Exploring the Beta-Binomial Model: Analyzing Overdispersed Binomial Data

Holling Cancer Center Talk

Dr. Brian Neelon and Chun-Che Wen

May 09, 2024

Overdispersed Binomial Data

Binomial model

- Binomial distribution assumes the Bernoulli trials are independent of one another (e.g. coin toss).
- If this assumption is violated, then data have a greater variation than assumed under the binomial data.
- As known as **overdispersion (OD)**
- Traditional normal and Poisson model may not be appropriate
  - Poisson model assumes unlimited upper bound.
- We will present a few examples in the next slides

Potential Applications - (1)

Timeline Followback (TLFB) Data

Number of abstinent days for prior week, ranging from 0 to 7

Subject ID: 1

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
			×	×	×	×

TLFB= 4

Subject ID: 2

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
×						×

TLFB= 2

Potential Applications - (2)

Adverse Childhood Experiences (ACEs) Questionnaire Survey

Participants document the cumulative count of adverse childhood events, ranging from 0 to 10 occurrences

1. Did you feel that you didn't have enough to eat, had to wear dirty clothes, or had no one to protect or take care of you?	☐
2. Did you lose a parent through divorce, abandonment, death, or other reason?	☐
3. Did you live with anyone who was depressed, mentally ill, or attempted suicide?	☐
4. Did you live with anyone who had a problem with drinking or using drugs, including prescription drugs?	☐
5. Did your parents or adults in your home ever hit, punch, beat, or threaten to harm each other?	☐
6. Did you live with anyone who went to jail or prison?	☐
7. Did a parent or adult in your home ever swear at you, insult you, or put you down?	☐
8. Did a parent or adult in your home ever hit, beat, kick, or physically hurt you in any way?	☐
9. Did you feel that no one in your family loved you or thought you were special?	☐
10. Did you experience unwanted sexual contact (such as fondling or oral/anal/vaginal intercourse/penetration)?	☐
Your ACE score is the total number of checked responses	

Other Potential Applications

- Number of readmissions within at 30 days period
- Number of tissue factor + cells among all CD3+/CD4+ cells

Inference Approaches

- Frequentist inference
  - **Parametric Beta-Binomial (BB) Model** (Focus)
  - Quasi-likelihood Method (R Code)
- Bayesian inference (Flexible modeling)

Beta-Binomial Distribution

Derivation of Beta-Binomial Distribution

$$Y|\pi \sim \text{Bin}(m, \pi)$$
$$\pi \sim \text{Beta}(a, b),$$

- $m =$ the total number of Bernoulli trials
- $\pi =$ **random variable**, representing the probability of success
- a and b are the two shape parameters in beta distribution.
- Recall: the mean and variance of Beta(a, b) distribution
  - $\text{E}(\pi) = \frac{a}{a+b}$
  - $\text{Var}(\pi) = \frac{ab}{(a+b)^2(a+b+1)} = \left(\frac{a}{a+b}\right)\left(\frac{b}{a+b}\right)\boxed{\left(\frac{1}{a+b+1}\right)}$

Derivation of Beta-Binomial Distribution (Cont.)

Marginal pdf of Y:

$$
\begin{aligned}
f(Y=y) &= \int_0^1 \Pr(Y|\pi)\Pr(\pi)d\pi = \int_0^1 \Pr(Y,\pi)d\pi \\
&= \int_0^1 \underbrace{\binom{m}{y}\pi^y(1-\pi)^{m-y}}_{\text{Bin}(m,\pi)} \underbrace{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\pi^{a-1}(1-\pi)^{b-1}}_{\text{Beta}(a,b)} d\pi \\
&= \binom{m}{y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 \pi^y(1-\pi)^{m-y}\pi^{a-1}(1-\pi)^{b-1}d\pi \\
&= \binom{m}{y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\underbrace{\int_0^1 \pi^{(y+a)-1}(1-\pi)^{(m-y+b)-1}d\pi}_{\text{kernel of Beta}(y+a,m-y+b)} \\
&= \binom{m}{y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(y+a)\Gamma(m-y+b)}{\Gamma(m+a+b)} \\
&= \binom{m}{y}\frac{B(y+a,m-y+b)}{B(a,b)}
\end{aligned}
$$

Note: $B(c,s) = \frac{\Gamma(c)\Gamma(d)}{\Gamma(c+d)} = \frac{(c-1)!(d-1)!}{(c+d-1)!}$

Mean and Variance of BB Distribution

$$
\begin{aligned}
\mathrm{E}(Y) &= \mathrm{E}[\mathrm{E}(Y|\pi)] = m\frac{a}{a+b} = m\mu \\
\mathrm{Var}(Y) &= \mathrm{E}[\mathrm{Var}(Y|\pi)] + \mathrm{Var}[\mathrm{E}(Y|\pi)] \\
&= m\mu(1-\mu)\left[1 + (m-1)\boxed{\frac{1}{a+b+1}}\right] \\
&= m\mu(1-\mu)\left[1 + (m-1)\frac{\theta}{1+\theta}\right] \quad \left(\text{let } \theta = \frac{1}{a+b},\ \text{Agresti and Arostegui}\right) \\
&= m\mu(1-\mu)\left[1 + (m-1)\frac{1}{1+\phi}\right] \quad (\text{let } \phi = a+b,\ \text{SAS} - \text{Proc fmm}) \\
&= m\mu(1-\mu)\left[1 + (m-1)\rho\right] \quad \left(\text{let } \rho = \frac{1}{a+b+1}, \text{R } (\texttt{VGAM} \text{ and } \texttt{aod} \text{ package})\right)
\end{aligned}
$$

Note 1: m can be varied by observations, $Y_i \sim \mathrm{BB}(m_i, \mu_i, \rho)$, $i = 1, \ldots, n$

Note 2: $\frac{\theta}{1+\theta} = \rho$ the correlation between each pair of the Bernoulli trials.

