



JOHNS HOPKINS  
BLOOMBERG SCHOOL  
of PUBLIC HEALTH

## Lecture 11

Lecture 10 [ Log-linear regression  
Examples plus  
Case study of excess deaths due to Hurricane Maria

Lecture 11 [

# Review of Lecture 10

- ▶ Use of “marginal” and “conditional” to describe logistic models

- ▶ Lecture 4:  $Y_i$   $i = 1, \dots, n$  independent

- Marginal model: here we were correlating  $Y$  (binary) with a single  $X$  (binary), i.e. evaluating the unadjusted relationship  $\beta_x^m$
- Conditional model: We added information about another covariate  $C$  (possible confounding variable), this makes the interpretation of the log odds ratio for  $X$  conditional, i.e. among persons with the same value of  $C$ , the relative odds of  $Y$  comparing those with and without  $X$  are  $\exp(\beta_x^c)$

- ▶ Lecture 9 and 10:

- Now we are in the case of correlated data: longitudinal or clustered
- ✗ Marginal model: defines that the goal is to make comparisons across subsets of the population or among the same population at different time points, i.e. how does odds of  $Y$  differ when I look at individuals with  $X = 1$  or  $X = 0$
- ✗ Conditional model: Among persons from the same cluster, how does odds of  $Y$  differ when I look at units with  $X = 1$  or  $X = 0$  (only among persons from the same cluster).



# Log-linear models for count variables

- ▶ Count variable
  - ▶ Takes on values of non-negative integers
  - ▶ 0, 1, 2, ..., 3321, ..... 10001, ....
- ▶ Counts of outcomes of interest occurring within a given time range or group of eligible persons
  - ▶ Number of non-accidental deaths per day in Chicago
  - ▶ Number of days of work missed due to illness within a year
  - ▶ Number of myocardial infarctions (MIs) among patients at risk for MI
- ▶ Variability tends to increase as mean increases
- ▶ Effects of predictors tend to be multiplicative (reflecting relative changes not absolute change)



# Poisson process

- ▶ Poisson process defines how observations of events of interest occur over time or space
- ▶ Imagine a range of time  $[0, T]$  and breaking that range of time into small bins  $[t, t+dt]$
- ▶  $\Pr(\text{Event occurs in } [t, t+dt]) = \lambda dt$
- ▶  $\Pr(2 \text{ or more events occur in } [t, t+dt]) \sim 0$
- ▶ Memoryless property: chance of an event in one interval is independent of the chance of an event in a future interval
- ▶ In a Poisson process, the event times in an interval  $[0, T]$  are uniformly distributed, that is, have equal chance of occurring anywhere in the part of the interval.



# Poisson process

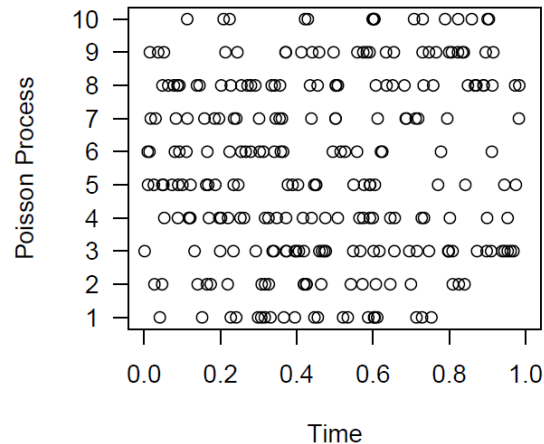
- ▶ The number of events  $X$  occurring in the interval  $[0, T]$  follows a Poisson distribution

- ▶ Probability mass function:  $P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}$ ,  $x = 0, 1, 2, \dots$

See page 3 of Lecture 10 handout for derivation.

- ▶ The mean and variance of  $X$  is  $\lambda T$

10 Realizations of Poisson Process



= 25

# Log-linear model

- ▶ First formulation -> we will assume exposure time is the same for all observations!

- ▶ General form:

$$Y_i \sim P(\mu_i), i = 1, \dots, n \text{ independent}$$

$$\text{Var}(Y_i) = \mu_i$$

$$\log(E(Y_i)) = \log(\mu_i) = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}$$

- ▶ Interpretation:

$\beta_0$  = log of the expected # of events  
given  $X_{i1}, \dots, X_{ip} = 0$

$\beta_1$  = difference in the log expected # of events  
per unit increase in  $X_1$ , holding  $X_2, \dots, X_p$

= Log of the ratio of expected # of fixed  
events comparing  $X_1 = x$  to  $X_1 = x-1$ , all other  
 $X$  fixed

distribution of  $Y_i$

Log link  
inverse link  
exp

# Log-linear model

- ▶ First formulation -> we will assume exposure time is the same for all observations!
- ▶ Hypothetical example: a study of insulin-dependent diabetic patients followed for 4 weeks after acquiring an insulin pump. The patients record and report the total number of hypoglycemic episodes during the 4 week follow-up.
- ▶ The goal of the analysis is to compare the total number of hypoglycemic episodes for male and female diabetic patients



