第五章 泊松分布的统计推断

The Poisson distribution

- Support a RV Y has a Poisson distribution with parameter λ. We say $Y \sim Po(\lambda)$
- The pmf of Y at a value y is given by

$$f_Y(y) = \frac{e^{-\lambda}\lambda^y}{y!}, \qquad y = 0, 1, \ldots$$

- The moment generating function of Y is

$$M_Y(t) = exp(\lambda(e^t - 1))$$

- Thus, the mean is

$$E(Y) = \lambda$$

  and the variance is

$$var(Y) = \lambda$$

- Useful S-PLUS/R commands for Poisson RVs are dpois,ppois,qpois,rpois.

Additivity of Poisson RVs

- Suppose $Y_1 \sim Po(\lambda_1)$ which is independent of $Y_2 \sim Po(\lambda_2)$
  Then $Y_1 + Y_2 \sim Po(\lambda_1 + \lambda_2)$
- More generally let $\{Y_i : i = 1, ..., n\}$ be a set of n independent RVs, with $Y_i \sim Po(\lambda_i)$ for each i.
  Then $\sum_{i=1}^{n} Y_i \sim Po(\sum_{i=1}^{n} \lambda_i)$
- This has implication for factors in poisson GLMs.

Inference for λ

- Again suppose the RV $Y \sim Po(\lambda)$
- Then, the ML estimate of λ is $\hat{\lambda} = y$, and via standard arguments, an approximate 95% CI for λ is

$$y \pm 1.96\sqrt{\lambda}$$

- We can estimate $\sqrt{\lambda}$ by $\sqrt{y}$
- Problem: CI can contain values that are less than zero.

Inference for $log\lambda$

- Let $\theta = log\lambda$. Then an approximate 95% CI for θ is

$$\hat{\theta} \pm 1.96\sqrt{\frac{1}{\lambda}}$$

where $\hat{\theta} = logy$. We approximate this CI by

$$logy \pm 1.96\sqrt{\frac{1}{y}}$$

- Based on the CI for θ,an approximate 95% CI for λ is

$$[e^{logy-1.96\sqrt{\frac{1}{y}}}, e^{logy+1.96\sqrt{\frac{1}{y}}}]$$

$$\text{ie.,}[ye^{\frac{-1.96}{\sqrt{y}}}, ye^{\frac{1.96}{\sqrt{y}}}]$$

GLMs for Poisson data

- Suppose $\{Y_1, ..., Y_n\}$ are a set of independent $Po(\lambda_i)$ RVs
- Most commonly used link is the canonical link,

$$\eta_i = g(\lambda_i) = log\lambda_i$$

- Regardless of the link function $g(\dot{})$chosen, we fit the model

$$\eta_i = x_i^T\beta, i = 1, ..., n$$

- We estimate the parameters β using IWIS(with the log link, we need to be careful about the choice of starting values)

Model fit criteria for Poisson GLMs

- The Deviance is

$$D(y, \hat{\mu}) = 2\sum_{i=1}^{n}[y_i log(y_i/\hat{\mu}_i) + (y_i - \hat{\mu}_i)].$$

- When we include an intercept in our model $\sum_i y_i = \sum_i \hat{\mu}_i$ and the deviance reduces to

$$D(y, \hat{\mu}) = 2\sum_{i=1}^{n} y_i log(y_i/\hat{\mu}_i).$$

- Expanding the deviance as a Taylor series in $(y_i - \hat{\mu}_i)/\hat{\mu}_i$, we find that

$$D(y, \hat{\mu}) \approx \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i},$$

which is the Pearson χ^2 statistic.

