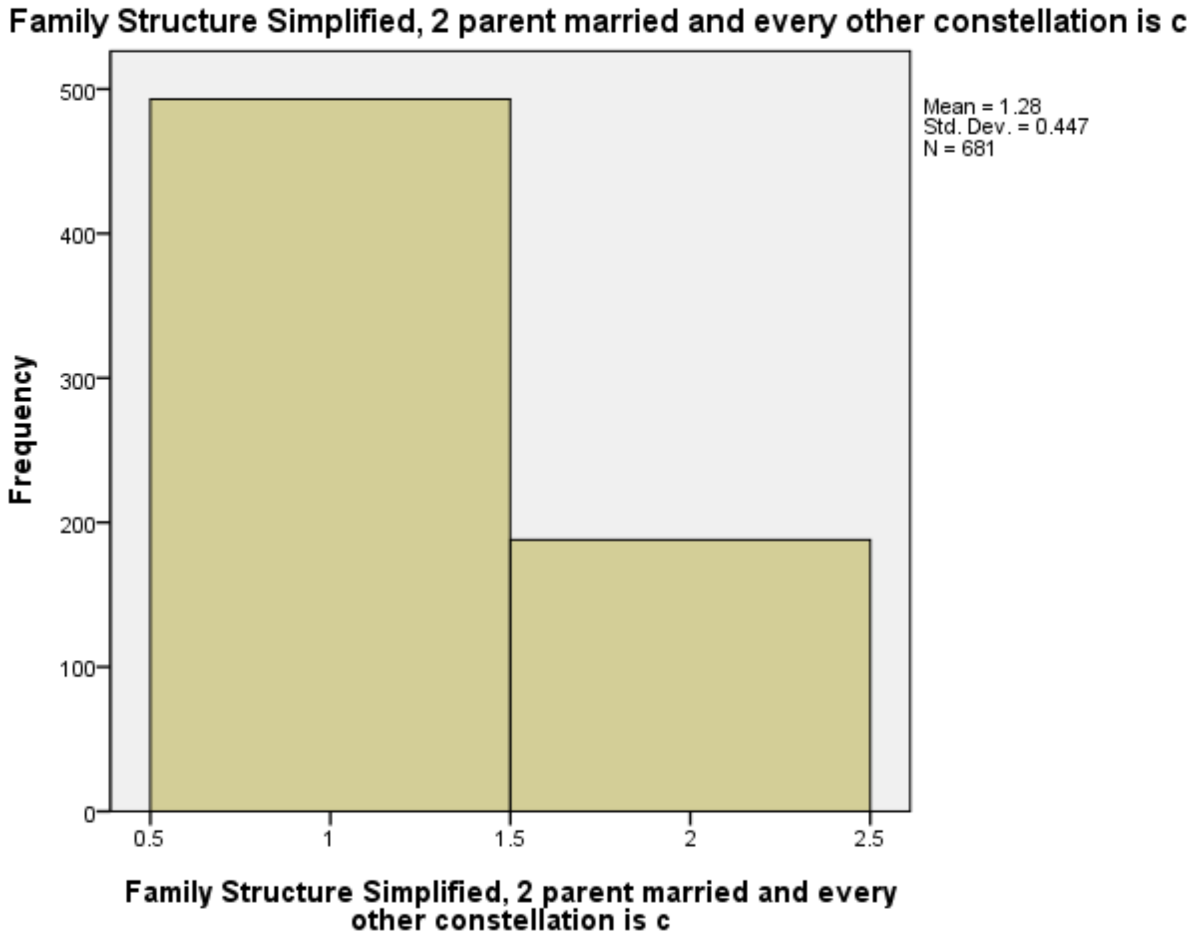
Regression with Non-normal Outcomes: Applied Examples of Dichotomous, Categorical, Poisson, and Negative Binomial Outcomes

