UNAP
F I N E S I

IV Semestre

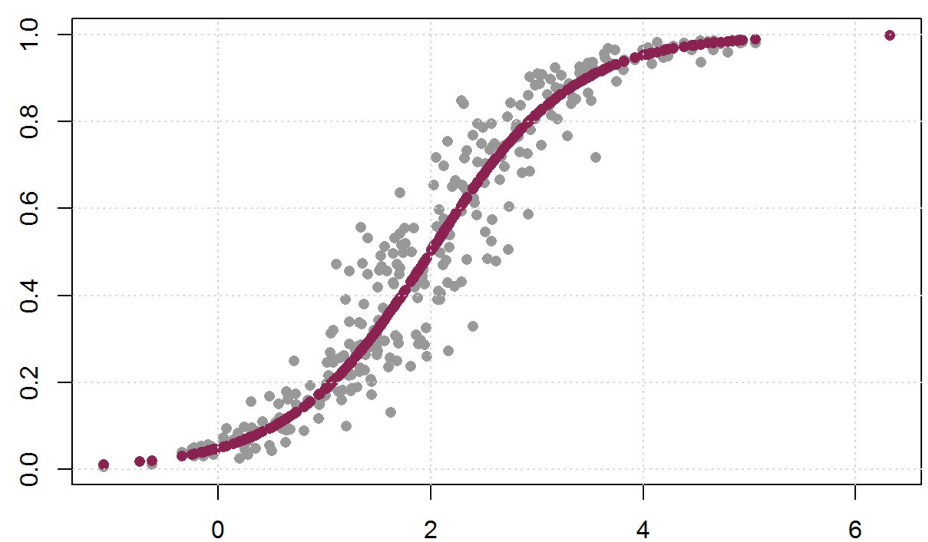
Modelos Discretos

Regresión Logística Simple

Mtr. Alcides Ramos Calcina

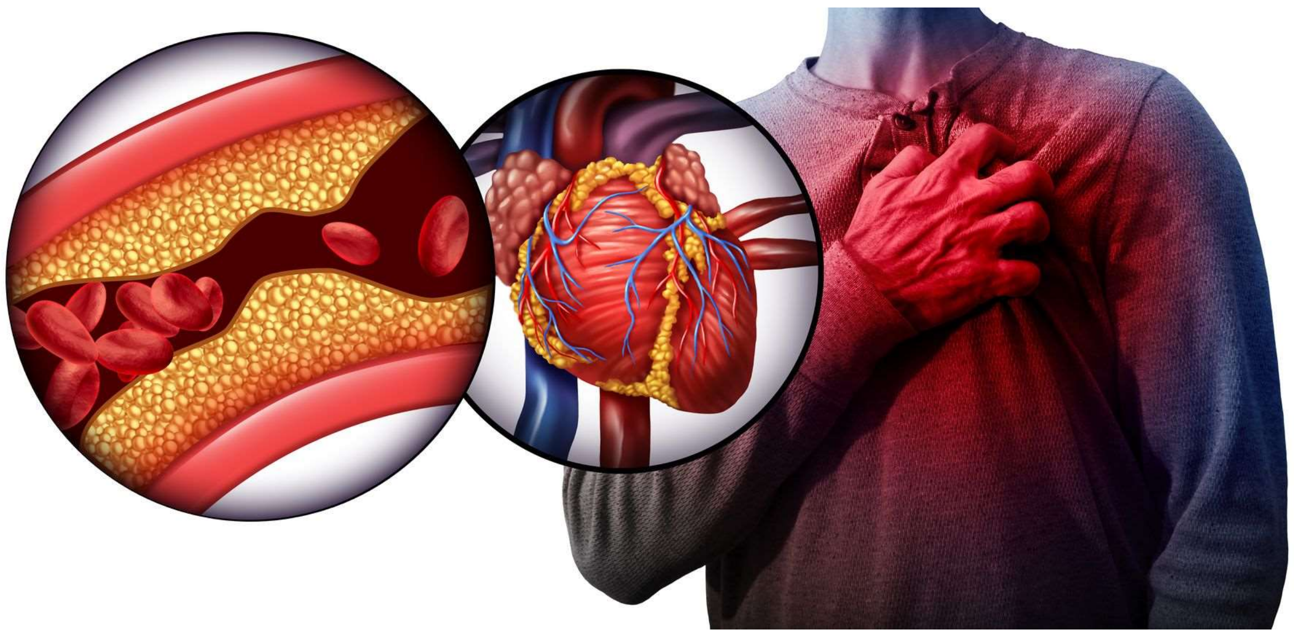
EJEMPLOS

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



Ejemplo 1

La Tabla 1 muestra la edad en años (EDAD) y la presencia o ausencia de evidencia de enfermedad cardiaca coronaria (ECC) para 100 sujetos seleccionados que participan en un estudio. La variable de respuesta es ECC, que se codifica con un valor de 0 para indicar que ECC está ausente, o 1 para indicar que está presente en el individuo.



Ejemplo 1

Tabla 1

Edad y estado de la enfermedad cardíaca coronaria (ECC) de 100 sujetos.