Example: Doctors and smoking

(Data taken from Dobson, 2000)

- In October 1951, R.Doll and A.B.Hill sent questionnaires to the population of all physicians listed in the British registry of doctors who resided in England and Wales(59,000 physicians)
- The questionnaire included questions such as:
  -Are you a smoker or non-smoker(never consistently smoked as much as one cigarette a day for as long as one year)?
  -What is your age?
- There was a 68% response rete.
  Doll.R and A.B.Hill.,The mortality of doctors in relation to their smoking habits.Brit Med J 1954;1:1451-1455
  Doll.R and A.B.Hill.Mortality in relation to smoking: 10 years' observation of British doctors. Brit Med J 1964;1:1399-1410,1460-1467.

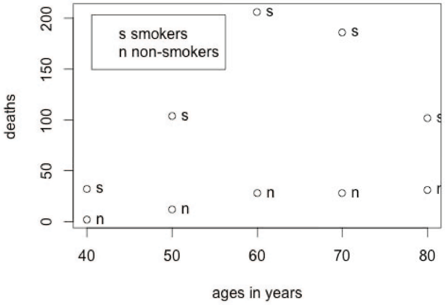
The follow-up

- In susequent years Doll and Hill kept track of the number of doctors who died of lung cancer.
  (From Registrar records and confirmed by medical records).
- They also calculated the person years at risk(a measure of time which indicates the sum of each individual's time at risk before death in this case).
- Scientific questions of interest:
  1.Is the death rate higher for smokers compared to non-smokers?
  2.Does the death rate depend on the age of the doctor?

The data

age	smoking	deaths	person years
35 to 44	smoker	32	52,407
45 to 54	smoker	104	43,248
55 to 64	smoker	206	28,612
65 to 74	smoker	186	12,663
75 to 84	smoker	102	5,317
35 to 44	non-smoker	2	18,790
45 to 54	non-smoker	12	10,673
55 to 64	non-smoker	28	5,710
65 to 74	non-smoker	28	2,585
75 to 84	non-smoker	31	1,462

A plot of the age versus the counts

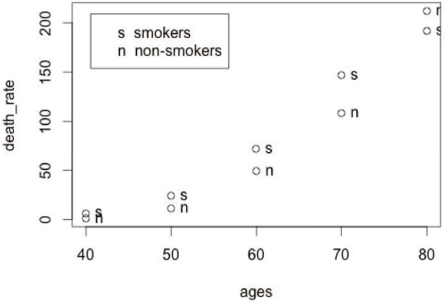


Calculating the death rate

- Suppose we calculate the number of deaths per 10,000 person years, that is,

$$\textit{death rate} = \frac{\textit{number\_of\_deaths}}{\textit{person\_years}/10{,}000}$$

- Is the death rate related to smoking status and age?



A linear model

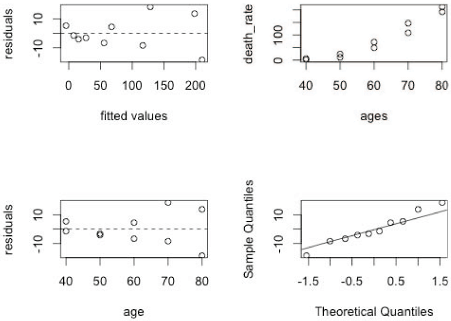
- We consider the following linear model for the death rate

```
lm.model <- lm(death.rate ~ ages + I(ages^2)
          + smoking,data = doctors)
```

- Here is the model summary

```
Residuals:
    Min      1Q  Median      3Q     Max
-18.350  -5.989  -2.269   5.158  18.384

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 123.91669   87.16751   1.422  0.20498
ages         -7.34186    3.01916  -2.432  0.05104 .
I(ages^2)     0.10340    0.02504   4.130  0.00615 **
smoking      11.83466    8.37939   1.412  0.20755
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.25 on 6 degrees of freedom
Multiple R-squared:  0.9811, Adjusted R-squared:  0.9717
F-statistic: 103.8 on 3 and 6 DF,  p-value: 1.465e-05
```

Diagnostic plots for the linear model



- Comments on this model:

Modeling strategies

- Considering a linear model with ’death rate’ as the response, ignores the facts that the number of deaths are **counts**
- Also,if the data are counts we may have a **mean-variance** relationship. (Note: a linear model with a transformed response,e.g.,log,does not fit well for this dataset).
- We will use a Poisson GLM.
- Problem:how do we incorporate the person years?

