Quantitative Data Analysis II

SOC 781

Mlogit review &

Count models

Today we will…

- Mlogit example
- Cover count models

Data: dependent variables

- Pain/Distress: back or joint aches almost every day; K6; DSM-III drug and alcohol misuse
  - binary indicators

- DVa: Languishing, Moderate, Flourishing
  - three category Mental Health Continuum
- DVb: Lpd, L, Mpd, M, Fpd, F
  - six category MHC combined with pain/distress indicator
- DVc: Lp&d, Lp, L, Mp&d, Mp, M, Fp&d, Fp, F
  - nine category MHC combined with pain&distress and pain indicators

Data: independent variables

- SES index: R’s (and S’s, if applicable) edu, occup, HHinc, assets
  - standardized → percentile ranks
  - comparable across waves, avoids shift in edu composition (Hendi, 2015)

- Race/ethnicity: NHwhite vs. other
  - not enough minorities to disaggregate

- Period (survey wave); Age (years); Sex (M,F)
  - findings consistent across sex

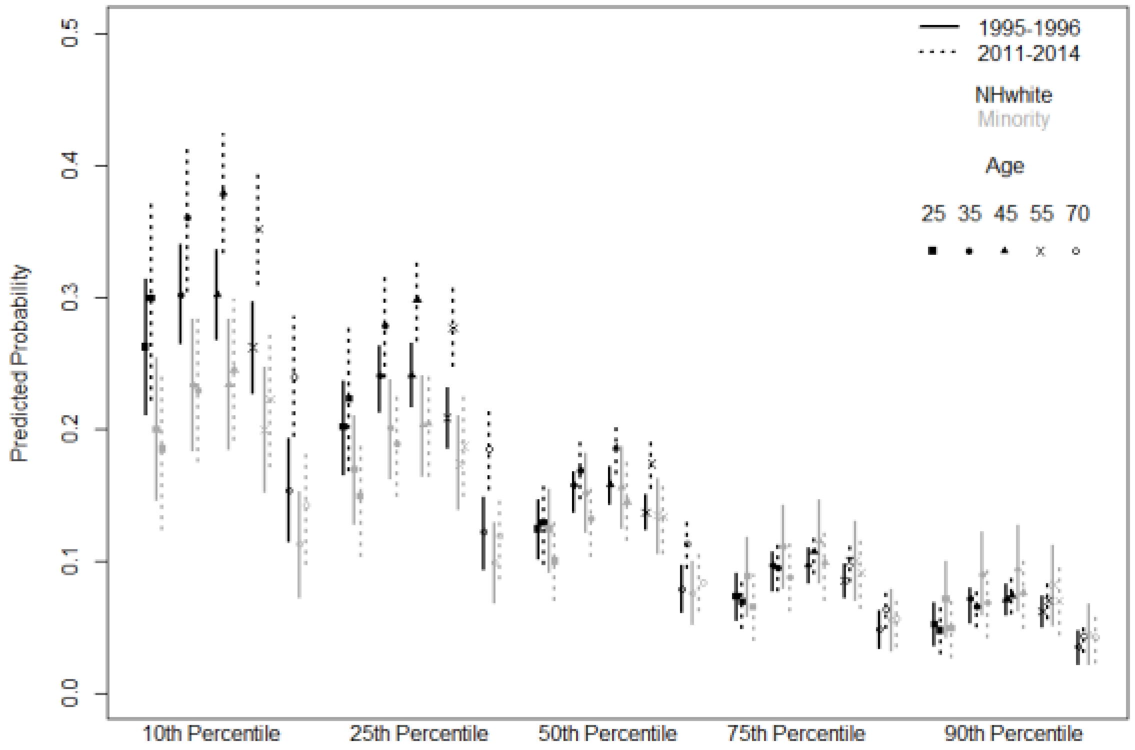
- Multinomial logistic regression models
  - age-race/ethnicity-SES-period predicted probabilities

Competing hypotheses

- H1: Substantial isolated increase in languishing
    - would support DoD narrative
- H2: Relatively consistent languishing but isolated increase in PD
    - suggests something else is going on
        - links between stress and health well-established

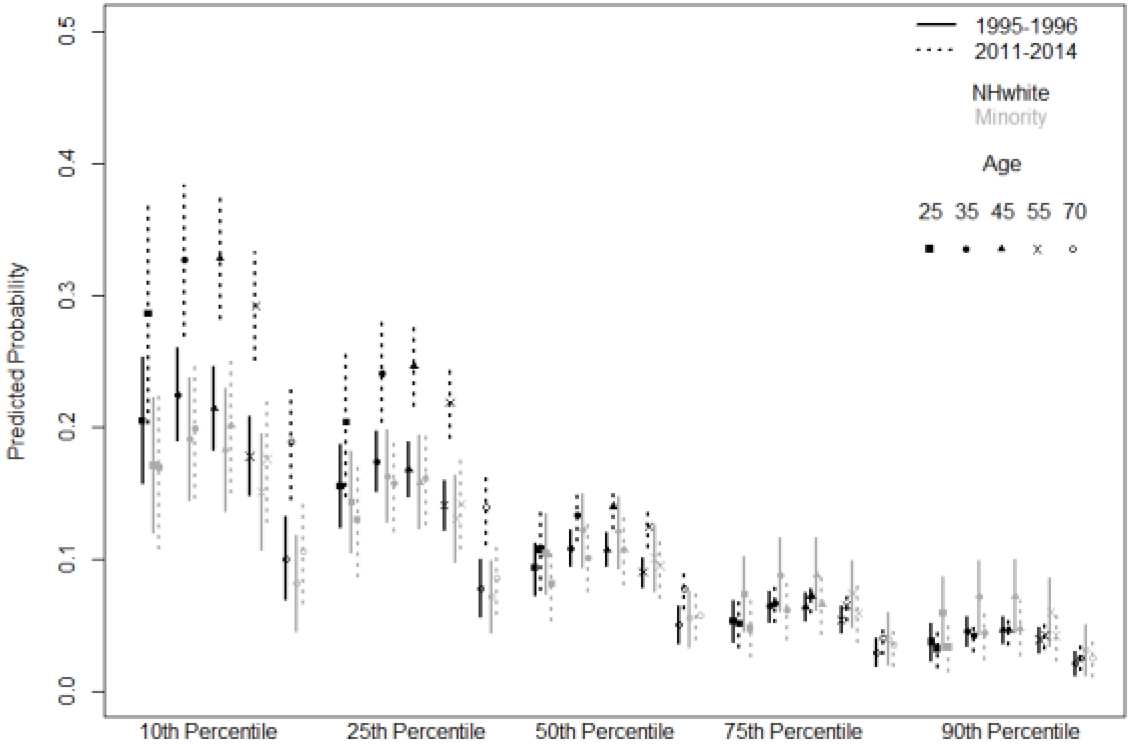
RQ1: Has "despair" and/or distress uniquely risen?

H1: Substantial isolated increase in languishing?



DVa (L, M, F); only L shown; 84% CIs

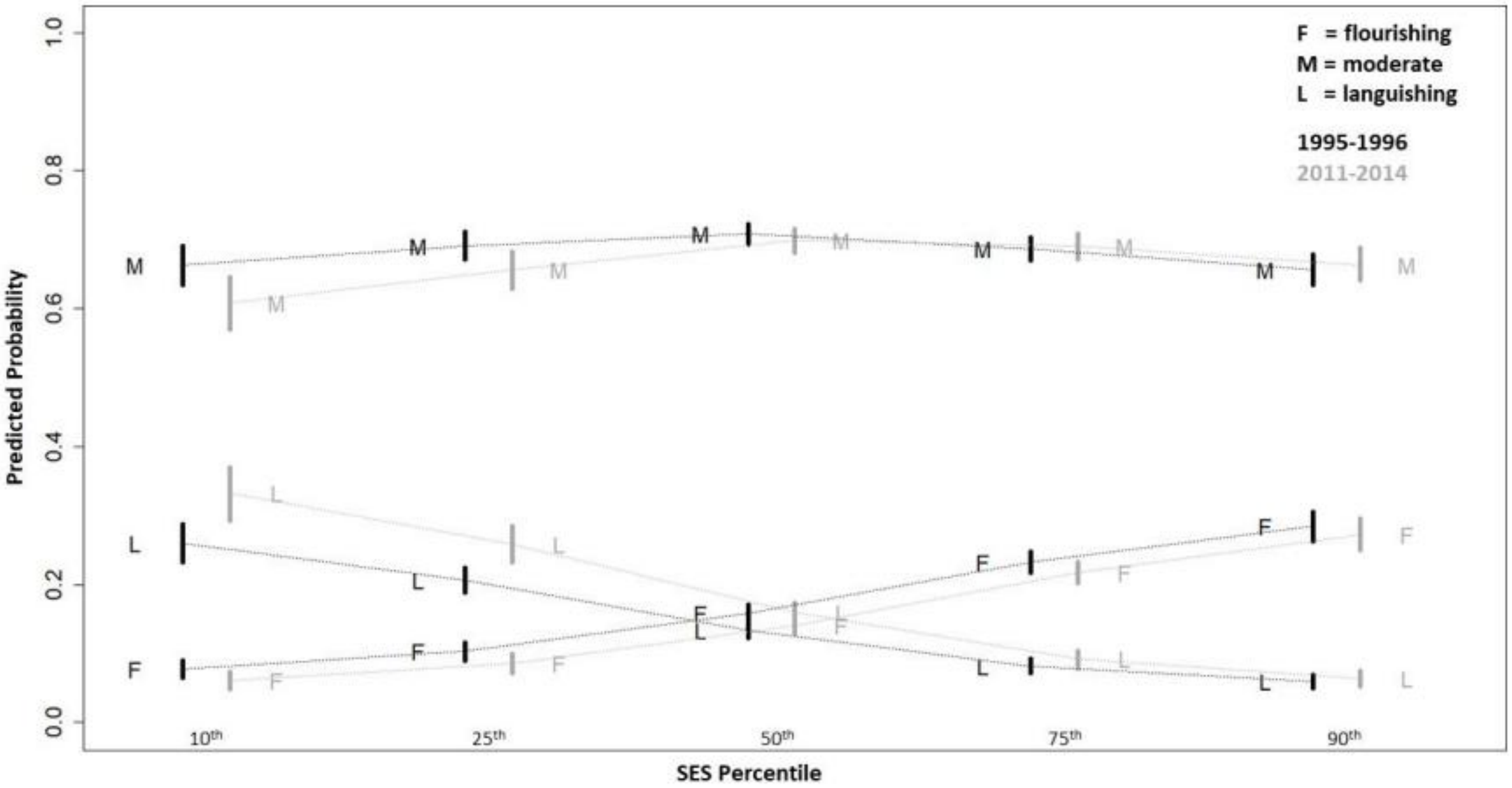
H2: Consistent languishing, but substantial isolated increase in pain/distress?



