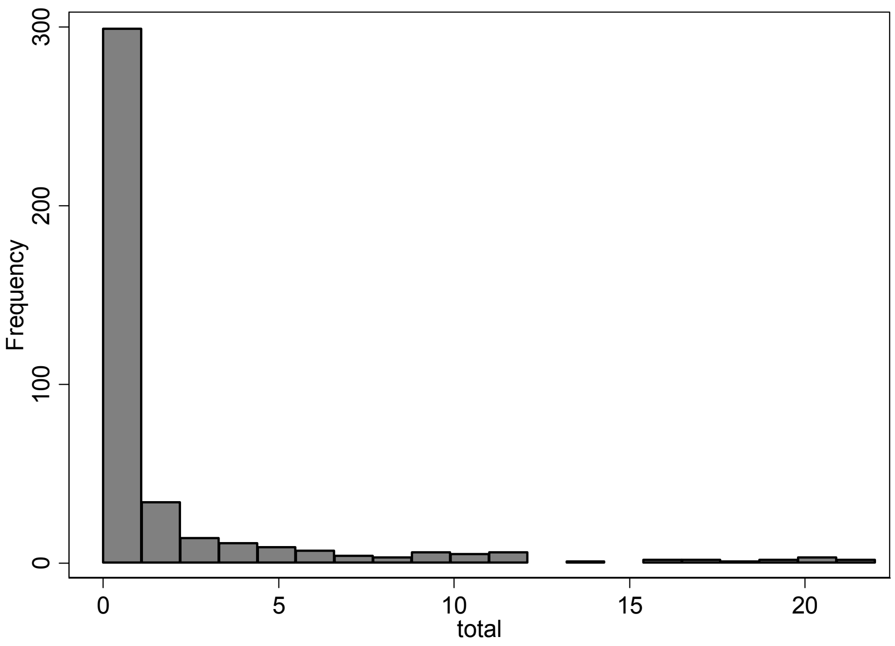
Zeros

Modeling with a preponderance of zeros

Frauke Kreuter, UCLA Statistics

Data

- Characteristics
  - Large number of zeros
  - Skewed
- Examples
  - Convictions
  - Drinking
  - Drug abuse

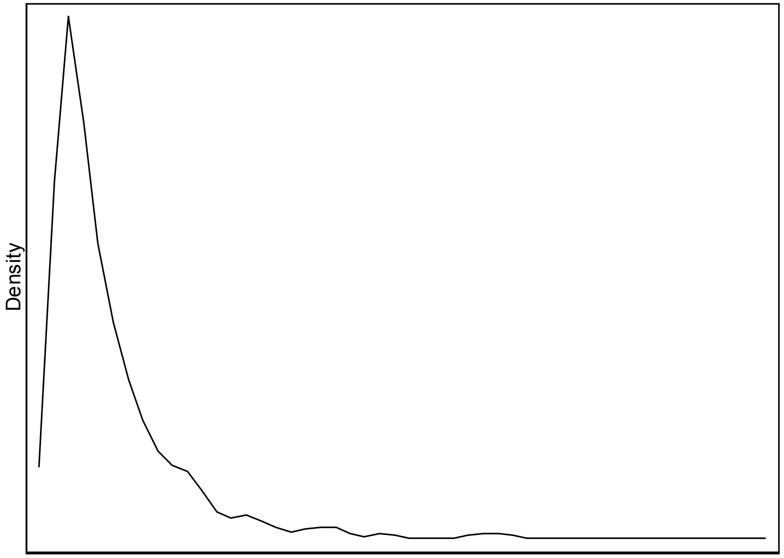


“Classic” examples

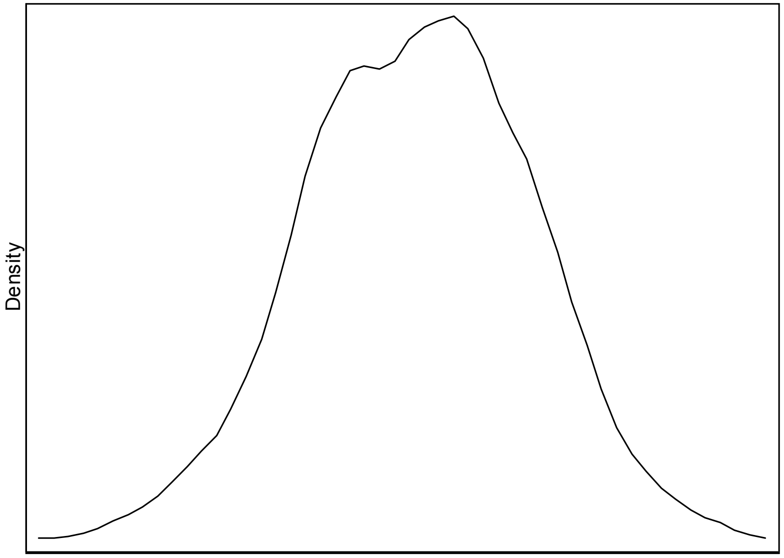
- Manufacturing applications
- Economics
- Medicine
- Public health
- Environmental science
- Education

