

Week 10

Event Counts II

Rich Frank

University of New Orleans

November 1, 2012

- Let me start with a few things before we plunge into count models.

Wald statistic

- Are the estimated parameters far away from what they would be under the null hypothesis?

- The Wald statistic is calculated as follows:

$$W = [\mathbf{Q}\hat{\boldsymbol{\beta}} - \mathbf{r}]'[\mathbf{Q}\widehat{Var}(\hat{\boldsymbol{\beta}})\mathbf{Q}']^{-1}[\mathbf{Q}\hat{\boldsymbol{\beta}} - \mathbf{r}]$$

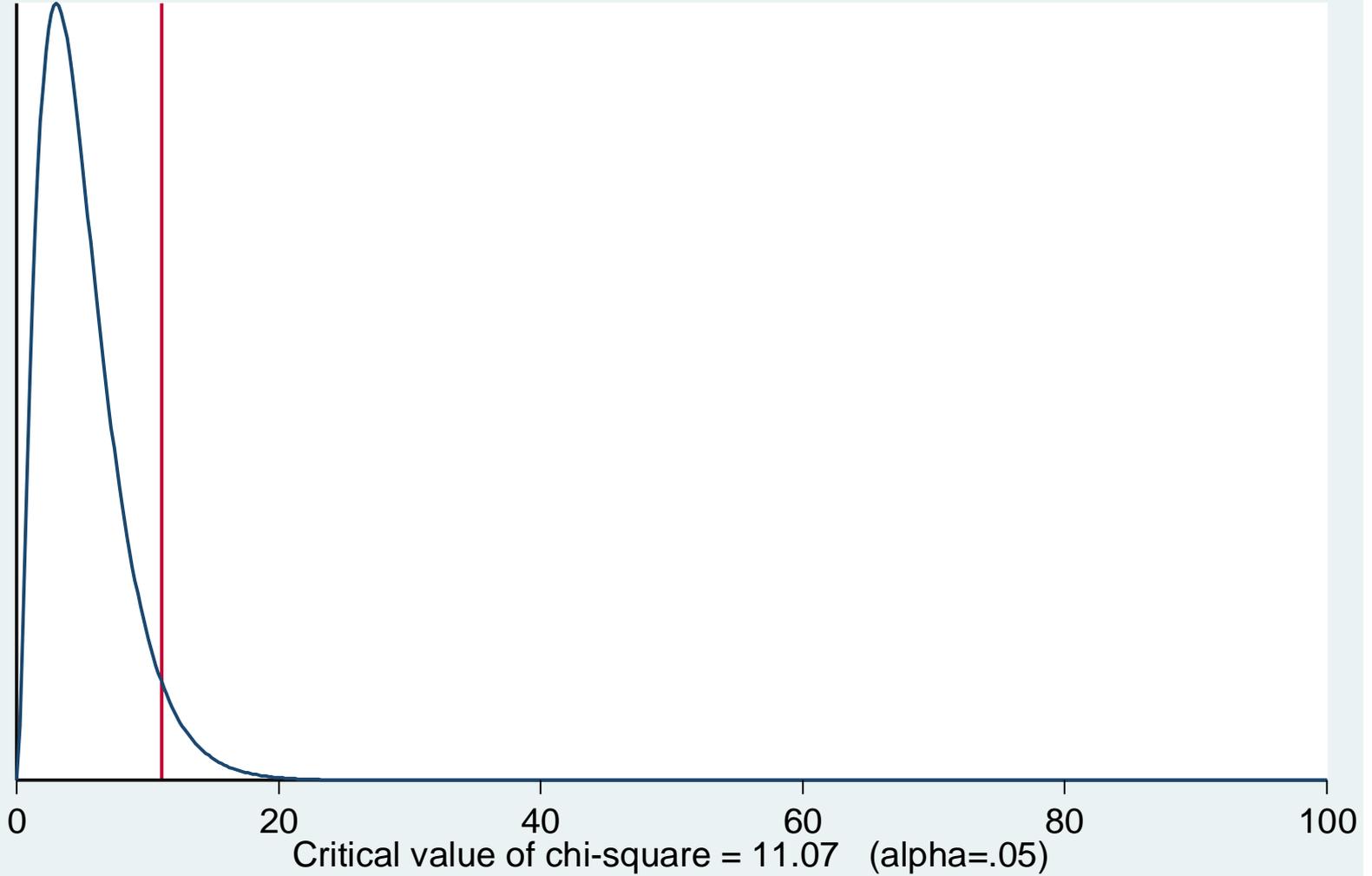
- W is *distributed* chi-square.

- In Stata: `test`

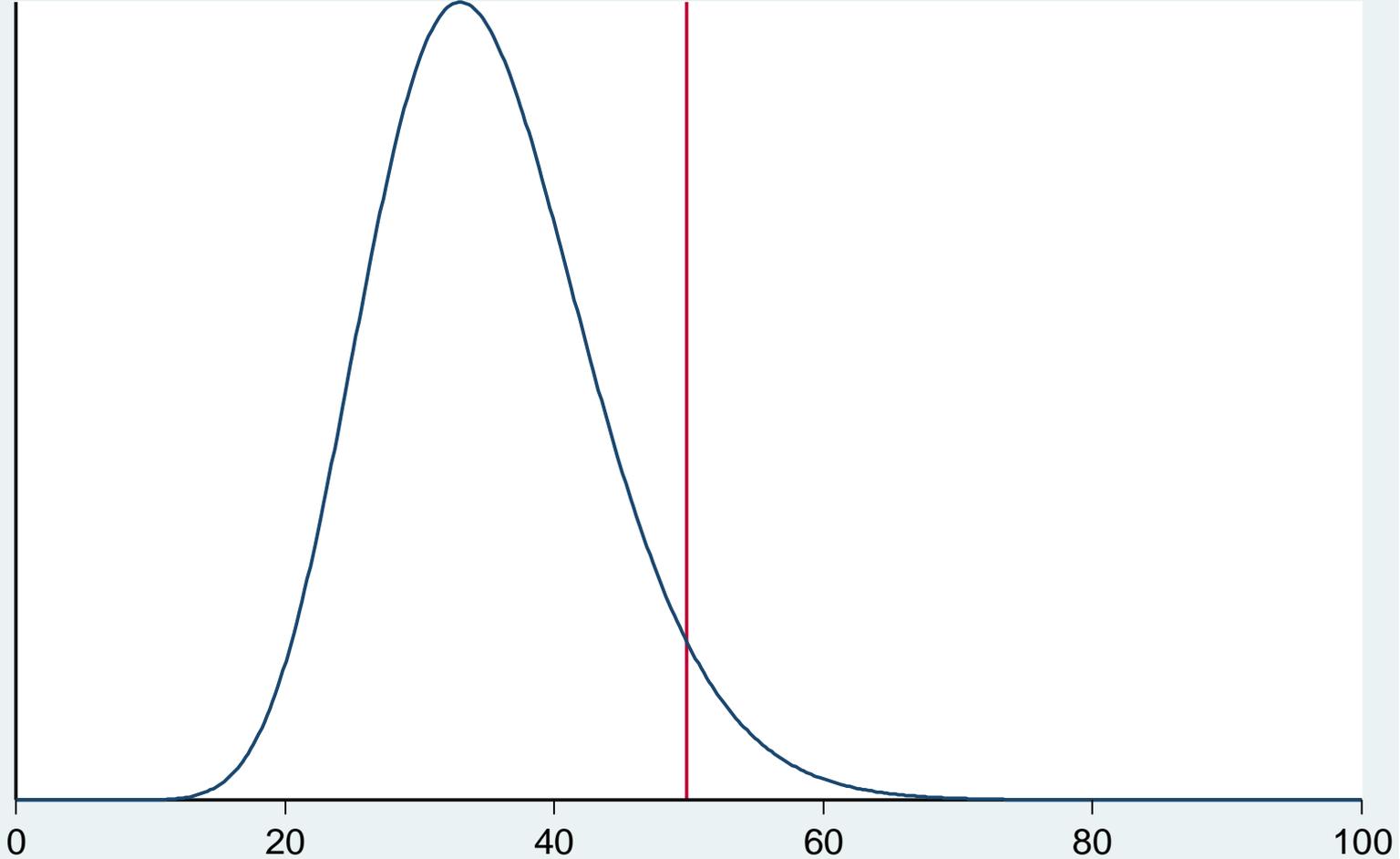
When do you reject the null?

- If the $p < .05$ then you reject the null that the independent variable has **no** effect on the dependent variable (at the 95% confidence level).
- The χ^2 statistic is but a distribution.
- Like the standard normal, we reject the null if the estimated χ^2 is greater than a certain level.

