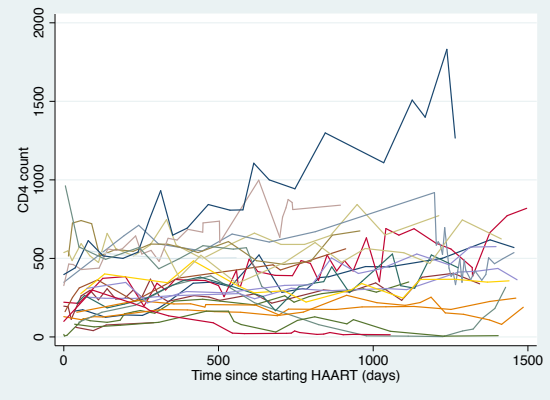


Longitudinal Data
Fall 2015

Chapter 4

Longitudinal Outcome, Baseline Covariates

Instructors

Nick Jewell (*jewell@berkeley.edu*)

GSI

Robin Mejia (*mejia@nasw.org*)

Theme of Chapter

- In situations where:
  - Only "baseline covariates"
  - No expectation of any trends with time,

- One available method of regression is to reduce the set of outcomes on a subject, *i*, to one summary measure:

$$Y_i = g(Y_{i1}, Y_{i2}, \ldots, Y_{in_i}) \overset{ex}{=} \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

Theme of Chapter, cont.

- Perform appropriate regression on the summary measure, for example:

$$E[Y_i \mid \mathbf{X}_{i1}] = b_0 + b_1 X_{i11} + b_2 X_{i12} + \ldots + + b_p X_{i1p}$$

- If different subjects have different numbers of observations (i.e., n_i are not all the same), one might want to use weighted regression, for instance weighting by $wt_i = n_i$.

HIV Data Example

- HIV example: the association of average CD4 count in the future with baseline viral load.

$$E[Y_i \mid X_i] = b_0 + b_1 X_i$$

where Y_i is average CD4 over future measurements and X_i is the baseline viral load measurement.

Sample of HIV+ Data

Table 1.1: Extract of Data from SFGH/HAART Stu

[1] Id. Number	days	CD4 count	log(viral load)	gender	age
1	39	45	2.70	1	32.0
1	137	119	5.22	1	32.0
1	147	113	.	1	32.0
1	179	74	5.20	1	32.0
1	187	95	.	1	32.0
1	298	137	3.87	1	32.0
1	335	.	5.07	1	32.0
1	354	167	5.14	1	32.0
1	411	.	4.66	1	32.0
1	1684	427	.	1	32.0
2	0	196	5.68	1	44.0
2	7	369	3.93	1	44.0
2	13	353	4.11	1	44.0
2	27	474	3.55	1	44.0
2	55	425	3.10	1	44.0
2	111	493	2.70	1	44.0
2	139	464	2.70	1	44.0
2	167	448	2.70	1	44.0
2	195	427	2.70	1	44.0

STATA commands to process data

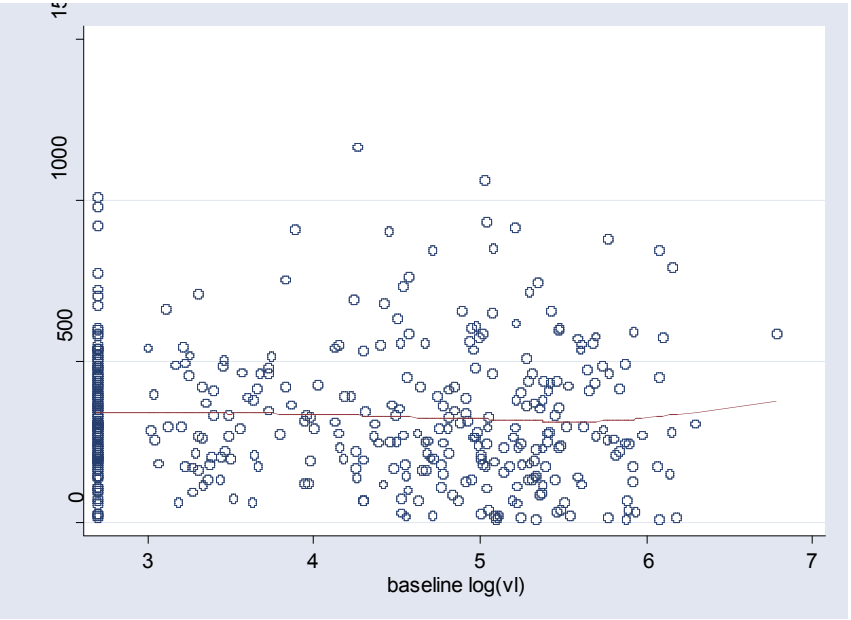
```
label variable cd4 "CD4 count"
label variable etime "Time since HAART (days)"
label variable logvl500 "Truncated log(viral load)"
*** Drop if any visits before baseline
drop if etime < 0

*** Get time of first visit
capture drop basetime
egen basetime = min(etime), by(id)

*** Define Elapsed Time
capture drop elapse
gen elapse = etime-basetime

*** Get Baseline Log(Viral Load)
capture drop logvlb2
gen logvlb2 = logvl500
replace logvlb2 =. if etime != basetime
capture drop logvlbase
egen logvlbase = mean(logvlb), by(id)
```

```
***** Average CD4 in future versus baseline viral load
drop if cd4 == .
capture drop cd4notb
gen cd4notb = cd4
replace cd4notb = . if elapse == 0
capture drop avecd4
egen avecd4 = mean(cd4notb), by(id)
*** Track Number of Observations per id
sort id elapse
capture drop cntid
quietly by id: gen cntid = _n
quietly by id: gen totid = _N
** List data for regressions below
list avecd4 logvlbase totid if _n ==1
```



	id	avecd4	logvlb~e	totid
2.	1	278.2174	2.69897	24
30.	2	552.2917	5.679697	25
53.	3	32.54546	5.937819	12
68.	4	84.83334	5.36135	7
90.	5	129.7333	5.31848	31
120.	6	360.1818	4.780245	22
132.	7	437.5807	5.4669	32
155.	8	197.1111	5.01174	10

Unweighted Regression

```
. regress avecd4 logvlbase if cntid==1

      Source |       SS       df       MS              Number of obs =     406
-------------+------------------------------           F(  1,   404) =    0.84
       Model |  36036.2407     1  36036.2407           Prob > F      =  0.3590
    Residual |  17266901.1   404  42739.8542           R-squared     =  0.0021
-------------+------------------------------           Adj R-squared = -0.0004
       Total |  17302937.3   405   42723.302           Root MSE      =  206.74

------------------------------------------------------------------------------
      avecd4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   logvlbase |  -8.007322   8.720353    -0.92   0.359    -25.15026    9.135612
       _cons |   360.9837   37.15651     9.72   0.000     287.9395     434.028
------------------------------------------------------------------------------
```

Weighted Regression (by number of observations used to estimate future ave(CD4) in future)

```
. ** Weighted by Number, not robust
. capture drop wt1

. gen wt1 = totid-1

. regress avecd4 logvlbase if cntid==1 [aweight=wt1]
(sum of wgt is   7.5850e+03)

      Source |       SS       df       MS              Number of obs =     406
-------------+------------------------------           F(  1,   404) =    0.00
       Model |  98.0869708     1  98.0869708           Prob > F      =  0.9604
    Residual |  16091955.4   404  39831.5728           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared = -0.0025
       Total |  16092053.5   405  39733.4655           Root MSE      =  199.58

------------------------------------------------------------------------------
      avecd4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   logvlbase |   .4245909   8.556156     0.05   0.960    -16.39556    17.24474
       _cons |   334.2507   37.60463     8.89   0.000     260.3255    408.1758
------------------------------------------------------------------------------
```

Reduction of repeated binary measures data to a simpler outcome structure:

Analyzing Count Data

When is reducing longitudinal outcome data to a summary measure relevant?

