UNAP
FINESI

IV Semestre

Modelos Discretos

Regresión Logística Múltiple

Mtr. Alcides Ramos Calcina

Regresión Logística Múltiple

Mtr. Alcides Ramos Calcina

Introducción

- La regresión logística múltiple es una extensión de la regresión logística simple.
- Se basa en los mismos principios que la regresión logística simple (explicados anteriormente) pero ampliando el número de predictores.
- Los predictores pueden ser tanto continuos como categóricos.
- Consideremos ahora la variable respuesta dicotómica Y y un conjunto de predictoras $X_1, X_2, \ldots, X_m$ medidas en n individuos.

1. El modelo

El modelo logístico multivariante establece que

$$\text{logit}\left(p\right) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m$$

donde los β_i son los parámetros desconocidos del modelo.

De manera análoga al caso univariante, otra forma de expresar este mismo modelo es:

$$p = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m}}$$

1. El modelo

En definitiva, este modelo es muy parecido en su definición al modelo logístico simple, la única diferencia es que ahora entran en juego mas variables con la esperanza de que nos ayuden a entender mejor porqué varía la respuesta de unos individuos a otros.

Según la definición del modelo multivariante, para un individuo concreto, cuanto mayor sea el valor de

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m$$

llamado componente sistemático del modelo, mayor será la probabilidad de que presente la característica de interés.

2. Estimación de los parámetros

La estimación de los parámetros $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_m$

de un modelo de regresión logística se efectúa por medio del método de estimación por máxima verosimilitud.

Para ello partimos de un modelo de Bernoulli donde *Y* solo puede tomar valores de (0,1).

$$P[Y=1 \mid X] = p \quad \text{y} \quad P[Y=0 \mid X] = 1-p$$

Calculamos su función de verosimilitud bajo el supuesto de independencia

$$P[Y=y_1, \cdots, Y=y_n] = \prod_{i=1}^{n} p_i^{y_i} \left(1-p_i\right)^{1-y_i}$$

2. Estimación de los parámetros

Aplicando *Log* para simplificar

$$Log\left(P[Y]\right) = \sum_{i=1}^{n} y_i Log\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n} Log\left(1-p_i\right)$$

Expresaremos las probabilidades p_i como una función de las $X\acute{s}$

$$p = P[Y=1 \mid X] = G\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m\right)$$

Para el caso del modelo logit

$$G\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m\right) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m}}$$

Al introducir la función G en el *Log* tenemos:

$$Log\left(P[Y]\right) = \sum_{i=1}^{n} y_i \left(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m\right) - \sum_{i=1}^{n} Log\left(1 - e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m}\right)$$

2. Estimación de los parámetros

derivando respecto a los m+1 parámetros e igualando a cero cada una de las m+1 derivadas obtendremos un sistema de ecuaciones, evidentemente más complejo que en el caso de una sola predictora, que se tiene que resolver por procedimientos iterativos.

$$\frac{\partial}{\partial \beta_i} = \left(\sum_{i=1}^{n} y_i \left(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m \right) - \sum_{i=1}^{n} Log\left(1 - e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m}\right) \right) = 0$$

Para encontrar los verdaderos valores se suele utilizar el algoritmo de Newton-Raphson.

Como resultado de este procedimiento de inferencia tenemos dos resultados:

- Los coeficientes β_i, que son los estimadores de máxima verosimilitud.
- El logaritmo de la verosimilitud (Log Lik), este jugará un papel muy importante a la hora de ver si el modelo es significativo o no.

