Machine Learning for Business

Module 7: Poisson regression
Day 4, 9.00 – 12.00

Asst. Prof. Dr. Santitham Prom-on

Department of Computer Engineering, Faculty of Engineering
King Mongkut's University of Technology Thonburi

BIG DATA EXPERIENCE CENTER

Regression

$$f(x) = b_0 + b_1 x_1 + b_2 x_2 + \ldots$$

Poisson regression

So far we have looked at 2 kinds of regression

- Linear regression (simple, multiple)
  - Continuous response, normally distributed with constant variance
  - Mean a linear function of the covariates
- “Logistic regression”
  - Response (number of successes in n trials) has a binomial distribution Bin(n, p)
  - Mean is *np* where log-odds of *p* is a linear function of the covariates

Example - Count data

```
ceb.data <- read.csv("4-ceb.csv")
head(ceb.data)
```

```
# index duration resident education mean  var  n
#     1      0-4     Suva      none 0.50 1.14  8
#     2      0-4     Suva     lower 1.14 0.73 21
#     3      0-4     Suva     upper 0.90 0.67 42
#     4      0-4     Suva      sec+ 0.73 0.48 51
#     5      0-4    urban      none 1.17 1.06 12
#     6      0-4    urban     lower 0.85 1.59 27
```

Poisson regression (cont)

- Now we consider “Poisson regression”
  - Response is a count, assumed to have a Poisson distribution with (positive) mean μ
  - Assume that log μ is a linear function of the covariates
  - Alternatively,

    μ = exp(linear function of covariates)
- Poisson is a standard distribution when response is a count.

Poisson distribution

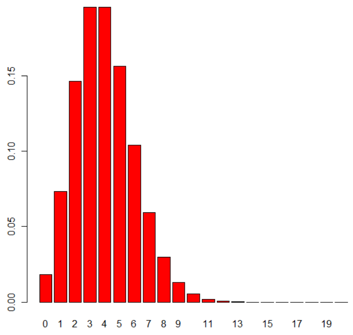
$$\Pr(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!}$$

Count Y can have values 0, 1, 2, . . .

(thus, mean μ must be positive)

Poisson distribution (cont)

```
x <- 0:20
poisson.mean <- 4
poisson.probs <- dpois(x,poisson.mean)
names(poisson.probs) <- x
barplot(poisson.probs, col="red")
poisson.probs
```



The Poisson Regression Model

- The response Y with covariates $x_1, \ldots, x_k$ has a Poisson distribution, with mean μ

- Mean μ is related to the covariates by

$$\log(\mu) = \beta_0 + \beta_1 x_1 + . . . + \beta_k x_k$$

- As for logistic regression, the parameters are estimated by maximum likelihood

Interpretation of β's

- If $\beta_j > 0$, mean ***increases*** with x_j
- If $\beta_j < 0$, mean ***decreases*** with x_j

- Unit increase in x_j changes the mean by a factor of $\exp(\beta_j)$

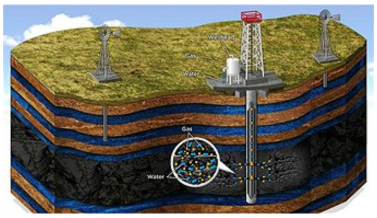
  (like the odds in logistic regression)

Estimation of β's

- To estimate the β's , we use the method of *maximum likelihood,* as in logistic regression
- Basic idea:
  - Using the ***Poisson*** distribution, we can work out the probability of getting any particular set of responses y.
  - In particular, we can work out the probability of getting the data we actually observed – this is the likelihood
  - Choose β's to maximise this probability (or, equivalently, the log-likelihood)

Example: Mining accident data

- This example features the number of accidents per mine in a 3 month period in 44 coal mines in West Virginia. The variables are
  - COUNT: the number of accidents (response)
  - INB: inner burden thickness
  - EXTRP: percentage of coal extracted from mine
  - AHS: the average height of the coal seam in the mine
  - AGE: the age of the mine

