Semiparametric Generalized Linear Models

North American Stata Users' Group Meeting
Chicago, Illinois

Paul Rathouz
Department of Health Studies
University of Chicago
prathouz@uchicago.edu

Liping Gao
MS Student in Statistics

July 24, 2008

Stata Software Development

Masha Kocherginsky
Department of Health Studies
University of Chicago

Philip Schumm
Department of Health Studies
University of Chicago

Example: AHEAD Study

- Assets and Health Dynamics Among the Oldest Old
- National longitudinal study of individuals (and spouses/partners) aged ≥ 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, $n = 6441$
- Models for:
  - instrumental activities of daily living
  - immediate word recall

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.

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 model for ordinal data

Review: Log-linear and Proportional Odds Models

- Log-linear model:

$$\log\{\mathrm{E}(Y|X;\beta)\} = \log(\mu) = \beta_0 + X^{\mathrm{T}}\beta_{\mathrm{LL}}$$

$$\mathrm{var}(Y|X:\beta,\phi) = \phi\mu$$

  Rest of distribution (higher moments) are **unspecified**

  Interpretation: $\beta_{\mathrm{LL}} \quad \longrightarrow$ log ratio of **means**

- Proportional odds model:

$$\mathrm{logit}\{\Pr(Y \geq c;\alpha,\beta)\} = \alpha_c + X^{\mathrm{T}}\beta_{\mathrm{PO}}, \quad \alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_C$$

  for $Y \in \{0,1,\ldots,c,\ldots,C\}$

  Distribution is **fully-specified**

  Interpretation: $\beta_{\mathrm{PO}} \quad \longrightarrow$ log ratio of **cumulative odds**

Fitted log-linear and proportional odds models for `numiadl`, AHEAD Data

	LLM						
	Poisson		Quasi		POM		
	$\widehat{\beta}$	$\widehat{\mathrm{se}}(\widehat{\beta})$	$\widehat{\mathrm{se}}(\widehat{\beta})$	Z	$\widehat{\beta}$	$\widehat{\mathrm{se}}(\widehat{\beta})$	Z
(Intercept)	-3.62	0.277	0.337	-10.69	–	–	–
age	0.05	0.003	0.004	12.74	0.07	0.005	12.71
`sex`:female	0.16	0.043	0.052	3.04	0.26	0.064	4.01
`iwr`	-0.21	0.012	0.014	-14.72	-0.26	0.018	-14.42
`netwc`:1-24k	-0.26	0.063	0.077	-3.28	-0.45	0.113	-4.01
`netwc`:25k-74k	-0.46	0.065	0.079	-5.62	-0.65	0.111	-5.81
`netwc`:75k-199k	-0.69	0.067	0.081	-8.50	-0.93	0.110	-8.46
`netwc`:200k-up	-0.76	0.074	0.090	-8.48	-0.92	0.116	-7.90
log-Likelihood	-5179.6				-4951.5		
Scale			1.48				

AHEAD Data: log-linear and proportional odds models for number of IADL difficulties

- Log-linear model:
  - regression coefficients have convenient interpretation as the **log-ratio of mean** number of IADL difficulties corresponding to unit differences in covariates
  - valid quasi-likelihood inferences, but no likelihood function
- Proportional odds model:
  - similar conclusions as the log-linear model
  - regression coefficients have less-convenient interpretation as **log odds ratios** for "high" versus "low" number of IADL difficulties
  - but, likelihood inferences obtain

Generalized Linear (GL) and Quasilikelihood (QL) Models

- Broad class of mean regression models with high level of flexibility
  - linear predictor
  - link function
  - 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 model correct
  - empirical or "sandwich" variance estimator valid when variance model incorrect

- Practical power of QL with empirical variance estimation has lead to advances in:
  - longitudinal data analysis
  - models for missing and covariate data
  - models for covariates measured with error

Drawbacks of Quasilikelihood Mean Models

- No likelihood-based inferences
- 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

