Pontificia Universidad JAVERIANA Bogotá

MAESTRÍA EN EPIDEMIOLOGÍA CLÍNICA

BIOESTADÍSTICA AVANZADA

MÓDULO I

Semana 2

Regresión lineal múltiple

Material de contenido y aplicación

Carlos Javier Rincón R.

Introducción

En esta semana se revisará el modelo de regresión lineal múltiple. Se presentan nuevamente los supuestos de la regresión, como se realiza la estimación de los coeficientes basado en una notación de matrices y se presentan las variables indicadoras que se utilizan cuando las variables independientes son de naturaleza cualitativa (nominal u ordinal). Se presenta un ejemplo de aplicación que deben ser replicados por los estudiantes en sus computadores personales utilizando el programa R/RStudio, para afianzar los conceptos anteriores.

Supuestos y planteamiento del modelo

La regresión lineal múltiple, es un método que evalúa la relación entre una variable dependiente y de naturaleza continua y una combinación lineal de p variables independientes, con $p \geq 2$. La expresión general de esta relación se representa en el siguiente modelo:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_j x_j + \ldots + \beta_p x_p + e$$

donde y es el desenlace, x_i es una de las p variables independientes con j variando de 1 hasta p, β_j es uno de los coeficientes del modelo, ahora con j variando de 0 hasta p y e es el término del error. Aunque y debe ser una variable de naturaleza continua, las variables x_j no tienen ninguna restricción en su naturaleza, es decir pueden ser variables categóricas nominales u ordinales o continuas; incluso x_j puede ser una variable que se expresa en función de otra variable, por ejemplo $x_j = log(w_j)$ o $x_j = w_j^2$.

A continuación, se presentan los mismos cinco supuestos del modelo presentados en la sección anterior, ahora en el contexto de una regresión lineal múltiple:

1. **Existencia:** Para todo x fijo, y es una variable aleatoria con media $\mu_{y|x}$ y varianza $\sigma^2_{y|x}$. Ahora x, representa el conjunto de variables independiente $x = (x_1, x_2, \ldots, x_p)$.

2. **Independencia:** Las observaciones del desenlace, y, son estadísticamente independientes. Es decir, que los valores observados del desenlace en un sujeto no se ven afectados por los valores observados del desenlace en los otros sujetos.

3. **Linealidad:** La media o **valor esperado** del desenlace dado un conjunto de valores fijos de las variables independientes, se puede expresar como una combinación lineal de estas variables independientes. Es decir, que el valor esperado del desenlace se puede expresar como:

$$\mu_{y|x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \; [1]$$

Ahora, incorporando el término del error, el desenlace y se expresa como:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + e \; [2]$$

4. **Homocedasticidad:** Las varianzas de y para valores específicos de las variables independientes $x = (x_1, x_2, \ldots, x_p)$. son iguales, esto lo denotamos como $\sigma^2_{y|x} = \sigma^2$. De manera equivalente y basado en la expresión [2], tenemos que la varianza de lo errores para valores específicos de x son iguales, es decir: $\sigma^2_{y|x} = \sigma^2$.

5. **Normalidad:** En el supuesto 1 de existencia se establece que y es una variable aleatoria

$$y|x \sim N(\mu_{y|x}, \sigma^2_{y|x})$$

dado un conjunto de x fijos. Esta variable aleatoria tiene una distribución normal que denotamos de la siguiente forma:

basado en la expresión del modelo [2], tenemos en consecuencia que $e|x \sim N(0, \sigma^2)$.

Estimación de los coeficientes de regresión

Dado el planteamiento del modelo, a continuación se buscará estimar los coeficientes ($\hat{\beta}'s$) que determina la combinación lineal de las variables x que expresan la media del desenlace y. Para esto, se requiere de las observaciones de las variables $y_i, x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{ip}$ en una muestra de n sujetos; el subíndice i denota los valores observados en un sujeto particular i con i variando entre 1 hasta n y j corresponde a una variable independiente específica con j variando entre 1 y p. La expresión del modelo se presenta a continuación:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_j x_{ij} + \ldots + \hat{\beta}_p x_{ip}$$

donde $\hat{y}_i$ es la estimación del desenlace en el sujeto i, $\hat{\beta}_j$ es la estimación de los coeficientes de regresión y x_{ij} son los valores observados en el sujeto i de la variable j.