Jeremy Yorgason

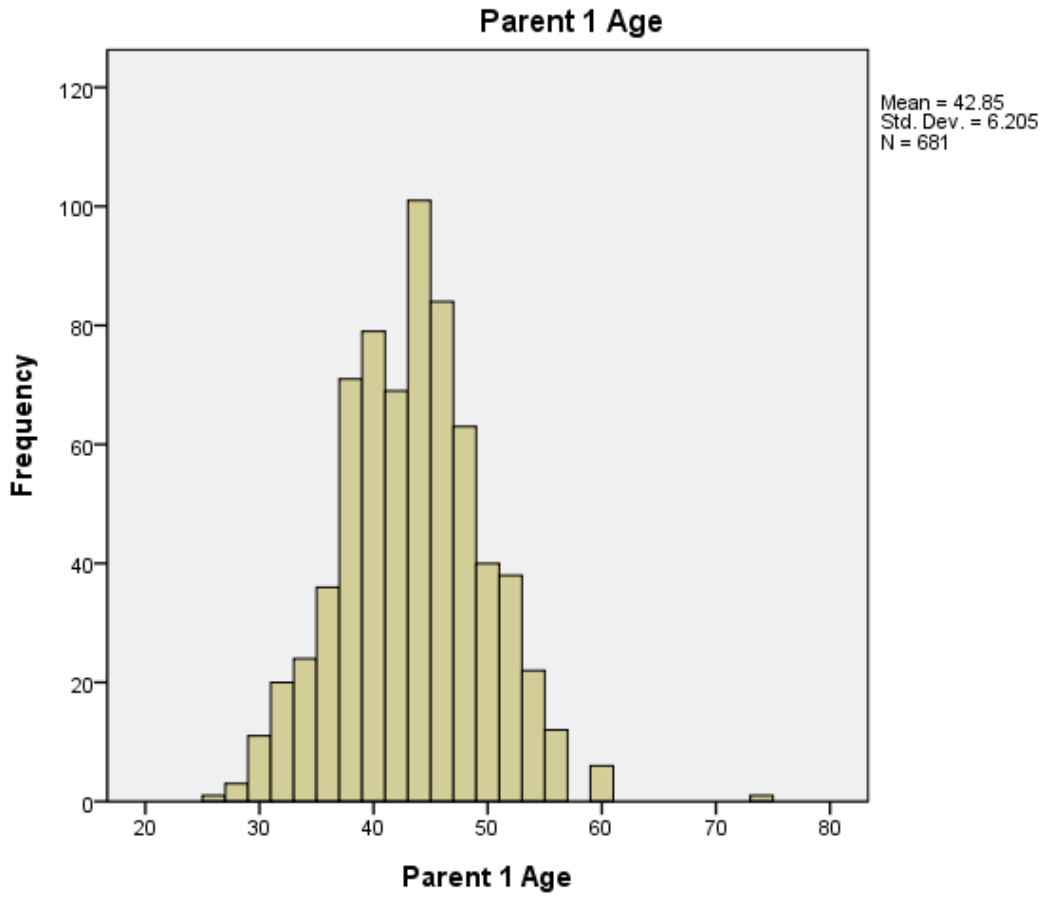
Lance Erickson

Methods Workshop: Family Studies Center

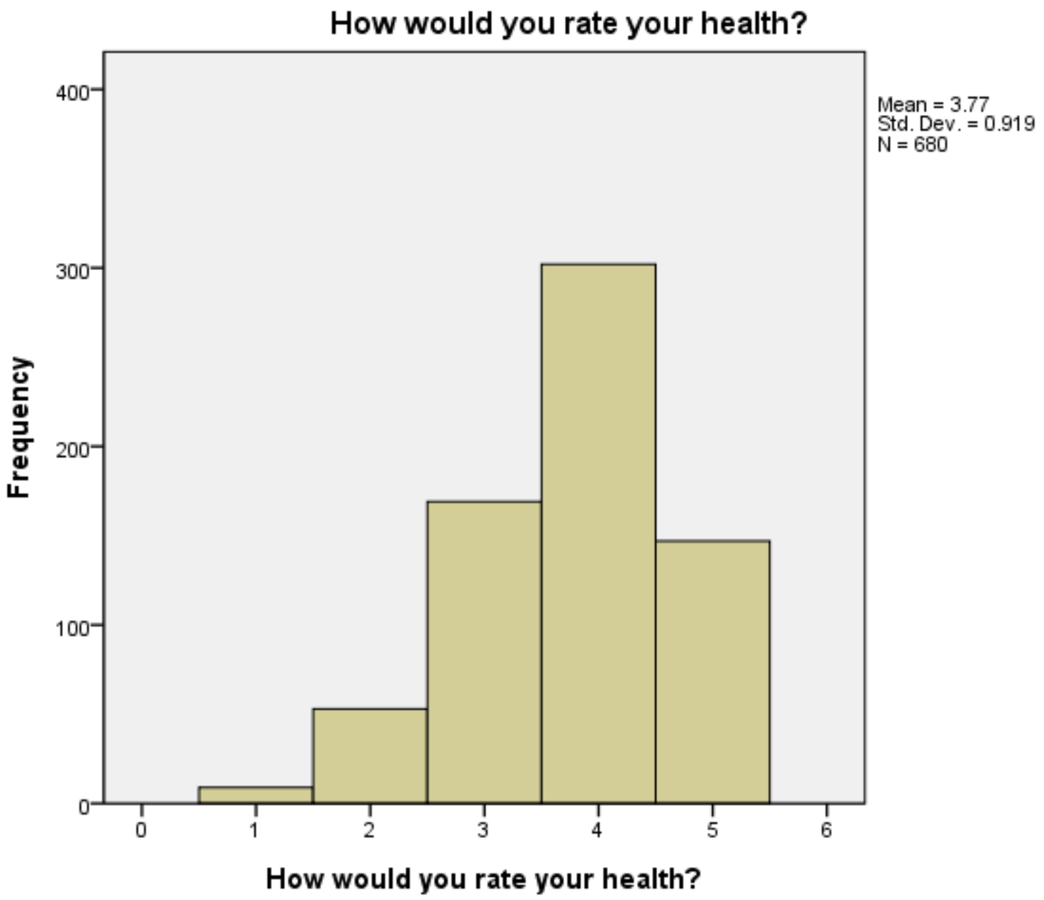
Game: Name that Distribution



Game: Name that Distribution

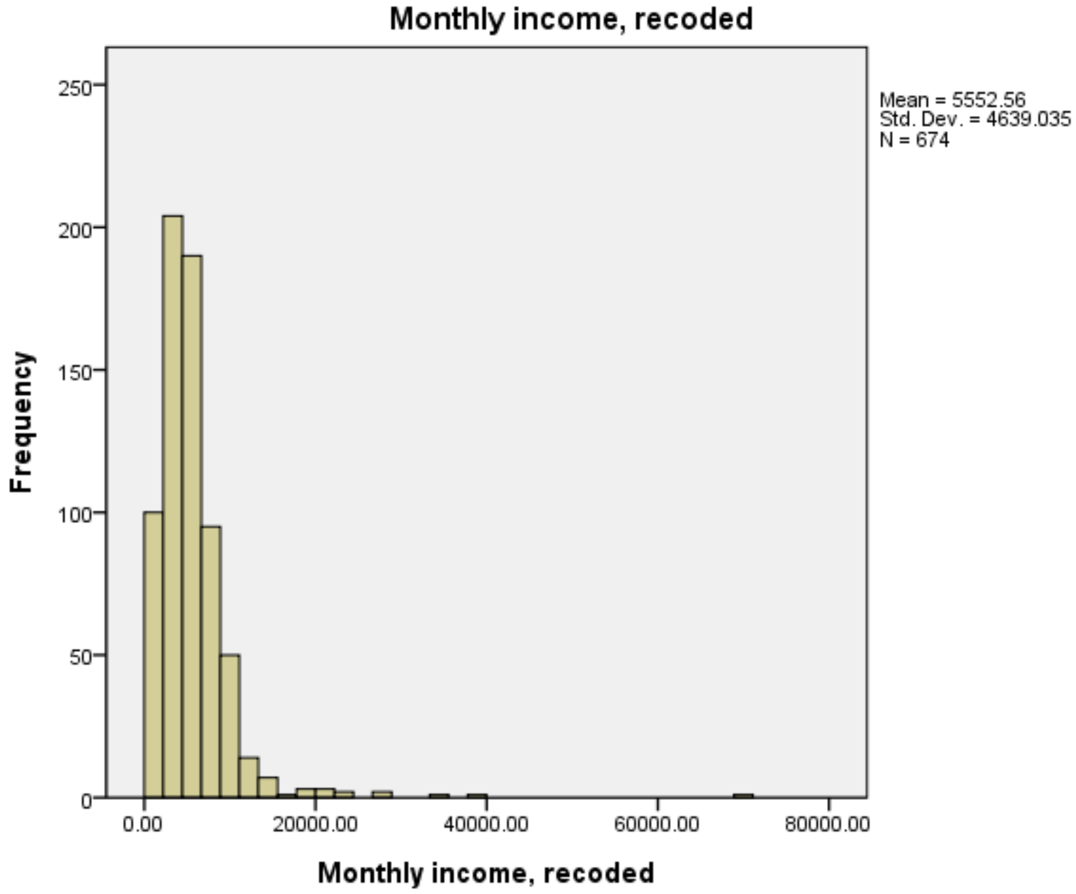


Game: Name that Distribution



Game: Name that Distribution

Game: Name that Distribution



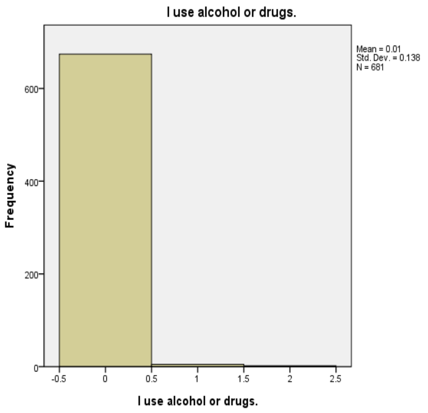
Game: **Name that Distribution**

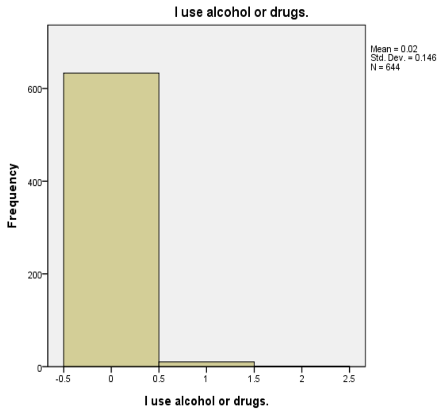
Drug and Alcohol Use among Adolescents

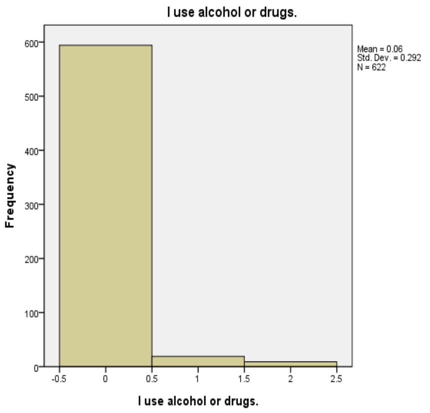
0 = not true

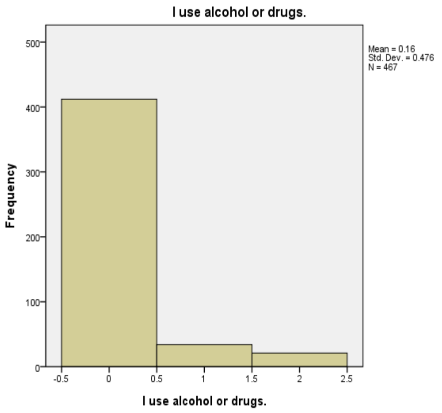
1 = somewhat true

2 = often true