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

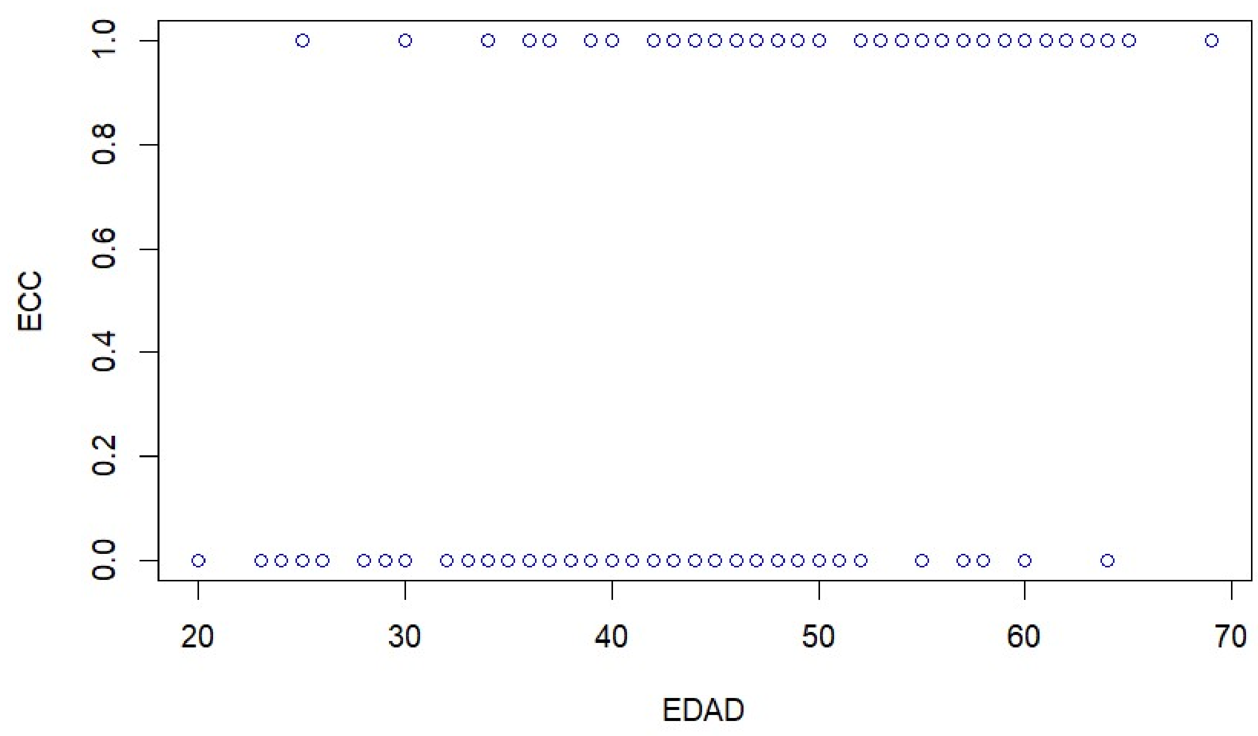
Fuente: David W. & Stanley, (2000)

Solución

Es de interés explorar la relación entre la edad y la presencia o ausencia de ECC en esta población de estudio. Usaremos el diagrama de dispersión para dar una impresión de la variable independiente.

```
plot(EDAD, ECC, col = "blue")
```

En este diagrama de dispersión, todos los puntos caen en una de las dos líneas paralelas que representan la ausencia de ECC (y = 0) y la presencia de ECC (y = 1). Existe cierta tendencia a que los individuos sin evidencia de ECC sean más jóvenes que aquellos con evidencia de ECC.



