Regression with Biomarkers Subject to Lower Limits of Detection

Kevin Donovan

University of North Carolina at Chapel Hill

8/6/2020

Motivation

Adolescent Medicine Trials Network for HIV/AIDS Interventions (ATN)

- Study: ATN 071
- Objective: Analyze associations between selected biomarkers and various neurocognitive outcomes in Youth with HIV
- Challenge: Measurement error for many biomarkers due to lower limits of detection (LOD) (ex. viral load: 50 copies/mL)

Motivation

Study sample

- 128 Youth with HIV
- Neurocognitive outcomes:
  Motor, Attention, Executive Functioning, Verbal Nonmemory, Visuospatial Memory (composite Z scores)
- Composites = means of test performance Z scores
- Biomarkers: grouped by biological association
  1. Macrophage Activation (10 biomarkers, 3 with LOD)
  2. Vascular Inflammation (4 biomarkers, 0 with LOD)
  3. Lymphocyte Activation (4 biomarkers, 3 with LOD)

Motivation

Analytic Strategy

- Proportion of sample values under LOD:
  1. GM-CSF: 41% (LOD $=$ 0.01 pg/mL)
  2. IL-1β: 26% (LOD $=$ 0.01 pg/mL)
  3. IL-6: 20% (LOD $=$ 0.1 pg/mL)
  4. INFγ: 38% (LOD $=$ 0.1 pg/mL)
  5. IL-10: 27% (LOD $=$ 0.1 pg/mL)
  6. sIL-2rα: 1% (LOD $=$ 0.1 pg/mL)
  7. Viral Load: 62% (LOD $=$ 50 copies/mL)
- Fit linear regression models for each pair of neurcognitive domain scores and biomarker group (15 total models)
  - Outcome $=$ neurocognitive domain scores;
    Covariates $=$ biomarkers in group and potential confounders (ex. viral load, gender, race)

Methodology

In literature

- Substitution: use LOD/function of LOD value $(/\sqrt{2})$ [1]
- Maximum likelihood (ML) with observed only:
  likelihood generally derived with very limited number of covariates [1–3]
- Bayesian estimation [4]
- **ML using Expectation-Maximization (EM)**: allows for arbitrary number of LOD/non-LOD covariates [5]

Methodology

May et al. 2011

Model Overview:
Assume random sample of n independent observations of (Y_i, X_i) for $i = 1, \ldots, n$
X_i is a set of p covariates with $X_i \sim N_p(\mu_X, \Sigma_X)$,
$Y_i | X_i \sim N(\mu_i, \sigma^2)$ and $\mu_i = \beta_0 + \beta_1 X_{i,1} + \ldots + \beta_p X_{i,p}$ for $i = 1, \ldots, n$

For simplicity, assume covariates $X_{i,q}, X_{i,q+1}, \ldots, X_{i,p}$ are subject to LODs for $0 < q \leq p$

Let $Y = (Y_1, \ldots, Y_n)$ and X be the n by p martix of covariates.
If all covariates were fully observed (all inside LODs), the conditional log likelihood function would be

$$L(\theta|Y, X) = \sum_{i=1}^{n} \log[f(Y_i|X_i, \theta)],$$

where $\theta = (\beta, \sigma^2)$ and $\beta = (\beta_0, \ldots, \beta_p)$

Covariates with LODs may be unobserved $\implies$ model joint log likelihood instead:

$$L(\gamma|Y, X) = \sum_{i=1}^{n} \log[f(Y_i|X_i, \theta)f(X_i|\alpha)]$$

where $\gamma = (\theta, \alpha)$, $\alpha = (\mu_X, \Sigma_X)$, $X_i = (X_{i,1} \ldots,$ and $X_{i,p})$

LODs $\implies$ X_i can be partitioned into $X_i = (X_{i,obs}, X_{i,cens})$
where $X_{i,obs}$ denotes covariates with fully observed values,
$X_{i,obs}$ denotes covariates with values outside of their LODs

To maximize $L(\gamma|Y, X)$ with respect to γ, use E-M algorithm due to missing covariate values.