Distributions

- Skewed



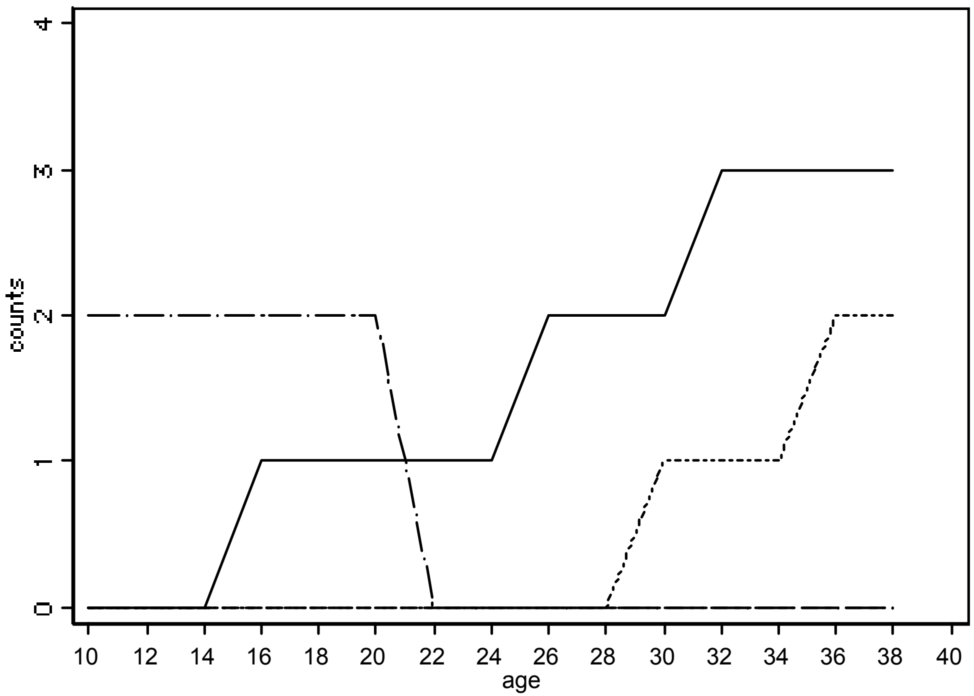
- Log transformed



Modeling strategies

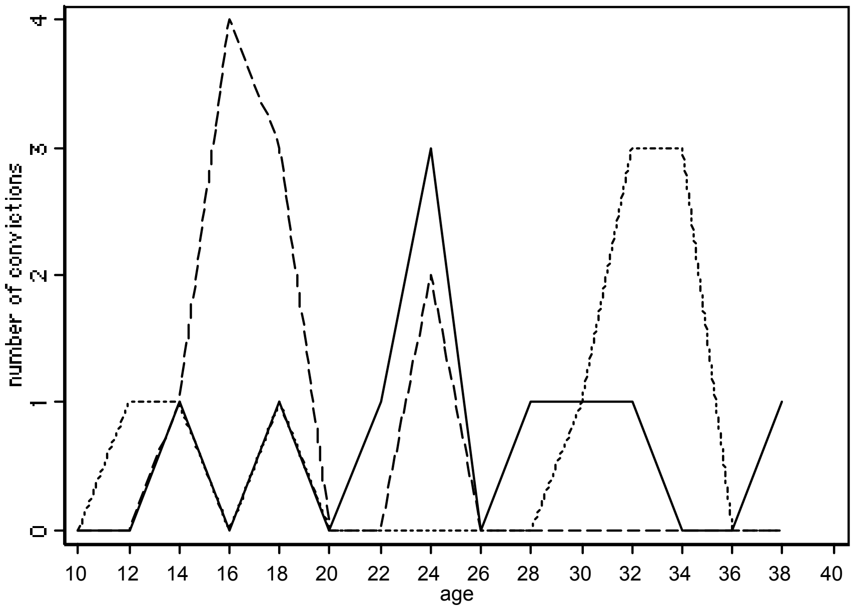
- Continuous
  - Censored normal
  - Two-part model
  - Mixture models
    - Zero inflated models
    - Two-class model
- Counts
  - Two-part modeling
  - Zero inflated models
  - Mixture models
    - Zero inflated models
    - Two-class model

Zeros in longitudinal data



- Onset
  - Zero until onset
  - Once behavior shown we want to model counts
- "Offset"
  - Zero after a certain point
  - E.g. abstinence, "jail time"

Zeros in longitudinal data



- Criminal careers
  - In and out of zero (random zeros);
    - Example: Solid line Respondent 1
  - Measurement errors
  - Zero throughout (partly structural zeros)
    - Here 60% 0 age10-40

Separate process for zeros

Example – Olsen & Schafer (2001)

- Data n=1961
  - Panel of Adolescent Prevention Trial
  - Middle school and high school students
  - Grade 7 trough 11
- Variables
  - Self reported recent alcohol use
  - Parental monitoring, rebelliousness, gender