BB distribution in terms of μ and $(\theta,\ \phi,\ \rho)$

$$Y \sim \text{BB}(m, a = \text{Shape 1}, b = \text{Shape 2})$$

$$f(Y = y) = \binom{m}{y} \frac{B(y + a, m - y + b)}{B(a, b)}$$

- Agresti and Arostegui (2007) & Najera-Zuloaga (2017)

$$\boxed{a = \mu/\theta \text{ and } b = (1 - \mu)/\theta},$$

- Proc FMM

$$\boxed{a = \mu\phi \text{ and } b = (1 - \mu)\phi},$$

  where $0 < \phi$ is called scale parameter.

- R `VGAM` and `aod` pacakges

$$\boxed{a = \mu(1 - \rho)/\rho \text{ and } b = (1 - \mu)(1 - \rho)/\rho},$$

Summary - (1)

In our slides,

- $\theta = \frac{1}{a+b} > 0$
  - When $\theta \to \infty$, OD; when $\theta \to 0$, $Y \to$ Bin
- $\phi = a + b > 0$
  - When $\phi \to 0$, OD; when $\phi \to \infty$, $Y \to$ Bin
- $\rho = \frac{1}{a+b+1}$, $0 < \rho < 1$
  - When $\rho \to 1$, OD; when $\rho \to 0$, $Y \to$ Bin

Recall: As $\theta \to 0$, $\phi \to \infty$, $\rho \to 0$ in the beta distribution, $\text{Var}(\pi) \to 0$ and beta distribution converges to a degenerate distribution at μ. Then $\text{Var}(Y) \to m\mu(1-\mu)$ and BB distribution coverges to the Bin(m, μ).

Summary - (2)

Software	Parameters	Correspond to our slides
Agresti	μ, θ	μ, θ
R (`VGAM` package)	μ, ρ	μ, ρ
R (`aod` package)	μ, ϕ	μ, ρ
R (`PROreg` package) Arostegui (2007) Najera-Zuloaga (2017)	μ, ϕ	μ, θ
SAS (Proc FMM)	μ, ϕ	μ, ϕ

Beta-Binomial Regression (BBR) Model

You can use one of (1) μ and θ, (2) μ and ϕ, and (3) μ and ρ these three parameterizations.

- We choose the logit link function in regression model:

$$\text{logit}(\mu_i) = \alpha + \boldsymbol{x}_i^T \boldsymbol{\beta}$$

- The interpretations of the α and $\boldsymbol{\beta}$ are the same as logistic regression
- In theory, you can also choose probit or cloglog links

Implementation

Data Description: Orobanche

Orobanche is a genus of parasitic plants with chlorophyll that grow on the roots of flowering plants.

- Total sample size: 21
- Outcome (Y): germinated/seeds (Number of germinated/ number of seeds)
- Covariates (X): Host type (Bean and Cucumber) and Variety (0.a73 and 0.a75)
- Regression Model:

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 \text{Cucumber}_i + \beta_2\ 0.a75_i + \beta_3(\text{Cucmber}_i \times 0.a075_i)$$

```
library(dispmod)
data(orobanche) # Data
head(orobanche) # Present the first 6 obs
```

```
##   germinated seeds slide host variety
## 1         10    39     1 Bean   0.a75
## 2         23    62     2 Bean   0.a75
## 3         23    81     3 Bean   0.a75
## 4         26    51     4 Bean   0.a75
## 5         17    39     5 Bean   0.a75
## 6          5     6     1 Cuke   0.a75
```

Binomial Model

```
# Binomial Model
mod0<-glm(cbind(germinated, seeds-germinated) ~ host*variety,
      data = orobanche, family = binomial)
```

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-0.4122	0.1842	-2.2383	0.0252
Cucumber	0.5401	0.2498	2.1619	0.0306
O.a75	-0.1459	0.2232	-0.6539	0.5132
Cucumber x O.a75	0.7781	0.3064	2.5392	0.0111

$$Y_i \sim \text{Bin}(m_i, \mu_i)$$
$$\text{logit}(\mu_i) = \beta_0 + \beta_1 \text{Cucumber}_i + \beta_2\ 0.a75_i + \beta_3(\text{Cucmber}_i \times 0.a075_i)$$

- Reference group: Host: Bean and Variety: 0.a73
- $\exp(\beta_1)$=exp(0.54)=1.72=OR(Bean + 0.a075 vs ref.)
- $\exp(\beta_1+\beta_2+\beta_3)$=exp(1.17)=3.23=OR(Cucumber + 0.a075 vs ref.)