DVb (Lpd, L, Mpd, M, Fpd, F); only Lpd shown; 84% CIs

RQ1: Has “despair” and/or distress uniquely risen?

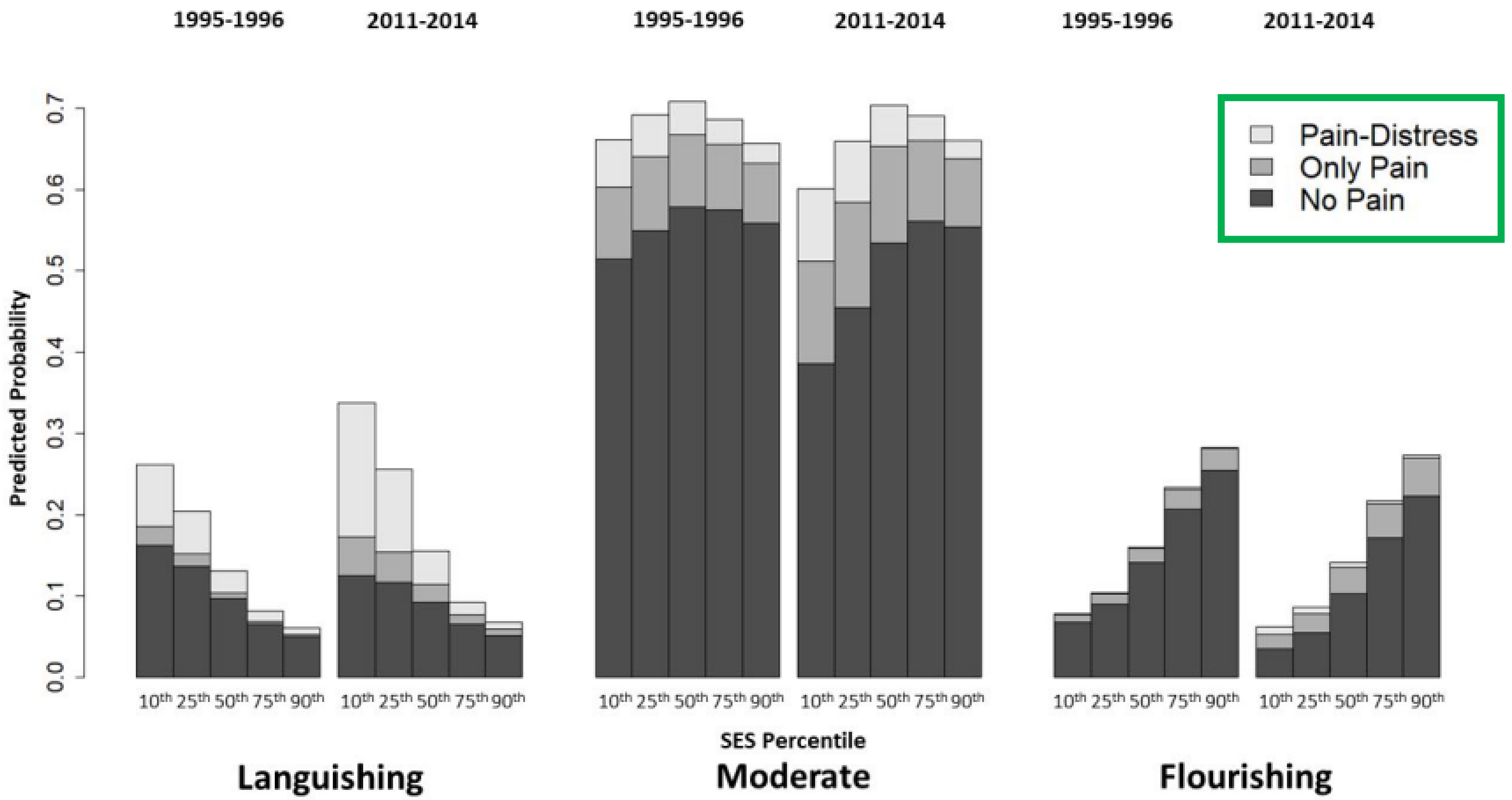
H1: Substantial isolated increase in languishing among Nhwhites?



DVa (L, M, F); based on average marginal effects across age; 84% CIs

RQ1: Has “despair” and/or distress uniquely risen?

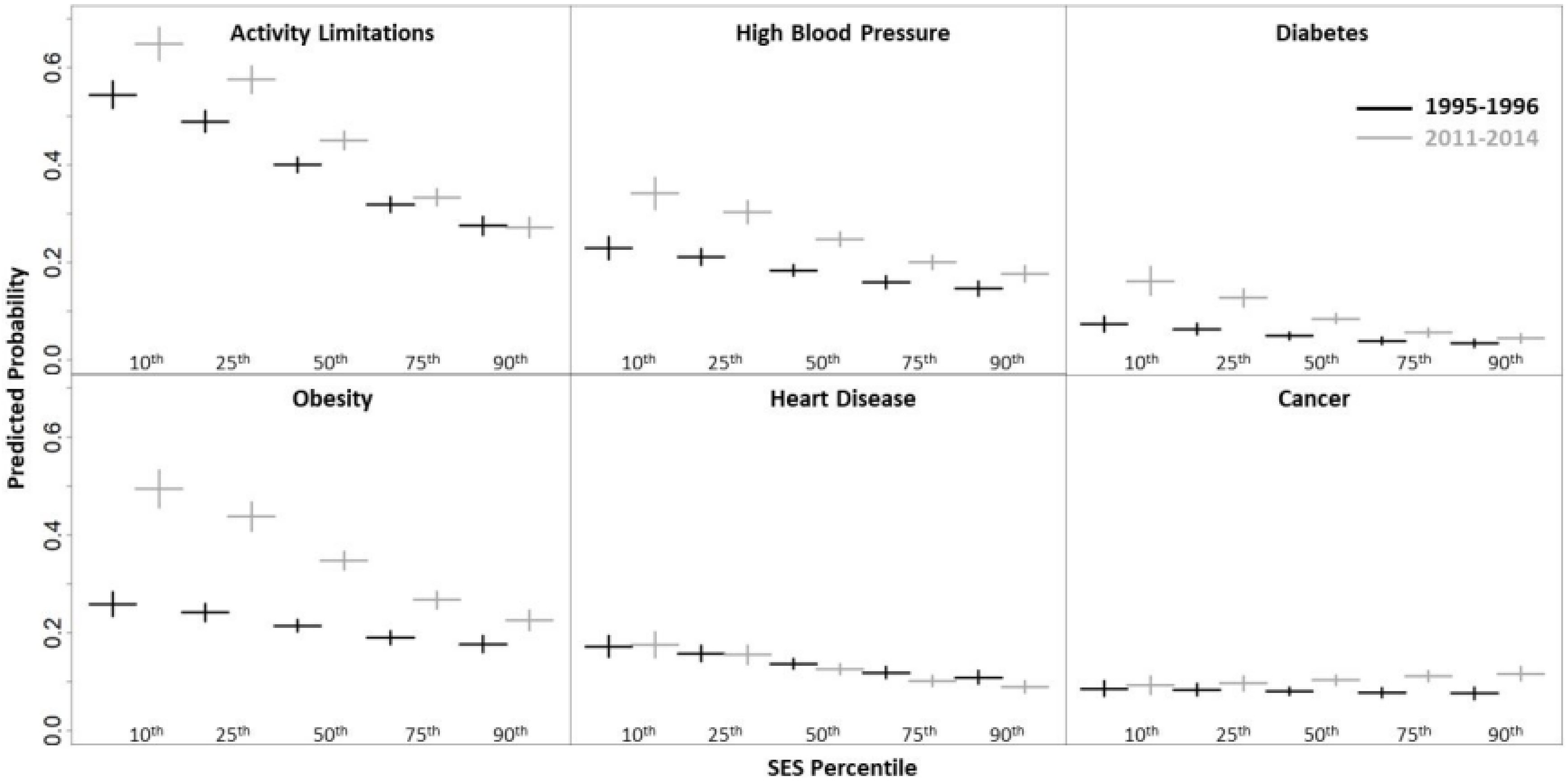
H2: Consistent languishing, but substantial isolated increase in pain/distress among Nhwhites?



