7 Logistic regression

Regresion logistica

- Para variables dicotomicas, modelamos las probabilidad

$\pi(x_1, x_2, \ldots, x_q)$ Probabilidad de $y = 1$

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q.$$

$$\pi(x_1, x_2, \ldots, x_q) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q)}.$$

$\exp(\beta_j)$ Efecto de la variable j-esima

Modelo lineal general

- Tienen una distribución del error centrada en la media (normal, binomial,..)
- Funcion de enlace (link function)

$$g(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q.$$

- Una función conocida que modele la varianza alrededor de la media

Caso 1

The erythrocyte sedimentation rate (ESR) is the rate at which red blood cells (erythrocytes) settle out of suspension in blood plasma, when measured under standard conditions. If the ESR increases when the level of certain proteins in the blood plasma rise in association with conditions such as rheumatic diseases, chronic infections and malignant diseases, its determination might be useful in screening blood samples taken from people suspected of suffering from one of the conditions mentioned. The absolute value of the ESR is not of great importance; rather, less than 20mm/hr indicates a 'healthy' individual. To assess whether the ESR is a useful diagnostic tool, Collett and Jemain (1985) collected the data shown in Table 7.1. The question of interest is whether there is any association between the probability of an ESR reading greater than 20mm/hr and the levels of the two plasma proteins. If there is not then the determination of ESR would not be useful for diagnostic purposes.

View(plasma)

Caso 1

We begin by looking at the ESR data from Table 7.1. As always it is good practise to begin with some simple graphical examination of the data before undertaking any formal modelling. Here we will look at conditional density plots of the response variable given the two explanatory variables; such plots describe how the conditional distribution of the categorical variable ESR changes as the numerical variables fibrinogen and gamma globulin change. The required R code to construct these plots is shown with Figure 7.1. It appears that higher levels of each protein are associated with ESR values above 20 mm/hr.

- Visualizacion

```
 > data("plasma", package = "HSAUR2")
> layout(matrix(1:2, ncol = 2))
> cdplot(ESR ~ fibrinogen, data = plasma)
> cdplot(ESR ~ globulin, data = plasma)
```

- Analisis: regresion logistica

```
>  plasma_glm_1 <- glm(ESR ~ fibrinogen, data = plasma,
+ family = binomial())
 > confint(plasma_glm_1, parm = "fibrinogen")
 > exp(coef(plasma_glm_1)["fibrinogen"])
 > exp(confint(plasma_glm_1, parm = "fibrinogen"))
```

- Dos variables explicativas

```
> plasma_glm_2 <- glm(ESR ~ fibrinogen + globulin,
+ data = plasma, family = binomial())

> summary(plasma_glm_2)

> anova(plasma_glm_1, plasma_glm_2, test = "Chisq")

R> plot(globulin ~ fibrinogen, data = plasma, xlim = c(2, 6),
+ ylim = c(25, 55), pch = ".")
R> symbols(plasma$fibrinogen, plasma$globulin, circles = prob,
+ add = TRUE)
```

- predicciones

```
> prob <- predict(plasma_glm_2, type = "response")
```

```
 > plot(globulin ~ fibrinogen, data = plasma, xlim = c(2, 6),
+ ylim = c(25, 55), pch = ".")
> symbols(plasma$fibrinogen, plasma$globulin, circles = prob,
+ add = TRUE)
```

Caso 2

In a survey carried out in 1974/1975 each respondent was asked if he or she agreed or disagreed with the statement "Women should take care of running their homes and leave running the country up to men". The responses are summarised in Table 7.2 (from Haberman, 1973) and also given in Collett (2003). The questions of interest here are whether the responses of men and women differ and how years of education affect the response.

View(womensrole)

- Modelo con dos variables explicativas

```
 > data("womensrole", package = "HSAUR2")
> fm1 <- cbind(agree, disagree) ~ gender + education
> womensrole_glm_1 <- glm(fm1, data = womensrole,
+ family = binomial())

summary(womensrole_glm_1)

> role.fitted1 <- predict(womensrole_glm_1, type = "response")
```

```
 > myplot <- function(role.fitted) {
  + f <- womensrole$gender == "Female"
  + plot(womensrole$education, role.fitted, type = "n", ylab =
"Probability of agreeing", xlab = "Education", ylim = c(0,1))
  + lines(womensrole$education[!f], role.fitted[!f], lty = 1)
  + lines(womensrole$education[f], role.fitted[f], lty = 2)
  + lgtxt <- c("Fitted (Males)", "Fitted (Females)")
  + legend("topright", lgtxt, lty = 1:2, bty = "n")
  + y <- womensrole$agree / (womensrole$agree +
  + womensrole$disagree)
  + text(womensrole$education, y, ifelse(f, "\\VE", "\\MA"),
  + family = "HersheySerif", cex = 1.25)
  + }
```

```
> myplot(role.fitted1)
```