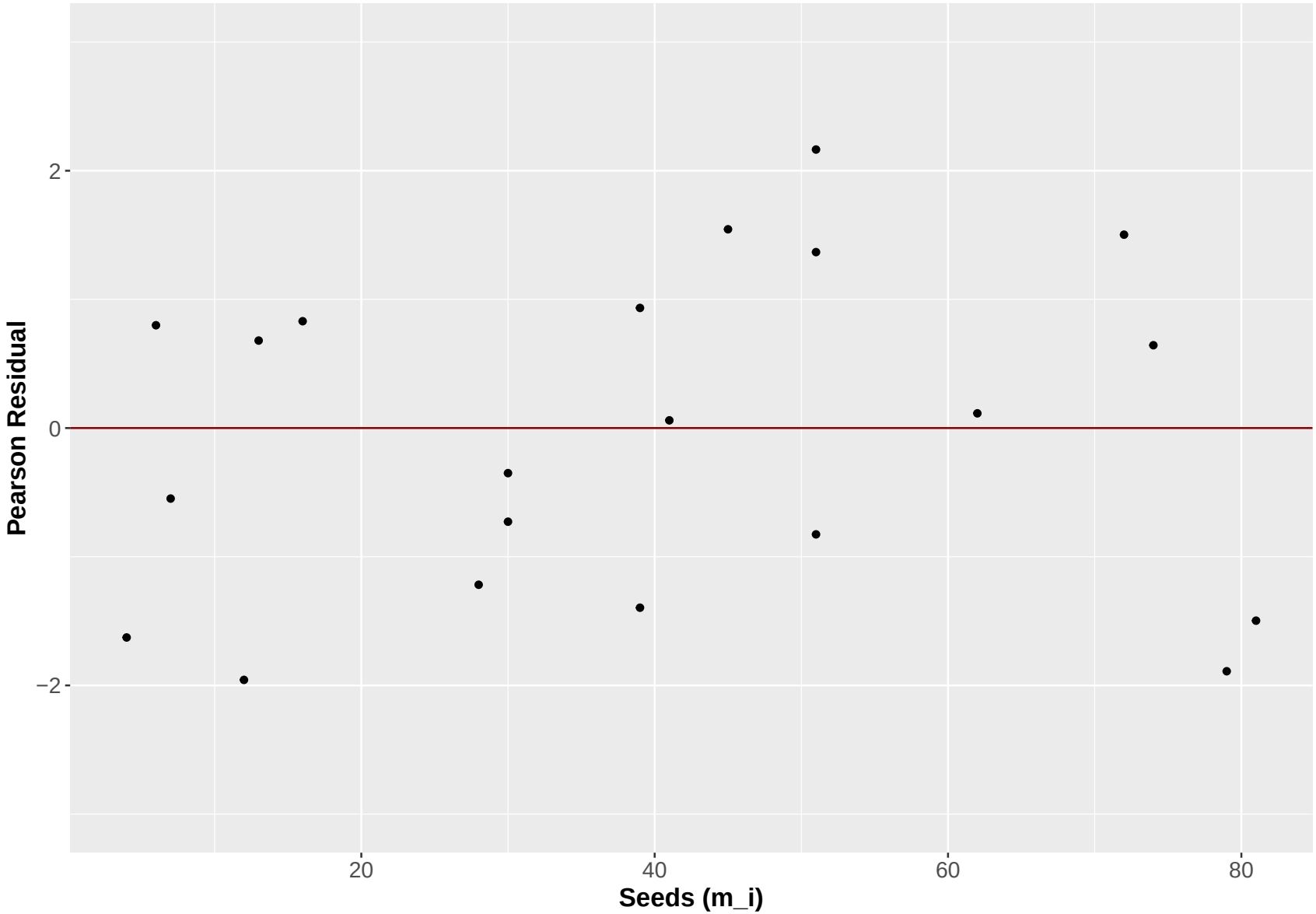
Model Diagnostic

Liang and McCullagh (1993) suggests plotting Pearson residual from a binomial GLM against m_i

- If the variability in the residuals changes as a function of m_i, then BB might be preferred.
- If the variability in the residual does not vary as a function of m_i, then quasi-binomial model that assumes the constant overdispersion might be preferred.
  - The details of quasi-binomial method is provided in Appendix

Model Diagnostic (conti.)

X-axis: Number of seeds (m_i) and Y-axis: Pearson Residuals from a binomial GLM



R Package: "aod" (Analysis of Overdispersed Data)

```
# Parametric BB Fit
library(aod)       # betabin()
mod1<-betabin(cbind(germinated, seeds-germinated) ~ host*variety,
              link="logit", # default
              random=~1, data = orobanche)
```

	Estimate	Std. Error	z value	Pr(> |z|)
Intercept	-0.4456	0.2183	-2.0411	0.0412
Cucumber	0.5235	0.2968	1.7636	0.0778
O.a75	-0.0961	0.2737	-0.3512	0.7255
Cucumber x O.a75	0.7962	0.3779	2.1068	0.0351

```
summary(mod1)@Phi # Same as rho in our slide
```

```
##                    Estimate Std. Error  z value   Pr(> z)
## phi.(Intercept) 0.01235803 0.01130974 1.092689 0.1372651
```

- Instead of using $ to extract element, use @
- Phi corresponds to $\rho = \frac{1}{a+b+1}$ $(0 < \rho < 1)$ in our slide
- Significance of test $H_0 : \rho = 0$ vs $H_0 : \rho > 0$. $P - value = 0.14$ suggests we cannot conclude that ρ is significantly different from 0
  - This could be due to lack of power under the z-test, given the small sample comprising only 21 slides.