3. Pruebas de significancia

Como en el caso univariante se prueba la significancia de las variables independientes del modelo mediante las siguientes pruebas:

- ***Prueba de verosimilitud.*** Con la significancia de los m+1 parámetros, bajo la hipótesis para determinar si las variables independientes influyen significativamente en la probabilidad del suceso del modelo relacionado a la variable del resultado del siguiente modo:

  $H_0 : \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_m = 0$

  H_1 : Para algún $\beta_i \neq 0$

- ***Estadístico Wald.*** Evalúa el coeficiente estimado en la población y se define como un cociente entre el coeficiente y el error estándar del coeficiente en la hipótesis:

  $H_0 : \beta_i = 0, \ \forall i = 1, 2, 3, \ldots, m$

  H_1 : Algún $\beta_i \neq 0$

3. Pruebas de significancia

- Para evaluar la bondad de ajuste del modelo se utiliza la prueba de Hosmer-Lemeshow, como se vio anteriormente, consiste en calcular para cada observación del conjunto de datos las probabilidades de la variable dependiente que predice el modelo, se agrupa en aproximadamente 10 grupos iguales a partir de las probabilidades esperadas y se compara con las frecuencias observadas mediante una prueba χ^2 con g-2 grados de libertad, donde g es el número de grupos formados como se explicó en el modelo simple.

- El modelo se ajusta bien si no hay evidencias para rechazar la hipótesis nula.

4. Comparación y selección del modelo

a) Criterio de Información

Para juzgar el ajuste del modelo

- Criterio de información de Aike (AIC): $AIC = -2LL + 2k$
- Criterio de información de Bayes (BIC): $BIC = -2LL + 2k \times \log(n)$

donde n es el número de casos del modelo.

Dado un conjunto de modelos candidatos para los datos, el modelo preferido es el que tiene el valor mínimo en el AIC y BIC.

4. Comparación y selección del modelo

b) Metodos paso a paso (Stepwise Methods)

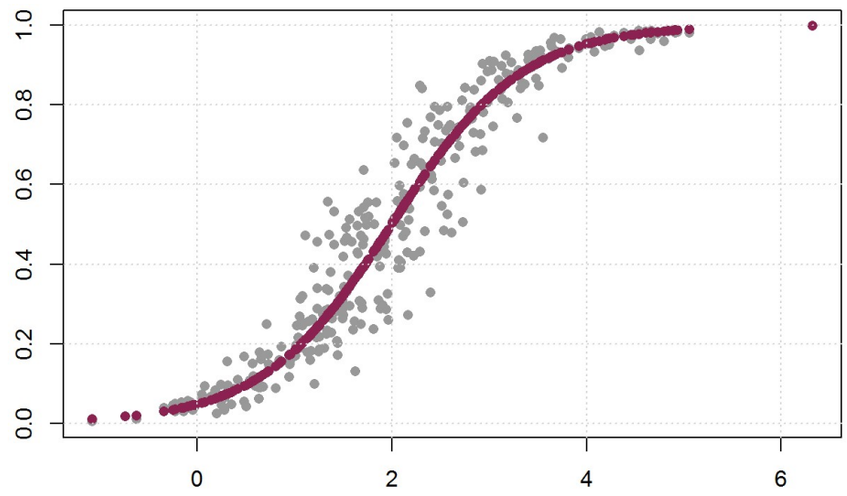
- Si usamos el método hacia delante (***forward***) el ordenador empieza con un modelo que incluye solo la constante y entonces añade predictores individuales al modelo basándose en la variable que mejore el AIC o BIC.

- El método hacia atrás (***backward***) usa el mismo criterio, pero empieza el modelo incluyendo todas las variables predictoras. Si se puede, esa variable se saca del modelo, y se analizan de nuevo el resto de variables.

- La eliminación bidireccional es una combinación de los dos primeros métodos que prueban qué variables deben incluirse o excluirse (buscando siempre que se mejore el *AIC* o *BIC*).

EJEMPLOS

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m$$



Ejemplo 2

Flores (2007) realizo un análisis estadístico de los factores de riesgo que influyen en la enfermedad de Angina de pecho, para lo cual obtuvo datos del servicio de Cardiología del Policlínico Juan José Rodríguez, en Perú, en el cuarto trimestre del año 2001, estudio la ocurrencia o no ocurrencia de la enfermedad en un grupo de 149 pacientes, de los cuales 68 tiene la enfermedad y 78 no la presentan, los datos se muestran en la Tabla 2.