- Model: two-parts
  - U-part
    - Logit
    - use, no-use
  - Y-part
    - Log-normal
    - y>0

Two-part model

$$u_{it} = \begin{cases} 1 & if & y_{it} & > & 0 \\ 0 & if & y_{it} & = & 0 \end{cases}$$

$$y_{it} = \begin{cases} m_{it} & if & y_{it} & > & 0 \\ . & if & y_{it} & = & 0 \end{cases}$$

$$m_{it} = \eta_{0i} + \eta_{1i} a_{it} + \varepsilon_{it}$$

Raw data

i	*Grade 7*	*Grade 8*	*Grade 9*	*Grade 10*	*Grade 11*
1	0	0	0	0	0
2	0	0	1.7	2.3	3
3	0	1	0	1	1.7

U-part

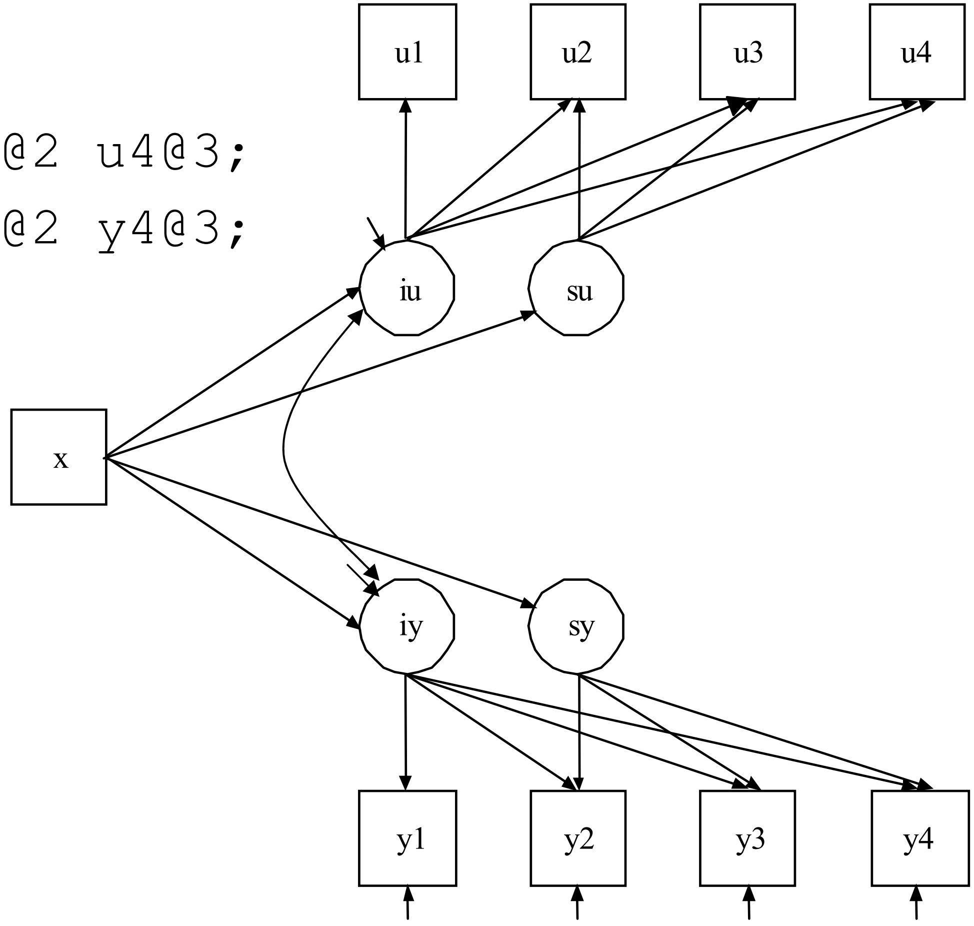
i	*Grade 7*	*Grade 8*	*Grade 9*	*Grade 10*	*Grade 11*
1	0	0	0	0	0
2	0	0	1	1	1
3	0	1	0	1	1

Y-part

i	*Grade 7*	*Grade 8*	*Grade 9*	*Grade 10*	*Grade 11*
1	.	.	.	.	.
2	.	.	1.7	2.3	3
3	.	1	.	1	1.7