Chi-square Distribution (df=5)



Chi-square Distribution (df=35)



Critical value of chi-square = 49.8 (alpha=.05)

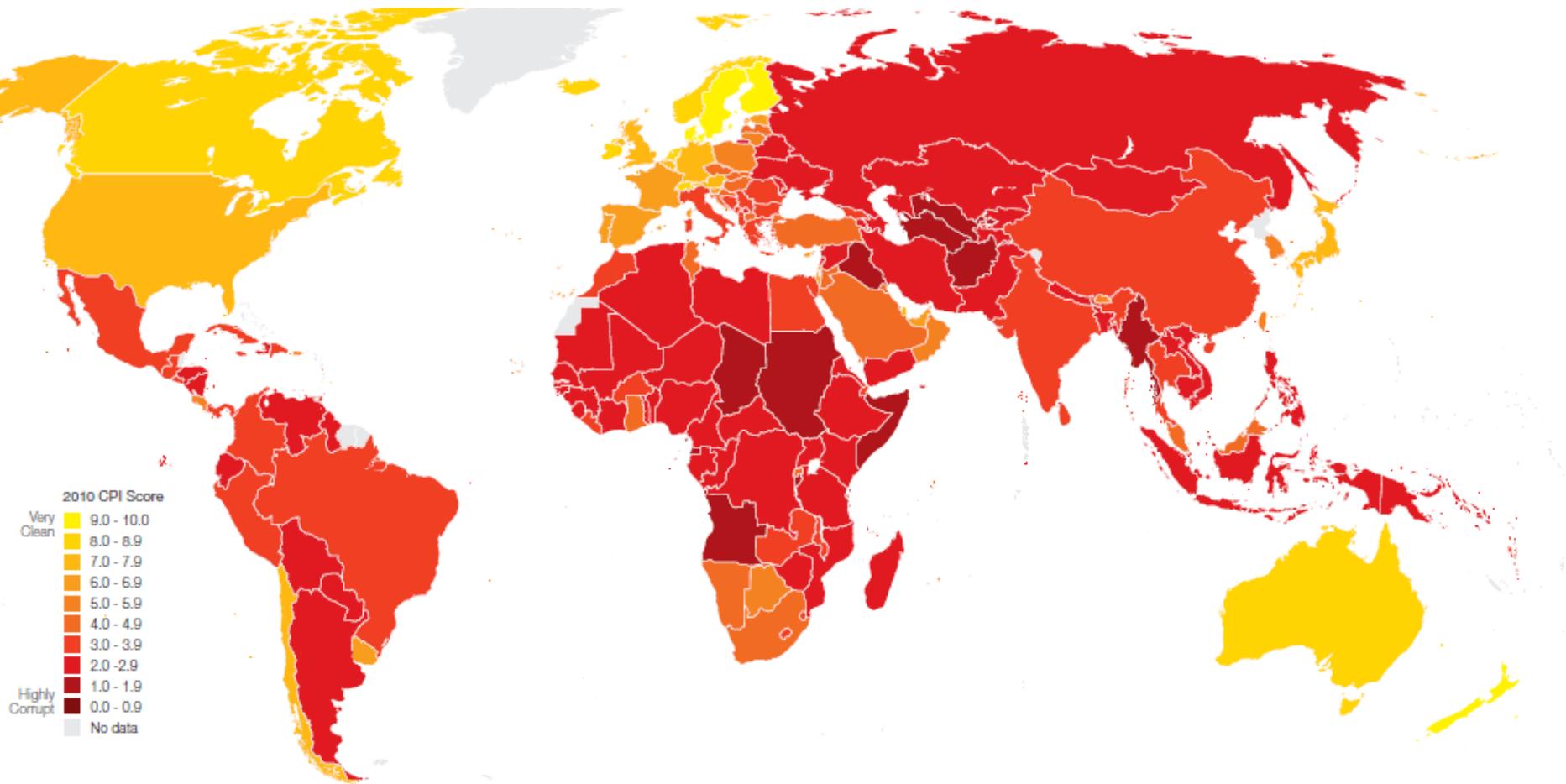
What are the effects of pooling your data?

- Pooling across different units?
- Pooling across different time episodes?
- Shellman (2004) talks about how your results can depend on the time interval you chose.
- See Greene (2012: Chs. 20-21) for an in-depth analysis of time-series approaches.

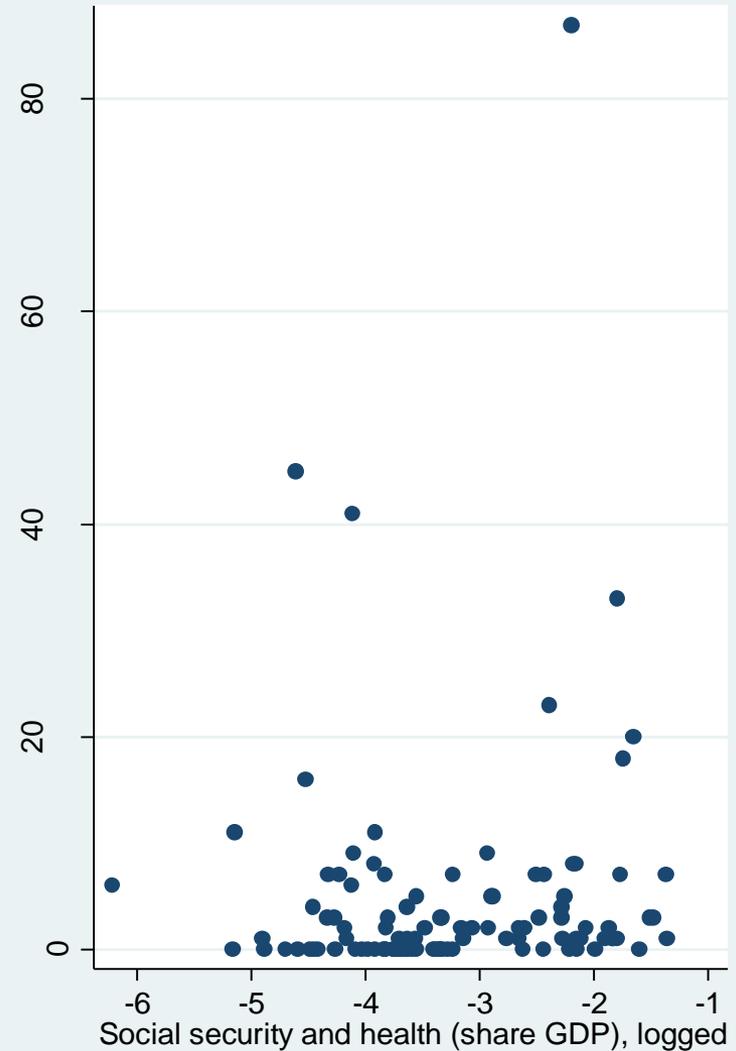
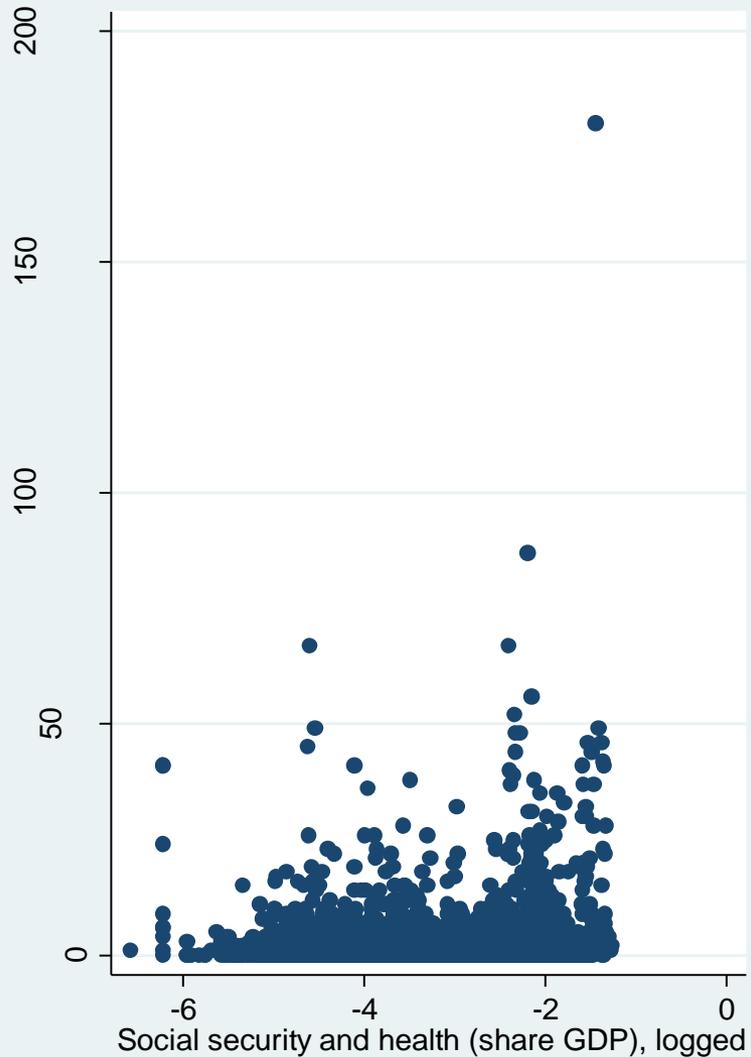
Pooling across space

