Math-661: Assignment 5 – Sample Solution

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.**

```
smoke = c(rep("nonsmoker", 5), rep("smoker", 5))
age = rep(c("35-44", "45-54", "55-64", "65-74", "75-84"),2)
exposure = c(18793, 10673, 5710, 2585, 1462, 52407, 43248, 28612, 12663, 5317)
death = c(2, 12, 28, 28, 31, 32, 104, 206, 186, 102)
coronary = data.frame(smoke, age, exposure, death)

coronary.fit = glm(death ~ as.factor(age)+as.factor(smoke),
           offset = log(exposure), family=poisson, data=coronary)

> summary(coronary.fit)
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            -7.9194     0.1918 -41.298  < 2e-16 ***
as.factor(age)45-54     1.4840     0.1951   7.606 2.82e-14 ***
as.factor(age)55-64     2.6275     0.1837  14.301  < 2e-16 ***
as.factor(age)65-74     3.3505     0.1848  18.131  < 2e-16 ***
as.factor(age)75-84     3.7001     0.1922  19.250  < 2e-16 ***
as.factor(smoke)smoker  0.3545     0.1074   3.302  0.00096 ***
---
(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
```

```
> 1-pchisq(12.134, 4)
[1] 0.01638213
```

The goodness-of-fit test suggests that the model does not adequately fit the data (p-value = 0.0164). This may be due to the fact that this model assumes the effect of smoking to be the same across all age groups. Based on this model, given age, the estimated death rate for smokers is $e^{0.35} = 1.425$ times that of nonsmokers, i.e., the ratio of smokers' to nonsmokers' coronary death rates is 1.425 for each age group.

If we calculate the sample coronary death rates per 1000 person-years for smokers and non-smokers across age groups, we get:

	Death rate (per 1000 person-years)		
Age	Nonsmokers	Smokers	Ratio of death rates
35 – 44	0.106	0.611	5.76
45 – 54	1.124	2.405	2.14
55 – 64	4.903	7.200	1.47
65 – 74	10.832	14.688	1.36
75 – 84	21.204	19.184	0.905

We note that the sample death rates increase with age among both smokers and nonsmokers. The death rates are consistently higher among smokers in the first four age groups, but are slightly lower in the oldest age group. We also note that the ratio of death rates for smokers versus nonsmokers is highest among the youngest age group and decreases with age. Thus, the assumption of constant ratio of death rates among smokers and non-smokers is not valid.

Figure 1 shows the observed and fitted ratios of coronary death rates for smokers versus non-smokers. We note that the observed ratio varies across age group, while the fitted ratio is constant across age groups.

```
rate.nonsm = (death/exposure)[1:5]
rate.sm = (death/exposure)[6:10]
ratio = rate.sm/rate.nonsm

fitted.nonsm = (fitted(coronary.fit)/exposure)[1:5]
fitted.sm = (fitted(coronary.fit)/exposure)[6:10]
fitted.ratio = fitted.sm/fitted.nonsm

plot(ratio, xlab="age groups", ylab="ratio", type="l")
lines(fitted.ratio, lty=2)
legend("topright", c("observed", "fitted"), lty=1:2, bty="n")
```

(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.**

As we noted in (a), the ratio of death rates between smokers and non-smokers is not constant. It is therefore sensible to add an interaction term for age and smoking in the model.

We use scores of 0 to 4 for the age groups and fit a Poisson loglinear model with interaction of age and smoking:

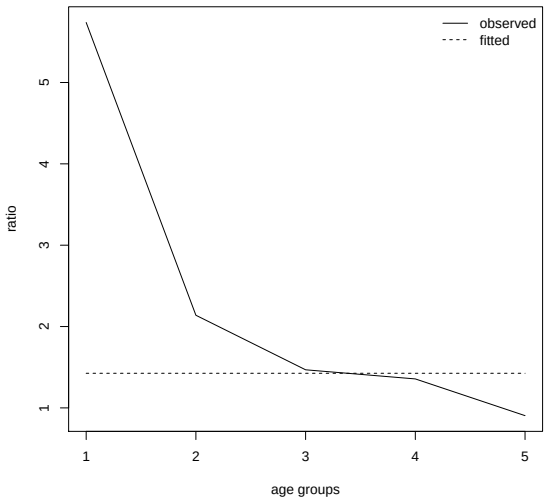

Figure 1: Plot of observed and fitted ratios of coronary death rates for smokers versus nonsmokers

```
age.score = rep(0:4,2)
inter.fit = glm(death ~ age.score*as.factor(smoke),
        offset = log(exposure), family=poisson, data=coronary)

fitinter.nonsm = (fitted(inter.fit)/exposure)[1:5]
fitinter.sm = (fitted(inter.fit)/exposure)[6:10]
fitinter.ratio = fitinter.sm/fitinter.nonsm

plot(0:4, log(ratio), xlab="age scores", ylab="log ratio", type="l")
lines(0:4, log(fitinter.ratio), lty=2)
legend("topright", c("observed", "fitted"), lty=1:2, bty="n")
```