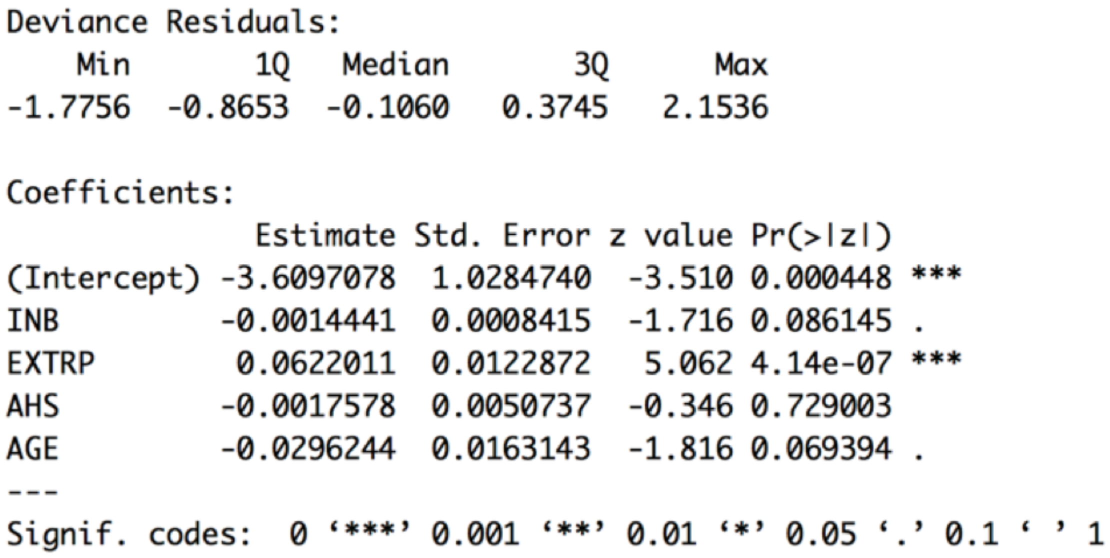


R: read coal mine data

```
mines.df <- read.csv("4-mines.csv")
head(mines.df)

#   COUNT INB EXTRP AHS AGE
# 1     2  50    70  52 1.0
# 2     1 230    65  42 6.0
# 3     0 125    70  45 1.0
# 4     4  75    65  68 0.5
# 5     1  70    65  53 0.5
# 6     2  65    70  46 3.0
```

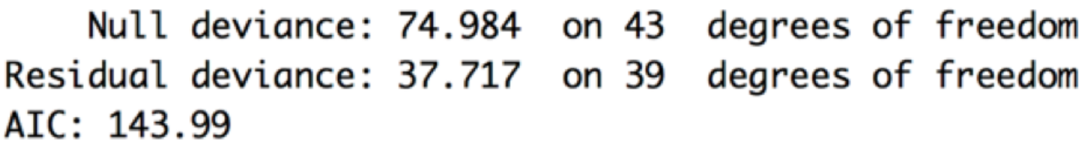
R: Modeling with Poisson regression

```
mines.glm <- glm(COUNT ~ INB+EXTRP+AHS+AGE,
                 family = poisson, data = mines.df)

summary(mines.glm)
```



```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7756  -0.8653  -0.1060   0.3745   2.1536

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.6097078  1.0284740  -3.510 0.000448 ***
INB         -0.0014441  0.0008415  -1.716 0.086145 .
EXTRP        0.0622011  0.0122872   5.062 4.14e-07 ***
AHS         -0.0017578  0.0050737  -0.346 0.729003
AGE         -0.0296244  0.0163143  -1.816 0.069394 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

R: validate goodness of fit

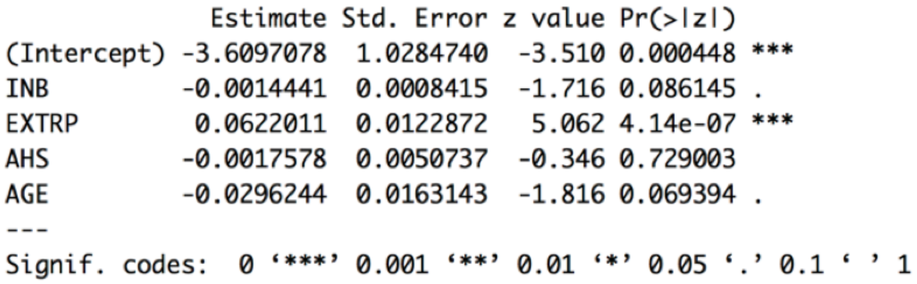


