UNAP
FINESI

IV Semestre

Modelos Discretos

Regresión Logística con Variables Independientes Categóricas

Mtr. Alcides Ramos Calcina

Introducción

- Cuando se introduce una variable categórica como predictor, un nivel se considera el de referencia (normalmente codificado como 1) y el resto de niveles se comparan con él.

- En el caso de que el predictor categórico tenga más de dos niveles, se generan lo que se conoce como *variables dummy*, que son variables creadas para cada uno de los niveles del predictor categórico y que pueden tomar el valor de 0 o 1.

- Suponga que una de las variables independientes es tipo de material que ha sido codificado como madera, metal y otro. En este caso, se necesitan dos variables de diseño, digamos, D1 y D2 (véase la tabla siguiente).

Introducción

Tipo material	D1	D2
Otro	0	0
Madera	1	0
Metal	0	1

1) En general, si una variable escala nominal tiene k posibles valores, entonces se necesitan $k-1$ variables de diseño.

2) Para ilustrar la notación usada para las variables de diseño en estas notas, suponga que la j-ésima variable independiente x_j tiene k_j niveles. Las k_j-1 variables de diseño serán denotadas por D_{jl} y los coeficientes para estas variables de diseño como β_{jl}.

Introducción

- Por consiguiente, el logit para un modelo probabilístico con k variables y la j-ésima variable discreta será:

$$g\left(x\right)=\text{logit}\left(p_j\right)=\log\left(\frac{p_j}{1-p_j}\right)=\beta_0+\beta_1X_1+\cdots+\sum_{r=1}^{k_j-1}\beta_{jr}D_{jr}+\cdots+\beta_mX_m$$

Variable independiente dicotómica

Mtr. Alcides Ramos Calcina

Variable Independiente Dicotómica

Consideraremos el caso donde el modelo logístico contiene sólo una variable independiente y que ésta es nominal y dicotómica (es decir, medida en dos niveles). Asumamos que la variable independiente x está codificada como:

- $x = 1$: si la persona si está expuesta a un factor de riesgo.
- $x = 0$: si la persona no está expuesta a un factor de riesgo.

Para simplificar la notación, haremos uso de:

$$p(x) = P[Y = 1 \mid X = x]$$

para representar la probabilidad condicional de $Y = 1$ dado $X = x$ cuando se utiliza la regresión logística.

Variable Independiente Dicotómica

Las dos ecuaciones del modelo serán:

$$g(0) = \log\left(\frac{p(0)}{1-p(0)}\right) = \beta_0 + \beta_1(0) = \beta_0$$

$$g(1) = \log\left(\frac{p(1)}{1-p(1)}\right) = \beta_0 + \beta_1(1) = \beta_0 + \beta_1$$

despejando p obtenemos

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{e^{g(x)}}{1+e^{g(x)}} \qquad x \in \{0,1\}$$

Variable Independiente Dicotómica

Los posibles valores de las probabilidades logísticas pueden ser convenientemente organizarse en una tabla de 2×2 como, se muestra en la siguiente tabla:

Valores de $p(x)$ para $x = 0, 1$.

Valores de x e y	$x = 1$ (expuesto)	$x = 0$ (no expuesto)
$y = 1$ (caso)	$p(1) = \dfrac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}}$	$p(0) = \dfrac{e^{\beta_{01}}}{1 + e^{\beta_0}}$
$y = 0$ (control)	$1 - p(1) = \dfrac{1}{1 + e^{\beta_0 + \beta_1}}$	$1 - p(0) = \dfrac{1}{1 + e^{\beta_0}}$
Total	1	1

