SO, NOW WHAT?
ANALYZING CATA DATA USING GENERALIZED LINEAR MODELS

Miles A. Zachary
West Virginia University

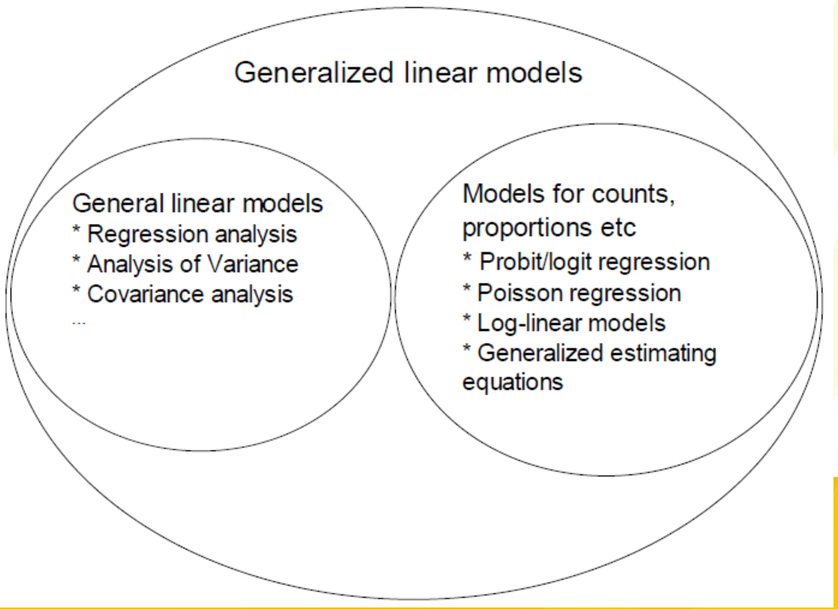
ANALYZING CATA DATA

- Most CATA software produces data in the form of counts
  - i.e., number of words related to a given construct or dimension
  - Easily exported as a spreadsheet
- Count data is discrete, not continuous
  - Continuous data can take any value between two points
  - Discrete data is more restrictive (e.g., counts are non-negative)
- Discrete data present several challenges to traditional analyses
  - Violate assumptions (i.e., normality and homoscedasticity)
    - Often biases standard error estimates and test statistics
  - Predicted values can be out of range or result in strange interpretations

SO, NOW WHAT?

GENERALIZED LINEAR MODELING

- Generalized linear models are a general class of statistical models that include a number of common special cases



GENERALIZED LINEAR MODELING

- GLMs have three (3) basic components

1. Linear Predictor (*η*) – the linear combination of explanatory variables (e.g., X_1, X_2,…X_n)

2. Family Distribution (y ~) – the probability distribution that theoretically produces the dependent variable
    - Any exponential distribution can be specified
    - E.g., normal (Gaussian), Bernoulli, Poisson, negative binomial, gamma, etc.

3. Link Function (*g(·)*) – relates the mean of the distribution to the linear predictor; linearizes the relationship between the DV and IVs
    - E.g., identity, logit/probit, log, complementary log-log, power, etc.
    - The canonical link is the natural link function of a given distribution

