Semiparametric generalized linear models with discrete (or continuous?) data: Bayesian implementation in Stata

2024 Stata Biostatistics and
Epidemiology Virtual Symposium

Paul Rathouz
Department of Population Health
Dell Medical School at the University of Texas at Austin
paul.rathouz@austin.utexas.edu

with

Entegar Alam
Department of Statistics and Data Science
University of Texas at Austin

Peter Mueller
Department of Statistics and Data Science
University of Texas at Austin

22 February 2024

Example: AHEAD Study

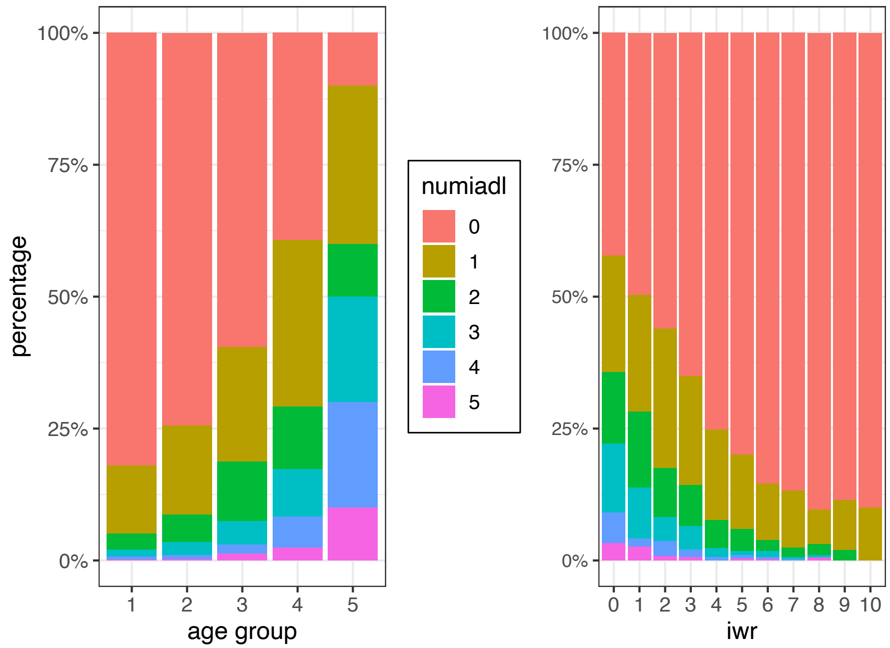
- Assets and Health Dynamics Among the Oldest Old
- National longitudinal study of individuals (and spouses/partners) aged $\geq$ 70 years
- Objectives:
  - monitor transitions in physical, functional, and cognitive health
  - study relationship of late-life changes in health to patterns of dissaving and income flows
- Baseline (complete) data from 1993, $n = 6,651$
- Models for:
  - instrumental activities of daily living
  - immediate word recall
  - mean of (scored) ordinal variable

AHEAD Variables: Baseline Wave

Variable	Description
numiadl	Number of instrumental activities of daily living tasks for which the subject has some difficulty, range: 0 to 5.
age	Age (years) at interview of the subject, range 70 to 103.
sex	Sex of subject ($1 =$ female, $0 =$ male).
iwr	Immediate word recall. Number of words out of 10 that subjects can list immediately after hearing them read. A measure of cognitive function.
netwc	Categorical values of net worth.

AHEAD Data: Two Strong Predictors (of numiadl)

Age and **Immediate Word Recall (iwr)**



Distribution of `numiadl`, AHEAD Data

`numiadl`	count	freq	cumul
0	4,915	73.90	73.90
1	1,099	16.52	90.42
2	362	5.44	95.87
3	169	2.54	98.41
4	69	1.04	99.44
5	37	0.56	100.00
Total	6,651	100.00	

As `numiadl` is skewed with an excess of zeros, suggest analysis with

- **Over-dispersed (quasi-Poisson) log-linear model for count data
- **Proportional odds (ordinal logistic) model for ordinal data

  **discussed in prior work (Rathouz and Gao, 2009)

- **A new SPGLM / GLDRM with log link**:

$$\log\{\mathrm{E}(Y|X;\beta)\} = X^{\mathrm{T}}\beta$$

Generalized Linear Quasilikelihood (QL) Models

Mean Model: For link $g(\cdot)$ and linear preditor η

$$\mathrm{E}(Y|X;\beta) = \mu(X,\beta) \equiv \mu \quad \text{with} \quad g(\mu) = \eta = X^{\mathrm{T}}\beta$$

Variance Model: For given X, variance of $(Y|X)$ is

$$\mathrm{var}(Y|X;\beta,\phi) = \phi v(\mu) \quad \leftarrow \quad \boxed{v(\mu) \text{ is variance model}}$$

In QL, β is orthogonal to ϕ

Interpretation of β does not depend on form of $v(\mu)$ or on ϕ

QL estimator $\widehat{\beta}$ will be CAN even in presence of:

- misspecification of $\mathrm{var}(Y|X;\beta,\phi)$
- poor estimation of $\mathrm{var}(Y|X;\beta,\phi)$

(although standard errors will be incorrect)

This is what is meant by a "working model"

Quasilikelihood (QL) Models (cont.)