```
    Null deviance: 74.984  on 43  degrees of freedom
Residual deviance: 37.717  on 39  degrees of freedom
AIC: 143.99
```

```
1 - pchisq(37.717,39)
#[1] 0.5283455
```

This is the probability that the result is fitted to the distribution.
It has chi-square distribution.

Interpretation of coefficients

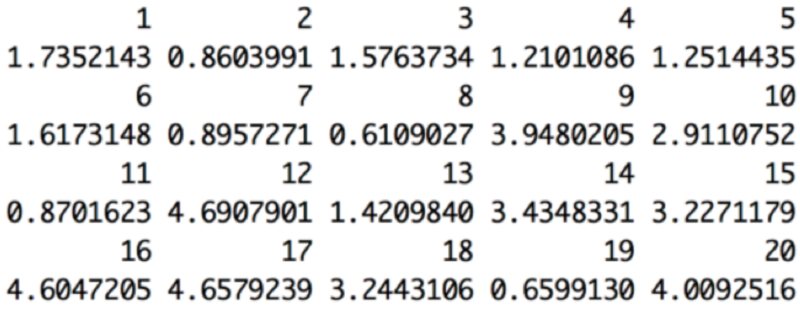


```
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -3.6097078  1.0284740  -3.510 0.000448 ***
INB         -0.0014441  0.0008415  -1.716 0.086145 .  
EXTRP        0.0622011  0.0122872   5.062 4.14e-07 ***
AHS         -0.0017578  0.0050737  -0.346 0.729003    
AGE         -0.0296244  0.0163143  -1.816 0.069394 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- As the inner burden thickness increases, the number of accidents goes down (but only weakly significant)
- As the extraction percentage goes up the accidents go up
- As the age of the mine increases, the number of accidents goes down (but only weakly significant)

Prediction

```
res <- predict(mines.glm,
               mines.df,
               type = "response")
res
```



```
        1         2         3         4         5 
1.7352143 0.8603991 1.5763734 1.2101086 1.2514435 
        6         7         8         9        10 
1.6173148 0.8957271 0.6109027 3.9480205 2.9110752 
       11        12        13        14        15 
0.8701623 4.6907901 1.4209840 3.4348331 3.2271179 
       16        17        18        19        20 
4.6047205 4.6579239 3.2443106 0.6599130 4.0092516 
```

Prediction error

```
# measure Mean Absolute Error (MAE)
mean(abs(mines.df$COUNT - res))

# [1] 0.9307389
```

