Multinomial Regression for Correlated Data Using the Bootstrap in R

Jennifer Thompson, MPH & Rameela Chandrasekhar, PhD,
Vanderbilt University

August 9, 2015

Purpose

- Multinomial logistic regression: Useful for outcomes with >2 levels without inherent order
- Model fits (# levels - 1) coefficients for each variable
- Some methods exist in R, including:
  - VGAM package: `vglm()` with `family = multinomial()`
  - multgee package: `nomLORgee()`
- To our knowledge, neither method allows us to easily get SEs/confidence intervals for predicted probabilities

Proposed Method: Clustered Bootstrapped Multinomial Regression

- Given data set with N subjects and m_n records per subject, use clustered bootstrap sampling to create B data sets
  - Sample N subject IDs with replacement
  - Take all m_n records from each sampled ID
- Fit multinomial model on each of B data sets

Proposed Method: Clustered Bootstrapped Multinomial Regression

- *Coefficients:* Estimates = means of B estimates
- *CIs:* Percentile method; (2.5th, 97.5th)
- *P-values:* Wald test
- *Predicted probability* of an outcome level: Estimates straightforward; for CIs, use method in Liu's *Survival Analysis: Models and Applications* Appendix B
- Functions collected in **ClusterBootMultinom** package on Github (github.com/jenniferthompson/ClusterBootMultinom)

Motivating Example

- Cohort of critically ill patients with data collected daily in the ICU
- Outcome: Mental status, assessed daily while in hospital; could be normal, delirious or comatose
- Cannot assume that coma is worse than delirium
- Exposure: Levels of a biomarker measured on study days 1, 3, and 5, if patient remained in the hospital
- Most confounders also measured daily in the ICU
- **Main question:** After adjusting for confounders, are biomarker levels associated with mental status on the day following biomarker measurement?
- Final data: 767 unique patients with $>=1$ day of complete data; 1946 total patient-days

Create Data Sets

```
# library(devtools)
# install_github(
#   'jenniferthompson/ClusterBootMultinom')

library(ClusterBootMultinom)

## Set number of bootstraps
nboot <- 25
```

Using `create.sampdata()`,

- Create B (here, 25) data sets, plus extra in case of nonconvergence
- Each has all records from 767 IDs sampled with replacement from set of original IDs

```
boot.datasets <-
  create.sampdata(org.data = our.data,
                  id.var = 'id',
                  n.sets = ceiling(nboot * 1.25))
```

Run Models on Bootstrapped Data Sets

Using **`multi.bootstrap()`**:

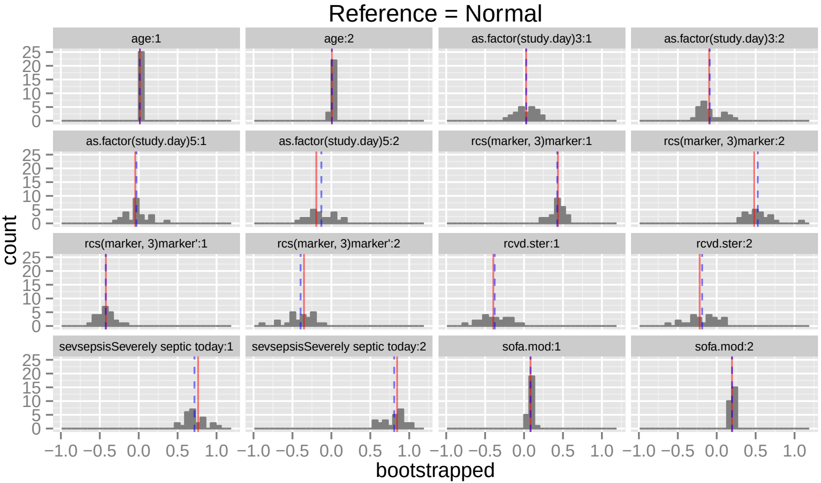
- Run model on original data set
- If that model converges, run same model on bootstrapped data sets until B converged models
- Save errors and warnings to .txt file
- To calculate CIs for predicted probabilities for all outcome levels, run models twice, using highest & lowest outcome levels as reference

```
## mod.formula <-
##  as.formula(mental.tmw ~ age + rcvd.ster +
##   sevsepsis + sofa.mod + as.factor(study.day) +
##   rcs(marker, 3))
```