- Broad class of mean regression models with high level of flexibility
  - linear predictor + link function w non-linear extensions
  - continuous, count, categorical outcomes
- QL estimation "works" (is consistent) if mean model is correct:
  - even if distributional model is wrong
  - even if variance model is wrong
- QL estimation:
  - efficient with correct standard errors when variance correct
  - empirical or "sandwich" or robust estimator when variance incorrect
- Practicality of QL with empirical variance $\longrightarrow$ advances in:
  - longitudinal data analysis
  - models for missing response and covariate data
  - models for covariates measured with error

Drawbacks of Quasilikelihood (QL) Mean Models

- No likelihood-based inferences
  - poor performance in small sample sizes
  - excessive reliance on sandwich estimator
- No inferences about cumulative response distribution
- Difficult to marry with latent-variable or random-effect models
- Application of Bayes' Theorem hampered:
  - posterior prediction of random effects
  - biased- or outcome-dependent sampling models
  - missing data models

An Alternative: A New Class of Semiparametric GLMs

Generalized Linear Density Ratio Model (GLDRM)

Mean Model: For link $g(\cdot)$ and linear preditor η

$$\mathrm{E}(Y|X;\beta) = \mu(X,\beta) \equiv \mu \quad \text{with} \quad g(\mu) = \eta = X^{\mathrm{T}}\beta \qquad (1)$$

Distributional Model: For given X, density of $(Y|X)$ is

$$f(y|X;\beta,f_0) = \frac{f_0(y)\exp(\theta y)}{\int_{\mathcal{Y}} f_0(u)\exp(\theta u)\ du} \quad \leftarrow \quad \boxed{\text{exponential tilting}}$$

where canonical parameter θ is implicitly defined to satisfy mean model (1)

That is, $\theta = \theta(\mu, f_0) = \theta(X, \beta, f_0)$

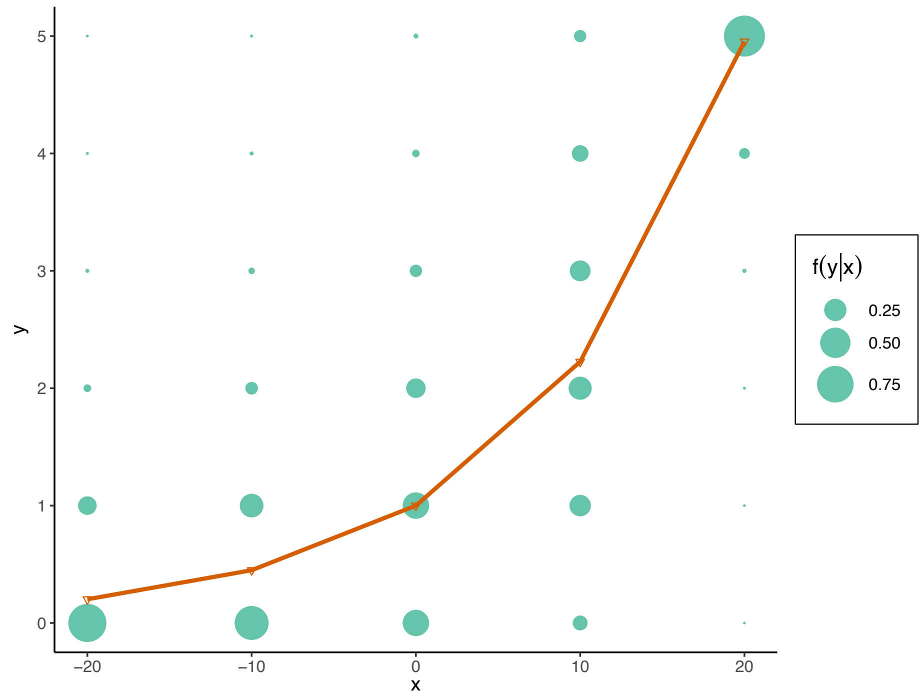
Key idea: Reference distribution $f_0(\cdot)$ is non-parametric, estimated with point mass on observed support for Y

Yields semi-parametric generalized linear model (SPGLM)

How Does this Tilting Work?

Tilting Redistributes mass according to a canonical parameter (θ) while maintaining the support of Y

Simulated example



Robustness and ML Estimation of β and f_0

- In GLDRM, β (or any model for μ) is orthogonal to f_0
- Interpretation of β does not depend on f_0
- For finite support (i.e., finite dimension f_0) ...
- ML estimator $\widehat{\beta}$ will be CAN even in presence of:
  - misspecification of f_0
  - poor estimation of f_0
  - misspecification of tilting model

  (although standard errors will be incorrect)
- Implication: Tilting model and f_0 form a "working model" for distribution of $f(Y|X)$ (as QL exploits a working model for the mean $\mathrm{E}(Y|X)$)
- Both β and f_0 admit Fisher score and information
- Suggest iterative ML estimation: $\widehat{\beta} \rightarrow \hat{f}_0 \rightarrow \widehat{\beta} \rightarrow \hat{f}_0 \cdots$

More Advantages to a Full Likelihood Model

- Full likelihood inferences (ML-SPGLM)
- Natural extension to Bayesian inference model using priors $\beta \sim N(\cdot, \cdot)$ and $f_0 \sim \mathrm{Dir}(\cdot)$ (Dir-SPGLM)
- Model for mean as well as full distribution (conditional on $X = x$), e.g., quantiles or exceedance probabilities

$$\Pr(Y \geq y | X = x, \beta, f_0) \quad \leftarrow \quad \boxed{\text{exceedance probability}}$$