Let $\hat{\gamma}^{(t)}$ denote the estimate of γ at iteration t.
For simplicity, suppose $i = m, m+1, \ldots, n$ where $0 < m \leq n$ have covariate values outside of LODs.

Let $Y = (Y_1, \ldots, Y_n)$ and X be the n by p martix of covariates.
If all covariates were fully observed (all inside LODs), the conditional log likelihood function would be

$$L(\theta|Y,X) = \sum_{i=1}^{n} \log[f(Y_i|X_i,\theta)],$$

where $\theta = (\beta, \sigma^2)$ and $\beta = (\beta_0, \ldots, \beta_p)$

Covariates with LODs may be unobserved $\implies$ model joint log likelihood instead:

$$L(\gamma|Y,X) = \sum_{i=1}^{n} \log[f(Y_i|X_i,\theta) f(X_i|\alpha)]$$

where $\gamma = (\theta, \alpha)$, $\alpha = (\mu_X, \Sigma_X)$, $X_i = (X_{i,1} \ldots,$ and $X_{i,p})$

E-M Algorithm

Expectation (E) step:

$$
\begin{aligned}
Q(\gamma|\hat{\gamma}^{(t)}) &= \mathrm{E}[L(\gamma|Y,X)|Y,X_{obs}] \\
&= \sum_{i=1}^{m-1} \log[f(Y_i|X_i,\theta)f(X_i|\alpha)] \\
&+ \sum_{i=m}^{n} \mathrm{E}(\log[f(Y_i|X_i,\theta)f(X_i|\alpha)]|Y_i,X_{i,\text{obs}})
\end{aligned}
$$

where X_{obs} denotes the observed covariate values for all n observations.

$\mathrm{E}(\log[f(Y_i|X_i,\theta)f(X_i|\alpha)]|Y_i,X_{i,\text{obs}})$ often does not have a closed form [5] $\Longrightarrow$

Monte Carlo approximation is used for this term for $m \leq i \leq n$.

Monte Carlo Approximation

Suppose Z has density function $f(z)$.
Given $Z_1, \ldots, Z_n$ i.i.d from $f(z)$ for $n > 0$,
the Monte Carlo approximation of $E(Z)$ is $\sum_{i=1}^{n} Z_i/n$

Thus in the E step above, define Monte Carlo approximation

$$\hat{\mathrm{E}}(\log[f(Y_i|X_i,\theta)f(X_i|\alpha)]|Y_i, X_{i,\mathrm{obs}}) = \sum_{j=1}^{r_i} \log[f(Y_i|Z_i,\theta)f(Z_i|\alpha)]/r_i$$

for $i = m, \ldots, n$ where $Z_{i,j}$ i.i.d from truncated distribution of $X_{i,cens}|X_{i,obs}, Y_i, c_l < X_{i,cens} < c_u$

The corresponding approximated E step is

$$\hat{Q}(\gamma|\hat{\gamma}^{(t)}) = \sum_{i=1}^{m-1} \log[f(Y_i|X_i,\theta)f(X_i|\alpha)] + \sum_{i=m}^{n} \hat{\mathrm{E}}(\log[f(Y_i|X_i,\theta)f(X_i|\alpha)]|Y_i, X_{i,\mathrm{obs}})$$

Plugging-in the Monte Carlo expectation and rearranging terms results in

$$\begin{aligned}\hat{Q}(\gamma|\hat{\gamma}^{(t)}) &= \sum_{i=1}^{m-1} \log[f(Y_i|X_i,\theta)f(X_i|\alpha)] \\ &+ \sum_{i=m}^{n}\sum_{j=1}^{r_i} \log[f(Y_i|Z_{i,j},\theta)f(Z_{i,j}|\alpha)]/r_i \\ &= \sum_{i=1}^{m-1} \log[f(Y_i|X_i,\theta)] + \sum_{i=m}^{n}\sum_{j=1}^{r_i} \log[f(Y_i|Z_{i,j},\theta)]/r_i \\ &+ \sum_{i=1}^{m-1} \log[f(X_i|\alpha)] + \sum_{i=m}^{n}\sum_{j=1}^{r_i} \log[f(Z_{i,j}|\alpha)]/r_i\end{aligned}$$

thus the maximization step can be applied to obtain θ and α separately.