Example: Outcome Dependent Sampling

- $S = I(\text{unit sampled into study})$ or $S = I(\text{unit has complete data})$
- Suppose known or estimable: $p(S = 1|Y, X)$
- Bayes' Theorem:

$$f(Y|X, S = 1) = \frac{f(Y|X)p(S = 1|Y, X)}{\int f(u|X)p(S = 1|u, X)\ du}$$

- Difficult if $f(Y|X)$ specified as QL model; easy if $f(Y|X)$ a fully-specified probability model

Alternative Approach: Ordinal Data Models

- Proportional odds (POM) or ordinal probit models
- Fully-specified probability models (likelihood inferences)
- Easily combined with random effects / latent variables
- Semi-parametric specification (baseline odds function estimated, not assumed)
- However:
  - regression coefficients are for log cumulative odds (not mean)
  - more difficult for applied audiences to grasp
  - tying to graphical data presentations more difficult
- Desired:

> A regression model parameterized in terms of the **mean response**, with similar **level of flexibility** as the POM

Outline

- ✓ AHEAD data example
- A new class of GLMs
  - flexibility similar to POM
  - parametric model for mean response:
    - linear predictor ($\eta = X^{\mathrm{T}}\beta$)
    - link function
  - non-parametric baseline distribution (when $\eta = 0$)
  - response distribution for $\eta \neq 0$ via exponential tilting
- Some model properties
- Simulations including comparison to the POM
- Return to AHEAD data examples

Notation and Basic Model

Data: Y = scalar response on support $\mathcal{Y} \subset \mathcal{R}$

X = predictor vector ($p \times 1$)

Mean Model:

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

for known (user-specified), strictly monotone **link** $g(\cdot)$ mapping $(m, M) \subset \mathcal{R}$ into $\mathcal{R}$, where $m = \inf(\mathcal{Y})$, and $M = \sup(\mathcal{Y})$

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

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

where θ is a function of μ and $f_0(\cdot)$ is a **baseline density** on $\mathcal{Y}$

Idea: Estimate both β **and** f_0 from data ... but first fix f_0 ...

Model Given Fixed Baseline Density $f_0(\cdot)$

- The model

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

can be re-written as

$$f(y|X;\beta,f_0) = \exp\{\theta y - b(\theta) + \log f_0(y)\},$$

where

$$b(\theta;f_0) \equiv b(\theta) = \log \int_{\mathcal{Y}} f_0(u)\exp(\theta u)\ du,$$

- For fixed f_0, this is a **natural exponential family model** with:
  - canonical parameter θ
  - cumulant generating function $b(\cdot)$
- In particular, $\mathrm{var}(Y|X;\beta,f_0) = b''(\theta)$

Fixed Baseline Density $f_0(\cdot)$ (cont.)

- Combining the **distributional model**

$$f(y|X;\beta,f_0) = \exp\{\theta y - b(\theta) + \log f_0(y)\},$$

  with the **mean regression model**

$$\mathrm{E}(Y|X;\beta) = g^{-1}(\eta) = g^{-1}(X^{\mathrm{T}}\beta),$$

  this becomes a **generalized linear model** with linear predictor η, link function $g(\cdot)$ and error distribution $f(y|X;\beta,f_0)$

- **Special cases of Baseline Density $f_0(\cdot)$:**
  - Bernoulli data (n trials): f_0 is Bin$\{n,(1/2)\}$
  - Poisson data: f_0 is Poi(1)

Canonical Link Function for Fixed f_0

- $f(y|X;\beta,f_0)$ has mean μ and canonical parameter θ
- Induces **canonical link function** $g_c(\cdot)$ such that

$$g_c(\mu;f_0) \equiv g_c(\mu) = \theta \quad \forall \mu \in (m,M),$$

  depending in general on f_0

- Because

$$\mu = \mathrm{E}(Y|X) = b'(\theta) = \frac{\int_{\mathcal{Y}} y f_0(u)\exp(\theta u)\ du}{\int_{\mathcal{Y}} f_0(u)\exp(\theta u)\ du}\ ,$$

  $g_c(\cdot)$, as an **implicit function** of μ, is the solution in θ to $b'(\theta)=\mu$

- With regularity conditions, $g_c(\mu;f_0)$ **exists** and is a **unique mapping** from (m,M) onto $(-\infty,+\infty)$