DVc (Lp&d, Lp, L, Mp&d, Mp, M, Fp&d, Fp, F); based on average marginal effects across age

RQ2: Has morbidity increased among poor whites?

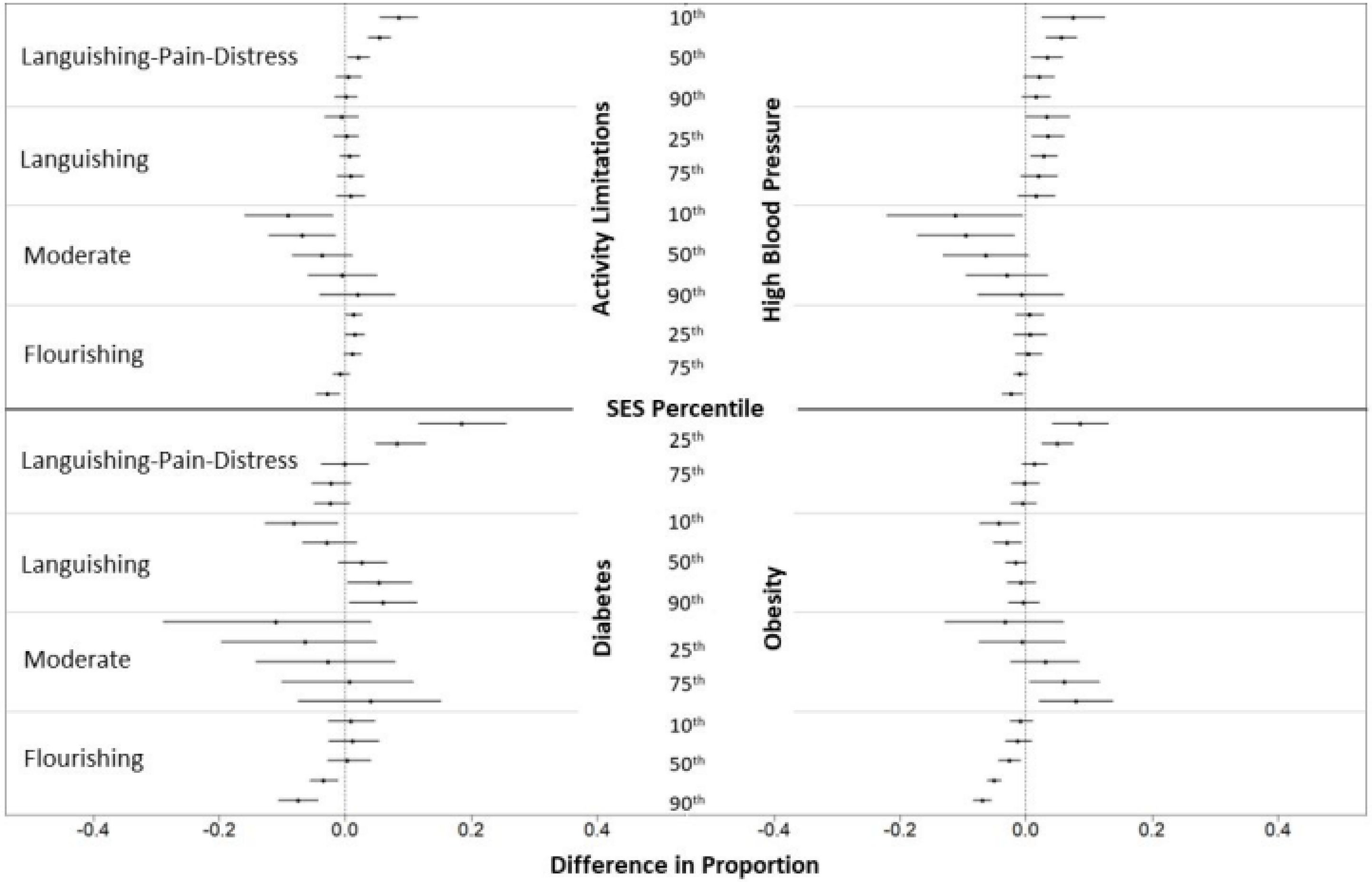
Trends in socioeconomic gradient in morbidity among NHwhites



based on binary logits for each respective outcome; 84% CIs; trends in gradients for racial/ethnic minorities relatively consistent (also worse off except HD and cancer); cancer pattern flipped, racial/ethnic gap for cancer also flipped

RQ3: Were increases driven by weathering?

Trends in LpdLMF contribution to morbidity prevalence among NHwhites



Difference in proportion computed based on the SES-specific contribution in 2011-2014 subtracted from the SES-specific contribution in 1995-1996 for each respective chronic condition

Count outcomes

- DV reflects how many times something has happened
  - Cannot fall below zero

- Possible quirks to consider…
  - often does NOT resemble normal distribution
    - and “clumps” around certain numbers
  - sometimes there are a lot of zeros
    - rare events
  - sometimes zeros not included
    - e.g., only asked how many times among those who had event
  - sometimes going from 0 to 1 is a lot different than from 1 to 2, or 2 to 3…
    - e.g., once you’ve done it once its easier to do it again
  - sometimes need to consider “exposure”
    - e.g., risk of event occurring

Poisson regression: assumptions

- $\text{Log}(y)$is linearly related to unit-changes in covariates
- Events are independent
  - when the event occurs, it does not affect the prob. of occurring again
- Variance of y equals the mean of y
  - if var(y)> μ = overdispersion
  - if var(y)< μ = underdispersion

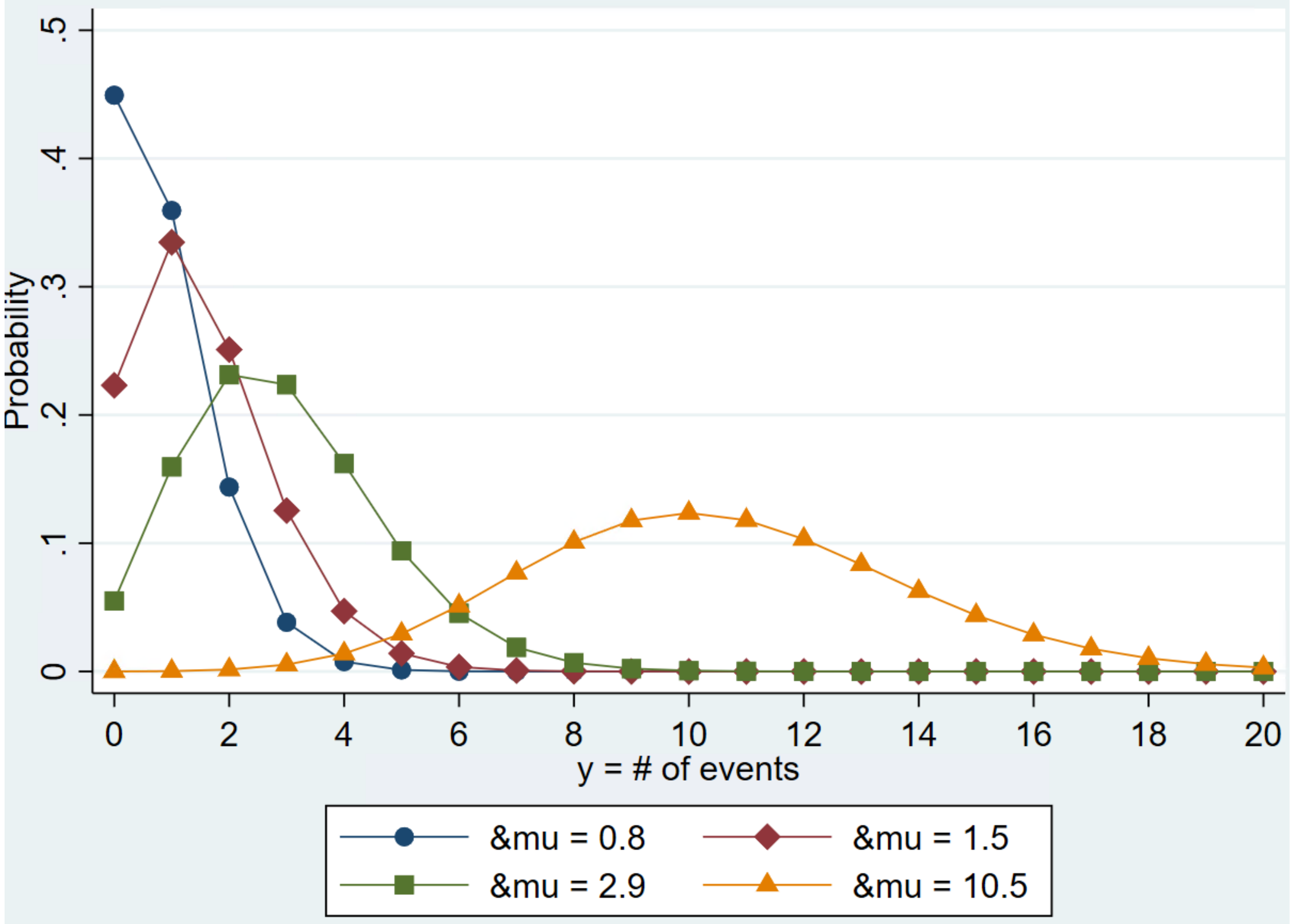
Poisson regression

- $\text{Log}(y) = b_0 + b_1x_1 + b_2x_2 + \cdots$
- y is the count, bs are estimates of the reg. coef., and xs are predictors
- bs are the log of the expected count
  - NOT log odds

