

# 1. Exercise 1 – Agresti 7.36

Table 1 is based on a study involving British doctors.

Age	Person-Years		Coronary Deaths	
	Nonsmokers	Smokers	Nonsmokers	Smokers
35 – 44	18,793	52,407	2	32
45 – 54	10,673	43,248	12	104
55 – 64	5,710	28,612	28	206
65 – 74	2,585	12,663	28	186
75 – 84	1,462	5,317	31	102

Table 1: Data on Coronary Death Rates

- (a) Fit a main effects model for the log rates using age and smoking as factors. In discussing lack-of-fit, show that this model assumes a constant ratio of nonsmokers' to smokers' coronary death rates over age, and evaluate how the sample ratio depends on age.

```
smokers<-data.frame(  c( 35 , 44 , 18793 , 52407 , 2 , 32 ),
  c( 45 , 54 , 10673 , 43248 , 12 , 104 ),  c( 55 , 64 , 5710 , 28612 , 28 , 206 ),
  c( 65 , 74 , 2585 , 12663 , 28 , 186 ),    c( 75 , 84 , 1462 , 5317 , 31 , 102 ) )

smokers<-t(as.matrix(smokers));smokers<-as.data.frame(smokers)
rownames(smokers)<-paste0( smokers$V1,"-", smokers$V2); smokers<-smokers[,-c(1:2)]
names(smokers)<-c(paste0("PY",c("nonsmokers", "smokers")),
  paste0("CD",c("nonsmokers", "smokers")))

smokers<-data.frame(rep(1:5,2),c(rep("NS",5),rep("S",5))      ,
  c(smokers$PYnonsmokers,smokers$PYsmokers),
  c(smokers$CDnonsmokers,smokers$CDsmokers))

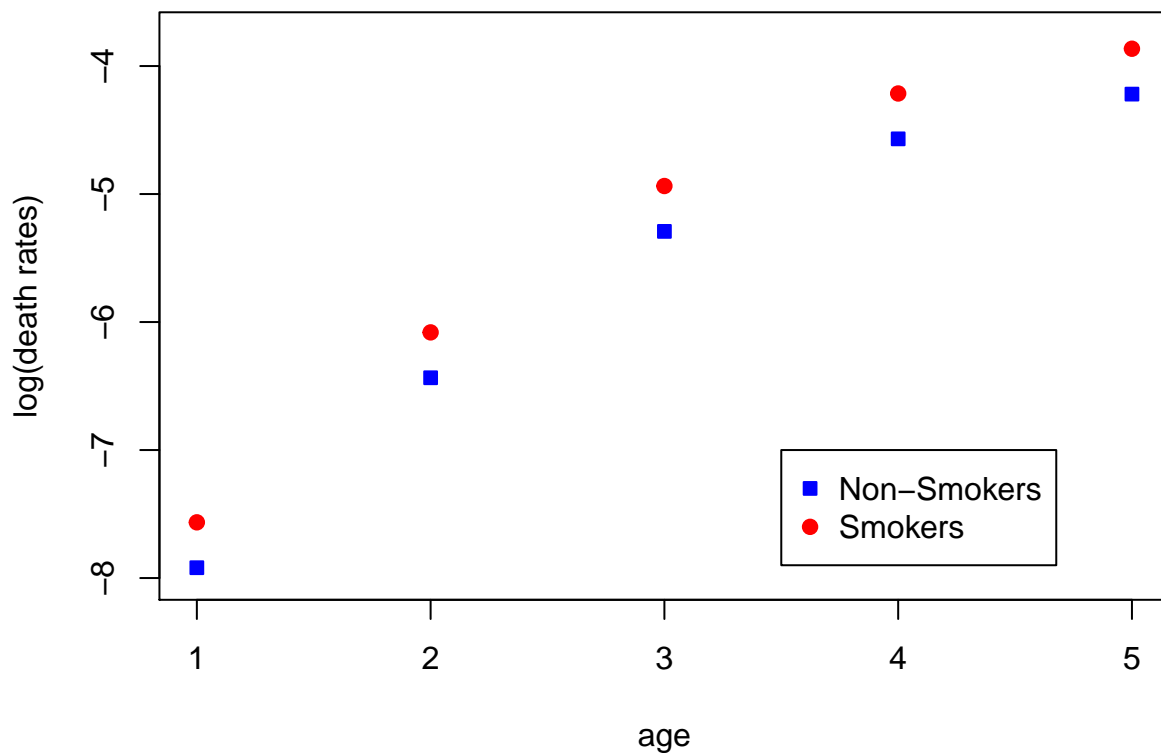
names(smokers)<-c("age", "smoker", "PersonYears", "Deaths")
smokers$age<-as.factor(smokers$age);smokers$ageQI<-as.numeric(smokers$age)
smokers$ratios<-smokers[,4]/smokers[,3];str(smokers)