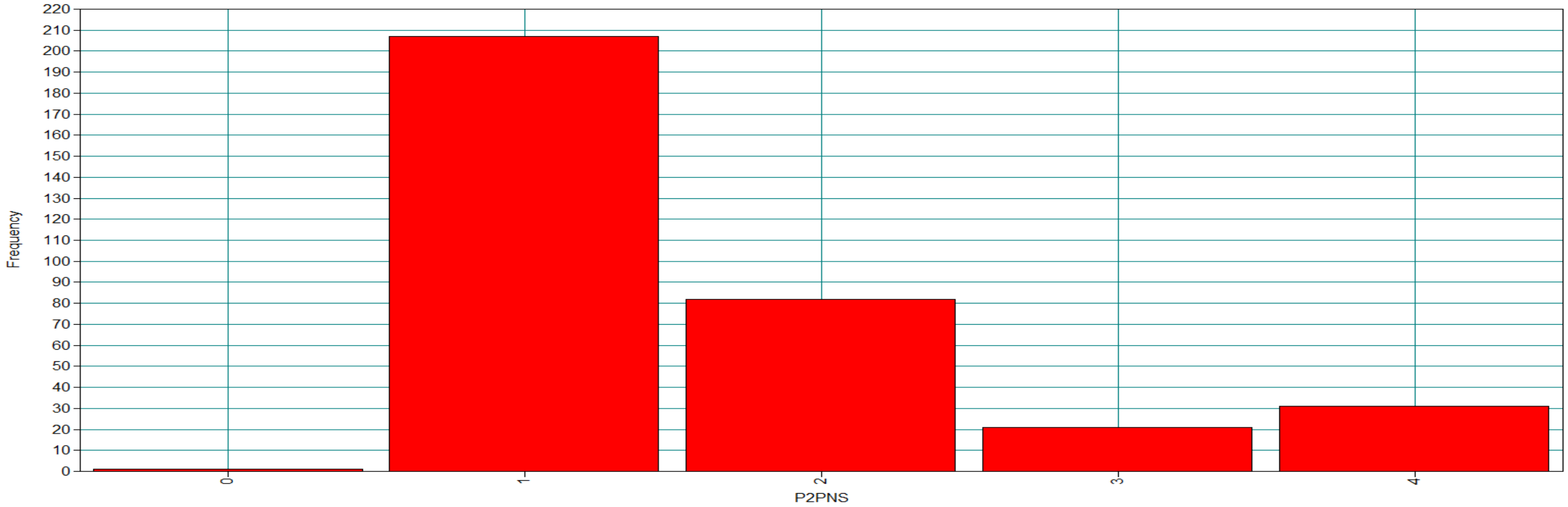






Game: Name that Distribution

Saved, Planned how much to save for retirement



Saved & Planned

Saved Only

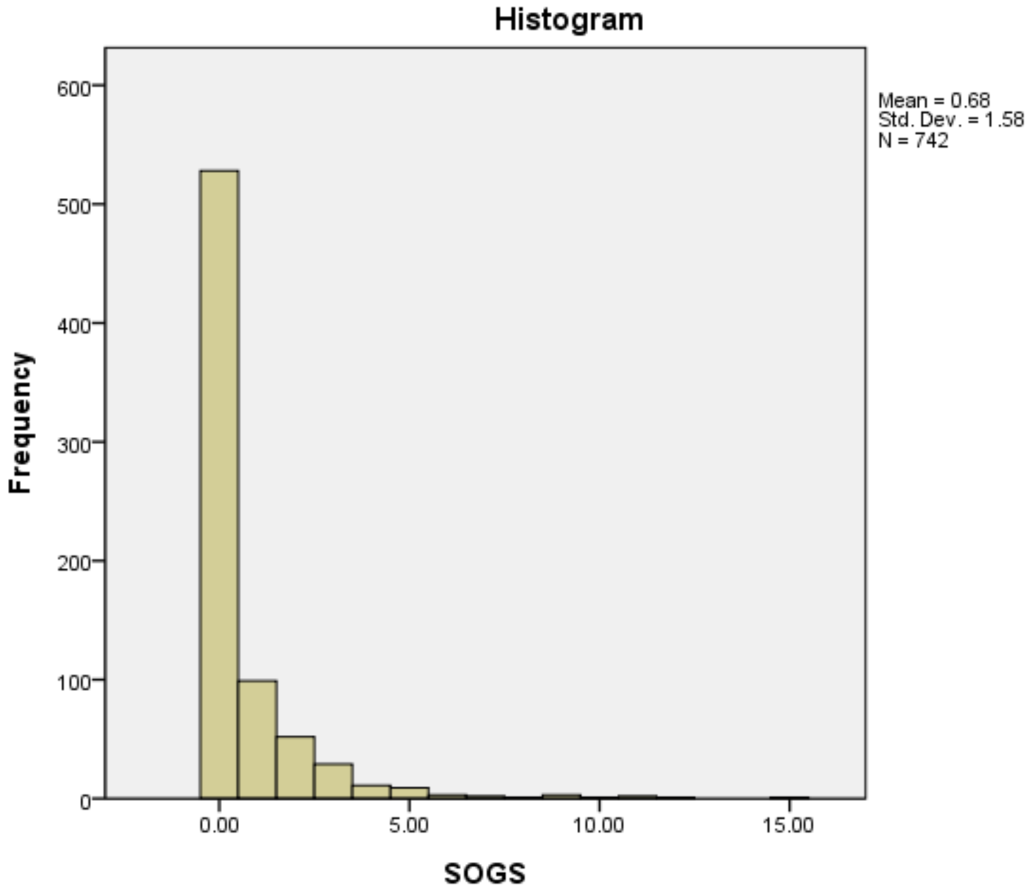
Planned Only

Didn’t Save or Plan

Game: <u>Name that Distribution</u>

Gambling Behavior in College

Mean= .68
Variance = 2.50



If the **<u>VARIANCE</u>** is larger than the **<u>mean</u>**, you should specify a negative binomial distribution (this one is actually "zero inflated" as well).

Examples

- Logistic regression
- Multinomial logistic regression
- Ordered logistic regression
- Count outcomes
  - Poisson
  - Negative binomial
  - Zero-inflated Poisson or negative binomial