- Model is easy to specify in some sense, plug-and-play (some object!)
- Let's see how it works with AHEAD

AHEAD: Fitted Log-linear ("Poisson") Model for Mean

For full data ($n = 6441$) and a small ($n = 100$) random sample

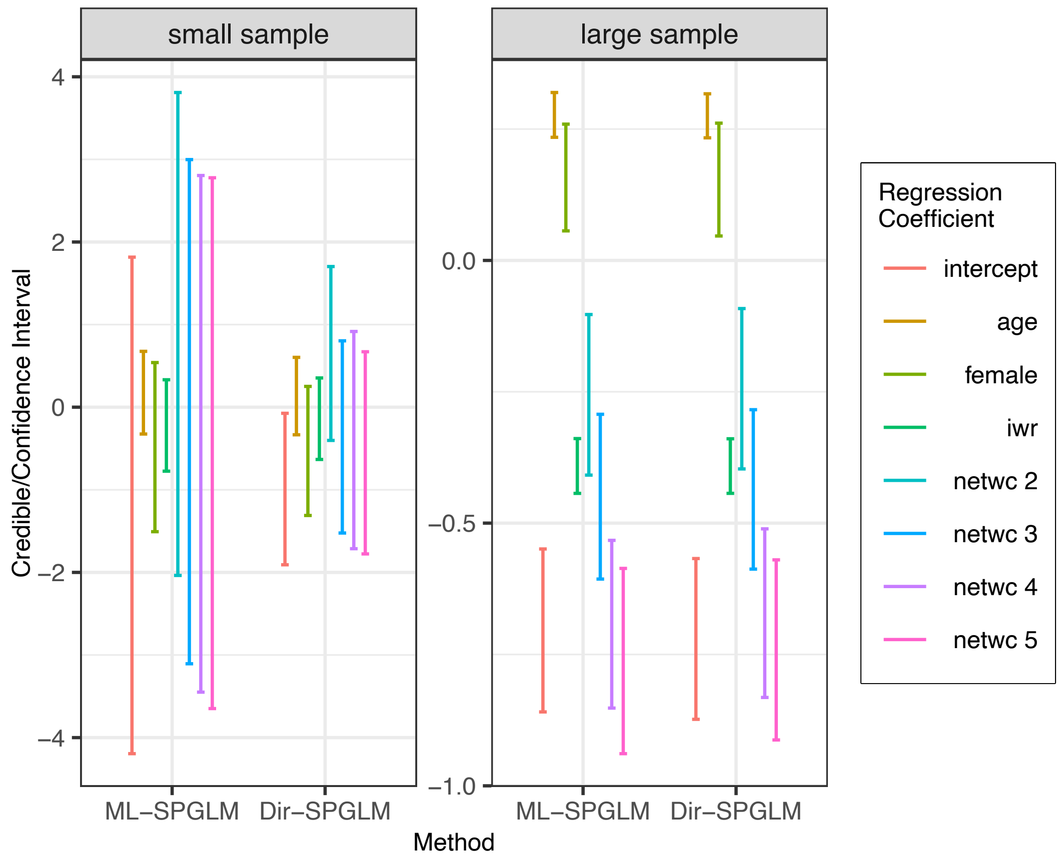
Using maximum likelihood (ML) and Bayesian MCMC inference with $10{,}000$ samples (including $3{,}000$ burn-in)

Mean model parameters have standard log-linear interpretation

Comparable results for large sample; Bayes more efficient for $n = 100$ (6 to 69% reduction in CI length)

Will examine bias in simulations

AHEAD: Fitted Log-linear ("Poisson") Model for Mean



Highlights of Current (ML) State

- Theory for both finite (ML) and infinite (SP ML) support (Note: Infinite support means continuous response)
- Good small sample performance (for mean / β parameters)
- Good computational performance for support cardinality k up to about $k = 1{,}500$
- Two current limitations
  - No good inferences for f_0 (except estimation)
  - Derived parameters (e.g., exceedance probabilities) are a challenge

A Bayesian Approach to Inference

Why?

- An alternative computational and inferential framework
- Develop inferences about reference distribution f_0
- Inferences about any derived model parameter, e.g. exceedance probabilities,

$$\Pr(Y \geq y | X = x, \beta, f_0) \quad \leftarrow \quad \boxed{\text{exceedance probability}}$$

or

$$\Pr(Y \geq y | X_{\text{age}} = \text{age}, \beta, f_0) \quad \leftarrow \quad \boxed{\text{average over other } X\text{'s}}$$

- Basis for (future) hierarchical modeling (random effects, latent variables)
- Allows principled answers to design questions (owing to unified model for data and parameters)

Goal for Today

Bayesian estimation and inference for case of finite support: Challenges and Results

Bayesian Inference Model

- Finite support case: $y \in \mathcal{Y} = \{s_1, \ldots, s_k\}$, where $s_l < s_{l+1}$ (we just use the observed (empirical) support)
- $f_0 \in \mathrm{simplex}(k-1)$
- Priors:
  - $\beta \sim N_p(0, \mathcal{I}_p)$
  - $f_0 \sim \mathrm{Dir}(\alpha H) \equiv \mathrm{Dir}(\alpha H(s_i), \ldots, \alpha H(s_k))$, where α is a user-specified concentration parameter
  - H is chosen to be the empirical frequency distribution of marginal y, so that
    * prior: $\mathrm{E}(f_0) = H$, and
    * average (over $\mathcal{Y}$) the mean $\mathrm{E}(f_0)$ distribution of the prior of f_0, is specified as $\mathrm{mean}(y)$.