Run Models on Bootstrapped Data Sets

```
## Run with Normal as reference level
boot.models.n <-
  multi.bootstrap(
    org.data = our.data,
      ## original data set
    data.sets = boot.datasets,
      ## list of bootstrapped data sets
    ref.outcome = grep('Normal',
                       levels(our.data$mental.tmw)),
      ## outcome level to use as reference
    multi.form = mod.formula,
      ## model formula
    n.boot = nboot,
      ## number of successful model fits desired
    xvar = 'Marker')
      ## text for status updates
```

Check Distribution of Coefficients using `boot.coef.plot()`

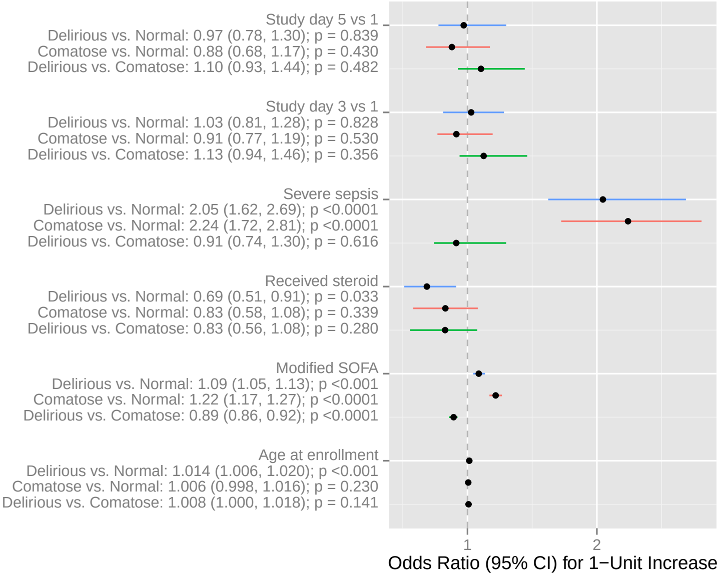


Calculate Odds Ratios

Use **`multi.plot.ors()`** to show ORs, 95% CIs for each outcome comparison.

```r
## Plot odds ratios and CIs for non-biomarker variables
covariate.ors <-
  multi.plot.ors(
    coef.list = list(boot.matrix.n, boot.matrix.c),
      ## List of matrices with bootstrapped coefs
    label.data = or.labels,
      ## data frame containing labels for each variable
    remove.vars = 'marker',
      ## this plot is just for confounders
    round.vars = 'age', round.digits = 3,
      ## round results for age to 3 instead of 2 places
    out.strings.list = list(out.comp.n, out.comp.c),
      ## list of strings describing comparisons
    delete.row = 'Normal vs. Comatose')
      ## One comparison will be redundant
```

```
covariate.ors$or.plot
```



Create Design Matrices

To get predicted probabilities for outcomes vs. a continuous covariate, we need to adjust all other covariates to specific values.

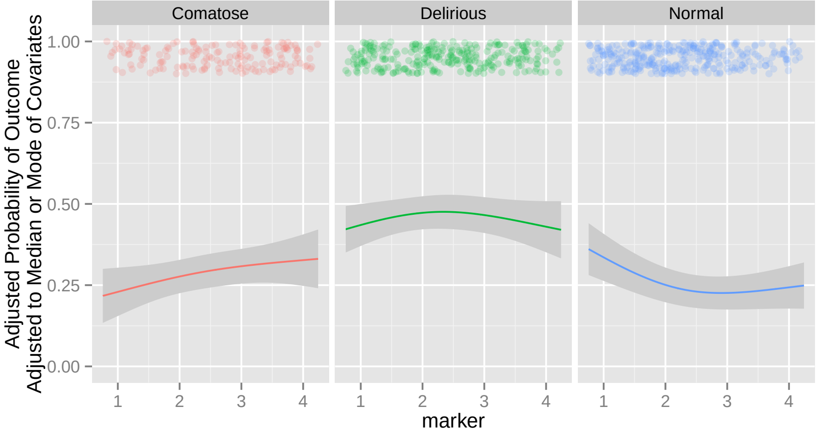
- Pass `multi.plot.probs()` [# outcome levels - 1] numeric vectors
- Functions assume covariate in question is **last** variable in model formula; its X values will become columns at the end of design matrices
- Example has $\beta_{0_{1,2}} + 6$ other β per outcome level, excluding biomarker; set each to median/mode, representing "average" patient

```
##             [,1] [,2]     [,3]     [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14]
## design.out1    1    0 60.84873  0.00000    0    0    1    0    6     0     0     0     0     0
## design.out2    0    1  0.00000 60.84873    0    0    0    1    0     6     0     0     0     0
```