- Repeated yes/no measurements on subjects.
- Interested in the association of baseline variables and the outcome of interest.
- Does not make sense to do if there are time-dependent covariates of interest.
- Summing the number of events and then modeling counts as function of baseline covariates.

Example - Water Intervention Trials

- Subjects randomized to either device which filters out pathogens or a similar looking placebo devices.
- Subjects record daily whether or not they have a gastro-intestinal episode (yes/no)
- Purpose is to determine the amount of GI illness attributable to drinking water.

What the data look like?

	id	date	hcgi	group
1.	A7283	14780	.	6
2.	A7283	14781	0	6
3.	A7283	14782	0	6
4.	A7283	14783	0	6
5.	A7283	14784	0	6
6.	A7283	14785	0	6
7.	A7283	14786	0	6
17.	A7283	14796	0	6
225.	C1632	14738	.	7
226.	C1632	14739	.	7
227.	C1632	14740	.	7
228.	C1632	14741	0	7
229.	C1632	14742	0	7
230.	C1632	14743	0	7
231.	C1632	14744	1	7
232.	C1632	14745	0	7
233.	C1632	14746	0	7
234.	C1632	14747	0	7
235.	C1632	14748	0	7
237.	C1632	14750	0	7
238.	C1632	14751	1	7

How to reduce data?

- Sum up the number of episodes to make an overall count.
- In the water trial example, calculate the number of GI episodes per person.
- In notation, if Y_{ij} is the jth measurement on the ith person and Y_{ij} = *0* (no) or *1* (yes), then make a new variable Y_i,

$$Y_i = \sum_{j=1}^{n_i} Y_{ij}$$

- Note, we will ultimately allow for different number of time intervals (n_i) among subjects.

What the data look like after reduction?

	id	hcgi	daysatrisk	group
1.	A7283	0	111	6
2.	C1632	3	89	7
3.	C2412	3	7	7
4.	C2515	5	29	7
5.	C2771	1	104	6
6.	C4722	0	112	6
7.	D1959	2	79	7
8.	D3531	0	111	6
9.	E1000	2	11	6
10.	E8776	0	112	6
11.	F4246	0	110	7
12.	G3700	0	112	7
13.	G4393	1	103	6
14.	H1438	0	112	6
15.	H1961	3	85	7
16.	H6003	1	106	7
17.	H6995	0	112	7

Poisson Distribution

- Used to model counts that can vary from 0 to ∞.
- Events occur independently and at a rate λ.

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

- Note that *y!=y*(y-1)* ... 2*1*.
- λ is sometimes called rate, and is both the mean ($E[Y_i]$) and variance ($var[Y_i]$)

Other examples of Poisson Data

- Number of automobile fatalities in a given region over a year.
- Number of AIDS cases for a given risk group for a month
- Number of earthquakes of a given size range in a region by decade
- Number of people infected by an infective

Example of simple Poisson Calculation

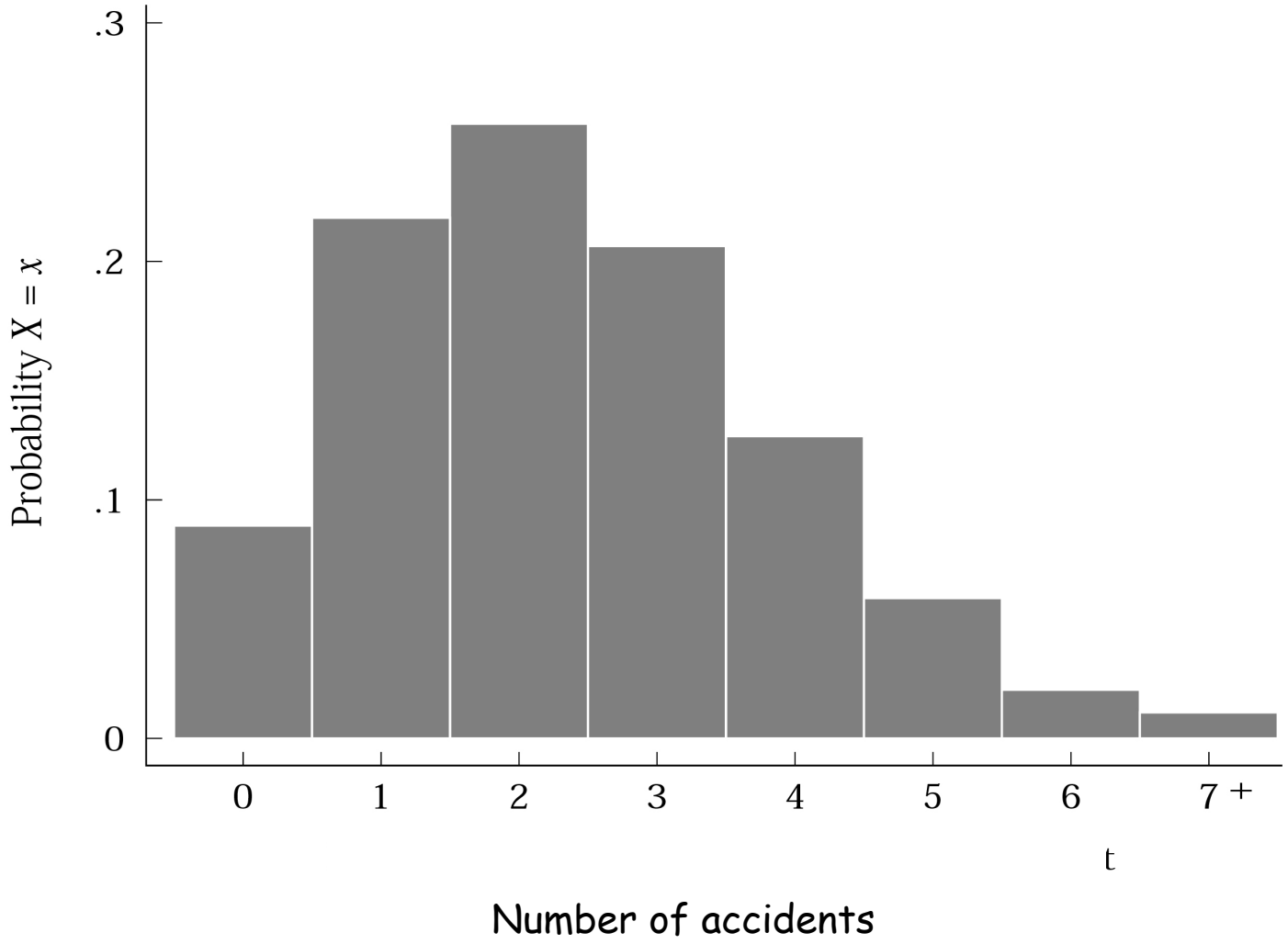
- Assume the accident rate is 0.00024/person/year.
- Then, the mean rate (λ) for a town of 10,000 people is 10,000*0.00024 = 2.4/year.

$$P(Y=0) = \frac{2.4^{0} e^{-2.4}}{0!} = 0.097$$

$$P(Y=1) = \frac{2.4^{1} e^{-2.4}}{1!} = 0.2177$$

$$P(Y>1) = 1 - P(Y=0) - P(Y=1) = 0.6916$$

Poisson with λ=2.4



Estimating the Poisson Parameter, λ

- Assume one observed *m* independent counts (all over same length of time for now) all from the same Poisson distribution.
- The maximum likelihood estimate of λ is:

$$\hat{\lambda} = \frac{Y_1 + Y_2 + \cdots + Y_m}{m}$$

- Just the average number of counts.