Example: Onions

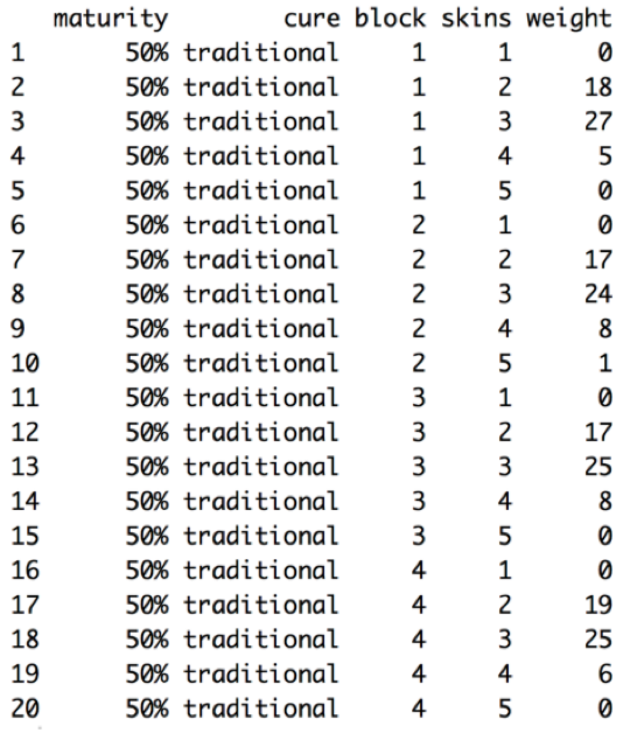
On each onion, the following were measured:

- Maturity: the maturity of the onion. Levels are 50%,. 70%, 90%, 95% and 100%
- Cure: the method of curing: either “traditional”, “shears or “partial”
- Block: the area of land the onions were grown in, one of 1, 2, 3 or 4
- Skins: the number of skins

- To make then data set more compact, the data were grouped.
- The number of onions having the same values of the above variable is recorded as “weight”

R: read the onions data

```
onions.df <- read.csv("4-onions.csv")
head(onions.df, n = 20)
```



```
   maturity        cure block skins weight
1       50% traditional     1     1      0
2       50% traditional     1     2     18
3       50% traditional     1     3     27
4       50% traditional     1     4      5
5       50% traditional     1     5      0
6       50% traditional     2     1      0
7       50% traditional     2     2     17
8       50% traditional     2     3     24
9       50% traditional     2     4      8
10      50% traditional     2     5      1
11      50% traditional     3     1      0
12      50% traditional     3     2     17
13      50% traditional     3     3     25
14      50% traditional     3     4      8
15      50% traditional     3     5      0
16      50% traditional     4     1      0
17      50% traditional     4     2     19
18      50% traditional     4     3     25
19      50% traditional     4     4      6
20      50% traditional     4     5      0
```

Doing it in R

Must treat “block” as a factor

```
> onions.glm<-glm(skins ~ factor(block)*maturity*cure,
family=poisson, weight=weight, data=onions.df)
> anova(onions.glm, test="Chisq")
Analysis of Deviance Table
Model: poisson, link: log
Response: skins
Terms added sequentially (first to last)
                            Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                          227     680.81
factor(block)                3     1.62       224     679.19      0.65
maturity                     4    70.46       220     608.72 1.814e-14
cure                         2     1.07       218     607.66      0.59
factor(block):maturity      12     2.43       206     605.23      1.00
factor(block):cure           6     0.81       200     604.43      0.99
maturity:cure                8     2.43       192     602.00      0.96
factor(block):maturity:cure 24     3.06       168     598.93      1.00
```

Takes care of the repetitions

No evidence cure has an effect, p-value is 0.65. Strong maturity effect

Fit simpler model

```
> model2<-glm(skins ~ maturity, family=poisson,
weight=weight, data=onions.df)
> summary(model2)
Coefficients:
            Coefficients:
           Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.76547    0.02784  27.493  < 2e-16 ***
maturity50%  0.28302    0.03687   7.676 1.64e-14 ***
maturity70%  0.21786    0.03740   5.825 5.70e-09 ***
maturity90%  0.15282    0.03795   4.026 5.66e-05 ***
maturity95%  0.09813    0.03844   2.552   0.0107 *
Null deviance: 680.81  on 227  degrees of freedom
Residual deviance: 610.35  on 223  degrees of freedom
AIC: 8967.2
```

Baseline is 100%

Number of skins goes down as onions get more mature

Offsets

- Often Poisson data are concerned with rates, as in death rates or accident rates in a population.
- In this case, we look at the number of deaths from a particular cause over a period of time. The number of deaths will be related to the population size, and the time, which we have to take into account.
- We could treat the data as binomial, but a more common approach is to use ***offsets***.