Figure 1 shows the observed and fitted log ratio of coronary death rates based on the model with quantitative interaction of age and smoking. We see that the fitted log ratio is linear. However, the observed log ratio is not quite linear.

```
> summary(inter.fit)
Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -7.82032    0.23382 -33.446  < 2e-16 ***
age.score                          1.04685    0.07743  13.520  < 2e-16 ***
as.factor(smoke)smoker             1.03469    0.24849   4.164 3.13e-05 ***
age.score:as.factor(smoke)smoker -0.24899    0.08359  -2.979  0.00289 **
---
(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
```

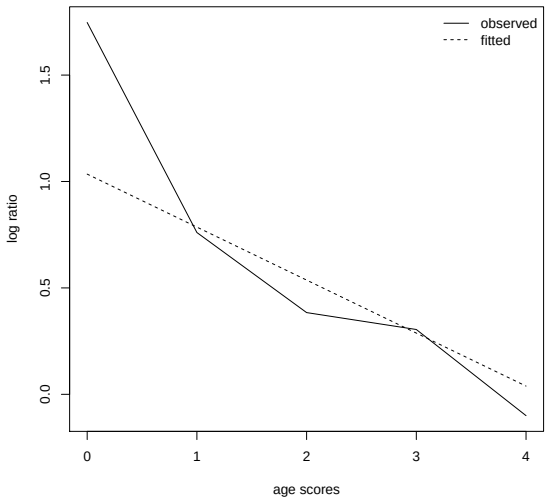

Figure 2: Plot of observed and fitted log ratio of coronary death rates based on model with quantitative interaction

All the regression coefficients are statistically significant:

- Among non-smokers, the estimated equation for the Poisson loglinear model is:

$$\log\left(\hat{\mu}_i/t_i\right) = -7.82 + 1.047 \text{ age}$$

  - the death rate among the youngest non-smoking subjects is estimated to be $\exp(-7.82) = 4 \times 10^{-4}$, i.e., 0.4 deaths per 1000 person years.
  - the death rate is estimated to increase by a factor of $\exp(1.047) = 2.85$ for every 1 unit increase in age score (i.e., falling in the next age group).
- Among smokers, the estimated equation for the Poisson loglinear model is:

$$\log\left(\hat{\mu}_i/t_i\right) = -7.82 + 1.035 + (1.047 - 0.249) \text{ age}$$

  - the death rate among the youngest subjects is estimated to be $\exp(-7.82 + 1.035) = 0.0011$, i.e., 1.1 deaths per 1000 person-years.
  - the death rate is estimated to increase by a factor of $\exp(1.047 - 0.249) = 2.22$ for every 1 unit increase in age score (i.e.,falling in the next age group).

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.**

```
response = c(0:10, 12:18, 20, 22:25, 27, 30, 50, 60)
male = c(65, 11, 13, 14, 26, 13, 15, 7, 21, 2, 24, 6, 3, 0, 3, 3, 0, 0,
          7, 0, 0, 1, 1, 0, 3, 1, 1)
female = c(128, 17, 23, 16, 19, 17, 17, 3, 15, 2, 13, 10, 3, 1, 10, 1, 1,
          1, 6, 1, 1, 0, 3, 1, 1, 0, 0)
gss = data.frame(response=rep(response, 2), gender=c(rep("male", 27), rep("female", 27)),
           counts=c(male, female))

fit.pois = glm(response ~ gender, family=poisson, weight=counts, data=gss)
> summary(fit.pois)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.45936    0.02738  53.302  < 2e-16 ***
gendermale   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
```

The sample mean and variance for the 310 females is

$$\frac{\sum_i n_i y_i}{\sum_i n_i} = \frac{1334}{310} = 4.303 \qquad \frac{\sum_i n_i (y_i - \bar{y})^2}{\sum_i n_i - 1} = \frac{10601.5}{309} = 34.31$$

The sample mean and variance for the 240 males is

$$\frac{\sum_i n_i y_i}{\sum_i n_i} = \frac{1406}{240} = 5.858 \qquad \frac{\sum_i n_i (y_i - \bar{y})^2}{\sum_i n_i - 1} = \frac{13097.18}{239} = 54.80$$

Thus, among both males and females, the sample variances are substantially larger than the sample means.