## Example: Same exposure time

$$\text{Log}(E(Y_i)) = \text{Log}(\mu_i) = \beta_0 + \beta_1 \text{male}_i$$

```
set.seed(1346)
N = 100
male = rbinom(N,1,0.5)
Y= rpois(N,exp(log(12)+0.2*male))
summary(glm(Y~male,family="poisson"))$coefficients
```

link = "log"

	Estimate	Std. Error	z value	Pr(> z )
## (Intercept)	2.5176965	0.04016096	62.690141	0.0000000000
## male	0.1956729	0.05421405	3.609266	0.0003070652

- $\hat{\beta}_0$  is the logarithm of the mean number of hypoglycemic episodes during the 4-week follow-up among females. The mean number of hypoglycemic episodes among females during the follow-up is  $\exp(\hat{\beta}_0) = \exp(2.52) = 12.4$ .
- $\hat{\beta}_0 + \hat{\beta}_1$  is the logarithm of the mean number of hypoglycemic episodes during the 4-week follow-up among males. The mean number of hypoglycemic episodes among males during the follow-up is  $\exp(\hat{\beta}_0 + \hat{\beta}_1) = \exp(2.52 + 0.20) = 15.2$ .



## Example: Same exposure time

$$\text{Log}(E(Y_i)) = \text{Log}(\mu_i) = \beta_0 + \beta_1 \text{male}_i$$

```
set.seed(1346)
N = 100
male = rbinom(N,1,0.5)
Y= rpois(N,exp(log(12)+0.2*male))
summary(glm(Y~male,family="poisson"))$coefficients
```

$$Y_i = \lambda_i \times 4$$

4 weeks

##	Estimate	Std. Error	z value	Pr(> z )
## (Intercept)	2.5176965	0.04016096	62.690141	0.0000000000
## male	0.1956729	0.05421405	3.609266	0.0003070652

- $\hat{\beta}_1$  is the difference in the log mean number of hypoglycemic episodes during the 4 week follow-up comparing males to females OR the log relative mean number of hypoglycemic episodes during the 4 week follow-up comparing males to females.
- $\exp(\hat{\beta}_1) = \exp(0.20) = 1.22$  represents the relative mean number of hypoglycemic episodes comparing males to females. The mean number of hypoglycemic episodes during the 4-week follow-up is 22% greater for males compared to females.

$$\frac{15.2}{12.4} = 1.22$$

# Log-linear model

- ▶ Second formulation -> we will NOT assume exposure time is the same for all observations!
- ▶ Hypothetical example: a study of insulin-dependent diabetic patients followed up to 4 weeks after acquiring an insulin pump.
- ▶ Now suppose that not all patients were able to be followed for the entire 4-week period; patients were followed from 10 to 28 days. Patients report the number of hypoglycemic episodes within the duration of the patient's specific follow-up.
- ▶ The goal of the analysis is to compare the total number of hypoglycemic episodes for male and female diabetic patients



## Example: Variable exposure time

$$Y_i \sim P(\mu_i) = P(N_i \lambda_i), i = 1, \dots, n \text{ independent}$$

$$\begin{aligned} \text{Log}(E(Y_i)) &= \text{Log}(\mu_i) \\ &= \text{Log}(N_i \lambda_i) \\ &= \text{Log}(N_i) + \text{Log}(\lambda_i) \\ &= \text{Log}(N_i) + \beta_0 + \beta_1 \text{male}_i \end{aligned}$$

- for patient  $i$ , the expected number of hypoglycemic episodes is  $N_i \lambda_i$  where  $N_i$  is the total follow-up time in days for patient  $i$  and  $\lambda_i$  is the risk of a hypoglycemic episode per unit time / per day.
- $\beta_0$  is the logarithm of the risk of a hypoglycemic episode in a day for females.
- $\beta_0 + \beta_1$  is the logarithm of the risk of a hypoglycemic episode in a day for males.
- $\exp(\beta_1)$  is the relative risk of a hypoglycemic episode in a day comparing males to females OR the relative expected number of hypoglycemic episodes comparing males and females who have the same duration of follow-up.