Las variables observadas son:

- y : Angina de pecho (0 si no la presenta, 1 si la presenta).
- edad : Edad en años cumplidos del paciente.
- sexo : Sexo (0 femenino, 1 masculino).
- coles : Colesterol (valor normal 140-200 mg/dl)
- trigl : Triglicéridos (valor normal 45-150 mg/dl).
- gluc : Glucosa (valor normal 70-110 mg/dl.
- hta : Hipertensión arterial (0 no presenta, 1 si presenta).
- obes : Obesidad (0 no tiene, 1 si tiene).

Ejemplo 2

Tabla 2

Estado de la enfermedad Angina de Pecho y factores de riesgo.

dato	y	edad	sexo	coles	trigl	hiper	gluc	obes	dato	y	edad	sexo	coles	trigl	hiper	gluc	obes
1	1	45	M	226	337	1	105	1	49	1	71	F	240	250	1	80	0
2	1	69	F	156	174	0	112	0	50	1	70	M	192	270	0	87	1
3	1	64	M	235	266	0	87	0	51	1	73	F	200	52	1	72	1
4	1	66	M	240	214	1	77	0	52	1	46	F	185	90	0	98	1
5	1	68	F	152	138	1	86	1	53	1	88	M	236	105	1	91	0
6	1	75	M	152	138	0	86	1	54	1	86	M	185	125	0	85	1
7	1	67	M	225	302	0	120	0	55	1	73	M	265	191	1	76	0
8	1	50	M	173	159	1	69	0	56	1	64	M	166	67	1	82	1
9	1	77	F	170	46	1	85	0	57	1	72	M	202	254	0	105	0
10	1	72	M	224	98	0	82	1	58	1	57	F	240	112	1	109	1
11	1	68	M	217	181	1	79	1	59	1	67	F	135	181	1	66	0
12	1	71	M	203	163	1	84	0	60	1	67	F	222	421	0	102	1
13	1	54	F	183	123	0	94	1	61	1	66	F	306	105	1	93	1
14	1	77	F	169	236	1	207	0	62	1	49	M	243	110	0	86	0
15	1	69	M	170	148	0	78	0	63	1	81	M	224	102	0	88	0
16	1	53	M	220	188	1	152	1	64	1	63	F	224	110	1	92	0
114	0	27	F	174	165	0	95	0	141	0	72	F	198	46	0	99	1
115	0	29	F	163	132	0	95	1	142	0	70	M	199	89	0	88	1
116	0	66	F	192	126	0	75	0	143	0	43	F	158	92	1	65	1
117	0	69	F	165	137	0	85	1	144	0	84	M	140	102	0	98	0
118	0	52	F	162	145	0	73	0	145	0	84	M	172	106	0	89	1
119	0	74	M	195	117	0	75	1	146	0	34	F	165	111	1	87	0
120	0	65	F	185	115	0	95	0	147	0	23	F	195	125	1	95	1
121	0	56	F	200	123	0	85	0	148	0	61	M	195	125	1	96	0
122	0	42	F	198	106	0	74	0	149	0	55	F	144	150	0	87	0
123	0	46	F	165	156	0	66	0									

Fuente: (Flores, 2007))

Solución

Se importa los datos desde Excel a través del siguiente comando:

```
library(readxl)
datos <- read_excel("C:/…/Ejmplo2_Angina.xlsx")
attach(datos)
datos
# A tibble: 149 x 9
    dato     y  edad sexo  coles trigl hiper  gluc  obes
   <dbl> <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>
 1     1     1    45 M       226   337     1   105     1
 2     2     1    69 F       156   174     0   112     0
 3     3     1    64 M       235   266     0    87     0
 4     4     1    66 M       240   214     1    77     0
 5     5     1    68 F       152   138     1    86     1
 6     6     1    75 M       152   138     0    86     1
 7     7     1    67 M       225   302     0   120     0
 8     8     1    50 M       173   159     1    69     0
 9     9     1    77 F       170    46     1    85     0
# … with 139 more rows
```

Solución

Estimación del modelo logístico.