A Special Problem: f_0 Is Actually an Equivalence Class

- We long-ago noted that the model as specified is not fully identified with respect to f_0

$$f(y|X;\beta,f_0) = \frac{f_0(y)\exp(\theta y)}{\int_{\mathcal{Y}} f_0(u)\exp(\theta u)\; du} \quad \leftarrow \quad \boxed{\text{exponential tilting}}$$

- Problem: Can replace given $f_0(y)$ with any $\widetilde{f}_0(y) \propto f_9(y)\exp(\widetilde{\theta}y)$, in which case model becomes

$$f(y|X;\beta,f_0) = \frac{\widetilde{f}_0(y)\exp\{(\theta-\widetilde{\theta})y\}}{\int_{\mathcal{Y}} \widetilde{f}_0(u)\exp\{(\theta-\widetilde{\theta})u\}\; du} \ ,$$

  so θ is just replaced with $(\theta-\widetilde{\theta})$

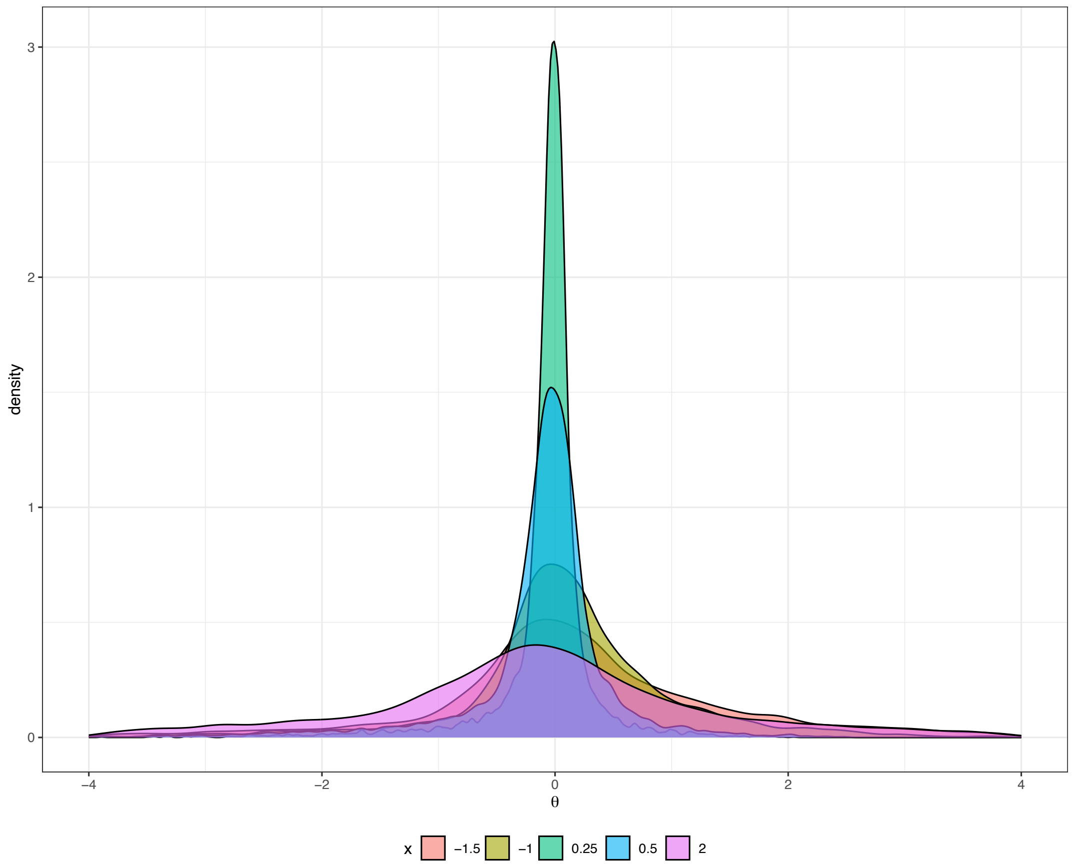
- Under ML, we solve this problem by pre-specifing f_0 to yield some given mean, μ_0 (default is empirical marginal mean of y)

f_0 Is Actually an Equivalence Class (cont.)

- Viewed differently, f_0 is an equivalence class of all exponential tilts of a given (or, in Bayesian case, sampled) "index" f_0
- In our Bayesian MCMC approach, we solve this problem by:
  - specifying Dirichlet H to be the empirical distribution of y
  - after all MCMC samples are generated, tilting each posterior $f_0(y)$ to be $f_0^*(y)$ with mean μ_0, the empirical mean of y
- Additional note: Priors on (β, f_0) induce a prior on $\theta = \theta(x, \beta, f_0)$ for a given x:

  If $g(\cdot)$ and μ_0 are chosen such that, as $||x||_2 \to 0$,
  - $\mu = g^{-1}(\eta) \xrightarrow{P} \mu_0$,
  - then, $\theta \xrightarrow{P} 0$,
  - and, (scaled at same rate as x) θ is asymptotically normal, as in picture (next slide)

Induced Prior on θ (just so you know . . .)



