Lecture 10

Models for longitudinal / clustered binary responses
Introduction to count outcomes and log-linear regression

Lecture 9 Review:

- Marginal Models for clustered binary outcomes
  - Define the distribution of Y

$$Y_{ij} \sim Bernoulli(\mu_{ij})$$
$$E(Y_{ij}) = \Pr(Y_{ij} = 1) = \mu_{ij}$$
$$Var(Y_{ij}) = \mu_{ij}(1 - \mu_{ij})$$
$$\text{Corr}(Y_{ij}, Y_{ik}) = f(\alpha, j, k)$$

  - Linear model: Identity link

$$\mu_{ij} = X'_{ij}\beta$$

  - Logistic model: logit link

$$\text{logit}(\mu_{ij}) = X'_{ij}\beta$$

- Estimation uses generalized estimating equations
- Add robust variance estimate to protect against misspecification of variance of the outcome and correlation model

Lecture 9 Review:

- Marginal Model: OR = 1.72

	Time = 0	**Time = 1**
Y = 0	61	46
Y = 1	50	65

- Conditional Model: OR = 10.89

	Y1 = 0	**Y1 = 1**
Y0 = 0	39	22
Y0 = 1	7	43

Two example studies

- Placebo-controlled trial to improve respiratory function
  - 111 patients
  - Baseline + 4 follow-ups
  - Compare the change in odds from baseline to follow-up across the active treatment vs. placebo groups.

- Matched case-control study looking at effect of exogenous estrogens on the risk of endometrial cancer
  - 63 matched sets: one case + 4 controls
  - Alive in same community at the time of diagnosis for the case, age within 1 year, same marital status and entered community at roughly the same time
  - Do women who use estrogens, have a history of gall-bladder disease or hypertension at increased risk of endometrial cancer?

Conditional Models

- Random effects logistic regression model:

Conditional Models

$$
\begin{aligned}
logit[Pr(Y_{ij} = 1 | post_{ij}, trtmnt01_i, b_i)] &= \beta_{0i}^c + \beta_1^c I(post_{ij} > 0) + \beta_2^c I(post_{ij} > 0) trtmnt01_i \\
&= \beta_0^c + b_i + \beta_1^c I(post_{ij} > 0) + \beta_2^c I(post_{ij} > 0) trtmnt01_i
\end{aligned}
$$

where $b_i \sim N(0, \sigma^2)$ and the covariates are independent of b_i.

Interpretation:

- β_{0i}^c: defines a patient specific log-odds of a good respiratory response at baseline
- $\beta_{0i}^c = \beta_0^c + b_i$, where $b_i \sim N(0, \sigma^2)$: β_0^c is the log-odds of a good respiratory response for the average patient (i.e. $b_i = 0$)
- $\beta_{0i}^c = \beta_0^c + b_i$, where $b_i \sim N(0, \sigma^2)$: b_i represents the deviation from this average log-odds of a good respiratory response for patient i

Example: Logistic regression with random intercept

$$
\begin{aligned}
logit[Pr(Y_{ij} = 1 | post_{ij}, trtmnt01_i, b_i)] &= \beta_{0i}^c + \beta_1^c I(post_{ij} > 0) + \beta_2^c I(post_{ij} > 0) trtmnt01_i \\
&= \beta_0^c + b_i + \beta_1^c I(post_{ij} > 0) + \beta_2^c I(post_{ij} > 0) trtmnt01_i
\end{aligned}
$$

where $b_i \sim N(0, \sigma^2)$ and the covariates are independent of b_i.

$$
\mu_{ij}^c = \frac{exp(\beta_0^c + b_i + \beta_1^c I(post_{ij} > 0) + \beta_2^c I(post_{ij} > 0) trtmnt01_i)}{1 + exp(\beta_0^c + b_i + \beta_1^c I(post_{ij} > 0) + \beta_2^c I(post_{ij} > 0) trtmnt01_i)}
$$

- Slopes are log [ratio of individual odds]!