Calculate & Plot Predicted Probabilities, CIs of Each Mental Status by Marker Level

```
## Predicted probabilities for
##  outcome levels vs. biomarker
marker.prob.results <-
  multi.plot.probs(
    xval = 'marker',
    data.set = our.data,
    design.mat = list(design.out1, design.out2),
    mod.objs = list(boot.models.n$org.model,
                    boot.models.c$org.model),
    coef.list = list(boot.matrix.n,
                     boot.matrix.c),
    vcov.list = list(boot.vcov.n,
                     boot.vcov.c))
```

```
marker.prob.results$prob.line.plot
```



Bootstrapped Wald P−Values:
Delirious vs. Normal: p < 0.0001
Comatose vs. Normal: p = 0.004
Delirious vs. Comatose: p = 0.187

Simulation Study

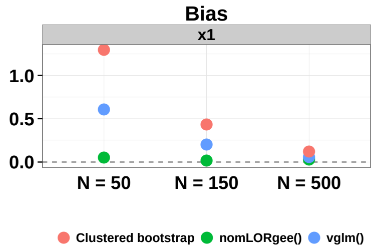
- Compared our method with
  1. `vglm()` from VGAM, without accounting for correlation
  2. `nomLORgee()` from multgee package, which accounts for correlation
- Simulated 1000 data sets with correlated multinomial data, based on example from SimCorMultRes package
  - data sets included ID, time (cluster size = 3), one X $\sim N(2.5, 3)$, outcome with $I = 4$ levels
  - all $\beta_{0i,\ldots,I-1} = 1$, all $\beta_{1i,\ldots,I-1} = 2$
  - correlation within patient = 0.9
  - N = 50, 150, 500

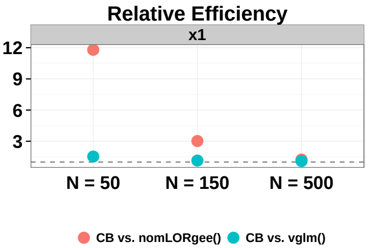
Model Convergence

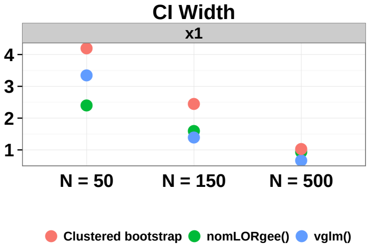
Proportions of models which did not converge:

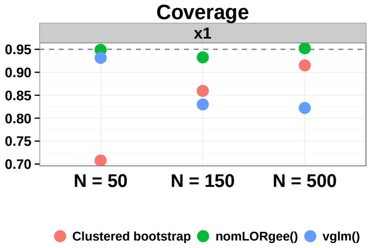
Method	N = 50	N = 150	N = 500
`nomLORgee()`	0.49	0.18	0.08
`vglm()`	0.14	0.01	0.01
Clustered bootstrap	0.33	0.03	0.01

Bias, Relative Efficiency & CIs









Future Work & Acknowledgements

- Future directions
  - Additional CI methods
  - Extending package to include more nonlinear terms, other flexibilities

- Clinical investigators & coauthors:
  - Tim Girard, MD, MSCI
  - Pratik Pandharipande, MD, MSCI
  - Wes Ely, MD, MPH
- R package resources:
  - Hilary Parker - Writing an R Package from Scratch
  - Hadley Wickham - devtools, roxygen2, *R Packages*
  - Karl Broman - R package primer
  - Jeremy Stephens, VUMC computer systems analyst

- Email: `jennifer.l.thompson@vanderbilt.edu`

- Package: github.com/jenniferthompson/ClusterBootMultinom