Logistic regression

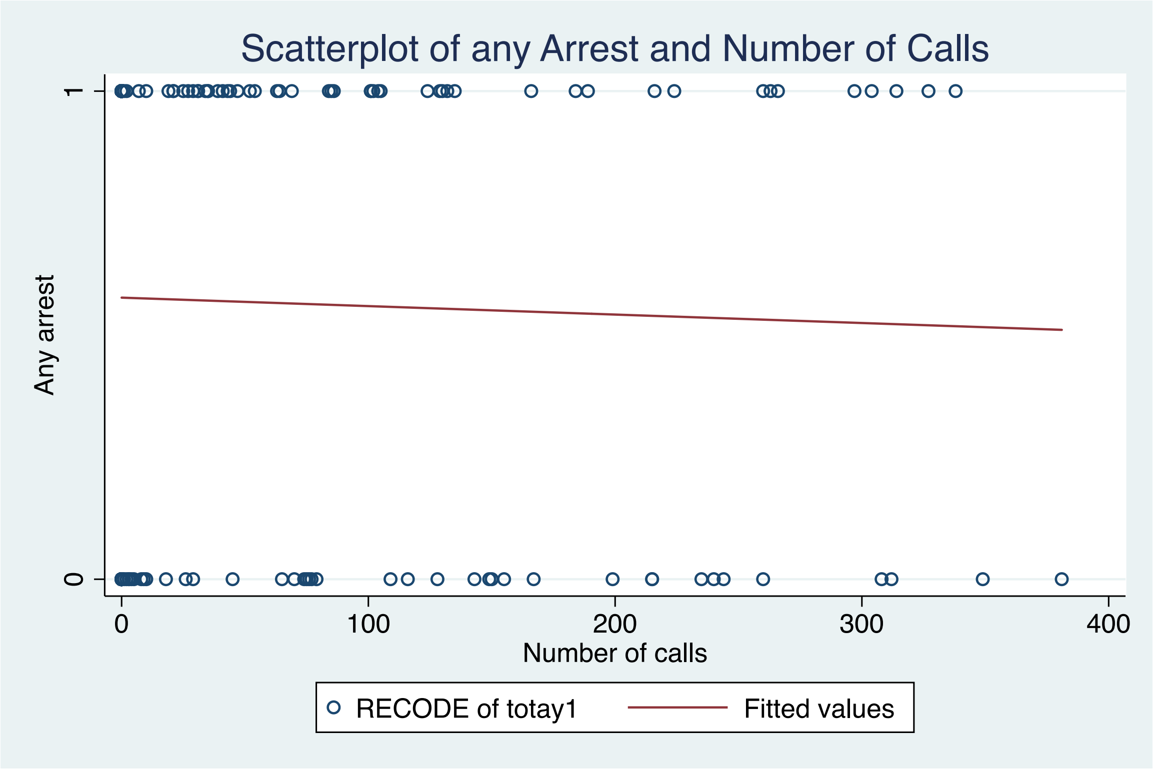
- Use when the outcome is dichotomous

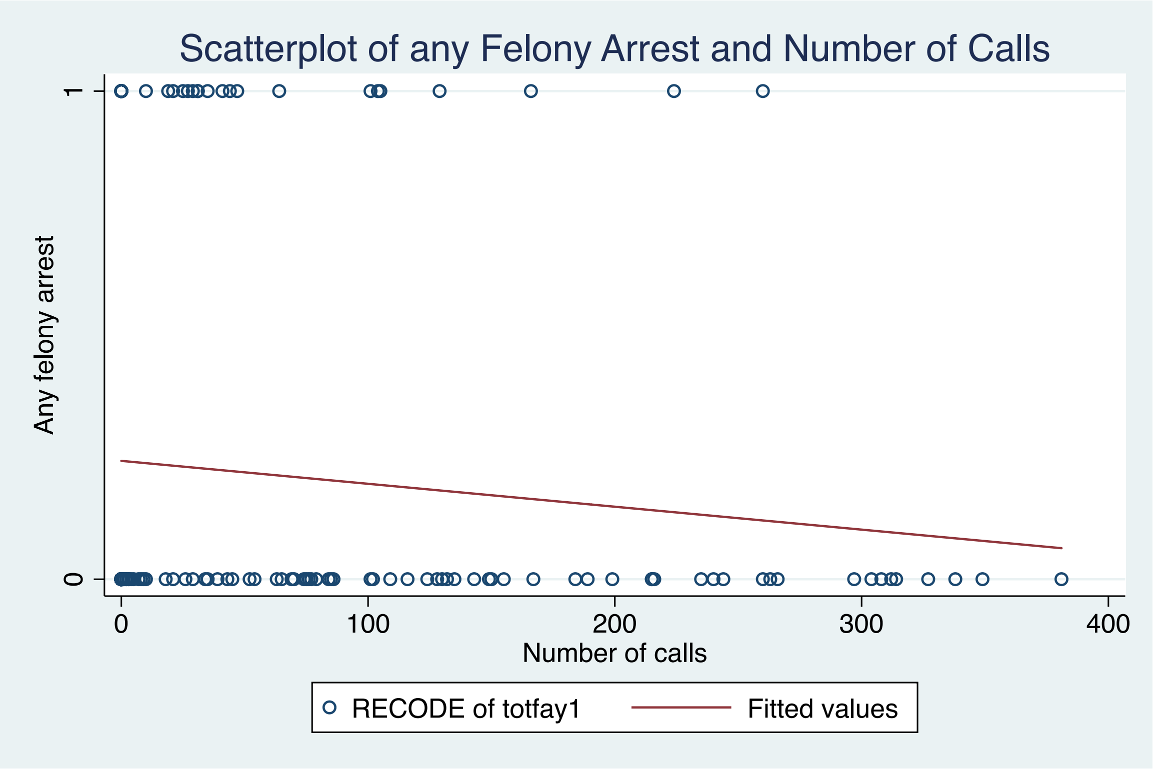
RealVictory

- Program to reduce recidivism in adolescents
- 6 cognitive-behavioral training sessions
  - Help individuals examine their attitudes and assess whether their actions are meeting their needs
- Set goals
- Twice daily phone calls for follow-up

Bahr SJ, Cherrington DJ, Erickson LD. 2016. "An evaluation of the impact of goal setting and cell phone calls on juvenile rearrests." *Int J Offender Ther Comp Criminol* 60: 1816-1835.

Was there an arrest in the next year?





```
quietly {
     eststo t4a:  logistic anyay1  treat    age    male    white totprior felprior ib1.program
     eststo t4a1: logistic anyay1  numcalls age    male    white totprior felprior ib1.program if treat==1
     eststo t4b:  logistic anyfay1 treat    age    male    white totprior felprior ib1.program
     eststo t4b1: logistic anyfay1 numcalls age /*male*/ white totprior felprior ib1.program if treat==1
}

estout t4a t4a1 t4b t4b1, ///
     eform nolz cells(b(star fmt(3))) starlevels(* .10 ** .05 *** .001) ///
     stats(N, fmt(0)) collabel(none) eqlabel(none) drop(_cons) ///
     mlabel("Model 1""Model 2^a""Model 3""Model 4^a,b") ///
     mgroups("Any Arrest" "Any Felony Arrest", pattern(1 0 1 0) span) ///
     prehead("Table 4." ///
             "The Relationship of Treatment and Number of Calls with Any Arrest and Any Felony Arrest:" ///
             "Odds Ratios from Logistic Regression") ///
     varlabel(treat "Treatment" exposure "    Posttreatment exposure days^c" age "    Age" ///
              male "    Male" white "    White" totprior "        Any" felprior "        Felony" ///
              1.program "        Juvenile probation" 2.program "        Rural" ///
              3.program "        Secure care" _cons "Constant" numcalls "Number of calls^c") ///
     refcat(age "Controls" totprior "    Number of previous arrests" 1.program "    Program site", label(" ")) ///
     postfoot("Note:" ///
              "^a Only treatment group included in analysis." ///
              "^b No females in the treatment group had a felony arrest." ///
              "^c In 100s. * p < .05, ** p < .01, ***p < .001, one-tailed tests.") ///
     varwidth(31)
```

Table 3. The Relationship of Treatment and Number of Calls with Any Arrest and Any Felony Arrest: Odds Ratio From Logistic Regression.

	Any arrest		Any felony arrest	
	Model 1	Model 2[a]	Model 3	Model 4[a,b]
Treatment	0.876		1.079	
Number of calls[c]		0.900		0.641*
Controls				
Age	0.867	0.926	0.786*	0.714*
Male	0.716	1.367	1.374	
White	0.835	0.730	1.397	1.247
Number of pretreatment arrests				
Any	1.022*	1.009	1.030**	1.023
Felony	0.987	0.980	0.973	0.992
Program site				
Probation	1.000	1.000	1.000	1.000
Rural	0.209***	0.201**	0.778	0.663
Secure care	0.159***	0.182**	2.878**	3.646*
N	256	136	256	136

Note. Odds Ratios (OR) are the antilog of the model coefficients and represent the change in odds of experiencing an arrest for a one-unit increase in the independent variable. For example, the OR for number of calls in Model 4 is .641, which indicates that the odds of arrest would decrease by a factor of .641 if the number of calls increases by one unit (100 calls in this case). A commonly used alternative interpretation transforms the OR into a percentage—$(1 - 0.641) \times 100 = 35.9$—indicating, in this case, that an increase of 100 calls reduces the odds of arrest by 35.9%.

[a]Only treatment group included in analysis.

[b]No females in the treatment group had a felony arrest.

[c]In 100s.

*$p < .05$. **$p < .01$. ***$p < .001$; one-tailed tests.

Logistic in Mplus

```
!Mplus Input syntax;
Categorical are p2retsav;
ANALYSIS:
Estimator = ML;
MODEL:
p2retsav on P1Mat P1FinStr Income2;
```

Mplus Output

```
P2RETSAV   ON
        P1MAT          -0.862        0.479   -1.801        0.072
        P1FINSTR       -0.192        0.033   -5.884        0.000
        INCOME2         0.135   0.076    1.767        0.077
Thresholds
        P2RETSAV$1  -2.217        0.570   -3.890        0.000
LOGISTIC REGRESSION ODDS RATIO RESULTS
P2RETSAV   ON
        P1MAT           0.422
        P1FINSTR               0.826
        INCOME2                1.145
```

Multinomial logistic regression

- Use when you have a nominal dependent variable with more than two categories

- Who do you know that would watch your house if you were hospitalized for two weeks?
    - Friend
    - Relative
    - No one

```
mlogit house female age married income health outprim attach veteran, cluster(zip) base(0) rrr
mlogit house female age married income health outprim attach veteran, cluster(zip) base(1) rrr
```

Table 2. ***Community Attachment and Veteran Status as Predictors of Anticipated Help: Odds Ratios from Multinomial Logistic Regression***

	Friend vs. No One	Relative vs. No One	Relative vs. Friend
Watch house			
Female	1.11	1.37	1.24
Age	.90*	.90*	1.00
Married	1.49	2.23	1.50
Household income (in $1k)	1.00	1.00	.99*
Self-rated health	1.07	.99	.93
Leaves community for primary care	.83	.61	.74
Community attachment	1.34	1.40	1.05
Veteran	1.81	1.60	.88

Note: Source—*Utah Rural Community Study*. N = 569.

* $p < .05$; ** $p < .01$; *** $p < .001$; two-tailed tests.

```
!Mplus input for a multinomial model;
Nominal = p2pns;
DEFINE:
 !Creating a new variable that represents planning and saving for retirement;
 !The reference category is the last group by default;
 !In this model, the last group includes those that didn't plan or save;
 P2pns = ;
 !Save and Plan = 1;
 If p2retsav==1 and P2savpln ==1 THEN p2pns =1;
 !Save but not plan = 2;
 If p2retsav==1 and P2savpln ==2 THEN p2pns =2;
 !Plan but not save = 3;
 If p2retsav==2 and P2savpln ==1 THEN p2pns =3;
 !Not plan or save = 4;
 If p2retsav==2 and P2savpln ==2 THEN p2pns =4;
 MODEL:
 p2pns#1 p2pns#2 p2pns#3 on p2edu p2age p2finstr;
```

Mplus Output

```
P2PNS#1        ON
        P2EDU          1.650   0.284     5.816          0.000
        P2AGE          0.664   0.140     4.754          0.000
        P2FINSTR      -0.076   0.100    -0.755          0.450
P2PNS#2    ON
        P2EDU          0.598   0.160     3.749          0.000
        P2AGE          0.070   0.035     1.980          0.048
        P2FINSTR      -0.212   0.037    -5.670          0.000
P2PNS#3    ON
        P2EDU          0.369   0.167     2.209          0.027
        P2AGE          0.046   0.037     1.263          0.207
        P2FINSTR      -0.154   0.035    -4.373          0.000
LOGISTIC REGRESSION ODDS RATIO RESULTS
P2PNS#1    ON
        P2EDU          5.206
        P2AGE          1.942
        P2FINSTR       0.927
P2PNS#2    ON
        P2EDU          1.819
        P2AGE          1.072
        P2FINSTR       0.809
P2PNS#3    ON
        P2EDU          1.446
        P2AGE          1.047
        P2FINSTR       0.857
```