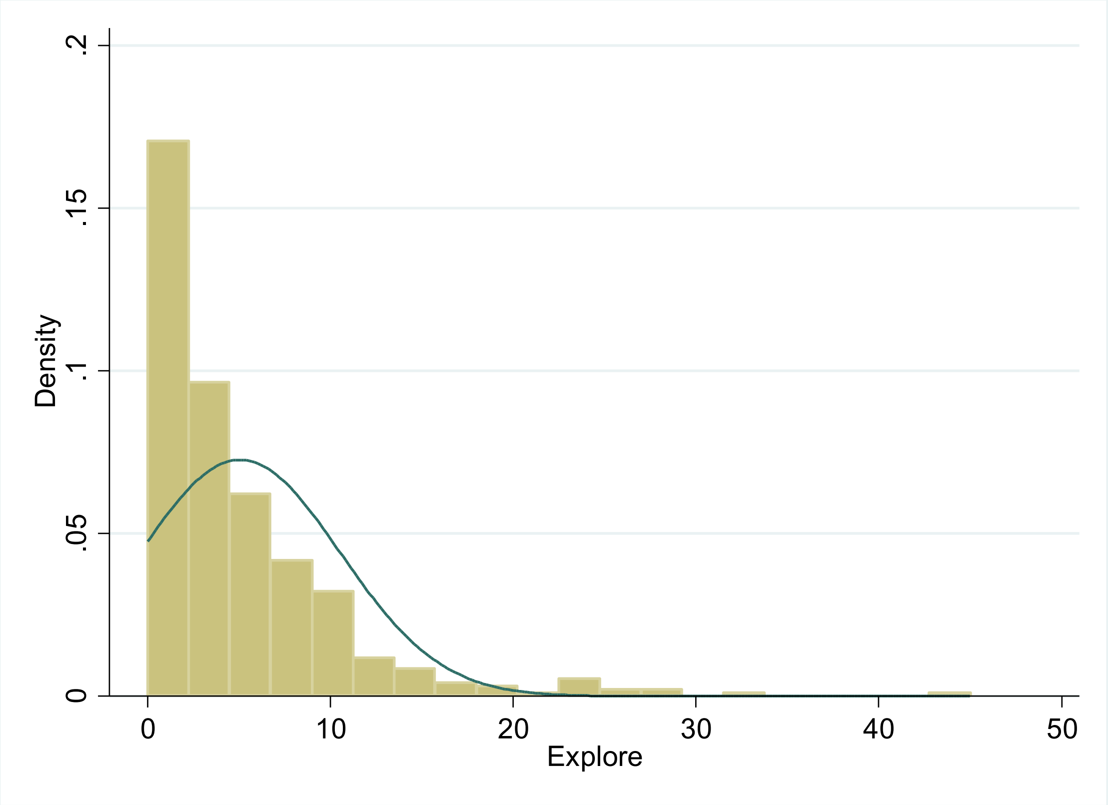
GENERALIZED LINEAR MODELING

Model	Family Distribution Component	Linear Predictor Component	Link Function (Canonical)
Linear Regression	Normal (Gaussian)	Continuous	Identity
Logistic Regression	Binomial	Mixed	Logit
Loglinear Regression	Poisson	Categorical	Log
Poisson Regression	Poisson	Mixed	Log
Negative Binomial Regression	Negative Binomial	Mixed	Generalized Log

GENERALIZED LINEAR MODELING: AN EXAMPLE

- RQ: How is retained earnings related to exploration rhetoric in shareholder letters?
  - Firms likely use exploration rhetoric to justify higher RE
  - "We're just one kid in a garage with a good idea away from going out of business" – Bill Gates (when questioned about RE policy)
- Sample: Shareholder letters for S&P 500 firms from 2005-2011
  - Shareholder letters are a common means through which firms communicate with shareholders
- CATA: Shareholder letters were content-analyzed using CATScanner 1.0
  - Exploration rhetoric measured using Uotila and colleagues (2009) measure
  - Exported as .csv file then imported into Stata 13

GENERALIZED LINEAR MODELING: AN EXAMPLE



CASE 1: CROSS-SECTIONAL STUDY WITH COUNT DV

- Test relationship between RE and exploration rhetoric using simple OLS and Poisson regression

$$E\left(Y_{Exp}\right) = \eta = \beta_0 + \ \beta_1 X_{Size} + \beta_2 X_{Age} + \beta_3 X_{HT} + \beta_4 X_{RE} + \varepsilon$$

Where, *Y ~ Normal*

$$\ln[E\left(Y_{Exp}\right)] = \eta = \beta_0 + \ \beta_1 X_{Size} + \beta_2 X_{Age} + \beta_3 X_{HT} + \beta_4 X_{RE} + \varepsilon$$

Where, *Y ~ Poisson*

CASE 1: CROSS-SECTIONAL STUDY WITH COUNT DV

Source	SS	df	MS
Model	224.907938	4	56.2269844
Residual	5743.38128	199	28.8612125
Total	5968.28922	203	29.4004395

Number of obs =	204
F(4, 199) =	1.95
Prob > F =	0.1039
R-squared =	0.0377
Adj R-squared =	0.0183
Root MSE =	5.3723

explore	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
size	-5.28e-06	5.68e-06	-0.93	0.354	-.0000165	5.92e-06
age	.0062887	.0092117	0.68	0.496	-.0118764	.0244538
hightech	1.407856	.7691013	1.83	0.069	-.1087779	2.924491
z_re	.1226746	.0782502	1.57	0.119	-.0316315	.2769806
_cons	4.785301	.7404076	6.46	0.000	3.325249	6.245352