Poisson distribution

$$\Pr(y|\mu) = \frac{e^{-\mu}\mu^{y}}{y!}$$

- Var(y) = μ : assumes equidispersion (often over, rarely under)
- As μ increases probability of zero count decreases rapidly
- As mean of distribution μ increases, mass of distribution shifts right
- As μ increases distribution approximates normality



Count outcome: example

```
gen numpart =.
replace numpart = numwomen if female==0
replace numpart = nummen if female==1
replace numpart=. if numpart>750
tab numpart if nmiss==0
```

- Number of opposite-sex (heterosexual) sex partners since 18-years-old

numpart	Freq.	Percent	Cum.
0	1,710	6.13	6.13
1	7,005	25.10	31.23
2	3,044	10.91	42.13
3	2,724	9.76	51.90
4	1,981	7.10	58.99
5	2,137	7.66	66.65
6	1,262	4.52	71.17
7	621	2.23	73.40
8	653	2.34	75.74
9	228	0.82	76.56
10	1,691	6.06	82.61
11	104	0.37	82.99
12	516	1.85	84.84
13	88	0.32	85.15
14	81	0.29	85.44
15	697	2.50	87.94
16	63	0.23	88.16
17	41	0.15	88.31
18	92	0.33	88.64
19	18	0.06	88.71
20	904	3.24	91.94
21	36	0.13	92.07
22	38	0.14	92.21
23	28	0.10	92.31
24	34	0.12	92.43
25	365	1.31	93.74
26	13	0.05	93.79
27	17	0.06	93.85
28	22	0.08	93.93
29	5	0.02	93.94
30	381	1.37	95.31
31	6	0.02	95.33
32	12	0.04	95.37
33	17	0.06	95.44
34	6	0.02	95.46
35	84	0.30	95.76
36	14	0.05	95.81
37	3	0.01	95.82
38	2	0.01	95.83
39	1	0.00	95.83
40	165	0.59	96.42
41	2	0.01	96.43
42	5	0.02	96.45
43	1	0.00	96.45
45	32	0.11	96.56
46	1	0.00	96.57
47	2	0.01	96.57
48	2	0.01	96.58
49	2	0.01	96.59
50	336	1.20	97.79
51	5	0.02	97.81
52	6	0.02	97.83
53	1	0.00	97.84
54	2	0.01	97.84
55	5	0.02	97.86
56	3	0.01	97.87
57	1	0.00	97.88
58	1	0.00	97.88
59	1	0.00	97.88
60	52	0.19	98.07
61	1	0.00	98.07
62	2	0.01	98.08
63	3	0.01	98.09
65	10	0.04	98.13
66	1	0.00	98.13
68	2	0.01	98.14
69	1	0.00	98.14
70	23	0.08	98.22
73	3	0.01	98.23
74	1	0.00	98.24
75	36	0.13	98.37
77	1	0.00	98.37
80	14	0.05	98.42
82	1	0.00	98.42
83	1	0.00	98.43
84	1	0.00	98.43
85	6	0.02	98.45
86	2	0.01	98.46
87	2	0.01	98.47
89	1	0.00	98.47
90	7	0.03	98.50
96	1	0.00	98.50
99	2	0.01	98.51
100	230	0.82	99.33
101	5	0.02	99.35
103	1	0.00	99.35
105	1	0.00	99.36
110	2	0.01	99.36
120	11	0.04	99.40
121	1	0.00	99.41
122	4	0.01	99.42
125	3	0.01	99.43
130	3	0.01	99.44
137	1	0.00	99.44
138	1	0.00	99.45
140	2	0.01	99.46
145	1	0.00	99.46
147	1	0.00	99.46
150	32	0.11	99.58
165	1	0.00	99.58
167	1	0.00	99.58
170	2	0.01	99.59
175	3	0.01	99.60
200	57	0.20	99.81
201	1	0.00	99.81
210	1	0.00	99.81
222	2	0.01	99.82
240	1	0.00	99.82
250	7	0.03	99.85
253	1	0.00	99.85
255	1	0.00	99.86
270	1	0.00	99.86
280	1	0.00	99.86
300	20	0.07	99.94
301	1	0.00	99.94
336	1	0.00	99.94
350	3	0.01	99.95
365	1	0.00	99.96
380	1	0.00	99.96
400	1	0.00	99.96
403	1	0.00	99.97
450	1	0.00	99.97
500	6	0.02	99.99
700	1	0.00	100.00
750	1	0.00	100.00
Total	27,908	100.00	

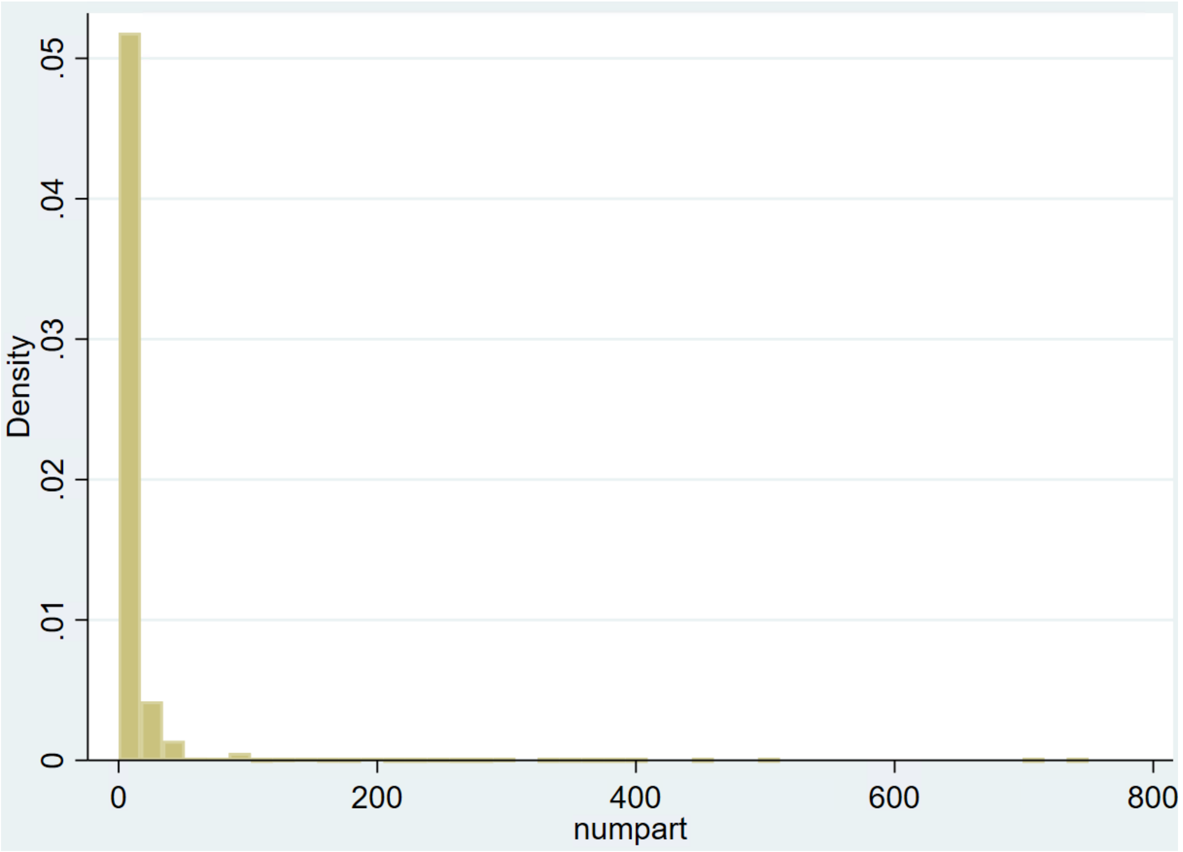
Count outcome: example

```
sum numpart if nmiss==0, detail
Mean          8.961481
Std. Dev.     22.9092

Variance      524.8315
```