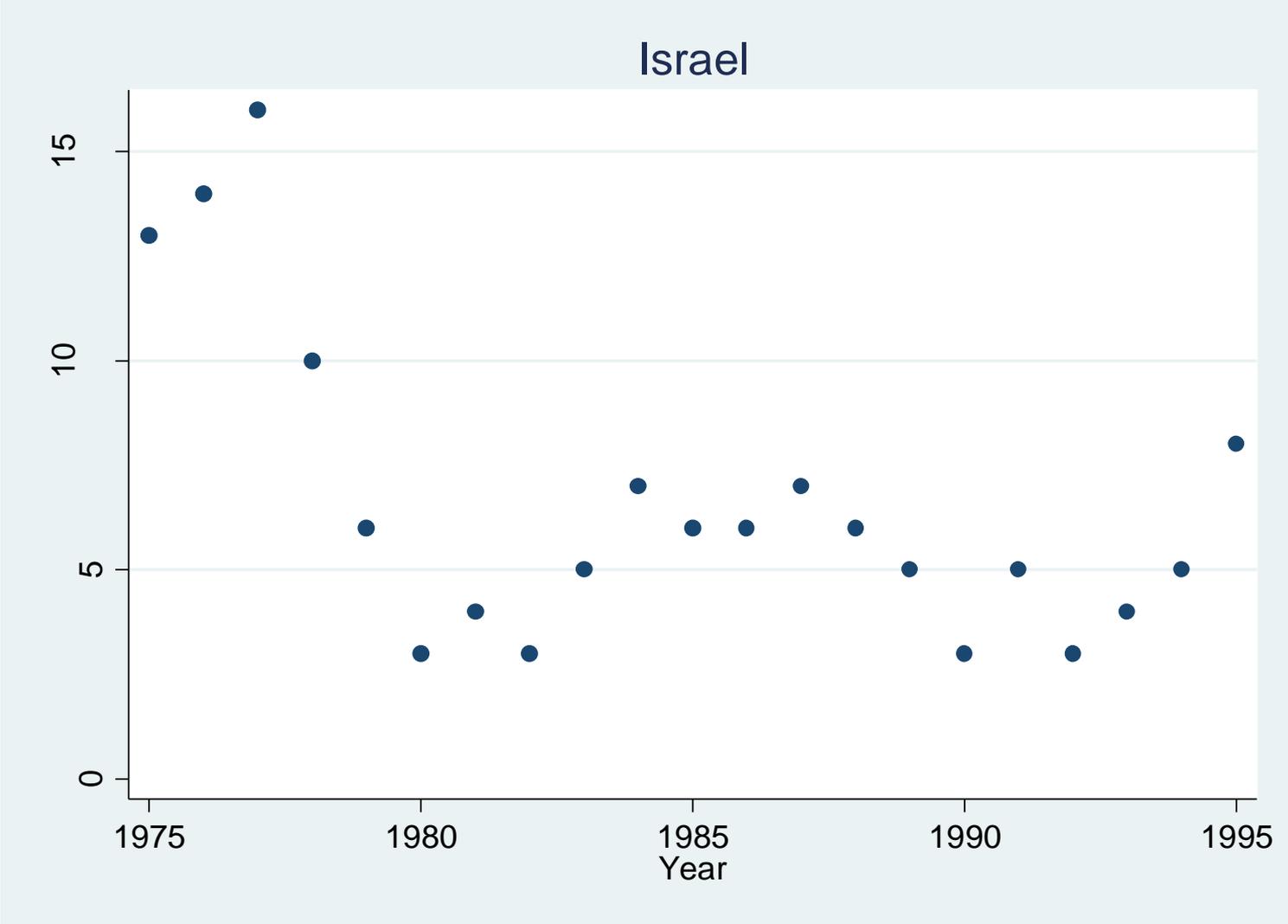
- Heger et al. (2009) talk about heterogeneity across space.
- Are observations independent of each other?
- If not, they violate the i.i.d. assumption.

Example: Transparency International's Corruption Index

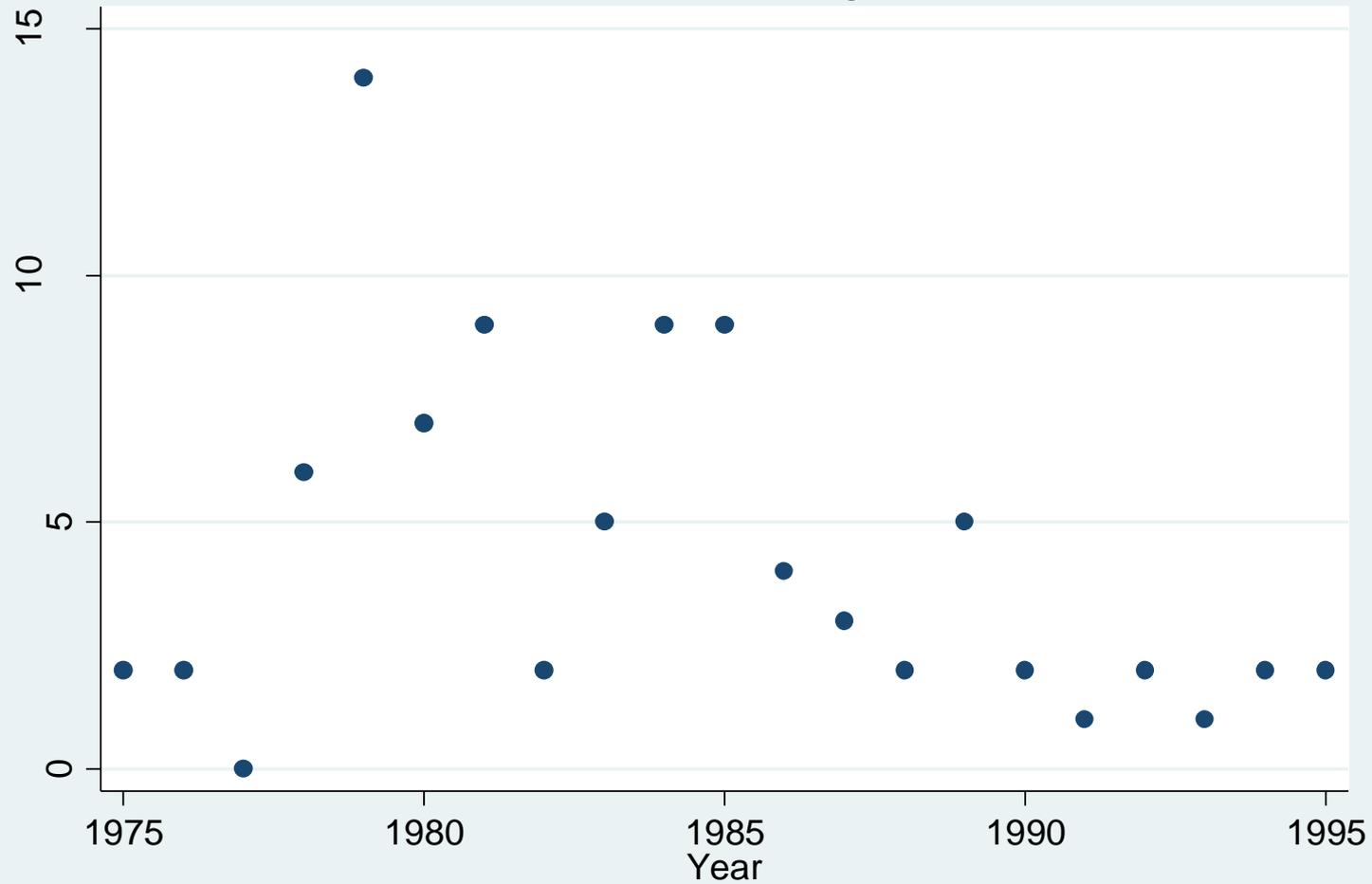


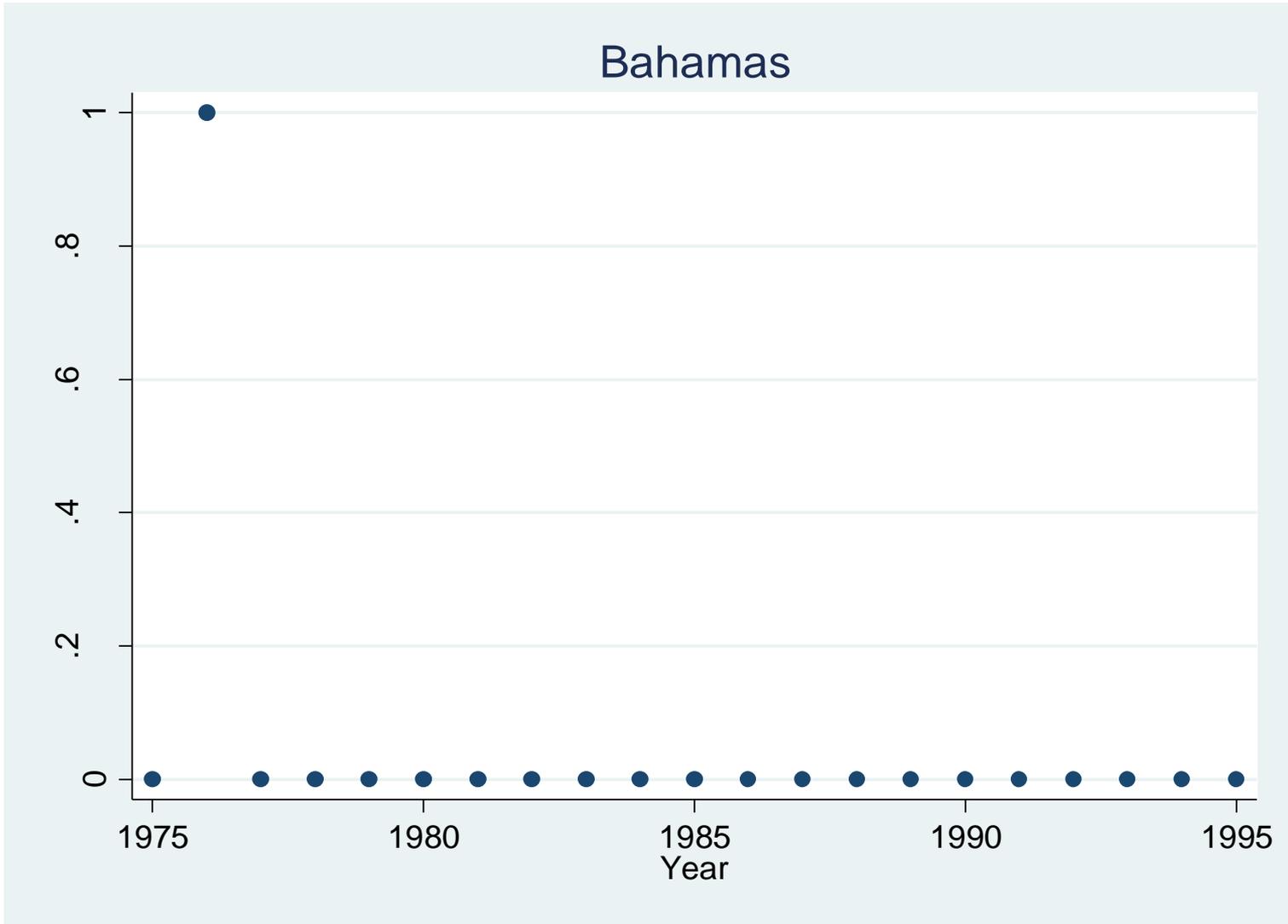
Burgoon (2006) pooled and cross-sectional terrorism data