Robustness and ML Estimation of f_0

- In SPGLM, β is **orthogonal** to f_0
- **Interpretation** of β does not depend on 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 $f(Y|X)$
- Both β and f_0 admit Fisher score and information
- Suggest iterative ML estimation: $\widehat{\beta} \rightarrow \hat{f}_0 \rightarrow \widehat{\beta} \rightarrow \hat{f}_0 \cdots$
- Yields a **semiparametric generalized linear model** (SPGLM)

SPGLM versus Proportional Odds Model (POM)

- **Semi-parametric models**:
  - finite-dimensional regression model in β $(p \times 1)$
  - non-parametric baseline density f_0
- Same number of parameters (similar **level of flexibility**):
  - $p - 1$ slope parameters capturing effects of X
  - card$(\mathcal{Y}) - 1$ baseline density parameters
- **Stochastic ordering**: Suppose for given $x_1 \neq x_2$ that

$$x_1^{\mathrm{T}}\beta = \eta_1 < \eta_2 = x_2^{\mathrm{T}}\beta$$

then for all $y \in \mathcal{Y}$ such that $m < y < M$,

$$\Pr(Y \leq y|X = x_1) > \Pr(Y \leq y|X = x_2)$$

- Model choice? Analytic goals, personal preferences

Outline

- ✓ AHEAD data example
- ✓ A new class of GLMs
  - flexibility similar to POM
  - parametric model for mean response:
    - linear predictor ($\eta = X^{\mathrm{T}}\beta$)
    - link function
  - non-parametric baseline distribution (when $\eta = 0$)
  - response distribution for $\eta \neq 0$ via exponential tilting
- ✓ Some model properties
- Simulations including comparison to the POM
- Return to AHEAD data examples

Simulation Study

Compare: log-linear model (LLM), SPGLM with log-link, POM

Examine: regression parameter tests and estimators

likelihood values

cdf estimation

Data generating mechanisms: $X_1 \sim N(0,1)$, $\mathrm{E}(Y) \approx 0.5$

SPGLM: $\eta = \beta_0 + \beta_1 X_1$

- $f_0 =$ 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 $y = 0$

POM with $\eta = \beta_1 X_1$ and 0-inflated truncated Poisson(1) on $\{0, 1, \ldots, 5\}$ as baseline distribution

1st Result: β estimation identical under LLM, SPGLM

Simulation results for Type I error, power and maximum likelihood values

True f_0	Model	Type I Error	Power	logL (se)
Truncated	SPGLM	0.056	0.62	-229.3 (11.6)
Poisson	LLM	0.055	0.61	-231.0 (11.7)
	POM	0.049	0.56	-229.6 (11.6)
0-inflated	SPGLM	0.047	0.47	-235.3 (14.0)
Poisson	LLM	0.091	0.58	-245.9 (15.5)
	POM	0.042	0.41	-235.5 (14.1)
POM	SPGLM	0.047	0.62	-227.7 (11.7)
	LLM	0.091	0.62	-229.3 (11.7)
	POM	0.042	0.66	-227.4 (11.7)

Notes: 1000 replicates, $n = 250$

SPGLM: $\beta_1 = 0.2$; POM: $\beta_1 = 0.3$

Simulation results for cdf estimation

True f_0	Model	$\widehat{\Pr}(Y > 1 \mid X = 0)$ est. (se)	$\widehat{\Pr}(Y > 3 \mid X = 0)$ est. (se)
Truncated	True	0.0892	0.0017
Poisson	SPGLM	0.0877 (0.017)	0.0017 (0.0007)
	LLM	0.0889 (0.013)	0.0018 (0.0006)
	POM	0.0907 (0.018)	0.0022 (0.0028)
0-inflated	True	0.1258	0.0073
Poisson	SPGLM	0.1241 (0.021)	0.0073 (0.0024)
	LLM	0.0900 (0.016)	0.0018 (0.0007)
	POM	0.1281 (0.021)	0.0080 (0.0055)
POM	True	0.0892	0.0017
	SPGLM	0.0839 (0.017)	0.0016 (0.0007)
	LLM	0.0876 (0.013)	0.0017 (0.0005)
	POM	0.0881 (0.018)	0.0016 (0.0025)