Offsets (cont)

- We imagine that the number of deaths in a fixed period in a population is Poisson, with mean of the form

  μ = (population size/100000) × mean for a population of size 100,000)

Taking logs, we get

log (μ) = log(population size/100000) + log(mean of standardised population)

Offsets (cont)

- The last term (the rate per 100,000) is modeled in terms of the covariates in the usual way.

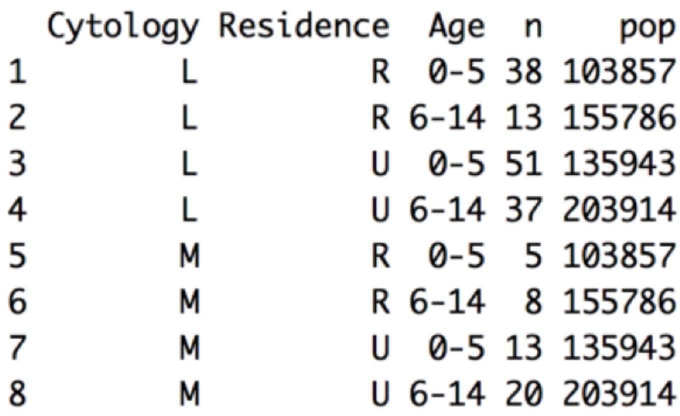
The term log(pop size/100000) is just another variable in the model, except that its regression coefficient is fixed at 1. Such a variable is called an ***offset***.

Example

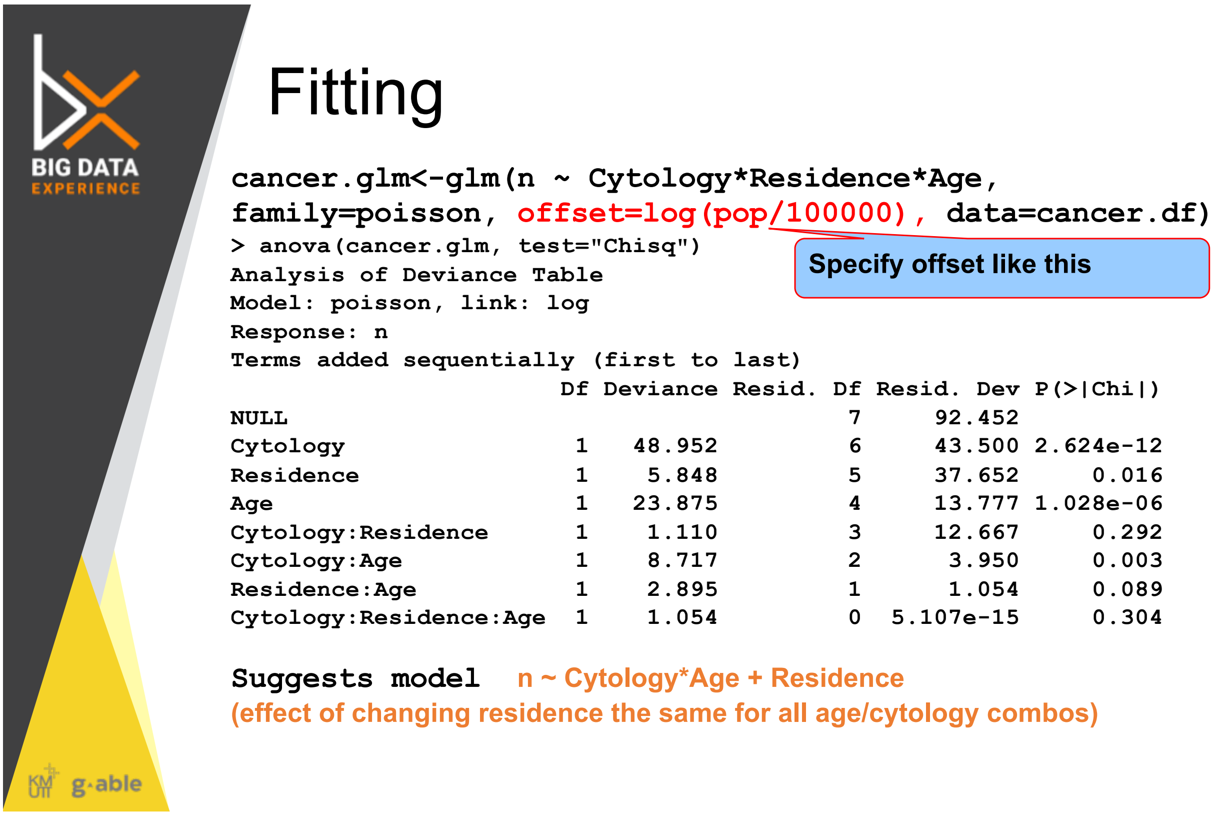
- Deaths from childhood cancers 1951 – 1960 in Northumberland and Durham, classified by
  - Cytology (Lymphoblastic / Myeloblastic)
  - Residence (Rural/Urban)
  - Age (0-5, 6-14)