Variable Independiente Dicotómica

- Para interpretar el efecto de la variable independiente en un modelo, es expresar la diferencia de logit deseada en términos del modelo. Observe que, en este caso, esta diferencia es igual a β_1:

$$g(1) - g(0) = (\beta_0 + \beta_1) - \beta_0 = \beta_1$$

- Los odds de los resultados cuando están presente individuos con $x = 1$ está definido como

$$Odd(1) = \frac{p(1)}{1 - p(1)}$$

- y cuando no están presente individuos con x = 0 está definido como

$$Odd(0) = \frac{p(0)}{1 - p(0)}$$

Variable Independiente Dicotómica

- Por tanto, el *Odd Ratio* (*OR*) se define como

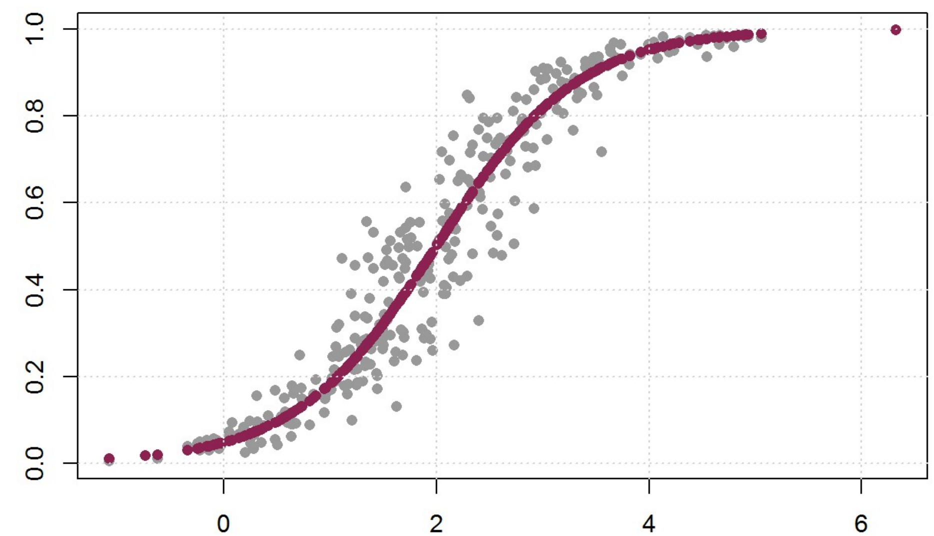
$$Odd\ \ Ratio = OR = \frac{Odd(1)}{Odd(0)} = \frac{\dfrac{p(1)}{1-p(1)}}{\dfrac{p(0)}{1-p(0)}}$$

- Teniendo en cuenta las expresiones de la tabla anterior, obtenemos:

$$OR = \frac{\left(\dfrac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)\Big/\left(\dfrac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\dfrac{e^{\beta_0}}{1+e^{\beta_0}}\right)\Big/\left(\dfrac{1}{1+e^{\beta_0}}\right)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{(\beta_0+\beta_1)-\beta_0} = e^{\beta_1}$$

EJEMPLOS

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



Ejemplo 3

Consideremos los datos del Ejemplo 1, a partir de la variable edad creamos una nueva variable, GEDAD (grupo etario), que toma los siguientes valores:

- 1 : Si EDAD ≥ 55
- 0 : de otro modo

Ahora nuestro interés es estudiar si el grupo etario (GEDAD) es un factor influyente en la presencia o no de enfermedades coronarias (ECC). Los datos se muestran en la siguiente tabla.

Ejemplo 3

Tabla

Grupo etario y estado de la enfermedad cardíaca coronaria de 100 sujetos.