## 'data.frame':    10 obs. of  6 variables:
##  $ age          : Factor w/ 5 levels "1","2","3","4",...: 1 2 3 4 5 1 2 3 4 5
##  $ smoker       : Factor w/ 2 levels "NS","S": 1 1 1 1 1 2 2 2 2 2
##  $ PersonYears: num  18793 10673 5710 2585 1462 ...
##  $ Deaths      : num   2 12 28 28 31 32 104 206 186 102
##  $ ageQI        : num   1 2 3 4 5 1 2 3 4 5
##  $ ratios       : num  0.000106 0.001124 0.004904 0.010832 0.021204 ...

smokers.fit<- glm(Deaths ~ age+smoker, offset = log(PersonYears), family=poisson, data=smokers)
summary(smokers.fit)

##
## Call:
## glm(formula = Deaths ~ age + smoker, family = poisson, data = smokers,
##      offset = log(PersonYears))
##
## Deviance Residuals:
##      1       2       3       4       5       6       7       8       9      10
## -2.18005  -1.30797  -0.13786   0.22886   1.91906   0.90176   0.51036   0.05133  -0.08734  -0.91239
##
```

```
## Coefficients:
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -7.9194     0.1918 -41.298 < 2e-16 ***
## age2          1.4840     0.1951   7.606 2.82e-14 ***
## age3          2.6275     0.1837  14.301 < 2e-16 ***
## age4          3.3505     0.1848  18.131 < 2e-16 ***
## age5          3.7001     0.1922  19.250 < 2e-16 ***
## smokerS       0.3545     0.1074   3.302 0.00096 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
## Null deviance: 935.091  on 9  degrees of freedom
## Residual deviance:  12.134  on 4  degrees of freedom
## AIC: 79.202
##
## Number of Fisher Scoring iterations: 4
par(mar=c(5.1,4.1,2.1,2.1))
plot( ( smokers.fit$linear.predictors[1:5] ) - log(smokers$PersonYears)[1:5],
      ylim=c(-8,-3.75) , pch=22 ,col="blue", bg ="blue",
      ylab="log(death rates)",xlab="age")
points(smokers.fit$linear.predictors[6:10] - log(smokers$PersonYears)[6:10],
       pch=21 , bg ="red",col="red")
legend(3.5,-7,c("Non-Smokers", "Smokers"), pch = c(22,21), col=c("blue","red") , pt.bg=c("blue","red"))
```