Simulation Study: Conclusions

- SPGLM and the Poisson LLM are similar in terms of bias and efficiency
- More accurate standard errors with the SPGLM
- SPGLM "automatically" accounts for over-dispersion
- SPGLM and POM have similar log-likelihood values, Type I errors and power and so would be comparable data analysis options in applications
- SPGLM more stable in estimation of tails of baseline cdf? Further study needed

AHEAD Variables: Baseline Wave (reminder slide)

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: Log-linear Models for `numiadl`

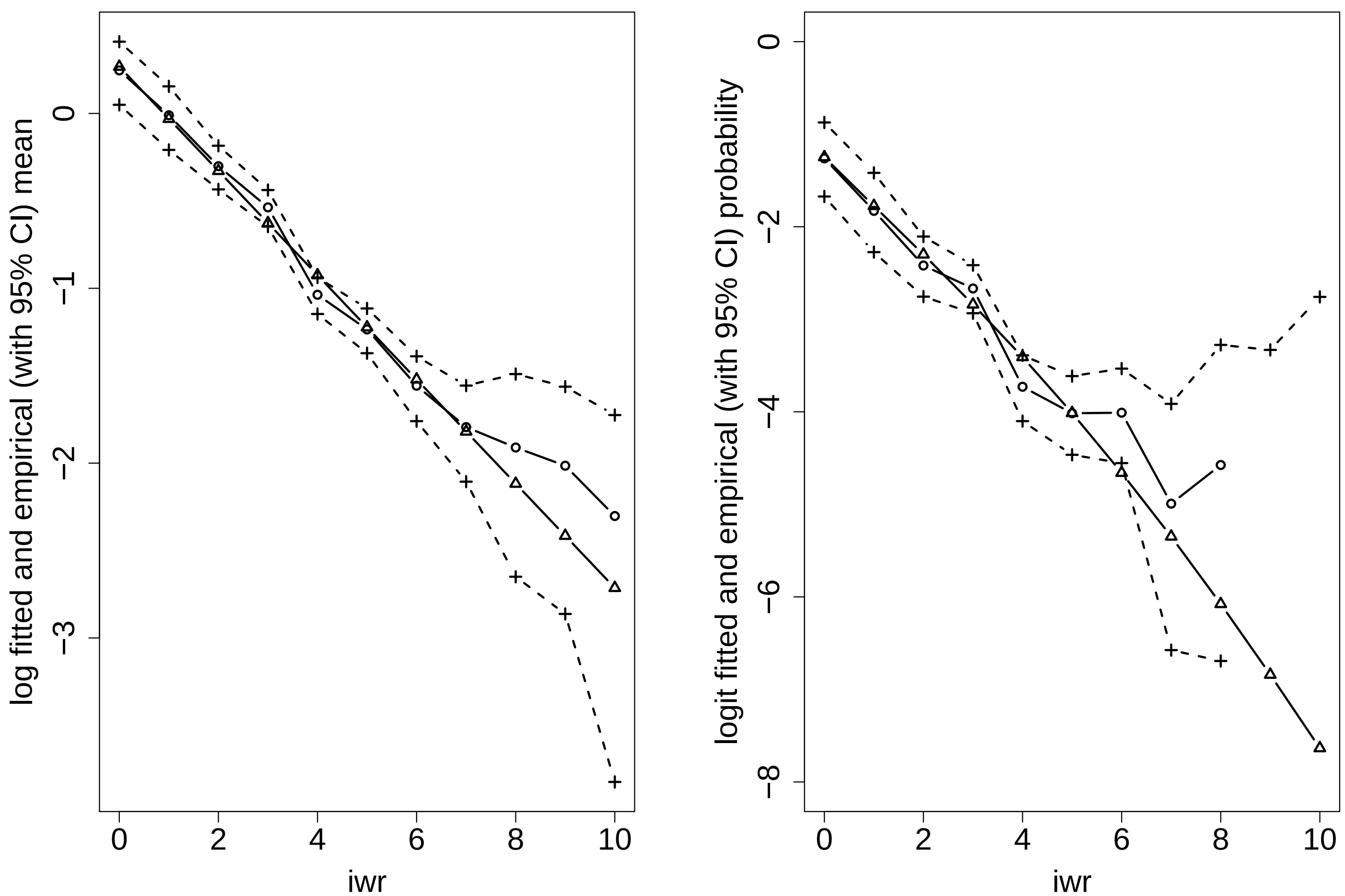
`numiadl`	count	freq	cumul
0	4,915	73.90	73.90
1	1,099	16.52	90.42
2	362	5.44	95.87
3	169	2.54	98.41
4	69	1.04	99.44
5	37	0.56	100.00
Total	6,651	100.00	