R: read in cancer data

```
cancer.df <- read.csv("4-cancer.csv")
cancer.df
```



```
  Cytology Residence  Age  n    pop
1        L         R  0-5 38 103857
2        L         R 6-14 13 155786
3        L         U  0-5 51 135943
4        L         U 6-14 37 203914
5        M         R  0-5  5 103857
6        M         R 6-14  8 155786
7        M         U  0-5 13 135943
8        M         U 6-14 20 203914
```



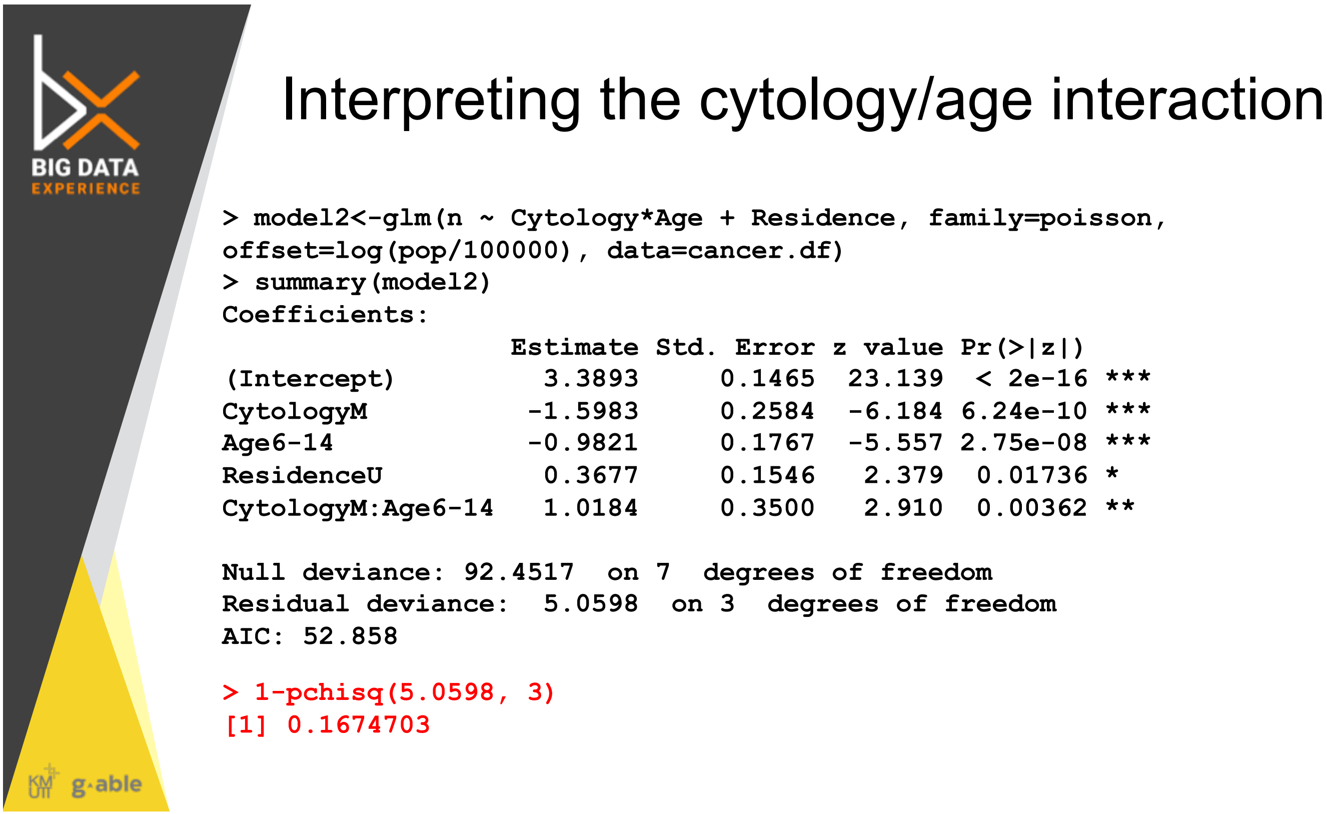
Fitting

```
cancer.glm<-glm(n ~ Cytology*Residence*Age,
family=poisson, offset=log(pop/100000), data=cancer.df)
> anova(cancer.glm, test="Chisq")
Analysis of Deviance Table
Model: poisson, link: log
Response: n
Terms added sequentially (first to last)
                       Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                       7     92.452
Cytology                1   48.952         6     43.500 2.624e-12
Residence               1    5.848         5     37.652     0.016
Age                     1   23.875         4     13.777 1.028e-06
Cytology:Residence      1    1.110         3     12.667     0.292
Cytology:Age            1    8.717         2      3.950     0.003
Residence:Age           1    2.895         1      1.054     0.089
Cytology:Residence:Age  1    1.054         0  5.107e-15     0.304
```