We note the rate of death is modeled by:

$$\log\left(\frac{\mu_i}{t_i}\right) = \mathbf{x}_i^T \boldsymbol{\beta}.$$

R is using a nequivalent offset for calculating the linear predictors,

$$\log(\mu_i) = \log(t_i) + \mathbf{x}_i^T \boldsymbol{\beta}.$$

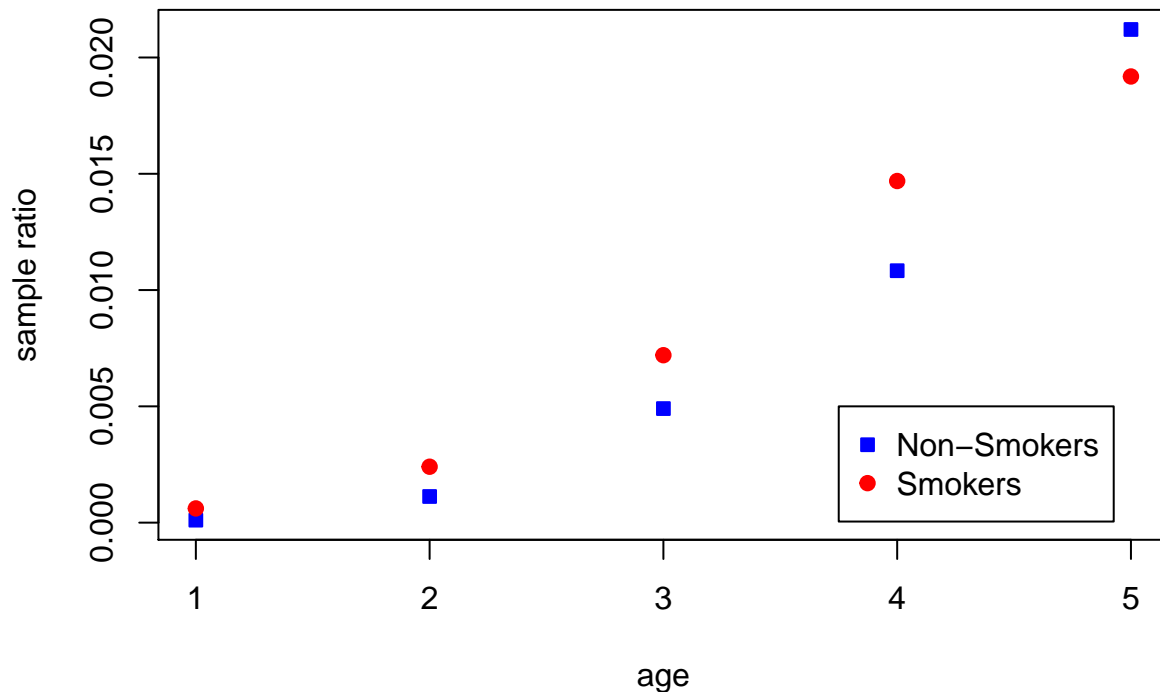
So we move the  $\log(t_i)$  term back to the LHS to show the constant ratio of coronary deaths between nonsmokers to smokers. (Note: even if we kept the  $\log(t_i)$  term on the RHS, the model would still have a constant ratio between nonsmokers to smokers coronary death counts).

We see the lack-of-fit with the model:

```
1-pchisq(12.134 , 4 )
```

```
## [1] 0.01638213
```

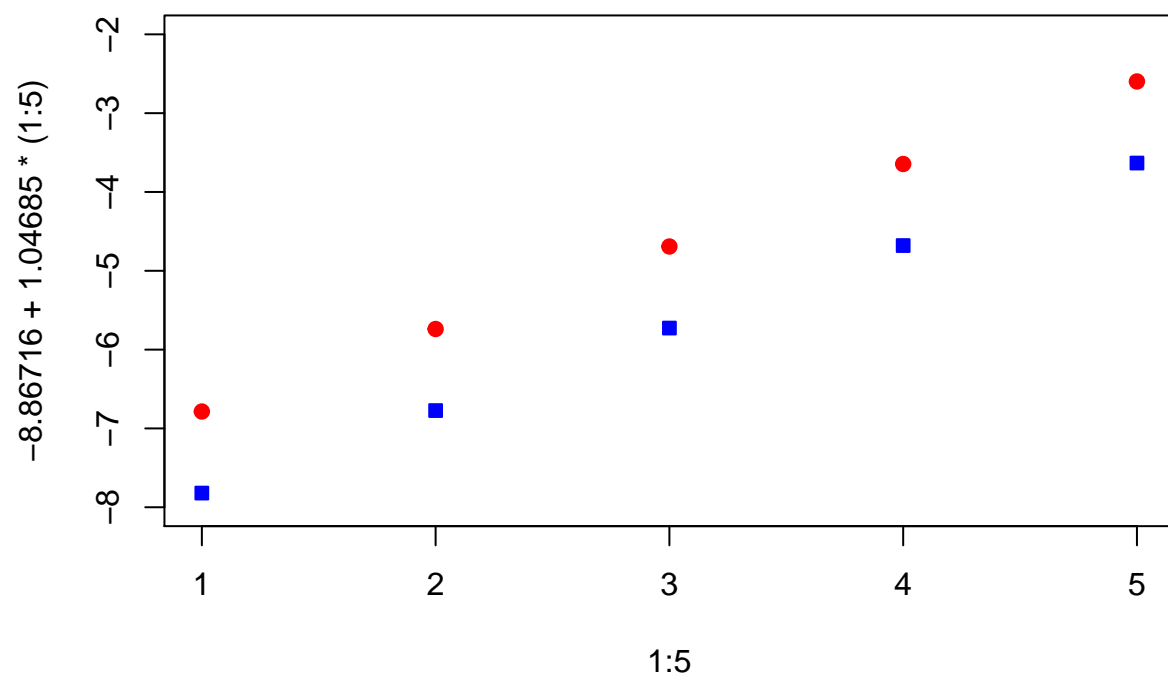
```
par(mar=c(5.1,4.1,1.1,2.1))
plot(1:5,smokers$ratios[1:5] , pch=22 ,col="blue", bg ="blue", ylab="sample ratio",xlab="age")
points(1:5,smokers$ratios[6:10],pch=21 , bg ="red",col="red")
legend(3.75,0.005,c("Non-Smokers", "Smokers"), pch = c(22,21), col=c("blue","red") , pt.bg=c("blue","red"))
```