Number of goals scored per team per game in the 1966 world cup football competition.

Number of goals	Observed	Expected
0	18	15.9
1	20	22.2
2	15	15.4
3	7	7.1
4	2	2.5
5	2	0.7
6+	0	0.2

Calculations--1966 World Cup

- 32 games, so 64 counts of number of goals/ game for one team.
- Distribution as shown on previous slide.
- Total number of goals = 89.
- So $\hat{\lambda} = \frac{89}{64} = 1.39$ goals/game/team.

- Estimated prob. of no goals for one team

$$P_{\hat{\lambda}}(Y = 0) = \frac{\left(1.39^0 e^{-1.39}\right)}{0!} = 0.25$$

- So, expected number of times in 64 team scores we see zero goals

$$m * P_{\hat{\lambda}}(Y = 0) = 64 * 0.25 = 15.9.$$

Two Sample Problem

- As an example, use the water intervention trials.
- Let *X=0* (placebo) or *1* (active device)
- Consider the model for the mean rate

$$\lambda(x) = \exp(\beta_0 + \beta_1 x)$$

- Why model the log rate and not the rate?

Interpretation of Regression Coefficients

- The mean rate when *X=0* is:

$$\lambda(0) = \exp(\beta_0)$$

- The mean rate when *X=1* is:

$$\lambda(1) = \exp(\beta_0 + \beta_1)$$

- So, β_1 = *log(λ(1))-log(λ(0))* or,

  - $exp(\beta_1) = \lambda(1)/\lambda(0)$ = ratio of means, sometimes called incident rate ratio (IRR).

Interpretation and Estimation of Regression Coefficients in the Two Sample Problem

- In general (assuming no interaction terms), the non-intercept coefficients represent the log(IRR) for a unit increase of the corresponding covariate.

- Estimating rates in two groups

$$\hat{\lambda}(0) = \frac{\sum_{i=1}^{m} Y_i(1 - X_i)}{\sum_{i=1}^{m} (1 - X_i)} \qquad \hat{\lambda}(1) = \frac{\sum_{i=1}^{m} Y_i X_i}{\sum_{i=1}^{m} X_i}$$

or the average in the two groups.

Interpretation and Estimation of Regression Coefficients in the Two Sample Problem

- Finally,

$$\hat{\beta}_0 = \log\left(\hat{\lambda}(0)\right)$$

$$\hat{\beta}_1 = \log\left(\hat{\lambda}(1)\right) - \log\left(\hat{\lambda}(0)\right)$$

- In general (many covariates), use MLE (involves iterative maximization algorithm).

Different Follow-up Periods

- More typically, people are followed for different time periods, as in the water intervention trials.
- Suppose rate per unit time is assumed to be a constant, λ, and follow-up for an individual is of length, T.
- Then, counts of events in period $[0,T]$ is then Poisson with rate parameter $\lambda * T$.

Estimating the Poisson Parameter, λ, with differing follow-up periods.

- Again, assume one observed *m* independent counts $(Y_1,\ldots,Y_m)$ with corresponding follow-up periods $(T_1, T_2, \ldots, T_m)$.

- The MLE estimate of the rate parameter is:

$$\hat{\lambda} = \frac{Y_1 + Y_2 + \cdots + Y_m}{T_1 + T_2 + \cdots T_m} = \frac{\frac{1}{m}\sum_{i=1}^{m} Y_i}{\frac{1}{m}\sum_{i=1}^{m} T_i} = \frac{\text{average number of events}}{\text{average years of observation}}$$

Poisson Regression (allowing for differing follow-up periods)

- Consider the same covariate (*X=0,1* corresponds to placebo, active) and underlying "model":

$$\lambda(x) = \exp(\beta_0 + \beta_1 x)$$

- Then, over a follow-up period of length *T*, the mean rate is:

$$\lambda(x,T) = T\exp(\beta_0 + \beta_1 x),$$

or,
$$\log(\lambda(x,T)) = \log(T) + \beta_0 + \beta_1 x.$$

- Thus, *log(T)* is special covariate with coefficient fixed at 1, sometimes called an *offset*.

What the data look like?

	id	date	hcgi	group
1.	A7283	14780	.	6
2.	A7283	14781	0	6
3.	A7283	14782	0	6
4.	A7283	14783	0	6
5.	A7283	14784	0	6
6.	A7283	14785	0	6
7.	A7283	14786	0	6
17.	A7283	14796	0	6
225.	C1632	14738	.	7
226.	C1632	14739	.	7
227.	C1632	14740	.	7
228.	C1632	14741	0	7
229.	C1632	14742	0	7
230.	C1632	14743	0	7
231.	C1632	14744	1	7
232.	C1632	14745	0	7
233.	C1632	14746	0	7
234.	C1632	14747	0	7
235.	C1632	14748	0	7
237.	C1632	14750	0	7
238.	C1632	14751	1	7

- group2 (*X*) is tx variable (*X=0* implies filtered, 1 sham)
- hcgiyrs is the follow-up time in years (*T*).

Example - Water Intervention Trial in STATA

```
poisson hcgieps group2, exposure(hcgiyrs)


Poisson regression                                Number of obs   =         42
                                                  LR chi2(1)      =       2.38
                                                  Prob > chi2     =     0.1233
Log likelihood = -94.329365                       Pseudo R2       =     0.0124

------------------------------------------------------------------------------
     hcgieps |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      group2 |   .4646641    .307828     1.51   0.131    -.1386677    1.067996
       _cons |   1.372013         .25     5.49   0.000     .8820223    1.862004
 ln(hcgiyrs) |          1  (exposure)
------------------------------------------------------------------------------
```

$$\log(\hat{\lambda}(x,T)) = \log\left[\hat{E}(Y \mid X = x, T)\right] = \log(T) + \hat{\beta}_0 + \hat{\beta}_1 x.$$

```
. poisson, irr
hcgieps |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      group2 |   1.591479   .4899019     1.51   0.131     .8705173    2.909542
------------------------------------------------------------------------------
```

$$\exp(\hat{\beta}_1)$$

Extending the Poisson model - Negative Binomial Regression

- Poisson regression assumes that all subjects with the same covariates have the same underlying rate.

- However, situations often arise when we expect that individual's rates vary around a mean rate (at random); that is, individuals have their own inherent propensity for events to occur

- For instance, we might thing of the population within a specific treatment group as having a mean rate of HCGI, but the individual's rates vary around this mean.

Extending the Poisson model - Negative Binomial Regression

- Let subject i have a rate of HCGI of λ_i, a distribution of diarrhea rates in some interval of time (say a day) of:

$$P(Y_i = y) = \frac{\lambda_i{}^{y} e^{-\lambda_i}}{y!}$$

and the mean rate of all subjects might be: $E(\lambda_i) = \lambda$.

- A model for the marginal counts might assume a particular distribution of λ_i in the population.

