Poisson regression model:

Poisson Distribution

Without offsets:

Data:(Yi,Zi), Yi:count Zi:covariate vector

Assumption:

Likelihood:

With offsets:

Data:(Yi,Zi,Xi), Xi offset

Assumption (log-linear model):

Likelihood:

Example:

Y:number of incidences of dengue fever(登革熱發生人數); X:population in different areas or time

install.packages("ggplot2")

install.packages ("sandwich")

install.packages ("msm")

require("ggplot2")

require("sandwich")

require("msm")

#p <- read.csv("http://www.ats.ucla.edu/stat/data/poisson\_sim.csv")

p <- read.csv("J:/poisson\_sim.csv")

p <- within(p, {

prog <- factor(prog, levels=1:3, labels=c("General", "Academic", "Vocational"))

id <- factor(id)

})

summary(p)

summary(m1 <- glm(num\_awards ~ prog + math, family="poisson", data=p))

Poisson regression (with offset)

Source: http://rfunction.com/archives/223

Example. Below a Poisson generalized linear model (GLM) is created, where the variable Y is the number of events. We assume Yi~Poisson( )

We also have N as the size of each group, and x1 and x2 are predictors

R code

Y <- c(15, 7, 36, 4,

16, 12, 41, 15)

N <- c(4949, 3534, 12210, 344,

6178, 4883, 11256, 7125)

x1 <- c(-0.1, 0, 0.2, 0,

1, 1.1, 1.1, 1)

x2 <- c(2.2, 1.5, 4.5, 7.2,

4.5, 3.2, 9.1, 5.2)

glm(Y ~ offset(log(N)) + (x1 + x2), family=poisson)

#glm(Y ~ offset(log(N)) + I(x1+x2), family=poisson)

Finding the optimal cutoff point

Finding the optimal cutoff point for a biomarker is important in biomedical research.

Examples:

Here we consider two types of outcomes, binary and survival.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3522617/

以下摘錄自這篇文章"Cutoff Finder: A Comprehensive and Straightforward Web Application..."(以上為link)

"

1. Fit of mixture model: A mixture model of two Gaussian distributions is fitted to the histogram of the biomarker. This procedure is implemented using the function *flexmix* from the R package flexmix [[16]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3522617/#pone.0051862-Leisch1). The optimal cutoff is determined as the value where the probability density functions of the mixing distribution coincide.
2. Significance of correlation with binary variable: This method correlates the dichotomized biomarker with a binary outcome variable using logistic regression. Logistic regression is executed using the function *glm* from R package stats [[15]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3522617/#pone.0051862-R1). The optimal cutoff is defined as the point with the most significant (Fisher’s exact test) split. Odds ratios (ORs) as well as sensitivity and specificity including 95% confidence intervals are calculated. Confidence intervals for proportions are estimated using Wilson’s method as it is implemented in the R package binom [[17]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3522617/#pone.0051862-DoraiRaj1).
3. Based on ROC curve: Two methods determine the cutoff point by minimizing the distance on the ROC curve to the left top edge of the diagram. The first method minimizes the Euclidean distance between these points. The second method minimizes the Manhattan distance between the points. Here, the sum of sensitivity and specificity is maximized, equivalent to maximization of Youden’s statistics J = sensitivity+specificity –1 [[18]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3522617/#pone.0051862-Youden1).
4. Minimum sensitivity or specificity: For each or these two methods, the user enters a percentage value. The cutoff point is chosen as the first threshold for the biomarker, where the sensitivity (or specificity) exceeds this predefined value.
5. Significance of correlation with survival variable: This method fits Cox proportional hazard models to the dichotomized variable and the survival variable. Survival analysis is executed using the functions *coxph* and *survfit* from the R package survival [[19]](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3522617/#pone.0051862-Therneau1). The optimal cutoff is defined as the point with the most significant (log-rank test) split. Hazard ratios (HRs) including 95% confidence intervals are calculated."

(Without considering the outcome variable)

Finite Mixture model:

Model:

Parameters:

EM algorithm to estimate parameters:

hist(faithful$waiting)

install.packages("mixtools")

library(mixtools)

wait = faithful$waiting

mixmdl = normalmixEM(wait)

plot(mixmdl,which=2)

lines(density(wait), lty=2, lwd=2)

summary(mixmdl)

ROC curve and logistic regression:

Logistic regression model

Binary Data:

Covariates

Model:

Likelihood:

mydata <- read.csv("http://www.ats.ucla.edu/stat/data/binary.csv")

mylogit <- glm(admit ~ gre, data = mydata, family = "binomial")

summary(mylogit)

prob=predict(mylogit,type=c("response"))

mydata$prob=prob

install.packages("pROC")

library(pROC)

g <- roc(admit ~ prob, data = mydata)

plot(g)

mydata$prob=prob

How to find the optimal cutoff point

Minimize the distance:

Given the minimum value of specificity or sensitivity:

Cross Validation:

Five folds or leave one out:

報告要求:

1.報告檔案

描述問題 利用的統計方法 結論(包含解釋)

2. Codes