Chapter 12 Poisson Regression and Related Loglinear Models

12.1 Introduction

- Categorical data often appear as discrete counts that are considered to be distributed as Poisson:

→ colony counts for bacteria or viruses
→ accidents or equipment failure
→ incidence for diseases

- Interested in estimating a rate or incidence and determining its relationship to a set of explanatory variables:

→ bacteria counts per unit volume
→ cancer deaths per person-months of exposure

12.2 Methodology for Poisson Regression

- Suppose response variable *Y* is distributed as Poisson, with expected value and variance both equal to μ. If we have a single explanatory variable *x*, the regression model is:

$$g(\mu) = \alpha + x\beta$$

where *g* is a link function , in terms of a GLM (usually the log function).

$$\log(\mu) = \alpha + x\beta$$

Can be re-written

$$\mu = e^{\alpha}e^{x\beta}$$

If the explanatory variable x is increased by one unit, it has a multiplicative effect on μ.

- Frequently, discrete counts represent information collected over time or in space, and interest lies in modeling rates.

If the exposure time or volume is denoted as *N*, the rate is *Y*/*N* with expected value μ /*N*. Modeling this rate with a loglinear model is written:

$$\log \frac{\mu}{N} = \alpha + x\beta$$

or

$$\log(\mu) = \alpha + x\beta + \log(N)$$

The term log(*N*) is called an *offset* and must be accounted for in the estimation process. Note that:

$$\mu = \exp\left\{\alpha + x\beta + \log(N)\right\} = Ne^{\alpha}e^{x\beta}$$

So that the mean is proportional to *N*.

- More generally, we can write the model in matrix terms:

$$\mu(x) = \{N(x)\}\{g(\beta \mid x)\}$$

Where $\mu(x)$ is the expected value of the number of events $n(x)$, x is the vector of explanatory variables, and $N(x)$ is the known total exposure to risk in the units in which the events occur.

The rate for incidence is $\lambda(x) = \mu(x) / N(x)$

The loglinear model is written as

$$\log\left\{\frac{\mu(x)}{N(x)}\right\} = \boldsymbol{x}'\boldsymbol{\beta}$$

for counts $n(x)$ with independent Poisson distributions.

The loglinear Poisson model is often written

$$\log\{n_i\} = \log\{N_i\} + \boldsymbol{x}_i'\boldsymbol{\beta}$$

in the generalized linear models framework.

Probability distribution:	Poisson distribution
Link function:	log
Offset:	$\log\{N_i\}$

The offset is a qualitative variable whose regression coefficient is known to be 1.

Poisson Regression

This method also applies when n_1 and n_2 are Poisson with expected values $N_1\lambda_1$ and $N_2\lambda_2$ where N_1 and N_2 can be person-time units of exposure. In this case the null hypothesis is $\lambda_1 = \lambda_2$ and one notes n_1 given $(n_1 + n_2) = n$ is Bin$(n,N_1/N)$

Let n_c= number with disease for control
n_v= number with disease for vaccine

Assume n_c is Poisson $(\lambda_\mathrm{c}N_\mathrm{c})$
n_v is Poisson $(\lambda_\mathrm{v}N_\mathrm{v})$
n_c and n_v independent

where N_c is extent of exposure for controls, and N_v is extent of exposure for vaccine

n_v given (n_v + n_c) = n as a conditional distribution is

$$\text{Bin}\left(n = n_v + n_c, P = \frac{\lambda_v N_v}{\lambda_v N_v + \lambda_c N_c}\right)$$

Then $P = \dfrac{\frac{\lambda_v}{\lambda_c}\left(\frac{N_v}{N_c}\right)}{\frac{\lambda_v}{\lambda_c}\left(\frac{N_v}{N_c}\right)+1} = \dfrac{RC}{RC+1}$ where $R = \dfrac{\lambda_v}{\lambda_c}$, $C = \dfrac{N_v}{N_c}$.

Use (n_v, n_c) or $p = n_v/(n_v + n_c)$ to produce a 100(1 – α)% confidence interval (P_L, P_U) for P.

Use $\dfrac{P}{(1-P)C}$ or $\dfrac{N_c P}{(1-P)N_v}$ as estimator for R and

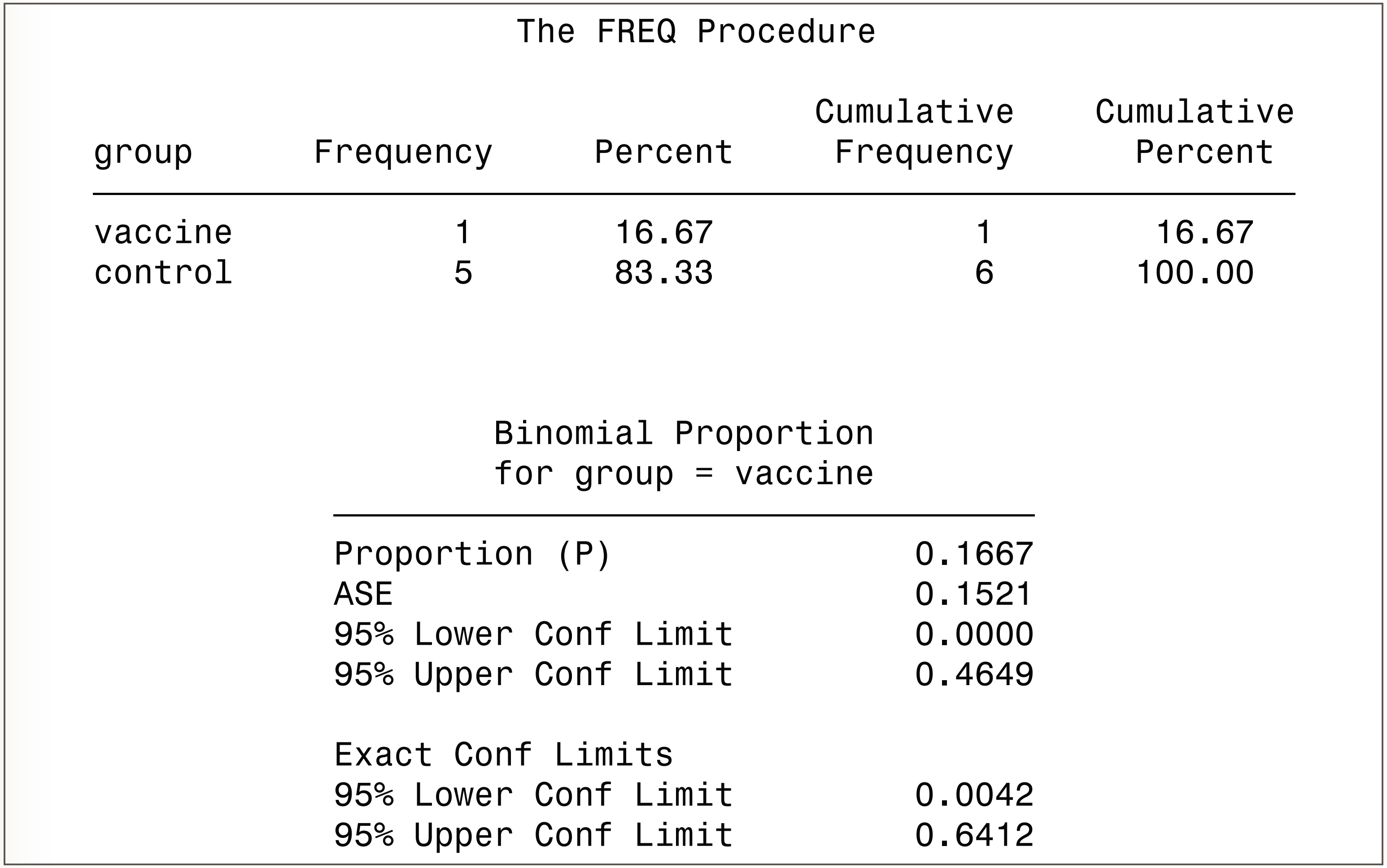
$\left\{\dfrac{P_L}{(1\text{-}P_L)C}, \dfrac{P_U}{(1\text{-}P_U)C}\right\}$ as 100(1$-\alpha$)% confidence interval

for $R = \lambda_v / \lambda_c$.

