Instituto Cántabro de ESTADÍSTICA

Estadística con R

Modelo Probabilístico Lineal

Modelo Probabilístico Lineal

- Forma de la función: $Y_i=b_0+b_1X_i+e_i$
- Variable dependiente, endógena o a explicar dicotómica : Y_i,
  - Si $Y_i=0 \Rightarrow e_i = -b_0- b_1X_i$ con probabilidad p.
  - Si $Y_i=1 \Rightarrow e_i =1 -b_0- b_1X_i$ con probabilidad $1-p$.

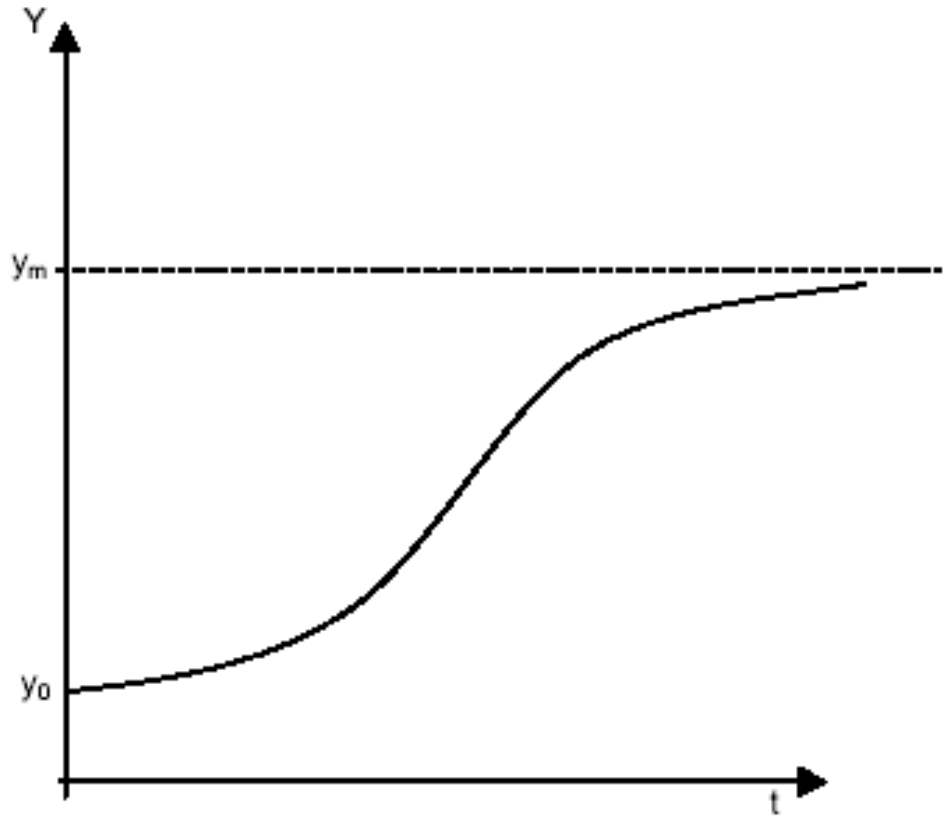
- Variable(s) independiente, exógena o explicativa cuantitativas ó cualtitativas: X_i
- Error aleatorio: e_i
- Parámetros ó coeficientes: b_0 (termino independiente) y b_1.

$$Var(e_i) = (1-\beta_0-\beta_1 X_i)(\beta_0+\beta_1 X_i) = p\cdot(1-p)$$

Estimadores MCO de estos modelos

- La perturbación aleatoria (*ei*) no sigue una distribución Normal, ya que el carácter binario (1 ó 0) de la variable endógena afecta a la distribución de la perturbación, teniendo ésta una distribución Binomial
- La perturbación aleatoria no tiene una varianza constante (es heteroscedástica)
- Las predicciones realizadas sobre la variable endógena no siempre se encuentran en el intervalo [0,1], ya que pueden ser mayores que cero y menores que uno.

Modelo LOGIT: Funcion de distribución logistica



Logit (L_i)

- La probabilidad de que $Y_i=0$ (p) se define ahora mediante la siguiente expresión:

$$p = \frac{1}{(1+e^{-z})}$$

donde $Z = b_0 + b_1X_1$, siendo b_i son los coeficientes a estimar y X_i es el vector de variable(s) independiente(s)

- La probabilidad de que $Y_i=1$ (p) :

$$(1-p) = \frac{1}{(1+e^{z})} \qquad \frac{p}{(1-p)} = \frac{(1+e^{z})}{(1+e^{-z})} = e^{z}$$

- Tomando logaritmos:

$$L_i = \ln\left[\frac{p_i}{(1-p_i)}\right] = \ln(e^{z}) = b_0 + b_1X_i$$

Estimacion Logit por GLM

- La estimación de los parámetros se realiza por máximo verosimilitud
- Para valorar el ajuste se utiliza el estadístico AIC (Akaike Information Criterion), o estadísticos derivado del anterior.
- En el caso general, la AIC es

