#count data

# regression

clusters <- read.csv("c:\\temp\\clusters.csv")

attach(clusters)

names(clusters)

plot(Distance,Cancers,pch=21,bg="lightblue")

model1 <- glm(Cancers~Distance,poisson)

summary(model1)

model2 <- glm(Cancers~Distance,quasipoisson)

summary(model2)

xv <- 0:100

yv <- 0.186865-0.006138\*xv

y <- exp(yv)

lines(xv,y,col="red")

y <- predict(model2,list(Distance=xv), type="response")

lines(xv,y,col="red")

# categorical explanatory vaariables

count <- read.csv("c:\\temp\\cells.csv")

attach(count)

names(count)

table(cells)

tapply(cells,smoker,mean)

tapply(cells,weight,mean)

tapply(cells,sex,mean)

tapply(cells,age,mean)

model1 <- glm(cells~smoker\*sex\*age\*weight,poisson)

summary(model1)

model2 <- glm(cells~smoker\*sex\*age\*weight,quasipoisson)

summary(model2)

model3 <- update(model2, ~. -smoker:sex:age:weight)

model4 <- update(model3, ~. -sex:age:weight)

anova(model4,model3,test="F")

model5 <- update(model4, ~. -smoker:sex:age)

anova(model5,model4,test="F")

model6 <- update(model5, ~. -smoker:age:weight)

anova(model6,model5,test="F")

Despite 1-star significance for one of the interaction terms, this was not significant either, so we leave it out.

model7 <- update(model6, ~. -smoker:sex:weight)

anova(model7,model6,test="F")

model8 <- update(model7, ~. -smoker:age)

anova(model8,model7,test="F")

model9 <- update(model8, ~. -sex:weight)

anova(model9,model8,test="F")

model10 <- update(model9, ~. -age:weight)

anova(model10,model9,test="F")

model11 <- update(model10, ~. -smoker:sex)

anova(model11,model10,test="F")

model12 <- update(model11, ~. -sex:age)

anova(model12,model11,test="F")

model13 <- update(model11, ~. -smoker:weight)

anova(model13,model11,test="F")

tapply(cells,list(smoker,weight),mean)

tapply(cells,list(sex,age),mean)

barplot(tapply(cells,list(smoker,weight),mean),beside=T)

weight <- factor(weight,c("normal","over","obese"))

barplot(tapply(cells,list(smoker,weight),mean),beside=T)

barplot(tapply(cells,list(smoker,weight),mean),beside=T)

legend(locator(1),c("non smoker","smoker"),fill=gray(c(0.2,0.8)))

# comlex contingency tables

induced <- read.csv("c:\\temp\\induced.csv")

attach(induced)

names(induced)

model <- glm(Count~Tree\*Aphid\*Caterpillar,family=poisson)

model2 <- update(model , ~ . - Tree:Aphid:Caterpillar)

anova(model,model2,test="Chi")

model3 <- update(model2 , ~ . - Aphid:Caterpillar)

anova(model3,model2,test="Chi")

# the wrong way of doing it

wrong <- glm(Count~Aphid\*Caterpillar,family=poisson)

wrong1 <- update(wrong,~. - Aphid:Caterpillar)

anova(wrong,wrong1,test="Chi")

tapply(Count,list(Tree,Caterpillar),sum)

# ancova with count data

species <- read.csv("c:\\temp\\species.csv")

attach(species)

names(species)

plot(Biomass,Species,pch=21,bg=(1+as.numeric(pH)))

model <- lm(Species~Biomass\*pH)

summary(model)

abline(40.60407,-2.80045,col="red")

abline(40.60407-22.75667,-2.80045-0.02733,col="green")

abline(40.60407-11.57307,-2.80045+0.23535,col="blue")

model <- glm(Species~Biomass\*pH,poisson)

summary(model)

model2 <- glm(Species~Biomass+pH,poisson)

anova(model,model2,test="Chi")

plot(Biomass,Species,pch=21,bg=(1+as.numeric(pH)))

xv <- seq(0,10,0.1)

length(xv)

acidity <- rep("low",101)

yv <- predict(model,list(Biomass=xv,pH=acidity),type="response")

lines(xv,yv,col="green")

acidity <- rep("mid",101)

yv <- predict(model,list(Biomass=xv,pH=acidity),type="response")

lines(xv,yv,col="blue")

acidity <- rep("high",101)

yv <- predict(model,list(Biomass=xv,pH=acidity),type="response")

lines(xv,yv,col="red")

# frquency distributions

case.book <- read.csv("c:\\temp\\cases.csv")

attach(case.book)

names(case.book)

frequencies <- table(cases)

frequencies

mean(cases)

windows(7,4)

par(mfrow=c(1,2))

barplot(frequencies,ylab="Frequency",xlab="Cases",

col="red",main="observed")

barplot(dpois(0:10,1.775)\*80,names=as.character(0:10),

ylab="Frequency",xlab="Cases",col="blue",main="expected")

var(cases)/mean(cases)

negbin <- function(x,u,k)

(1+u/k)^(-k)\*(u/(u+k))^x\*gamma(k+x)/(factorial(x)\*gamma(k))

xf <- numeric(11)

for (i in 0:10) xf[i+1] <- negbin(i,0.8,0.2)

barplot(xf)

mean(cases)^2/(var(cases)-mean(cases))

expected <- dnbinom(0:10,size=0.8898,mu=1.775)\*80

both <- numeric(22)

both[1:22 %% 2 != 0] <- frequencies

both[1:22 %% 2 == 0] <- expected

labels <- character(22)

labels[1:22 %% 2 == 0] <- as.character(0:10)

barplot(both,col=rep(c("lightgray","darkgray"),11),names=labels,ylab="Frequency",xlab="Cases")

legend(locator(1),c("Observed","Expected"), fill=c("lightgray","darkgray"))

cs <- factor(0:10)

levels(cs)[6:11] <- "5+"

levels(cs)

ef <- as.vector(tapply(expected,cs,sum))

of <- as.vector(tapply(frequencies,cs,sum))

sum((of-ef)^2/ef)

1 - pchisq(2.581842,3)