- If this underlying distribution (mixture) of rates within covariate groups is gamma distributed, then the *marginal* distribution of rates is negative binomial.

Gamma Distribution

- Specifically, if the rates are gamma (shape = ν, scale= λ/ν) distributed in the population.

$$f(\lambda_i;\nu,\lambda) = \lambda_i^{\nu-1} \frac{e^{-\lambda_i \nu / \lambda}}{\left(\frac{\lambda}{\nu}\right)^{\nu} \Gamma(\nu)}$$

which has mean, λ, and variance λ^2/ν, then Y is distributed $NB(\lambda, \nu)$,

Negative Binomial Distribution

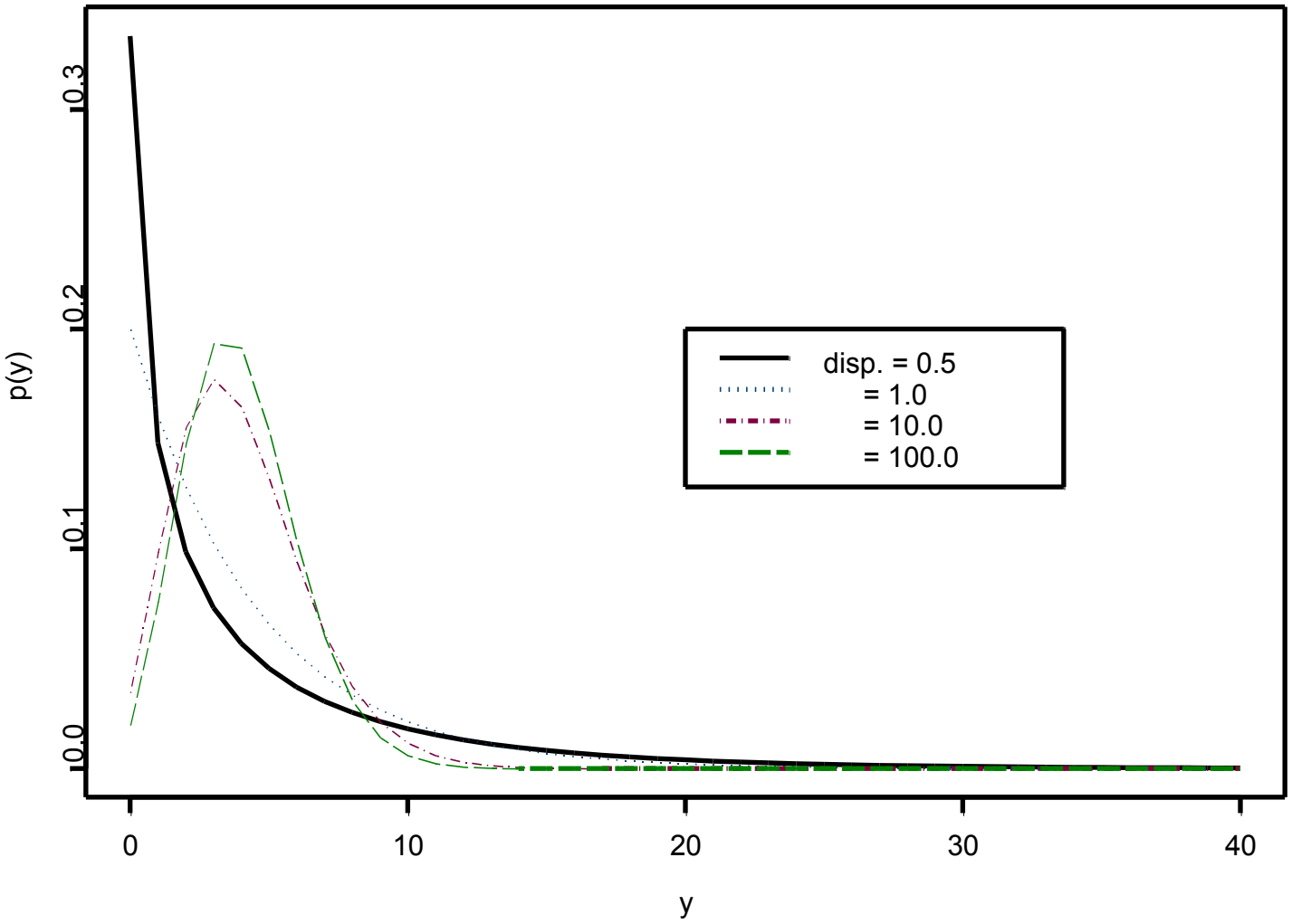
- Used to model counts that can vary from 0 to ∞.
- Two-parameter distribution

$$P(Y = y) = \left(\frac{\nu}{\nu + \lambda}\right)^{\nu} \frac{\Gamma(\nu + y)}{\Gamma(y+1)\Gamma(\nu)} \left(\frac{\lambda}{\nu + \lambda}\right)^{y}$$

- Note that $\Gamma(x)$ represents the gamma function - if x is an integer, then $\Gamma(x+1)=x!$.
- λ is the mean, and $(\lambda + \lambda^2/\nu)$ is the variance, so as ν is smaller, then variance is larger.
- ν is sometimes referred to as the dispersion parameter

Negative Binomial Distribution

λ=4.0



Negative Binomial Regression

- Consider the same covariate (*x=0,1* - placebo, active) and the same underlying model:

$$\lambda(x) = \exp(\beta_0 + \beta_1 x)$$

- We can fit this same model (for the mean) using MLE and the negative binomial distribution.

Example 2 - Negative Binomial Regression in STATA

```
. . nbreg hcgieps group2, exposure(hcgiyrs)

Negative binomial regression                    Number of obs   =         42
                                                LR chi2(1)      =       0.16
Dispersion     = mean                           Prob > chi2     =     0.6870
Log likelihood = -68.382747                     Pseudo R2       =     0.0012

------------------------------------------------------------------------------
     hcgieps |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      group2 |   .2937139   .7206237     0.41   0.684    -1.118683     1.70611
       _cons |   2.048749   .5712323     3.59   0.000     .9291542    3.168344
ln(hcgiyrs) |          1  (exposure)
-------------+----------------------------------------------------------------
    /lnalpha |   1.262019   .3526833                      .5707725    1.953266
-------------+----------------------------------------------------------------
       alpha |   3.532547    1.24587                      1.769633    7.051679
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) =   51.89 Prob>=chibar2 = 0.000 nbreg hcgieps
group2, exposure(hcgiyrs)
```

$\hat{\beta}_1$ $\hat{\beta}_0$

alpha here is simply 1/ ν

so that the larger alpha is the more extra-Poisson variation exists

```
. lincom group2, irr

 ( 1)  [hcgieps]group2 = 0

------------------------------------------------------------------------------
     hcgieps |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |     1.3414   .9666446     0.41   0.684     .3267099    5.507497
------------------------------------------------------------------------------
```

Difference in results?