Example

	Fail	Success	Total
Vaccine	1	19,999	20,000
Control	5	9,995	10,000
Total	6	29,994	30,000

```
proc freq order=data;
  tables group;
  exact bin;
run;
```



```
                         The FREQ Procedure

                                               Cumulative    Cumulative
group         Frequency      Percent           Frequency       Percent
------------------------------------------------------------------------
vaccine               1        16.67                   1         16.67
control               5        83.33                   6        100.00


                        Binomial Proportion
                        for group = vaccine
            -----------------------------------
            Proportion (P)               0.1667
            ASE                          0.1521
            95% Lower Conf Limit         0.0000
            95% Upper Conf Limit         0.4649

            Exact Conf Limits
            95% Lower Conf Limit         0.0042
            95% Upper Conf Limit         0.6412
```

$$C = \frac{N_v}{N_c} = \frac{20{,}000}{10{,}000} = 2$$

$$P = \tfrac{1}{6} = 0.1667$$

$P_L = 0.0042$ (exact 95% lower limit from SAS output)

$P_U = 0.6412$ (exact 95% upper limit from SAS output)

So, the point estimate for $R = \lambda_v/\lambda_c$ is

$$\frac{P}{(1-P)C} = \frac{1/6}{(5/6)2} = 0.10$$

and the exact 95% confidence interval for *R* is

$$\left\{\frac{P_L}{(1-P_L)C}, \frac{P_U}{(1-P_U)C}\right\} = \left\{\frac{0.0042}{(0.9958)2}, \frac{0.6412}{(0.3588)2}\right\}$$

$$= \left\{0.0021, 0.8935\right\}$$

Since the upper limit is less than 1, we can conclude that the rate ratio is significantly less than 1, and thus the vaccine is protective against failure relative to control.

Consider a very large, randomized field study to compare failure rates for a new vaccine versus a control condition.

Suppose the data are

	Fail	Success	Total
Vaccine	n_1	$N_1 - n_1$	N_1
Control	n_2	$N_2 - n_2$	N_2
Total	n	$N - n$	N

Note that N_1, N_2 are very large; but n_1, n_2 are very small; e.g., $n \le 20$ and $N \ge 1000$. Consider Fisher's test probability function for n_1

$$\Pr\{n\} = \frac{N_1!N_2!(N-n)!n!}{n_1!n_2!(N_1-n_1)!(N_2-n_2)!N!}$$

$$\approx \frac{n!}{n_1!n_2!}\left(\frac{N_1}{N}\right)^{n_1}\left(\frac{N_2}{N}\right)^{n_2}$$

Since $N_1 \gg n_1$, $N_2 \gg n_2$, $\Pr\{n_1\}$ is binomial $\mathrm{Bin}\left(n, \frac{N_1}{N}\right)$.

Thus, the left-hand (lower) tail p-value for Fisher's test in this case is

$$p = \sum_{j=0}^{n_1} \frac{n!}{j!(n-j)!} \left(\frac{N_1}{N}\right)^j \left(\frac{N_2}{N}\right)^{n-j}$$

For moderately large n so that $\frac{nN_1}{N}, \frac{nN_2}{N} \geq 10$, then n_1 is Approximately normal $N\left(\frac{nN_1}{N}, \frac{nN_1N_2}{N^2}\right)$.

Example

	Fail	Success	Total
Vaccine	1	19,999	20,000
Control	5	9,995	10,000
Total	6	29,994	30,000

$$p = \binom{6}{0}\left(\frac{1}{3}\right)^6 + \binom{6}{1}\left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^5 = \frac{13}{3^6} = \frac{13}{729} \approx 0.02$$

12.3 Simple Poisson Counts Example

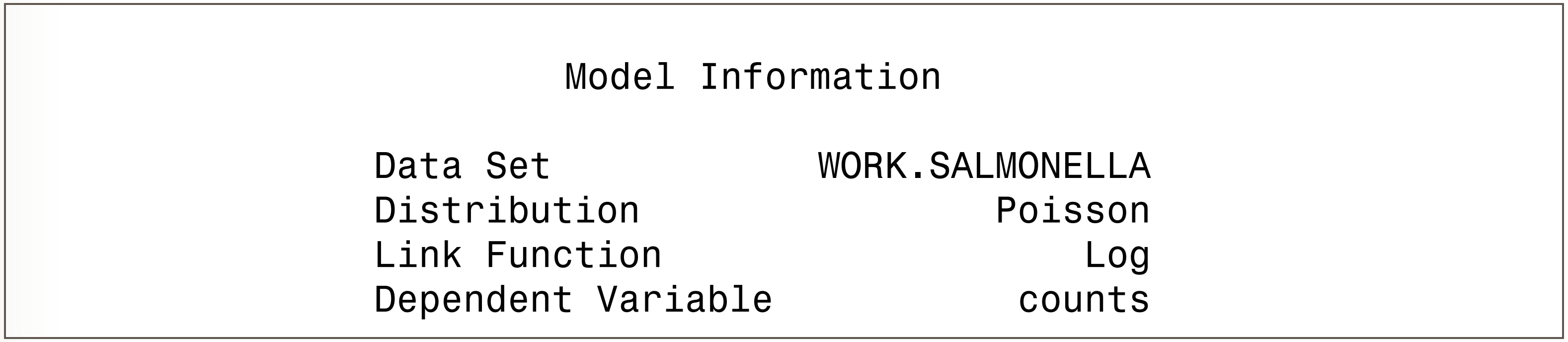
Salmonella Counts

Science Lab	Counts
A	63 64 65 68 69 70 72 73
	75 80 82 83 83 84 84 85 90 91
B	168 171 174 175 185 189 190
	191 195 197 198 198 203 205 205
	207 210 214 216 218

- It is reasonable to consider bacteria counts to be distributed as Poisson.

```
proc genmod;
   class lab;
   model counts=lab/ dist=poisson link=log type3;
run;
```

Model Information



```
                Model Information

         Data Set                WORK.SALMONELLA
         Distribution                    Poisson
         Link Function                       Log
         Dependent Variable               counts
```

Class Variable Information

```
                  Class Level Information

                Class       Levels    Values

                lab              2    A B
```

Assessment of Fit

```
            Criteria For Assessing Goodness Of Fit

Criterion                  DF           Value        Value/DF

Deviance                   36         40.3451          1.1207
Scaled Deviance            36         40.3451          1.1207
Pearson Chi-Square         36         40.0077          1.1113
Scaled Pearson X2          36         40.0077          1.1113
Log Likelihood                     21324.9700
```

Type 3 Analysis

```
                 LR Statistics For Type 3 Analysis

                                         Chi-
       Source                DF        Square     Pr > ChiSq

       lab                    1       1007.19         <.0001
```

Maximum Likelihood Parameter Estimates

```
            Analysis of Maximum Likelihood Parameter Estimates

                      Standard       Wald 95%             Wald
Parameter DF Estimate    Error  Confidence Limits Chi-Square  Pr > ChiSq

Intercept  1   5.2753   0.0160   5.2440    5.3067     108783      <.0001
lab     A  1  -0.9351   0.0313  -0.9965   -0.8738     892.34      <.0001
lab     B  0   0.0000   0.0000   0.0000    0.0000          .           .
Scale      0   1.0000   0.0000   1.0000    1.0000

NOTE:  The scale parameter was held fixed.
```

12.4 Poisson Regression for Incidence Densities

- Interested in fitting a model to the log rate, or incidence densities, of melanoma exposure (involves including an offset).