Example: Random intercept logistic model in R using glmer

```
ri.fit = glmer(r~post + postXtrt+(1|id),data=data,family="binomial",nAGQ=7)
summary(ri.fit)
```

```
## Random effects:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 6.49     2.55
## Number of obs: 555, groups:  id, 111
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.4212     0.3667   -1.15     0.25
## post         -0.0834     0.3683   -0.23     0.82
## postXtrt      1.9452     0.4850    4.01  6.1e-05 ***
```

- Intercept: For the average or typical patient (i.e. $b_i = 0$), the probability of a good response is $\frac{\exp(-0.42)}{1+\exp(-0.42)} = 0.40$

- You can compute baseline probability of a good response for any patient by: $\frac{\exp(-0.42+b_i)}{1+\exp(-0.42+b_i)}$

Example: Interpretation

```
ri.fit = glmer(r~post + postXtrt+(1|id),data=data,family="binomial",nAGQ=7)
summary(ri.fit)
```

```
## Random effects:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 6.49     2.55
## Number of obs: 555, groups:  id, 111
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.4212     0.3667   -1.15     0.25
## post         -0.0834     0.3683   -0.23     0.82
## postXtrt      1.9452     0.4850    4.01  6.1e-05 ***
```

Example: Interpretation

```
ri.fit = glmer(r~post + postXtrt+(1|id),data=data,family="binomial",nAGQ=7)
summary(ri.fit)
```

```
## Random effects:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 6.49     2.55
## Number of obs: 555, groups:  id, 111
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.4212     0.3667   -1.15     0.25
## post         -0.0834     0.3683   -0.23     0.82
## postXtrt      1.9452     0.4850    4.01  6.1e-05 ***
```

Comparison of marginal and conditional slope terms

Compare the marginal (β) and conditional (β^c) parameter estimates.

```
cbind(summary(fit.exch)$coeff[,1],summary(ri.fit)$coeff[,1])
```

```
##                [,1]     [,2]
## (Intercept) -0.1989 -0.42120
## post        -0.0410 -0.08343
## postXtrt     1.0083  1.94525
```

- Recall our discussion of confounding: Assume b_i is independent of covariates (as we do in random effects models)

Marginal model: $logit[Pr(Y_{ij}|X_{ij})] = \beta_0 + \beta_1 X_{ij}$

Conditional model: $logit[(Pr(Y_ij|X_{ij}, b_i))] = \beta_0^c + \beta_1^c X_{ij} + b_i$

In general:

- β = change in log population odds per unit change in X
- β^c = change in cluster-specific log odds per unit change in X

Estimation: Random effects logistic regression model

- Basic Idea:

$$f(Y\,|\beta^c, D) = \int f(Y, b_i|\beta^c, D)db_i = \int f(Y\,|b_i, \beta^c, D)f(b_i|D)db_i$$

$$L(y|\beta^c, D) = \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} (\mu_{ij}^c(\beta^c, b_i))^{y_{ij}} (1 - \mu_{ij}^c(\beta^c, b_i))^{1-y_{ij}} f(d_i|D)db_i$$

$$= \prod_{i=1}^{m} \int Pr(y_{i1}, ..., y_{in_i}|\beta^c, b_i)Pr(b_i|D)db_i$$

It can be shown that:

$$\frac{\partial log(L(y|\beta^c, D))}{\partial \beta^c} = \sum_{i=1}^{m}\sum_{j=1}^{n_i} X_{ij}^{'}(y_{ij} - E_{b_i|y}(\mu_{ij}^c(b_i, \beta^c)))$$

Estimation: Random effects logistic regression model

$$L(y|\beta^c, D) = \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} (\mu_{ij}^c(\beta^c, b_i))^{y_{ij}} (1 - \mu_{ij}^c(\beta^c, b_i))^{1-y_{ij}} f(d_i|D) db_i$$

$$= \prod_{i=1}^{m} \int Pr(y_{i1}, ..., y_{in_i}|\beta^c, b_i) Pr(b_i|D) db_i$$

- Solving the likelihood function requires estimation of the integral
- This is typically estimated via numerical methods
- Gaussian quadrature
- Adaptive gaussian quadrature
- Requires a number of quadrature points: nAGQ = 7