Fitting a Poisson GLM with an offset

- Consider log(person years at risk) as an **explanatory variable**
- We will use the default log link.
- Let $i = 1$ denote the non-smoking group, and $i = 2$denote the smoking group.
- Let $j = 1, ..., 5$ denote the different age groups.
- Our model is that$\{Y_{ij} : i = 1, 2; j = 1, ..., 5\}$ are an independent set of $Po(\lambda)$ RVs where

$$log\lambda_{ij} = \alpha + \delta logpy_{ij} + \beta a_j + gamma_i$$

- Here py_{ij} denotes the person years at risk for age group j in smoking group i, and a_j denotes the midpoint of the age for age group j. We assume that $\gamma_1 = 0$

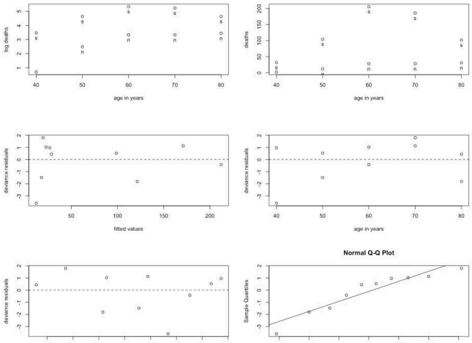
The Poisson GLM summary

```
Call:
glm(formula = deaths ~ log(person.years) + ages + smoking, family = poisson,
    data = doctors)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6005  -1.2152   0.4824   1.0132   1.8007

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -28.55467    2.86426  -9.969  < 2e-16 ***
log(person.years)   2.43524    0.22702  10.727  < 2e-16 ***
ages                0.17702    0.01542  11.478  < 2e-16 ***
smoking            -1.69914    0.35482  -4.789 1.68e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 644.269  on 9  degrees of freedom
Residual deviance:  25.576  on 6  degrees of freedom
AIC: 88.644
Number of Fisher Scoring iterations: 4
```

The analysis of deviance table

```
Analysis of Deviance Table

Model: poisson, link: log

Response: deaths

Terms added sequentially (first to last)


                 Df Deviance Resid. Df Resid. Dev
NULL                                 9     644.27
log(person.years)  1    66.31         8     577.96
ages             1   527.71          7      50.25
smoking          1    24.67          6      25.58
```

Model diagnostics



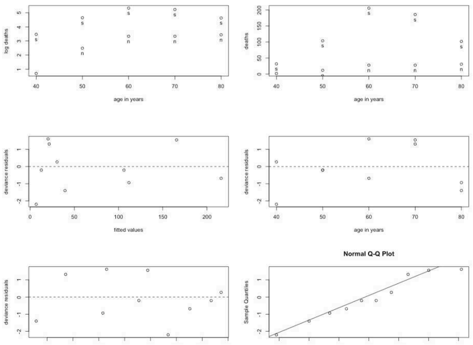
Incorporating an interaction term

```
Call:
glm(formula = deaths ~ log(person.years) + ages * smoking, family = poisson,
    data = doctors)

Deviance Residuals:
      1        2        3        4        5        6        7        8        9
 0.2701  -0.2083  -0.6852   1.5530  -0.9369  -2.2005  -0.2061   1.6127   1.3148
     10
-1.3995
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -29.905630   2.836330 -10.544  < 2e-16 ***
log(person.years)   2.427880   0.222166  10.928  < 2e-16 ***
ages                0.198594   0.016559  11.993  < 2e-16 ***
smoking             0.085936   0.667389   0.129  0.89754
ages:smoking       -0.027244   0.008575  -3.177  0.00149 **
---
Signif. codes:  0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 644.269  on 9  degrees of freedom
Residual deviance:  15.048  on 5  degrees of freedom
AIC: 80.116
Number of Fisher Scoring iterations: 4
```