Highlights (some technicalities) of Posterior Simulation

- MCMC posterior simulation with Metropolis-Hastings (MH) transition probabilities
- Each β_j and each $f_0(s_k)$ are updated one at a time
- β_j's use a random walk proposal using inverse FI matrix
- $f_0(s_k)$'s use a random walk proposal based on weighted empirical distribution of y, which essentially retilts (untilts!) each observation back to $\theta_i = 0$) given current θ_i

Bayesian Implementation in Stata

Bayesian Implementation in Stata
Progress

- Target: -bayesmh- with the -llevaluator()- option
- Likelihood, and related calculations, in mata
  - Normal prior for coefficients β: equation {resp:}
  - Reference distribution f_0, coded as a second equation:
    {f0:f01},...,{f0:f0k}
  - Dirichlet prior for f_0:
    prior({f0:}, dirichlet(2,2,2,2,2))
  - Identity, log, and (generalized) logit link functions

Bayesian Implementation in `Stata`
Challenges

- Recall the distributional model is:

$$f(y|X;\beta,f_0) = \frac{f_0(y)\exp(\theta y)}{\int_{\mathcal{Y}} f_0(u)\exp(\theta u)\ du} \quad \leftarrow \quad \boxed{\text{exponential tilting}}$$

  where canonical parameter θ is implicitly defined to satisfy the mean

$$g^{-1}(X^{\mathrm{T}}\beta) = \mu = \frac{\int_{\mathcal{Y}} u f_0(u)\exp(\theta u)\ du}{\int_{\mathcal{Y}} f_0(u)\exp(\theta u)\ du}$$

  – Function `-getTheta()-` programmed in `mata`

  – Requires careful (stressful) handling of boundaries and large values

- Above, integrals replaced with sums over finite support given by parameters in equation {`f0:`}

  Open question: How to handle when the support gets large?

Simulation Investigations

Simulation Investigations

Compare: Dir-SPGLM to ML-SPGLM for data on support $\{0, \ldots, 5\}$

Examine: regression parameters (β)

bias and (relative) efficiency of estimates

coverage probabilities for confidence / credible intervals

Examine: reference distribution parameters (f_0)

bias and (relative) efficiency of estimates

credible interval coverage probablities (new inferences)

Simulation Investigations (cont.)

Data generating mechanisms: $X_1 \sim N(0,1)$

$$\log(\mu) = \eta = \beta_0 + \beta_1 X_1$$

- $f_0 =$ truncated Poisson(1) on $\{0, 1, \ldots, 5\}$
- $f_0 =$ 0-inflated truncated Poisson(1) on $\{0, 1, \ldots, 5\}$ with $3\times$ the mass at $y = 0$

Here: $n = 25$ (also did $n = 250$), $2{,}000$ replicates.

Dir-SPGLM MCMC $10{,}000$ posterior samples, discarding the first $3{,}000$ and using the remaining $7{,}000$ for inference

Simulation Results: β Inferences

n	Scenario	Parm	Method	Truth	Est$_a$	RRMSE$_a$	RL$_a$	Est$_m$	RRMSE$_m$	RL$_m$	CP
25	1	β_0	ML-SPGLM	-0.7	-0.78	1.00	1.00	-0.74	1.00	1.00	0.97
			Dir-SPGLM		-0.76	0.82	0.91	-0.74	0.89	0.93	0.97
		β_1	ML-SPGLM	0.2	0.20	1.00	1.00	0.19	1.00	1.00	0.93
			Dir-SPGLM		0.16	0.79	0.92	0.17	0.83	0.93	0.97
	2	β_0	ML-SPGLM	-0.7	-0.81	1.00	1.00	-0.77	1.00	1.00	0.97
			Dir-SPGLM		-0.77	0.75	0.87	-0.75	0.85	0.90	0.97
		β_1	ML-SPGLM	0.2	0.18	1.00	1.00	0.18	1.00	1.00	0.94
			Dir-SPGLM		0.14	0.73	0.86	0.14	0.79	0.88	0.97
250	1	β_0	ML-SPGLM	-0.7	-0.71	1.00	1.00	-0.71	1.00	1.00	0.97
			Dir-SPGLM		-0.71	1.00	0.99	-0.71	0.98	0.99	0.96
		β_1	ML-SPGLM	0.2	0.20	1.00	1.00	0.20	1.00	1.00	0.96
			Dir-SPGLM		0.20	0.99	0.99	0.20	1.00	0.98	0.96
	2	β_0	ML-SPGLM	-0.7	-0.71	1.00	1.00	-0.71	1.00	1.00	0.97
			Dir-SPGLM		-0.71	0.98	0.99	-0.71	0.97	0.99	0.96
		β_1	ML-SPGLM	0.2	0.20	1.00	1.00	0.20	1.00	1.00	0.96
			Dir-SPGLM		0.20	0.97	0.98	0.20	0.96	0.98	0.96