COD	GEDAD	ECC	COD	GEDAD	ECC	COD	GEDAD	ECC	COD	GEDAD	ECC
1	0	0	26	0	0	51	0	1	76	1	1
2	0	0	27	0	0	52	0	1	77	1	1
3	0	0	28	0	0	53	0	0	78	1	1
4	0	0	29	0	1	54	0	1	79	1	1
5	0	1	30	0	0	55	0	0	80	1	0
6	0	0	31	0	0	56	0	1	81	1	0
7	0	0	32	0	1	57	0	0	82	1	1
8	0	0	33	0	0	58	0	0	83	1	1
9	0	0	34	0	0	59	0	1	84	1	1
10	0	0	35	0	0	60	0	0	85	1	1
11	0	0	36	0	0	61	0	1	86	1	0
12	0	0	37	0	1	62	0	1	87	1	1
13	0	0	38	0	0	63	0	0	88	1	1
14	0	0	39	0	1	64	0	0	89	1	1
15	0	0	40	0	0	65	0	1	90	1	1
16	0	1	41	0	0	66	0	0	91	1	0
17	0	0	42	0	0	67	0	1	92	1	1
18	0	0	43	0	0	68	0	0	93	1	1
19	0	0	44	0	0	69	0	0	94	1	1
20	0	0	45	0	1	70	0	1	95	1	1
21	0	0	46	0	0	71	0	1	96	1	1
22	0	0	47	0	0	72	0	1	97	1	0
23	0	1	48	0	1	73	0	1	98	1	1
24	0	0	49	0	0	74	1	0	99	1	1
25	0	0	50	0	0	75	1	1	100	1	1

Ejemplo 3

Cargamos los datos a R.

```
library(readxl)
datos <- read_excel("Ejmplo3_GAGE.xlsx")
View(datos)
attach(datos)
```

Realizamos una tabla de contingencia, es decir un cruce de la variable edad dicotomizada GEDAD con la variable de respuesta ECC.

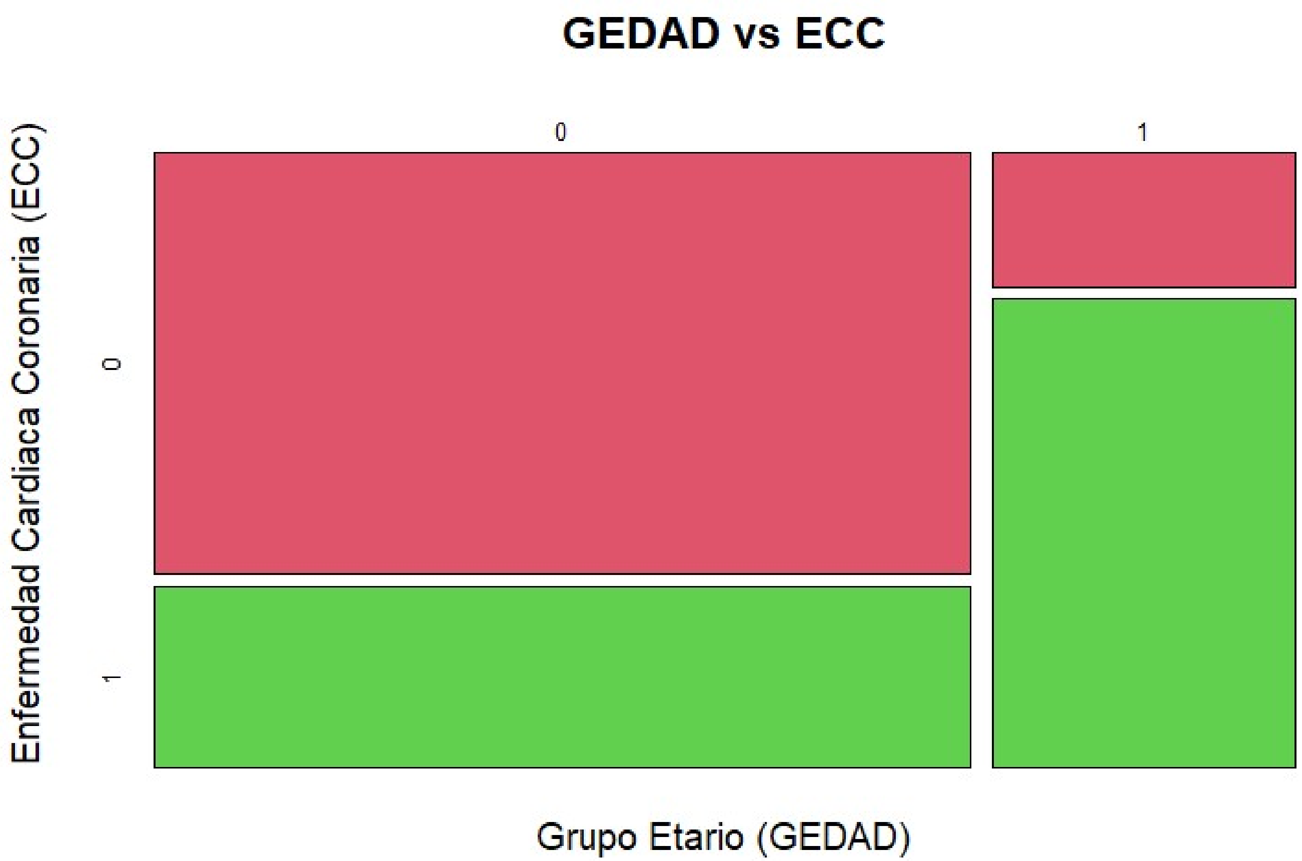
```
# Convertimos las variables a factor
datos$GEDAD <- factor(datos$GEDAD, labels = c("< 55", ">= 55"))
datos$ECC <- factor(datos$ECC, labels = c("Ausente", "Presente"))
table(ECC, GEDAD)
   GEDAD
ECC  0  1
  0 51  6
  1 22 21
```