The analysis of deviance table

```
Analysis of Deviance Table

Model: poisson, link: log

Response: deaths

Terms added sequentially (first to last)


                  Df Deviance Resid. Df Resid. Dev
NULL                                  9     644.27
log(person.years)  1    66.31         8     577.96
ages               1   527.71         7      50.25
smoking            1    24.67         6      25.58
ages:smoking       1    10.53         5      15.05
```

Diagnostic plots



Making scientific conclusions based on the Poisson GLM

age	smoking	obs.deaths	person years	est.deaths
35 to 44	smoker	32	52,407	30.50
45 to 54	smoker	104	43,248	106.14
55 to 64	smoker	206	28,612	215.99
65 to 74	smoker	186	12,663	165.62
75 to 84	smoker	102	5,317	111.76
35 to 44	non-smoker	2	18,790	6.90
45 to 54	non-smoker	12	10,673	12.73
55 to 64	non-smoker	28	5,710	20.31
65 to 74	non-smoker	28	2,585	21.61
75 to 84	non-smoker	31	1,462	39.46

a better offset model

- Remember that our previous model was that $\{Y_{ij} : i = 1, 2; j = 1, \ldots, 5\}$ are an independent set of $Po(\lambda_{ij})$ RVs where

$$\log \lambda_{ij} = \alpha + \delta \log py_{ij} + \beta a_j + \gamma_i$$

- Suppose we wish to fix $\delta = 1$ in this model.
- The R code for this model is

```
glm.model3<-glm(deaths~offset(log(person.years))+
              ages+smoking,
            data=doctors,family=poisson)
summary(glmmodel3)
anova(glm.model3)
```

- The summary does not include the offset in the coefficient table-in your interpretation, you have to remember it is there!

Fitting the model in R

- Another way to fit the offset model in R is:

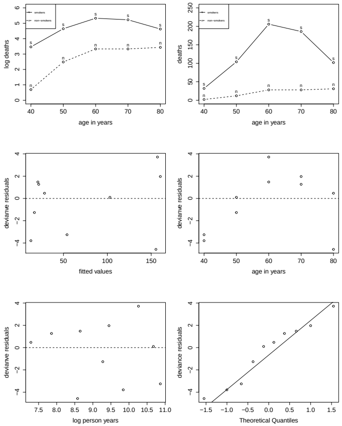
```
glm.model3<-glm(deaths~ages+smoking,
                offset=log(person.years),
                data=doctors, family=poisson)
```

The model summary

```
glm(formula=deaths~offset(log(person.years))+ages+smoking,
    family=poisson,data=doctors)

Deviance Residuals:
    Min        1Q    Median       3Q       Max
-4.5712  -2.7562    0.2857   1.4261    3.7183

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.625830   0.209721 -50.667  < 2e-16 ***
ages          0.083583   0.002904  28.777  < 2e-16 ***
smoking       0.406370   0.107195   3.791  0.00015 ***
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 935.067  on 9  degrees of freedom
Residual deviance:  69.182  on 7  degrees of freedom
AIC: 130.25

Number of Fisher Scoring iterations: 4
```

Model diagnostics

	Df	Deviance	Resid. Df	Resid. Dev
NULL			9	935.07
ages	1	850.06	8	85.01
smoking	1	15.83	7	69.18



Comment about this model

Models for contingency tables

- Poisson GLMs are often used as a probability model for contingency tables.

- We already showed that if $Y_i \sim B(m_i, p_i)$ where m_i is large and p_i is small then Y_i is approximately Poisson with mean $\mu_i = m_i p_i$. Another way of writing this is

$$\log \mu_i = \log m_i + \log p_i.$$

- Thus if we have a set of explanatory variables x_i, binomial regression of Y_i using x_i will give similar results to poisson regression of Y_i using x_i with a **log link** and an **offset** of $\log m_i$.
- We now generalize this result(after we consider an example).