Solución

Estimación del modelo logístico.

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9718  -0.8456  -0.4576   0.8253   2.2859

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.30945    1.13365  -4.683 2.82e-06 ***
EDAD         0.11092    0.02406   4.610 4.02e-06 ***
---
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: 107.35  on 98  degrees of freedom
AIC: 111.35
Number of Fisher Scoring iterations: 4
```

Por lo tanto, se ve que las estimaciones de máxima verosimilitud de β_0 y β_1 son:

$$\hat{\beta}_0 = -5.30945 \qquad \hat{\beta}_1 = 0.11092$$

La ecuación de regresión logística es:

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -5.30945 + 0.11092X$$

la estimación de la probabilidad

$$\hat{p}_i = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}}$$

Solución

Para los datos del ejemplo, se tiene: $\hat{p}_i = p(\text{presencia ECC}) = \dfrac{e^{-5.30945+0.11092(EDAD)}}{1+e^{-5.30945+0.11092(EDAD)}}$

Interpretación de los parámetros:

- La "pendiente" β_1, su estimación es 0.11092. Eso es positivo, lo que significa que EDAD está asociado positivamente con el evento, presencia de ECC. Al exponencializar la pendiente para convertir el registro de la proporción de log probabilidades en solo la proporción de probabilidades.

$$\exp\left(\hat{\beta}_1\right) = e^{\hat{\beta}_1} = e^{0.11092} = 1.11731$$

Entonces, la razón de posibilidades es 1.12. Esto significa que, por cada aumento de 1 un año en la EDAD, la probabilidad de tener la enfermedad cardiaca coronaria (ECC) aumenta en un 12% (1.12 – 1 = 0.12).

Solución

En R lo obtenemos del siguiente modo:

```
exp(cbind(OR = coef(modelo), confint(modelo)))
                     OR        2.5 %     97.5 %
(Intercept) 0.004944629 0.0004412621 0.0389236
EDAD        1.117306795 1.0692223156 1.1758681
```

Es importante señalar en estos resultados, el intervalo de confianza para el OR de edad, va desde 1.07 hasta 1.18 (relación de 7% a 18%).

Solución

Continuando con el ejemplo, suponga que queremos hacer una predicción para un sujeto en esta población con una edad de X = 30. Sustituya la ecuación para obtener su logit

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -5.30945 + 0.11092(30) = -1.98182$$

Haciendo uso de R tenemos la siguiente estimación:

```
predict(modelo, data.frame(EDAD = 30))
        1
-1.981819
```

Probablemente preferiría una probabilidad o un porcentaje en lugar de un logit.

Solución

Por consiguiente, el valor de probabilidad es: $$\hat{p}_i = \frac{e^{(-1.98185)}}{1+e^{(-1.98185)}} = 0.12112$$

Estamos pronosticando un 12.1% de probabilidad de la presencia de la enfermedad cardiaca coronaria (ECC) cuando la edad es igual a X = 30.

También se pude obtener en R el pronóstico en términos de probabilidad.