Two part modeling - M*plus*

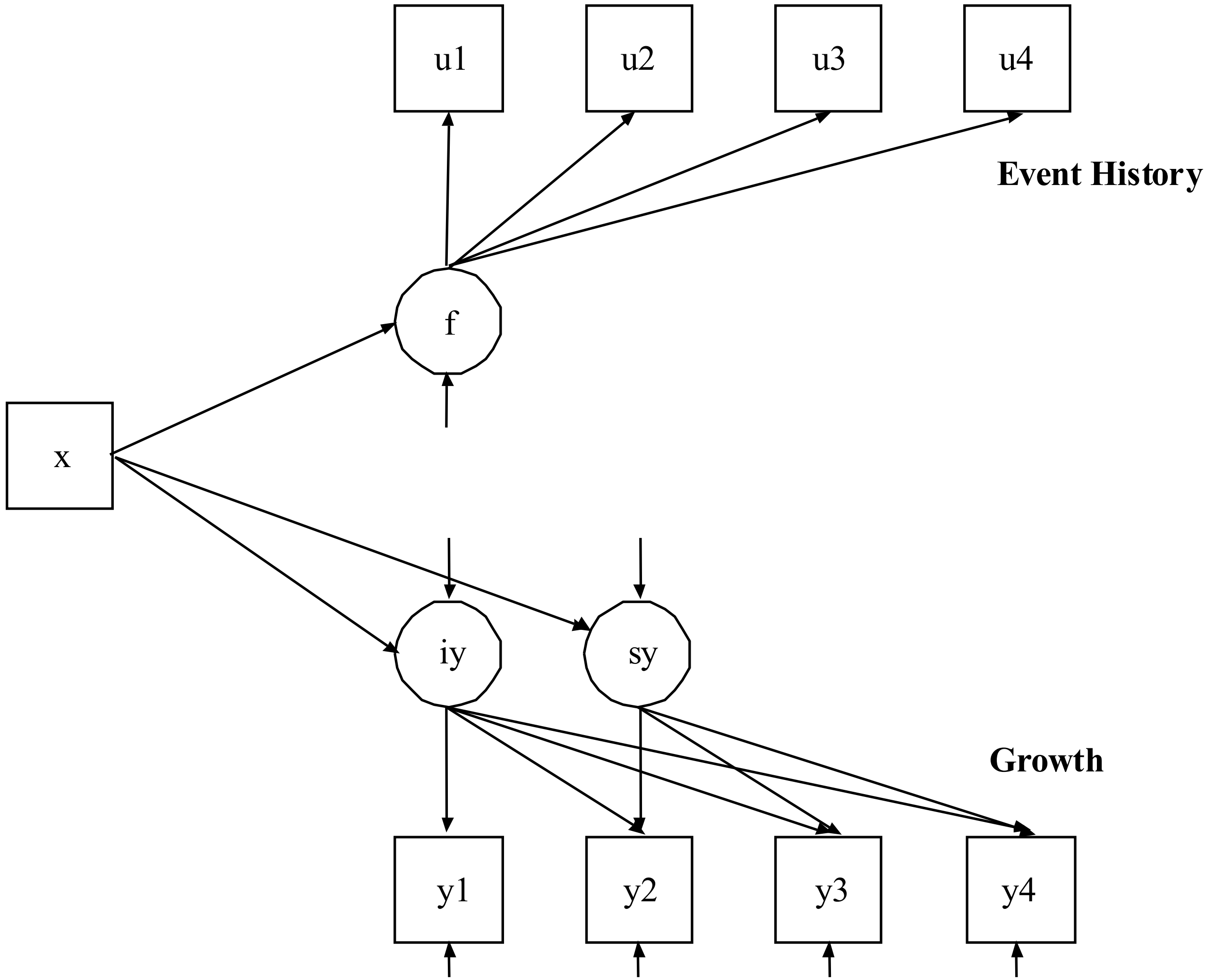
```
MODEL:
  iu su | u1@0 u2@1 u3@2 u4@3;
  iy sy | y1@0 y2@1 y3@2 y4@3;

  iu-sy on x;

  su@0;   sy@0;
  iu with su@0;
  iy with sy@0;
  iu with sy@0;
  su with iy-sy@0;
```



Example – Olsen & Schafer (2001)

- U-part
  - In grade 7 unsupervised girls higher odds for drinking, effect diminishes over time
  - Low monitoring in grade 7 no effect for boys, but unsupervised boys higher odds in grade 11
- Y-part
  - Reduced monitoring increase the amount of alcohol consumption in grade 7
  - For girls effect increases over time, for boys it vanishes by grade 11

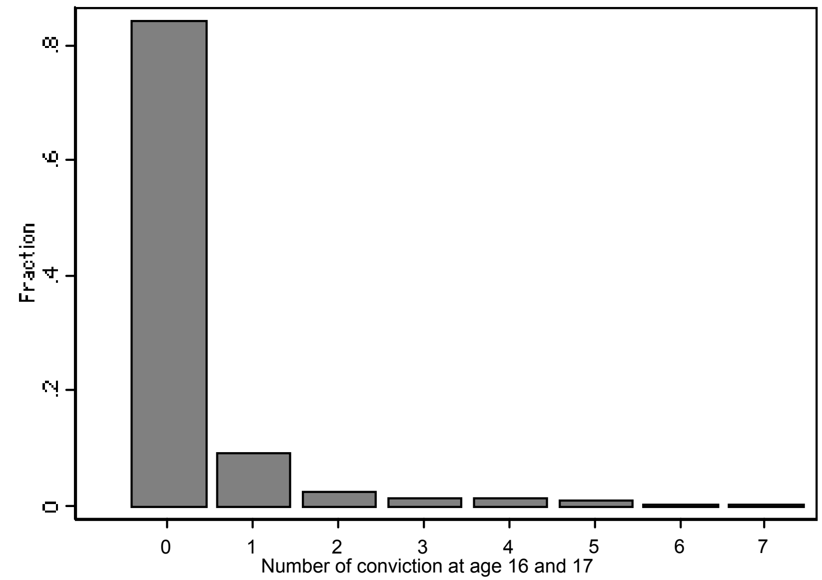
Example – Albert & Shih (2003)