Contingency table example

- The table on the next page displays the educational attainment of Americans by age categories in 1984.
- The counts are presented in thousands.
- The data was collected by the U.S.Bureau of the Census.
- Americans under age 25 are not included because many have not completed their education.
  (Refs:Moore and McCabe(1989), Introduction to the Practice of Statistics. Original source: World Almanac and Book of Facts,1986)

Views of the data

- Raw couts:

Age Group	Not complete High Sch.	Complete High Sch.	College 1-3yrs	College 3+yrs
25-34	5416	16431	8555	9771
35-44	5030	1855	5576	7596
45-44	5777	9435	3124	3904
55-64	7606	8795	2524	3109
64+	13746	7558	2503	2483

- As percentages:

Age Group	Not complete High Sch.	Complete High Sch.	College 1-3yrs	College 3+yrs
25-34	4.1	12.6	6.5	7.5
35-44	3.8	1.4	4.3	5.8
45-44	4.4	7.2	2.4	3.0
55-64	5.8	6.7	1.9	2.4
64+	10.5	5.8	1.9	1.9

Notation

- In this example we have $J = 5$ rows and $K = 4$ columns.
- There are $n = 130794$ people classified in this survey.
- Let Y_{jk} denote the frequency for the (j, k) cell of the table.
- We have that the **constraint** that

$$\sum_{j=1}^{J} \sum_{k=1}^{K} Y_{jk} = n.$$

A poisson model

- Suppose that $\{Y_{jk} : j = 1, \ldots, J; k = 1, \ldots, K\}$ are independent Poisson random variables with mean μ_{jk}.
- Taking expectations we have that

$$E(n) = \sum_{j=1}^{J} \sum_{k=1}^{K} \mu_{jk} = \mu, \quad say.$$

- Using the additivity result for Poisson RVs, the sum of all the entries of the table is a Poisson RV with mean μ.

Interpreting as a multinomial model

- It is possible to show that the joint distribution of $\{Y_{jk}\}$ conditional on n is **multinomial** with the conditional pmf

$$f(\{y_{jk}\}|n) = n! \prod_{j=1}^{J} \prod_{k=1}^{K} \frac{p_{jk}^{y_{jk}}}{y_{jk}!},$$

where we define

$$p_{jk} = \frac{\mu_{jk}}{\mu},$$

for each j and k.

- Note that $0 < p_{jk} < 1$ for each j, k and $\sum_j \sum_k p_{jk} = 1$.
- Thus we can interpret p_{jk} as the **probability** of an observation being in the (j, k) cell of the table.

Taking logs to obtain the log linear model

- Conditional on knowing n we have

$$E(Y_{jk}) = \mu_{jk} = np_{jk}.$$

and taking logs

$$\log \mu_{jk} = \log n + \log p_{jk}.$$

- Again, we can model the relationship between the Y_{jk}'s and the explanatory variables x_i (the variables in the rows and columns) using a Poisson GLM with offset $\log n$ and the default log link

-In our example we model the relationship between the counts, age groups and educational attainment.

Modeling the counts in terms of the age group

- In the following code **Age.Group** is a **factor** variable.