La estimación de los errores (e en la expresión del modelo [2]) se conocen como los residuales y se obtiene de la diferencia:

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip}$$

Basado en la expresión anterior, la estimación de los $\beta's$ se realiza por el método de **mínimos cuadrados** y corresponden a los valores que minimizan la suma de cuadrados de los residuales, es decir encontrar los valores de $\hat{\beta}$ que minimizan la siguiente expresión:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip})^2$$

Para expresar como se obtiene la solución de los $\beta's$ estimados, se requiere recurrir a la representación de las observaciones en término de **matrices**.

Es decir, las observaciones del desenlace y se expresan en un vector Y de dimensiones $n \times 1$, n filas y una columna, así:

$$Y_{(n\times 1)} = \begin{bmatrix} y_1 \\ y_2 \\ : \\ y_n \end{bmatrix}$$

y las observaciones de las variables independientes x se expresan en una matriz X de dimensiones n×(p+1), así:

$$X_{(n\times p+1)} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ : & : & : & \dots & : \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

Basado en esta notación, las estimaciones de los coeficientes del modelo son igual a:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

donde X' indica la transpuesta de la matriz X y $()^{-1}$ indica la inversa de una matriz. El resultado es un vector de dimensiones $(p+1) \times 1$ con la estimación de los coeficiente que minimiza la suma de cuadrado de los errores:

$$\hat{\beta}_{(p+1\times 1)} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ : \\ \beta_p \end{bmatrix}$$

En la práctica no será necesario realizar la estimación de los coeficientes a partir de la multiplicación de matrices ya que para esto tenemos herramientas como el programa R que nos facilita esta tarea.

Los $\hat{\beta}'s$ obtenidos por el método de mínimos cuadrados cumplen con las siguientes dos propiedades:

1. Bajo el supuesto de normalidad de y, los $\hat{\beta}'s$ también tienen una distribución normal permitiendo realizar inferencias sobre los $\beta's$ (pruebas de hipótesis e intervalos de confianza).
2. Se obtiene el máximo valor del coeficiente de correlación de Pearson entre y_i y $\hat{y}_i$.

Ahora, realizada la estimación de los coeficientes del modelo, nos enfocamos en su interpretación. Tenemos que:

- $\hat{\beta}_0$ representa el valor de la media del desenlace y cuando todas las variables independientes son igual a cero.
- $\hat{\beta}_j$ representa el cambio en la media del desenlace y por una unidad de cambio en la variable independiente x_j *controlando por el resto de variables independientes incluidas en el modelo*.

Como se mencionó anteriormente, las variables independientes pueden ser variables categóricas o variables cuantitativas. Si la variable $x_j$es cuantitativa, la interpretación dada previamente sobre $\hat{\beta}_j$ tiene sentido, pero si la variable es categórica como sexo (hombre - mujer) o nivel de ingresos (bajo - medio - alto), *"... una unidad de cambio en la variable independiente"* no tiene mucho sentido. Para darle sentido a esta interpretación debemos introducir las **variables indicadoras**. Por ejemplo, si la variable x_j es **Nivel de ingreso** con categorías de respuesta: bajo, medio y alto, esta variable se debe introducir al modelo como dos variables indicadoras que se

$$x_{j2} = \begin{cases} 1 \text{ si } x_j = medio \\ 0 \text{ si } x_j \neq medio \end{cases}$$

definen así:

y

$$x_{j3} = \begin{cases} 1 \text{ si } x_j = alto \\ 0 \text{ si } x_j \neq alto \end{cases}$$

Cuando x_{j2} y x_{j3} son iguales a cero, indica que el sujeto presentó un nivel de ingreso bajo y esta categoría queda definida como el grupo de comparación.

Ahora, al estimar el coeficiente β_{j2} y β_{j3} asociados a las dos variables indicadoras, la frase *"una unidad de cambio en la variable independiente"* implica una comparación entre la categoría de la variable indicadora y el grupo de comparación. Para nuestro ejemplo, β_{j2} indica el cambio o la diferencia en la media del desenlace entre el grupo de ingreso "medio" con relación al grupo de ingreso "bajo", controlando por el resto de las variables independientes incluidas en el modelo; y β_{j3} indica el cambio o la diferencia en la media del desenlace entre el grupo de ingreso "alto" con relación al grupo de ingreso "bajo", controlando por el resto de las variables independientes incluidas en el modelo.

Del ejemplo anterior, se deduce que si una variable categórica independiente tiene k categorías de respuesta, se requieren $k - 1$ variables indicadoras para definirla completamente, siendo la categoría de comparación la que no se le asigna una variable indicadora.

Ejemplo de aplicación

Utilizaremos una base de datos de 40 fetos donde se registran las siguientes 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).