Separate sources for zeros

Modeling count variables (t=1)

- Models used
  - Poisson
    - With parameter λ
    - Use GLM to model log(λ)
  - Negative binomial



- Problem
  - Zero inflation / overdispersion
  - Model assumptions don’t hold

Zero Inflated Poisson (ZIP)

- Zero outcome can arise from one of two sources, one where outcome is always zero, another where a poisson process is at work (Lambert 1992)
- The poisson process can produce zero or another outcome
- Covariates can predict group membership, and outcome of the poisson process

The model

$$\Pr(y_{it}=0)=\Pr(group1)+\Pr(y_{it}=0\,|\,group2)*\Pr(group2)$$

$$\Pr(y_{it}=1)=\Pr(y_{it}=1\,|\,group2)*\Pr(group2)$$

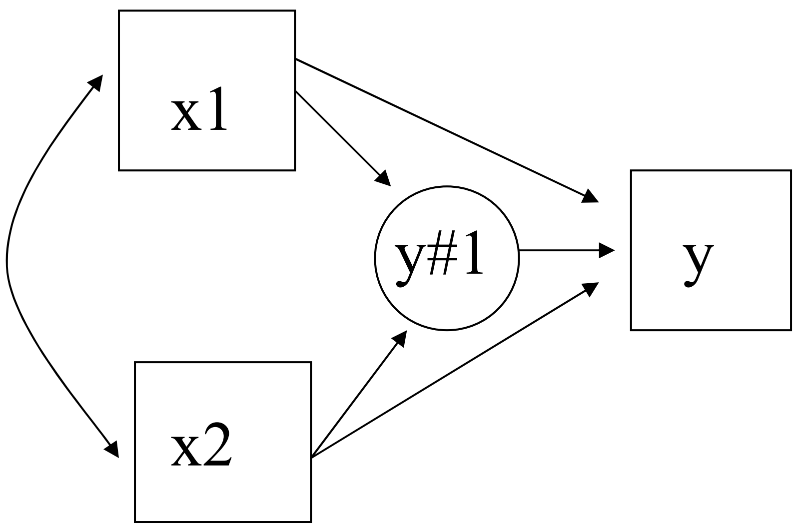
$$\Pr(group1)=\frac{e^{L}}{1+e^{L}}$$

Logistic regression model

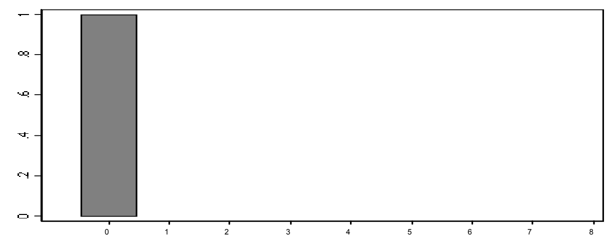
$$\Pr(Y=y)=\frac{e^{-\lambda}\lambda^{y}}{y!}$$

Poisson model

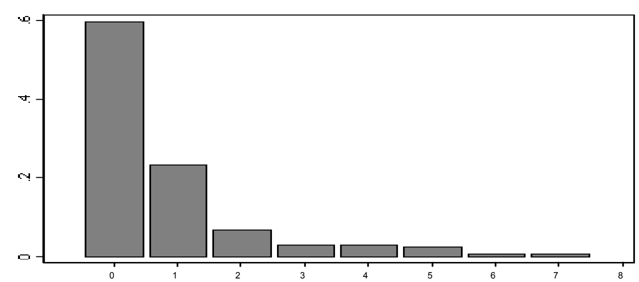
Two class models



y#1 = 1: “zero class”
P(y=0)=1



y#1 = 2: “convicted”
Poisson distribution



y#1 is a two class variable

Mplus example - ZIP

```
!Input file

VARIABLE:
   NAMES ARE ...           ;
   missing = .             ;
   USEV    = ...           ;
   COUNT   = total (i);

MODEL:
   total       ON x1 x2 ;
   total#1     ON x1 x2 ;
```

```
MODEL RESULTS

                   Estimates

 TOTAL    ON
    x1             0.453
    x2             0.434

 TOTAL#1 ON
    x1            -0.260
    x2            -0.952

 Intercepts
    TOTAL#1        0.816
    TOTAL          1.031
```