- (b) Explain why it is sensible to add a quantitative interaction of age and smoking. For this model, show that the log ratio of coronary death rates changes linearly with age. Assign scores to age, fit the model, and interpret.

```
smokersQI.fit<- glm(Deaths ~ ageQI*smoker, offset = log(PersonYears),family=poisson, data=smokers)
summary(smokersQI.fit)
```

```
##
## Call:
## glm(formula = Deaths ~ ageQI * smoker, family = poisson, data = smokers,
##      offset = log(PersonYears))
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -3.8784  -2.1219  -0.2482   1.7184   3.5269
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -8.86716    0.30567  -29.009  < 2e-16 ***
## ageQI         1.04685    0.07743   13.520  < 2e-16 ***
## smokerS       1.28369    0.32583    3.940 8.16e-05 ***
## ageQI:smokerS -0.24899    0.08359   -2.979  0.00289 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 935.091  on 9  degrees of freedom
## Residual deviance:  59.895  on 6  degrees of freedom
## AIC: 122.96
##
## Number of Fisher Scoring iterations: 4
plot(1:5,-8.86716 + 1.04685*(1:5) , ylim=c(-8,-2), pch=22 ,col="blue", bg ="blue")
points(1:5,-8.86716 + 1.04685*(1:5) +1.284  -0.249,pch=21 , bg ="red",col="red")
```



## 2. Exercise 2

One question in the 1990 General Social Survey asked subjects how many times they had sexual intercourse in the preceding month. Table 2 shows responses classified by gender.

Response	Male	Female	Response	Male	Female	Response	Male	Female
0	65	128	9	2	2	20	7	6
1	11	17	10	24	13	22	0	1
2	13	23	12	6	10	23	0	1
3	14	16	13	3	3	24	1	0
4	26	19	14	0	1	25	1	3
5	13	17	15	3	10	27	0	1
6	15	17	16	3	1	30	3	1
7	7	3	17	0	1	50	1	0
8	21	15	18	0	1	60	1	0