SAS Code: "PROC FMM" (Finite mixture model)

```
proc fmm data=one;
 model y/m= cuke oa75 int/dist=bb ;
run;
```

Parameter Estimates for Beta-Binomial Model				
Effect	Estimate	Standard Error	z Value	Pr > |z|
Intercept	-0.4446	0.2183	-2.04	0.0417
cuke	0.5221	0.2968	1.76	0.0786
oa75	-0.09739	0.2737	-0.36	0.7219
int	0.7979	0.3780	2.11	0.0348
Scale Parameter	79.9003	74.2909		

```
# Scale Parameter = phi = a+b (in our slide)
cat("rho=",1/(1+79.003)) # compared to aod pacakge
```

```
## rho= 0.01249953
```

Summary Table

	Binomial	Beta-Binomial (aod)	Beta-Binomial (FMM)	QL-Bin*	QL-BB (aod)*
Intercept	-0.4122	-0.4456	-0.4446	-0.4122	-0.4653
Cucumber	0.5401	0.5235	0.5221	0.5401	0.5102
O.a75	-0.1459	-0.0961	-0.0874	-0.1459	-0.0701
Cucumber x O.a75	0.7781	0.7962	0.7979	0.7781	0.8196
OD Par.	NA	$\rho = 0.0124$	$\phi = 79.003$	$\psi = 1.8618$	$\psi = 0.0249$

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 \text{Cucumber}_i + \beta_2\ 0.a75_i + \beta_3(\text{Cucmber}_i \times 0.a075_i)$$

Odds ratio results from beta-binomial model:

- Reference group: Host: Bean and Variety: 0.a73
- exp(β_1)=exp(0.5235)=1.69=OR(Cucumber + 0.a073 vs ref.)
- exp(β_2)=exp(-0.0.91)=0.91=OR(Bean + 0.a075 vs ref.)
- exp($\beta_1+\beta_2+\beta_3$)=exp(1.2236)=3.40=OR(Cucumber + 0.a075 vs ref.)

*Quasi-likelihood (QL) methods of binomial/beta-binomial types of variance. Details and code are presented in Appendix.

Longitudinal BB Data

Smoking Cessation Study (two-arm randomized trial)

- At each weekly visit, participants reported the number of abstinent days during the past week (7 days).
- Timeline followback (TLFB) assessments of abstinence days were measured weekly over the course of the 12-week study.

Extend to BB Mixed Model

Y_{ij}= number of abstinent days for the prior week for subject i at j-th visit

$$Y_{ij} \sim \text{BB}(m = 7, \mu_{ij}, \rho)$$
$$\text{logit}(\mu_{ij}) = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \boldsymbol{u}_{ij}^T \boldsymbol{b}_i$$
$$\boldsymbol{b}_i \sim \text{N}_q(\boldsymbol{0}, \boldsymbol{G})$$

- Fixed effect model

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 t_{ij}$$

- Random intercept model

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 t_{ij} + b_i = (\beta_0 + b_i) + \beta_1 t_{ij}$$

- Random slope model

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 t_{ij} + b_{1i} + b_{2i} t_{ij} = (\beta_0 + b_{1i}) + (\beta_1 + b_{2i}) t_{ij}$$

Data Description: Smoking Cessation Study

- y = number of abstient days for the prior week
- $m = 7$ days (prior week)
- t = time variable, ranging 1 to 12

```
library(PROreg) # BBmm()
head(dat)
```

```
##      id y m t
## 1 9002 0 7 1
## 2 9002 0 7 2
## 3 9002 0 7 3
## 4 9002 0 7 4
## 5 9002 0 7 5
## 6 9002 0 7 6
```

- Unfortunately, we can't share the data. Please see "Simulated BB Mixed Mode.R" for simulated data.

BB Random Intercept Model

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 t_{ij} + b_i = (\beta_0 + b_i) + \beta_1 t_{ij}$$

```
model <- BBmm(fixed.formula = y~t,
              random.formula = ~id,m=7,data=dat)
```

- $m = 7$ days (prior week)
- Need to factorize **id** variable in your dataset

```r
summary(model)
```

```
## Call:    BBmm(fixed.formula = y ~ t, random.formula = ~id, m = 7, data = dat)
##
## Fixed effects coefficients:
##
##              Estimate    StdErr t.value   p.value
## (Intercept) -3.692900  0.112806 -32.737 < 2.2e-16 ***
## t            0.355487  0.015093  23.553 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ----------------------------------------------------------------
## Random effects dispersion parameter(s):
##
##    Estimate    StdErr
## id 2.524921 0.1723668
##
## ----------------------------------------------------------------
## Logarithm of beta-binomial dispersion parameter log(phi):
##
##     Estimate    StdErr
## 1 -0.9077821 0.0821816
##
## ----------------------------------------------------------------
## Deviance of the model: 1009.56 ; with 1278 degrees of freedom.
## Deviance of the null model 3275.383 ; with 1279 degrees of freedom.
## Deviance goodness-of-fit test p-value: 0
##
## Number of observations: 1282
## Number of iterations: 10
## Balanced data, maximum score number: 7
## Number of random effects in each random component: 145
## Number of analysed dimensions: 1
```

BB Random Intercept Model (Alternative)