New Melanoma Cases
Among White Males: 1969-1971

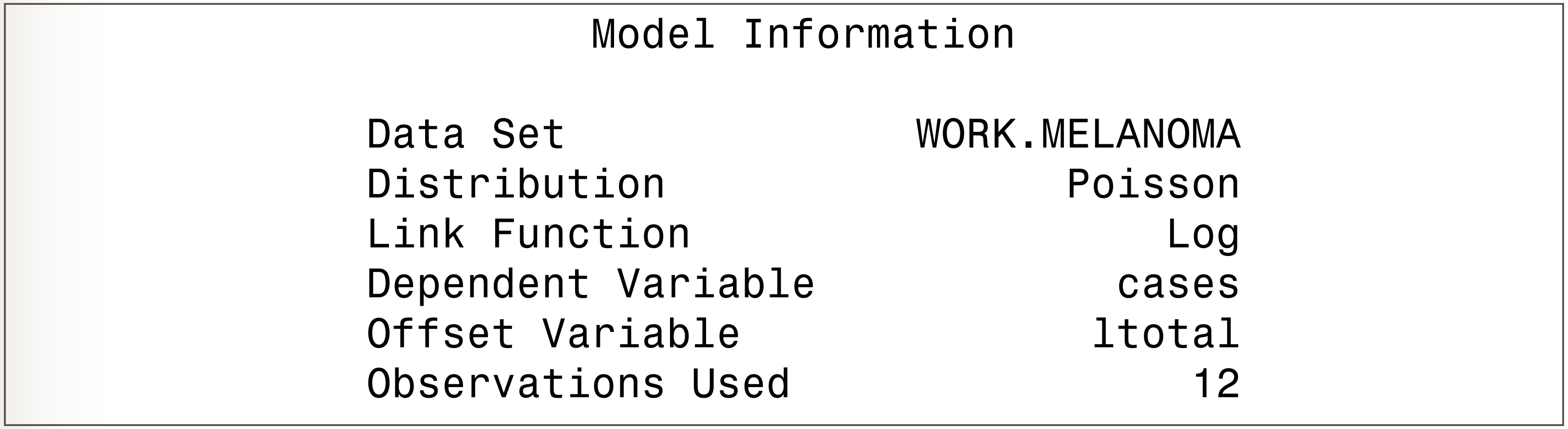
Region	Age Group	Cases	Total
Northern	< 35	61	2880262
Northern	35-44	76	564535
Northern	45-54	98	592983
Northern	55-64	104	450740
Northern	65-74	63	270908
Northern	> 75	80	161580
Southern	< 35	64	1074246
Southern	35-44	75	220407
Southern	45-54	68	198119
Southern	55-64	63	134084
Southern	35-74	45	70708
Southern	> 75	27	34233

- Fit a loglinear model to the ratio of cancer incidence to exposure:

```
data melanoma;
   input age $ region $ cases total;
   ltotal=log(total);
   datalines;
35-44 south 75  220407
45-54 south 68  198119
  ⋮     ⋮    ⋮     ⋮
<35   north 61  2880262
;

proc genmod data=melanoma;
   class region (ref=‘north’) age (ref=‘<35’)/param=ref;
   model cases = age region / dist=poisson
                      link=log offset=ltotal;
run;
```

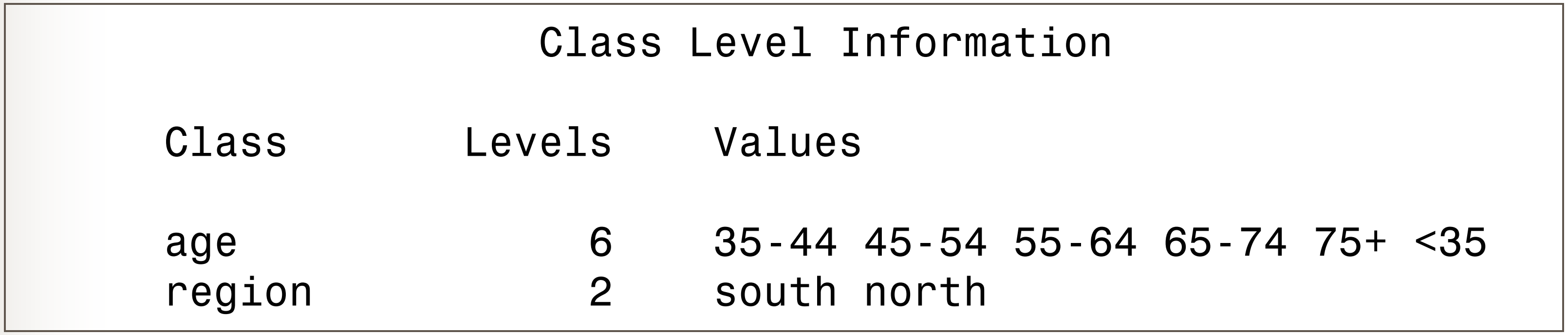
Model Information



```
                      Model Information

             Data Set                      WORK.MELANOMA
             Distribution                        Poisson
             Link Function                           Log
             Dependent Variable                    cases
             Offset Variable                      ltotal
             Observations Used                        12
```

Class Variable Information

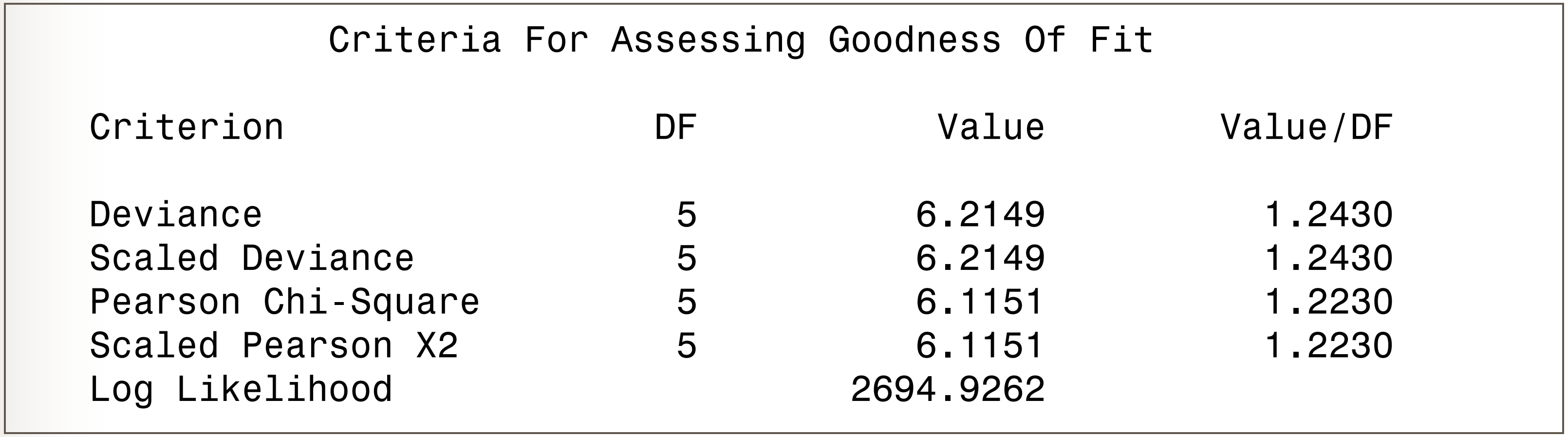


```
                 Class Level Information

     Class         Levels    Values

     age                6    35-44 45-54 55-64 65-74 75+ <35
     region             2    south north
```