Ejemplo 3

Visualizamos de forma grafica.

```
mosaicplot(tabla2x2, main = "ECC vs GEDAD", color = 2:4)
```

Mediante el gráfico se nota claramente que, los pacientes con enfermedad coronaria para los que tienen edad < 55 años fueron menor que los pacientes sin la enfermedad, y en los pacientes con edad ≥ 55 años se obtuvieron resultados 20% - 80% aproximadamente.



Ejemplo 3

Estimación del modelo

```
modelo <- glm(ECC ~ GEDAD, data = datos, family = "binomial")
summary(modelo)
Call:
glm(formula = ECC ~ GEDAD, family = "binomial", data = datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7344  -0.8469  -0.8469   0.7090   1.5488

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.8408     0.2551  -3.296  0.00098 ***
GEDAD         2.0935     0.5285   3.961 7.46e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 136.66  on 99  degrees of freedom
Residual deviance: 117.96  on 98  degrees of freedom
AIC: 121.96

Number of Fisher Scoring iterations: 4
```

Ejemplo 3

Se tiene las estimaciones de máxima verosimilitud para los coeficientes.

$$\hat{\beta}_0 = -0.84078 \qquad \text{y} \qquad \hat{\beta}_1 = 2.09355$$

Las ecuaciones estimadas de regresión logística son:

$$\hat{g}(0) = \log\left(\frac{\hat{p}(0)}{1-\hat{p}(0)}\right) = \hat{\beta}_0 = -0.84078$$

$$\hat{g}(1) = \log\left(\frac{\hat{p}(1)}{1-\hat{p}(1)}\right) = \beta_0 + \beta_1 = -0.84078 + 2.09355 = 1.25277$$

Ejemplo 3

y la estimación de la probabilidad

$$\hat{p}(0)=\frac{e^{g(0)}}{1+e^{g(0)}}=\frac{e^{-0.84078}}{1+e^{-0.8407}}=0.3014$$

Cuando el paciente es menor de 55 años (GEDAD = 0), la probabilidad de tener la enfermedad cardiaca coronaria (ECC) es de 30.14%.

$$\hat{p}(1)=\frac{e^{\hat{g}(1)}}{1+e^{\hat{g}(1)}}=\frac{e^{1.25277}}{1+e^{1.25277}}=0.7778$$

Cuando el paciente tiene 55 años o más (GEDAD = 1), la probabilidad de tener la enfermedad cardiaca coronaria (ECC) es de 77.78%.