Special case: Matched Case-Control Study

- Consider the likelihood function for the logistic regression model with random intercept

$$logit[\mu_{ij}^c] = X_{ij}^{'}\beta^c + b_i \quad b_i \sim N(0, \sigma^2).$$

$$L(y|\beta^c, \sigma^2) = \prod_{i=1}^{m} \int \frac{exp\left[(\sum_{j=1}^{n_i} y_{ij} X_{ij})' \beta^c + y_i^+ b_i \right]}{\prod_{j=1}^{n_i} (1 + exp(X_{ij}^{'}\beta^c + b_i))} f(b_i|\sigma_2) db_i$$

$y_i^+ = \sum_{j=1}^{n_i} y_{ij}$ is sufficient for b_i, i.e. $Pr(y_{ij}|y_i^+, b_i)$ does not depend on b_i

Special case: Matched Case-Control Study

- Data:
  - Case: $Y_{i1} = 1, X_{i1}$
  - Control: $Y_{io} = 0, X_{i0}$

- Model:

$$Pr(Y_{ij} = 1|X_{ij}, b_i) = \frac{exp(X_{ij}^{'}\beta^c + b_i)}{1 + exp(X_{ij}^{'}\beta^c + b_i)}$$

- Goal is to estimate parameters for X without making assumptions about distribution of b

$$CL(Y_i|\beta^c) = \prod_{i=1}^{m} [Pr(Y_{i0} = 0|X_{i0}, y_i^+ = 1)Pr(Y_{i1} = 1|X_{i1}, y_i^+ = 1)]$$

Special case: Matched Case-Control Study

$$
\begin{aligned}
Pr(Y_{i1}=1|X_{i1},Y_i^+=1,b_i) &= \frac{Pr(Y_{i1}=1 \text{ and } Y_i^+=1|b_i)}{Pr(Y_i^+=1|b_i)} \\
&= \frac{Pr(Y_{i1}=1 \text{ and } Y_{i0}=0|b_i)}{Pr(Y_{i1}=1 \text{ and } Y_{i0}=0|b_i)+Pr(Y_{i1}=0 \text{ and } Y_{i0}=1|b_i)} \\
&= \frac{Pr(Y_{i1}=1|b_i)\times Pr(Y_{i0}=0|b_i)}{Pr(Y_{i1}=1|b_i)\times Pr(Y_{i0}=0|b_i)+Pr(Y_{i1}=0|b_i)\times Pr(Y_{i0}=1|b_i)} \\
&= \frac{\left(\frac{exp(X_{i1}\beta^c+b_i)}{1+exp(X_{i1}\beta^c+b_i)}\times\frac{1}{1+exp(X_{i0}\beta^c+b_i)}\right)}{\frac{exp(X_{i1}\beta^c+b_i)}{1+exp(X_{i1}\beta^c+b_i)}\times\frac{1}{1+exp(X_{i0}\beta^c+b_i)}+\frac{1}{1+exp(X_{i1}\beta^c+b_i)}\times\frac{exp(X_{i0}\beta^c+b_i)}{1+exp(X_{i0}\beta^c+b_i)}} \\
&= \frac{exp(X_{i1}\beta^c+b_i)}{exp(X_{i1}\beta^c+b_i)+exp(X_{i0}\beta^c+b_i)} \\
&= \frac{exp(X_{i1}\beta^c)}{exp(X_{i1}\beta^c)+exp(X_{i0}\beta^c)} \\
&= \frac{exp((X_{i1}-X_{i0})\beta^c)}{1+exp((X_{i1}-X_{i0})\beta^c)}
\end{aligned}
$$