- Stack all observations (i, j): $\text{logit}(\boldsymbol{\mu}) = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b}$

```
library(lme4)
# Create random effect design matrix
randformula = "y ~ t  + (1 | id)"
tmp <- lFormula(eval(randformula), dat)
Z <- t(as.matrix(tmp$reTrms$Zt))
```

$$\boldsymbol{Z} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \end{pmatrix} \} \text{ repeat } n_1 \text{ times}$$
$$\begin{matrix} \dots & \dots & \dots & \dots & \dots \end{matrix}$$
$$\begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \end{pmatrix} \} \text{ repeat } n_2 \text{ times}$$
$$\begin{matrix} \dots & \dots & \dots & \dots & \dots \end{matrix}$$
$$\begin{pmatrix} 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} \} \text{ repeat } n_n \text{ times}$$

$$; \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

BB Random Intercept Model (Alternative)

```
model2 <-BBmm(y=y,X=X,Z=Z,m=7,nRandComp=n)
```

- y = number of abstinent days, Y_{ij}
- $\boldsymbol{X}$ = fixed effect design matrix
- $\boldsymbol{Z}$ = random effect design matrix
- m = number of Bernoulli trials
- *nRandComp* = the number of random effects in each random component of the model (number of subjects)
- If you specify $\boldsymbol{Z}$, do not use `ramdom.formula` argument at the same time.

```r
summary(model2)  # same as model 1
```

```
## Call:    BBmm(X = X, y = y, Z = Z, nRandComp = n, m = 7)
##
## Fixed effects coefficients:
##
##    Estimate    StdErr t.value   p.value
##   -3.692900  0.112806 -32.737 < 2.2e-16 ***
## t  0.355487  0.015093  23.553 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------
## Random effects dispersion parameter(s):
##
##   Estimate    StdErr
## 1 2.524921 0.1723668
##
## ------------------------------------------------------------
## Logarithm of beta-binomial dispersion parameter log(phi):
##
##     Estimate    StdErr
## 1 -0.9077821 0.0821816
##
## ------------------------------------------------------------
## Deviance of the model: 1009.56 ; with 1278 degrees of freedom.
## Deviance of the null model 3275.383 ; with 1279 degrees of freedom.
## Deviance goodness-of-fit test p-value: 0
##
## Number of observations: 1282
## Number of iterations: 10
## Balanced data, maximum score number: 7
## Number of random effects in each random component: 145
## Number of analysed dimensions: 1
```

BB Random Intercept + Slope Model

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 t_{ij} + b_{1i} + b_{2i} t_{ij} = (\beta_0 + b_{1i}) + (\beta_1 + b_{2i}) t_{ij}$$

- Stack all observations (i, j):

$$\text{logit}(\boldsymbol{\mu}) = \boldsymbol{X}\beta + \underbrace{\boldsymbol{Z}_1\boldsymbol{b}_1}_{\text{Rand. Int.}} + \underbrace{\boldsymbol{Z}_2\boldsymbol{b}_2}_{\text{Rand. Sp.}}$$

```
randformula = "y ~ t + (1 + t | id)"
tmp <- lFormula(eval(randformula), dat)
Z <- t(as.matrix(tmp$reTrms$Zt))
Z1<-Z[,seq(1,dim(Z)[2],by=2)] # Random intercept matrix
Z2<-Z[,seq(2,dim(Z)[2],by=2)] # Random slope matrix
Zstar<-cbind(Z1,Z2)
```

$$\boldsymbol{Z}_1 = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} \begin{matrix} \\ \} \text{ repeat } n_1 \text{ times} \\ \\ \\ \} \text{ repeat } n_2 \text{ times} \\ \\ \\ \} \text{ repeat } n_n \text{ times} \end{matrix} ; \boldsymbol{b}_1 = \begin{bmatrix} b_{11} \\ b_{12} \\ \vdots \\ b_{1n} \end{bmatrix}$$

$$\boldsymbol{Z}_2 = \begin{pmatrix} t_{11} & 0 & 0 & \dots & 0 \\ t_{12} & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ t_{1n_1} & 0 & 0 & \dots & 0 \\ 0 & t_{21} & 0 & \dots & 0 \\ 0 & t_{22} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & t_{2n_2} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & t_{n_1} \\ 0 & 0 & 0 & \dots & t_{n_2} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & t_{n_n} \end{pmatrix} \begin{matrix} \\ \\ \\ \} \text{ repeat } n_1 \text{ times} \\ \\ \\ \\ \} \text{ repeat } n_2 \text{ times} \\ \\ \\ \\ \\ \} \text{ repeat } n_n \text{ times} \end{matrix} ; \boldsymbol{b}_2 = \begin{bmatrix} b_{21} \\ b_{22} \\ \vdots \\ b_{2n} \end{bmatrix}$$

```
model2 <-BBmm(y=y,X=X,Z=Zstar,m=7,nRandComp=c(n,n))
```

- y = number of abstinent days, Y_{ij}
- $\boldsymbol{X}$ = fixed effect design matrix
- $\boldsymbol{Z}$ = random effect design matrix (Intercept+Slope)
- m = number of Bernoulli trials
- *nRandComp* = the number of random effects in each random component of the model (number of subjects)
- Note: For the random slope model, we recommend to specify the model in terms of design matrix, such as $\boldsymbol{X}$ and $\boldsymbol{Z}$ as mentioned above

```r
summary(model3)
```