```
predict(modelo, data.frame(EDAD = 30), type = "response")
        1
0.1211251
```

Solución

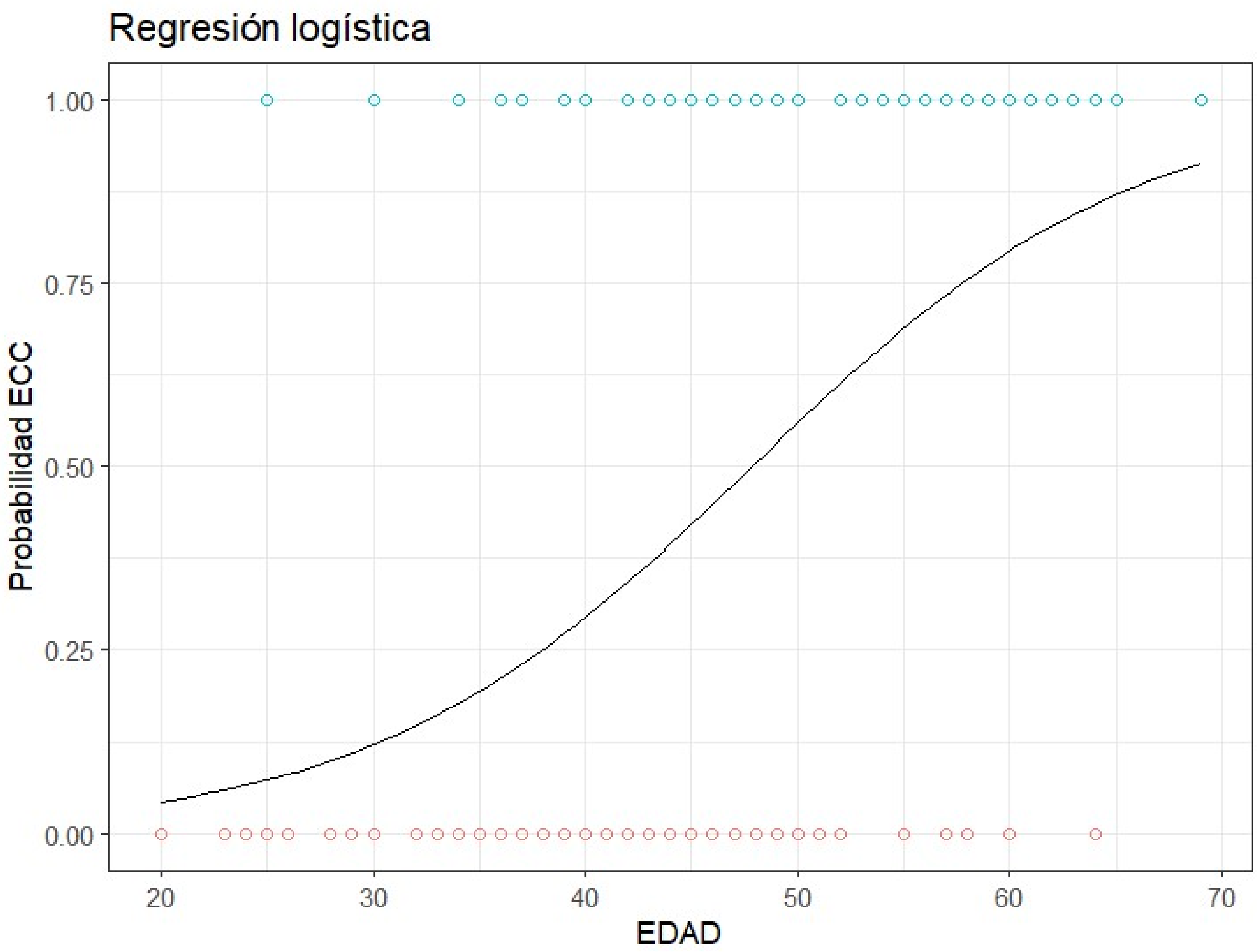
Gráfica del modelo

Para ejecutar en R, es necesario instalar y/o activar la librería **ggplot2**.

```
# library(ggplot2)
ggplot(data = datos, aes(x = EDAD, y = ECC)) +
  geom_point(aes(color = as.factor(ECC)), shape = 1) +
  stat_function(fun      =      function(x){predict(modelo,      newdata      =
data.frame(EDAD = x), type = "response")}) +
  theme_bw() + labs(title = "Regresión logística", y = "Probabilidad ECC")
+
  theme(legend.position = "none")
```

Solución

Se muestra el siguiente grafico



Mtr. Alcides Ramos Calcina

FINESI

Modelos Discretos

IV Semestre

GRACIAS

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