- Interaccion entre educacion y genero

```
 > fm2 <- cbind(agree,disagree) ~ gender * education
> womensrole_glm_2 <- glm(fm2, data = womensrole,
+ family = binomial())
```

```
> summary(womensrole_glm_2)
```

```
 > role.fitted2 <- predict(womensrole_glm_2, type = "response")
> myplot(role.fitted2)
```

- Verificacion de los residuales

```
 > res <- residuals(womensrole_glm_2, type = "deviance")
> plot(predict(womensrole_glm_2), res,
+ xlab="Fitted values", ylab = "Residuals",
+ ylim = max(abs(res)) * c(-1,1))
> abline(h = 0, lty = 2)
```

Caso 3

Giardiello et al. (1993) and Piantadosi (1997) describe the results of a placebo-controlled trial of a non-steroidal anti-inflammatory drug in the treatment of familial andenomatous polyposis (FAP). The trial was halted after a planned interim analysis had suggested compelling evidence in favour of the treatment. The data shown in Table 7.3 give the number of colonic polyps after a 12-month treatment period. The question of interest is whether the number of polyps is related to treatment and/or age of patients.

View(polyps)

The data on colonic polyps in Table 7.3 involves *count* data. We could try to model this using multiple regression but there are two problems. The first is that a response that is a count can take only positive values, and secondly such a variable is unlikely to have a normal distribution. Instead we will apply a GLM with a log link function, ensuring that fitted values are positive, and a Poisson error distribution, i.e.,

$$\mathsf{P}(y) = \frac{e^{-\lambda}\lambda^{y}}{y!}.$$

This type of GLM is often known as *Poisson regression*. We can apply the model using

```
 > data("polyps", package = "HSAUR2")
> polyps_glm_1 <- glm(number ~ treat + age, data = polyps,
+ family = poisson())
```

```
> summary(polyps_glm_1)
```

```
 > polyps_glm_2 <- glm(number ~ treat + age, data = polyps,
+ family = quasipoisson())
> summary(polyps_glm_2)
```

Caso 4

The last of the data sets to be considered in this chapter is shown in Table 7.4. These data arise from a study reported in Kelsey and Hardy (1975) which was designed to investigate whether driving a car is a risk factor for low back pain resulting from acute herniated lumbar intervertebral discs (AHLID). A *case-control study* was used with cases selected from people who had recently had X-rays taken of the lower back and had been diagnosed as having AHLID. The controls were taken from patients admitted to the same hospital as a case with a condition unrelated to the spine. Further matching was made on age and gender and a total of 217 matched pairs were recruited, consisting of 89 female pairs and 128 male pairs. As a further potential risk factor, the variable **suburban** indicates whether each member of the pair lives in the suburbs or in the city.

View(backpain)

A frequently used design in medicine is the matched case-control study in which each patient suffering from a particular condition of interest included in the study is matched to one or more people without the condition. The most commonly used matching variables are age, ethnic group, mental status etc. A design with m controls per case is known as a $1 : m$ matched study. In many cases m will be one, and it is the $1 : 1$ matched study that we shall concentrate on here where we analyse the data on low back pain given in Table 7.4. To

With matched pairs data the form of the logistic model involves the probability, φ, that in matched pair number i, for a given value of the explanatory variable the member of the pair is a case. Specifically the model is

$$\text{logit}(\varphi_i) = \alpha_i + \beta x.$$

The odds that a subject with $x = 1$ is a case equals $\exp(\beta)$ times the odds that a subject with $x = 0$ is a case.

The model generalises to the situation where there are q explanatory variables as

$$\text{logit}(\varphi_i) = \alpha_i + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_q x_q.$$

```
> library("survival")
> backpain_glm <- clogit(I(status == "case") ~
+ driver + suburban + strata(ID), data = backpain)
```

```
> print(backpain_glm)
```