JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Lecture 9

Review of logistic regression model assumptions
Models for $\underbrace{\text{longitudinal / clustered binary}}$ responses

violation independence

Review of logistic regression assumptions

- And solutions to violations
- Mean model is correctly specified
  - Violation impact estimation of association parameters
  - Plot average predicted vs. observed proportions within quintiles or deciles of predicted values
  - Plot average predicted vs. observed proportions as a function continuous exposure
  - Summary tables of average predicted vs. observed proportions by level of categorical exposure
  - SOLUTION: change your mean model

β = logOR

$\hat{p}_{obs}$

$\hat{P}_{model}$

- Observations are independent

inferences

  - Violation impacts estimation of standard errors, confidence intervals, hypothesis tests
  - SOLUTIONS:
    - Marginal logistic regression model fit using generalized estimating equations
    - Conditional logistic regression model

Review of logistic regression assumptions

- Variance is correctly specified
  - Logistic model assumes: Var(Y) = p(1-p)
  - Under or over-dispersion
  - Compute Var(Y) and compare with predicted variance, overall or by select variables
  - SOLUTION:
    - Bootstrap
    - GLM: family = "quasibinomial" assumes Var(Y) = ϕ x p x (1-p) where ϕ = 1/(n-k) sum of squared Pearson residuals

$Y_i \sim \text{Bernoulli}(\mu_i)$

$E(Y_i) = \mu_i = Pr(Y_i = 1)$

$Var(Y_i) = \mu_i(1-\mu_i)$

Observed variance in Y > Bernoulli

over dispersion

- There are no "influencial" observations
  - DFFITS or DFBETAS

Two example studies

Datasets online library

resperatory

- 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?

endometrial

Two approaches to modeling

- Marginal models: Goal is to focus on comparing subsets of the population while accounting for the correlation in responses to make proper inferences

$Y_{ij} = \begin{matrix} 1 \\ 0 \end{matrix}$

$i =$ cluster $j =$ individual

$X_{ij} = \begin{matrix} 0 \\ 1 \end{matrix}$ exposure

$$\log\left[\frac{Pr(Y_{ij}=1 \mid X_{ij}=1)}{Pr(Y_{ij}=0 \mid X_{ij}=1)}\right] - \log\left[\frac{Pr(Y_{ij}=1 \mid X_{ij}=0)}{Pr(Y_{ij}=0 \mid X_{ij}=0)}\right]$$

correlation → nuisance

- Conditional models

Compare odds of an event occurring across exposure groups among individuals within the same cluster

$$\log\left[\frac{Pr(Y_{ij}=1 \mid X_{ij}=1, b_i)}{Pr(Y_{ij}=0 \mid X_{ij}=1, b_i)}\right] - \log\left[\frac{Pr(Y_{ij}=1 \mid X_{ij}=0, b_i)}{Pr(Y_{ij}=0 \mid X_{ij}=0, b_i)}\right]$$

Longitudinal design

$$Y_{ij} = \begin{cases} 0 \\ 1 \end{cases} \quad \begin{array}{l} i = \text{individual}, 1, \ldots, m \\ j = \#\text{ of assessments } 1, \ldots, n_i \end{array}$$

$$Y_{ij} \sim \text{Bernoulli}(\mu_{ij}) \rightarrow Var(Y_{ij}) = \mu_{ij}(1-\mu_{ij})$$

$$g(\mu_{ij}) = X_{ij}'\beta \rightarrow \underbrace{Corr(Y_{ij}, Y_{ik}) = f(\alpha, j, k)}$$

$$Var(Y_i) = \begin{bmatrix} \mu_{i1}(1-\mu_{i1}) & & * \\ & \ddots & \\ * & & \mu_{in_i}(1-\mu_{in_i}) \end{bmatrix} \leftarrow Cov(Y_{ij}, Y_{ik})$$

$\hookrightarrow n_i \times 1$ vector

$n_i \times n_i$ matrix

$$Var(Y) = \begin{bmatrix} Var(Y_1) & & 0 \\ & \ddots & \\ 0 & & Var(Y_m) \end{bmatrix}$$

$\sum_{i=1}^{m} n_i \times 1$

Marginal model: GLM Review

- Requires specification of 3 components

$$Y_i = \begin{cases} 0 \\ 1 \end{cases} \quad i = 1, \dots, n \quad \text{independent}$$

1) Distribution $\quad Y_i \sim \text{Bernoulli}(\mu_i), \; Var(Y_i) = \mu_i(1-\mu_i)$

2) link function $\quad g(\mu_i) = X_i'\beta \qquad g^{-1}(X_i'\beta) = \mu_i$

3) linear predictor

$$\underbrace{Var(Y)}_{\substack{n\times 1 \\ \text{vector}}} = \underbrace{\begin{bmatrix} \mu_1(1-\mu_1) & & 0 \\ & \ddots & \\ 0 & & \mu_n(1-\mu_n) \end{bmatrix}}_{n\times n}$$

Marginal model: GLM review

- For a logistic regression model, we derived the likelihood function, log likelihood function and score equations.