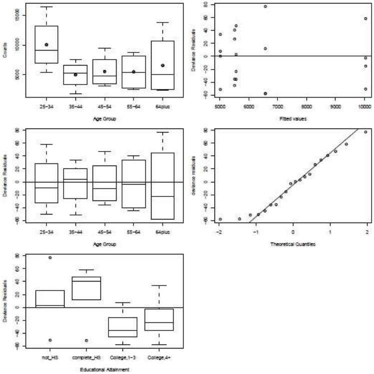
```
## calculate 'n'
n<-sum(ed$Count)

##create the offset vector
the.offset<-log(rep(n,length(ed$Count)))

## Model the counts in terms of the age groups
ed.model<-glm(Count~offset(the.offset)+Age.Group,
              data=ed,family=poisson)

## Summarize the model
summary(ed.model)
anova(ed.model)
```

The model summary

```
glm(formula = Count ~ offset(the.offset) + Age.Group,
family = poisson, data = ed)

Deviance Residuals:
Min 1Q Median 3Q Max
-57.836 -37.991 -1.253 28.492 77.059
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.566723 0.004989 -514.45 <2e-16 ***
Age.Group35-44 -0.694617 0.008646 -80.34 <2e-16 ***
Age.Group45-54 -0.591303 0.008358 -70.75 <2e-16 ***
Age.Group55-64 -0.600608 0.008383 -71.64 <2e-16 ***
Age.Group64plus -0.424006 0.007933 -53.45 <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)
Null deviance: 41770 on 19 degrees of freedom
Residual deviance: 32473 on 15 degrees of freedom
AIC: 32693
Number of Fisher Scoring iterations: 5
```

The model fit

```
          Df Deviance Resid. Df Resid. Dev
NULL                         19      41770
Age.Group  4     9297        15      32473
```



Comments on the age group model

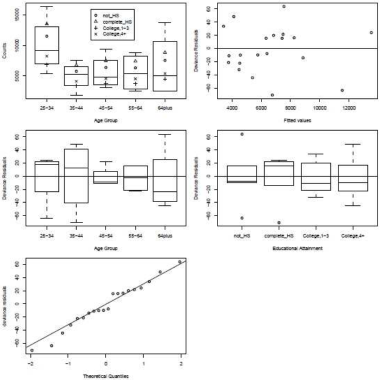
-

Adding educational attainment

```
glm(formula = Count ~ offset(the.offset) + Age.Group + Education,
family = poisson, data = ed)

Deviance Residuals:
    Min       1Q   Median      3Q     Max
-70.785 -21.595  -8.823 20.311 63.850

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -2.427713   0.006623 -366.57   <2e-16 ***
Age.Group35-44       -0.694617   0.008646  -80.34   <2e-16 ***
Age.Group45-54       -0.591303   0.008358  -70.75   <2e-16 ***
Age.Group55-64       -0.600608   0.008383  -71.64   <2e-16 ***
Age.Group64plus      -0.424006   0.007933  -53.45   <2e-16 ***
Educationcomplete_HS  0.159531   0.007022   22.72   <2e-16 ***
EducationCollege,1-3 -0.522560   0.008455  -61.80   <2e-16 ***
EducationCollege,4+  -0.335589   0.007990  -42.00   <2e-16 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for poisson family taken to be 1)
Null deviance: 41770 on 19 degrees of freedom
Residual deviance: 23365 on 12 degrees of freedom
```

The model fit

```
AIC: 23590
Number of Fisher Scoring iterations: 5
          Df Deviance Resid. Df Resid. Dev
NULL                         19      41770
Age.Group  4     9297        15      32473
Education  3     9108        12      23365
```