Table 2: Data from the 1990 General Social Survey

- (a) Fit a Poisson GLM with log link and a dummy variable for gender (1=males, 0=females) and explain if the model seems appropriate.

```
setwd("G:\\math\\661")
dat<-read.csv("sex.csv")
dat<-data.frame(
  (rep(dat$Response,2)),
  c(dat$Male,dat$Female),
  as.factor(c(rep(1,nrow(dat)),rep(0,nrow(dat)))) )
names(dat)<-c("response","counts","gender")
str(dat)

## 'data.frame': 54 obs. of 3 variables:
## $ response: int 0 1 2 3 4 5 6 7 8 9 ...
## $ counts : int 65 11 13 14 26 13 15 7 21 2 ...
## $ gender : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...

cbind(head(dat),tail(dat))
```

```
## response counts gender response counts gender
## 1 0 65 1 24 0 0
## 2 1 11 1 25 3 0
## 3 2 13 1 27 1 0
## 4 3 14 1 30 1 0
## 5 4 26 1 50 0 0
## 6 5 13 1 60 0 0
```

```
dat.fit<-glm(response ~ gender, family=poisson, weights=counts, data=dat)
summary(dat.fit)
```

```
##
## Call:
## glm(formula = response ~ gender, family = poisson, data = dat,
## weights = counts)
##
## Deviance Residuals:
```

```
##      Min      1Q   Median      3Q      Max
## -33.191   0.000    3.437    6.126   13.430
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.45936    0.02738  53.302 < 2e-16 ***
## gender1      0.30850    0.03822   8.071 6.95e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 4050.8  on 44  degrees of freedom
## Residual deviance: 3985.7  on 43  degrees of freedom
## AIC: 5271.3
##
## Number of Fisher Scoring iterations: 6
tab<-cbind(dat[which(dat$gender == 0),],dat[which(dat$gender == 1 ),-1])
tab<-tab[,c(1,2,4)];tab[19,2]<-sum(tab[19:nrow(tab),2]);
tab[19,3]<-sum(tab[19:nrow(tab),3]);tab<-tab[1:19,]
names(tab)[2:3]<-c("Female","Male")
head(tab)

##      response Female Male
## 28          0    128   65
## 29          1     17   11
## 30          2     23   13
## 31          3     16   14
## 32          4     19   26
## 33          5     17   13

c(sum(tab[,2]),sum(tab[,1]*tab[,2]));    sum(tab[,1]*tab[,2])/sum(tab[,2])

## [1] 310 1297
## [1] 4.183871
sum( tab[,2]*((tab[,1]- 4.183871 )^2) ) / ( sum(tab[,2]) -1)

## [1] 29.76867
c(sum(tab[,3]),sum(tab[,1]*tab[,3]));    sum(tab[,1]*tab[,3])/sum(tab[,3])

## [1] 240 1297
## [1] 5.404167
sum( tab[,3]*((tab[,1]- 4.183871 )^2) ) / ( sum(tab[,3]) -1)

## [1] 31.30203
1-pchisq(3985.7,43)

## [1] 0
```

The sample mean for the 1297 women is 4.183871 with a variance of 29.76867. The sample mean for the 1297 men is 5.404167 with a variance of 31.30203. In both groups the sample variances are about

6-7 times the size of the sample means. This is suggesting overdispersion relative to the Poisson. We also see that the model does not give a good fit to the data ( $p$ -value  $\approx 0$ ).

- (b) Interpret the regression coefficient of gender for the model in (a) and provide a 95% Wald confidence interval for the ratio of means for males versus females.

```
exp(0.30850 -1.96*0.03822);exp(0.30850 +1.96*0.03822)
```

```
## [1] 1.263125
```

```
## [1] 1.467281
```

When gender is male, the count of sexual intercourse is estimated to be 1.36 times that of females ( $e^{0.30850}$ ). The Wald 95% confidence interval for the ratio of means for males versus females is:

$$\exp(0.30850 \pm 1.96 \cdot 0.03822) = (1.263125, 1.467281)$$

- (c) Fit a negative binomial model. Is there evidence of overdispersion? What is the estimated difference in log means, its standard error, and the 95% Wald confidence interval for the ratio of means.

```
library(MASS)
nb.fit<-glm.nb(response ~ gender, weights=counts, data=dat)
summary(nb.fit)
```

```
##
## Call:
## glm.nb(formula = response ~ gender, data = dat, weights = counts,
##       init.theta = 0.5018752366, link = log)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -17.0366   0.0000   0.9873   1.5894   3.4336
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.45936    0.08472  17.226  <2e-16 ***
## gender1      0.30850    0.12724   2.425  0.0153 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.5019) family taken to be 1)
##
##      Null deviance: 606.53  on 44  degrees of freedom
## Residual deviance: 600.60  on 43  degrees of freedom
## AIC: 2883
##
## Number of Fisher Scoring iterations: 1
##
##
##              Theta:  0.5019
```



```
##          Std. Err.:  0.0387
##
##  2 x log-likelihood:  -2876.9770
1-pchisq(600,43)

## [1] 0
```

We note that  $\widehat{Var}(Y) = \hat{\mu} + \hat{\gamma}\hat{\mu}^2$  is actually overestimating the sample variances. For females the sample variance is 29.76867, and the negative binomial model is estimating  $4.303205 + \left(\frac{1}{0.5019}\right) * 4.303205^2 = 41.19815$ . Likewise for males, the sample variance is 31.30203, and the negative binomial model is estimating  $5.858303 + \left(\frac{1}{0.5019}\right) * 5.858303^2 = 74.23789$ .