Simulation Results: f_0 Inferences

n	Scenario	Parm	Method	Truth	Est_a	RRMSE_a	Est_m	RRMSE_m	CP
25	1	Scenario 1 results better than Scenario 2							
	2	$f_0(0)$	ML-SPGLM	0.471	0.397	1.00	0.409	1.00	N/A
			Dir-SPGLM		0.458	0.58	0.461	0.66	0.89
		$f_0(1)$	ML-SPGLM	0.232	0.288	1.00	0.259	1.00	N/A
			Dir-SPGLM		0.246	0.58	0.240	0.83	0.88
		$f_0(2)$	ML-SPGLM	0.172	0.236	1.00	0.240	1.00	N/A
			Dir-SPGLM		0.178	0.60	0.164	0.58	0.91
		$f_0(3)$	ML-SPGLM	0.085	0.070	1.00	0.062	1.00	N/A
			Dir-SPGLM		0.076	0.58	0.070	0.38	0.88
		$f_0(4)$	ML-SPGLM	0.031	0.007	1.00	0.000	1.00	N/A
			Dir-SPGLM		0.029	0.77	0.019	0.51	0.90
		$f_0(5)$	ML-SPGLM	0.009	0.000	1.00	0.000	1.00	N/A
			Dir-SPGLM		0.010	1.44	0.006	0.46	0.93

Simulation Investigations: Conclusions for Small Sample Sizes

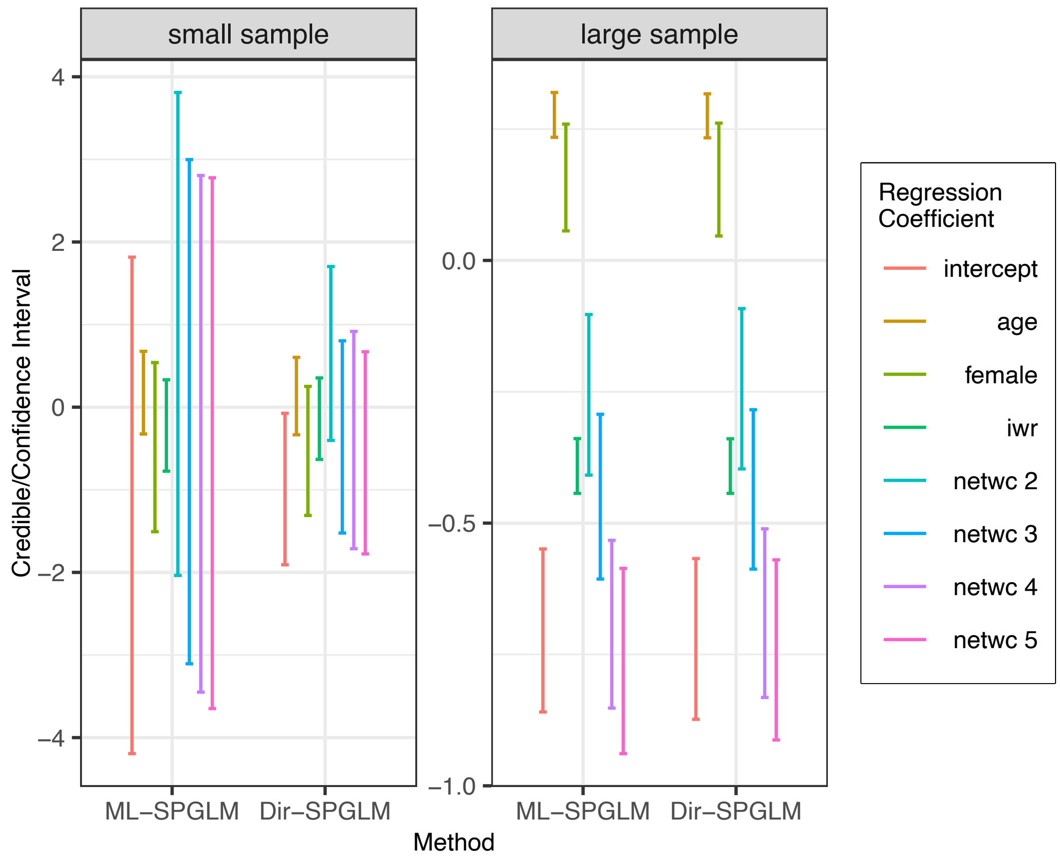
- For small sample sizes, Dir-SPGLM exhibits comparable bias with increased efficiency vs ML-SPGLM
- Confidence / credible interval coverage for β values comparable and acceptable
- New inferences: Credible interval coverage for f_0 acceptable
- (Not shown) Exceedance value inferences (see AHEAD)

Return to AHEAD

AHEAD Study: Estimation and Predictive Inferences

- Already seen comparable results (Dir-SPGLM vs ML-SPGLM) for regression coefficients β (reminder next slide)
- How about $\text{reference distribution } f_0$ in estimation (large sample) and prediction (small training sample) modes?

AHEAD: Fitted Log-linear ("Poisson") Model for Mean



AHEAD: Full Data Inferences on f_0: ML vs Dir

TSS	Par	Method	Est	CI
Large	$f_0(0)$	ML-SPGLM	0.725	N/A
		Dir-SPGLM	0.725	[0.719, 0.731]
	$f_0(1)$	ML-SPGLM	0.187	N/A
		Dir-SPGLM	0.186	[0.176, 0.196]
	$f_0(2)$	ML-SPGLM	0.059	N/A
		Dir-SPGLM	0.059	[0.054, 0.064]
	$f_0(3)$	ML-SPGLM	0.022	N/A
		Dir-SPGLM	0.022	[0.019, 0.025]
	$f_0(4)$	ML-SPGLM	0.006	N/A
		Dir-SPGLM	0.006	[0.005, 0.008]
	$f_0(5)$	ML-SPGLM	0.002	N/A
		Dir-SPGLM	0.002	[0.001, 0.002]

AHEAD Study: Training to Test: Predictive Inference

Training data: Randomly sample $n = 100$; fit model; test on remaining $n = 6{,}341$