CASE 1: CROSS-SECTIONAL STUDY WITH COUNT DV

```
Poisson regression                              Number of obs     =        204
                                                LR chi2(4)        =      40.19
                                                Prob > chi2       =     0.0000
Log likelihood = -753.46525                     Pseudo R2         =     0.0260
```

explore	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
size	.9999989	4.99e-07	-2.22	0.026	.9999979	.9999999
age	1.001169	.0007355	1.59	0.112	.9997284	1.002612
hightech	1.292262	.0785241	4.22	0.000	1.147169	1.455706
z_re	1.019222	.0054934	3.53	0.000	1.008512	1.030046
_cons	4.748374	.2904699	25.47	0.000	4.211868	5.353219

CASE 2: PANEL STUDY WITH COUNT DV

- Test relationship between RE and exploration rhetoric using simple panel linear regression and generalized estimating equations
  - Compensates for autocorrelation in longitudinal data

$$E\left(Y_{Exp}\right) = \eta = \beta_0 + \beta_1 X_{Size} + \beta_2 X_{Age} + \beta_3 X_{HT} + \beta_4 X_{RE} + \varepsilon$$

Where, *Y ~ Normal*

$$\ln[E\left(Y_{Exp}\right)] = \eta = \beta_0 + \beta_1 X_{Size} + \beta_2 X_{Age} + \beta_3 X_{HT} + \beta_4 X_{RE} + \varepsilon$$

Where, *Y ~ Poisson*

CASE 2: PANEL STUDY WITH COUNT DV

```
Random-effects GLS regression                   Number of obs      =      1204
Group variable: id                              Number of groups   =       207

R-sq:  within  = 0.0000                         Obs per group: min =         1
       between = 0.0414                                        avg =       5.8
       overall = 0.0338                                        max =         7

                                                Wald chi2(4)       =      9.60
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0476

------------------------------------------------------------------------------
     explore |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        size |   2.06e-07   2.82e-06     0.07   0.942    -5.31e-06    5.73e-06
         age |   .0086418   .0065193     1.33   0.185    -.0041359    .0214194
    hightech |   1.493096    .573825     2.60   0.009     .3684194    2.617772
        z_re |   .0242243   .0353619     0.69   0.493    -.0450838    .0935325
       _cons |   4.396266   .5416948     8.12   0.000     3.334563    5.457968
-------------+----------------------------------------------------------------
     sigma_u |  3.6253216
     sigma_e |   3.877597
         rho |  .46641433   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

CASE 2: PANEL STUDY WITH COUNT DV

GEE population-averaged model		Number of obs =	1115
Group and time vars:	id year	Number of groups =	182
Link:	log	Obs per group: min =	2
Family:	Poisson	avg =	6.1
Correlation:	AR(1)	max =	7
		Wald chi2(4) =	88.61
Scale parameter:	1	Prob > chi2 =	0.0000

explore	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
size	.9999997	2.17e-07	-1.18	0.239	.9999993	1
age	1.001268	.0004534	2.80	0.005	1.00038	1.002157
hightech	1.32367	.0526632	7.05	0.000	1.224374	1.431019
z_re	1.009486	.0022934	4.16	0.000	1.005001	1.013991
_cons	4.670786	.1817241	39.62	0.000	4.327855	5.040891

CASE 3: RCM STUDY WITH COUNT DV

- Test relationship between RE and exploration rhetoric using a linear RCM model and a generalized linear mixed model
  - Estimating fixed effects and random intercept and slope effects for RE by firm

$$E(Y_{Exp}) = \eta = \beta_0 + \beta_1 X_{Size} + \beta_2 X_{Age} + \beta_3 X_{HT} + \beta_4 X_{RE} + \varepsilon$$

$$\beta_0 = \gamma_{ID} + u_{ID}$$

$$\beta_4 = \gamma_{ID} + u_{ID}$$

Where, *Y ~ Normal*

$$\ln[E(Y_{Exp})] = \eta = \beta_0 + \beta_1 X_{Size} + \beta_2 X_{Age} + \beta_3 X_{HT} + \beta_4 X_{RE} + \varepsilon$$

$$\beta_0 = \gamma_{ID} + u_{ID}$$

$$\beta_4 = \gamma_{ID} + u_{ID}$$

Where, *Y ~ Poisson*

CASE 3: RCM STUDY WITH COUNT DV

Log likelihood = -3523.256	Wald chi2(4) = 10.01
	Prob > chi2 = 0.0402

explore	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
size	-5.61e-08	2.80e-06	-0.02	0.984	-5.54e-06	5.43e-06
age	.0077793	.0066405	1.17	0.241	-.0052359	.0207944
hightech	1.456369	.5865977	2.48	0.013	.3066585	2.606079
z_re	.0776071	.0577411	1.34	0.179	-.0355635	.1907776
_cons	4.525331	.5598816	8.08	0.000	3.427983	5.622679

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
id: Independent				
var(z_re)	.0484966	.0385793	.0101994	.2305935
var(_cons)	13.12277	1.663861	10.23528	16.82484
var(Residual)	14.75399	.6779342	13.48333	16.14438

LR test vs. linear regression: chi2(2) = 415.30 Prob > chi2 = 0.0000

CASE 3: RCM STUDY WITH COUNT DV