Comments on this model

-

The saturated interaction model

```
glm(formula = Count ~ offset(the.offset) + Age.Group * Education,
family = poisson, data = ed)
Deviance Residuals:
[1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Coefficients:
                                     Estimate Std. Error  z value  Pr(>|z|)
(Intercept)                          -3.18427    0.01359 -234.341   < 2e-16 ***
Age.Group35-44                       -0.07394    0.01958   -3.776  0.000159 ***
Age.Group45-54                        0.06453    0.01891    3.412  0.000646 ***
Age.Group55-64                        0.33958    0.01778   19.099   < 2e-16 ***
Age.Group64plus                       0.93139    0.01604   58.055   < 2e-16 ***
Educationcomplete_HS                  1.10981    0.01567   70.831   < 2e-16 ***
EducationCollege,1-3                  0.45716    0.01736   26.327   < 2e-16 ***
EducationCollege,4+                   0.59006    0.01694   34.831   < 2e-16 ***
Age.Group35-44:Educationcomplete_HS  -2.10735    0.03136  -67.201   < 2e-16 ***
Age.Group45-54:Educationcomplete_HS  -0.61927    0.02290  -27.038   < 2e-16 ***
Age.Group55-64:Educationcomplete_HS  -0.96457    0.02215  -43.545   < 2e-16 ***
Age.Group64plus:Educationcomplete_HS -1.70795    0.02123  -80.464   < 2e-16 ***
Age.Group35-44:EducationCollege,1-3  -0.35411    0.02607  -13.583   < 2e-16 ***
Age.Group45-54:EducationCollege,1-3  -1.07193    0.02819  -38.024   < 2e-16 ***
Age.Group55-64:EducationCollege,1-3  -1.56025    0.02880  -54.183   < 2e-16 ***
Age.Group64plus:EducationCollege,1-3 -2.16042    0.02782  -77.665   < 2e-16 ***
Age.Group35-44:EducationCollege,4+   -0.17786    0.02485   -7.158   8.2e-13 ***
Age.Group45-54:EducationCollege,4+   -0.98194    0.02676  -36.691   < 2e-16 ***
```

The analysis of deviance table and some conclusions

```
Age.Group55-64:EducationCollege,4+   -1.48470    0.02720   -54.575   < 2e-16 ***
Age.Group64plus:EducationCollege,4+  -2.30134    0.02761   -83.343   < 2e-16 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for poisson family taken to be 1)
Null deviance: 4.1770e+04 on 19 degrees of freedom
Residual deviance: 5.1823e-12 on 0 degrees of freedom
AIC: 249.04
Number of Fisher Scoring iterations: 2

                    Df Deviance Resid. Df Resid. Dev
NULL                                   19      41770
Age.Group            4     9297        15      32473
Education            3     9108        12      23365
Age.Group:Education 12    23365         0  5.182e-12
```

Type of zeros

- Suppose that $\{Y_i = 1, \cdots, I\}$ are independent Poisson random variables with mean μ_i.Define $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_I)^T$

- Let our observed data be $\{y_i i = 1, \cdots, I\}$,or equivalently$\boldsymbol{y} = (y_1, \cdots, y_I)^T$
- We can have **two** types of zeros
  - **Structural zeros** occur when $\mu_i = 0$.In that case y_i must be zero. e.g.,in a survey of cancers broken down by gender some cancers are gender specific(e.g.,postate,ovarian)and lead to structural zeros.
  - **Sampling zeros** occur when $\mu_i > 0$,but when $y_i = 0$.

Existence of MLEs

(Haberman,1973;adapted from Ageresti,1996)

(1) The log-likelyhood function is a strictly concave function of log μ.

(2) If an MLE of μ exists then it is unique and satisfies the likelihood equations $\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^T\mu$.Conversely,if $\hat{\mu}$ satisfies the Poisson model and also the likelihood equations then it is the MLE of μ.

(3) If **all**$y_i > 0$ the the MLEs of the model parameters exist.

(4) Suppose that the MLEs exist for a loglinear model that equates certain observed and fitted counts in certain marginal tables.Then those marginal counts have uniformly positive counts.

Implications

- For saturated models:
  - By (2)and (3),when all $y_i > 0$,the MLE of $\boldsymbol{\mu}$ is $\boldsymbol{y}$.
  - By (4),the parameter estimates **do not exist** when any $y_i = 0$.
- For unsaturated models:
  - By (2)and (3),when all $y_i > 0$,the MLEs exist.
  - By (4),the parameter estimates **do not exist** when any count is zero in a sufficient set of marginal tables.
- We illustrate with examples.