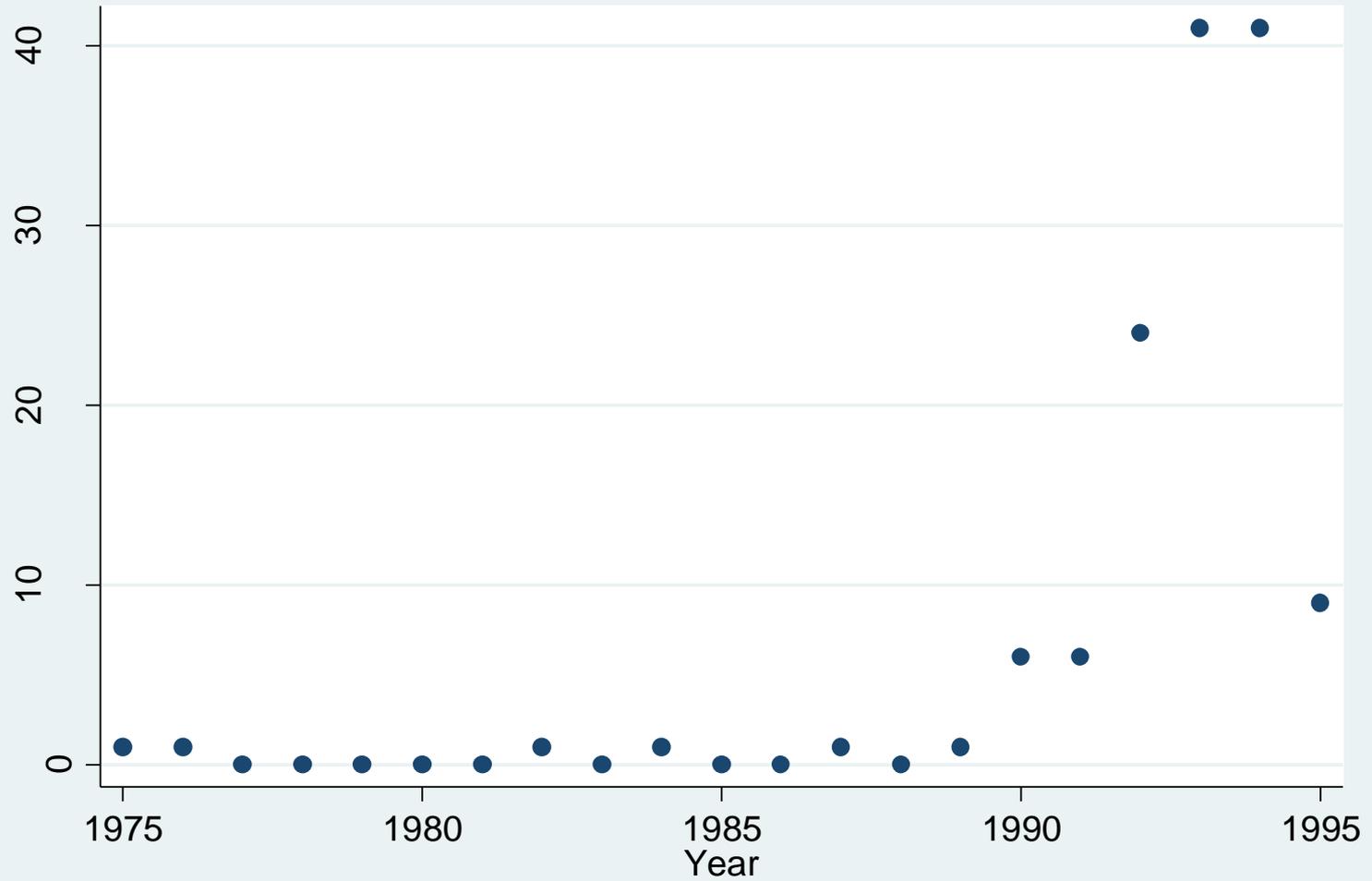


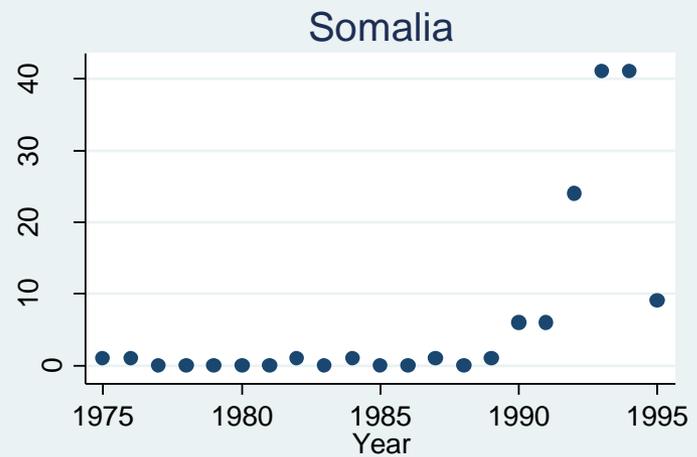
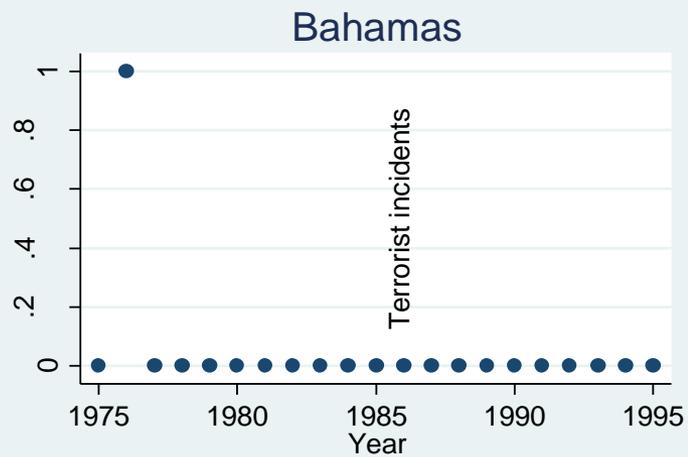
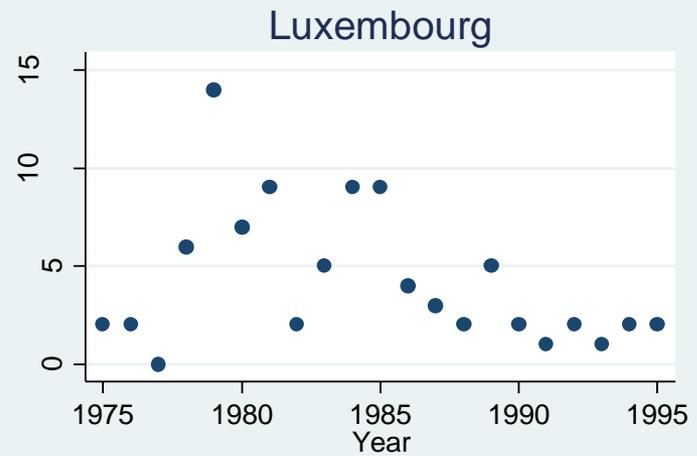
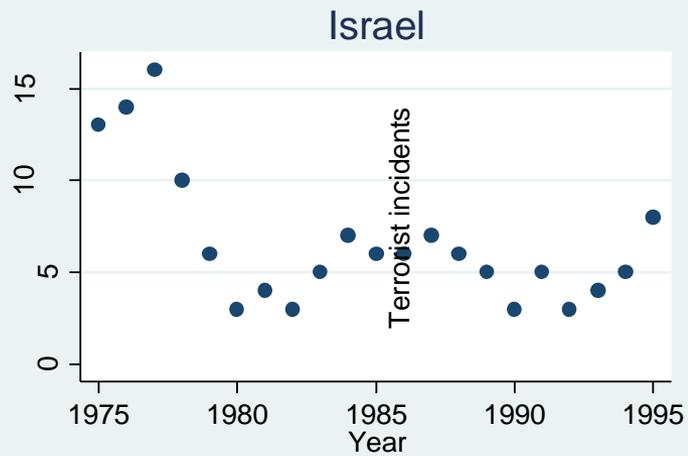
Luxembourg





Somalia





- What do we do with cross-sectional time series (CSTS) data?
- Think about the error term, ε .
- Are the errors likely to be heteroskedastic?
- How do we control for heteroskedasticity?
- All comes down to what you know (or think you know) *theoretically* about the data generating process.