Los datos se puede crear a partir del siguiente código:

```
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,
       22.5,26.4,19.5,26.4,24.1,21.3,17.1,24.7,22.8,17.7,15.1,
       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,
       58,73.7,82.7,59,82.5,68.1,55,49,76.9,72.6,62.2,47.8,64.9,
       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,
       31.9,25.7,33.3,28.3,25.5,22.2,30.07,31,23.9,17.8,24.5,30.2)
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(DBP,LCC,DOF,edad,sexo)
```

Las tres primeras variables son cuantitativas, y edad y sexo son variables categóricas *ordinal* y *nominal* respectivamente. El primer paso será establecer la categoría de respuesta que definimos como grupo comparador en las dos variables categóricas. Esto se puede realizar convirtiendo las variables a una estructura tipo factor y definiendo el orden de las categorías con el parámetro `levels`, donde la primera categorías que coloquemos será la categoría de comparación. Así:

```
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"))
```

Ahora, se ajusta una regresión lineal múltiple donde el desenlace es **DBP** y el resto de variables registradas en la base de datos son las variables independientes del modelo utilizando la función `lm`, así:

```
rlm<-lm(DBP~LCC+DOF+edad+sexo,bd)
summary(rlm)
```

```
##
## Call:
## lm(formula = DBP ~ LCC + DOF + edad + sexo, data = bd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8099 -0.4848 -0.0172  0.6444  2.7829
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.77683    1.61292   1.102   0.2784
## LCC                  0.17987    0.03676   4.894 2.36e-05 ***
## DOF                  0.25012    0.10293   2.430   0.0205 *
## edad31 a 35 años     1.10406    0.51638   2.138   0.0398 *
```

```
## edadmayor a 35 años  0.37090    0.57045   0.650   0.5199
## sexoM                0.46963    0.41212   1.140   0.2624
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.256 on 34 degrees of freedom
## Multiple R-squared:  0.8249, Adjusted R-squared:  0.7992
## F-statistic: 32.04 on 5 and 34 DF,  p-value: 6.09e-12
```

En la sección `Residuals` vemos las estadísticas descriptivas (mínimo, percentil 25, mediana, percentil 75 y máximo) obtenidas sobre los residuales.

En la sección `Coefficients` se reporta la estimación de los coeficientes del modelo. Para cada una de estas estimaciones se presentan sus interpretaciones:

- $\hat{\beta}_0$: La media del diámetro biparietal es igual a 1.78 mm dado que la longitud cráneo caudal y diámetro occipitofrontal son iguales a cero milímetros en el grupo de madres menores o iguales a 30 años y en feto de sexo “Hombre”.
- $\hat{\beta}_{LCC}$: La media del diámetro biparietal aumenta en 0.18 mm por cada aumento en un milímetro en la longitud cráneo-caudal, controlando por el resto de variables independientes.
- $\hat{\beta}_{DOF}$: La media del diámetro biparietal aumenta en 0.25 mm por cada aumento en un milímetro en el diámetro occipitofrontal, controlando por el resto de variables independientes.
- $\hat{\beta}_{31\ a\ 35}$ La media del diámetro biparietal aumenta 1.10 mm en el grupo de madres de 31 a 35 años en relación al grupo de madres de 30 o menos años, controlando por el resto de variables independientes.
- $\hat{\beta}_{mayor\ a\ 35}$ La media del diámetro biparietal aumenta 0.37 mm en el grupo de madres mayores a 35 años en relación al grupo de madres de 30 o menos años, controlando por el resto de variables independientes.
- $\hat{\beta}_{Mujer}$ La media del diámetro biparietal aumenta 0.47 mm en el grupo de fetos mujeres en relación al grupo de fetos hombres, controlando por el resto de variables independientes.

Utilizando la expresión matricial, podemos reconstruir la estimación de los coeficientes del modelo. Primero, creamos las variables indicadoras de las dos variables categóricas edad y sexo. Podemos utilizar la función `dummy_cols` del paquete fastDummies que debe ser instalado previamente (escribir en la consola: install.packages("fastDummies") y ejecutar este comando). Las variables indicadoras se crean, utilizando esta función, de la siguiente forma:

```
library(fastDummies) # llamar esta librería previamente instalada.
bd<-dummy_cols(bd,select_columns = c("edad","sexo"))
```