- Log-linear models under Poisson, over-dispersed Poisson (quasi-Poisson) and SPGLM
- Proportional odds model (POM)

Fitted log-linear and proportional odds models for `numiadl`, AHEAD Data

	SPGLM			LLM		POM		
				Pois.	Quasi			
	$\widehat{\beta}$	$\widehat{\mathrm{se}}(\widehat{\beta})$	Z	$\widehat{\beta}$	$\widehat{\mathrm{se}}(\widehat{\beta})$	$\widehat{\beta}$	$\widehat{\mathrm{se}}(\widehat{\beta})$	Z
(Intercept)	-3.61	0.337	-10.69	-3.62	0.337	–	–	–
`age`	0.05	0.004	12.74	0.05	0.004	0.07	0.005	12.71
`sex`:female	0.12	0.052	3.04	0.16	0.052	0.26	0.064	4.01
`iwr`	-0.21	0.014	-14.72	-0.21	0.014	-0.26	0.018	-14.42
`netwc`:1-24k	-0.26	0.078	-3.28	-0.26	0.077	-0.45	0.113	-4.01
`netwc`:25k-74k	-0.45	0.080	-5.62	-0.46	0.079	-0.65	0.111	-5.81
`netwc`:75k-199k	-0.69	0.081	-8.50	-0.69	0.081	-0.93	0.110	-8.46
`netwc`:200k-up	-0.76	0.090	-8.48	-0.76	0.090	-0.92	0.116	-7.90
log-Likelihood		-4951.2		-5179.6			-4951.5	
Scale					1.48			

AHEAD: Fitted values for log-linear model for `numiadl` as a function of `iwr`: Mean and $\Pr(\texttt{iadl} \geq 3)$



AHEAD: Log-linear Models for `numiadl`

- Extremely close estimates and standard errors under SPGLM and quasi-Poisson model fits
- Likelihood values for SPGLM and POM are equivalent
- Hypothesis tests for effects of predictors on `numiadl` under SPGLM and POM are very comparable
- SPGLM fitted mean and CDF as a function of `iwr` very good
- **Conclusion:**

> From data perspective, SPGLM and POM are equally appropriate likelihood-based approaches to modelling these data, the main difference between the two being in the interpretation of the regression coefficients

AHEAD: Logistic-linear Models for `iwr`

- `iwr` is number of successes out of 10 trials
- Logistic-linear models under Binomial, quasi-Binomial and SPGLM

`iwr`	count	freq	cumul
0	154	2.39	2.39
1	195	3.03	5.42
2	526	8.17	13.58
3	1,001	15.54	29.13
4	1,450	22.51	51.64
5	1,355	21.04	72.68
6	954	14.81	87.49
7	445	6.91	94.40
8	196	3.04	97.44
9	105	1.63	99.07
10	60	0.93	100.00
Total	6,441	100.00	