Interpretation

- So you have thought about the effects of time and space, and you finally have results.
- What to do now?
- Interpretation
 - Where data meet theory
 - It's crucial! Both to understanding the substantive importance of your results, but also in conveying them to others.

What is the Variance Inflation Factor?

- A measure of how severe the multicollinearity is among your independence variables.
- Specifically, it looks at how much larger the standard error would be if an IV was not correlated with other variables
 - See Gujarati 2003: Ch 10 for a more in-depth discussion of multicollinearity.

$$\text{VIF} = \frac{1}{(1 - r^2_{23})}$$

where r_{23} is the correlation coefficient between two X variables (X_2, X_3) (Gujarati 2003: 351).

- As you can see as the correlation increases towards 1, the VIF rises.

- Not seen very often in ML (in my experience).
- Why are we not able to run it for ML models?
 - ...Because only works after `regress`.
- So how does Burgoon (2006) do it?

Point predictions

- A bit more challenging to do by hand with the NB than with the Poisson ($e^{\beta X}$).

$$\begin{aligned}\hat{P}(y | x) &= \\ &= \frac{\Gamma(y + \hat{\alpha}^{-1})}{y! \Gamma(\hat{\alpha}^{-1})} \left(\frac{\hat{\alpha}^{-1}}{\hat{\alpha}^{-1} + \hat{\mu}} \right)^{\hat{\alpha}^{-1}} \left(\frac{\hat{\mu}}{\hat{\alpha}^{-1} + \hat{\mu}} \right)^y\end{aligned}$$

Where $\hat{\mu} = e^{XB}$

Probabilities of terrorist attacks

```
. prvalue
```

```
nbreg: Predictions for terrorinlead
```

```
Confidence intervals by delta method
```

```

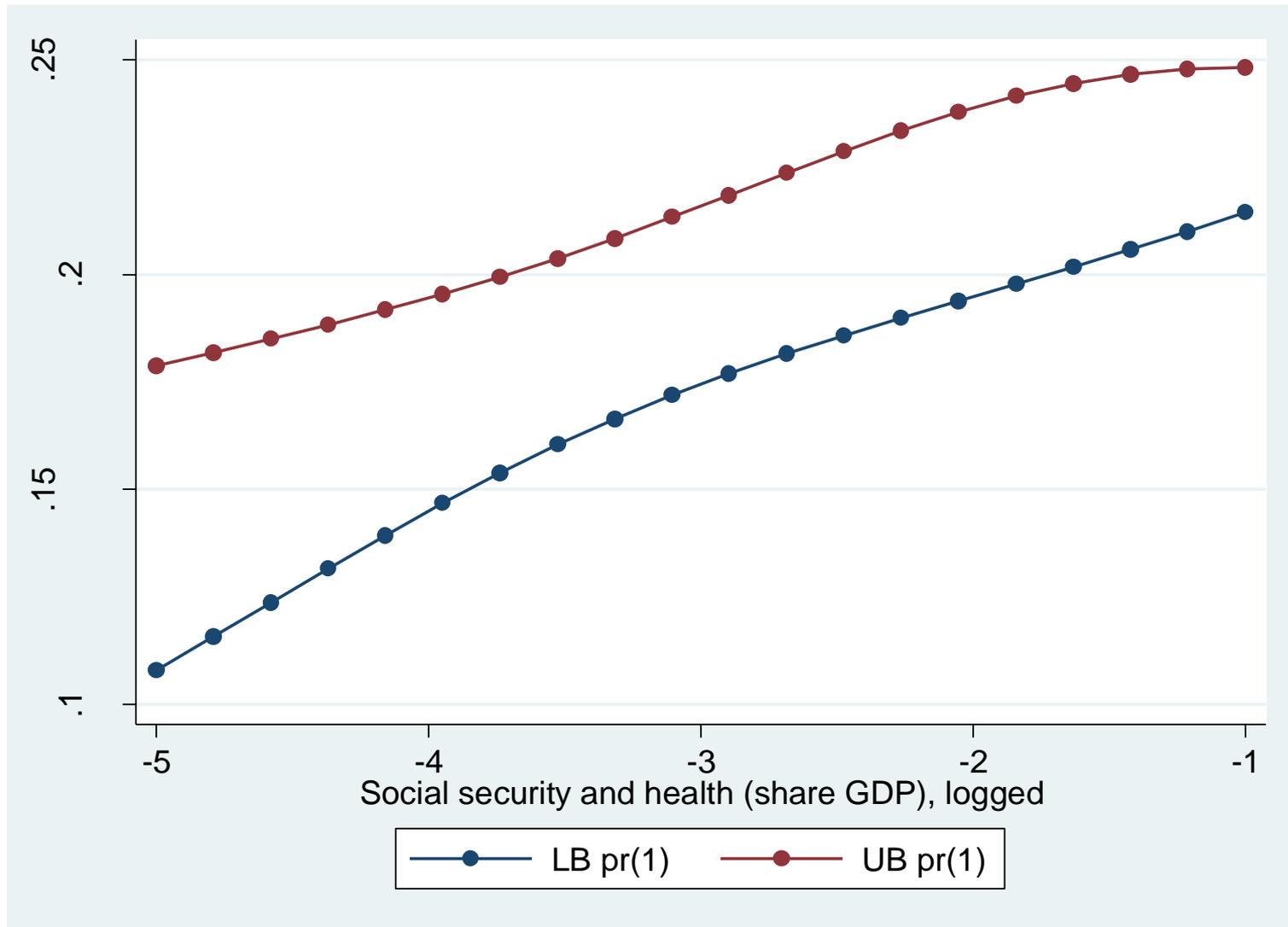
                                95% Conf. Interval
Rate:                          1.7938 [ 1.5203, 2.0673]
Pr (y=0|x):                     0.3832 [ 0.3498, 0.4166]
Pr (y=1|x):                     0.2189 [ 0.2105, 0.2274]
Pr (y=2|x):                     0.1372 [ 0.1358, 0.1385]
Pr (y=3|x):                     0.0884 [ 0.0833, 0.0936]
Pr (y=4|x):                     0.0578 [ 0.0516, 0.0640]
Pr (y=5|x):                     0.0381 [ 0.0322, 0.0441]
Pr (y=6|x):                     0.0253 [ 0.0201, 0.0305]
Pr (y=7|x):                     0.0168 [ 0.0126, 0.0211]
Pr (y=8|x):                     0.0112 [ 0.0079, 0.0146]
Pr (y=9|x):                     0.0075 [ 0.0049, 0.0102]

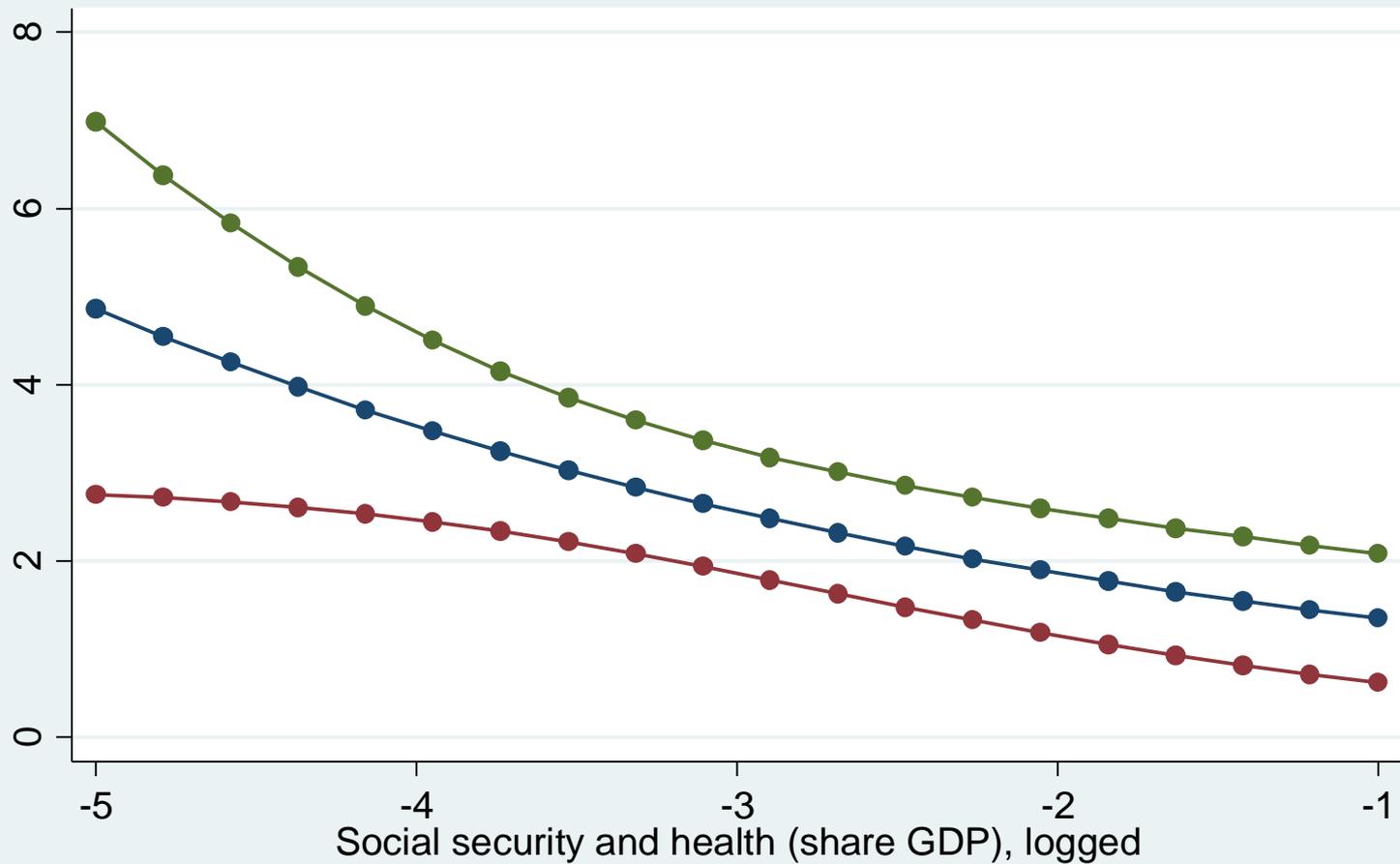
      transferslog      govleft      democ      poplog      govcap
x=      -3.2036522      .29904441      2.0553682      16.170913      .84046542

      conflict      tradelog      terrorinc      europe      africa
x=      .03878583      4.0498434      3.6003373      .22709387      .24620573

      asia      america
x=      .1742552      .23496346
```