- Recall the score equation:

$$\begin{aligned} U(\beta) &= X'(Y - \mu(\beta)) \\ &= \left(\frac{\partial \mu}{\partial \beta}\right)' V^{-1}(Y - \mu(\beta)) \\ &= \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta}\right)' V_i^{-1}(Y_i - \mu_i(\beta)) \end{aligned}$$

where $\frac{\partial \mu}{\partial \beta} = VX$, $V = diag\left[\mu(\beta)(1 - \mu(\beta))\right]$, $V_i = \mu_i(\beta)(1 - \mu_i(\beta))$.

Marginal Model: Longitudinal GLM

- You need to include one additional element in the model specification

Marginal Model: Longitudinal GLM

- In linear models, we could easily write out the joint distribution for Y_i , the vector of responses for cluster i

$$Y_i \sim MVN(\mu_i, V_i)$$

- In general, it is hard to write out the joint distribution of a Bernoulli random variable, Poisson random variable, etc.
- We don't use maximum likelihood estimation here
- Derive estimates of β using multivariate version of the score equation (estimating equation)

Marginal Model: Generalized Estimating Equations

Zeger
Liang

- Estimation procedure is called generalized estimating equations (GEE)
- Weighted least squares when Y_i is multivariate normal is a special case.
- GEE:

$$\sum_{i=1}^{m} \left[\frac{\partial \mu_i}{\partial \beta}\right]' V_i^{-1}(Y_i - \mu_i(\beta)) = 0$$

Y_i = vector
V_i = matrix

- GLM:

$$\underbrace{\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta}\right)' V_i^{-1}(Y_i - \mu_i(\beta))}$$

Y_i = scalar
V_i = scalar

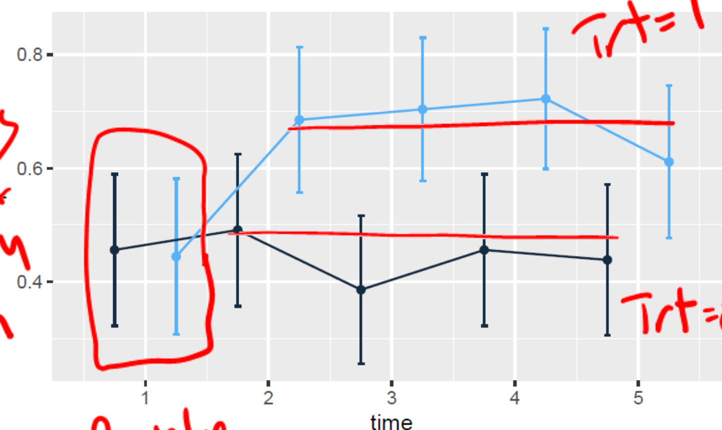
Example: Exploratory data analysis, mean model

- Placebo-controlled trial of respiratory function
- Baseline (time 1) and 4 follow-ups (times 2 through 5)
- Treatment is assigned after baseline respiratory function is recorded

$Pr(Y_{ij} = 1 \mid time = j)$

	1	2	3	4	5
0	0.4561404	0.4912281	0.3859649	0.4561404	0.4385965
1	0.4444444	0.6851852	0.7037037	0.7222222	0.6111111

Probability of good respiratory function



Trt=1

Trt=0

Baseline

Example: Mean model specification and interpretation

- Model specification:

$$logit[Pr(Y_{ij}=1|\underbrace{post_{ij}}, \underbrace{trtmnt01_i})] = \boxed{\beta_0} + \underbrace{\beta_1 post_{ij}} + \underbrace{\beta_2 post_{ij} \times trtmnt01_i}$$

- β_0: log odds of a good respiratory response at baseline
- β_1: log odds ratio of a good respiratory response comparing follow-up to baseline among patients receiving the placebo
- $\beta_1 + \beta_2$: log odds ratio of a good respiratory response comparing follow-up to baseline among patients receiving the active treatment
- β_2: treatment effect! Does the relative improvement in the odds of a good response comparing follow-up to baseline differ for the patients receiving active treatment vs. placebo

Example: Exploratory data analysis, correlation structure

- How do we assess the degree of correlation in the data?

- Linear models:
  - Pairwise correlation coefficients between each follow-up time
  - Autocorrelation function, $Corr(Y_{ij}, Y_{ik}) = f(\alpha, j, k)$ → bounded between -1, 1
  - Use the above to propose a model for the correlation structure

- Logistic models:
  - $Corr(Y_{ij}, Y_{ik}) = f(\alpha, \mu_{ij}, \mu_{ik})$ and is constrained by μ_{ij}, μ_{ik}
  - Alternative to the correlation, we can measure association over time using odds ratios