Maximization (M) step:

Above approx. E step $\implies$
$\hat{\theta}$ and $\hat{\alpha}$ obtained by weighted maximum likelihood estimation

Regression model $\implies$
weighted least squares procedure and weighted sample mean/covariance for $\hat{\theta}$ and $\hat{\alpha}$ respectively

weight $= 1$ for subjects with no covariates outside of LODs,
$= r_i$ else.

Often $r_i = r$ for all i

M step algorithm:

Per May et al. (2011)

For simplicity, assume all observations with missing covariate values have values under the LOD for covariates $X_{i,q}, \dots, X_{i,p}$:

1. provide starting values $\hat{\gamma}^{(0)}$ by fitting the model on only the observed data (i.e., a "complete case" analysis). Set the starting values for the sampling procedure to be the means of the covariates subject to a LOD in the observed data only.
2. at step t, for observation i with censored covariates $X_{i,\text{cens}}$, sample $X_{i,q,\text{cens}}$. Denote this sampled value by $Z_{i,q,1}$.
3. for observation i conditional on $X_{i,\text{obs}}$, Y_i, and $Z_{i,q,1}$, sample $X_{i,q+1,\text{cens}}$. Denote this sampled value by $Z_{i,q+1,1}$.
4. repeat step 2 and 3 until you have sampled values for all $X_{i,\text{cens}}$, resulting in vector $Z_{i,1} = (Z_{i,q,1}, \dots, Z_{i,p,1})$.

5. repeat steps 2-4 R times (e.g., 25) until you have $Z_{i,1}, \ldots, Z_{i,R}$
6. repeat steps 2-5 for all observations with censored covariates. This results in a new dataset, consisting of one observation (row) (Y_i, X_i) for a subject with fully observed covariate data (i.e., none under a LOD) and R observations (rows) $(Y_i, X_{i,obs}, Z_{i,1}), \ldots, (Y_i, X_{i,obs}, Z_{i,R})$ for a subject with at least one covariate under a LOD.
7. compute $\hat{\gamma}^{(t+1)}$ using weighted least squares with the sampled dataset detailed in step 6, with weight $1/R$ for observations with sampled data.
8. repeat steps 2-7 until the absolute difference between $\hat{\gamma}_{t-1}$ and $\hat{\gamma}_t$ is less than some predetermined threshold ϵ

Standard Error Calculation:

- Analytical calculation: Louis's method for computing the information matrix [6]
- Bootstrap standard errors

Sampling methods:

- Recall steps 2-6 involve sampling of unobserved biomarker values to compute Monte Carlo approximation
- Sampling is most computationally/time intensive portion of algorithm
- Options:
  1. Adaptive Rejection Metropolis Sampling (ARMS) [7], used by May et al. (2011)
  2. Slice sampling [8]

Standard Error Calculation:

- Analytical calculation: Louis's method for computing the information matrix [6]
- Bootstrap standard errors

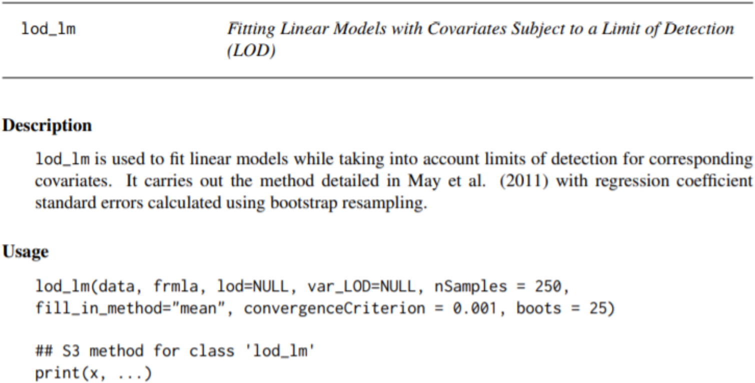
Sampling methods:

- Recall steps 2-6 involve sampling of unobserved biomarker values to compute Monte Carlo approximation
- Sampling is most computationally/time intensive portion of algorithm
- Options:
  1. Adaptive Rejection Metropolis Sampling (ARMS) [7], used by May et al. (2011)
  2. Slice sampling [8]

lodr Package

Before: no implementation available
Now: linear regression implementation with R package lodr

Computationally intensive sampling $\Longrightarrow$ C++ used for sampling, M step, and bootstrap



lod_lm *Fitting Linear Models with Covariates Subject to a Limit of Detection (LOD)*

Description

lod_lm is used to fit linear models while taking into account limits of detection for corresponding covariates. It carries out the method detailed in May et al. (2011) with regression coefficient standard errors calculated using bootstrap resampling.