Ahora creamos las matrices **Y** y **X** descritas en la sección de contenidos, así:

```
Y<-matrix(bd$DBP,nrow = 40,ncol = 1,dimnames = list(1:40,"y"));Y
```

```
##       y
## 1  18.0
## 2  19.0
## 3  19.2
## 4  21.0
## 5  26.0
## 6  24.4
## 7  23.0
## 8  16.1
## 9  18.8
## 10 17.0
## 11 22.0
## 12 20.2
## 13 22.5
## 14 17.2
## 15 20.3
## 16 19.0
## 17 22.0
```

```
## 18 20.3
## 19 19.5
## 20 19.5
## 21 20.8
## 22 23.8
## 23 22.0
## 24 22.0
## 25 21.8
## 26 19.6
## 27 18.6
## 28 22.5
## 29 26.4
## 30 19.5
## 31 26.4
## 32 24.1
## 33 21.3
## 34 17.1
## 35 24.7
## 36 22.8
## 37 17.7
## 38 15.1
## 39 19.3
## 40 23.3
```

```
X<-matrix(c(rep(1,40),bd$LCC,bd$DOF,bd$`edad_31 a 35 años`,
            bd$`edad_mayor a 35 años`,bd$sexo_M),nrow = 40,
          ncol = 6,dimnames=list(1:40,c("int","lcc","dof",
                                        "31a35","mayor35","sexoM")));X
```

```
##    int  lcc   dof 31a35 mayor35 sexoM
## 1    1 50.0 23.00     0       1     1
## 2    1 58.9 26.10     1       0     0
## 3    1 55.5 24.50     1       0     0
## 4    1 64.0 30.00     0       1     0
## 5    1 82.0 34.00     1       0     0
## 6    1 79.1 29.30     0       1     0
## 7    1 72.3 24.90     1       0     1
## 8    1 49.0 21.30     1       0     0
## 9    1 56.9 23.00     0       0     1
## 10   1 61.0 24.00     0       1     1
## 11   1 63.4 26.30     1       0     1
## 12   1 54.5 23.40     1       0     0
## 13   1 74.8 27.30     0       0     0
## 14   1 51.0 22.10     0       1     1
## 15   1 61.3 26.10     1       0     1
## 16   1 74.0 29.00     0       0     1
## 17   1 67.0 29.00     0       1     1
## 18   1 60.1 23.60     1       0     0
## 19   1 62.1 23.10     1       0     1
## 20   1 56.8 25.50     1       0     1
## 21   1 72.2 27.40     0       0     0
## 22   1 73.2 29.00     1       0     0
## 23   1 64.0 24.00     1       0     0
## 24   1 65.1 28.50     1       0     1
## 25   1 64.9 26.30     1       0     0
## 26   1 60.4 27.00     0       0     1
## 27   1 58.0 25.30     0       0     0
## 28   1 73.7 22.80     0       0     1
## 29   1 82.7 31.90     1       0     1
## 30   1 59.0 25.70     1       0     0
## 31   1 82.5 33.30     1       0     1
```

```
## 32   1 68.1 28.30     0       1     1
## 33   1 55.0 25.50     0       0     1
## 34   1 49.0 22.20     1       0     1
## 35   1 76.9 30.07     0       1     1
## 36   1 72.6 31.00     0       0     1
## 37   1 62.2 23.90     0       1     0
## 38   1 47.8 17.80     0       1     0
## 39   1 64.9 24.50     1       0     0
## 40   1 71.4 30.20     0       1     1
```

Basado en estas dos matrices y utilizando la función solve del paquete MASS que permite obtener la inversa de una matriz, tenemos que el vector de β^ es igual a:

```
library(MASS)
beta<-solve(t(X)%*%X)%*%t(X)%*%Y;beta
```

```
##                 y
## int     1.7768315
## lcc     0.1798653
## dof     0.2501184
## 31a35   1.1040591
## mayor35 0.3709045
## sexoM   0.4696323
```

Se puede verificar que coincide con las estimaciones de los coeficientes reportadas previamente y que se pueden reportar con la función `coefficients`, así:

```
coefficients(rlm)
```

```
##          (Intercept)                  LCC                  DOF    edad31 a 35 años
##            1.7768315            0.1798653            0.2501184            1.1040591
## edadmayor a 35 años                sexoM
##            0.3709045            0.4696323
```

Lecturas complementarias

1. David G. Kleinbaum, Lawrence L. Kupper, Azhar Nizam, Eli S. Rosenberg. 8. Multiple Regression Analysis: General Considerations. En: Applied Regression Analysis and Other Multivariable Methods. Fifth Edition. Boston, MA: Cengage Learning; 2014. p. 136-64.