecto - para efecto prácticos del ejercicio no consideramos los resultados obtenidos en el ejemplo de práctica de la sección anterior).

Ajustamos una regresión lineal simple $DBP = \beta_0 + \beta_1 DOF$ y una regresión lineal múltiple $DBP = \beta_0 + \beta_1 DOF + \beta_2 LCC + \beta_3 edad + \beta_4 sexo$:

```
m.c<-lm(DBP~DOF,data=bd) # regresión lineal simple
summary(m.c)
```

```
##
## Call:
## lm(formula = DBP ~ DOF, data = bd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1568 -1.0887 -0.3213  0.9755  4.8436
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.23947    2.18677  -0.110    0.913
## DOF          0.80677    0.08101   9.959 3.82e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.844 on 38 degrees of freedom
## Multiple R-squared:  0.723,  Adjusted R-squared:  0.7157
## F-statistic: 99.19 on 1 and 38 DF,  p-value: 3.822e-12
```

```
m.a<-lm(DBP~DOF+LCC+edad+sexo,data=bd) # regresión lineal múltiple
summary(m.a)
```

```
##
## Call:
## lm(formula = DBP ~ DOF + LCC + edad + sexo, data = bd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0099 -0.6101 -0.1220  0.8183  4.8845
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.55734    2.05576  -0.758 0.453938
## DOF                  0.38877    0.13385   2.905 0.006422 **
## LCC                  0.17913    0.04752   3.770 0.000623 ***
## edad31 a 35 años     1.04573    0.66413   1.575 0.124616
## edadmayor a 35 años  0.81194    0.74371   1.092 0.282624
## sexoM                0.03695    0.52534   0.070 0.944336
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.615 on 34 degrees of freedom
## Multiple R-squared:  0.8099, Adjusted R-squared:  0.782
## F-statistic: 28.98 on 5 and 34 DF,  p-value: 2.4e-11
```

Se obtiene una estimación cruda igual a $\beta_{DOF} = 0.80677$ y una estimación ajustada igual a $\beta_{DOF|LCC,edadysexo} = 0.38877$ indicando un cambio relativo de 51.9% lo que se puede considerar como clínicamente significativo y por lo tanto se deben incluir todas las variables al modelo.

Lecturas recomendadas

1. David G. Kleinbaum, Lawrence L. Kupper, Azhar Nizam, Eli S. Rosenberg. 16. Selecting the Best Regression Equation. En: Applied Regression Analysis and Other Multivariable Methods. Fifth Edition. Boston, MA: Cengage Learning; 2014. p. 438-80.
2. David G. Kleinbaum, Lawrence L. Kupper, Azhar Nizam, Eli S. Rosenberg. 11. Confounding and Interaction in Regression. En: Applied Regression Analysis and Other Multivariable Methods. Fifth Edition. Boston, MA: Cengage Learning; 2014. p. 226-56.