- With Poisson regression, the IRR was estimated to be 1.59 (95% CI: 0.87—2.91)
- Negative Binomial Regression yields an estimated IRR of 1.34 (95% CI: 0.33—5.51)

Example - Water Intervention Trial in STATA

```
poisson hcgieps group2, exposure(hcgiyrs)


Poisson regression                                Number of obs   =         42
                                                  LR chi2(1)      =       2.38
                                                  Prob > chi2     =     0.1233
Log likelihood = -94.329365                       Pseudo R2       =     0.0124

------------------------------------------------------------------------------
     hcgieps |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      group2 |   .4646641    .307828     1.51   0.131    -.1386677    1.067996
       _cons |   1.372013        .25     5.49   0.000     .8820223    1.862004
 ln(hcgiyrs) |          1  (exposure)
------------------------------------------------------------------------------
```

$$\log(\hat{\lambda}(x,T)) = \log\left[\hat{E}(Y \mid X = x, T)\right] = \log(T) + \hat{\beta}_0 + \hat{\beta}_1 x.$$

```
. poisson, irr
hcgieps |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      group2 |   1.591479   .4899019     1.51   0.131     .8705173    2.909542
------------------------------------------------------------------------------
```

$\exp(\hat{\beta}_1)$

Example from Social Ep

Social Epidemiology Eco. Data Set

- Health outcomes are counts of mortality from twelve causes for 3138 United States counties between 1995 and 2004.
- Covariates include rates of crime, numbers of disasters, inflation rate, median income.
- Want to investigate how both 1) mean counts of mortality outcomes are related to "stressor" as well as how 2) variability (or dispersion) is related.

Hypothesis

- Though stressors are well known to be associated with sum changes in mortality of various kinds, more so in poorer than richer neighborhoods.
- We argue that among poor neighborhoods, you get a greater diversity in responses.
- Perhaps, while richer neighborhoods have similar “vulnerabilities” to changing circumstances, there is a greater variability in this response among poor neighborhoods.

Implied Parameter of Interest

- Thus, we investigate the impact of over, over-dispersion in via modeling both the mean and the dispersion

- STATA has options to model both the mean and the dispersion, the two parameters in a negative binomial distribution, called *Generalized Negative Binomial Regression*:

$$Y \sim NB\{\lambda(X), \nu(X)\}$$

Data Assembled Under Supervision of Sandro Galea (Columbia Univ)

- Data assembled and merged from
  - 2000 US Census
  - National Center for Health Statistics,
  - NACCHO
  - CDC
  - US Economic Census
  - National Center for Education Statistics
  - FBI
  - SHELDUS
  - EPA

- In the end, one has county level repeated measures data (repeated over years) on sources of mortality and possible causes.

Median Income versus Rate (mean) and dispersion of Fatal Accidents among counties in 1995

- As as reminder, NB distribution (as function of Median Income (*X*)) is:

$$P(Y = y \mid X) = \left(\frac{\nu(X)}{\nu(X) + \lambda(X)}\right)^{\nu(X)} \frac{\Gamma(\nu(X) + y)}{\Gamma(y+1)\Gamma(\nu(X))} \left(\frac{\lambda(X)}{\nu(X) + \lambda(X)}\right)^{y}$$

where:

$$\log\{\lambda(X)\} \equiv \log\{E(Y \mid X)\} = \beta_0 + \beta_1 X$$

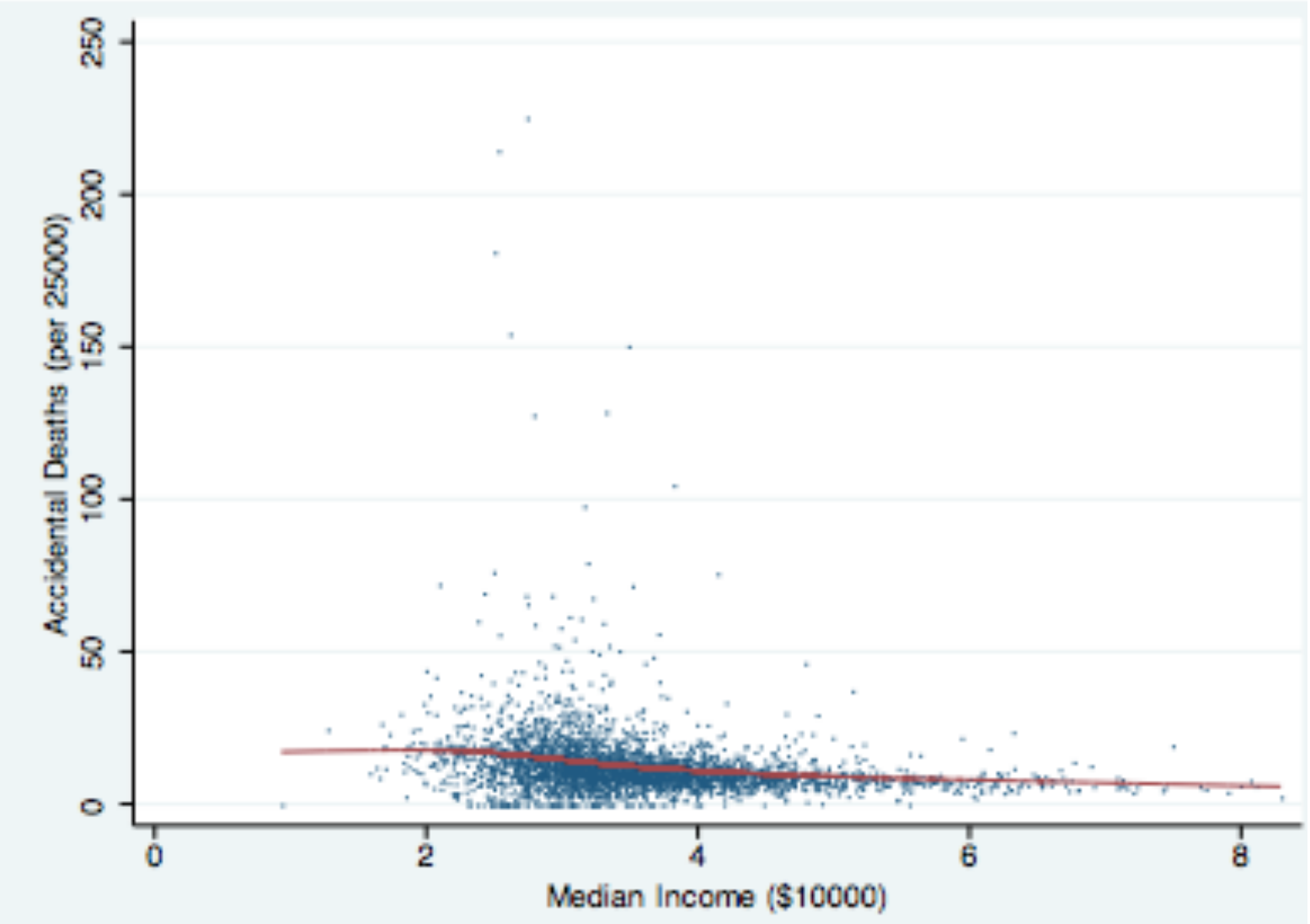
$$\log\left\{\frac{1}{\nu(X)}\right\} = -\log\{\nu(X)\} = \alpha_0 + \alpha_1 X$$

$$Var(Y \mid X) = \exp(\beta_0 + \beta_1 X) + \exp\{2 * (\beta_0 + \beta_1 X) + (\alpha_0 + \alpha_1 X)\}$$

Question of Interest

- How is median county income (*med10000,* in $10,000's) associated with the distribution of fatal accidents (*rnacc*).
- Use only the year 1995 (we will examine this data again the context of repeated years).
- We have the population of the county, and the natural log of this (*logpop00*) is used as the offset in our analyses.