Based on the Poisson model, the estimated expected count is $\exp(1.459) = 4.303$ for female and $\exp(1.459 + 0.3085) = 5.856$ for male. The goodness-of-fit test indicates a lack of fit to the data

```
> 1-pchisq(fit.pois$deviance, fit.pois$df.residual)
[1] 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.**

The Poisson GLM indicates that the estimated expected count for males is $\exp(0.3085) = 1.36$ times that for females.

A 95% Wald confidence interval for the ratio of means for males versus females is

$$\exp(0.3085 \pm 1.96 \times 0.03822) = (1.263, 1.467)$$

(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.**

```
fit.nb = glm.nb(response ~ gender, weight=counts, data=gss)

> summary(fit.nb)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.45936    0.08472  17.226   <2e-16 ***
gendermale   0.30850    0.12724   2.425   0.0153 *
---
(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
```

The dispersion parameter is estimated to be $\hat{\gamma} = 1/0.5019 = 1.99$ with a 95% CI of $(1.73, 2.35)$. Thus, there is evidence that $\gamma > 1$ and there is overdispersion.

The estimated difference in log means between males and females is 0.3085, which the same as with the Poisson model. However the standard error has increased from 0.03822 to 0.12724, and the 95% Wald CI for the ratio of means is wider

$$\exp(0.3085 \pm 1.96 \times 0.12724) = (1.061, 1.747)$$

The AIC indicates that the negative binomial model provides a better fit to the data than the Poisson GLM, but the goodness-of-fit test provides evidence that this model does not provide adequate fit to the data

```
> 1-pchisq(fit.nb$deviance, fit.nb$df.residual)
[1] 0
```

(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.**

The function `zeroinfl()` seems to require ungrouped data, so let's first reorganize the data in an ungrouped format

```
gss.ungrp=as.data.frame(lapply(gss, function(x,p) rep(x,p), gss$counts))
library(pscl)
fit.zip = zeroinfl(response ~ gender|1, data=gss.ungrp)

> summary(fit.zip)
Count model coefficients (poisson with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.99107    0.02747  72.493   <2e-16 ***
gendermale   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 ***
---
Log-likelihood: -1835 on 3 Df
```

The fitted ZIP distribution is a mixture with probability $e^{-0.6166}/(1 + e^{-0.6166}) = 0.35$ for the degenerate distribution and $1 - 0.35 = 0.65$ for the Poisson model, with the latter having expected count of $\exp(1.992) = 7.32$ for females and $\exp(1.992 + 0.092) = 8.03$ for males.

The regression coefficient of gender indicates that the log expected count for male is 0.09 higher than that of female.

(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=gss.ungrp)
```

```
> summary(fit.zipnb)
Count model coefficients (negbin with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.89133    0.06990  27.059  < 2e-16 ***
gendermale   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 ***
---
Theta = 1.5461
Number of iterations in BFGS optimization: 9
Log-likelihood: -1410 on 4 Df
```

The fitted ZINB distribution is a mixture with probability $e^{-0.8439}/(1+e^{-0.8439}) = 0.30$ for the degenerate distribution and $1 - 0.30 = 0.70$ for the negative binomial model, with the latter having expected count of $\exp(1.89) = 6.63$ for females and $\exp(1.89 + 0.146) = 7.67$ for males. The dispersion parameter is estimated to be $1/exp(0.436) = 0.65$.

The regression coefficient of gender indicates that the log expected count for male is 0.146 higher than that of female.

(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 $y_i > 20$.**

The estimated expected counts under the different models are summarized in Table 3:

```
n = by(gss$counts, gss$gender, sum)
muhat  = cumsum(exp(fit.pois$coefficients))
nfem.poi = c(dpois(0:20, lambda=muhat[1]), 1-ppois(20, lambda=muhat[1]))*n[1]
nmale.poi = c(dpois(0:20, lambda=muhat[2]), 1-ppois(20, lambda=muhat[2]))*n[2]
nfem.nb = c(dnbinom(0:20, size=fit.nb$theta, mu=muhat[1]),
      1-pnbinom(20, size=fit.nb$theta, mu=muhat[1]))*n[1]
nmale.nb = c(dnbinom(0:20, size=fit.nb$theta, mu=muhat[2]),
      1-pnbinom(20, size=fit.nb$theta, mu=muhat[2]))*n[2]

phi.zip =  exp(fit.zip$coefficients$zero)/(1+exp(fit.zip$coefficients$zero))
mu.zip = cumsum(exp(fit.zip$coefficients$count))
nfem.zip = c(phi.zip+(1-phi.zip)*dpois(0,lambda=mu.zip[1]),
      (1-phi.zip)*dpois(1:20, lambda=mu.zip[1]),
      (1-phi.zip)*(1-ppois(20, lambda=mu.zip[1])))*n[1]
nmale.zip = c(phi.zip+(1-phi.zip)*dpois(0,lambda=mu.zip[2]),
     (1-phi.zip)*dpois(1:20, lambda=mu.zip[2]),
      (1-phi.zip)*(1-ppois(20, lambda=mu.zip[2])))*n[2]

phi.zinb =  exp(fit.zinb$coefficients$zero)/(1+exp(fit.zinb$coefficients$zero))
mu.zinb = cumsum(exp(fit.zinb$coefficients$count))
nfem.zinb = c(phi.zinb+(1-phi.zinb)*dnbinom(0, size=fit.zinb$theta, mu=mu.zinb[1]),
```