El modelo propuesto sería:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1.edad + \beta_2.sexo + \beta_3.coles + \beta_4.trigl + \beta_5.hiper + \beta_6.gluc + \beta_7.obes$$

Procedemos a estimar el modelo logístico tomando en cuenta las ocho variables independientes.

Solución

```
modelo <- glm(y ~ edad + sexo + coles + trigl + hiper + gluc + obes, data = datos, family =
"binomial")
summary(modelo)

Call:
glm(formula = y ~ edad + sexo + coles + trigl + hiper + gluc +
    obes, family = "binomial", data = datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1341  -0.4749  -0.1341   0.3987   2.5673
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.914184   2.752528  -5.418 6.01e-08 ***
edad          0.083862   0.022998   3.646 0.000266 ***
sexoM         0.457323   0.540590   0.846 0.397569
coles         0.035098   0.010388   3.379 0.000728 ***
trigl         0.012232   0.005011   2.441 0.014651 *
hiper         2.364831   0.562555   4.204 2.63e-05 ***
gluc          0.001077   0.011515   0.094 0.925489
obes          0.099239   0.519772   0.191 0.848582
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 205.422  on 148  degrees of freedom
Residual deviance:  99.923  on 141  degrees of freedom
AIC: 115.92
Number of Fisher Scoring iterations: 6
```

Como se muestra en la salida de resultados, las variables que resultaron estadísticamente significativas son:

- ✓ edad,
- ✓ colesterol,
- ✓ triglicéridos e
- ✓ hipertensión,

por lo tanto, construiremos un nuevo modelo.

Solución

El nuevo modelo es: $$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1.edad + \beta_2.coles + \beta_3.trigl + \beta_4.hiper$$

```
modelo2 <- glm(y ~ edad + coles + trigl + hiper, data = datos, family = "binomial")
summary(modelo2)
Call:
glm(formula = y ~ edad + coles + trigl + hiper, family = "binomial",
    data = datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2099  -0.4976  -0.1224   0.4238   2.5477
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.896982   2.580619  -5.773 7.80e-09 ***
edad          0.088372   0.022516   3.925 8.68e-05 ***
coles         0.034719   0.010361   3.351 0.000806 ***
trigl         0.013370   0.004948   2.702 0.006884 **
hiper         2.354901   0.558593   4.216 2.49e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 205.42  on 148  degrees of freedom
Residual deviance: 100.78  on 144  degrees of freedom
AIC: 110.78

Number of Fisher Scoring iterations: 6
```

Por tanto, los coeficientes estimados son:

$$\hat{\beta}_0 = -14.896982$$
$$\hat{\beta}_1 = 0.088372$$
$$\hat{\beta}_2 = 0.034719$$
$$\hat{\beta}_3 = 0.013370$$
$$\hat{\beta}_4 = 2.354901$$

Solución

La ecuación de regresión logística es:

$$\log\left(\frac{p}{1-p}\right) = -14.8970 + 0.0884.edad + 0.0347.coles + 0.0134.trigl + 2.3549.hiper$$

Entonces, la ecuación de regresión logística para predecir la ocurrencia de la Enfermedad Angina de pecho es:

$$\hat{p}_i = \frac{e^z}{1+e^z}$$

Siendo:

$$z = -14.8970 + 0.0884.edad + 0.0347.coles + 0.0134.trigl + 2.3549.hiper$$

Solución

- ***Significancia de los coeficientes***

La prueba de hipótesis para coeficientes individuales del modelo, se efectúa mediante la estadística de **Wald**.