Interpreting the zip part

Odds of being in zero class: $e^{0.816} = 2.261$

Probability to be in zero class:

$$odds = \frac{\Pr(zero\_class)}{1 - \Pr(zero\_class)} = 2.26$$

$$\Pr(zero\_class) = \frac{2.26}{1 + 2.261} = .693$$

Interpreting the count part

The average rate of conviction (average number of crimes) given a person is in the non-zero class, and both covariates are equal to zero:

$$e^{1.031} = 2.804$$

The average rate of conviction for boys with $x_1=0$ and $x_2=0$

$$= 2.804*(1 - \Pr(zero\_class))$$

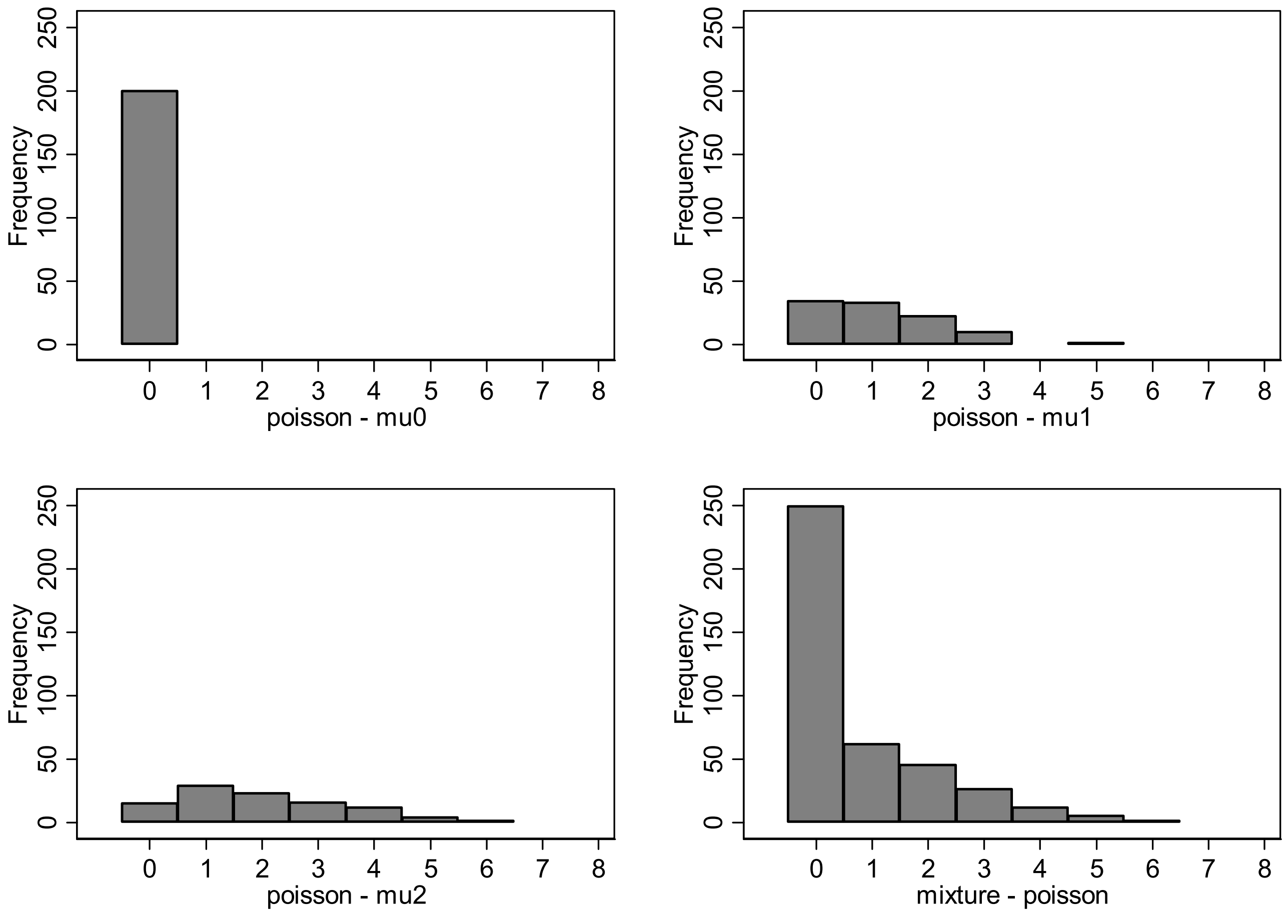
$$= 2.804*1 - .693$$

$$= .861$$

Zero counts in longitudinal data

- Single-class models
  - Censored, Two-part modeling
- Two-class models
  - Censored inflated, ZIP, Mover-stayer
- Multi-class models
  - LCGA with zip, GMM with zero class, combination of the above

Multi-class model



Example – Roeder et al.