- Overdispersion

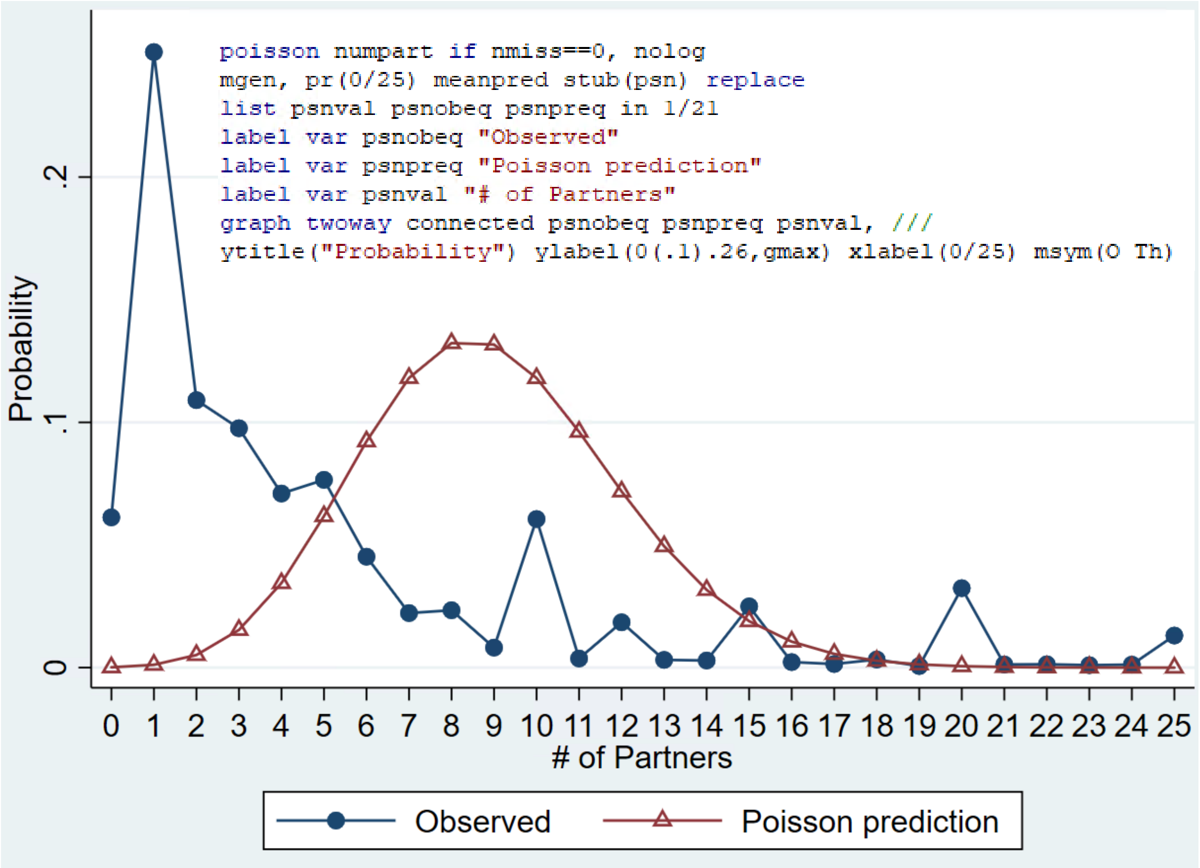
```
hist numpart if nmiss==0
```



- Compare observed and predicted
  - univariate not account heterogeneity

```
poisson numpart if nmiss==0, nolog
```

numpart	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	2.192935	.0019996	1096.68	0.000	2.189016	2.196855



Count outcome: example

- Does number of partners differ by sex?

```
sort female
by female: sum numpar if nmiss==0
```

-> female = 0

Variable	Obs	Mean	Std. Dev.	Min	Max
numpart	12,214	14.11544	31.74637	0	750

-> female = 1

Variable	Obs	Mean	Std. Dev.	Min	Max
numpart	15,694	4.950363	10.59282	0	365

- Appears so, but need to test
  - Poisson model

```
poisson numpart i.female if nmiss==0
```

numpart	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
1.female	-1.047808	.0043211	-242.49	0.000	-1.056278	-1.039339
_cons	$\boxed{2.647269}$	.0024084	1099.19	0.000	2.642549	2.65199

- Who's the constant?

exp(2.647)=14.11

- Female

exp(-1.048)=0.35

- What does this mean?
- (14.11 – 4.95) / 14.11 = 0.65
- 14.11 – (14.11*0.65) = 4.95

```
gen male=.
replace male=1 if female==0
replace male=0 if female==1
poisson numpart i.male if nmiss==0
```

numpart	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
1.male	1.047808	.0043211	242.49	0.000	1.039339	1.056278
_cons	1.599461	.0035877	445.82	0.000	1.592429	1.606493

- exp(1.599461) = 4.95

Factor and percentage changes

- Incident rate ratio (IRR) more informative than ORs

```
poisson numpart c.age##c.age i.female i.nonwhite c.educ i.married if nmiss==0, irr
```

numpart	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	1.125314	.0008983	147.90	0.000	1.123555	1.127076
c.age#c.age	.9988228	8.20e-06	-143.44	0.000	.9988067	.9988389
1.female	.3470288	.0015064	-243.80	0.000	.3440887	.349994
1.nonwhite	.9782051	.0050396	-4.28	0.000	.9683774	.9881325
educ	1.009942	.0006978	14.32	0.000	1.008575	1.011311
1.married	.5723129	.0024113	-132.46	0.000	.5676064	.5770584
_cons	1.132287	.0225371	6.24	0.000	1.088965	1.177332

- Who, and what's the constant?

```
listcoef, percent
```

	b	z	P>|z|	%	%StdX	SDofX
age	0.1181	147.905	0.000	12.5	645.1	17.010
c.age#c.age	-0.0012	-143.443	0.000	-0.1	-86.6	1708.765
1.female	-1.0583	-243.803	0.000	-65.3	-40.8	0.496
1.nonwhite	-0.0220	-4.277	0.000	-2.2	-0.9	0.398
educ	0.0099	14.317	0.000	1.0	2.9	2.915
1.married	-0.5581	-132.458	0.000	-42.8	-24.3	0.500
constant	0.1242	6.242	0.000	.	.	.

- Being female decreases the expected number of partners by a factor of 0.35, or 65%, holding all else constant

- Each additional year of education increases the number of expected partners by a factor of 1.01, or 1%, holding all else constant

Marginal effects

- Marginal effect: Δ in the predicted rate given a Δ in X
  - holding all other Xs constant
    - Is there a meaningful way to hold all other Xs constant?

- Average marginal effect (AME): the average of the marginal effect for all observations
  - Likely, no one is "average." What about underrepresented groups?

- Marginal effect at the mean (MEM): all other Xs held at their means
  - Many mean values are often meaningless (e.g., dummy Xs)

- Marginal effect at representative values (MER): all other Xs held at substantively meaningful values
  - What are "meaningful" values? Can become quickly overwhelmed with details

Average marginal effect (AME)

- Avg. Δ in predicted rate for Δ in X, holding all else constant

```
poisson numpart c.age##c.age i.female i.nonwhite c.educ i.married if nmiss==0, irr
mchange
```

	Change	p-value
age		
+1	0.096	0.000
+SD	0.046	0.000
Marginal	0.101	0.000
female		
1 vs 0	-9.272	0.000
nonwhite		
1 vs 0	-0.196	0.000
educ		
+1	0.089	0.000
+SD	0.262	0.000
Marginal	0.089	0.000
married		
1 vs 0	-4.961	0.000

- On average, a one-year increase in age is associated with a 0.096 increase in the rate of partners
- Being female decreases the expected number of partners by 9.272, on average

Marginal effect at the mean (MEM)

```
poisson numpart c.age##c.age i.female i.nonwhite c.educ i.married if nmiss==0, irr
mchange, atmeans
```

	Change	p-value
age		
+1	0.103	0.000
+SD	-1.413	0.000
Marginal	0.115	0.000
female		
1 vs 0	-12.072	0.000
nonwhite		
1 vs 0	-0.223	0.000
educ		
+1	0.101	0.000
+SD	0.298	0.000
Marginal	0.101	0.000
married		
1 vs 0	-5.701	0.000

- For respondents average on all characteristics, a one-year increase in age is associated with a 0.103 increase in the rate of partners
- Being female decreases the expected number of partners by 12.072, when all characteristics are held at global means

Marginal effect at representative values (MER)

```
mchange female, at(female=1 nonwhite=0 educ=12 married=0 age=40)
```

	Change	p-value
female		
1 vs 0	-14.219	0.000

	age	female	nonwhite	educ	married
at	40	1	0	12	0

- Among those who are white, HS educated, age 40, and married being female decreases the expected number of partners by 14.219

```
mchange female, at(female=1 nonwhite=0 educ=12 married=1 age=40)
```