Se plantea las hipótesis siguientes:

H_0: $\beta_i = 0$

H_1: Algún $\beta_i \neq 0$ para i = 1, …, 4

Nos fijamos en el estadístico Z (estadístico de Wald) y su valor de probabilidad asociado al estadístico, en esta se observa que el ***p*-valor** < 0.05 para las cuatro variables. Por tanto, se rechaza la hipótesis nula de que los coeficientes sean igual a cero, lo que indica una significancia individual de cada variable al modelo.

Solución

Con respecto a la **devianza**, se tiene:.

$$\chi^2 = 2LL(nulo) - 2LL(residual) = 205.42 - 100.78 = 104.64$$

Evaluamos la diferencia entre las devianzas si es estadísticamente significativa.

```
anova(modelo2, test = "Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: y
 Terms added sequentially (first to last)

       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                    148     205.42
edad    1   37.921       147     167.50 7.366e-10 ***
coles   1   32.429       146     135.07 1.236e-08 ***
trigl   1   12.207       145     122.86  0.000476 ***
hiper   1   22.080       144     100.78 2.615e-06 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como podemos ver las 4 devianzas son significativas; entonces, el modelo se ajusta perfectamente, es decir, las 4 variables explicativas del modelo son diferentes de cero.

También es importante mencionar que, el valor de AIC inicial que fue de 115.92 en el segundo modelo decrece a 110.78 por lo que mejora la capacidad predictiva del segundo modelo.

Solución

- ***Intervalo de confianza para los coeficientes***

Además del valor de las estimaciones de los coeficientes del modelo, conviene obtener sus correspondientes intervalos de confianza:

```
confint(modelo2, level = 0.95)
Waiting for profiling to be done...
                    2.5 %       97.5 %
(Intercept) -20.588016229 -10.37546865
edad          0.048055889   0.13724651
coles         0.016288312   0.05708576
trigl         0.004543259   0.02413293
hiper         1.321262178   3.53380764
```

En este caso, aplicando la inversa del logaritmo natural para las variables obtenemos los odds:

```
exp(confint(modelo2, level = 0.95))
```

Solución

- ***Evaluación del modelo***

Evaluación a través del estadístico **Pseudo R^2**.

```
PseudoR2(modelo2, c("CoxSnell", "Nagel"))
  CoxSnell Nagelkerke
 0.5045388  0.6744385
```

El R^2 de Nagelkerke es 0.674 es decir que, el 67.4% de las observaciones de Angina de pecho son explicadas por las variables edad, colesterol, triglicéridos e hipertensión.

Solución

aplicamos el test de **Hosmer-Lemeshow**.

```
hoslem.test(datos$y, fitted(modelo2))

    Hosmer and Lemeshow goodness of fit (GOF) test

data:  datos$y, fitted(modelo2)
X-squared = 2.7169, df = 8, p-value = 0.9508
```

Como el p-value = 0.9508 > 0.05, entonces aceptamos la hipótesis nula, es decir el modelo de regresión logística se ajusta a los datos y estaría apto para realizar predicciones de la ocurrencia de la enfermedad de Angina de pecho.

Solución

- ***Interpretación de OR***

Procederemos a calcular el OR del modelo:

```
exp(cbind(OR = coef(modelo2), confint(modelo2)))
Waiting for profiling to be done...
                        OR        2.5 %       97.5 %
(Intercept) 3.390961e-07 1.144823e-09 3.118827e-05
edad        1.092394e+00 1.049229e+00 1.147111e+00
coles       1.035328e+00 1.016422e+00 1.058747e+00
trigl       1.013460e+00 1.004554e+00 1.024426e+00
hiper       1.053709e+01 3.748149e+00 3.425415e+01
```

Por lo tanto, las variables Edad, Hipertensión, Colesterol y Triglicéridos son factores que incrementan la probabilidad de ocurrencia de la enfermedad de Angina de pecho.

FINESI

Modelos Discretos

IV Semestre

https://aulavirtual2.unap.edu.pe/

GRACIAS