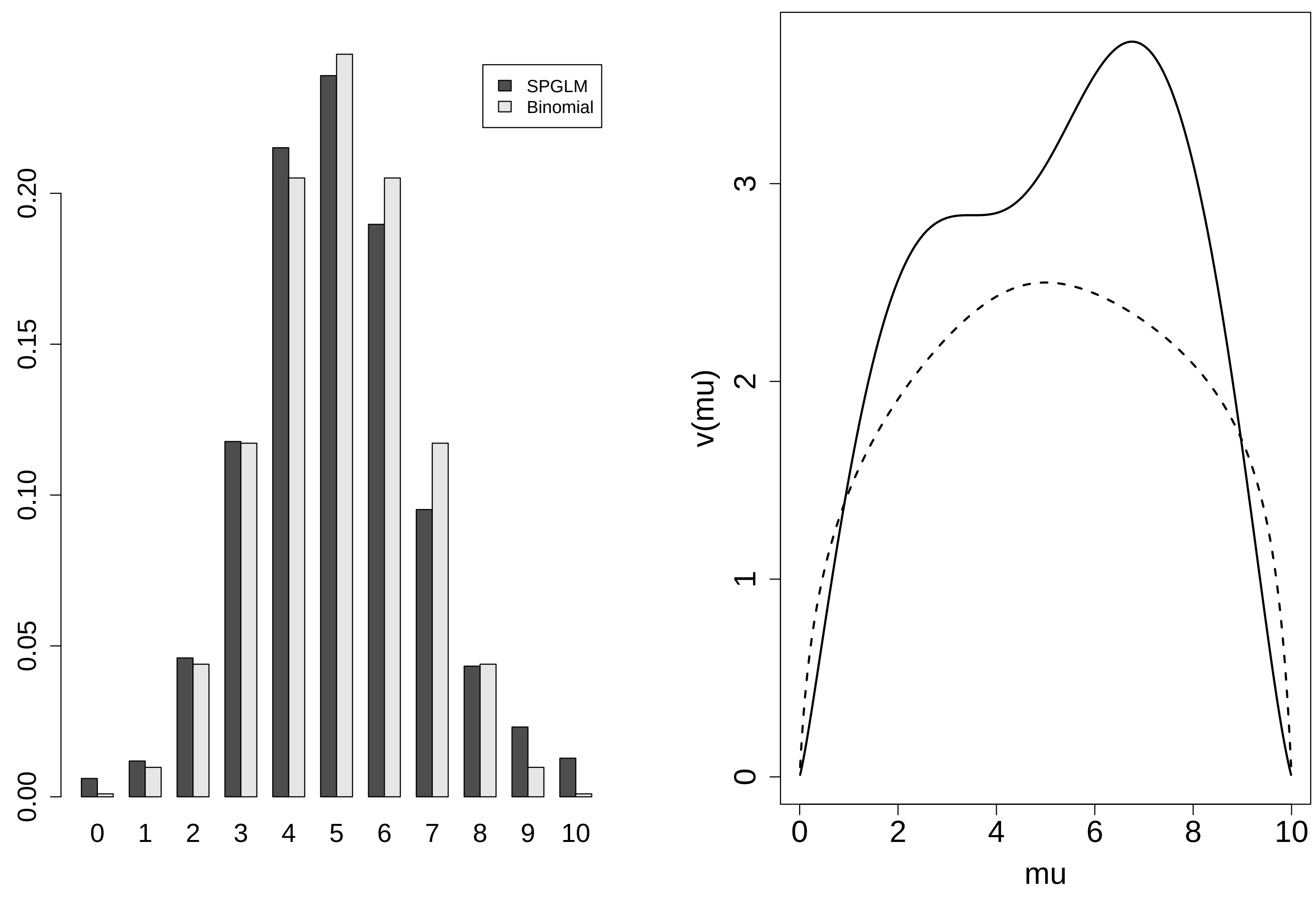
Logistic-linear models for `iwr`, AHEAD Data

	SPGLM			Logistic-linear				
				Binomial			Quasi	
	$\widehat{\beta}$	$\widehat{\mathrm{se}}(\widehat{\beta})$	Z	$\widehat{\beta}$	$\widehat{\mathrm{se}}(\widehat{\beta})$	Z	$\widehat{\mathrm{se}}(\widehat{\beta})$	Z
(Intercept)	2.22	0.134	16.54	2.22	0.120	18.50	0.134	16.58
`age`	-0.04	0.002	-23.53	-0.04	0.001	-26.48	0.002	-23.74
`sex`:female	0.21	0.019	11.44	0.21	0.017	12.70	0.019	11.38
`netwc`:1-24k	0.28	0.041	6.75	0.28	0.037	7.50	0.041	6.72
`netwc`:25k-74k	0.39	0.040	9.67	0.39	0.036	10.83	0.040	9.71
`netwc`:75k-199k	0.55	0.039	14.16	0.55	0.034	15.89	0.038	14.24
`netwc`:200k-up	0.69	0.040	17.32	0.69	0.035	19.51	0.039	17.49
log-Likelihood		-12552			-12812			
Scale							1.25	