There is evidence that  $\hat{\gamma} > 0$ ;  $\hat{\gamma} = \left(\frac{1}{0.5019}\right) = 1.992429$  and a 95% confidence interval for  $\gamma$  is given by:

$$\frac{1}{0.5019 \pm 1.96 \cdot 0.0387} = (1.730846, 2.347153).$$

So the extra parameter is picking up some of the dispersion compared with the Poisson. But we recall that the negative binomial approaches the Poisson as  $\gamma \rightarrow 0$ , so there might be more overdispersion unaccounted for. We test the fit of the negative binomial model:

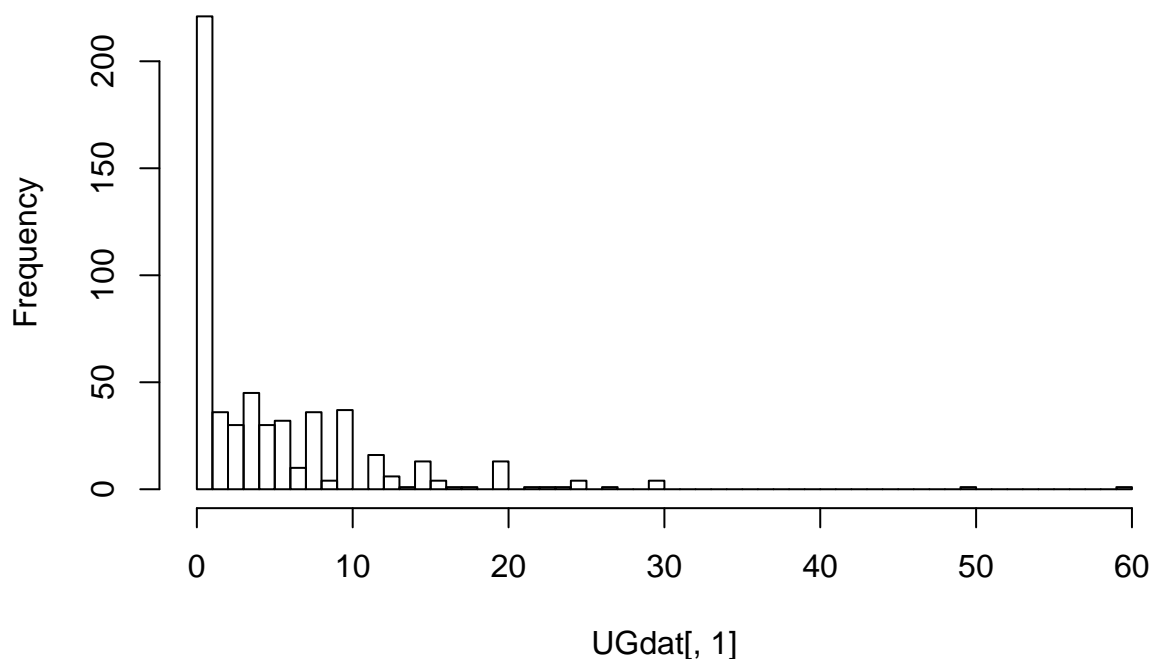
```
1-pchisq(600.60,43)

## [1] 0
```

and see the model does not fit well. We look at a histogram of the raw data and see an excessive amount of zeroes in the data, one reason for overdispersion.

```
UGdat<-as.data.frame(lapply(dat, function(x,p) rep(x,p), dat[["counts"]]))
hist(UGdat[,1], breaks = seq(0,60,by=1))
```

## Histogram of UGdat[, 1]



- (d) Consider a zero-inflated Poisson model with the zero-inflated component constant across subject (that is with intercept only for the model of  $\phi_i$ ). What are the mixing proportions for the degenerate distribution and the Poisson model? Interpret the regression coefficient of gender.

```
suppressWarnings(suppressMessages(library(psc1)))
```

```
fit.zip = zeroinfl(response ~ gender | 1 ,data=UGdat)
summary(fit.zip )
```

```
## Warning in deparse(x$call, width.cutoff = floor(getOption("width") * 0.85)): invalid 'cutoff' value
```

```
##
```

```
## Call:
```

```
## zeroinfl(formula = response ~ gender | 1, data = UGdat)
```

```
##
```

```
## Pearson residuals:
```

```
##      Min      1Q  Median      3Q      Max
## -1.1692 -1.1547 -0.4264  0.6238 12.2789
```

```
##
```

```
## Count model coefficients (poisson with log link):
```

