DISCUSSION 8

Coupon Data(coupons.txt)

We will reuse the data on redemption rate of coupons last week. The goal is to use simulations to determine whether 2 models (separate regions versus combined regions) provide the same significance.

1. Read the data in the file coupons.txt, then add a column with the number of total coupons. Make sure to use the original data set with one row for each group of 10 coupons. Fit a logistic regression on the full data set (all regions combined) including all variables and all possible 2-way and 3-way interactions, assuming no overdispersion. Report the deviance of the model, and test the goodness-of-fit of this model by comparing the deviance to the appropriate chi-square distribution.

2. Use a parametric bootstrap approach (instead of the chi-square distribution) to test the goodness-of-fit of the model in the previous question. Define a function to simulate the experiment, and repeat it 1000 times. Report the simulated null distribution and p-value. Compare these 2 approaches. Which one is more reliable?

3. Create a data frame which only contains the data from the region C (still one row for each group of 10 coupons). Fit a logistic model to these data using price, gender, and their interaction. Test the goodness-of-fit for this model using a simulation approach. Also apply the chi-square approach and determine how reliable it is in this case.

4. Fit the region C data with a quasi-binomial analysis. How does the statistical significance of the interaction and of the main effects change, quantitatively and qualitatively? You may summarize your answers in a table, to compare the results from the logistic model to the results from the quasi-binomial analysis.

5. Re-use the simulation in Question 3, to determine the p-value associated to the goodness-of-fit test on the region C data. But this time, do it 4 different times, based on each of 3 different numbers of simulated in silico experiments: based on 20 simulated experiments (4 different times), 200 and 2,000 simulated experiments (4 different times each). Report your 12 resulting p-values in a table.

6. Try different sample size for each group in the simulation in Question 2. The goal is to determine how large individual sample sizes need to be for the chi-square distribution (on the residual df) to be a good approximation of the residual deviance, under the null hypothesis that the logistic model is correct.
Solutions

library(lattice)
library(MASS)

# 1
data = read.table("coupons.txt", header=T)
data$total=10
data$prop=data$redeemed/10
str(data)

full = glm(prop ~ price*gender*region, weights=total, data=data, family=binomial)
summary(full)

# 2
fitted.prop = fitted(full) # or predict(full, type='response')
simulate.deviance = function(){
  sim.data = with(data, data.frame(
    price = price,
    gender = gender,
    region = region,
    rep = rep,
    total = 10,
    redeemed = rbinom(120, size=10, prob=fitted.prop)
  ))
  sim.fit = glm(redeemed/10 ~ price*gender*region,
    weight=total, data=sim.data, family=binomial)
  return( sim.fit$deviance )
}
simulate.deviance()

sim1000 = replicate(1000, simulate.deviance())
hist(sim1000, xlim=c(60,200), ylim=c(0,300))
points(187.74, 0, col="red", pch=16, cex=2)
curve(dchisq(x,df=108)*1000*10, add=T)
sum(sim1000>187.74)/1000
pchisq(187.74, 108, lower.tail=F)

# 3
dataC=subset(data, region=='C')
fitC = glm(prop ~ price*gender, weights=total, data=dataC, family=binomial)
summary(fitC)
fitted.prop.C=fitted(fitC)
simulate.deviance.C = function(){
sim.data=within(dataC, redeemed <- rbinom(40, size=10, prob=fitted.prop.C))
sim.fit=glm(redeemed/10 ~ price*gender, weight=total, data=sim.data, family=binomial)
return( sim.fit$deviance )
}
simulate.deviance.C()

sim1000.C = replicate(1000, simulate.deviance.C())
hist(sim1000.C)
points(38.15, 0, col="red", pch=16,cex=2)
curve(dchisq(x,df=36)*1000*5, add=T)
sum(sim1000.C>38.15)/1000
pchisq(38.15 , 36,lower.tail=F)

# 4
fitC.quasi = glm(prop ~ price*gender, weights=total, data=dataC, family=quasibinomial)
summary(fitC.quasi)
summary(fitC)

# 5
for (simno in c(20,200,2000)) {
  cat('simno = ', simno,'
  for (j in 1:4) {
    sim.C = replicate(simno, simulate.deviance.C())
    print(sum(sim.C>38.15)/simno)
  }
}

# 6
simulate.deviance = function(size=10){
  sim.data=within(data, {
    redeemed <- rbinom(120, size=size, prob=fitted.prop)
    total = size
  })
  sim.fit = glm(redeemed/size ~ price*gender*region,
                weight=total, data=sim.data, family=binomial)
  return( sim.fit$deviance )
}
simulate.deviance(20)

par(mfrow=c(2,2))
for (size in c(10,20,50,100)) {
  sim1000 = replicate(1000, simulate.deviance(size))
  hist(sim1000, freq=F, main=paste('size=',size), ylim=c(0, 0.03))
  curve(dchisq(x,df=108), add=T)
}