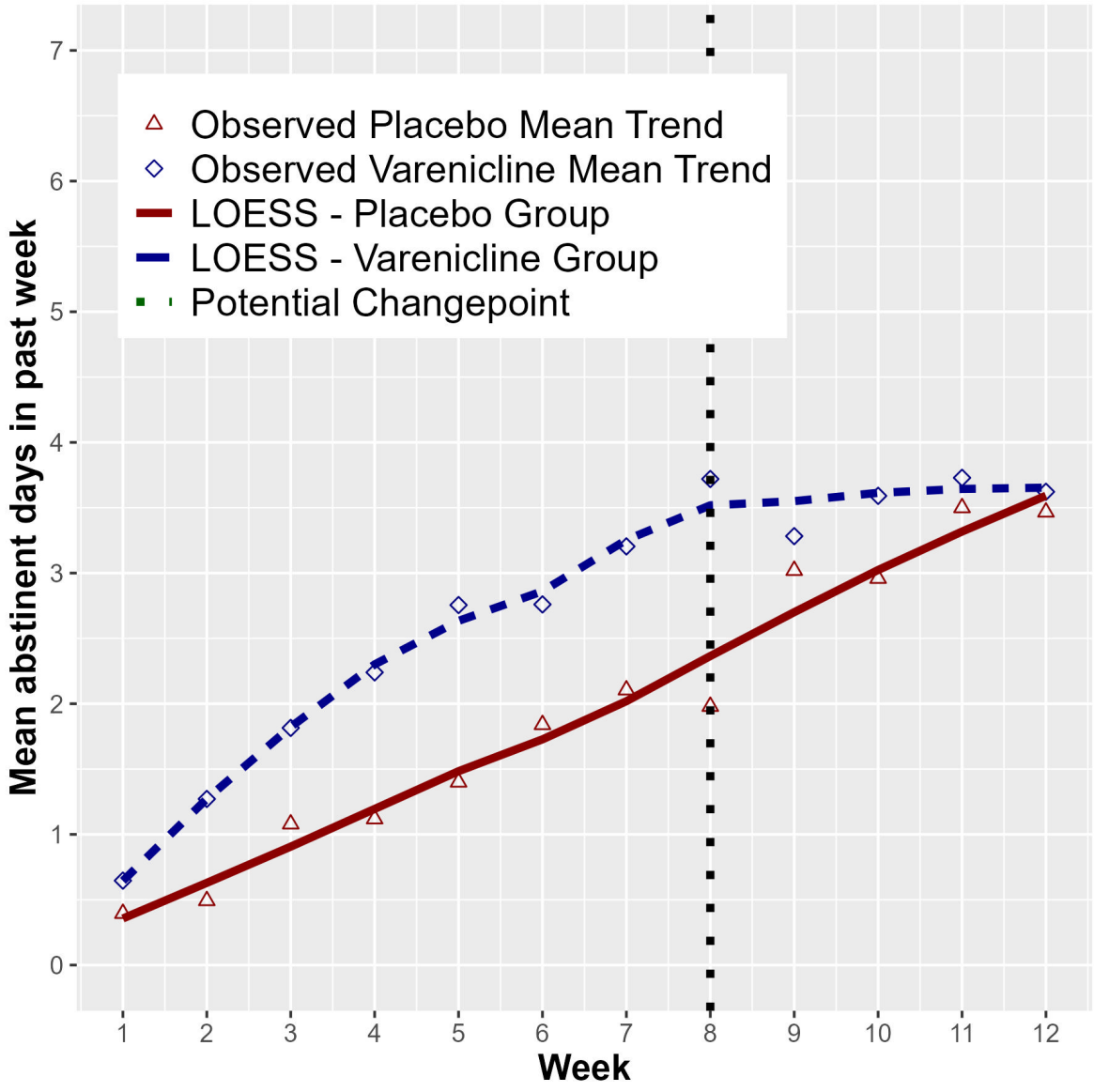
```
## Call:    BBmm(X = X, y = y, Z = Zstar, nRandComp = c(n, n), m = 7)
##
## Fixed effects coefficients:
##
##    Estimate    StdErr t.value   p.value
##   -3.869190  0.110300 -35.079 < 2.2e-16 ***
## t  0.400502  0.014742  27.168 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------
## Random effects dispersion parameter(s):
##
##    Estimate     StdErr
## 1 2.6254922 0.19944129
## 2 0.4026176 0.03202462
##
## ------------------------------------------------------------
## Logarithm of beta-binomial dispersion parameter log(phi):
##
##    Estimate   StdErr
## 1 -1.491936 0.103395
##
## ------------------------------------------------------------
## Deviance of the model: 846.6065 ; with 1277 degrees of freedom.
## Deviance of the null model 4161.672 ; with 1278 degrees of freedom.
## Deviance goodness-of-fit test p-value: 0
##
## Number of observations: 1282
## Number of iterations: 13
## Balanced data, maximum score number: 7
## Number of random effects in each random component: 145 145
## Number of analysed dimensions: 1
```