## Example: Variable exposure time

$$\log(E(\underline{Y_i})) = \log(\mu_i) = \log(\underline{N_i} \lambda_i) = \log(\underline{N_i}) + \underbrace{\beta_0 + \beta_1 \text{male}_i}_{\text{offset}}$$

```
##              Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -0.2752677 0.03603750 -7.638368 2.199923e-14
## male         0.1142061 0.05012278  2.278527 2.269520e-02
```

```
expected.Y = fit$fitted
predicted.lambda = exp(fit$coefficients[1] + male*fit$coefficients[2])
head(cbind(N,Y,male,expected.Y,predicted.lambda))
```

##	N	Y	male	expected.Y	predicted.lambda
## 1	17	19	1	14.47107 = $\hat{\mu}_i$	0.8512397 = $\hat{\lambda}_i$
## 2	22	18	0	16.70611 $\hat{\mu}_i = E(Y_i)$	0.7593688
## 3	19	16	1	16.17355	0.8512397
## 4	19	15	1	16.17355	0.8512397
## 5	22	13	0	16.70611	0.7593688
## 6	25	18	1	21.28099	0.8512397

$$\hat{\mu}_i = N_i \cdot \hat{\lambda}_i = 22 \cdot .76$$

## Example: Variable exposure time

$$\log(E(Y_i)) = \log(\mu_i) = \log(N_i \lambda_i) = \log(N_i) + \beta_0 + \beta_1 \text{male}_i \quad \log(\lambda_i)$$

##	Estimate	Std. Error	z value	Pr(> z )
## (Intercept)	-0.2752677	0.03603750	-7.638368	2.199923e-14
## male	0.1142061	0.05012278	2.278527	2.269520e-02

- Interpret  $\beta_0$  =  $\log(\lambda_i)$  for females  
= log of the risk of a hypoglycemic episode in a day for females  
 $\exp(-0.275) = \text{daily risk of an episode}$
- Interpret  $\beta_1$   
=  $\log(\lambda_i | \text{male}) - \log(\lambda_i | \text{female})$   
=  $\log\left[\frac{\lambda_i | \text{male}}{\lambda_i | \text{female}}\right] = \log(1.1)$   
 $\exp(1.1) \approx 1.12$  fixing time



# Estimation: Maximum likelihood estimation

The likelihood function is:

$$L(\beta|Y) = \prod_{i=1}^n \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}$$

The log-likelihood is:

$$\log L(\beta|Y) = \sum_{i=1}^n (-\mu_i) + y_i \log(\mu_i) - \log(y_i!)$$

The score equation is:

$$\frac{\partial \log L(\beta|Y)}{\partial \beta} = \sum_{i=1}^n \left( -\frac{\partial \mu_i}{\partial \beta} \right) + y_i \frac{\partial \log(\mu_i)}{\partial \beta}$$

$$= \sum_{i=1}^n (-\mu_i X_i') + y_i X_i'$$

$$= \sum_{i=1}^n X_i' (y_i - \mu_i)$$

$\mu_i(\beta)$

$$\hat{\beta} \sim N(\hat{\beta}, (X' \text{diag}(\hat{\mu}) X)^{-1})$$



# Robust variance estimation

Count data is almost always over-dispersed, i.e.  $Var(Y_i) > E(Y_i)$ .

Solution: Assume  $E(Y_i|X_i) = \mu_i = N_i e^{X_i^T \beta}$  and  $Var(Y_i|X_i) = \mu_i \phi$ .

We can estimate  $\phi$  by:

$$\hat{\phi} = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \bigg/ (n - p)$$

which is the Pearson residual estimate of  $\phi$ .

Alternatively, you can use the deviance estimator as:

$$\hat{\phi} = 2 \sum_{i=1}^n [Y_i \log(Y_i / \mu_i) - (Y_i - \mu_i)] \bigg/ (n - p)$$

Either is fine for computing the robust variance estimate.



# Example: Robust variance estimation

- ▶ Daily non-accidental deaths in Chicago, 1987 – 1994
- ▶ Log-linear model for daily deaths as a function of:
  - ▶ PM10
  - ▶ Current temperature + average of prior three days (natural spline 3 df)
  - ▶ Time: year, season, month
- ▶ Data are overdispersed; greater variance than expected by Poisson model



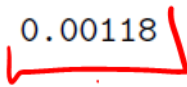


## Example: Robust variance estimation

```
fit.poisson.year = glm(total~ pm10+ns(temp,3)+ns(avgtemp,3)+as.factor(year),  
  data=data,family="poisson")
```

```
fit.robust.year = glm(total~ pm10+ns(temp,3)+ns(avgtemp,3)+as.factor(year),  
  data=data,family="quasipoisson")
```

##	Poisson beta	<u>Poisson SE</u>	Robust beta	<u>Robust SE</u>
## 1	0.00349	0.00104	0.00349	0.00116
## 2	0.00229	0.00107	0.00229	0.00117
## 3	0.00178	<u>0.00111</u>	0.00178	0.00118



expected Robust SE > Poisson SE



# Case Study

- ▶ Estimation of excess deaths after Hurricane Maria