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.99107    0.02747  72.493  <2e-16 ***
## gender1      0.09242    0.03830   2.413  0.0158 *
```

```
##
```

```
## Zero-inflation model coefficients (binomial with logit link):
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.61660    0.08944  -6.894 5.41e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 8
## Log-likelihood: -1835 on 3 Df
#mixing proportions
phi=as.numeric( exp(coef(fit.zip)[3])/(1+exp(coef(fit.zip)[3])) )
phi;1-phi

## [1] 0.3505542
## [1] 0.6494458
```

The mixing parameter  $\phi = 0.3505542$ , so the mixing proportion for the degenerate distribution is 0.6494458 and the mixing proportion for the Poisson distribution at  $y_i = 0$  is 0.3505542.

Males have an expected log count that is 0.09242 higher than females.

- (e) Consider a zero-inflated negative binomial model. What are the mixing proportions for the degenerate distribution and the negative binomial model? Interpret the regression coefficient of gender.

```
fit.zinb = zeroinfl(response ~ gender | 1, dist="negbin", data=UGdat)
summary(fit.zinb)
```

```
## Warning in deparse(x$call, width.cutoff = floor(getOption("width") * 0.85)): invalid 'cutoff' value
##
## Call:
## zeroinfl(formula = response ~ gender | 1, data = UGdat, dist = "negbin")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -0.8054 -0.7979 -0.2814  0.3961  8.2062
##
## Count model coefficients (negbin with log link):
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.89133    0.06990  27.059 < 2e-16 ***
## gender1      0.14584    0.09487   1.537 0.124254
## Log(theta)   0.43572    0.12576   3.465 0.000531 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.8439    0.1166  -7.238 4.54e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 1.5461
## Number of iterations in BFGS optimization: 9
```

```
## Log-likelihood: -1410 on 4 Df
#mixing proportions
phi=as.numeric( exp(coef(fit.zinb)[3])/(1+exp(coef(fit.zinb)[3])) )
phi,1-phi

## [1] 0.300723
## [1] 0.699277
```

The mixing parameter  $\phi = 0.300723$ , so the mixing proportion for the degenerate distribution is 0.6494458 and the mixing proportion for the Poisson distribution at  $y_i = 0$  is 0.699277.

Males have an expected count that is 1.157011 ( $e^{0.14584}$ ) than females. The predictor **gender1** in the part of the negative binomial regression model predicting how many times they had sexual intercourse in the preceding month is not statistically significant.

- (f) Provide a table with the observed counts and the fitted counts for each of the four models for  $y_i = 0, \dots, 20$  and  $y_i > 20$ .

```
tab<-cbind(dat[which(dat$gender == 0),],dat[which(dat$gender == 1 ),-1])
tab<-tab[,c(1,2,5)]
names(tab)[2:3]<-c("Female","Male");head(tab)
```

```
##      response Female Male
## 28         0     128    1
## 29         1      17    1
## 30         2      23    1
## 31         3      16    1
## 32         4      19    1
## 33         5      17    1
```

```
sex.n = by(dat$counts, dat$gender, sum)
sex.nb = glm.nb(response ~ gender, weights=counts, data=dat)
sex.muhat = unique(sex.nb$fitted.values)

nfemale.poi = dpois(unique(dat[,1]), lambda=sex.muhat[1])*sex.n[1]
nmale.poi = dpois(unique(dat[,1]), lambda=sex.muhat[2])*sex.n[2]
nfemale.negb = round(dnbinom(unique(dat[,1]), size=sex.nb$theta,
mu=sex.muhat[1])*sex.n[1],3)
nmale.negb = round(dnbinom(unique(dat[,1]), size=sex.nb$theta,
mu=sex.muhat[2])*sex.n[2],3)

tab$poiF<-round( nfemale.poi ,3)
tab$poiM<-round( nmale.poi ,3)