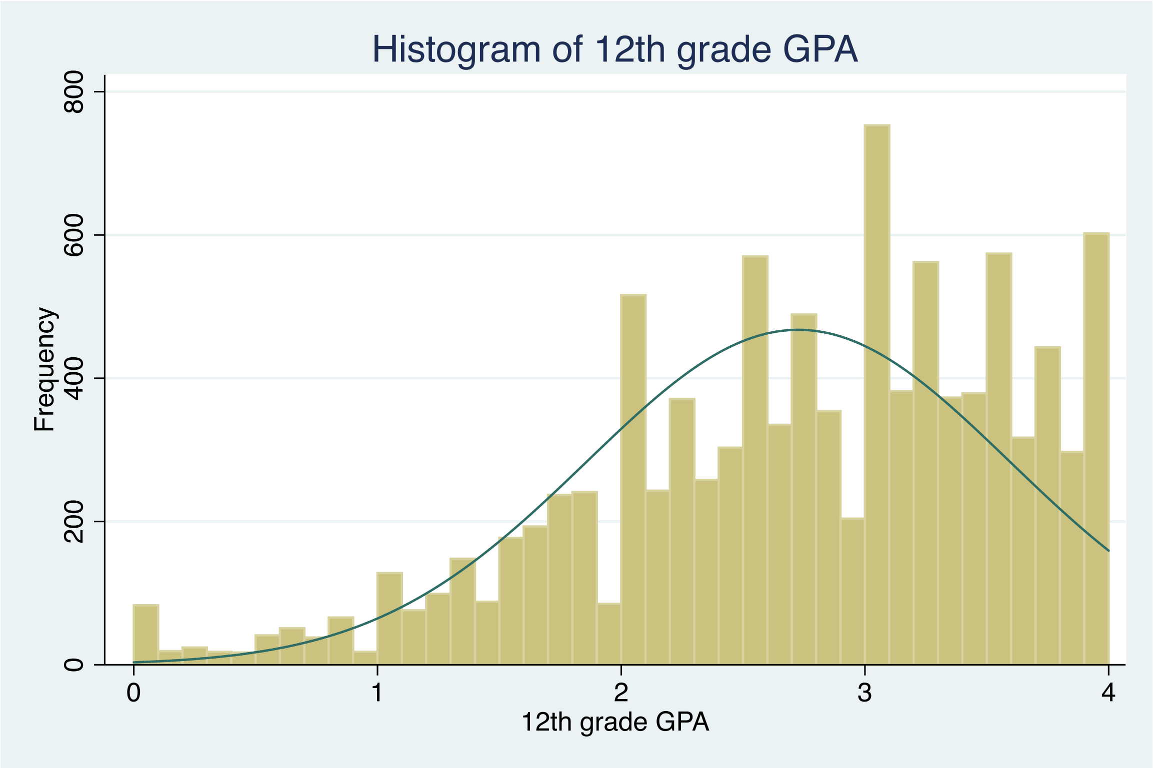
Ordered logistic regression

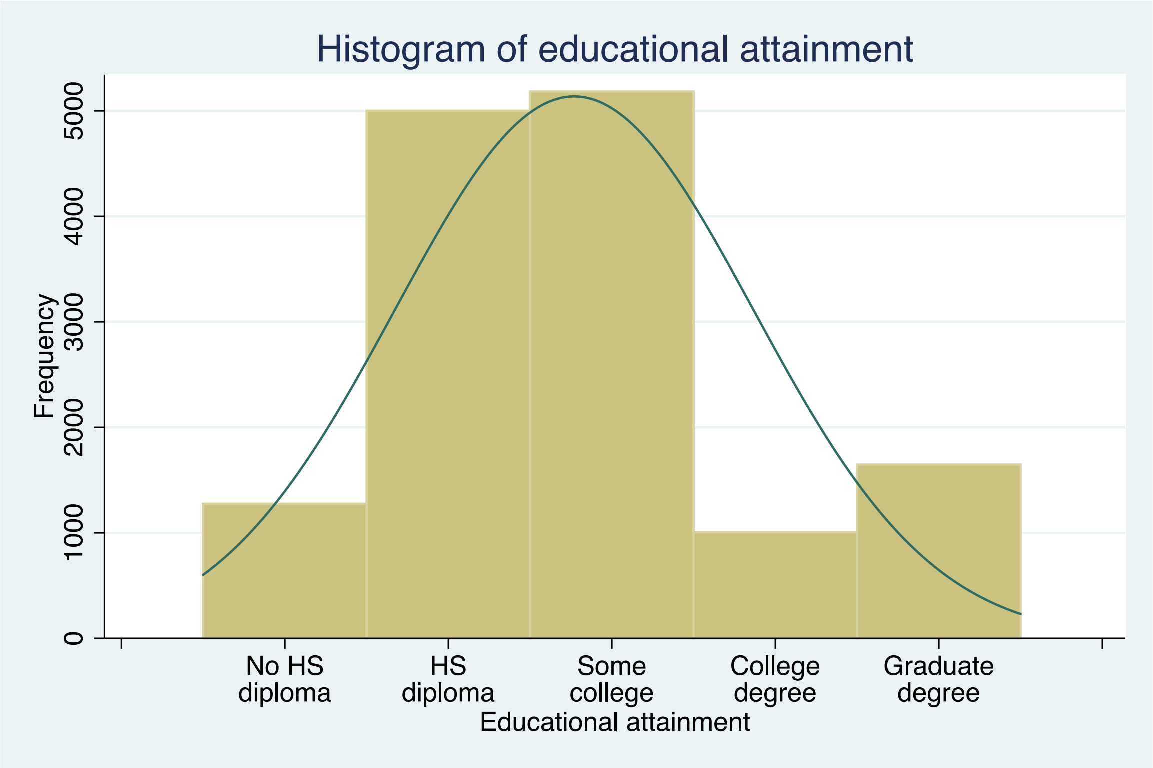
- When your dependent variable has ordered categories

Add Health

- Are natural mentoring relationship related to education?
  - 12th grade GPA
  - Educational attainment
    - Less than high school diploma
    - High school diploma
    - Some college
    - College degree
    - Graduate degree

Erickson LD, McDonald S, Elder GH, Jr,. 2009. "Informal mentors and educational achievement: Complementary or compensatory resources?" *Sociology of Education* 82: 344-367.


Histogram of 12th grade GPA
Frequency
0
200
400
600
800
0
1
2
3
4
12th grade GPA


Histogram of educational attainment
Frequency
0
1000
2000
3000
4000
5000
No HS diploma
HS diploma
Some college
College degree
Graduate degree
Educational attainment

```
svyset [pweight=gswgt3_2], psu(scid)

global vars calcage3 female private extra ib0.work ib1.race income nhooddis pta ptamis intact hpaed parelate
global vars1 tchrstud tchrstudmis class schsize numfr fgpa bcent10x appearance personality edasp pvt ogpa1

// 12-grade GPA
eststo t2a: svy, subpop(education): reg ogpa4 $vars $vars1
eststo t2b: svy, subpop(education): reg ogpa4 $vars $vars1 mentor
eststo t2c: svy, subpop(education): reg ogpa4 $vars $vars1 family friend teacher community

// Educational attainment
eststo t2d: svy, subpop(education): ologit degree $vars $vars1
eststo t2e: svy, subpop(education): ologit degree $vars $vars1 mentor
eststo t2f: svy, subpop(education): ologit degree $vars $vars1 family friend teacher community

estout t2a t2b t2c t2d t2e t2f, stats(N, fmt(%6.0fc)) eform(0 0 0 1 1 1) ///
    cells(b(star fmt(%9.3f))) nolz numbers mlabels(none) eqlabels(none) collabels(none) ///
    prehead("Table 2. Mentoring and 12th Grade GPA, Unstandardized Coefficients from OLS Regression") ///
    postfoot("Notes: *p < .05, **p < .01, *** p < .001; two-tailed tests")
```

Table 2. Influences on Educational Achievement and Attainment

	12th-Grade GPA[a]			Highest Degree Achieved[b]		
Variable	(1)	(2)	(3)	(4)	(5)	(6)
Mentoring						
Mentor		.103***			1.529***	
Social role						
Relative			.100***			1.501***
Friend			.059			1.400***
Teacher			.151***			1.987***
Community			.088**			1.303***
Constant	-.483	-.514	-.491			
N	[9,216]	[9,216]	[9,216]	[12,621]	[12,621]	[12,621]

[a]Unstandardized coefficients from the OLS regression.

[b]Odds ratios from the ordered logistic regression.

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$; two-tailed tests.