Special Modelings

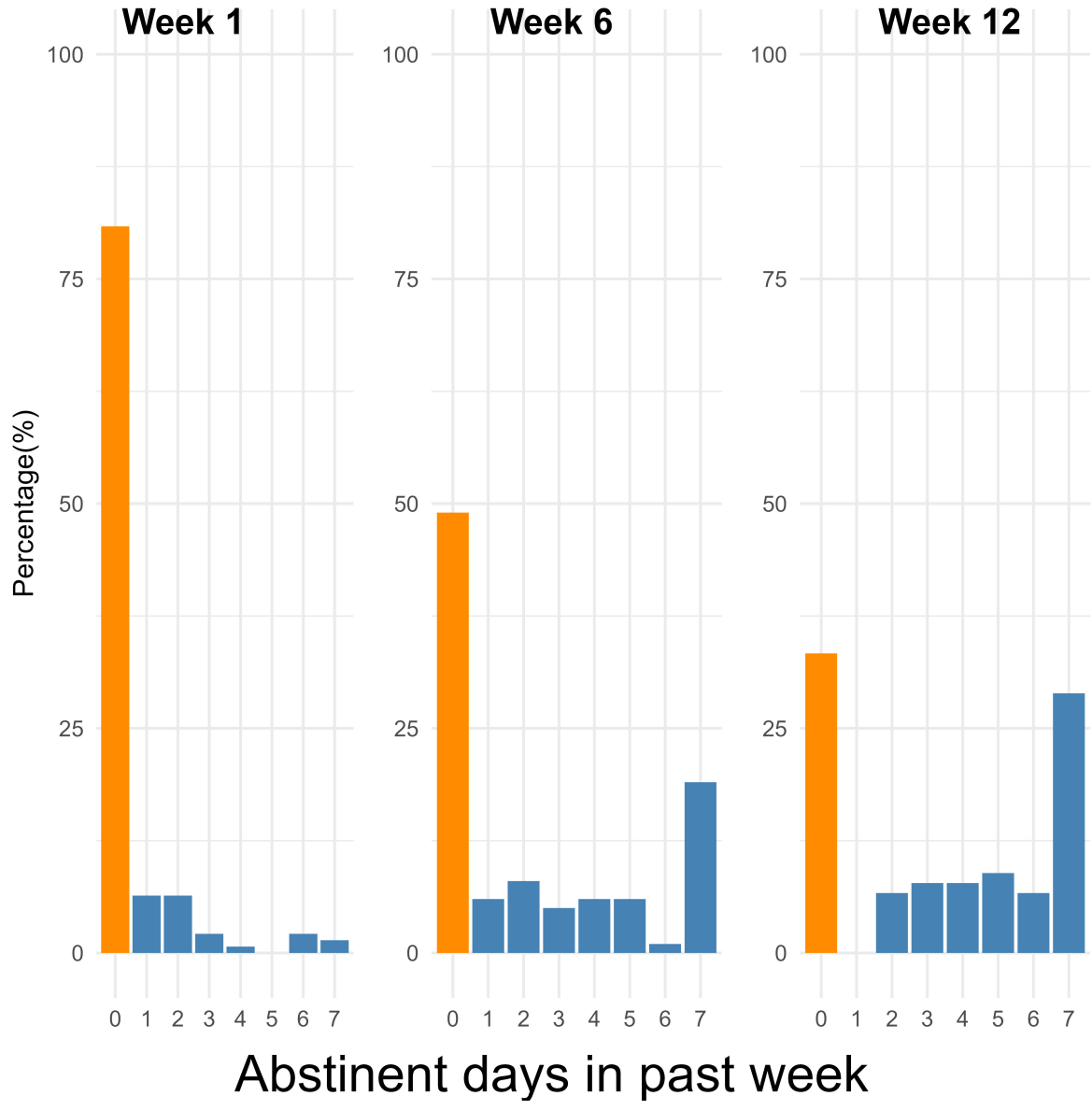
Bayesian Modeling

- Likelihood:
  - $Y_i \sim \text{BB}(m_i, \mu_i, \rho)$
  - $\text{logit}(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$
- Priors:
  - $\boldsymbol{\beta} \sim \text{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$
  - $\rho \sim \text{Unif}(0, 1)$
- Posterior $\propto$ Likelihood $\times$ Prior
  - Post. $= \prod_{i=1}^{n} p(y_i|\boldsymbol{\beta}, \rho) \times \pi(\boldsymbol{\beta}) \times \pi(\rho)$
- Metropolis-Hasting (MH) Steps