Ejemplo 3

La estimación de la razón de odds es:

$$OR = e^{\beta_1} = e^{2.09355} = 8.11364$$

Entonces, la razón de posibilidades es 8.11. Esto significa que, el paciente tiene 55 años o más (GEDAD = 1), la probabilidad de tener la enfermedad cardiaca coronaria (ECC) es 8 veces más probable que los pacientes con edad menor a 55 años.

En R obtenemos a su ves el IC para PR:

```
exp(cbind(OR = coef(modelo), confint(modelo)))
Waiting for profiling to be done...
                   OR     2.5 %     97.5 %
(Intercept) 0.4313725 0.2566283  0.7013384
GEDAD       8.1136364 3.0293727 24.7013080
```

Variable independiente politómica

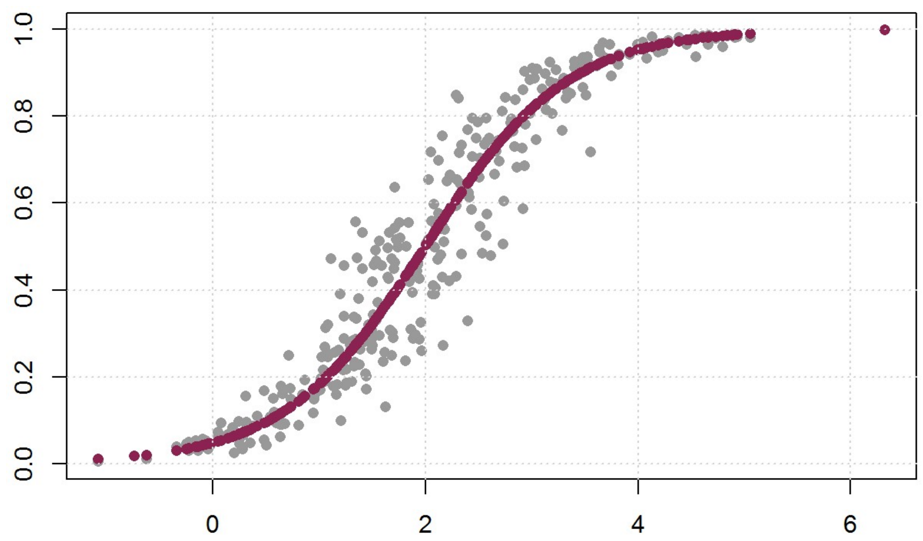
Mtr. Alcides Ramos Calcina

Variable independiente policotómica

Ahora supondremos que, en vez de dos categorías, la variable independiente tiene $k > 2$ valores diferentes. Analizaremos la situación a través del siguiente ejemplo.

EJEMPLOS

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



Ejemplo 4

Suponga un caso hipotético. En un proceso de admisión a una determinada universidad se tienen en cuenta solo a las variables admisión (y) codificada como:

y = 1, Admitido y

y = 0, No admitido

la procedencia (x), la cual está codificada en cuatro niveles:

1 = Otro,

2 = Religioso,

3 = Privado y

4 = Estatal.

Un resumen de los datos se muestra en la siguiente tabla

Ejemplo 4

Tabla
Admisión de un grupo de 400 estudiantes según su colegio de procedencia.