Assessment of Fit

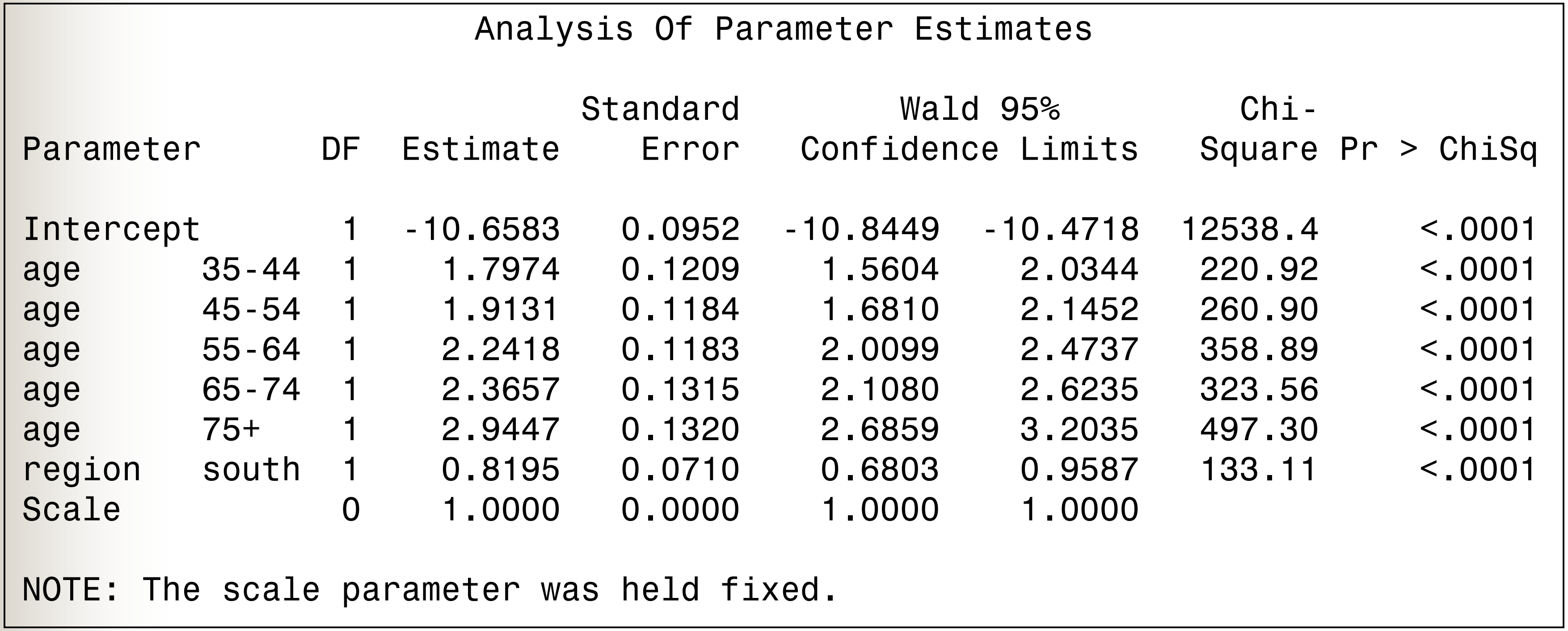


```
              Criteria For Assessing Goodness Of Fit

Criterion                    DF           Value        Value/DF

Deviance                      5          6.2149          1.2430
Scaled Deviance               5          6.2149          1.2430
Pearson Chi-Square            5          6.1151          1.2230
Scaled Pearson X2             5          6.1151          1.2230
Log Likelihood                        2694.9262
```

Estimated Model Parameters



```
                         Analysis Of Parameter Estimates

                                  Standard        Wald 95%           Chi-
Parameter        DF   Estimate     Error    Confidence Limits     Square Pr > ChiSq

Intercept         1   -10.6583    0.0952  -10.8449  -10.4718   12538.4      <.0001
age        35-44  1     1.7974    0.1209    1.5604    2.0344    220.92      <.0001
age        45-54  1     1.9131    0.1184    1.6810    2.1452    260.90      <.0001
age        55-64  1     2.2418    0.1183    2.0099    2.4737    358.89      <.0001
age        65-74  1     2.3657    0.1315    2.1080    2.6235    323.56      <.0001
age        75+    1     2.9447    0.1320    2.6859    3.2035    497.30      <.0001
region     south  1     0.8195    0.0710    0.6803    0.9587    133.11      <.0001
Scale             0     1.0000    0.0000    1.0000    1.0000

NOTE: The scale parameter was held fixed.
```

- You can exponentiate these parameters to express incidence density ratios in a manner similar to exponentiating parameters in logistic regression to obtain odds ratios

```
proc genmod data=melanoma;
   class region (ref=‘north’) age (ref=‘<35’)/param=ref;
   model cases = age region / dist=poisson link=log offset=ltotal;
   estimate ‘45-54 vs. <35’   age 0 1 0 0 0 / exp;
   estimate ‘South vs. North’ region 1 / exp;
run;
```

```
                         Contrast Estimate Results

                         Mean          Mean            L’Beta   Standard
Label                Estimate  Confidence Limits  Estimate      Error   Alpha

45-54 vs. <35          6.7740    5.3707    8.5440    1.9131     0.1184    0.05
Exp(45-54 vs. <35)                                   6.7740     0.8023    0.05
South vs. North        2.2693    1.9744    2.6083    0.8195     0.0710    0.05
Exp(South vs. North)                                 2.2693     0.1612    0.05

                           L’Beta
Label                 Confidence Limits   Chi-Square    Pr > ChiSq

45-54 vs. <35           1.6810    2.1452      260.90        <.0001
Exp(45-54 vs. <35)      5.3707    8.5440
South vs. North         0.6803    0.9587      133.11        <.0001
Exp(South vs. North)    1.9744    2.6083
```

12.5 Overdispersion in Lower Respiratory Infection Example

- Data: 284 children examined every 2 weeks for one year. Explanatory variables: passive smoking, SES, crowding. Outcome: total number of times of lower respiratory infection in the year.
- Reasonable to expect that the children experiencing colds are more likely to have other infections; therefore there may be some additional variance, or overdispersion (observed variance is larger than the nominal variance for a particular distribution).

- Overdispersion occurs with some regularity in analysis of proportions and discrete counts, since in assumed distributions (binomial, Poisson), variances are fixed by a single parameter (mean).

- To manage overdispersion, one can adjust the covariance matrix of a Poisson-based analysis with a scaling factor. That is, assume variance to be $\phi\mu$ instead of μ. The chi-square statistic divided by the d.f. is used as ϕ. The covariance matrix is multiplied by ϕ, and the scaled deviance and log likelihoods are divided by ϕ.

