

Ejercitación

Regresión logística

1

La prueba de Wald

Para cada i :

$$H_0 : \beta_i = 0 \quad H_A : \beta_i \neq 0$$

Estadístico (de Wald):

$$W = \frac{\hat{\beta}_i}{es(\hat{\beta}_i)} \sim_{H_0} n(0, 1)$$

2

Criterio optimalidad

Desviación Residual

$$D_R = -2 [LV_M - LV_S]$$

Desviación Nula

$$D_0 = -2 [LV_0 - LV_S]$$

$$0 \xleftarrow{\text{red}} D_R \leq D_0 \xrightarrow{\text{blue}}$$

**Mientras más se separen
mejor el modelo propuesto**

Prueba con las desviaciones

$$H_0: \beta_1 = \beta_2 = \dots \beta_p = 0$$

Ha: alguno(s) diferente(s) de 0

Bajo H_0 :

$$D_0 - D_R \sim \chi_p^2$$

Valor-p: $P(\{D_0 - D_R > \Delta_{\text{obs}}\})$

3

Análisis del ajuste

Coeficiente de
determinación
 R^2 (de Mc Fadden)
para la regresión
logística

4

Coeficiente de determinación R^2 (Mc Fadden) para la regresión logística

Modelo	Logaritmo de la verosimilitud
Saturado	Modelo con el ajuste perfecto (referencia) LV_S
Propuesto	El modelo que se está analizando LV_M
Nulo	El modelo con sólo término independiente LV_0

$$LV_0 \leq LV_M \leq LV_S \leq 0$$

$$R^2 = \frac{LV_0 - LV_M}{LV_0} = 1 - \frac{LV_M}{LV_0}$$

$$0 \leq R^2 \leq 1$$

Mal
ajuste



Buen
ajuste

Como el modelo es Bernoulli:

$$0 < V < 1$$

$$-\infty < LV < 0$$

$$0.2 < R^2 < 0.4$$

ajuste aceptable

5

Problema: Abundancia del lenguado en el estuario Tagus en Portugal

Interés ecológico: Determinar los factores ambientales que influyen en la selección de sitios de crianza por parte de esta especie

Ajuste sólo con salinity

```

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.66071    0.90167   2.951 0.003169 **
salinity     -0.12985    0.03494  -3.716 0.000202 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 87.492  on 64  degrees of freedom
Residual deviance: 68.560  on 63  degrees of freedom
AIC: 72.56

```

```
#calculo R2 McFadden
```

```
#ajuste del modelo de interes
```

```
r=glm(solea_solea~salinity,family=binomial)
```

```
#ajuste del modelo solo con intercepto
```

```
r0=glm(solea_solea~1,family=binomial)
```

```
#calculo
```

```
1-(logLik(r)/logLik(r0))
```

```

> #calculo
> 1-(logLik(r)/logLik(r0))
'log Lik.' 0.2163813 (df=2)

```

6

7

Preguntas comunes:

(1) Ajuste el modelo y valore su ajuste

(2) Analice el efecto de las variables de pronóstico (independientes) sobre la variable dependiente

(2.1) Utilice el coeficiente R^2 de McFadden para valorar la calidad del ajuste

Presente los resultados computacionales en las tablas siguientes:

variable	Estimación	Error estimación	Intervalo de confianza	valor-p (Wald)

Desviación	diferencia	Valor-p
Residual		
Nula		

Todos los análisis error 0.01

8

Problema:
Malaria

$$\mu_{y/x_1, \dots, x_k} = \pi = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + b_2 X_2)}}$$

malaria

Malaria antibody data

Description

The malaria data frame has 100 rows and 4 columns.

Usage

```
malaria
```

Format

This data frame contains the following columns:

subject subject code.

age age in years.

ab antibody level.

mal a numeric vector code, Malaria: 0: no, 1: yes.

Details

A random sample of 100 children aged 3–15 years from a village in Ghana. The children were followed for a period of 8 months. At the beginning of the study, values of a particular antibody were assessed. Based on observations during the study period, the children were categorized into two groups: individuals with and without symptoms of malaria.

9

Problema
Base de datos: diab

Para determinar el tratamiento y atención a los pacientes de diabetes es necesario conocer el tipo de diabetes (A,B). Se realizó un estudio para determinar la naturaleza de la diabetes tipo A. En el estudio se analizaron individuos diabéticos no obesos.

(3) Calcule la probabilidad de que se enferme de malaria si las variables independientes toman los valores:
RI=200
PG=150
PR=1.00

Utilice un error tipo I de 0.05

Variables		
Tipo	1	diabetes A
	0	diabetes B
RI	Respuesta a la insulina	
PG	Resistencia a la insulina	
PR	Peso Relativo	

10

Al profesor de Estadística le interesa identificar las asignaturas que de alguna forma se relacionan con los resultados de sus estudiantes. Utilice la regresión logística para identificar si los resultados en Matemáticas (Mat), Física (Fis), Literatura (Lit) y Educación Física (Ef) se relacionan (influyen) con los resultados de Estadística (Est).

Los datos de las diferentes variables se encuentran en el **marco de datos L**:

Las variables:

Mat, Fis, Lit, Ef, contienen los resultados en las asignaturas correspondientes y se evalúan entre 0 y 5.

La variable Est toma los valores 1 / 0: 1 si aprobado, 0 si no aprueba.

(3) ¿Cuál es la probabilidad de aprobar estadística una persona que tenga los resultados siguientes?:

Mat	Fis	Lit	Ef
1.2	2.4	4	3.9

Utilice un error tipo I de 0.05

Archivo: datos L

11

Base de datos cor
Factores que influyen sobre
la presencia de infarto.

Variables

Inf: 1 infarto, 0 no

café: 1 bebe

habitualmente, 0 no

CR: gasto en comidas
rápidas

ing: ingresos

(3) ¿Cuál es la probabilidad de tener infarto una persona que tome café y gaste en comida rápida 1500 y tenga ingresos de 4500?:

12

Enfermedad coronaria

Se selecciona una muestra de varones mayores de 40 años que al momento de comenzar el estudio no tienen criterio de enfermedad coronaria (EC). Se observa su evolución durante 5 años y se toma nota de quienes han desarrollado criterios de EC .

Variables:

EC: 1 (si) / 0 (No)

Fuma: 1 SI / 0 No

Edad

(4) Si una persona incrementa su edad en 10 años, cómo influiría esto en el riesgo (probabilidad) de presentar enfermedad coronaria

Utilice un error tipo I de 0.05