JOHNS HOPKINS
BLOOMBERG SCHOOL of PUBLIC HEALTH

Quiz 2
PS2
PS3 is posted

Lecture 11

Lecture 10 Handout

Log-linear regression
Examples plus

Lecture 11 Handout

Case study of excess deaths due to Hurricane Maria

Review of Lecture 10

X_i: 0 1

Y_i: 0 1

- Use of "marginal" and "conditional" to describe logistic models
  - Lecture 4: $Y_i = \{0, 1$ $i = 1, \ldots, n$ independent
    - Marginal model: here we were correlating Y (binary) with a single X (binary), i.e. evaluating the unadjusted relationship $\text{Logit}\{Pr(Y_i = 1 \mid X_i)\} = \beta_0^m + \beta_1^m X_i$
    - 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)

      $\beta_0^c + \beta_1^c X_i + \beta_2^c C$
  - 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).

Y_{ij} [$i = 1, \ldots, m$ individuals; $j = 1, \ldots, n_i$ assessments

$\beta_0^c + \beta_1^c X_{i1} + \boxed{b_i}$

[$i = 1, \ldots, m$ clusters; $j = 1, \ldots, n_i$ individuals within 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 for a given year
- 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 or population at risk

- Imagine a range of time [0,T] and breaking that range of time into small bins [t, t+dt]

- Pr(Event occurs in [t,t+dt]) = λ dt
  ↳ intensity

- Pr(2 or more events occur in [t, t+dt]) ~ 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

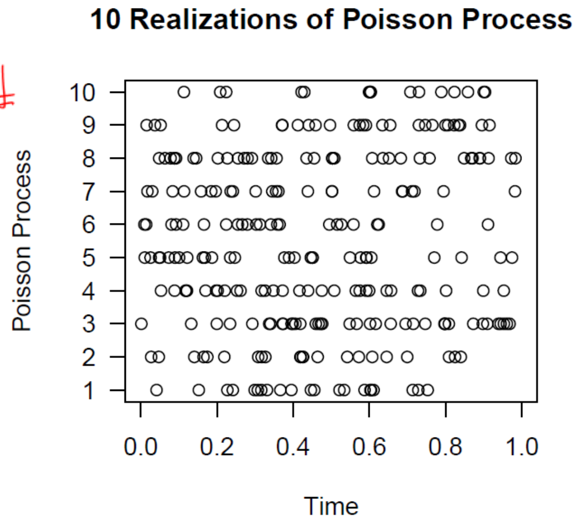
- The number of events X occurring in the interval [0,T] follows a Poisson distribution
- Probability mass function: $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$, $x = 0, 1, 2, \ldots$

See page 3 of Lecture 10 handout for derivation.

- The mean and variance of X is λ T

Expected # of events in [0, T]

$Var(X) = E(X)$



$E(X) = 25$

Log-linear model

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

- General form:

$$Y_i \sim P(\mu_i), i = 1, \dots, n \; independent$$

$\longrightarrow E(Y_i) = \mu_i \qquad Var(Y_i) = \mu_i$

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

link function: log inverse link: exp

- Interpretation:

β_0 = log of expected # of events when $X_1, \dots, X_p = 0$

β_1 = Difference in the log expected # of events comparing $X_1 = x$ vs $X_1 = x-1$, with all other Xs held fixed.

$$\beta_1 = \log\left[\frac{E(Y_i \mid X_1 = x, X_2, \dots, X_p)}{E(Y_i \mid X_1 = x-1, X_2, \dots, X_p)}\right]$$

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

$$Log(E(Y_i)) = Log(\mu_i) = \beta_0 + \beta_1 male_i$$

```
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
```

log link

```
##             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
```

β_0

- $\hat{\beta}_0$ 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

$$Log(E(Y_i)) = Log(\mu_i) = \beta_0 + \beta_1 male_i$$

```
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
```

```
##              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
```

$\beta_1 = \log[E(Y_i \mid male=1)] - \log[E(Y_i \mid male=0)]$

$= \log\left[\frac{E(Y_i \mid male=1)}{E(Y_i \mid male=0)}\right]$

- $\boxed{\hat{\beta}_1}$ 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.
- $exp(\hat{\beta}_1) = exp(0.20) = \boxed{1.22}$ 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.

$15.3/12.4 = 1.22$

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

Y_i = # of episodes N_i = days of follow-up

$E(Y_i) = \mu_i = N_i \boxed{\lambda_i}$

λ_i: risk of an episode per unit time

N_i: fixed and known

Example: Variable exposure time

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

$$\begin{aligned}
Log(E(Y_i)) &= Log(\mu_i) \\
&= Log(N_i \lambda_i) \\
&= \underbrace{Log(N_i)}_{\text{constant}} + Log(\lambda_i) \\
&= \underbrace{Log(N_i)}_{\text{offset}} + \beta_0 + \beta_1 male_i
\end{aligned}$$

- 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 λ_i is the risk of a hypoglycemic episode per unit time / per day.
- β_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

$$\log\big(E(Y_i)\big) = \log(\mu_i) = \log(N_i\lambda_i) = \log(N_i) + \beta_0 + \beta_1 male_i$$

```
##               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
```

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

$\hat{\lambda}_i$

```
##    N  Y male expected.Y predicted.lambda
## 1 17 19    1   14.47107        0.8512397
## 2 22 18    0   16.70611        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
```

glm (Y ~ male,
family = "poisson",
offset = log(N))

$E(Y_i) = E(N_i \hat{\lambda}_i)$

Example: Variable exposure time

$$\log(E(Y_i)) = \log(\mu_i) = \log(N_i\lambda_i) = \log(N_i) + \beta_0 + \beta_1 male_i$$

```
##               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 β_0

log risk of a hypoglycemic episode per day among females

$\exp(-.275) = \hat{\lambda}_{i,\text{females}}$

- Interpret β_1

$$\log\left[\frac{\lambda_i \mid male=1}{\lambda_i \mid male=0}\right] = \log\left[\frac{N\lambda_i \mid male=1}{N\lambda_i \mid male=0}\right]$$

$\exp(.11) \hat{=} 1.12$

the risk of a hypoglycemic episode for males is 12% greater than that for females

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:

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

LRT

wald

The score equation is:

$$\begin{aligned}
\frac{\partial logL(\beta|Y)}{\partial\beta} &= \sum_{i=1}^{n}\left(-\frac{\partial\mu_i}{\partial\beta}\right) + y_i\frac{\partial log(\mu_i)}{\partial\beta} \\
&= \sum_{i=1}^{n}(-\mu_i X_i') + y_i X_i' \\
&= \sum_{i=1}^{n} X_i'(y_i - \mu_i)
\end{aligned}$$

$\mu_i(\beta)$

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

Robust variance estimation

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

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

We can estimate ϕ by:

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

which is the Pearson residual estimate of ϕ.

Alternatively, you can use the deviance estimator as:

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

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")
```

$\Rightarrow E(Y_i) = Var(Y_i)$

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

estimation of β is the same: $Var(Y_i) = \phi E(Y_i)$

```
##    Poisson beta Poisson SE Robust beta Robust SE
## 1       0.00349    0.00104     0.00349   0.00116
## 2       0.00229    0.00107     0.00229   0.00117
## 3       0.00178    0.00111     0.00178   0.00118
```

$E(Y_i) = Var(Y_i)$

quasiPoisson

$Var(Y_i) = E(Y_i) \boxed{\phi}$

Case Study

- Estimation of excess deaths after Hurricane Maria