Specify offset like this

Suggests model n ~ Cytology*Age + Residence

(effect of changing residence the same for all age/cytology combos)



Interpreting the cytology/age interaction

```
> model2<-glm(n ~ Cytology*Age + Residence, family=poisson,
offset=log(pop/100000), data=cancer.df)
> summary(model2)
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         3.3893     0.1465  23.139  < 2e-16 ***
CytologyM          -1.5983     0.2584  -6.184 6.24e-10 ***
Age6-14            -0.9821     0.1767  -5.557 2.75e-08 ***
ResidenceU          0.3677     0.1546   2.379  0.01736 *
CytologyM:Age6-14   1.0184     0.3500   2.910  0.00362 **

Null deviance: 92.4517  on 7  degrees of freedom
Residual deviance:  5.0598  on 3  degrees of freedom
AIC: 52.858

> 1-pchisq(5.0598, 3)
[1] 0.1674703
```

Interpreting the cytology/age interaction

For Rural residence, rate per 100,000 is :

Baseline cell

	Age=0-5	Age=6-14
Cytology=L	Exp(3.3893) =29.6	Exp(3.3893-0.9821) =11.1
Cytology=M	Exp(3.3893-1.5983) =5.9	Exp(3.3893-0.9821-1.5983 +1.0184) =6.2

Urban residence increases these rates by a factor of exp(0.3677)=1.44

Activity

- Use the data “4-ceb.csv”
- Create a new column y = mean*n
- Because y comes from multiplication of Poisson distributed variable and the rate, taking log of y results in log(y) = log(mean) + log(n)
- Create a new variable os = log(n)
- Build models to predict y use os as an offset
  - y ~ 1
  - y ~ resident
  - y ~ education
  - y ~ duration
- Interpret the model result, which one affect the number of children

Thank you

Question?