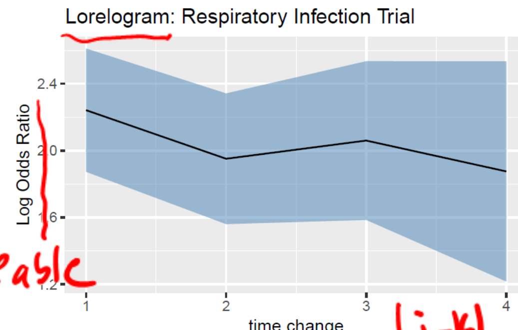
$$OR(Y_{ij}, Y_{ik}) = \frac{Pr(Y_{ij} = 1, Y_{ik} = 1)Pr(Y_{ij} = 0, Y_{ik} = 0))}{Pr(Y_{ij} = 1, Y_{ik} = 0)Pr(Y_{ij} = 0, Y_{ik} = 1))}$$

Y_{ik}

0 1

Y_{ij} 0

1



exchangeable or 1

|j-k|

Example: Fitting the model in R using gee

geepack
↳ geeglm

```
data$post = ifelse(data$time>1,1,0)
data$postXtrt = data$post * data$trtmnt01
fit.exch = gee(r~post+post:trtmnt01,data=data,
       family="binomial",corstr="exchangeable",id=id)
```

assume Bernoulli var, exchangeable corr

gee
↳ gee

```
##
## Coefficients:
##                  Estimate Naive S.E.    Naive z Robust S.E.   Robust z
## (Intercept)   -0.19885086  0.1915041 -1.0383635   0.1907707 -1.0423556
## post          -0.04097561  0.1943549 -0.2108288   0.2103911 -0.1947592
## post:trtmnt01  1.00825259  0.2457427  4.1028787   0.2624356  3.8419053
##
## Estimated Scale Parameter:  1.007704
## Number of Iterations:  2
##
## Working Correlation
##           [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] 1.0000000 0.4673692 0.4673692 0.4673692 0.4673692
## [2,] 0.4673692 1.0000000 0.4673692 0.4673692 0.4673692
## [3,] 0.4673692 0.4673692 1.0000000 0.4673692 0.4673692
## [4,] 0.4673692 0.4673692 0.4673692 1.0000000 0.4673692
## [5,] 0.4673692 0.4673692 0.4673692 0.4673692 1.0000000
```

βs

robust variance

model based se

robust se

V_i $\quad Corr(\hat{r}_{ij}, \hat{r}_{ik})$

r_{ij} = standardized Pearson residuals

Example: Interpretation of results

```
## Coefficients:
##                 Estimate Naive S.E.    Naive z Robust S.E.   Robust z
## (Intercept)  -0.19885086  0.1915041 -1.0383635   0.1907707 -1.0423556
## post         -0.04097561  0.1943549 -0.2108288   0.2103911 -0.1947592
## post:trtmnt01  1.00825259  0.2457427  4.1028787   0.2624356  3.8419053
## 
##                  Marg Marg LL Marg UL MargR LL MargR UL
## (Intercept)     0.820   0.559   1.202    0.560    1.200
## post            0.960   0.651   1.416    0.630    1.462
## post:trtmnt01   2.741   1.677   4.481    1.622    4.633
```

naive robust

- Interpretation of parameters:

$\exp(\beta_0)$ = the odds of good respiratory function at baseline are .82, 95% CI .56 to 1.20

$\exp(\beta_1)$ = Among persons receiving placebo, the odds of good resp function during the follow-up are decreased by 4% compared to baseline,

Example: Comparison across working correlation models

- Compare the results to the model fit assuming independence + gee

```
##                  Marg Marg LL Marg UL MargR LL MargR UL   Ind Ind LL Ind UL
## (Intercept)     0.820   0.559   1.202    0.560    1.200 0.820  0.559  1.202
## post            0.960   0.651   1.416    0.630    1.462 0.970  0.608  1.547
## post:trtmnt01   2.741   1.677   4.481    1.622    4.633 2.679  1.802  3.982
##                IndR LL IndR UL
## (Intercept)      0.560   1.200
## post             0.620   1.519
## post:trtmnt01    1.437   4.994
```

→ Hypothesis tests:
wald type
no likelihood ratio tests

WLS

Conditional Models

- Random effects logistic regression model:

Conditional Models

$$
\begin{aligned}
logit[Pr(Y_{ij}=1|post_{ij}, trtmnt01_i, b_i)] &= \beta^c_{0i} + \beta^c_1 I(post_{ij}>0) + \beta^c_2 I(post_{ij}>0)trtmnt01_i \\
&= \beta^c_0 + b_i + \beta^c_1 I(post_{ij}>0) + \beta^c_2 I(post_{ij}>0)trtmnt01_i
\end{aligned}
$$

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

Interpretation:

- β^c_{0i}: defines a patient specific log-odds of a good respiratory response at baseline
- $\beta^c_{0i} = \beta^c_0 + b_i$, where $b_i \sim N(0, \sigma^2)$: β^c_0 is the log-odds of a good respiratory response for the average patient (i.e. $b_i = 0$)
- $\beta^c_{0i} = \beta^c_0 + b_i$, 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

```r
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

```r
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)

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

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

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

Next time…

- Quick comments on estimation
  - Conditional logistic regression where we don't assume a distribution for b_i
  - Application to matched case control study

- Motivation and regression models for Poisson random variables