Ordered logistic regression in Mplus

- Specify as categorical, interpret as continuous

Categorical is Education;

Interpret coefficient as a regular regression coefficient

Negative binomial regression

- Use when the dependent variable is a count and the mean and variance of the dependent variable are not the same

Poisson, Negative Binomial, and Zero Inflated Distributions

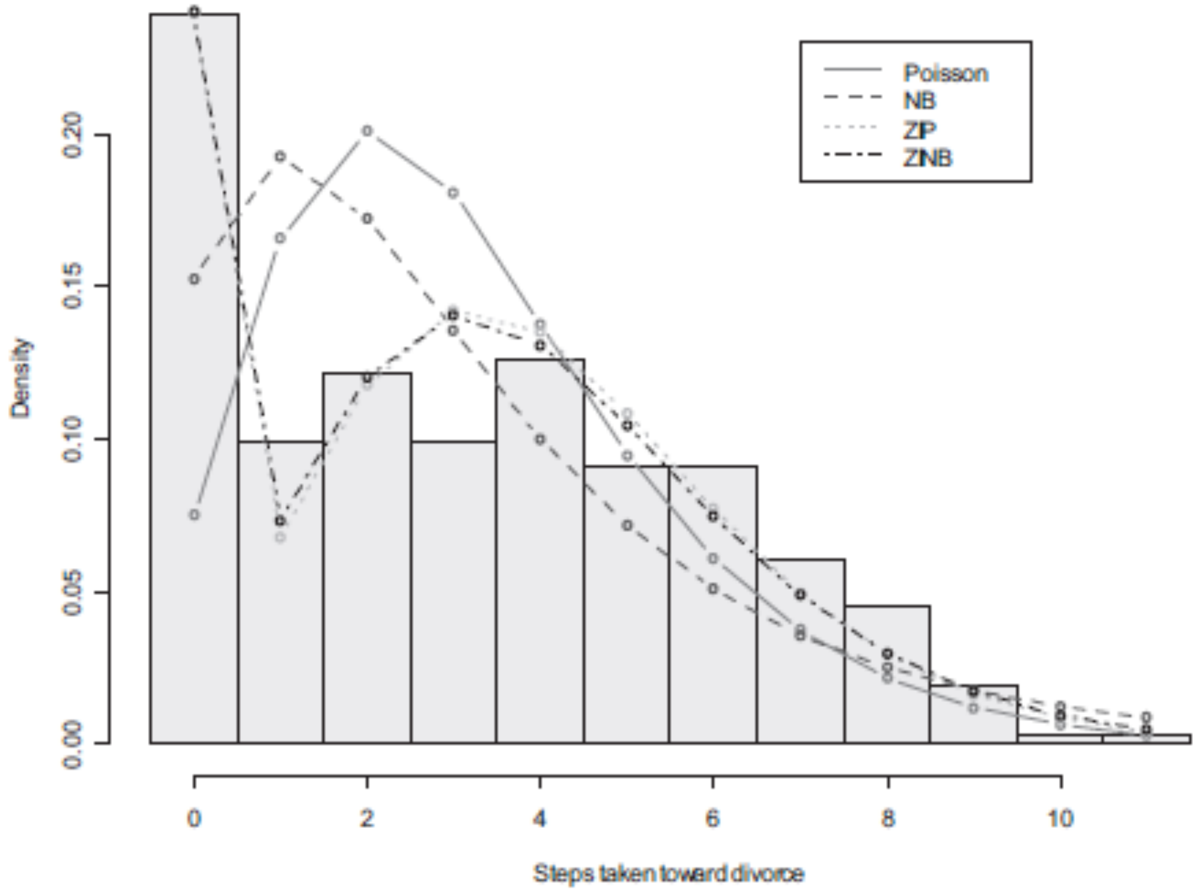


Figure 1. Histogram of Marital Status Inventory with predicted probabilities from regressions. NB = negative binomial; ZIP = zero-inflated Poisson; ZINB = zero-inflated negative binomial.

Poisson vs. Negative Binomial Model

- Steps to setting up these models

1. Determine if Poisson or Negative Binomial is better fit to the data (do a chi-square difference test - # of parameters difference and chi square distribution).
2. If lots of zeros, see if ZIP model is better than poisson or NB model (do a Vuong test, say in Stata).
3. See if ZINB model fits the data better than a ZIP model (especially if variance is greater than the mean – these models are nested, so you can once again do a chi-square difference test between the two models.).
   – Model diagnostics: If model doesn't converge, try changing the start values.
   – Use bootstrapping (1000 bootstraps recommended by Atkins and Gallop (2007) and compare the bootstrapped results with the original results
4. Interpret model output

Atkins, D.C., & Gallop, R.J. (2007). Rethinking how family researchers model infrequent outcomes: A tutorial on count regression and zero-inflated models. *Journal of Family Psychology, 21*, 726-735.

Interpreting Zero Inflated Model Output

- Zero Inflated Models
  - Interpret the "Zero" (#) portion of the model, with coefficients being logistic coefficients (pos = more likely to have a zero, neg = less likely to have a zero)
    - Exponentiate them and interpret as odds ratio
    - Or, transform into probabilities
  - Interpret predictors in the "count" portion of the model the same way you would with a count outcome in a poisson/negative binomial model.
    - That is, after transforming the output (100($e^{B*\delta}$ -1), then the coefficient represents the "percentage change in the expected counts"

Poisson vs. Negative Binomial in Mplus

- **Poisson**

Count is SOGSFREQ(p);

MODEL FIT INFORMATION

Number of Free Parameters	4
Loglikelihood	
H0 Value	-832.021
H0 Scaling Correction Factor for MLR	2.9153
Information Criteria	
Akaike (AIC)	1672.041
Bayesian (BIC)	1687.619
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	1674.929

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
SOGSFREQ ON				
REL	-0.010	0.006	-1.650	0.099
RACEDI	-0.072	0.149	-0.483	0.629
AGE	0.036	0.010	3.827	0.000
Intercepts				
SOGSFREQ	0.062	0.222	0.281	0.779

- **Negative Binomial**

Count is SOGSFREQ(nb);

MODEL FIT INFORMATION

Number of Free Parameters	5
Loglikelihood	
H0 Value	-675.264
H0 Scaling Correction Factor for MLR	0.8713
Information Criteria	
Akaike (AIC)	1360.528
Bayesian (BIC)	1380.000
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	1364.137

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
SOGSFREQ ON				
REL	-0.011	0.006	-1.774	0.076
RACEDI	-0.103	0.136	-0.752	0.452
AGE	0.048	0.015	3.122	0.002
Intercepts				
SOGSFREQ	-0.166	0.342	-0.487	0.626
Dispersion				
SOGSFREQ	1.310	0.167	7.867	0.000

ZIP vs. ZINB in Mplus

Zero Inflated Poisson

Count is SOGSFREQ(i);

MODEL FIT INFORMATION

Number of Free Parameters	8
Loglikelihood	
H0 Value	-701.169
H0 Scaling Correction Factor for MLR	1.5235
Information Criteria	
Akaike (AIC)	1418.337
Bayesian (BIC)	1449.492
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	1424.112

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
SOGSFREQ ON				
REL	0.000	0.006	0.055	0.956
RACEDI	0.010	0.125	0.083	0.934
AGE	0.022	0.009	2.514	0.012
SOGSFREQ#1 ON				
REL	0.026	0.011	2.303	0.021
RACEDI	0.201	0.245	0.823	0.411
AGE	-0.046	0.027	-1.678	0.093
Intercepts				
SOGSFREQ#1	-0.032	0.629	-0.051	0.959
SOGSFREQ	0.639	0.216	2.963	0.003

Zero Inflated Negative Binomial

Count is SOGSFREQ(nbi);