Exceedance probabilities: ML-SPGLM

$$p(y \geq y_0 \mid x) \mathrel{\widehat{=}} p(y \geq y_0 \mid x; \beta^{(\text{mle})}, f_0^{(\text{mle})})$$

And for Dir-SPGLM (posterior mean)

$$p(y \geq y_0 \mid x) \mathrel{\widehat{=}} (1/B) \sum_{b=1}^{B} p(y \geq y_0 \mid x, \beta^{(b)}, f_0^{(b)})$$

Set $y_0 = 2$ (moderate) and $y_0 = 4$ (severe)

ROC and AUC(ROC): from held out test data

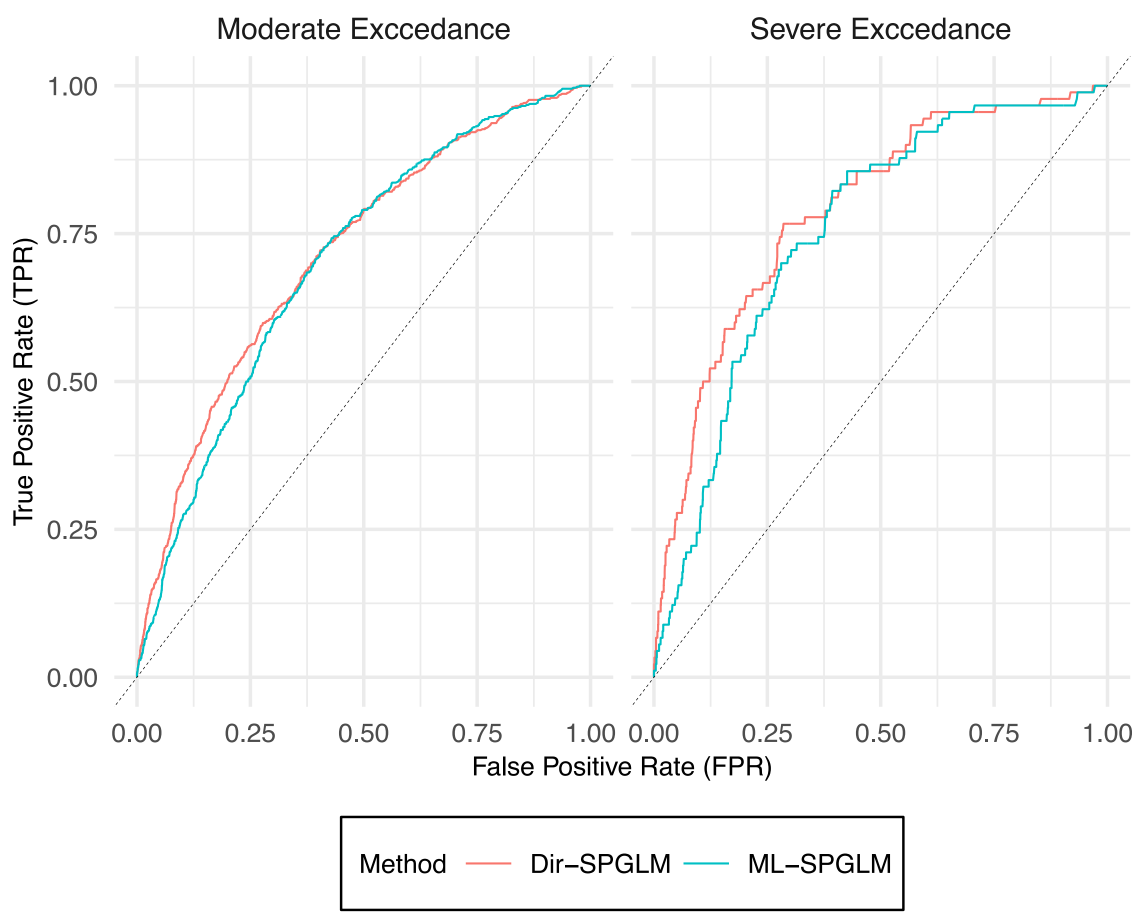
AHEAD: Small ($n = 100$) Training Inferences on f_0: ML vs Dir

TSS	Par	Method	Est	CI
Small	$f_0(0)$	ML-SPGLM	0.805	N/A
		Dir-SPGLM	0.815	[0.764, 0.859]
	$f_0(1)$	ML-SPGLM	0.136	N/A
		Dir-SPGLM	0.126	[0.069, 0.198]
	$f_0(2)$	ML-SPGLM	0.023	N/A
		Dir-SPGLM	0.021	[0.004, 0.057]
	$f_0(3)$	ML-SPGLM	0.021	N/A
		Dir-SPGLM	0.020	[0.003, 0.044]
	$f_0(4)$	ML-SPGLM	0.009	N/A
		Dir-SPGLM	0.010	[0.001, 0.028]
	$f_0(5)$	ML-SPGLM	0.006	N/A
		Dir-SPGLM	0.009	[0.001, 0.026]

AHEAD Study: Predictive Inference Results

AUC: $y_0 = 2$: 0.70 (ML-SPGLM); 0.71 (Dir-SPGLM)

AUC: $y_0 = 4$: 0.75 (ML-SPGLM); 0.79 (Dir-SPGLM)