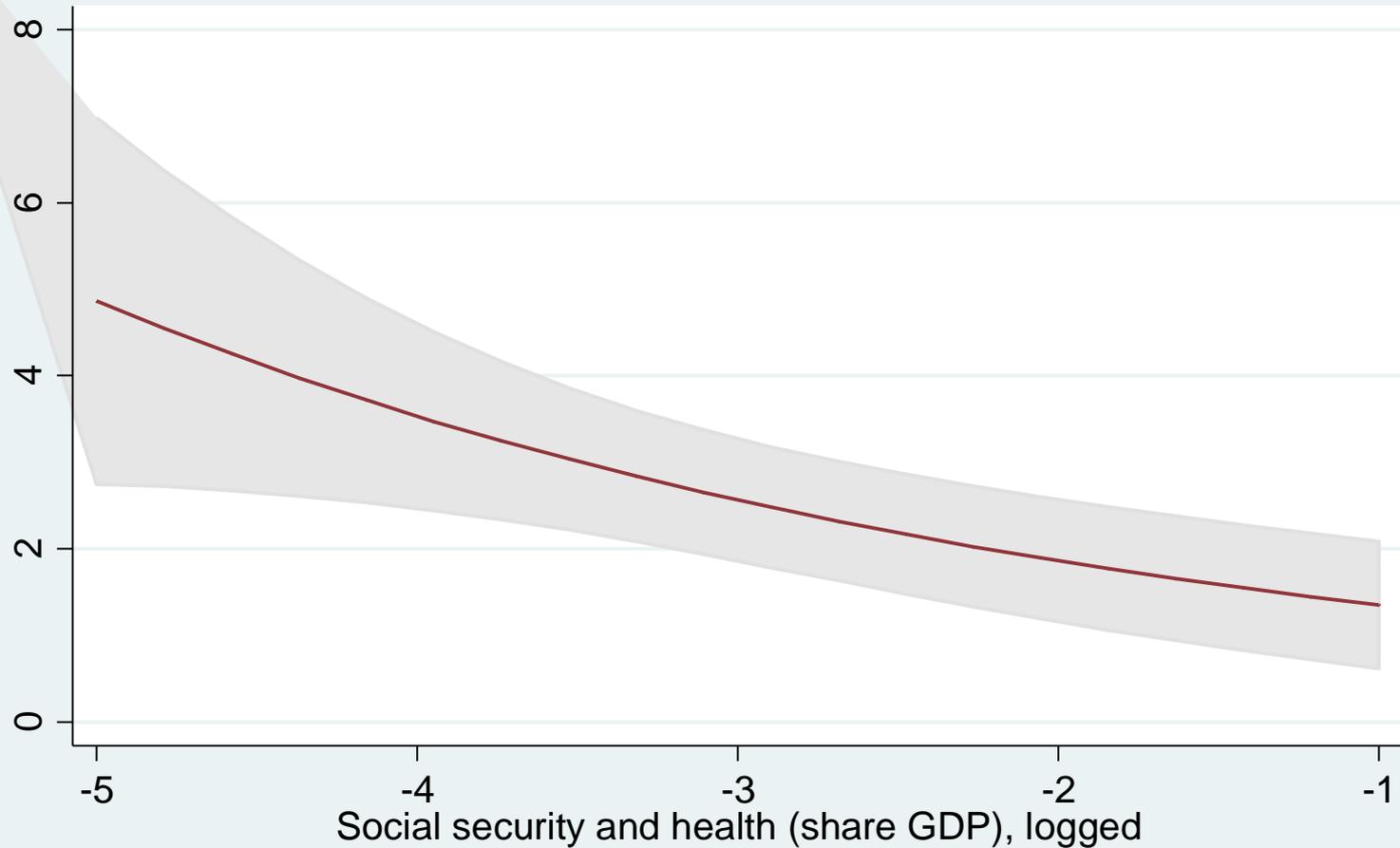
Continuous predicted probabilities





I think it looks prettier this way.

Predicted number of terrorist attacks



Lower bound/Upper bound predicted rate μ

Stata code

```
nbreg terrorinclead transferslog govleft democ poplog govcap ///
conflict tradelog terrorinc europe africa asia america , ///
dispersion(mean) robust cluster(cow)

prgen transferslog, x(govleft=0 democ=0 conflict=0 europe=0 africa=0 ///asia=0 america=0 )
rest(mean) from(-5) to (-1) gen(trns) ci n(20)

** Probability of having no attacks **
graph twoway connected trnsp1lb trnsp1ub trnsx, ///
    ytitle("Probability of a Zero Count") ///

** Predicted number of attacks
graph twoway connected trnsmu trnsmulb trnsmuub trnsx, ///
    ytitle("Predicted Count")

graph twoway (rarea trnsmulb trnsmuub trnsx, color(gs14)) ///
    (connected trnsmu trnsx, lpattern(solid) msize(zero)) ///
    , ytitle("Predicted Count") title("Predicted number of terrorist attacks")
```

Let's set some variables to specific countries!

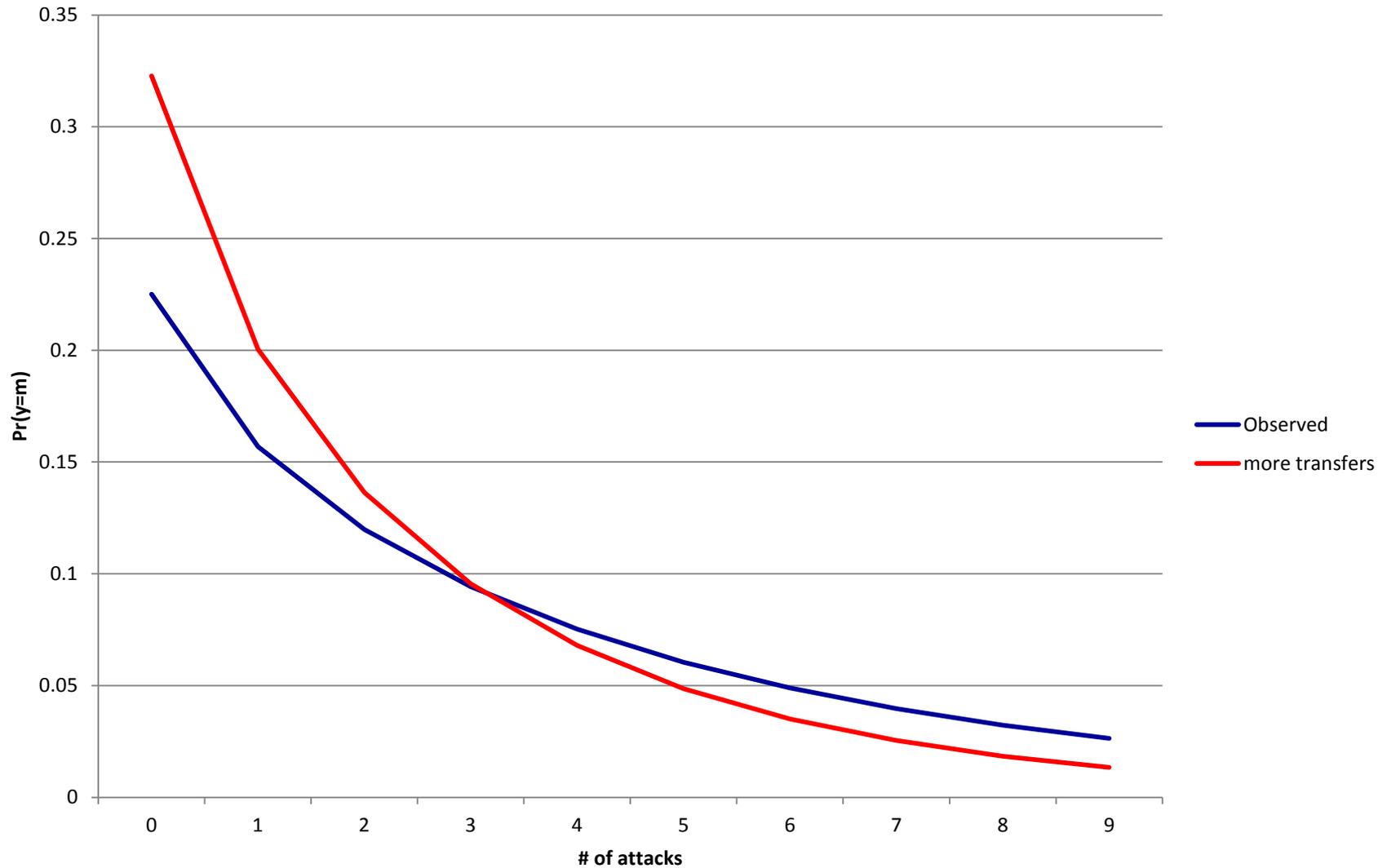
- Israel 1985?
- Bahamas 1990?