- “Cambridge study” (Farrington/West)
- 411 boys – 403 in the analysis
- Age 10 to age 40 (!)
- Number of convictions each year
  - Total 0 up to 7 in a given year
  - Analysis mostly done with 2 year intervals
- Daring and rearing as covariates

Raw data

i	*Age 10*	*Age 12*	*Age 14*	*Age 16*	*Age 18*
1	0	0	1	1	0
2	0	0	0	2	1
3	0	0	0	0	0

M*plus* for (zero inflated) count

```
VARIABLE:
NAMES = …
COUNT = cage10 - cage30(i);

MODEL:
   i s  q    | cage10@0   … cage30@10;
  ii si qi   | cage10#1@0 … cage30#1@10;
   s-qi@0;
```

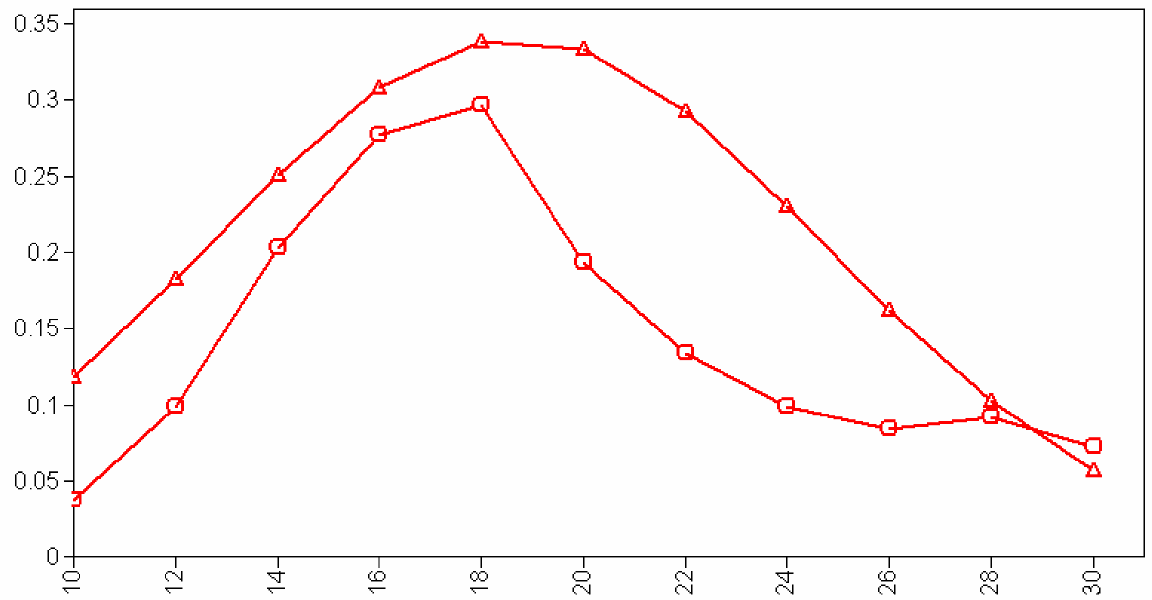
M*plus* for (zero inflated) count

H0 Value		-1519.762
Free Param.		4
Bayesian		3063.530
	est.	t
I	-4.265	-18.967
S	0.488	7.781
Q	-0.056	-8.023

H0 Value		-1496.937
Free Parameter		7
Bayesian (BIC)		3035.878
	est.	t
I	-3.231	-10.461
S	0.130	1.272
Q	-0.027	-2.802
II	@0	
SI	-1.776	-3.765
QI	0.132	4.866
CAGE10#1	1.532	2.861