	Change	p-value
female		
1 vs 0	-8.138	0.000

	age	female	nonwhite	educ	married
at	40	1	0	12	1

- Among those who are white, HS educated, age 40, and NOT married being female decreases the expected number of partners by 8.138

Ideal types

```
mtable, at(female=(0 1) married=0 nonwhite=0 educ=12 age=40) pr(0/100) width(3)
```

	female	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	0	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.005	0.008	0.014	0.022	0.031	0.043	0.055	0.066	0.076	0.082	0.086	0.085	0.080	0.073	0.063
2	1	0.001	0.004	0.015	0.038	0.071	0.107	0.135	0.146	0.138	0.116	0.087	0.060	0.038	0.022	0.012	0.006	0.003	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000

	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52
1	0.053	0.043	0.033	0.025	0.018	0.013	0.009	0.006	0.004	0.002	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

- Should all sum to 1 for males and females, respectively
    - cut off at 100
- Doesn’t make a whole lot of sense for DV with such a range

Ideal types: differences

```
mchange female, at(married=0 nonwhite=0 educ=12 age=40) pr(0/100) ///
stat(from to change p) brief
```

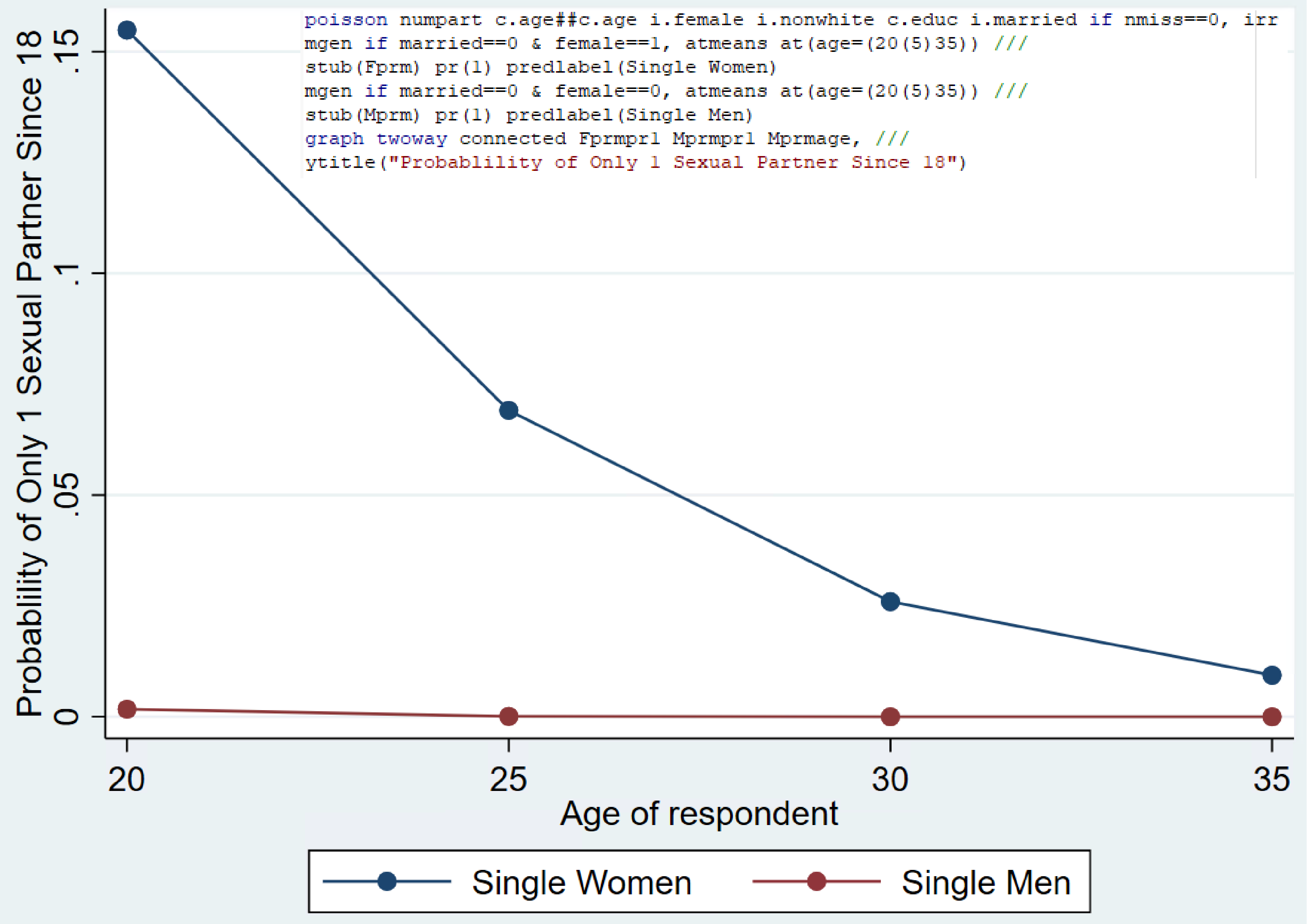
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
female																	
From	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.005	0.008	0.014	0.022	0.031	0.043
To	0.001	0.004	0.015	0.038	0.071	0.107	0.135	0.146	0.138	0.116	0.087	0.060	0.038	0.022	0.012	0.006	0.003
1 vs 0	0.001	0.004	0.015	0.038	0.071	0.107	0.135	0.146	0.137	0.115	0.085	0.056	0.030	0.008	-0.010	-0.025	-0.040
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
female																	
From	0.055	0.066	0.076	0.082	0.086	0.085	0.080	0.073	0.063	0.053	0.043	0.033	0.025	0.018	0.013	0.009	0.006
To	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1 vs 0	-0.053	-0.066	-0.076	-0.082	-0.085	-0.085	-0.080	-0.073	-0.063	-0.053	-0.043	-0.033	-0.025	-0.018	-0.013	-0.009	-0.006
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

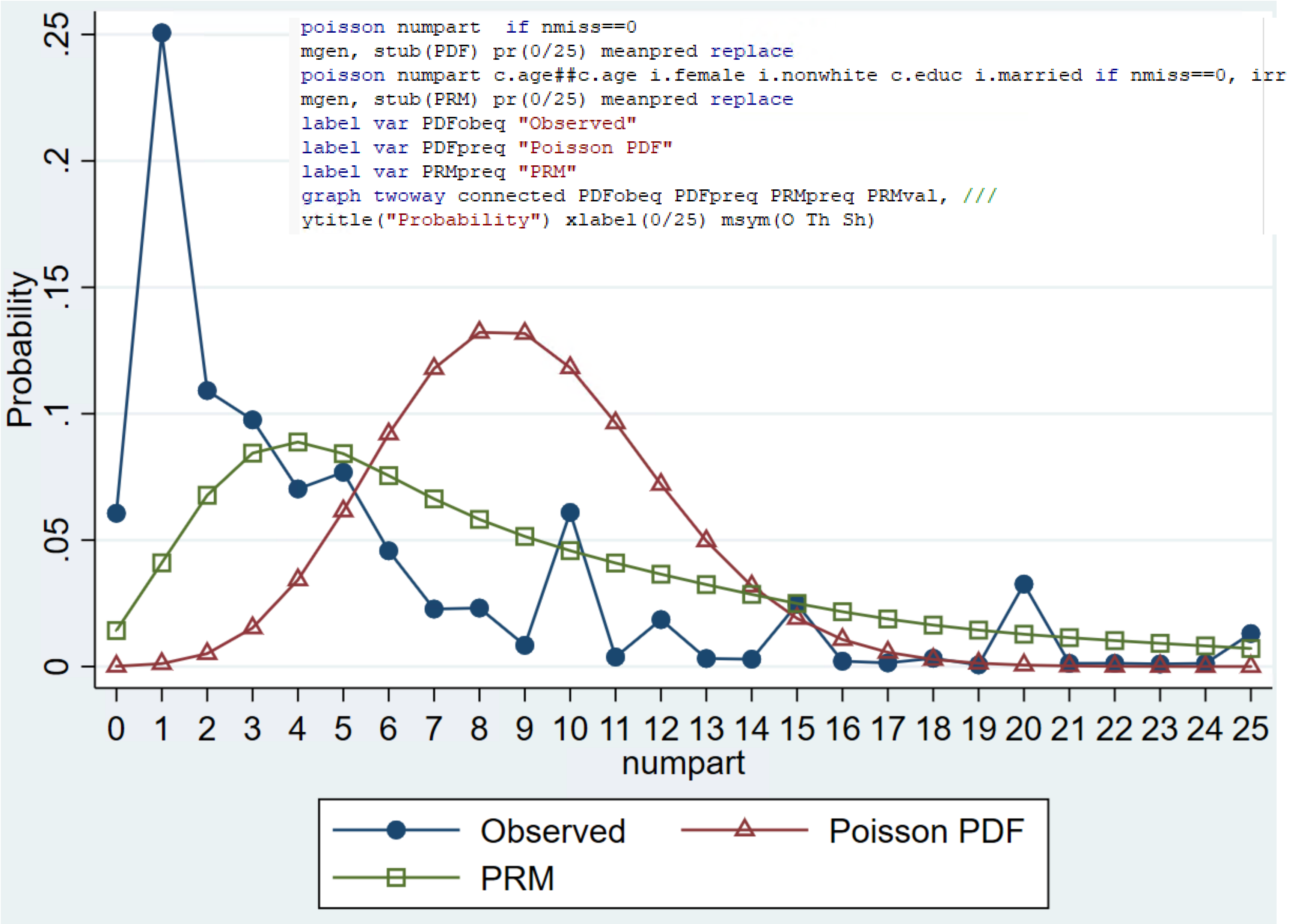
	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
female																	
From	0.004	0.002	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
To	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1 vs 0	-0.004	-0.002	-0.001	-0.001	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67
female																	
From	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
To	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1 vs 0	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Graphing predicted probabilities



