1. **Exercise 1 – Agresti 7.36**

Table 1 is based on a study involving British doctors.

	Person-Years		Coronary Deaths	
Age	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.

(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.

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 ≈ 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.

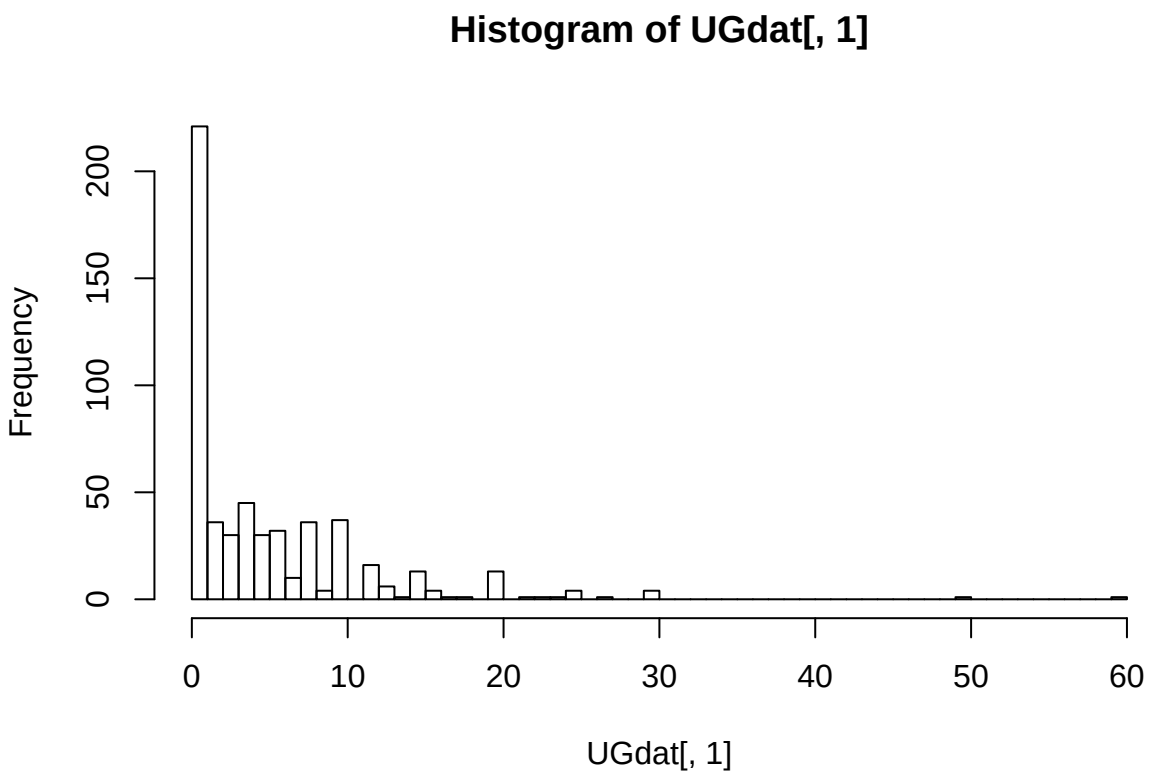
(c) Fit a negative binomial model. Is there evidence of overdispersion? What is the estimated difference in log means, its standard error, and the 95% Wald confidence interval for the ratio of means.

```
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
```

```r
UGdat<-as.data.frame(lapply(dat, function(x,p) rep(x,p), dat[["counts"]]))
hist(UGdat[,1], breaks = seq(0,60,by=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 ϕ_i). What are the mixing proportions for the degenerate distribution and the Poisson model? Interpret the regression coefficient of gender.

```r
suppressWarnings(suppressMessages(library(pscl)))

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

```
##
## 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
```

```r
#mixing proportions
as.numeric( exp(coef(fit.zip)[3])/(1+exp(coef(fit.zip)[3])) )
```

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

```r
1-as.numeric( exp(coef(fit.zip)[3])/(1+exp(coef(fit.zip)[3])) )
```

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

(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.

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

```
##
## 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
as.numeric( exp(coef(fit.zinb)[3])/(1+exp(coef(fit.zinb)[3])) )
```

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

```
1-as.numeric( exp(coef(fit.zinb)[3])/(1+exp(coef(fit.zinb)[3])) )
```

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

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