Special case: Matched Case-Control Study

$$CL(Y|\beta^c) = \prod_{i=1}^{m} \left[\frac{exp((X_{i1} - X_{i0})\beta^c)}{1 + exp((X_{i1} - X_{i0})\beta^c)} \right]^1$$

- Marginal logistic regression:
  - No intercept
  - Responses Y = 1
  - Covariates:

$$(X_{11} - X_{10}, X_{21} - X_{20}, ..., X_{m1} - X_{m0})$$

Example: Endometrial cancer example

Matched case-control study looking at effect of exogenous estrogens on the risk of endometrial cancer

- 63 matched sets: one case + 4 controls
- Alive in same community at the time of diagnosis for the case, age within 1 year, same marital status and entered community at roughly the same time
- Do women who use estrogens, have a history of gall-bladder disease or hypertension at increased risk of endometrial cancer?

- I will do the analysis using the case + 1 matched control
  - You will revisit this data in Problem Set 3 using all the participants.

Example: Endometrial cancer example

```
dat = read.table("./endometrial.txt")
names(dat) = c("set","case","age","ageg","est","gall","hyp","obesity","nonestdrug")
dat$est = dat$est - 1
dat$gall = dat$gall - 1
dat$hyp = dat$hyp - 1
dat$obesity[dat$obesity==3] = NA
dat$obesity = dat$obesity - 1
dat$nonestdrug = dat$nonestdrug - 1
dat$firstctrl = unlist(tapply(dat$set,dat$set,FUN=function(x) c(0,1,rep(0,length(x)-2))))

tapply(dat$est,dat$case,mean)
```

```
##      0      1
## 0.5040 0.8889
```

```
tapply(dat$gall,dat$case,mean)
```

```
##       0       1
## 0.09524 0.26984
```

```
tapply(dat$hyp,dat$case,mean)
```

```
##      0      1
## 0.3254 0.4127
```

Example: Endometrial cancer example

```
library(survival)
```

```
## Warning: package 'survival' was built under R version 3.6.3
```

```
## Fit the conditional logistic model with
## all three exposures using only 1st control
fit1=clogit(case~est+gall+hyp+ strata(set), data=subset(dat,case==1|firstctrl==1))
summary(fit1)$coeff
```

```
##            coef exp(coef)  se(coef)          z     Pr(>|z|)
## est   2.2479841 9.4686292 0.6255817  3.5934304 0.0003263528
## gall  0.6907726 1.9952565 0.6157373  1.1218625 0.2619209223
## hyp  -0.1333443 0.8751637 0.4455392 -0.2992874 0.7647207469
```

```
## Drop hypertension from the model
fit1=clogit(case~est+gall+strata(set),
            data=subset(dat,case==1|firstctrl==1))
summary(fit1)$coeff
```

```
##          coef exp(coef)  se(coef)        z     Pr(>|z|)
## est  2.209052  9.107077 0.6097099 3.623120 0.0002910712
## gall 0.694732  2.003172 0.6156339 1.128482 0.2591162174
```

Example: Endometrial cancer example

```
## Add the interactions
fit1.int=clogit(case~est*gall+strata(set),
            data=subset(dat,case==1|firstctrl==1))
summary(fit1.int)$coeff
```

```
##                coef  exp(coef)  se(coef)         z     Pr(>|z|)
## est       2.671060 14.4552809 0.7533387  3.545629 0.0003916766
## gall      2.292397  9.8986370 1.2224136  1.875304 0.0607509226
## est:gall -2.141460  0.1174832 1.3700403 -1.563064 0.1180376313
```

```
# Compute the synergistic effect
coeff.sum = sum(fit1.int$coefficients)
var.sum = t(c(1,1,1)) %*% vcov(fit1.int) %*% c(1,1,1)
exp(coeff.sum)
```

```
## [1] 16.81038
```

```
exp(coeff.sum-1.96*sqrt(var.sum))
```

```
##          [,1]
## [1,] 2.855665
```

```
exp(coeff.sum+1.96*sqrt(var.sum))
```

```
##          [,1]
## [1,] 98.95735
```