est	admit	proced	est	admit	proced	est	admit	proced	est	admit	proced
1	0	2	101	0	2	201	0	2	301	0	4
2	1	2	102	0	2	202	1	4	302	1	2
3	1	3	103	0	1	203	1	3	303	1	4
4	1	1	104	0	2	204	0	1	304	1	4
5	0	1	105	1	4	205	1	3	305	0	2
6	1	4	106	1	4	206	1	2	306	0	1
7	1	3	107	1	3	207	0	3	307	1	3
8	0	4	108	0	4	208	1	3	308	0	4
9	1	2	109	0	2	209	0	2	309	0	4
10	0	4	110	0	4	210	0	4	310	0	2
11	0	1	111	0	1	211	0	1	311	0	2
12	0	3	112	0	1	212	0	4	312	0	4
13	1	3	113	0	2	213	0	4	313	0	2
14	0	4	114	0	3	214	0	2	314	1	1
…	…	…	…	…	…	…	…	…	…	…	…
95	1	4	195	1	4	295	0	3	395	1	2
96	0	4	196	0	4	296	0	2	396	0	4
97	0	1	197	0	2	297	0	3	397	0	2
98	0	4	198	1	1	298	0	4	398	0	4
99	0	4	199	0	2	299	0	4	399	0	4
100	0	2	200	0	1	300	0	2	400	0	2

Ejemplo 4

La codificamos de la variable independiente categórica sería como se muestra en la tabla siguiente, así mismo, es importante mencionar que el software R realiza automáticamente este proceso, tomando como 0 a la primera categoría.

Variables de diseño para colefio de procedencia (con cuatro niveles).

Colegio de procedencia	D1	D2	D3
Otro	0	0	0
Religioso	1	0	0
Privado	0	1	0
Estatal	0	0	1

Ejemplo 4

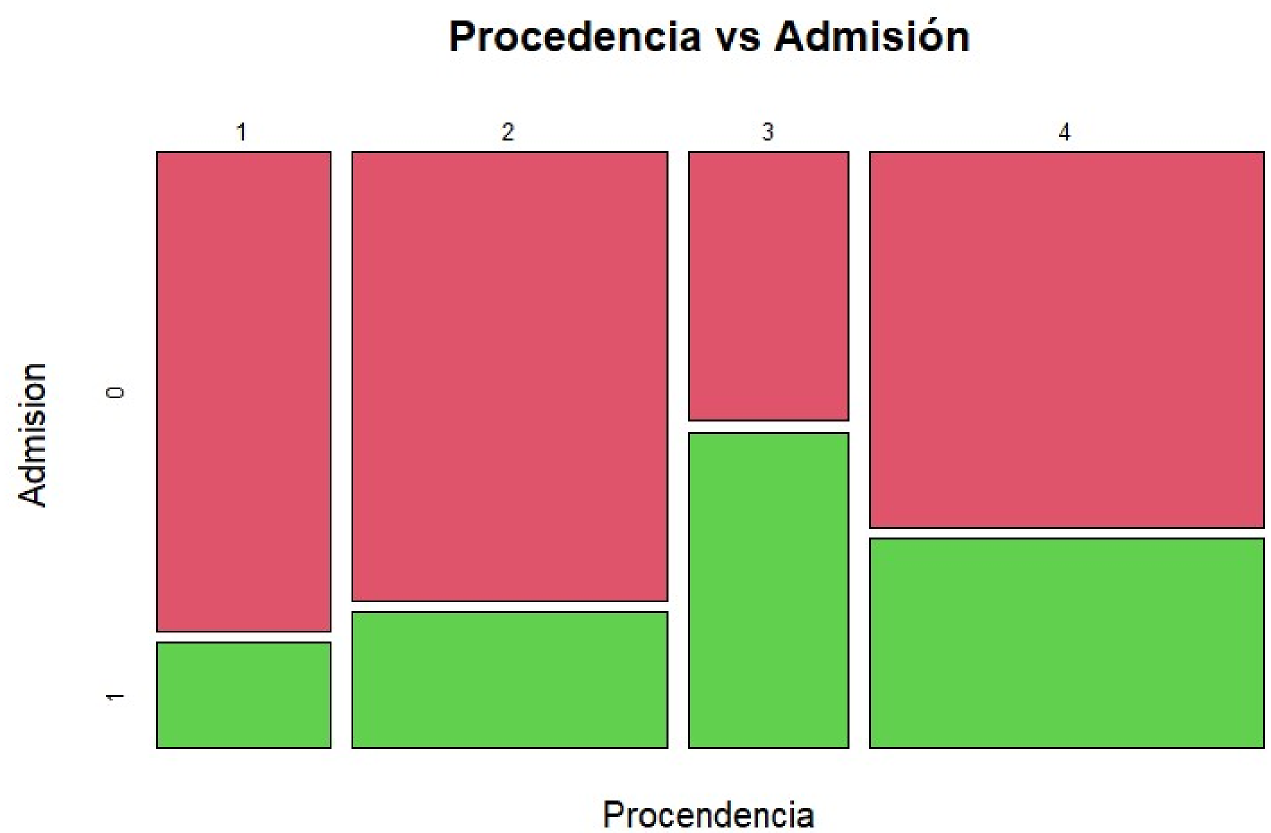
La tabulación cruzada de colegio de procedencia (proced) por admisión (admit):

```
datos$admit <- factor(datos$admit, labels = c("No admitido",
"Admitido"))
datos$proced <- factor(datos$proced, labels = c("Otro", "Religioso ",
"Privado", " Estatal"))
tabla <- table(admit, proced)
tabla
      admit
proced  0  1
     1 55 12
     2 93 28
     3 28 33
     4 97 54
```