Let's set some variables to specific countries!

- Israel 1985? Estimate $P(y=1)$ is .157.
- Let's say we increase transfers by 1 unit: $P(y=1)$ is .200.
- The probability of a terrorist attack increases by 27%

Pr ($y=m$)	Observed welfare transfers	adding 1 unit	% Change
<i>0</i>	0.2251	0.3227	0.0976
<i>1</i>	0.1569	0.2004	0.0435
<i>2</i>	0.1198	0.1364	0.0166
<i>3</i>	0.0942	0.0955	0.0013
<i>4</i>	0.0752	0.0679	-0.007
<i>5</i>	0.0604	0.0486	-0.012
<i>6</i>	0.0489	0.0350	-0.014
<i>7</i>	0.0397	0.0254	-0.014
<i>8</i>	0.0323	0.0184	-0.014
<i>9</i>	0.0264	0.0134	-0.013

Israel 1985



- Okay, now let's turn back to the underlying models—the Poisson and negative binomial models.
- A refresher before we look at zero-altered models.

Poisson

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

for $y = 0, 1, 2, \dots$

Negative Binomial

- The NB model adds an error, ε , that is assumed to be uncorrelated with the X 's.
- We then estimate a random variable $\tilde{\mu}$:

$$\tilde{\mu}_i = e^{(\beta x_i + \varepsilon_i)} = \mu_i e^{\varepsilon_i} = \mu_i \delta_i$$

Where $\delta_i = e^{\varepsilon_i}$

- In order to be able to identify our model (like we do with the logit/probit, and multinomial) we assume a value for the mean of the error term.

- The most convenient assumption is that $E(\delta_i) = 1$.
- This lets us have the same expected count as the Poisson despite allowing a new source of variation.

$$E(\tilde{\mu}_i) = E(\mu_i \delta_i) = \mu_i E(\delta_i) = \mu_i$$

- In an effort to ease interpretation and understanding this is the last time I am using the i subscript. All variables that vary by individual unit can be assumed to have an i subscript.

- The negative binomial distribution is given by the formula:

$$P(y | x) = \frac{\Gamma(y+v)}{y!\Gamma(v)} \left(\frac{v}{v+\mu}\right)^v \left(\frac{\mu}{v+\mu}\right)^v$$

- The expected value of Y in the NB distribution is the same as the Poisson:

$$E(y|x) = \mu$$

- But the variance is different and given by:

$$\text{Var}(y|x) = \mu \left(1 + \frac{\mu}{v}\right) = e^{\beta x} \left(1 + \frac{e^{\beta x}}{v}\right)$$

- We have seen that the Poisson fails if the estimated mean and variance are not equivalent.
- The NB relaxes the equidispersion assumption.
- However, what if our data have a large number of cases in which there are no instances of the event?

What if there are “extra” zeros in your data

- This could be a problem for several reasons:
- Could inflate variance.
- Could indicate a different data-generating process is at work.

What if there are “extra” zeros in your data

```
. tab terrorism incident if e(sample)
Transnation al terrorism incident | Freq. Percent Cum.
-----+-----
      0 |      765    43.00    43.00
      1 |      279    15.68    58.68
      2 |      170     9.56    68.24
      3 |      118     6.63    74.87
      4 |       82     4.61    79.48
      5 |       64     3.60    83.08
      6 |       41     2.30    85.39
      7 |       37     2.08    87.46
      8 |       33     1.85    89.32
      9 |       27     1.52    90.84
     10 |       16     0.90    91.74
     11 |       11     0.62    92.36
     12 |       10     0.56    92.92
     13 |        7     0.39    93.31
     14 |       14     0.79    94.10
     15 |       10     0.56    94.66
     16 |        7     0.39    95.05
     17 |        4     0.22    95.28
-----+-----
     87 |        1     0.06    99.94
    180 |        1     0.06   100.00
-----+-----
Total |    1,779   100.00
```

Example

- Suppose we ask people how often they go fishing.
- A certain number go fish by the lake or the bayou several times a week.
- However, there are a number of people who do not fish at all.
- Who are these non-fishers?

- If you could ask them why they did not fish last month, they could say several things:
- My family was in town, so I did not have enough time.
- Fish? I am a vegetarian and hate killing harmless animals!

Hurdles

- What gets you over that hurdle from being a non-fisher to a fisher?
- Or a state without terrorist attacks to one that does?
- Does it matter where you try and find fishers?



Truncated counts

- What if you only ask people how much they fish at the new fishing dock in City Park?
- Or survey of people about public buses given on a bus?
- You could be truncating your data.
- In the Poisson the probability of 0 and positive counts are:

$$P(y = 0 \mid x) = e^{-\mu}$$

$$P(y > 0 \mid x) = 1 - e^{-\mu}$$

Truncated Poisson and NB

- Truncated Poisson Likelihood (used in Stata with `ztp`):

$$L(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) = \prod_{i=1}^n \frac{e^{-\mu} \mu^y}{y! (1 - e^{-\mu})}$$

- Truncated Negative Binomial (used in Stata with `ztnb`):

$$L(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) = \prod_{i=1}^n \frac{\frac{\Gamma(y + \alpha^{-1})}{y! \Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu} \right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1} + \mu} \right)^y}{1 - (1 + \alpha\mu)(1 + \alpha\mu)^{-\alpha^{-1}}}$$

- Truncated counts are motivated by very specific conditions.
- When the data-gathering process involves observations are in the data only after the first count occurs.
- This means that there are necessarily no 0's in the data.

- However, more frequently in the data that you are likely to use (like Burgoon 2006), there are *too many* zeros not too few.
- This is when we look at *zero-inflated* models.
- Getting back to the fishing example...

- We can assume that the population is made up of two groups:
- People can be in group 1 (fished in the last month) with probability Ψ and in group 2 (not fished in past month) with probability $1 - \Psi$.
- We do not know Ψ , but we are interested in estimating it.

- We cannot tell whether people did not fish because:
 - 1) they work at the Gap and it is the Xmas shopping season so they had to work 80 hours a week or because...
 - 2) they hate fishing and fear large bodies of water.
- You can think of this difference as a type of *discrete, unobserved heterogeneity*.

- So the overall probability of being a 0 is a combination of the probabilities of 0's in both groups, weighted by the probability that someone is in each group.

$$\begin{aligned} P(y = 0 \mid \mathbf{x}) &= [\Psi * 1] + [(1 - \Psi) * e^{-\mu}] \\ &= \Psi + [(1 - \Psi) * e^{-\mu}] \end{aligned}$$

$$P(Y_i = y \mid \mathbf{x}) = (1 - \Psi) \frac{e^{(-\mu)} \mu^y}{y!} \text{ for } y > 0$$

- Now we can parameterize Ψ (the probability of not experiencing the transition).

$$\text{Let } \Psi = F(\mathbf{z}\boldsymbol{\gamma})$$

- Where:

- F is either the normal

$$\Psi = \Phi(\mathbf{z}\boldsymbol{\gamma})$$

- or logistic

$$\Psi = \frac{e^{\mathbf{z}\boldsymbol{\gamma}}}{1 + e^{\mathbf{z}\boldsymbol{\gamma}}}$$

- The z 's can be (but don't have to be) the same as the x 's.