- Alternatively, one could manage overdispersion using an assumption of a Negative Binomial distribution for the counts. This distribution has two parameters to accommodate overdispersion (compared to one for the Poisson distribution)

- Poisson Regression without scaling factor

```
data lri;
     input id count risk passive crowding ses
          agegroup race @@;
     logrisk = log(risk/52);
     datalines;
 1  0  42  1  0  2  2  0  96  1  41  1  0  1  2  0  191  0  44  1  0  0  2  0
   ...
   ;
proc genmod data=lri;
   class ses race agegroup / param=ref;
   model count = passive crowding ses race agegroup
                / dist=poisson offset=logrisk type3;
run;
```

Model Information: Poisson Regression

```
                        The GENMOD Procedure

                         Model Information

              Data Set                    WORK.LRI
              Distribution                 Poisson
              Link Function                    Log
              Dependent Variable             count
              Offset Variable              logrisk
              Observations Used                284
```

Fit Statistics: Poisson Regression

```
              Criteria For Assessing Goodness Of Fit

 Criterion                 DF           Value        Value/DF

 Deviance                 276        408.1549          1.4788
 Scaled Deviance          276        408.1549          1.4788
 Pearson Chi-Square       276        495.4494          1.7951
 Scaled Pearson X2        276        495.4494          1.7951
 Log Likelihood                     -260.4117
```

- Poisson Regression with scaling factor:

Add `scale=pearson` to request Pearson scaling factor

```
proc genmod data=lri;
   class ses race agegroup / param=ref;
   model count = passive crowding ses race
            agegroup / dist=poisson
            offset=logrisk type3 scale=pearson;
run;
```

Assessment of Fit: Poisson Regression with Scaling Factor

```
                 Criteria For Assessing Goodness Of Fit

    Criterion                    DF          Value        Value/DF

    Deviance                    276       408.1549          1.4788
    Scaled Deviance             276       227.3708          0.8238
    Pearson Chi-Square          276       495.4494          1.7951
    Scaled Pearson X2           276       276.0000          1.0000
    Log Likelihood                       -145.0676
```

Type 3 Analysis: Poisson Regression with Scaling Factor

```
                        LR Statistics For Type 3 Analysis

                                                        Chi-
  Source        Num DF     Den DF    F Value   Pr > F   Square   Pr > ChiSq

  passive            1        276       3.89   0.0494     3.89       0.0484
  crowding           1        276       5.86   0.0162     5.86       0.0155
  ses                2        276       1.22   0.2966     2.44       0.2950
  race               1        276       0.38   0.5408     0.38       0.5403
  agegroup           2        276       1.07   0.3443     2.14       0.3429
```

- Negative Binomial Regression:

```
proc genmod data=lri;
   class ses id race agegroup / param=ref;
   model count = passive crowding ses race agegroup
        / dist=nb offset=logrisk type3;
run;
```

Assessment of Fit: Negative Binomial Regression

```
                  Criteria For Assessing Goodness Of Fit

Criterion                     DF          Value          Value/DF

Deviance                     276       256.9688            0.9310
Scaled Deviance              276       256.9688            0.9310
Pearson Chi-Square           276       298.2410            1.0806
Scaled Pearson X2            276       298.2410            1.0806
Log Likelihood                        -242.2932
```

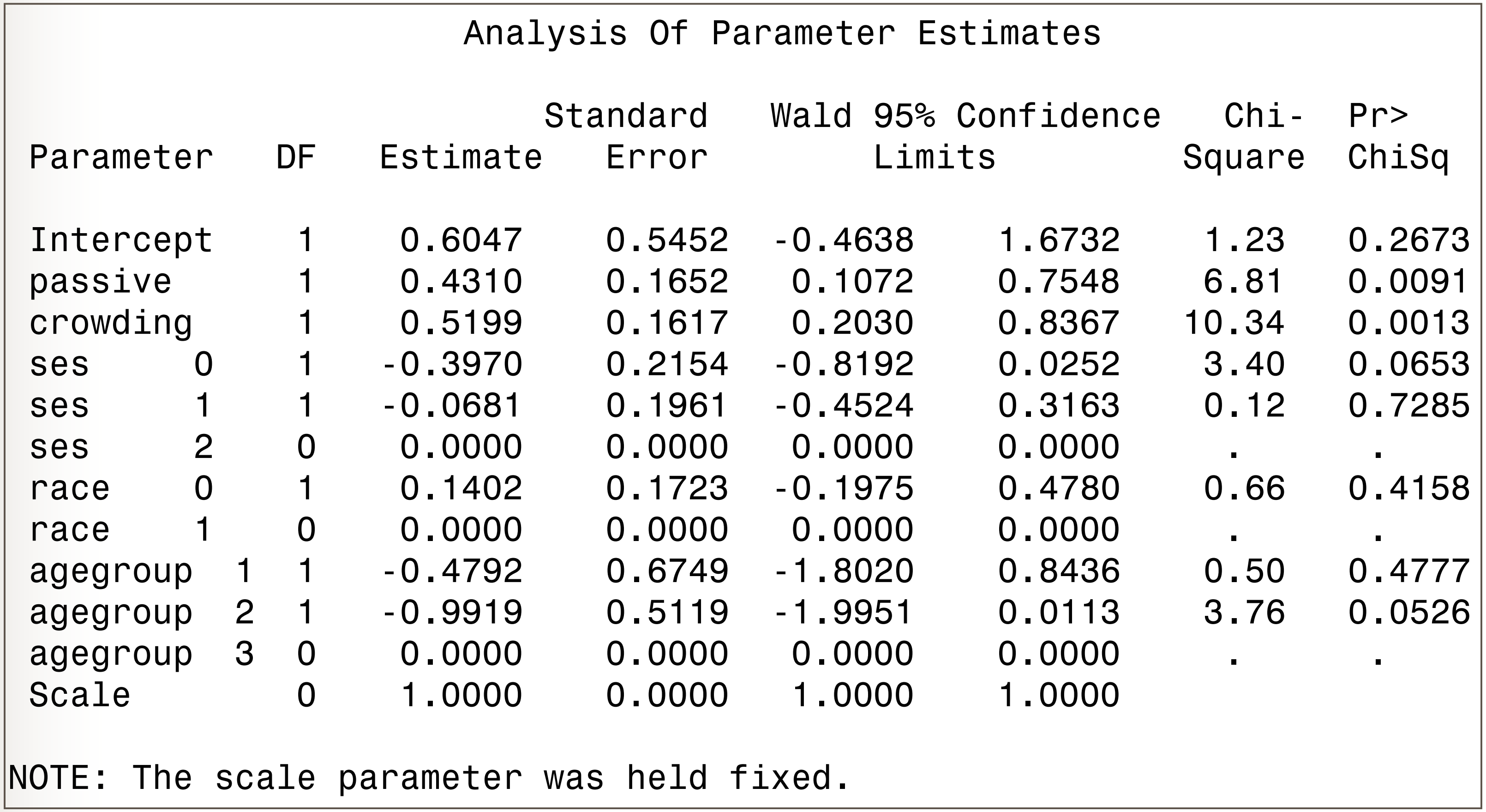
Type 3 Analysis: Negative Binomial Regression

```
                      LR Statistics For Type 3 Analysis

                                        Chi-
              Source           DF     Square     Pr > ChiSq

              passive           1       4.43         0.0353
              crowding          1       5.83         0.0158
              ses               2       2.39         0.3034
              race              1       0.26         0.6112
              agegroup          2       2.92         0.2328
```

Estimated Model Parameters: Poisson Regression



```
                      Analysis Of Parameter Estimates

                              Standard   Wald 95% Confidence   Chi-    Pr>
 Parameter   DF   Estimate     Error          Limits          Square   ChiSq

 Intercept    1     0.6047     0.5452   -0.4638     1.6732      1.23    0.2673
 passive      1     0.4310     0.1652    0.1072     0.7548      6.81    0.0091
 crowding     1     0.5199     0.1617    0.2030     0.8367     10.34    0.0013
 ses      0   1    -0.3970     0.2154   -0.8192     0.0252      3.40    0.0653
 ses      1   1    -0.0681     0.1961   -0.4524     0.3163      0.12    0.7285
 ses      2   0     0.0000     0.0000    0.0000     0.0000       .       .
 race     0   1     0.1402     0.1723   -0.1975     0.4780      0.66    0.4158
 race     1   0     0.0000     0.0000    0.0000     0.0000       .       .
 agegroup   1 1    -0.4792     0.6749   -1.8020     0.8436      0.50    0.4777
 agegroup   2 1    -0.9919     0.5119   -1.9951     0.0113      3.76    0.0526
 agegroup   3 0     0.0000     0.0000    0.0000     0.0000       .       .
 Scale        0     1.0000     0.0000    1.0000     1.0000

NOTE: The scale parameter was held fixed.
```