Usage

```
lod_lm(data, frmla, lod=NULL, var_LOD=NULL, nSamples = 250,
fill_in_method="mean", convergenceCriterion = 0.001, boots = 25)

## S3 method for class 'lod_lm'
print(x, ...)
```

lodr Package

Package mimics lm function in R

Supporting functions:

- lod_lm
- summary.lod_lm
- coef.lod_lm
- residuals.lod_lm

lodr Package

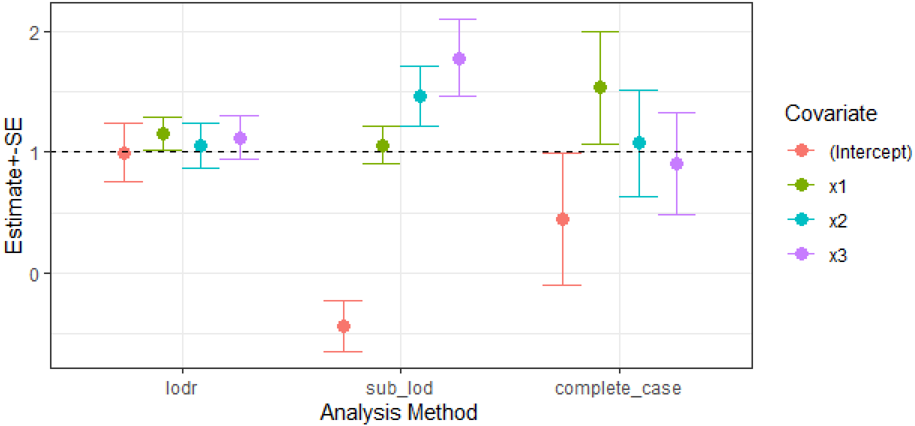
Simulation Example

$n = 100$; outcome variable Y with 3 covariates (X_1, X_2, X_3)
X_2 and X_3 subject to lower LODs of 0
All covariates generated from $N(0, 1)$

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$ where $\beta_0 = \ldots = \beta_3 = 1$

Simulation Results:

100 repetitions



Mean estimate and standard errors (SE) of regression analysis of simulated data across 100 simulations.

Dashed line indicates true parameter values

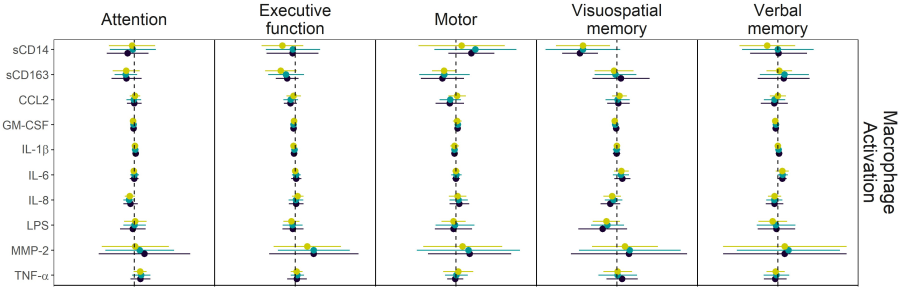
Method	Covariate	Mean Est.	Est. Bias	Mean SE	Mean DF
lodr	(Intercept)	1.00	-0.00	0.24	96.00
lodr	x1	1.15	0.15	0.13	96.00
lodr	x2	1.05	0.05	0.19	96.00
lodr	x3	1.12	0.12	0.18	96.00
sub_lod	(Intercept)	-0.44	-1.44	0.21	96.00
sub_lod	x1	1.06	0.06	0.16	96.00
sub_lod	x2	1.47	0.47	0.25	96.00
sub_lod	x3	1.78	0.78	0.32	96.00
complete_case	(Intercept)	0.44	-0.56	0.55	21.30
complete_case	x1	1.54	0.54	0.47	21.30
complete_case	x2	1.08	0.07	0.44	21.30
complete_case	x3	0.91	-0.09	0.42	21.30

