##1、三维列联表loglinearmodel

setwd("D:/R\_Files/CDA\_Weiyang")

data0 = read.csv("D:/R\_Files/CDA\_Weiyang/Datasets/subdata1.csv",header=T)

data1=data.frame(data0)

n<-data1$count

FansLevel<-data1$FansLevel

gender<-data1$gender

type<-data1$upType

bili.llm<-glm(formula=n~FansLevel\*gender\*type,family=poisson,data=data1)

##log linear model

summary(bili.llm)

bili.llmd1<-glm(formula=n~FansLevel\*gender\*type,family=poisson,data=data1)

bili.llmd1n<-glm(formula=n~FansLevel+gender+type+FansLevel\*gender+FansLevel\*type+gender\*type,family=poisson,data=data1)

summary(bili.llmd1)

drop1(bili.llmd1)

bili.llmd2<-glm(formula=n~FansLevel+gender+type+FansLevel\*type+gender\*type,family=poisson,data=data1)

summary(bili.llmd2)

drop1(bili.llmd2)

bili.llmd3<-glm(formula=n~FansLevel+gender+type+gender\*type,family=poisson,data=data1)

summary(bili.llmd3) ##最终模型

fits3<-fitted(bili.llmd3)

resids3<-residuals(bili.llmd3)

h3<-lm.influence(bili.llmd3)$hat

adjresids3<-resids3/sqrt(1-h3)

round(cbind(bili.llmd3$n,fits3,adjresids3),2)

##2、试试多分类累计逻辑回归(因为响应变量粉丝数Y是有序的)

library(VGAM)

little<-c(8686,4270,20433,5804,10616,4383)

middle<-c(925,443,1919,583,1280,541)

big<-c(146,72,311,87,225,79)

superbig<-c(117,46,234,59,170,44)

gender<-c(0,1,0,1,0,1)

type<c("anime","anime","game","game","knowledge","knowledge")

type1<-c(1,1,0,0,0,0)

type2<-c(0,0,1,1,0,0)

datagaf=data.frame(gender,type1,type2,little,middle,big,superbig)

cumlogitm<-vglm(cbind(little,middle,big,superbig)~gender+type1+type2+gender\*type1+gender\*type2,family=cumulative(parallel=TRUE))

summary(cumlogitm)

cumlogitm2<-vglm(cbind(little,middle,big,superbig)~gender+type1+type2+gender\*type1+gender\*type2,family=cumulative(parallel=TRUE))

summary(cumlogitm2) ##gender影响不显著

cumlogitm3<-vglm(cbind(little,middle,big,superbig)~gender+type1+type2,family=cumulative(parallel=TRUE))

cumlogitm4<-vglm(cbind(little,middle,big,superbig)~type1+type2,family=cumulative(parallel=TRUE))

summary(cumlogitm4)

anova.vglm(cumlogitm4,cumlogitm,type='I')

fitclm<-fitted(cumlogitm)

##3、列联表卡方检验

setwd("D:/R\_Files/CDA\_Weiyang")

data2 = read.csv("D:/R\_Files/CDA\_Weiyang/Datasets/genderandfanlevel.csv",header=T)

n<-data2$count

gen<-data2$gender

fan<-data2$fanlevel

gaf<-xtabs(n~gen+fan,data=data2) ##总表

gaf.chisq=chisq.test(gaf)

G2=with(gaf.chisq,2\*sum(observed\*log(observed/expected)))

G2 # G2=77.10952 #拒绝两者独立的假设

##三个子表

GAF<-array(c(8686,4270,925,443,146,72,117,46, ##anime

20433,5804,1919,583,311,87,234,59, ##game

10616,4383,1280,541,225,79,170,44), ##knowledge

dim=c(2,4,3),

dimnames=list(

gender=c("Man","Woman"),

fanslevel=c("<10w","10w-50w","50w-100w",">100w"),

channel=c("anime","game","knowledge")))

dfanime<-array(c(8686,4270,925,443,146,72,117,46),

dim=c(2,4),

dimnames=list(gender=c("Man","Woman"),

fanslevel=c("1","2","3","4")))

anime.chisq<-chisq.test(dfanime)

anime.chisq

G2anime=with(anime.chisq,2\*sum(observed\*log(observed/expected)))

tanime<-table(dfanime$fanslevel)

kruskal.test(fanslevel~gender,data=tanime)

## X-squared = 1.7898, df = 3, p-value = 0.6172 ,G2=1.832407

dfgame<-array(c(20433,5804,1919,583,311,87,234,59),

dim=c(2,4),

dimnames=list(gender=c("Man","Woman"),

fanslevel=c("<10w","10w-50w","50w-100w",">100w")))

game.chisq<-chisq.test(dfgame)

game.chisq ## X-squared = 2.6001, df = 3, p-value = 0.4575

G2game=with(game.chisq,2\*sum(observed\*log(observed/expected)))

G2game #2.596815

dfknowledge<-array(c(10616,4383,1280,541,225,79,170,44),

dim=c(2,4),

dimnames=list(gender=c("Man","Woman"),

fanslevel=c("<10w","10w-50w","50w-100w",">100w")))

knowledge.chisq<-chisq.test(dfknowledge)

knowledge.chisq #X-squared = 9.4242, df = 3, p-value = 0.02415

G2knowledge=with(knowledge.chisq,2\*sum(observed\*log(observed/expected)))

G2knowledge # 9.991834 #有辛普森悖论