Estimated Model Parameters: Poisson Regression with Scaling Factor

```
                      Analysis Of Parameter Estimates

                             Standard   Wald 95% Confidence    Chi-   Pr>
 Parameter    DF   Estimate    Error         Limits           Square  ChiSq

 Intercept     1     0.6047    0.7304  -0.8269     2.0362      0.69   0.4077
 passive       1     0.4310    0.2214  -0.0029     0.8649      3.79   0.0515
 crowding      1     0.5199    0.2166  -0.0953     0.9444      5.76   0.0164
 ses       0   1    -0.3970    0.2886  -0.9627     0.1687      1.89   0.1690
 ses       1   1    -0.0681    0.2627  -0.5830     0.4469      0.07   0.7956
 ses       2   0     0.0000    0.0000   0.0000     0.0000       .      .
 race      0   1     0.1402    0.2309  -0.3123     0.5928      0.37   0.5436
 race      1   0     0.0000    0.0000   0.0000     0.0000       .      .
 agegroup   1  1    -0.4792    0.9043  -2.2516     1.2931      0.28   0.5962
 agegroup   2  1    -0.9919    0.6858  -2.3361     0.3522      2.09   0.1481
 agegroup   3  0     0.0000    0.0000   0.0000     0.0000       .      .
 Scale         0     1.3398    0.0000   1.3398     1.3398

NOTE: The scale parameter was estimated by the square root of Pearson's
Chi-Square/DOF.
```

Estimated Model Parameters: Negative Binomial Regression

```
                 Analysis Of Maximum Likelihood Parameter Estimates

                                   Standard         Wald 95%       Wald Chi-
Parameter    DF    Estimate     Error      Confidence Limits   Square   Pr > ChiSq

Intercept    1      0.6751    0.6333    -0.5661      1.9163    1.14     0.2864
passive      1      0.4530    0.2144     0.0329      0.8732    4.47     0.0346
crowding     1      0.5017    0.2061     0.0978      0.9057    5.93     0.0149
ses      0   1     -0.3987    0.2933    -0.9736      0.1762    1.85     0.1740
ses      1   1     -0.0857    0.2775    -0.6296      0.4582    0.10     0.7574
ses      2   0      0.0000    0.0000     0.0000      0.0000     .           .
race     0   1      0.1178    0.2320    -0.3368      0.5725    0.26     0.6115
race     1   0      0.0000    0.0000     0.0000      0.0000     .           .
agegroup 1   1     -0.5652    0.8082    -2.1494      1.0189    0.49     0.4843
agegroup 2   1     -1.0131    0.6006    -2.1902      0.1641    2.85     0.0917
agegroup 3   0      0.0000    0.0000     0.0000      0.0000     .           .
Dispersion   1      0.9760    0.2593     0.4678      1.4843

NOTE: The negative binomial dispersion parameter was estimated by maximum
likelihood.
```

12.6 Exact Poisson Regression

- Exact Poisson regression is a useful strategy when you have small numbers of events because it does not depend on asymptotic results, but uses conditional distributions of the sufficient statistics of the parameters

- Example:

Medical Events for Rotavirus Vaccine Study

Region	Vaccine		Placebo	
	Events	Person Years	Events	Person Years
United States	3	7500	58	7250
Latin America	1	1250	10	1250

- Request exact analysis with CLTYPE=EXACT.
- ESTIMATE=ODDS option requests incidence density ratios

```
data rotavirus;
   input region $ treatment $ counts years_risk @@ ;
   logrisk = log(years_risk);
datalines;
US      Vaccine  3  7500   US      Placebo   58   7250
LA      Vaccine  1  1250   LA      Placebo   10   1250
;
run;

proc genmod order=data;
   class region treatment / param=ref;
   model counts = treatment region / dist=poisson
                         offset=log_risk type3;
   estimate ‘ treatment’ treatment 1 / exp;
   exact treatment / estimate=odds cltype=exact;
run;
```

Assessment of Fit

```
                   Criteria For Assessing Goodness Of Fit

Criterion                     DF          Value        Value/DF

Deviance                       1         0.2979          0.2979
Scaled Deviance                1         0.2979          0.2979
Pearson Chi-Square             1         0.3431          0.3431
Scaled Pearson X2              1         0.3431          0.3431
Log Likelihood                         189.6784
```

Parameter Estimates

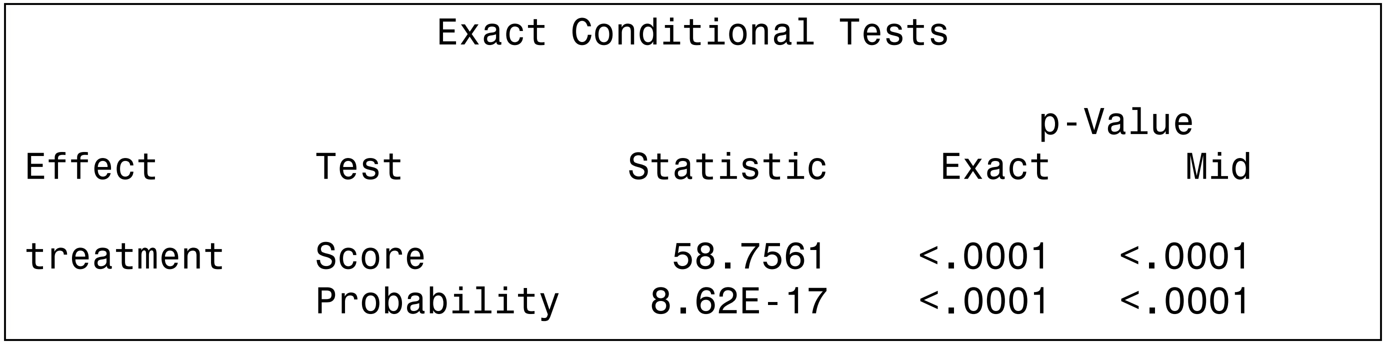
```
              Analysis Of Maximum Likelihood Parameter Estimates

                                  Standard      Wald 95%       Wald Chi-   Pr >
Parameter         DF   Estimate      Error  Confidence Limits     Square  ChiSq

Intercept          1    -4.7886     0.3028   -5.3820  -4.1951    250.11  <.0001
treatment Vaccine 1    -2.8620     0.5145   -3.8704  -1.8536     30.94  <.0001
region    US       1    -0.0467     0.3276   -0.6888   0.5953      0.02  0.8865
Scale              0     1.0000     0.0000    1.0000   1.0000

NOTE: The scale parameter was held fixed
```