Changepoint Model

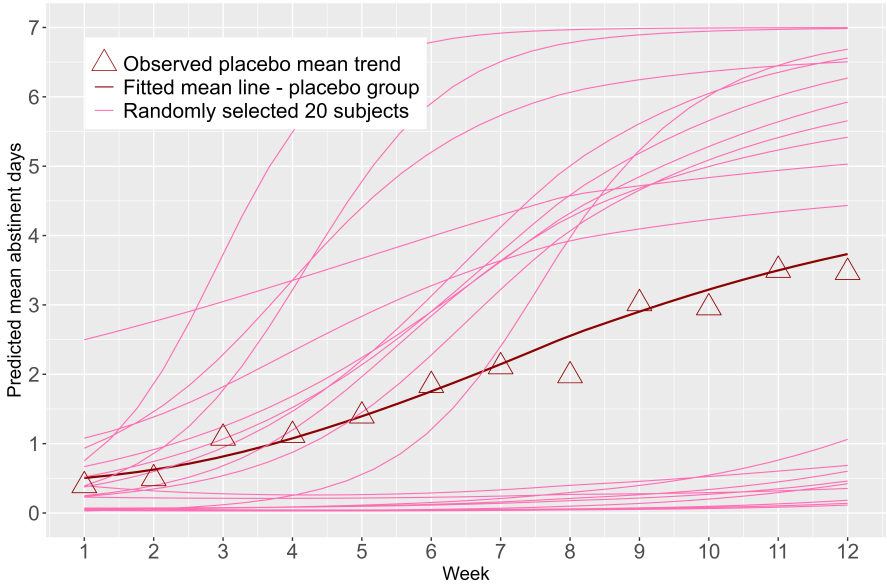


Zero-Inflated Model

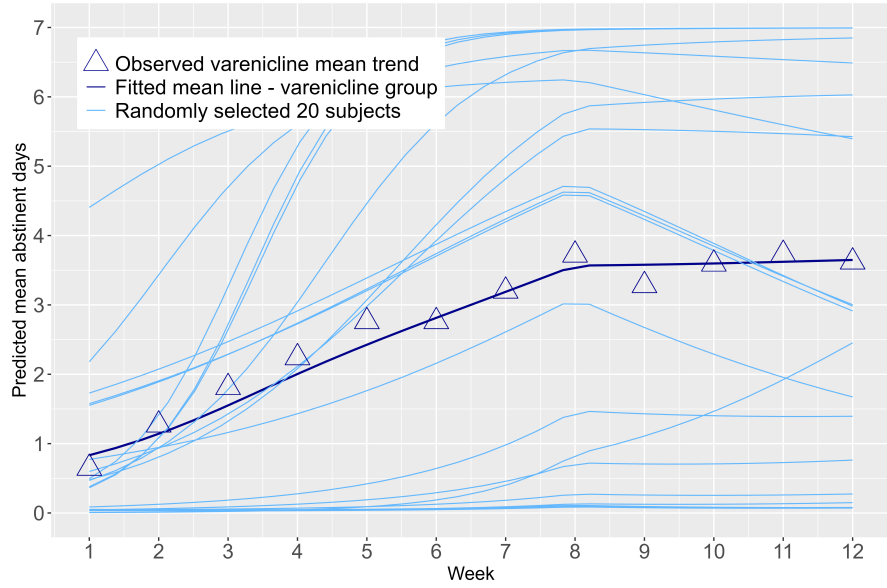


Mixture Model

Placebo Group



Varenicline Group



Appendix

Generate BB data in R

```r
library(VGAM) # rbetabinom()

n<-350            # Number of obs
m<-20             # Number of Bernolli trials (can by varied)

# Covariates
x1<-rnorm(n)
x2<-rbinom(n,1,0.5)
X<-cbind(1,x1,x2)

# Fixed effects
beta<-c(-0.5,1.0,-0.5)

# Correlation parameter (0<rho<1)
rho<-0.25

# Linear predictors
eta<-X%*%beta
# Logit link (can be probit/cloglog links)
pi<-exp(eta)/(1+exp(eta))

# Generate outcomes
y<-rbetabinom(n=n,size=m,prob=pi,rho=rho)
```

Quasi-likelihood (QL) Method - Binomial Type of Variance

- Instead of specifying distribution for Y, we specify the relationship b/t mean and variance functions.
- $\text{Var}(Y_i) = v(\mu_i) = m_i\mu_i(1-\mu_i)$, under binomial sampling and $v(\mu_i)$ is the variance function.
- Alternative variance function: $\text{Var}(Y_i) = \psi v(\mu_i)$, for some constant ψ
- $\text{Var}(Y_i) = \psi[m_i\mu_i(1-\mu_i)]$
- $\psi > 1$ represent overdispersion in binomial model

```
# Quasi-Likelihood Method
mod2<-glm(cbind(germinated, seeds-germinated) ~ host*variety,
      data = orobanche, family = quasibinomial)
```

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-0.4122	0.2513	-1.6404	0.1193
Cucumber	0.5401	0.3409	1.5844	0.1315
O.a75	-0.1459	0.3045	-0.4792	0.6379
Cucumber x O.a75	0.7781	0.4181	1.8609	0.0801

```
summary(mod2)$dispersion # psi above
```

```
## [1] 1.861832
```

QL Method - Correlated Bernoulli Trials

- Assume a common correlation (ψ) between each pair of m_i
- Suppose $\pi_i = P(Y_{it} = 1)$, for $t = 1, \ldots, m_i$ and $\text{corr}(Y_{it}, Y_{is}) = \psi$ for $s \neq t$
- Then, $\text{Var}(Y_{it}) = \pi_i(1 - \pi_i)$ and $\text{cov}(Y_{is}, Y_{it}) = \psi\pi_i(1 - \pi_i)$
- $\text{Var}(Y_i = \sum_{y=1}^{m_i} Y_{it}) = m_i\mu_i(1 - \mu_i)[1 + (m_i - 1)\psi]$
- $\psi > -1/(m_i - 1)$ to ensure the $\text{Var}(Y_i) > 0$

```
# Quasi-Likelihood Method
library(aod)
mod3<-quasibin(cbind(germinated, seeds-germinated)~ host*variety,
               data=orobanche)
mod3
```

```
## Quasi-likelihood generalized linear model
## ----------------------------------------
## quasibin(formula = cbind(germinated, seeds - germinated) ~ host *
##     variety, data = orobanche)
##
## Fixed-effect coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -0.4653     0.2439 -1.9081   0.0564
## hostCuke                0.5102     0.3347  1.5244   0.1274
## variety0.a75           -0.0701     0.3115 -0.2250   0.8219
## hostCuke:variety0.a75   0.8196     0.4352  1.8831   0.0597
##
## Overdispersion parameter:
##    phi
## 0.0249
##
```