	Observed		Poisson fit		Neg. Bin. fit		ZIP fit		ZINB fit	
Response	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
0	128	65	4.19	0.83	99.76	68.15	108.80	84.17	109.74	82.59
1	17	11	18.04	4.71	44.84	31.42	0.97	0.29	20.70	13.44
2	23	13	38.82	13.35	30.15	21.67	3.56	1.22	21.37	14.27
3	16	14	55.68	25.20	22.52	16.60	8.70	3.42	20.48	14.08
4	19	26	59.90	35.69	17.66	13.35	15.92	7.19	18.88	13.35
5	17	13	51.56	40.44	14.24	11.04	23.32	12.12	16.98	12.35
6	17	15	36.98	38.18	11.69	9.30	28.47	17.00	15.02	11.24
7	3	7	22.73	30.89	9.73	7.94	29.78	20.45	13.13	10.11
8	15	21	12.23	21.87	8.17	6.84	27.26	21.53	11.37	9.01
9	2	2	5.85	13.77	6.91	5.93	22.18	20.14	9.78	7.98
10	13	24	2.52	7.80	5.88	5.18	16.25	16.96	8.36	7.02
11	0	0	0.98	4.02	5.03	4.54	10.82	12.98	7.12	6.15
12	10	6	0.35	1.90	4.32	4.00	6.60	9.11	6.03	5.36
13	3	3	0.12	0.83	3.72	3.53	3.72	5.90	5.10	4.66
14	1	0	0.04	0.33	3.21	3.13	1.95	3.55	4.30	4.04
15	10	3	0.01	0.13	2.78	2.78	0.95	1.99	3.61	3.49
16	1	3	0.00	0.04	2.41	2.47	0.43	1.05	3.03	3.01
17	1	0	0.00	0.01	2.10	2.21	0.19	0.52	2.53	2.60
18	1	0	0.00	0.00	1.83	1.97	0.08	0.24	2.12	2.23
19	0	0	0.00	0.00	1.59	1.76	0.03	0.11	1.77	1.91
20	6	7	0.00	0.00	1.39	1.58	0.01	0.05	1.47	1.64
> 20	7	7	0.00	0.00	10.09	14.60	0.01	0.03	7.11	9.46

Table 3: Observed and fitted counts

```
      (1-phi.zinb)*dnbinom(1:20, size=fit.zinb$theta, mu=mu.zinb[1]),
      (1-phi.zinb)*(1-pnbinom(20, size=fit.zinb$theta, mu=mu.zinb[1])))*n[1]
nmale.zinb = c(phi.zinb+(1-phi.zinb)*dnbinom(0, size=fit.zinb$theta, mu=mu.zinb[2]),
      (1-phi.zinb)*dnbinom(1:20, size=fit.zinb$theta, mu=mu.zinb[2]),
      (1-phi.zinb)*(1-pnbinom(20, size=fit.zinb$theta, mu=mu.zinb[2])))*n[2]
```

We note that the fitted values for the ZINB model are closer to the observed counts. The AIC for the four models also suggest that this model provides a better fit to the data:

$$\begin{aligned} \text{Poisson GLM AIC} &= 5271.3 \\ \text{Neg. Bin. GLM AIC} &= 2883 \\ \text{ZIP GLM AIC} &= 3676 \\ \text{ZINB GLM AIC} &= 2828 \end{aligned}$$

If we use the Vuong non-nested hypothesis test, there is evidence that the ZINB model fits the data better than the negative binomial model:

```
> vuong(fit.nb2, fit.zipnb)
Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
```

```
---------------------------------------------------------------
              Vuong z-statistic             H_A    p-value
Raw                   -3.874915 model2 > model1 5.3331e-05
AIC-corrected         -3.738758 model2 > model1 9.2466e-05
BIC-corrected         -3.445344 model2 > model1 0.00028517
```