Exact Conditional Tests

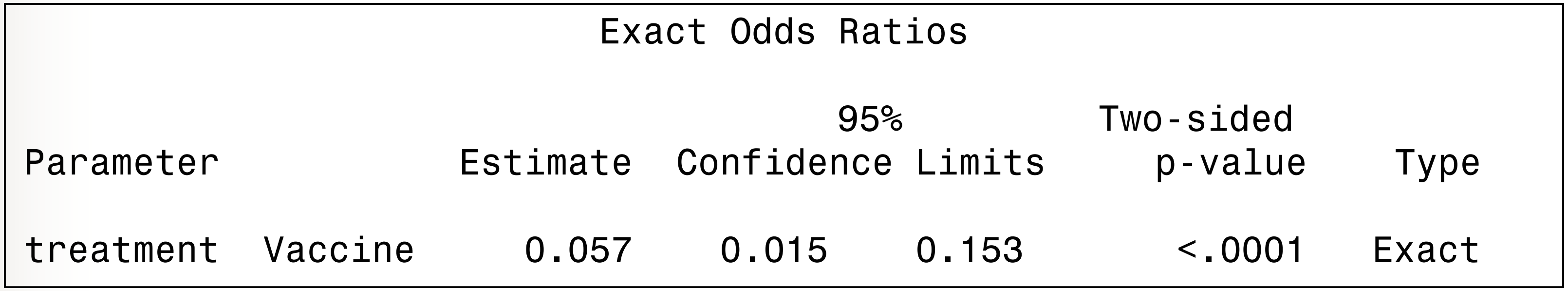


```
                       Exact Conditional Tests


                                                           p-Value
Effect         Test                Statistic       Exact         Mid

treatment      Score                 58.7561      <.0001      <.0001
               Probability          8.62E-17      <.0001      <.0001
```

Exact IDR Estimates and Confidence Intervals



```
                         Exact Odds Ratios


                                         95%           Two-sided
Parameter               Estimate   Confidence Limits     p-value     Type

treatment   Vaccine        0.057      0.015      0.153      <.0001    Exact
```

Occupational Health Study Pertaining to Byssinosis Complaints

- Mantel-Haenszel methods for associations of byssinosis complaints with workplace as dusty or not, employment duration as ≥ 10 years or not, and smoking history as yes or no

- Logistic Regression model for proportions with byssinosis complaints

Byssinosis Complaints

Workplace Conditions	Years of Employment	Smoking	Complaints		
			Yes	No	Proportion
Dusty	<10	Yes	30	203	0.129
Dusty	<10	No	7	119	0.056
Dusty	≥10	Yes	57	161	0.261
Dusty	≥10	No	11	81	0.120
Not Dusty	<10	Yes	14	1340	0.010
Not Dusty	<10	No	12	1004	0.012
Not Dusty	≥10	Yes	24	1360	0.017
Not Dusty	≥10	No	10	986	0.010

Poisson regression for log-linear models for WxExSxB counts

- Model 0 (full model) = W E W*E S W*S E*S W*E*S B W*B E*B W*E*B S*B W*S*B E*S*B W*E*S*B
- Model 1 (full model excluding 4-way interaction and E*S*B interaction)
- Model 2 (full model excluding 4-way interaction and all 3-way interactions with B)
- Model 3 = W E W*E S W*S B W*B E*B S*B
- Models 0, 1, 2 (but not Model 3) have identical results for corresponding logistic regression models with the variables being those that include B but with B deleted from them. Independent binomial distributions are assumed for the respective (W x E x S) cross-classifications.

Wald Test Statistics and *p*-values for Poisson Regression Loglinear Models for Byssinosis Data

	Model 0		Model 1		Model 2	
Variable	Q_W	*p*-value	Q_W	*p*-value	Q_W	*p*-value
Intercept	7990.10	<.0001	8117.64	<.0001	8102.94	<.0001
W	90.51	<.0001	93.08	<.0001	93.32	<.0001
E	1.12	0.2900	2.09	0.1485	4.00	0.0454
W x E	0.02	0.8745	0.00	0.9734	9.01	0.0027
S	57.65	<.0001	61.56	<.0001	70.15	<.0001
W x S	11.67	0.0006	11.74	0.0006	14.68	0.0001
E x S	1.95	0.1624	1.21	0.2704	1.82	0.1772
W x E x S	0.27	0.6002	0.42	0.5192	0.86	0.3532
B	1046.29	<.0001	1055.96	<.0001	1051.02	<.0001
W x B	156.95	<.0001	157.32	<.0001	247.62	<.0001
E x B	6.95	0.0084	10.79	0.0010	13.52	0.0002
W x E x B	2.96	0.0855	3.05	0.0808	X	N/A
S x B	8.57	0.0034	9.40	0.0022	10.29	0.0013
W x S x B	3.46	0.0629	3.34	0.0677	X	N/A
E x S x B	0.85	0.3555	X	N/A	X	N/A
W x E x S x B	0.69	0.4067	X	N/A	X	N/A
Residual	Not Applicable	Not Applicable	Q_L=1.60	0.4490	Q_L=8.10	0.0879
			(d.f.=2)		(d.f.=4)	

Estimates and Standard Errors for Poisson Regression Loglinear Models for Byssinosis Data

	Model 0		Model 1		Model 2	
Variable	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.
Intercept	4.3864	0.0491	4.3889	0.0487	4.3906	0.0488
W	-0.4668	0.0491	-0.4700	0.0487	-0.4168	0.0431
E	0.0519	0.0491	0.0637	0.0441	0.0846	0.0423
W x E	0.0077	0.0491	-0.0015	0.0441	-0.0685	0.0228
S	0.3726	0.0491	0.3768	0.0480	0.3906	0.0466
W x S	0.1676	0.0491	0.1645	0.0480	0.0875	0.0228
E x S	0.0686	0.0491	0.0247	0.0224	0.0299	0.0222
W x E x S	-0.0257	0.0491	0.0144	0.0224	0.0206	0.0221
B	-1.5873	0.0491	-1.5847	0.0488	-1.5811	0.0488
W x B	0.6148	0.0491	0.6117	0.0488	0.6671	0.0424
E x B	0.1294	0.0491	0.1414	0.0430	0.1563	0.0425
W x E x B	0.0844	0.0491	0.0752	0.0430	X	N/A
S x B	0.1437	0.0491	0.1478	0.0482	0.1526	0.0476
W x S x B	0.0913	0.0491	0.0881	0.0482	X	N/A
E x S x B	0.0453	0.0491	X	N/A	X	N/A
W x E x S x B	-0.0407	0.0491	X	N/A	X	N/A

Estimates, Standard Errors, Wald Test Statistics and *p*-values, with Likelihood Ratio Test Statistics and *p*-values for a further simplified Poisson Regression Loglinear Model for Byssinosis Data

	Model 3					
Variable	Estimate	s.e.	Q_W	*p*-value	Q_L	*p*-value
Intercept	4.3898	0.0489	8050.72	<.0001		
W	-0.4146	0.0431	92.64	<.0001	77.88	<.0001
E	0.0978	0.0411	5.67	0.0172	5.80	0.0160
W x E	-0.0610	0.0217	7.91	0.0049	7.96	0.0048
S	0.3924	0.0466	70.99	<.0001	84.86	<.0001
W x S	0.0844	0.0226	13.89	0.0002	14.22	0.0002
B	-1.5836	0.0489	1048.73	<.0001	2584.06	<.0001
W x B	0.6677	0.0424	248.11	<.0001	247.04	<.0001
E x B	0.1606	0.0424	14.35	0.0002	14.81	0.0001
S x B	0.1577	0.0474	11.06	0.0009	11.97	0.0005
Residual	Q_L=9.97 (d.f.=6) with p=0.1260					

- This model does not have a counterpart for logistic regression since logistic regression must include WxExS and everything WxExS contains as study design specifications for sample size

Model 3 has a multi–category logistic counterpart

- Independent multinomial distributions for (S x B) are assumed for the respective (W x E) cross-classifications

- Results are identical to those for a Poisson loglinear model except those for intercept, W, E, and (W x E), which do not exist

- The results for S and (W x S) represent the intercept for S and the association of W with S

- The results for B, (W x B), and (E x B) represent the intercept for B and the associations of W and E with B

- The result for (S x B) represents an intercept for the homogeneous association of S and B for the respective (W x E) cross-classifications

Wald Chi-Square tests and *p*-values from WLS analysis for proportions (*p*) with Byssinosis complaints (The model is $\mathbf{E}(\mathbf{p}) = \mathbf{X}\boldsymbol{\beta}$)

Source	d.f.	Chi-Square	*p*-value
Intercept	1	131.56	< 0.0001
Workplace	1	93.89	< 0.0001
Employ_years	1	14.47	0.0001
Workplace*employ_years	1	12.95	*0.0003*
Smoking	1	14.82	0.0001
Workplace*smoking	1	13.29	*0.0003*
Residual (E×S, W×E×S)	2	3.47	0.1761

For this method, independent binomial distributions are assumed for the respective (W×E×S) cross-classifications.