AHEAD: Logistic-linear Models for `iwr`

- SPGLM and quasi-Binomial yield extremely close results
- Likelihood suggests SPGLM fits substantially better than Binomial ($X^2 = 520$ on $K - 2 = 9$ df)
- Compare fitted f_0 and Binomial f_0
- Compare fitted variance functions $v(\mu) = b''\{g_c(\mu; f_0)\}$ under two models

Fitted $\hat{f}_0$ and variance function for log-logistic models for `iwr`, AHEAD Data



Summary

- A new class of GLMs for $(Y|X)$
- A user-specified parametric mean function
- Unspecified (non-parametric) reference distribution
- Similar mean models and inferences as commonly-used over-dispersed GLMs
- Comparable level of flexibility to the popular proportional odds model
- Better of both worlds (we hope!)

Aspirations for the Class of SPGLM Models

- A flexible alternative to QL models for mean response when full distribution is desirable but difficult to specify
- Modeling framework on which to build random effects or other latent variable models
- Methods for missing data and biased samples
- Extension to infinite support case

Extra Slides

Related Literature: Estimating $f_0(\cdot)$ with the data?

- When using tilting model

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

  for **multi-group** analysis (as in 1-way ANOVA):

  - each group j gets own θ_j (and own mean μ_j)
  - $f_0(\cdot)$ estimated from the data

- Then $f(y|X;\theta,f_0)$ is called a **density ratio model** (DRM)

- **Proposal:** Expand DRM to more general regression spaces via a user-specified regression model $g^{-1}(X^{\mathrm{T}}\beta)$ for μ, while still estimating f_0 from the data

- New model: **generalized linear density ratio model** (my first name) or **semiparametric generalized linear model**

Maximum Likelihood Estimation of SPGLM (sketch)

- Both β and f_0 admit Fisher score and information
- Orthogonality of β and f_0 suggest iterative estimation:

$$\widehat{\beta} \;\rightarrow\; \hat{f}_0 \;\rightarrow\; \widehat{\beta} \;\rightarrow\; \hat{f}_0 \;\cdots$$

- Constraints on f_0: (μ_0 an arbitrary reference mean)

$$f_0(y) \geq 0 \;\forall y \in \mathcal{Y}\;, \quad \sum_{y \in \mathcal{Y}} f_0(y) = 1\;, \quad \text{and} \quad \sum_{y \in \mathcal{Y}} y f_0(y) = \mu_0$$

- Complication in f_0 estimation: $\theta = g_c(\mu; f_0)$ depends on f_0 !
  - yields an extra term in f_0 score
  - an inconvenience when support $\mathcal{Y}$ is **finite**
  - open problem when $\mathcal{Y}$ is **infinite**: "Is MLE $\hat{f}_0$ restricted to observed support (as in, e.g., the Cox PH model)?"

Simulation results for β estimation under SPGLM data generating mechanisms and LLM and SPGLM models

		Mean		RMSE		CP	
True f_0	Model	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_0$	$\widehat{\beta}_1$
	True β	-0.7	0.2	-	-	-	-
Truncated Poisson	SPGLM	-0.708	0.199	0.090	0.087	0.957	0.959
	LLM	-0.707	0.199	0.090	0.086	0.956	0.959
0-inflated Poisson	SPGLM	-0.703	0.200	0.107	0.103	0.947	0.949
	LLM	-0.703	0.200	0.107	0.102	0.904	0.921

Notes: 1000 replicates, $n = 250$

Simulation results for f_0 estimation under SPGLM data generating mechanisms and models

	Truncated Poisson		0-inflated Poisson	
Support	True f_0	Bias (se)	True f_0	Bias (se)
0	0.367	-0.004 (0.030)	0.471	-0.005 (0.028)
1	0.368	0.003 (0.037)	0.232	0.003 (0.031)
2	0.185	0.002 (0.039)	0.172	0.005 (0.035)
3	0.062	0.002 (0.025)	0.085	0.001 (0.027)
4	0.016	-0.002 (0.017)	0.031	-0.002 (0.020)
5	0.003	-0.001 (0.009)	0.009	-0.002 (0.012)