Zero-inflated Poisson (ZIP)

- So to maximize the likelihood we

$$P(y = 0 \mid \mathbf{x}) = \Psi + (1 - \Psi) e^{-\mu}$$

$$P(y \mid \mathbf{x}) = (1 - \Psi) \frac{e^{(-\mu)} \mu^y}{y!} \text{ for } y > 0$$

- But what if we still think that the mean and the variance of the counts are still likely to be over-dispersed?
- Zero-inflated negative binomial (ZINB)!
- As we saw last week the expected value of Y using the NB distribution is the same as the Poisson:

$$E(y|x) = \mu$$

- But the variance is different:
- For the ZIP:

$$\text{Var}(y \mid \mathbf{x}, \mathbf{z}) = \mu(1 - \Psi)[1 + \mu\Psi]$$

For the ZINB:

$$\text{Var}(y \mid \mathbf{x}, \mathbf{z}) = \mu(1 - \Psi)[1 + \mu(\Psi + \alpha)]$$

Burgoon (2006) ZIP

```

Zero-inflated Poisson regression          Number of obs   =      1779
                                           Nonzero obs     =      1013
                                           Zero obs        =       766

Inflation model = probit                 LR chi2(12)     =    3258.45
Log likelihood = -5095.227                Prob > chi2     =     0.0000
    
```

-----	-----	-----	-----	-----	-----	-----
terrorincl~d	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	-----
-----	-----	-----	-----	-----	-----	-----
terrorincl~d						
transferslog	-.2317454	.0210745	-11.00	0.000	-.2730506	-.1904402
govleft	-.1325791	.0292379	-4.53	0.000	-.1898843	-.0752738
democ	.015904	.002779	5.72	0.000	.0104572	.0213508
poplog	.2573649	.0155079	16.60	0.000	.2269699	.2877599
govcap	.4330824	.0321101	13.49	0.000	.3701478	.4960171
conflict	-.2438625	.0621844	-3.92	0.000	-.3657416	-.1219833
tradelog	-.1021026	.0394219	-2.59	0.010	-.1793681	-.0248371
terrorinc	.0164768	.0004536	36.33	0.000	.0155879	.0173658
europe	.5573274	.0597876	9.32	0.000	.4401458	.6745089
africa	-.5659734	.0881433	-6.42	0.000	-.738731	-.3932157
asia	-.5459298	.059044	-9.25	0.000	-.6616539	-.4302057
america	.0755931	.0476529	1.59	0.113	-.0178048	.168991
_cons	-3.562389	.4046672	-8.80	0.000	-4.355523	-2.769256
-----	-----	-----	-----	-----	-----	-----
inflate						
transferslog	-.0494114	.0592194	-0.83	0.404	-.1654792	.0666564
govleft	.0014092	.0847792	0.02	0.987	-.164755	.1675735
democ	-.0219012	.0068304	-3.21	0.001	-.0352885	-.0085138
poplog	-.1708239	.0380789	-4.49	0.000	-.2454572	-.0961905
govcap	-.1912593	.1256435	-1.52	0.128	-.437516	.0549975
conflict	-.0546306	.2171013	-0.25	0.801	-.4801413	.3708801
tradelog	-.0664417	.0968829	-0.69	0.493	-.2563286	.1234453
terrorinc	-.2625691	.0265924	-9.87	0.000	-.3146892	-.210449
europe	.3971044	.1800846	2.21	0.027	.0441451	.7500638
africa	.4657872	.1725472	2.70	0.007	.1276009	.8039734
asia	.649158	.1699359	3.82	0.000	.3160898	.9822261
america	.1926071	.1616984	1.19	0.234	-.1243159	.5095302
_cons	2.858982	.9331775	3.06	0.002	1.029988	4.687976

```

-----
Vuong test of zip vs. standard Poisson:          z =      9.90  Pr>z = 0.0000
    
```

Burgoon (2006) ZINB

```

Zero-inflated negative binomial regression      Number of obs =      1779
                                                Nonzero obs   =      1013
                                                Zero obs     =       766

Inflation model = probit                      LR chi2(12)   =      698.27
Log likelihood = -3277.854                    Prob > chi2   =      0.0000
    
```

-----	-----	-----	-----	-----	-----	-----
terrorincl~d	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	-----
-----	-----	-----	-----	-----	-----	-----
terrorincl~d						
transferslog	-.3439978	.0518326	-6.64	0.000	-.4455878	-.2424079
govleft	-.2370396	.0710918	-3.33	0.001	-.3763769	-.0977023
democ	.0180253	.0067103	2.69	0.007	.0048733	.0311773
poplog	.1275457	.0364722	3.50	0.000	.0560616	.1990299
govcap	.3835305	.109199	3.51	0.000	.1695045	.5975566
conflict	-.1360413	.1556708	-0.87	0.382	-.4411504	.1690678
tradelog	-.1009643	.094663	-1.07	0.286	-.2865003	.0845717
terrorinc	.0676185	.0044828	15.08	0.000	.0588324	.0764045
europe	.3661464	.1314223	2.79	0.005	.1085633	.6237294
africa	-.6863947	.1567241	-4.38	0.000	-.9935683	-.3792211
asia	-.52394	.1346819	-3.89	0.000	-.7879117	-.2599682
america	-.204885	.1125723	-1.82	0.069	-.4255226	.0157526
_cons	-2.100327	.9456667	-2.22	0.026	-3.953799	-.246854
-----	-----	-----	-----	-----	-----	-----
inflate						
transferslog	-.2238388	.1036627	-2.16	0.031	-.4270139	-.0206636
govleft	.0486061	.1523741	0.32	0.750	-.2500417	.3472538
democ	-.0122005	.012639	-0.97	0.334	-.0369725	.0125715
poplog	-.287827	.0665332	-4.33	0.000	-.4182296	-.1574244
govcap	-.1846947	.2197949	-0.84	0.401	-.6154847	.2460953
conflict	-.0972352	.4298012	-0.23	0.821	-.93963	.7451596
tradelog	-.1301917	.1651943	-0.79	0.431	-.4539666	.1935831
terrorinc	-.7284495	.1308784	-5.57	0.000	-.9849663	-.4719326
europe	.4025133	.3231119	1.25	0.213	-.2307744	1.035801
africa	.2521182	.2934208	0.86	0.390	-.3229761	.8272125
asia	.5142099	.3012118	1.71	0.088	-.0761543	1.104574
america	-.202345	.2921164	-0.69	0.489	-.7748826	.3701925
_cons	4.314936	1.599341	2.70	0.007	1.180285	7.449587
-----	-----	-----	-----	-----	-----	-----
/lnalpha	-.2373784	.0653522	-3.63	0.000	-.3654664	-.1092903
-----	-----	-----	-----	-----	-----	-----
alpha	.7886928	.0515428			.6938729	.8964701
-----	-----	-----	-----	-----	-----	-----
Vuong test of zinb vs. standard negative binomial: z =				7.00	Pr>z = 0.0000	

Vuong statistic

- A means of testing which non-nested model is to be preferred.

$$V = \frac{\sqrt{N\bar{m}}}{s_m}$$

$$\text{Where } m = \ln \left[\frac{\widehat{P}_1(y | x)}{\widehat{P}_2(y | x)} \right]$$

And \bar{m} is the mean of m and s_m the standard deviation of m .

Vuong statistic

- In general, positive Vuong statistics suggest that the zero-inflated models are preferred while significant negative statistics indicate non-zero-inflated models are preferred.

Overall results

	(1)	(2)	(3)		(4)	
	Poisson	Negative Binomial	ZIP	ZIP Inflate	ZINB	ZINB Inflate
transferslog	-0.320**	-0.300**	-0.232***	-0.049	-0.344***	-0.224*
	0.104	0.107	0.021	0.059	0.052	0.104
govleft	-0.244*	-0.244*	-0.133***	0.001	-0.237***	0.049
	0.097	0.097	0.029	0.085	0.071	0.152
democ	0.026*	0.027*	0.016***	-0.022**	0.018**	-0.012
	0.011	0.012	0.003	0.007	0.007	0.013
poplog	0.241**	0.252**	0.257***	-0.171***	0.128***	-0.288***
	0.074	0.077	0.016	0.038	0.036	0.067
govcap	0.464**	0.426*	0.433***	-0.191	0.384***	-0.185
	0.178	0.189	0.032	0.126	0.109	0.22
conflict	-0.087	-0.028	-0.244***	-0.055	-0.136	-0.097
	0.177	0.165	0.062	0.217	0.156	0.43
tradelog	-0.046	-0.002	-0.102**	-0.066	-0.101	-0.13
	0.158	0.166	0.039	0.097	0.095	0.165
terrorinc	0.085***	0.084***	0.016***	-0.263***	0.068***	-0.728***
	0.013	0.013	0	0.027	0.004	0.131
europa	0.203	0.159	0.557***	0.397*	0.366**	0.403
	0.251	0.265	0.06	0.18	0.131	0.323
africa	-1.066***	-1.080***	-0.566***	0.466**	-0.686***	0.252
	0.26	0.252	0.088	0.173	0.157	0.293
asia	-0.779**	-0.829***	-0.546***	0.649***	-0.524***	0.514
	0.239	0.239	0.059	0.17	0.135	0.301
america	-0.199	-0.195	0.076	0.193	-0.205	-0.202
	0.204	0.204	0.048	0.162	0.113	0.292
Constant	-4.426*	-4.566**	-3.562***	2.859**	-2.100*	4.315**
	1.721	1.754	0.405	0.933	0.946	1.599
N	1779	1779	1779		1779	
chi2	258.442	406.513	3258.453		698.268	
p	0	0	0		0	
ll	-3380.86	-3351.12	-5095.23		-3277.85	
alpha	1.193	1.12				

* p<0.05, ** p<0.01, *** p<0.001

Interpretation

- You can interpret predicted probabilities in the same way that you can for poisson or negative binomial models.

In sum...

- We have examined four basic types of event count models:
 - Poisson (PRM)
 - Negative Binomial (NB)
 - Zero-Inflated Poisson (ZIP)
 - Zero-Inflated Negative Binomial (ZINB)
- The PRM is nested in the ZIP (*how?*).
- The NB is nested within the ZINB.

- Deciding which model is appropriate is straightforward given the alpha test and Vuong's test.
- There are a number of other count models out there that I have not covered in detail (in part because they are less common and in part because they are variations on the same theme).

- Now, let's go through the rest of the readings.

- Questions?