These results indicate very strong interactions of workplace with employment years and smoking in a strictly linear model for proportions. Such interactions correspond to employment years and smoking having very strong associations with Byssinosis complaints in the dusty workplace versus no association with Byssinosis complaints in the not dusty workplace.

Wald Chi-square tests and *p*-values from WLS analysis for a simplified model for proportions with Byssinosis complaints

Source	d.f.	Chi-Square	*p*-value
Intercept	1	131.50	< 0.0001
Workplace	1	93.98	< 0.0001
Employ_years (workplace=dusty)	1	13.90	0.0002
Smoking (workplace=dusty)	1	14.27	0.0002
Residual	4	4.68	0.3215

This model has a structure for which there is no variation of Byssinosis proportions within the non-dusty workplace and with employment years and smoking having additive effects within the dusty workplace. Predicted proportions are shown in the last table.

Predicted Proportions with Byssinosis Complaints from the Linear Model with WLS and Poisson Loglinear Model 2 and Model 3 (via predicted probabilities for the WxExSxB cross-classification)

Work Condition	Years of Employ-ment	Smoking	Byssinosis Complaints	WLS Predicted Proportion	Predicted Proportion from Model 2	Predicted W x E x S x B from Model 3	Model 3 Predicted Proportion
Dusty	<10	Yes	Yes	0.140	0.1376	0.0061	0.1377
			No			0.0382	
		No	Yes	0.046	0.0797	0.0017	0.0776
			No			0.0202	
	≥ 10	Yes	Yes	0.241	0.2297	0.0090	0.2314
			No			0.0299	
		No	Yes	0.146	0.1393	0.0025	0.1366
			No			0.0158	
Not Dusty	<10	Yes	Yes	0.012	0.0109	0.0027	0.0107
			No			0.2491	
		No	Yes	0.012	0.0060	0.0011	0.0059
			No			0.1844	
	≥ 10	Yes	Yes	0.012	0.0203	0.0052	0.0205
			No			0.2482	
		No	Yes	0.012	0.0111	0.0020	0.0108
			No			0.1837	

Standard errors for estimated proportions with Byssinosis Complaints

Work Condition	Years of Employment	Smoking	Model 0 $SE = \sqrt{p(1-p)/n}$	Model 2	Model 3	WLS s.e.
Dusty	<10	Yes	0.02196	0.01810	0.01797	0.020057
		No	0.02048	0.01498	0.01477	0.0189
	≥ 10	Yes	0.02975	0.02418	0.02445	0.024856
		No	0.03388	0.02354	0.02324	0.026402
Not Dusty	<10	Yes	0.00270	0.00195	0.00193	0.00158
		No	0.00342	0.00129	0.00126	0.00158
	≥ 10	Yes	0.00347	0.00303	0.00307	0.00158
		No	0.00315	0.00218	0.00215	0.00158

Comments

1. Mantel-Haenszel methods enable assessments without strictly requiring a formal model (which might not fit well) nor requiring assumptions for distributions
2. Interactions can be more evident in a strictly linear model than a log-linear model and vice versa
3. Poisson regression log-linear models (and their single multinomial counterparts) have broader scope than logistic regression models
4. Linear models are somewhat more straightforward to apply with WLS than with maximum likelihood, but are only realistic for situations with categorical explanatory variables and with all predictions in (0,1)
5. Weighted least squares methods need all counts to be at least 5 for all models whereas Poisson regression log-linear models only need this condition for marginal tables corresponding to the highest order interactions in the model

Poisson Regression Analysis of Relationship Between Event Rates and a Set of Explanatory Variables

Applications:

1. Epidemiologic studies: events are occurrences of rare diseases (or experiences) for populations with different sizes; explanatory variables are background covariables and risk factors; enumeration of number of events and determination of population size are through possibly different data sources.

2. Epidemiologic studies: events are occurrences of rare diseases (or experiences) for individuals with possibly different amounts of exposure to risk; explanatory variables are background covariables and risk factors.

3. Clinical trials: events are occurrences of rare disorders for individuals with possibly different levels of exposure to risk; explanatory variables are treatment and background covariables. The rare disorders in a vaccine study are the diseases to be prevented; in other studies, they can correspond to unfavorable side effects of a treatment.

Model specifications for expected total number of events for subject with total exposure *N* and x_1, x_2, …, x_t status for *t* explanatory variables is as follows:

Expected total:

$$\mu = N\lambda = N\exp\left(\alpha + \sum_{k=1}^{t}\beta_k x_k\right).$$

Incidence density:

$$\lambda = (\mu / N) = \exp\left(\alpha + \sum_{k=1}^{t}\beta_k x_k\right).$$

Log(Incidence density) $= \ln\lambda = \alpha + \sum_{k=1}^{t}\beta_k x_k$

The possible range for the α and the β_k is $(-\infty, \infty)$. The $\exp(\beta_k)$ are incidence density ratios for unit changes in the x_k, i.e., the amounts by which the incidence density λ is multiplied per unit change in x_k. When the total numbers of events in $n_1, n_2, \ldots, n_s$ for s populations with total exposures $N_1, N_2, \ldots, N_s$ approximately have independent Poisson distributions, maximum likelihood equations for estimation of α and the $\{\beta_k\}$ have the structure

$$\sum_{i=1}^{s} \hat{\mu}_i (1, x_{i1}, \ldots, x_{it}) = \sum_{i=1}^{s} n_i (1, x_{i1}, \ldots, x_{it}) \text{ where}$$

$$\hat{\mu}_i = \exp\left(\hat{\alpha} + \sum_{k=1}^{t} \hat{\beta}_k x_k \right) \text{ is the model predicted value for } \mu_i.$$

The maximum likelihood estimates $\hat{\alpha}$ and $\{\hat{\beta}_k\}$ are obtained by iterative solution of these equations. When the linear functions $\sum_{i=1}^{s} n_i(1, x_{i1}, \ldots, x_{it})$ are based on sufficient sample size to have an approximately multivariate normal distribution, then $\hat{\alpha}$ and $\{\hat{\beta}_k\}$ have an approximately multivariate normal distribution for which a consistent estimate of covariance structure is available.

Linear hypotheses concerning the $\{\hat{\beta}_k\}$ can be assessed with log-likelihood ratio chi-square statistics or Wald statistics. When sample sizes are sufficiently large (e.g., 80% of the n_i are ≥ 5 and all others are ≥ 2), model goodness of fit can be assessed with Pearson chi-squared statistic

$Q_P = \sum_{i=1}^{s} (n_i - \hat{\mu}_i)^2 / \hat{\mu}_i$ or the log-likelihood ratio statistic

$$Q_L = \sum_{i=1}^{s} 2n_i \log(n_i / \hat{\mu}_i).$$

Each of these criteria approximately has the chi-squared distribution with d.f. = (s – 1 – t) in this situation. More generally, goodness of fit can be assessed by using the log-likelihood ratio statistics to evaluate the impact of expansion of a model to include additional explanatory variable. If such expansions have negligible impact, goodness of fit is supported.