Raw Data (Fatal Accidents versus Median Income) - 1995



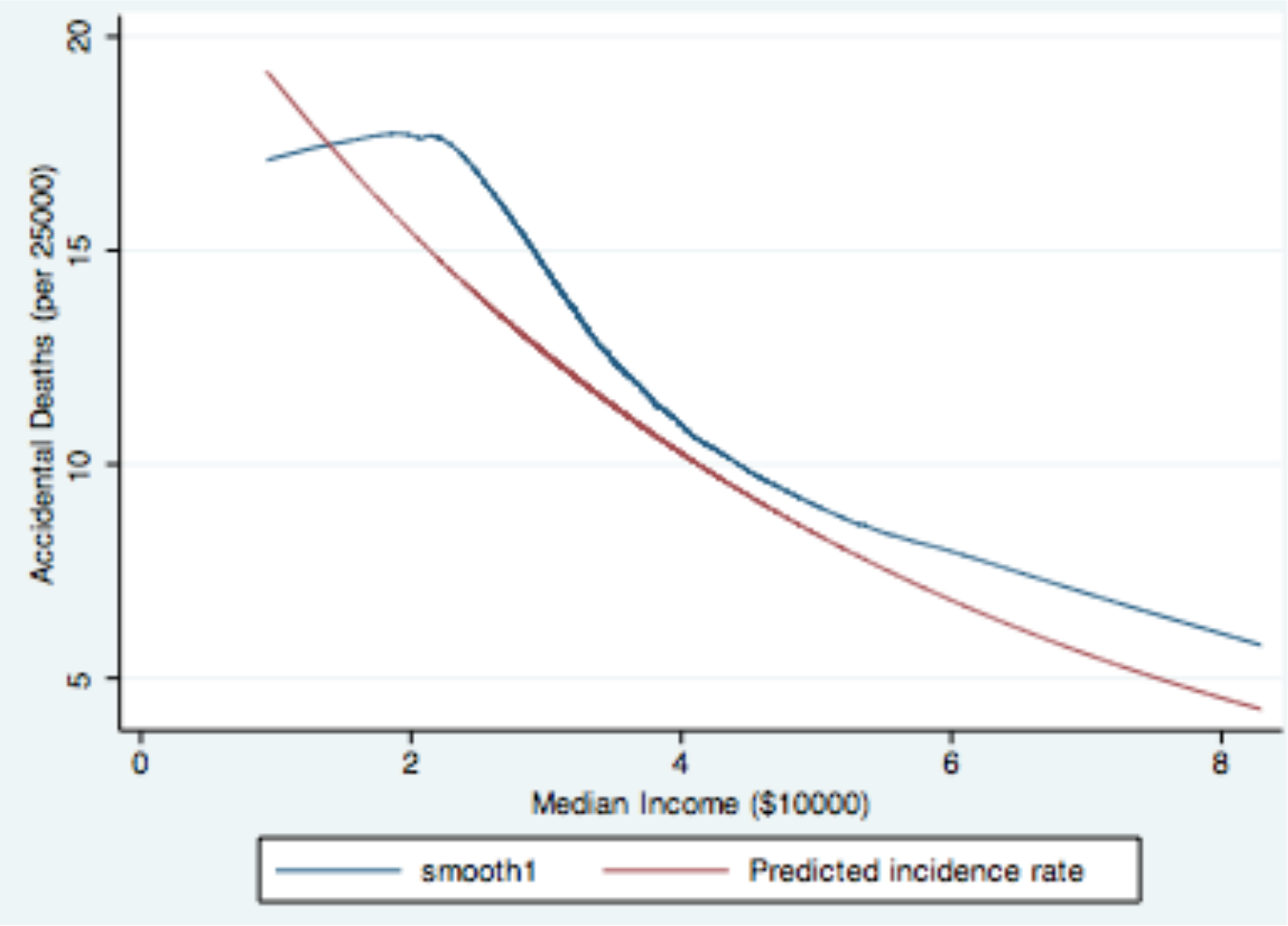
Standard Negative Binomial Regression

```
. nbreg rnacc med10000 if year==1995, offset(logpop00) irr
Negative binomial regression                    Number of obs   =       3138
                                                LR chi2(1)      =     586.14
Dispersion     = mean                           Prob > chi2     =     0.0000
Log likelihood = -10021.477                     Pseudo R2       =     0.0284

------------------------------------------------------------------------------
       rnacc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    med10000 |  -.2040556    .007626   -26.76   0.000    -.2190023   -.1891089
       _cons |  -6.982277   .0293075  -238.24   0.000    -7.039719   -6.924836
    logpop00 |   (offset)
-------------+----------------------------------------------------------------
    /lnalpha |  -2.438768   .0478215                      -2.532497    -2.34504
-------------+----------------------------------------------------------------
       alpha |   .0872683   .0041733                       .0794604    .0958434
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:  chibar2(01) = 4109.54 Prob>=chibar2 = 0.000
------------------------------------------------------------------------------
       rnacc |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |    .815417   .0062184   -26.76   0.000     .8033198    .8276964
```

$$\log\left\{\hat{\lambda}(X,pop)\right\} = \log(pop) - 6.98 - 0.20X,\ \log\left\{\frac{1}{\hat{\nu}}\right\} = -2.44, \hat{\nu} = 11.5$$

Smooth (Lowess) versus NB fit to predicted mean (E(Y|X=x))



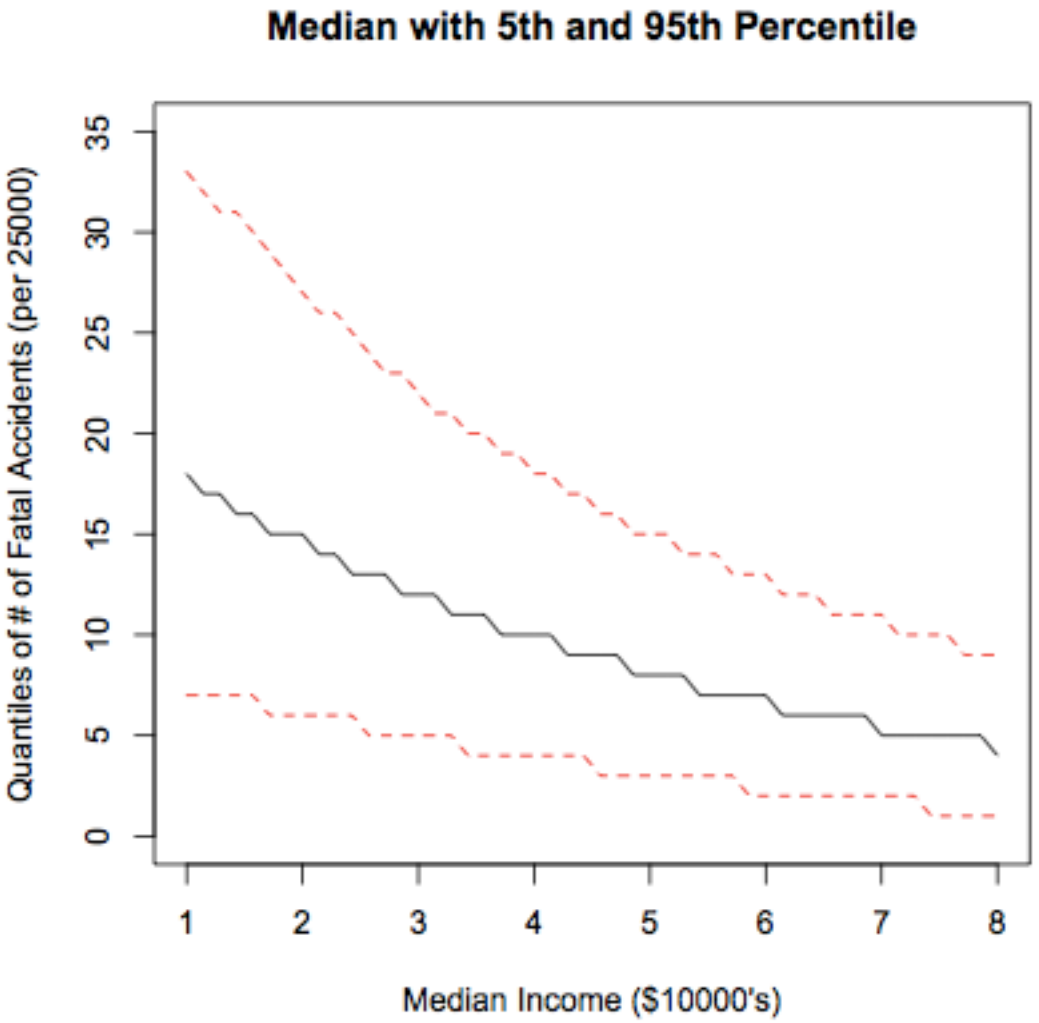
Generalized Negative Binomial Regr.

```
 gnbreg rnacc med10000 if year==1995, offset(logpop00) irr
Generalized negative binomial regression          Number of obs   =       3138
                                                  LR chi2(1)      =     576.38
                                                  Prob > chi2     =     0.0000
Log likelihood = -10016.806                       Pseudo R2       =     0.0280

------------------------------------------------------------------------------
       rnacc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
rnacc        |
    med10000 |   -.199518   .0072147   -27.65   0.000    -.2136587   -.1853774
       _cons |  -6.998786   .0285102  -245.48   0.000    -7.054665   -6.942907
    logpop00 |   (offset)
-------------+----------------------------------------------------------------
lnalpha      |
    med10000 |  -.1521521   .0493641    -3.08   0.002    -.2489039   -.0554003
       _cons |  -1.862365   .1919467    -9.70   0.000    -2.238574   -1.486156
------------------------------------------------------------------------------
```