Model fit: Informal assessment



Negative binomial regression model (NBRM)

- Adds parameter to reflect unobserved heterogeneity among observations
  - recall how model with covariates better reflects observed counts vs. univariate model
    - now adjust based on what we can’t observe (assumptions) to deal with overdispersion

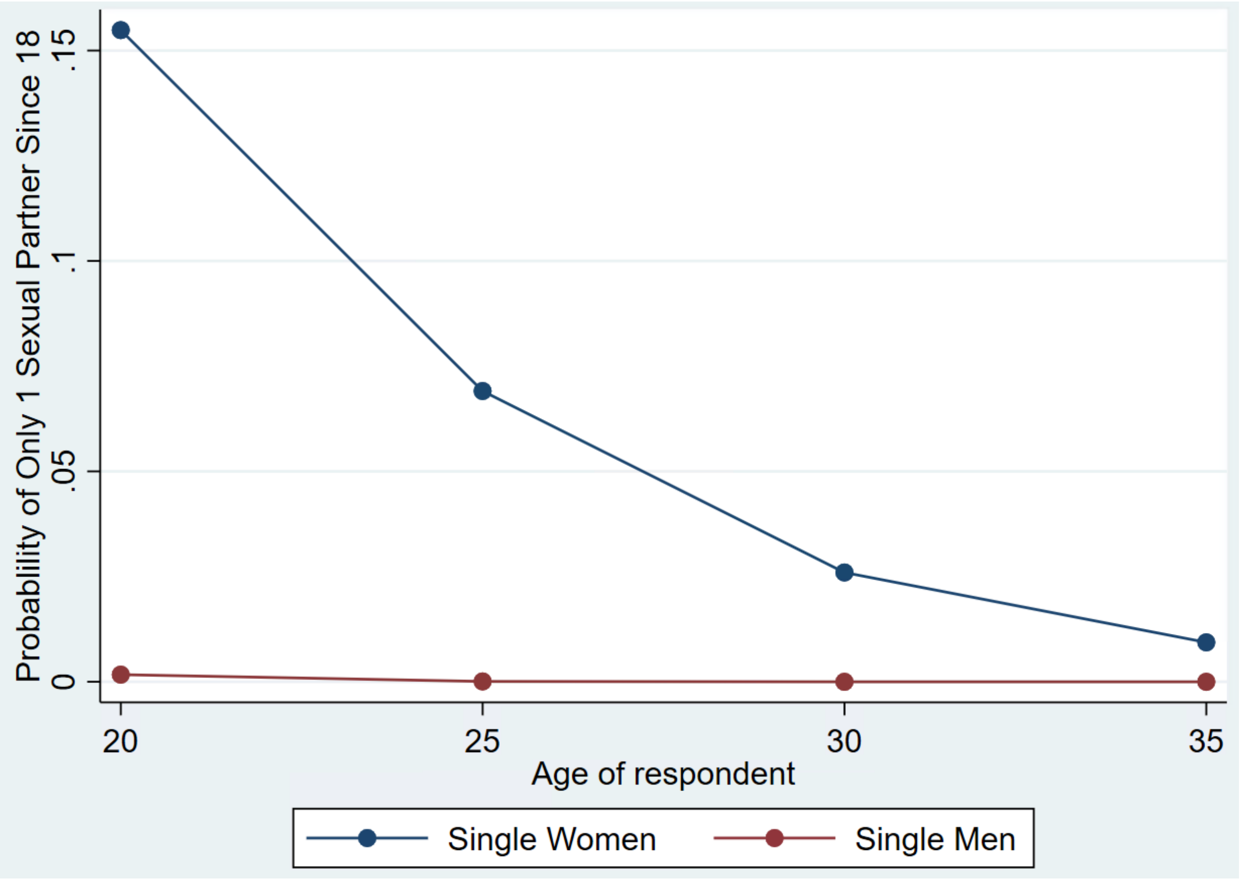
Variable	PRM	NBRM
numpart		
Age of respondent	1.125	1.117
	147.90	44.33
	0.000	0.000
c.age#c.age	0.999	0.999
	-143.44	-46.07
	0.000	0.000
1	0.347	0.346
	-243.80	-73.58
	0.000	0.000
1	0.978	0.932
	-4.28	-3.78
	0.000	0.000
Highest year of school~d	1.010	1.021
	14.32	8.33
	0.000	0.000
1	0.572	0.585
	-132.46	-36.02
	0.000	0.000
Constant	1.132	1.240
	6.24	3.33
	0.000	0.001

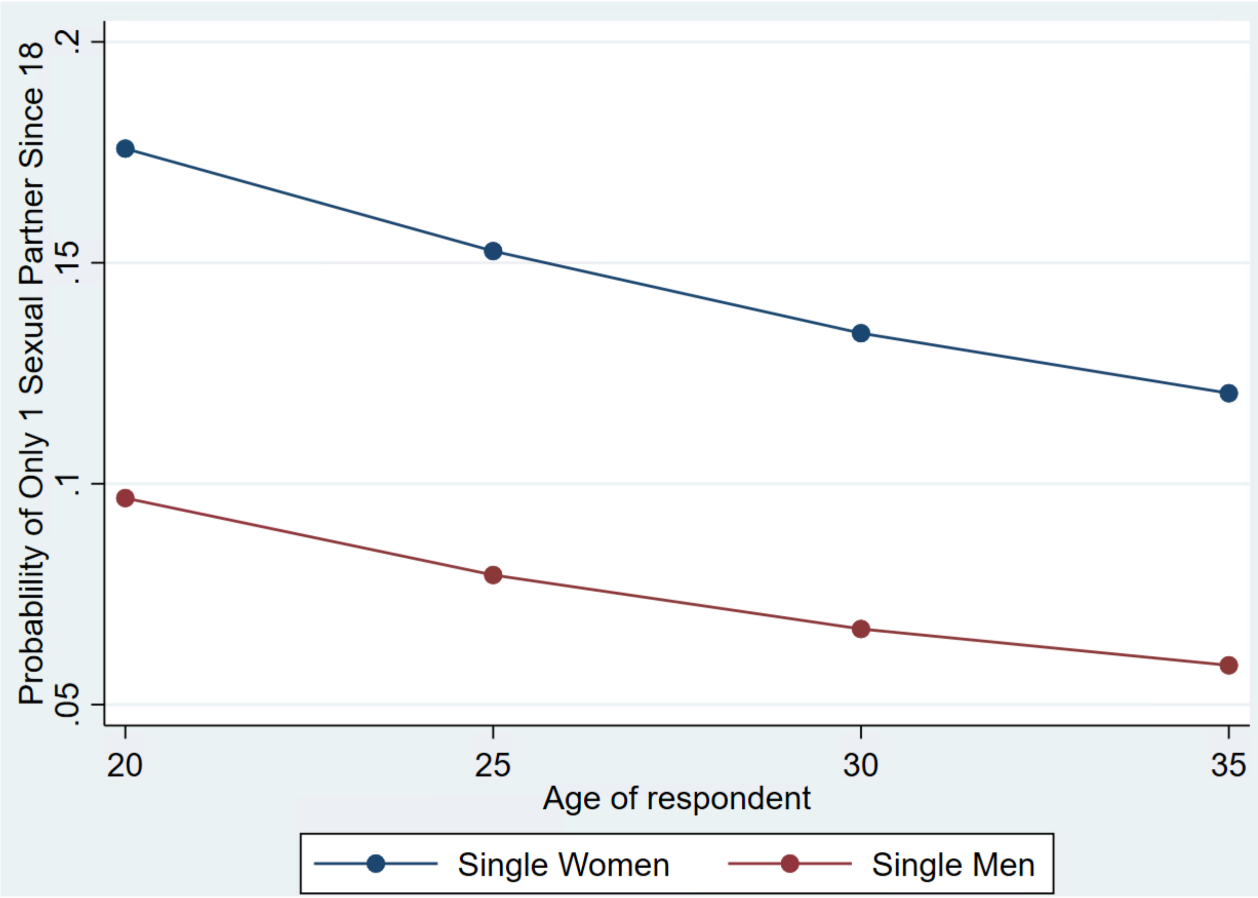
```
nbreg numpart c.age##c.age i.female i.nonwhite c.educ i.married if nmiss==0, irr
estimates store NBRM
/*compare with Poisson*/
quietly poisson numpart c.age##c.age i.female i.nonwhite c.educ i.married ///
 if nmiss==0, irr
 estimates store PRM
 estimates table PRM NBRM, b(%9.3f) t p(%9.3f) varlabel ///
stats(alpha N) eform vsquish
```