Example: Endometrial cancer example

- In summary, both estrogen use and history of gall bladder disease were found to increase the risk of endometrial cancer. Furthermore, these risk factors were found to be non-additive. That is, on the log odds scale, the risk associated with having both risk factors is only marginally greater than the risk associated with having a single risk factor. However, on the odds scale this translates to a substantive increase in risk. One way to interpret the findings is below.
  - The estimated odds of being a case for subjects with only estrogren use are 14.5 (95% CI: 3.1 to 71.4) times the odds of being a case for subjects with neither estrogen use or history of gallbladder disease.
  - The estimated odds of being a case for subjects with only a history of gall bladder disease are 9.9 (95%CI: 0.95 to 104.8) times the odds of being a case for subjects with neither estrogen use or history of gallbladder disease.
  - Finally, the estimated odds of being a case for subjects with both estrogen use and gall bladder disease are 16.8 (95% CI: 2.9 to 99.0) times the odds of being a case for subjects with neither estogren use or history of gallbladder disease. This is approximately double the odds ratio from either risk factor alone.

Log-linear models for count variables

- Count variable
  - Takes on values of non-negative integers
  - 0, 1, 2, …, 3321, ….. 10001, ….

- Examples
  - 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

- Notice anything? Counts of things occurring within a given time range or group of eligible persons

Log-linear models for count variables

- Characteristics of count variables
  - Non-negative integers
  - Variability tends to increase as mean increases
  - Effects of predictors tend to be multiplicative (reflecting relative changes not absolute change)

EXAMPLE: Numbers of Non-accidental Death per Day in Chicago, 1987-1994

Season	Mean	Variance	Variance/Mean
Winter (Dec-Feb)	122	177.6	1.45
Summer (June-Aug)	107	128.4	1.20

Poisson process

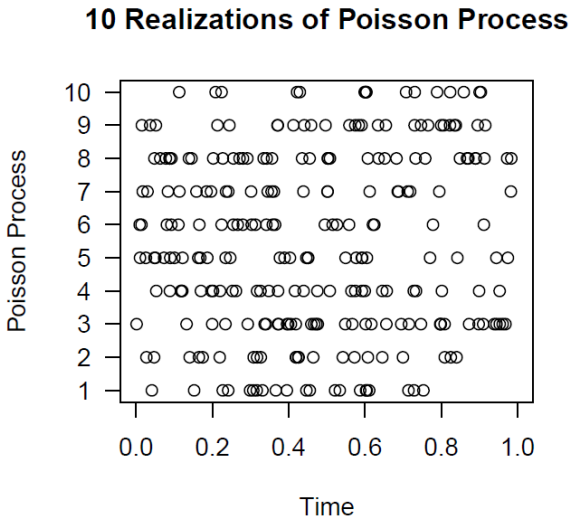
- Poisson process defines how observations of events of interest occur over time or space
- Imagine a range of time [0,T] and breaking that range of time into small bins [t, t+dt]
- Pr(Event occurs in [t,t+dt]) = λ dt
- Pr(2 or more events occur in [t, t+dt]) ~ 0
- Memoryless property: chance of an event 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: $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$

See page 3 of Lecture 10 handout for derivation.

- The mean and variance of X is λ T



Log-linear model

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

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

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

- Interpretation:

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
```

```
##               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

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

```
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
```

- $\hat{\beta}_1$ 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.

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

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

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

- 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))
```

```
##    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
```

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
```

- Interpret β_0

- Interpret β_1

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:

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

The score equation is:

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

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

Robust variance estimation

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

Solution: Assume $E(Y_i|X_i) = \mu_i = N_i e^{X_i^{!}\beta}$ 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:

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

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

```
##   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
```

Next time….

- Estimation of excess deaths after Hurricane Maria