  $AIC=-2\ln(L)/n+2k/n$

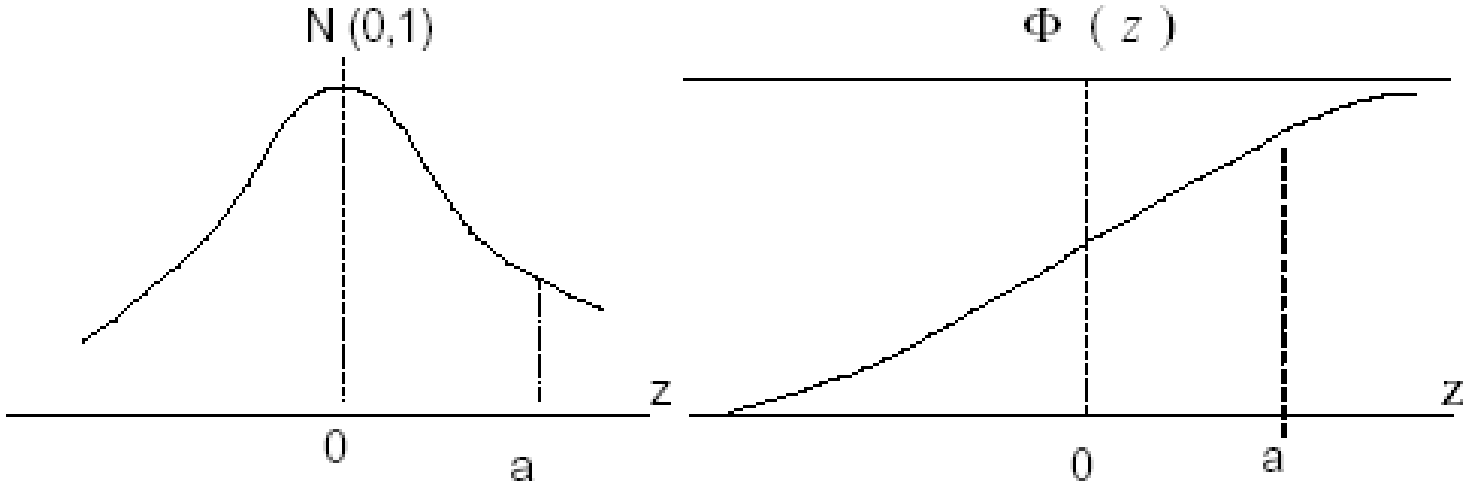
  donde *k* es el número de parámetros en el modelo estadístico, n el numero de datos , y *L* es el máximo valor de la función de verosimilitud para el modelo estimado.
- Dado un conjunto de modelos candidatos para los datos, el modelo preferido es el que tiene el valor mínimo en el AIC.
- El AIC no sólo recompensa la bondad de ajuste, sino también incluye una penalización (número de parámetros estimados). Esta penalización desalienta el sobreajute.

Instituto Cántabro de ESTADÍSTICA

Estimación Logit en R

- library(ISLR)
- attach(Default)
- fit=glm(default~.,data=Default,family=binomial)
- summary(fit)
- summary(fit$fitted.values)
- fit.pred=ifelse(fit$fitted.values>0.0333,1,0)
- table(fit.pred,Default$default)

Función Probit



Probit (I_i)

- Si $I_i=b_0+b_1X_i > s$ entonces $Y_i=1$
- Si $I_i=b_0+b_1X_i < s$ entonces $Y_i=0$
- La probabilidad de que este sea menor o igual al valor (s), se calcula a partir de la función de distribución acumulada de una distribución Normal estandarizada:

$$p_i = pr(Y=1) = pr(\beta_0 + \beta_1 X_i \leq s) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta_o+\beta_1 X_i} e^{-t/2} dt$$

- Forma más sencilla: $I_i^* = F^{-1}(I_i) = F^{-1}(p_i) = \beta_0 + \beta_1 X_i$

donde F^{-1} es la inversa de la función de distribución Normal.

Instituto Cántabro de ESTADÍSTICA

Estimación Probit en R

- fit2=glm(default~.,data=Default,family=binomial (link=probit))
- summary(fit2)
- fit2.probs=predict(fit2,type="response")
- summary(fit2.probs)
- fit2.pred=ifelse(fit2.probs>0.03348,1,0)
- table(fit2.pred,Default$default)

Instituto Cántabro de ESTADÍSTICA

Ejemplo

- A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, effect admission into graduate school. The response variable, admit/don't admit, is a binary variable.
- mydata <- read.csv("http://www.ats.ucla.edu/stat/data/binary.csv")
- head(mydata)
- alternativa (mtcars, modelo mpl para vs)

Instituto Cántabro de ESTADÍSTICA

Curva ROC

- library(ROCR)
- mylogit <- glm(admit ~ gre + gpa + rank, data = mydata, family = "binomial")
- mylogit.probs=predict(mylogit,type="response")
- predict.rocr <- prediction (mylogit.probs,mydata$admit)
- perf.rocr <- performance(predict.rocr,"tpr","fpr") #True y Tasa de falsos positivos
- auc <- as.numeric(performance(predict.rocr ,"auc")@y.values)
- plot(perf.rocr,type='o', main = paste('Area Bajo la Curva =',round(auc,2)))
- abline(a=0, b= 1)

Instituto Cántabro de ESTADÍSTICA

Minería de datos

```
# división de la muestra en entrenamiento y test
train=sample(seq(length(Default$default)),length(Default$default)*0.70,replace=FALSE)
# Estimación de modelo probit
glm.tr=glm(default[train]~.,data=Default[train,],family=binomial(link=probit))
#predicción
probs=predict.glm(glm.tr,newdata=Default[-train,],type="response")
pred=ifelse(probs>mean(probs),1,0)
table(pred,default[-train])
#gráfica curva ROC
library(ROCR)
predict.rocr  <- prediction (probs,Default$default[-train])
perf.rocr     <- performance(predict.rocr,"tpr","fpr") #True y Tasa de falsos positivos
auc <- as.numeric(performance(predict.rocr ,"auc")@y.values)
plot(perf.rocr,type='o', main = paste('Area Bajo la Curva =',round(auc,2)))
abline(a=0, b= 1)
```