$$\log\left\{\hat{\lambda}(X,pop)\right\} = \log(pop) - 7.0 - 0.20X,\ \log\left\{\frac{1}{\hat{\nu}(X)}\right\} = -1.86 - 0.15X$$

Results of Generalized Negative Binomial Model Fit



Note that not only is the mean generally higher at lower income places, but the probability for very large numbers of accidents is much greater.

Clever Use of NB in Infectious Disease

Example of when the difference of Poisson and NB Matters: Superspreading and the effect of individual variation on disease emergence

- Paper in Nature (vol 438: 355-9, 2005) Lloyd-Smith, et al., consider the distribution of secondary cases from infected individuals for several diseases (plague, measles, SARS, etc.).

- Point is to determine whether the data support:
  - homogenous transmission of disease in a population or,
  - clustering of transmission in a relatively small group of individuals (superspreaders).

Infectious Disease and NB

- The difference is important because:
  - it provides insight as to how diseases either spread or become extinct in populations and,
  - difference is important in how preventive measures (for stopping the spread the disease) should be employed, e.g.,
    - Vaccinate everyone with a imperfect vaccination, or
    - spend more resources on preventing transmission to potential superspreaders.
- How they determined what model fits better is to fit both a Poisson and NB model to data and look at the relative goodness of fit using AIC (Akiake Information Criterion).

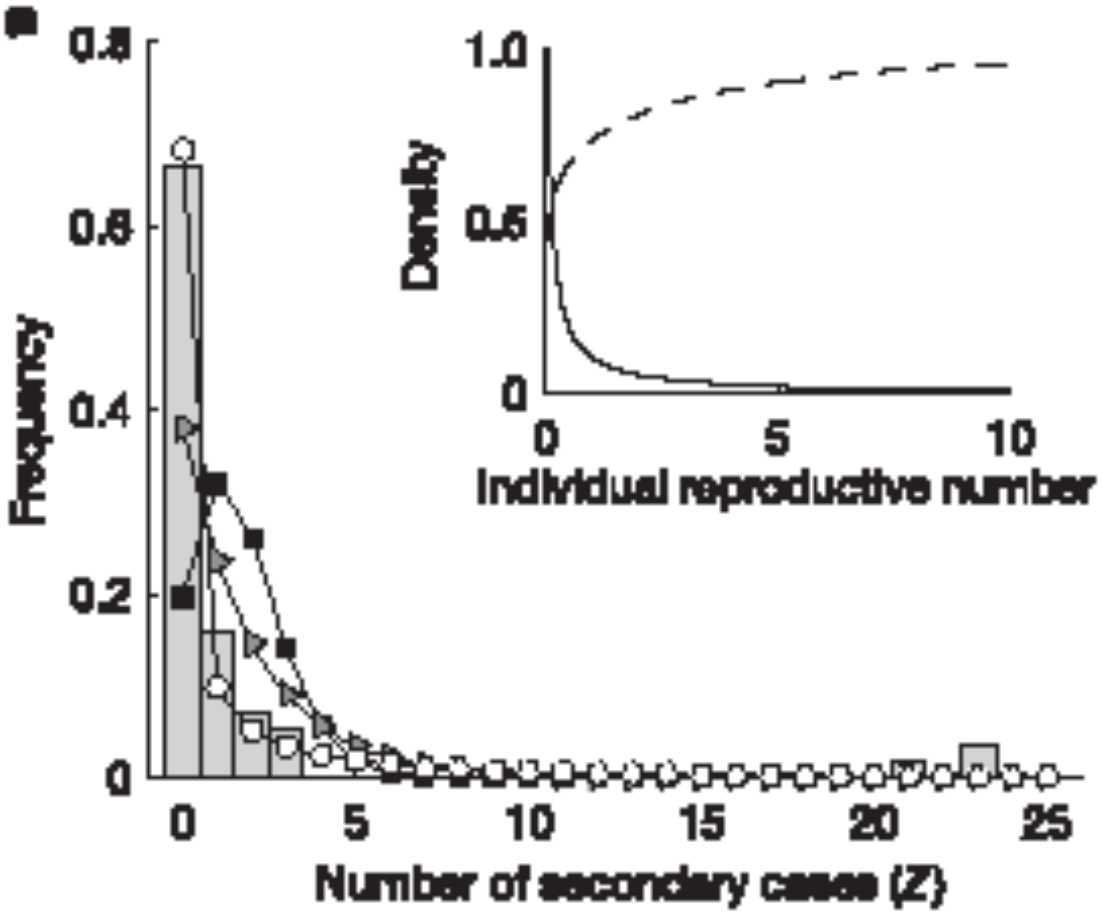
Using the Parameters of the NB to estimate the distribution of reproductive number.

- Again, one can interpret the negative binomial model as a mixture of Poissons, where the distribution of underlying rates of the outcome is gamma.