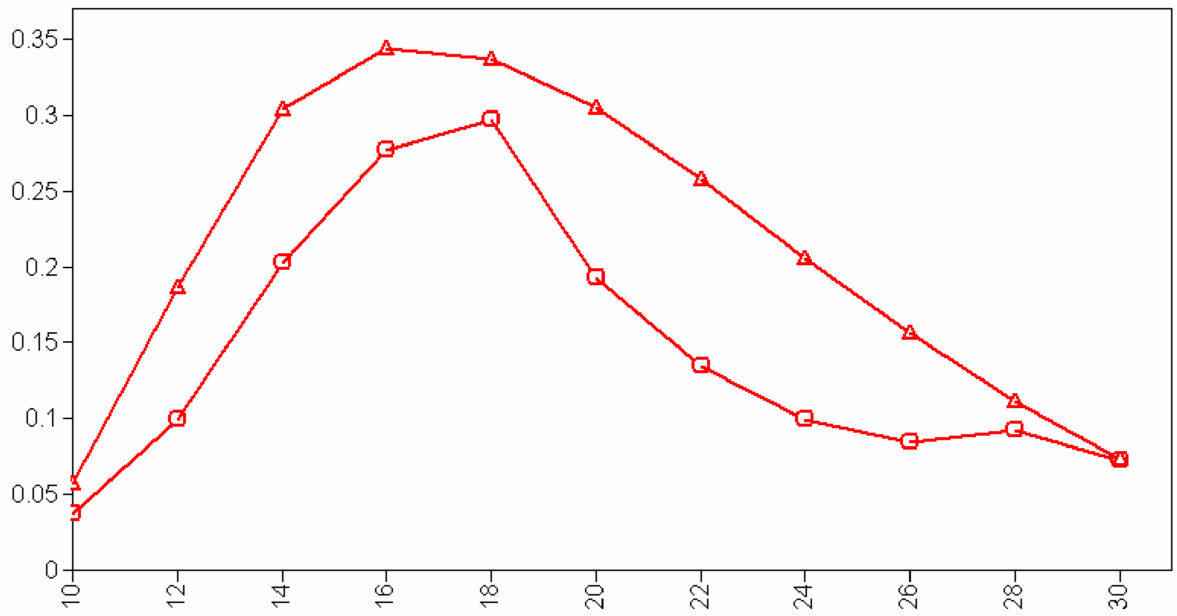
GMM

Predicted/ observed average conviction rate

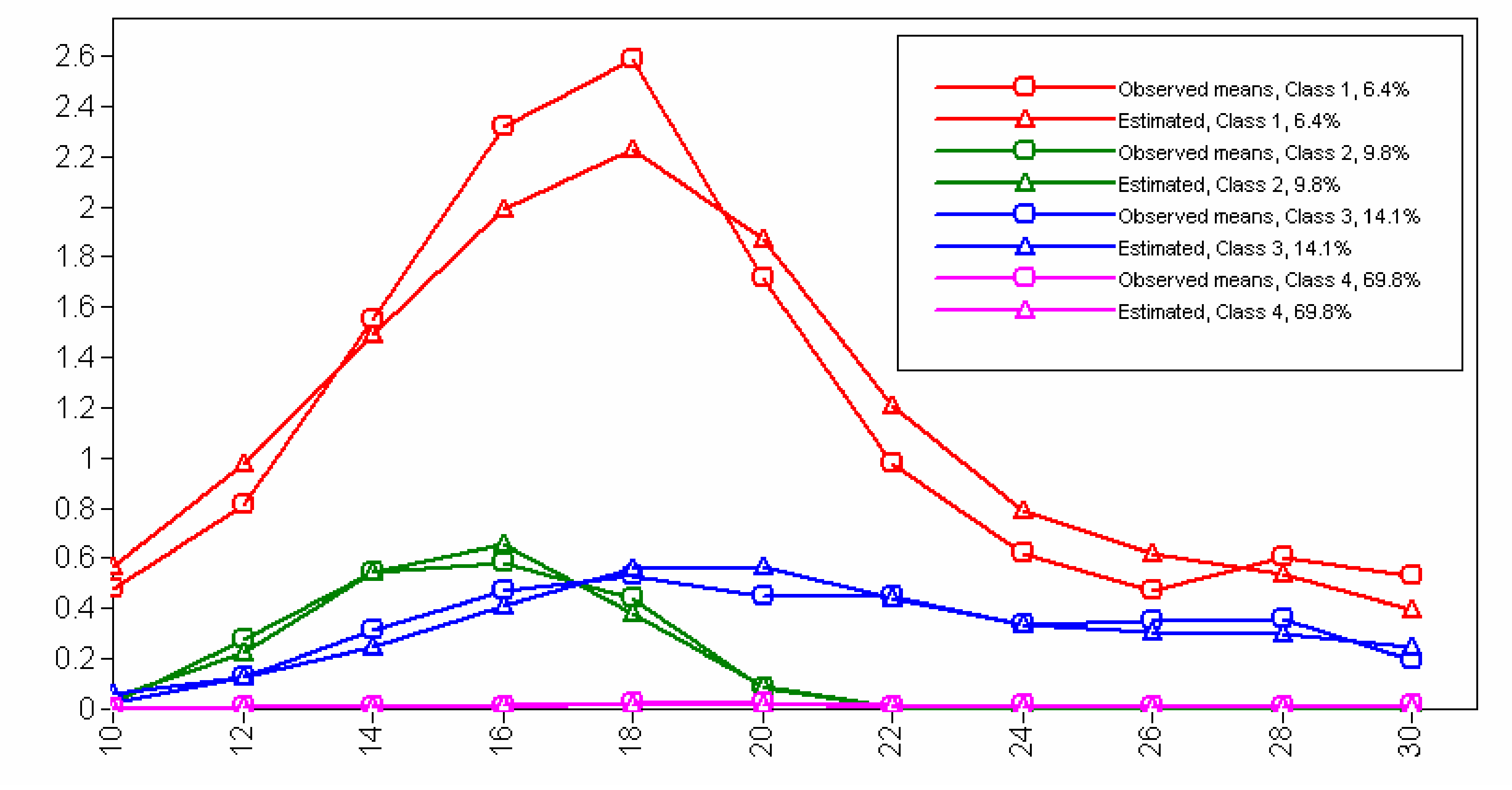


GMM-ZIP

Predicted/ observed average convicition rate



Outlook Latent class growth model with 4 classes



References (selection)

- Land, K.C., McCall, P.L., & Nagin, D.S. (1996): A Comparison of Poisson, Negative Binomial, and Semiparametric Mixed Poisson Regression Models. Sociological Methods & Research, 24, 387-442.
- Olsen, M.K. & Schafer, J.L. (2001): A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data. Journal of the American Statistical Association, 96, 730-745.
- Roeder, K., Lynch, K.G. & Nagin, D.S. (1999): Modeling uncertainty in latent class membership: A case study in criminology. Journal of the American Statistical Association, 94, 766-776.