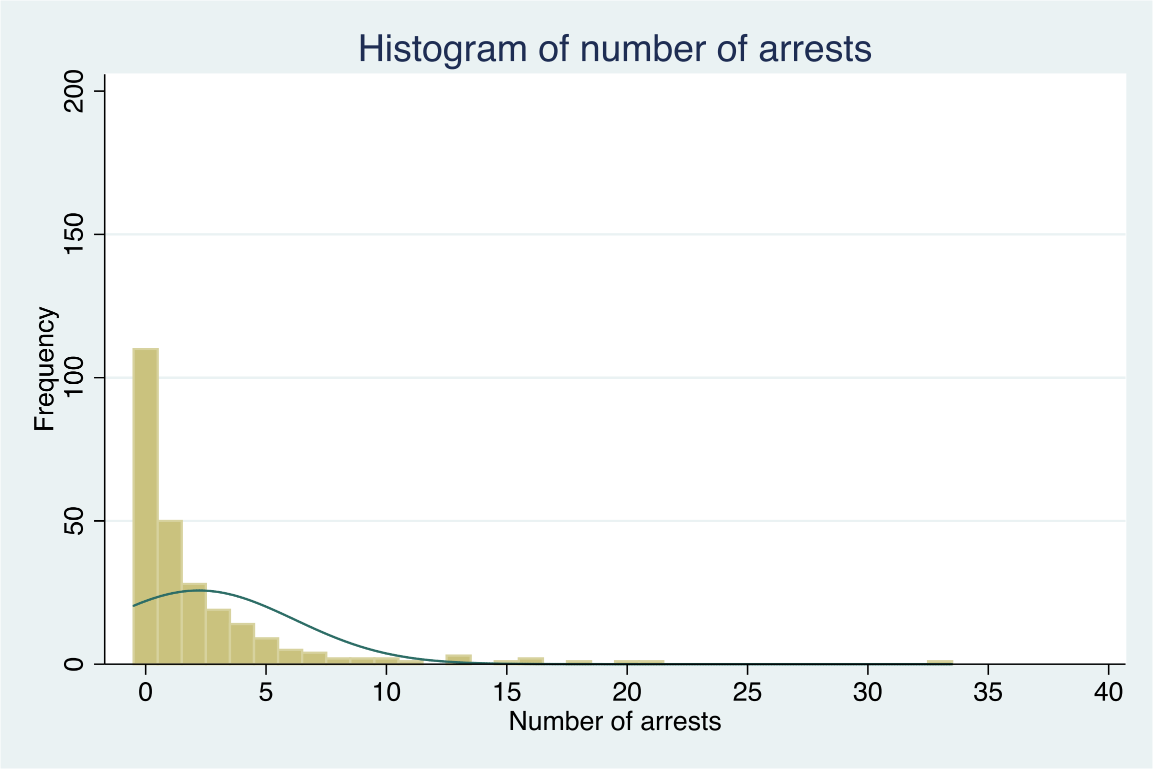
MODEL FIT INFORMATION

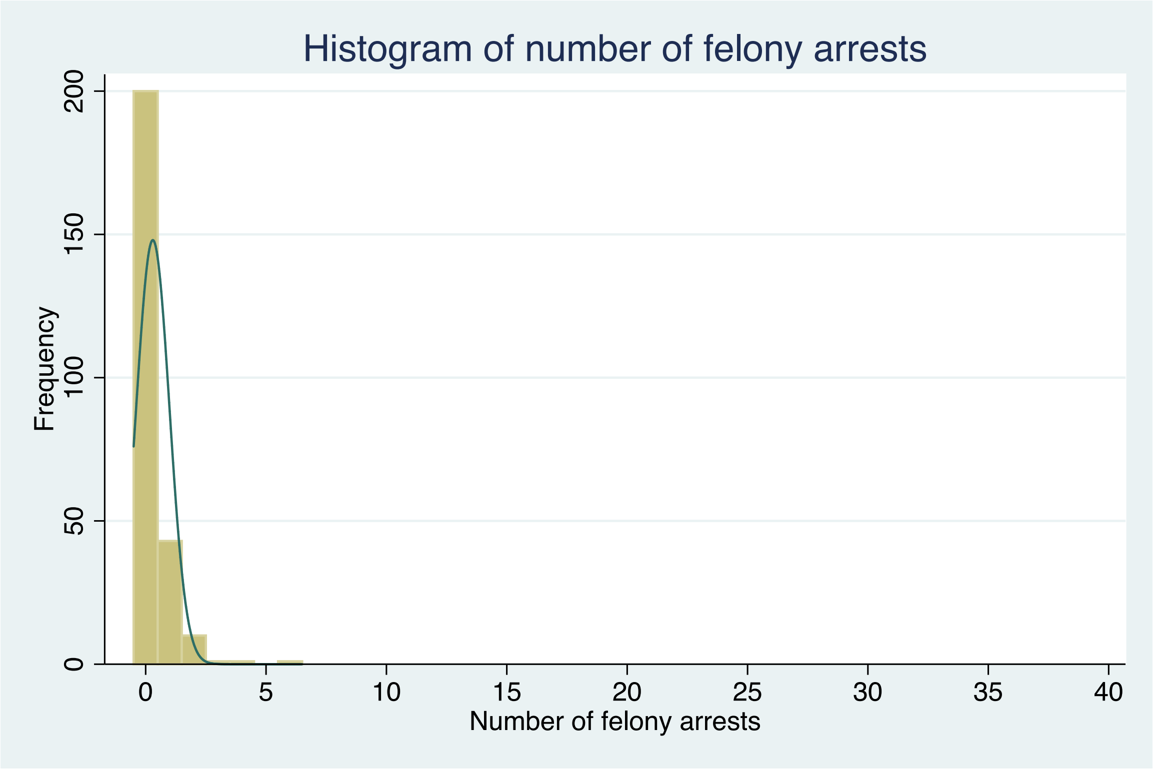
Number of Free Parameters	9
Loglikelihood	
H0 Value	-665.339
H0 Scaling Correction Factor for MLR	0.9423
Information Criteria	
Akaike (AIC)	1348.679
Bayesian (BIC)	1383.728
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	1355.175

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
SOGSFREQ ON				
REL	0.003	0.007	0.384	0.701
RACEDI	0.078	0.140	0.558	0.577
AGE	0.048	0.018	2.624	0.009
SOGSFREQ#1 ON				
REL	0.126	0.052	2.400	0.016
RACEDI	1.893	1.682	1.126	0.260
AGE	-0.007	0.060	-0.119	0.905
Intercepts				
SOGSFREQ#1	-6.124	3.865	-1.584	0.113
SOGSFREQ	-0.345	0.389	-0.888	0.374
Dispersion				
SOGSFREQ	0.924	0.203	4.551	0.000

How many arrests in the next year?





```
quietly {
     eststo t5a:  nbreg totay1  treat    age    male   white totprior felprior ib1.program            , irr
     eststo t5a1: nbreg totay1  numcalls age    male   white totprior felprior ib1.program if treat==1, irr
     eststo t5b:  nbreg totfay1 treat    age    male   white totprior felprior ib1.program            , irr
     eststo t5b1: nbreg totfay1 numcalls age /*male*/ white totprior felprior ib1.program if treat==1, irr
}


estout t5a t5a1 t5b t5b1, ///
     eform nolz cells(b(star fmt(3))) starlevels(* .10 ** .05 *** .001) ///
     stats(N, fmt(0)) collabel(none) eqlabel(none) drop(_cons) ///
     mlabel("Model 1""Model 2^a""Model 3""Model 4^a,b") ///
     mgroups("Any Arrest" "Any Felony Arrest", pattern(1 0 1 0) span) ///
     prehead("Table 5." ///
             "The Relationship of Treatment and Number of Calls with Total Arrests and Total Felony Arrests:" ///
             "Incidence Rate Ratios from Negative Binomial Regression") ///
     varlabel(treat "Treatment" exposure "    Posttreatment exposure days^c" age "    Age" ///
              male "    Male" white "    White" totprior "        Any" felprior "        Felony" ///
              1.program "        Juvenile probation" 2.program "        Rural" ///
              3.program "        Secure care" _cons "Constant" numcalls "Number of calls^c") ///
     refcat(age "Controls" totprior "    Number of previous arrests" 1.program "    Program site", label(" ")) ///
     postfoot("Note:" ///
              "^a Only treatment group included in analysis." ///
              "^b No females in the treatment group had a felony arrest." ///
              "^c In 100s. * p < .05, ** p < .01, ***p < .001, one-tailed tests.") ///
     varwidth(31)
```

Table 4. The Relationship of Treatment and Number of Calls with Total Arrests and Total Felony Arrests: Incident Rate Ratios (IRR) from Negative Binomial Regression.

	Total arrests		Total felony arrests	
	Model 1	Model 2[a]	Model 3	Model 4[a,b]
Treatment	1.064		0.944	
Number of calls[c]		0.923		0.644*
Controls				
Age	0.677***	0.689***	0.687***	0.694**
Male	1.360	1.371	1.537	
White	1.254	1.176	1.760**	1.453
Number of pretreatment arrests				
Any	1.022**	1.005	1.027**	1.018
Felony	0.941**	0.927*	0.961	0.988
Program site				
Probation	1.000	1.000	1.000	1.000
Rural	0.360***	0.218***	0.977	0.615
Secure care	0.357***	0.324**	2.158*	1.960
N	256	136	256	136

Note. Incident Rate Ratios (IRR) are the antilog of the model coefficient and represent the rate of change in arrests for a one-unit increase in the independent variable. For example, the IRR for number of calls in Model 4 is .644, which indicates that the incidence of arrests would decrease by a factor .644 if the number of calls increases by one unit (100 calls in this case). A commonly used alternative interpretation transforms the IRR into a percentage—$(1 - 0.644) \times 100 = 35.6$—indicating, in this case, that an increase of 100 calls reduces the rate of felony arrests by 35.6%.

[a]Only treatment group included in analysis.

[b]No females in the treatment group had a felony arrest.

[c]In 100s.