- Smaller standard errors
- Comparable coef. and p-values
  - likely because such large sample
- Typically, more impactful on predicted counts

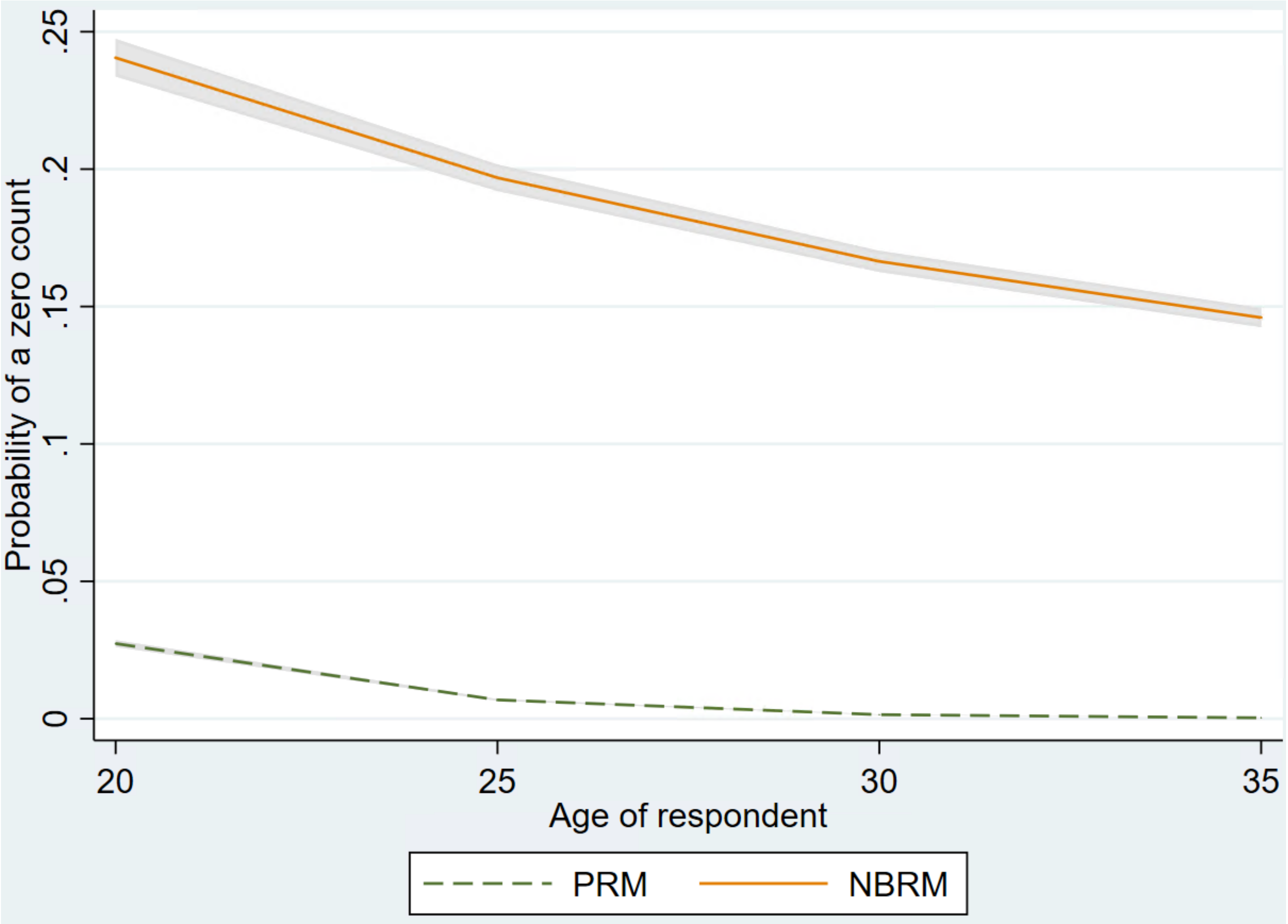
Negative binomial regression model (NBRM)

- Compare the probability of only 1 partner from Poisson (left) and NBRM (right)



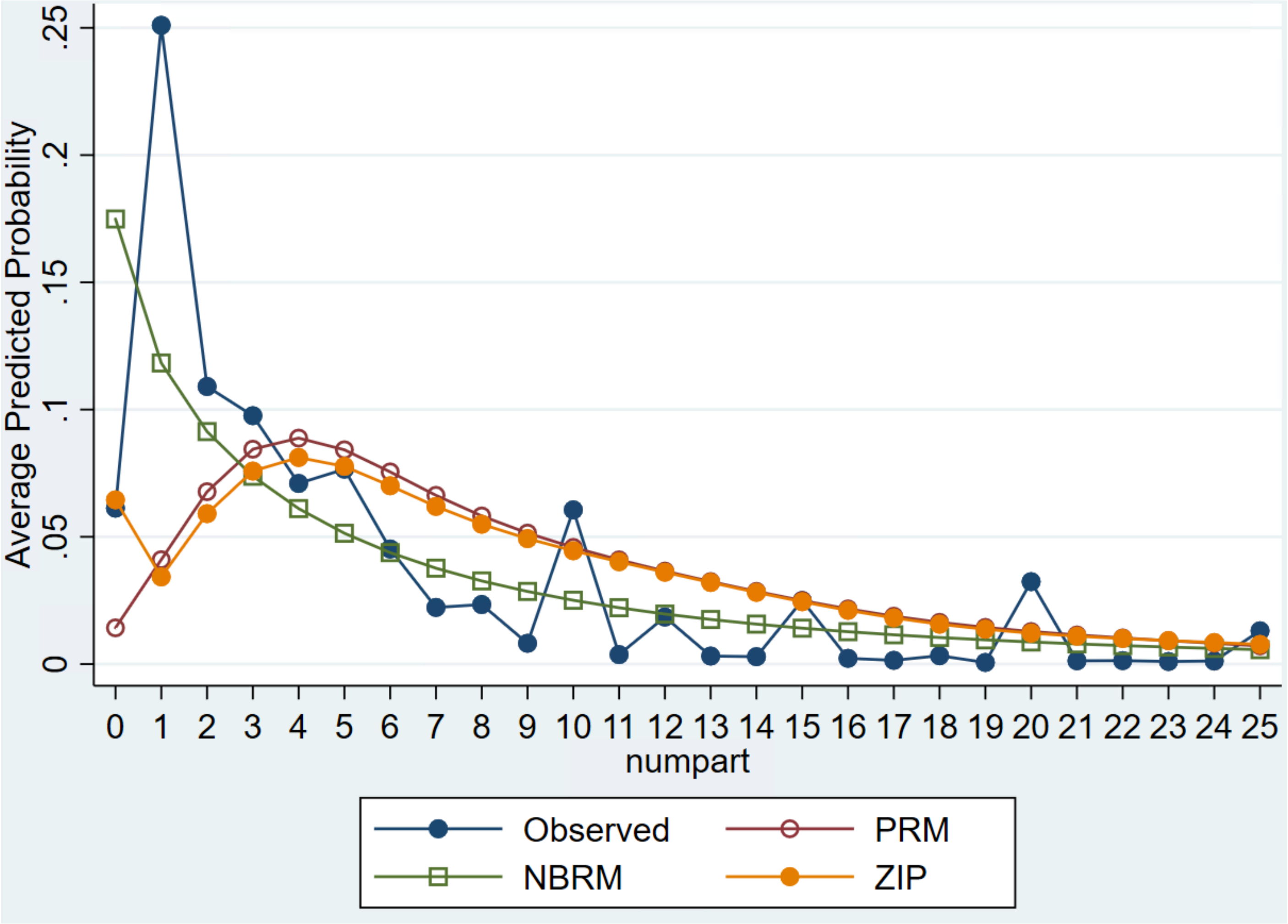


Negative binomial regression model (NBRM)



- Note how small the confidence intervals are for the PRM
- The PRM substantially underestimates the number of 0 partners
  - and does so with extra precision
- However, the probability of zero looks extremely high for the NBRM

Comparing models: informal



- Insights?
- Hmmm, maybe moving from 0-1 is a unique process?
  - Also moving from 1 to 2?
- Is 9-10, and 10-11 also unique?
  - Or is this clumping?
  - Same with 19-20 and 20-21?

Other techniques

- Zero-truncated count models
  - When sample excludes those who did not experience event
    - tpoisson and tnbreg
- Hurdle regression model
  - When moving from 0-1 is different than moving to others
- Zero-inflated count models
  - When lots of zeros
    - Zip and zinb
      - Zinb may be overkill

Exposure time

- One major issue with modeling count outcomes is exposure time may differ
- For example, does it make sense to control for age when modeling number of sexual partners?
- What about education?
  - Younger folks not have chance to complete
- How about marriage?
  - Younger folks not yet married
- Some techniques to deal with this, also event history modeling
  - Not getting into this, just be aware

Next Monday we will…

- Practice mlogit and count models
- Workshop projects
- Prep for presentations