tab$negbF<- nfemale.negb
tab$negbM<- nmale.negb
```

```

tab$zipF<- round(c( as.numeric(
  sex.n[1] * ( phi + (1-phi) * dpois(0,exp(1.99107)) ) ),
  sex.n[1]*( (1-phi)* dpois(unique(dat[,1])[-1],exp(1.99107)) ) ),3)

tab$zipM<-round(c( as.numeric(
  sex.n[2] * ( phi + (1-phi) * dpois(0,exp(1.99107 + 0.09242)) )),
  sex.n[2]*((1-phi)*dpois(unique(dat[,1])[-1],exp(1.99107 + 0.09242))))),3)

tab$zinbF<-round(c( as.numeric(
  sex.n[1]*( phi + (1-phi) * dnbinom(0,mu=exp(1.89133), exp(0.43572)))),
  sex.n[1]* ((1-phi) * dnbinom(unique(dat[,1])[-1], mu=exp(1.89133),
  exp(0.43572)) ) ),3)

tab$zinbM<-round(c( as.numeric(
  sex.n[2]*( phi + (1-phi) * dnbinom(0,mu=exp(1.89133+0.14584),
  exp(0.43572)) ) ),
  sex.n[2]* ((1-phi) * dnbinom(unique(dat[,1])[-1], mu=exp(1.89133 +
  0.14584),exp(0.43572)) ) ),3)

print(tab,row.names = F)

```

##	response	Female	Male	poiF	poiM	negbF	negbM	zipF	zipM	zinbF	zinbM
##	0	128	1	0.885	3.246	86.667	77.236	93.367	72.228	109.739	82.797
##	1	17	1	5.187	13.968	40.064	34.714	1.048	0.438	20.703	13.668
##	2	23	1	15.193	30.054	27.712	23.345	3.836	1.758	21.371	14.481
##	3	16	1	29.668	43.110	21.287	17.436	9.365	4.708	20.483	14.245
##	4	19	1	43.452	46.378	17.165	13.670	17.145	9.453	18.877	13.473
##	5	17	1	50.911	39.915	14.236	11.023	25.112	15.187	16.978	12.437
##	6	17	1	49.709	28.627	12.024	9.052	30.651	20.331	15.020	11.293
##	7	3	1	41.601	17.598	10.287	7.530	32.067	23.330	13.129	10.131
##	8	15	1	30.464	9.466	8.885	6.323	29.355	23.425	11.372	9.007
##	9	2	1	19.830	4.526	7.731	5.349	23.886	20.906	9.781	7.951
##	10	13	1	11.617	1.948	6.766	4.552	17.493	16.793	8.364	6.978
##	12	10	1	3.020	0.273	5.253	3.341	7.107	8.208	6.035	5.304
##	13	3	1	1.361	0.090	4.653	2.877	4.004	5.072	5.099	4.599
##	14	1	1	0.570	0.028	4.134	2.485	2.094	2.910	4.296	3.977
##	15	10	1	0.222	0.008	3.681	2.152	1.023	1.558	3.610	3.430
##	16	1	1	0.081	0.002	3.285	1.867	0.468	0.782	3.027	2.952
##	17	1	1	0.028	0.001	2.937	1.623	0.202	0.370	2.534	2.536
##	18	1	1	0.009	0.000	2.630	1.413	0.082	0.165	2.117	2.174
##	20	6	1	0.001	0.000	2.119	1.076	0.012	0.028	1.471	1.591
##	22	1	1	0.000	0.000	1.715	0.824	0.001	0.004	1.017	1.159
##	23	1	1	0.000	0.000	1.546	0.722	0.000	0.001	0.844	0.987
##	24	0	1	0.000	0.000	1.394	0.633	0.000	0.000	0.700	0.840
##	25	3	1	0.000	0.000	1.259	0.555	0.000	0.000	0.580	0.715
##	27	1	1	0.000	0.000	1.028	0.429	0.000	0.000	0.397	0.516
##	30	1	1	0.000	0.000	0.763	0.292	0.000	0.000	0.224	0.314
##	50	0	1	0.000	0.000	0.114	0.025	0.000	0.000	0.004	0.010
##	60	0	1	0.000	0.000	0.046	0.008	0.000	0.000	0.001	0.002