MLE existence examples

- Consider the following table of counts associated withthree factor levels.

x	A	B	C
y	0	7	12

- We fit the saturated Poisson GLM model:

```
## Here is the data
y <- c(0,7,12)
x <- factor(c("A","B","C"))

## Fit the model
model <- glm(y~x, family=poisson)
summary(model)

## Check for convergence of the IWLS
model$converged
[1] TRUE
```

- The IWLS algorithm has converged!
- The IWLS algorithm has converged! But, by Haberman’s results the MLEs do not exist!

Examining the model output

```
Deviance Residuals:
[1]  0  0  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -22.30   42247.17  -0.001        1
xB             24.25   42247.17   0.001        1
xC             24.79   42247.17   0.001        1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1.6739e+01  on 2  degrees of freedom
Residual deviance: 4.1223e-10  on 0  degrees of freedom
AIC: 14.144

Number of Fisher Scoring iterations: 20
```

Consolidating the disparity

- Very common for software packages to report convergence when the MLEs do not exist.
- Program get fooled by the nearly flat log likelihood.
- Indicators of problems:
  - Large number of Fisher Scoring iterations.
  - Large standard errors (since the log likelihood is nearly flat the inverse of the second derivative is very large).
  - Estimates are not robust (slight changes in the data may induce large changes in the estimates and estimated standard errors).
- Be very careful in fitting models with zero counts.

A non-saturated example

- Now consider these two tables:

x2	A	A	B	B	C	C
y2	0	0	7	8	12	13

x3	A	A	B	B	C	C
y3	0	1	7	8	12	13

- We fit a Poisson GLM with one factor to each table

```
## fit the model for table 1.
x2 <- factor(c("A","A","B","B","C","C"))
y2 <- c(0, 0, 7, 8, 12, 13)
model2 <- glm(y2 ~ x2, family=poisson)
summary(model2)

## fit the model for table 2.
x3 <- x2
y3 <- c(0, 1, 7, 8, 12, 13)
model3 <- glm(y3 ~ x3, family=poisson)
summary(model3)
```

The model summary for the first table

```
Deviance Residuals:
       1         2         3        4         5        6
-0.00002  -0.00002  -0.18466   0.18060  -0.14238   0.14049

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -22.30   29873.26  -0.001    0.999
x2B            24.32   29873.26   0.001    0.999
x2C            24.83   29873.26   0.001    0.999

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 35.07065  on 5  degrees of freedom
Residual deviance:  0.10673  on 3  degrees of freedom
AIC: 22.605

Number of Fisher Scoring iterations: 20
```

- Large standard errors.
- Large number of Fisher Scoring iterations.
- In the marginal table for factor “A”, the counts are zero – hence MLEs do not exist.

The second table

```
Deviance Residuals:
      1        2        3        4        5        6
-1.0000   0.6215  -0.1847   0.1806  -0.1424   0.1405

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.6931     1.0000  -0.693  0.48822
x3B           2.7081     1.0328   2.622  0.00874 **
x3C           3.2189     1.0198   3.156  0.00160 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 29.252  on 5  degrees of freedom
Residual deviance:  1.493  on 3  degrees of freedom
AIC: 25.991

Number of Fisher Scoring iterations: 5
```

- Standard errors look better.
- 'Regular' number of Fisher Scoring iterations.
- In the marginal table for factor “A”, the counts are nonzero. MLEs do exist for this marginal table.

Distribution of the deviance and Pearson χ^2 statistic

- The sampling distributions of both statistics converge to chi-squared as $n \to \infty$ for a fixed number of cells I.
- Distribution for χ^2 tends to be more robust for smaller n and more sparse tables, compared to $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$.
- There are many rules of thumb for the adequacy of the approximations. For reviews, see Cressie and Read (1989) and Lawal (1984).
- Differences of deviances/Pearson χ^2 statistics are closer to chi-squared (cf. Binomial GLMs)