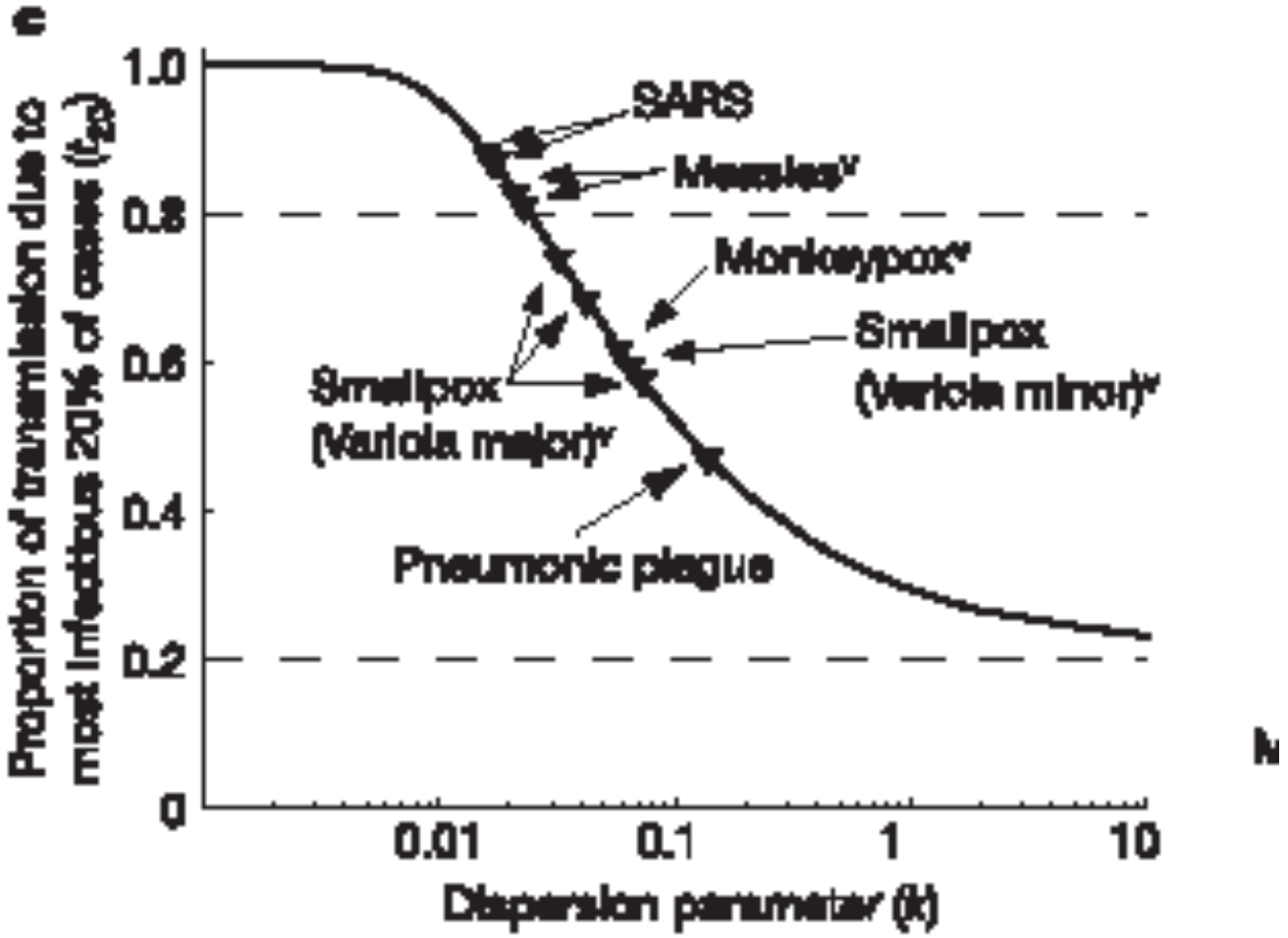
- Thus, estimating the mean and dispersion of the NB model (if the model is true) provides an estimate of the distribution of transmission rates in the population.

- And, thus the proportion of superspreaders.

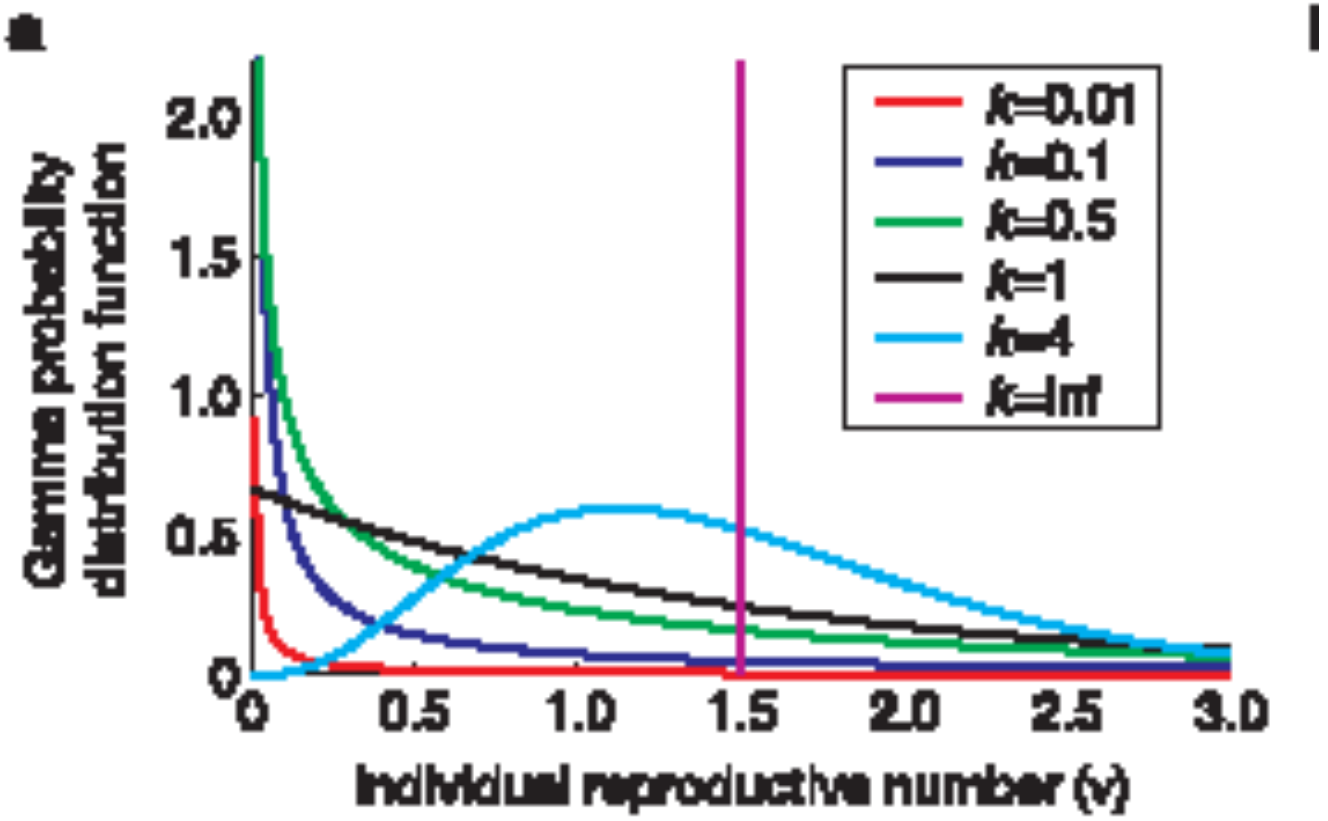
Transmission data from SARS outbreak in Singapore, 2003.



Dispersion parameter vs. Proportion of Transmission from 20% most infectious individuals (based on NB fit)



Example of different distributions of rates of transmission in a population for different dispersion parameters (k) all with same mean rate ($\lambda = 1.5$).



Probability of Extinction (no transmission) given mean and dispersion