AHEAD Study: Predictive Inference Uncertainty

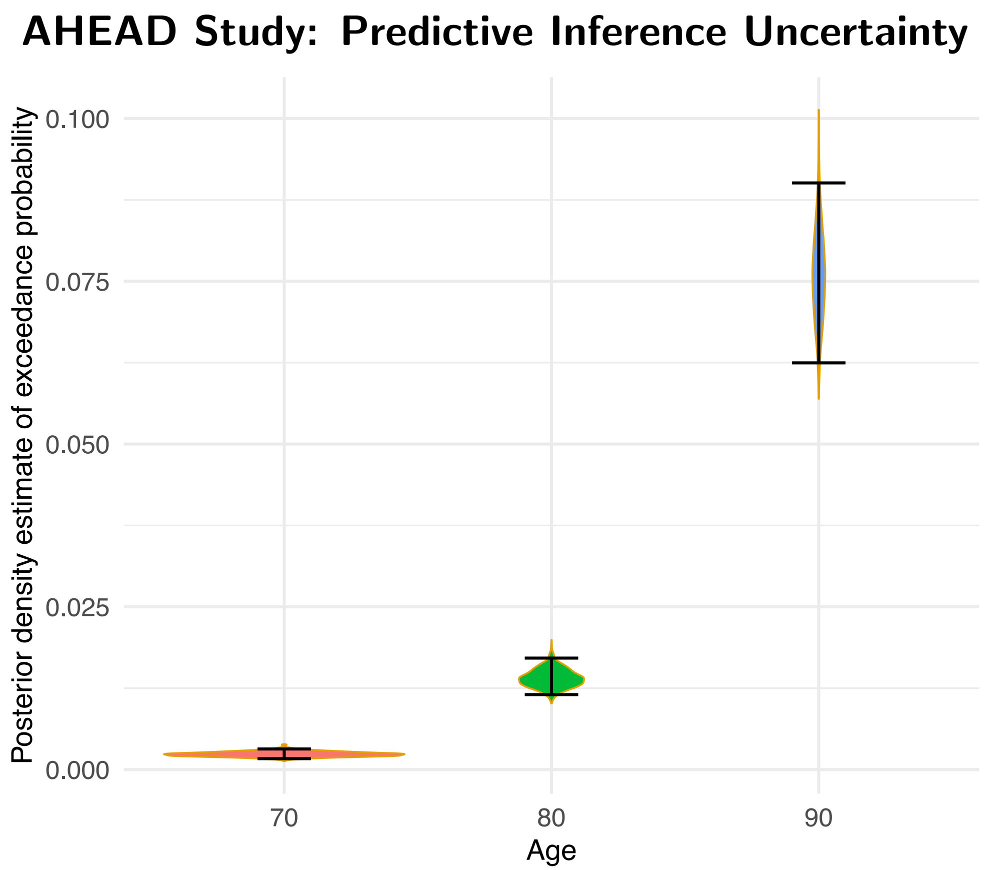
- Consider exceedance probabilty for $y_0 = 4$, for $\{x : x_{\mathrm{age}} = a\}$, and where we imagine the design is fixed, and $a \in \{70, 80, 90\}$ years

- Thus, the bth posterior value is

$$p\left(y \geq y_0 \mid x_{age} = t, \beta^{(b)}, f_0^{(b)}\right) = \frac{\sum\limits_{x:x_{age}=t} p\left(y \geq y_0 \mid x, \beta^{(b)}, f_0^{(b)}\right)}{\sum_x 1_{\{x:x_{age}=a\}}(x)} \, ,$$

- From the posterior distribution across $b = \{1, \ldots, B\}$, we obtain point esimates (e.g., median) and credible intervals (e.g,. at 2.5^{th} and 97.5^{th} percentiles)

- This would be difficult with ML owing to need to use the delta method or similar to leverage the joint sampling distribution of $(\widehat{\beta}, \widehat{f}_0)$



As expected, far more uncertainty where there is less covariate (x) data

Conclusions and Future Directions

Conclusions and Future Directions

- The SPGLM provides a flexible, full-likelihood alternative to the classic GLM family that has good small sample properties and comparable inferential performance to the QL family
- But, inferentially, this is restricted to mean model parameters (β)
- To extend inferential scope and to prepare for latent variable and other hierarchical models, we have introduced a Bayesian, Dirichlet-prior driven model that permits
  - inference on the reference distribution (f_0), and
  - on functionals of (β, f_0) such as exceedance probabilities and (later) quantiles (as functions of covariates x)

  that were not possible earlier

- Immediate next steps are to handle continuous responses using the Dirichlet Process Prior (DPP), work which we have undertaken already
- Random effects and other latent variable models for clustered or longitudinal responses
- As a tool planning clinical trials, allowing for uncertainty in that process.
  - Incorporate loss functions
  - Focus on a a "high" or "low" group for planning using exceedance probabilities
  - Use `Stata`'s bulit in Bayesian random effects structures

Funding and References

This research was supported by the National Institutes of Health Grant 2 R01 HL094786.

Rathouz and Guo (2009). *Biostatistics.*
Huang and Rathouz (2012). *Biometrika.*
Huang and Rathouz (2017). *Comm. Stat.*
Huang (2014). *JASA.* (includes discussion of PIT)
Wurm and Rathouz (2018). *R Journal*
(This work) Entejar Alam, Peter Muller, and Paul J. Rathouz. Dir-SPGLM: A Bayesian semiparametric GLM with data- driven reference distribution. 2024.