```
Integration method: mvaghermite                  Integration points =          7

                                                 Wald chi2(4)       =      21.02
Log likelihood =  -3307.718                      Prob > chi2        =     0.0003
```

explore	exp(b)	Std. Err.	z	P>|z|	[95% Conf. Interval]	
size	1	3.73e-07	0.29	0.773	.9999994	1.000001
age	1.001571	.0013784	1.14	0.254	.9988729	1.004276
hightech	1.555708	.1921759	3.58	0.000	1.22118	1.981877
z_re	1.042278	.016899	2.55	0.011	1.009678	1.075932
_cons	3.418806	.4101445	10.25	0.000	2.702453	4.325045
id						
var(z_re)	.0125007	.0040901			.0065831	.0237376
var(_cons)	.5948619	.0755231			.4638189	.7629286

LR test vs. Poisson regression: chi2(2) = 2568.50 Prob > chi2 = 0.0000

CASE 4: RCM GROWTH STUDY WITH COUNT DV

- Examine exploration rhetoric over time using a linear RCM model and a generalized linear mixed model
  - Estimating fixed effects and random intercept and slope effects for time by firm

$$E\left(Y_{Exp}\right) = \eta = \beta_0 + \ \beta_1 X_{Time} + \varepsilon$$

$$\beta_0 = \ \gamma_{ID} + u_{ID}$$

$$\beta_1 = \ \gamma_{ID} + u_{ID}$$

Where, *Y ~ Normal*

$$\ln[E\left(Y_{Exp}\right)] = \eta = \beta_0 + \ \beta_1 X_{Time} + \varepsilon$$

$$\beta_0 = \ \gamma_{ID} + u_{ID}$$

$$\beta_1 = \ \gamma_{ID} + u_{ID}$$

Where, *Y ~ Poisson*

CASE 4: RCM GROWTH STUDY WITH COUNT DV

```
                                                Wald chi2(1)      =     13.99
Log likelihood = -8230.6332                     Prob > chi2       =    0.0002

------------------------------------------------------------------------------
     explore |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |   .2381061   .0636484     3.74   0.000     .1133575    .3628546
       _cons |   5.088703   .2779249    18.31   0.000     4.543981    5.633426
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Independent              |
                   var(time) |   .4854223   .0989551      .3255362    .7238361
                  var(_cons) |   22.35206   2.052176      18.67099    26.75886
-----------------------------+------------------------------------------------
               var(Residual) |   24.59095   .8017026      23.06879    26.21355
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(2) =  1103.78   Prob > chi2 = 0.0000
```

CASE 4: RCM GROWTH STUDY WITH COUNT DV

```
Mixed-effects Poisson regression                Number of obs     =      2553
Group variable:                 id              Number of groups  =       540

                                                Obs per group: min =         1
                                                               avg =       4.7
                                                               max =         7

Integration method: mvaghermite                 Integration points =         7

                                                Wald chi2(1)      =     85.52
Log likelihood = -7366.7782                     Prob > chi2       =    0.0000

------------------------------------------------------------------------------
     explore |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |   1.041566   .0045869     9.25   0.000     1.032615    1.050596
       _cons |   3.598381   .1458448    31.59   0.000      3.32359    3.895893
-------------+----------------------------------------------------------------
id           |
   var(_cons)|   .6956614   .0501614                       .603978    .8012624
------------------------------------------------------------------------------
LR test vs. Poisson regression:  chibar2(01) =  8799.51 Prob>=chibar2 = 0.0000
```

Notice – no random slopes….

CONCLUSIONS

- Generalized linear models offer several advantages
  - Data are analyzed in their correct form
  - Unbiased standard errors and p-values
  - Easier and more direct interpretation of model coefficients
- But, GLIMs should be approached with caution and patience
  - Model estimates are sensitive to specification errors
  - Can be tedious to estimate all parameters
    - Iterative method of solving the likelihood equation (e.g., Newton-Raphson algorithm)
    - Particularly random effects with more complicated error structures
  - Convergence difficult when:
    - Low variance
    - Model is underidentified