Ejemplo 4

Gráficamente tenemos:

```
mosaicplot(tabla, main = "Procedencia vs Admisión", color = 2:7, ylab =
"Admision", xlab = "Procendencia")
```



Ejemplo 4

Estimación del modelo:

```
modelo <- glm(admit ~ proced, data = datos, family = "binomial")
summary(modelo)
Call:
glm(formula = admit ~ proced, family = "binomial", data = datos)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2479  -0.9408  -0.7255   1.1085   1.8546
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.5224     0.3186  -4.778 1.77e-06 ***
procedReligioso   0.3220     0.3847   0.837  0.40252
procedPrivado     1.6867     0.4093   4.121 3.77e-05 ***
procedEstatal     0.9367     0.3610   2.595  0.00947 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.98  on 399  degrees of freedom
Residual deviance: 474.97  on 396  degrees of freedom
AIC: 482.97
Number of Fisher Scoring iterations: 4
```

Ejemplo 4

Los coeficientes estimados son:

$$\hat{\beta}_0 = -1.5224 \qquad \hat{\beta}_{\text{Religioso}} = 0.3220 \qquad \hat{\beta}_{\text{Privado}} = 1.6867 \qquad \hat{\beta}_{\text{Estatal}} = 0.9367$$

A continuación, mostramos las estimaciones de los OR con su respectivo intervalo de confianza.

```
exp(cbind(OR = coef(modelo), confint(modelo)))
Waiting for profiling to be done...
                       OR     2.5 %     97.5 %
(Intercept)     0.2181818 0.1113753  0.3928625
procedReligioso 1.3799283 0.6613932  3.0210003
procedPrivado   5.4017857 2.4773658 12.4315196
procedEstatal   2.5515464 1.2914351  5.3723987
```

Ejemplo 4

- Para los de colegio religioso, la razón odds estimada es:

$$OR_{\text{Religioso}} = e^{\beta_{\text{Religioso}}} = e^{0.3220} = 1.3799$$

Esto nos indica que, el estudiante de colegio religioso tiene la probabilidad de ser admitido en 1.34 veces más que los estudiantes de otros colegios.

- Para los de colegio privado, la razón odds estimada es:

$$OR_{\text{Privado}} = e^{\beta_{\text{Privado}}} = e^{1.6867} = 5.4018$$

- Para los de colegio estatal, la razón odds estimada es:

$$OR_{\text{Estatal}} = e^{\beta_{\text{Estatal}}} = e^{0.9367} = 2.5515$$

FINESI

Modelos Discretos

IV Semestre

GRACIAS

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