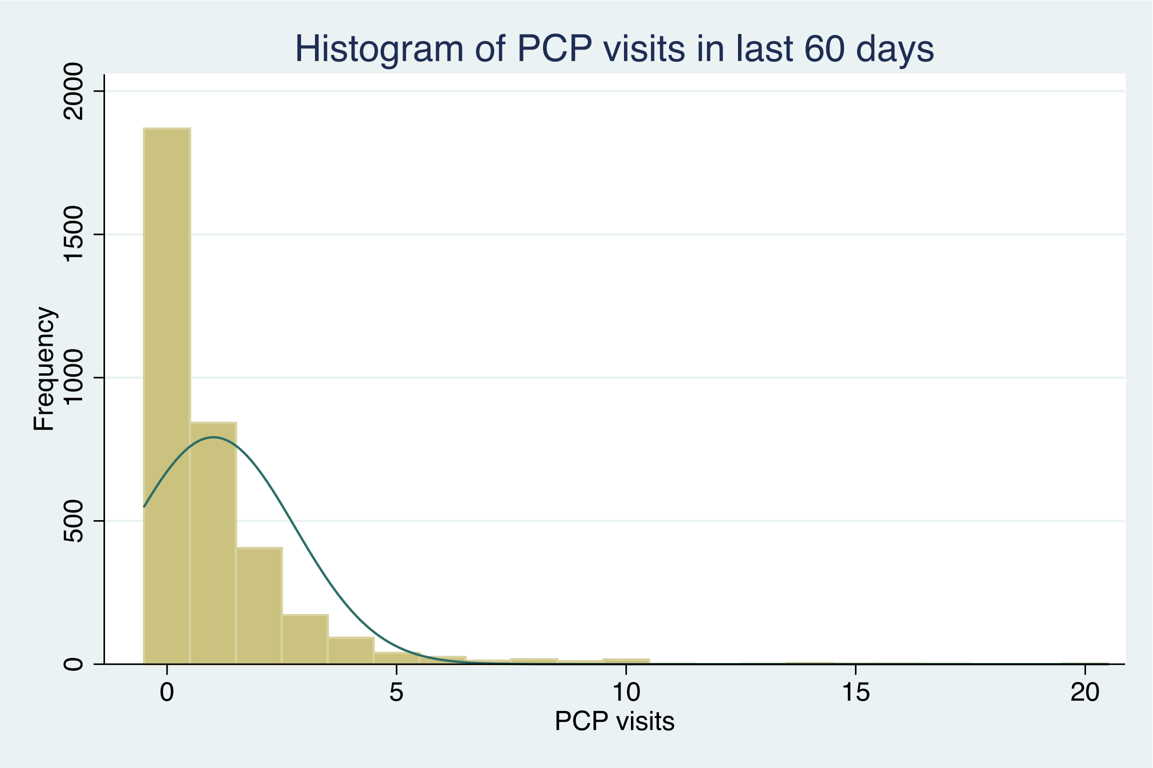
*$p < .05$. **$p < .01$. ***$p < .001$; one-tailed tests.

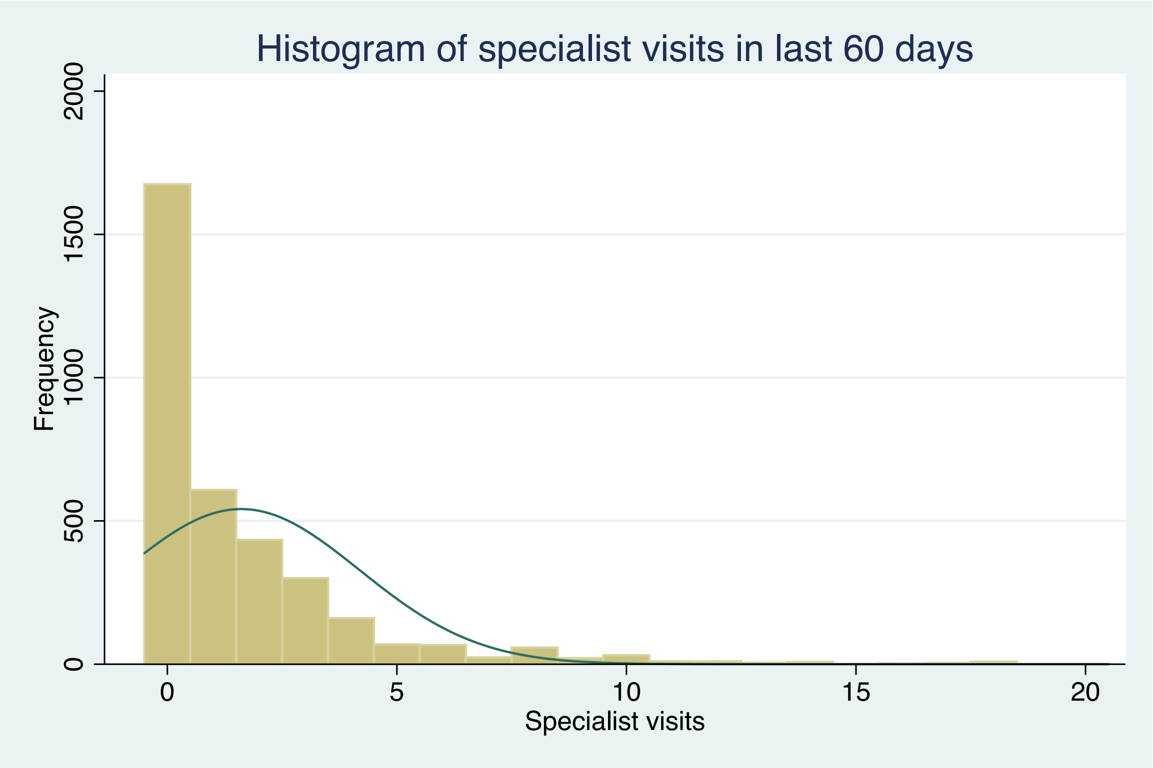
Zero-inflated Poisson

- Use when you have a count dependent variable with an equal mean and variance once you account for an excess of 0's.

Montana Health Matters

- Predictors of the number of doctor visits in the last 60 days
  - Primary care physical (PCP)
  - Specialist


Histogram of PCP visits in last 60 days
Frequency
0
500
1000
1500
2000
0
5
10
15
20
PCP visits


Histogram of specialist visits in last 60 days
Frequency
0
500
1000
1500
2000
0
5
10
15
20
Specialist visits

```
mi svyset zipcode [pweight=hhweight_n], strata(strata) singleunit(certainty) || houseid

local vars age female white married ib1.educ inc1k ib0.vetenrol pc_dist ib1.rural

eststo t2a: mi estimate, post cmdok: svy: zip drvsts `vars', inflate(`vars')
eststo t2b: mi estimate, post cmdok: svy: zip spvsts `vars', inflate(`vars')

estout t2a t2b, drop(_cons) ///
     cells(b(star fmt(3))) eform nolz unstack eqlabel(Main Inflate) collabel(none) mlabel(PCP Specialist) ///
     prehead("Table 2." ///
            "Understanding predictors of recent doctor visits:" ///
            "Exponentiated coefficients from zero-inflated Poisson model") ///
     varlabels(age "Age (in years)" female "Female" white "White" married "Married" 1.educ "     No HS degree" ///
              2.educ "     HS degree" 3.educ "     Some college" 4.educ "     College degree" ///
              5.educ "     Graduate degree" inc1k "Income (in $1,000)" 0.vetenrol "     Non-veteran" ///
              1.vetenrol "     VA enrolled" 2.vetenrol "     VA non-enrolled" 1.rural "     Urban" ///
              2.rural "     Rural" 3.rural "     Highly rural" pc_dist "Distance to PC" _cons "Constant") ///
     refcat(age "Pre-disposing characteristics" educ1 "Educational attainment" ///
           pc_dist "Accessibility" 1.rural "Rurality" , label(" ")) ///
     postfoot("Note: N = `:di %5.0fc `=e(N)''. *p < .05, **p < .01, *** p < .001; two-tailed tests" ///
             "Source: {it:Montana Health Matters}.") ///
     varwidth(30)
```

Table 2.
Understanding predictors of recent doctor visits:
Exponentiated coefficients from zero-inflated Poisson model

	PCP		Specialist	
	Main	Inflate	Main	Inflate
Pre-disposing characteristics				
Age (in years)	.998	.975***	.976***	.999
Female	1.317**	.828	.785*	1.156
White	.706**	1.039	1.167	1.014
Married	1.243*	1.201	.803*	1.113
No HS degree	1.000	1.000	1.000	1.000
HS degree	.808	.888	1.288	.930
Some college	1.123	1.424	1.279	1.068
College degree	.716	1.010	1.332	.946
Graduate degree	.797	.899	.837	.997
Income (in $1,000)	.995**	.997	1.000	.999
Non-veteran	1.000	1.000	1.000	1.000
VA enrolled	1.751***	.538*	.587*	1.213
VA non-enrolled	.898	.485**	.835	1.205
Accessibility				
Distance to PC	1.000	.999	.997	1.000
Rurality				
Urban	1.000	1.000	1.000	1.000
Rural	.916	1.090	1.042	1.095
Highly rural	1.010	.999	1.134	1.137

Note: N = 3,512. *p < .05, **p < .01, *** p < .001; two-tailed tests
Source: *Montana Health Matters*.