ATN 071 Analysis

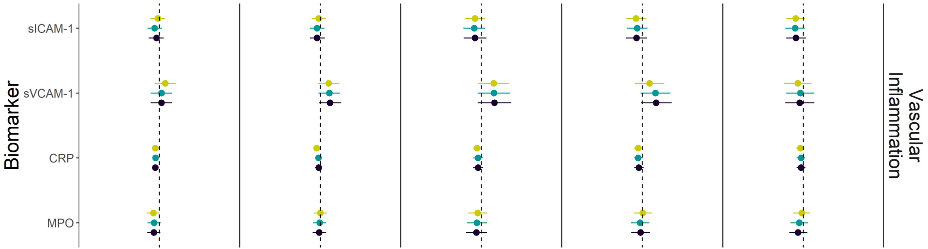
Recall: regression models run with neurocognitive composite as outcome, group biomarkers as covariates

Potential confounders: Age, gender, race, education, employment, income, recent substance use, depression symptoms, and concurrent infections

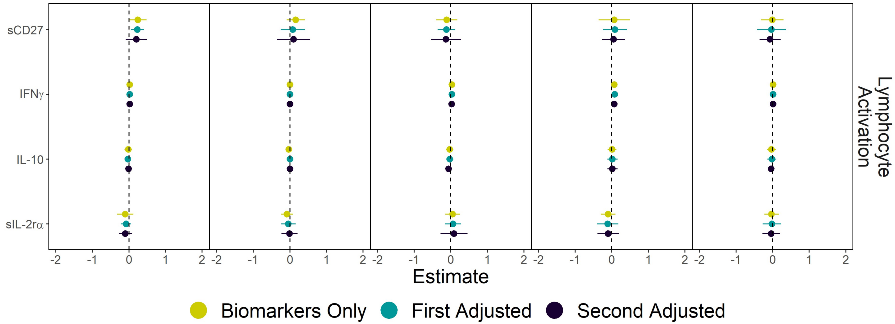
ATN 071 Analysis



ATN 071 Analysis



ATN 071 Analysis



Limitations and Future Research

- Method requires joint distribution of covariates specified $\Longrightarrow$ handling categorical covariates difficult **(ad hoc method applied)**
- Computational time for large datasets and simulation studies
- Implementation for generalized linear models in package
- Hypothesis testing and handling residuals

Bibliography I

1. Nie, L. *et al.* Linear regression with an independent variable subject to a detection limit. *Epidemiology (Cambridge, Mass.)* **21,** S17 (2010).
2. Lynn, H. S. Maximum likelihood inference for left-censored HIV RNA data. *Statistics in Medicine* **20,** 33–45 (2001).
3. Cole, S. R., Chu, H., Nie, L. & Schisterman, E. F. Estimating the odds ratio when exposure has a limit of detection. *International Journal of Epidemiology* **38,** 1674–1680 (2009).
4. Wu, H., Chen, Q., Ware, L. B. & Koyama, T. A Bayesian approach for generalized linear models with explanatory biomarker measurement variables subject to detection limit: an application to acute lung injury. *Journal of Applied Statistics* **39,** 1733–1747 (2012).
5. May, R. C., Ibrahim, J. G. & Chu, H. Maximum likelihood estimation in generalized linear models with multiple covariates subject to detection limits. *Statistics in Medicine* **30,** 2551–2561 (2011).

Bibliography II

6. Louis, T. A. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **44,** 226–233 (1982).
7. Gilks, W. R., Best, N. G. & Tan, K. Adaptive rejection Metropolis sampling within Gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **44,** 455–472 (1995).
8. Neal, R. M. Slice sampling. *Annals of Statistics,* 705–741 (2003).