Analyzing the household: adjusting and smoothing

Iván Mejía-Guevara
imejia@demog.berkeley.edu

Postdoctoral Scholar
CEDA
University of California, Berkeley

East-West Center Summer Seminar on Population,
June 10, 2010

Outline

1. Smoothing
2. Friedman's Super Smoother (supsmu)
3. Variance estimation for age profiles
4. Age profile confidence intervals

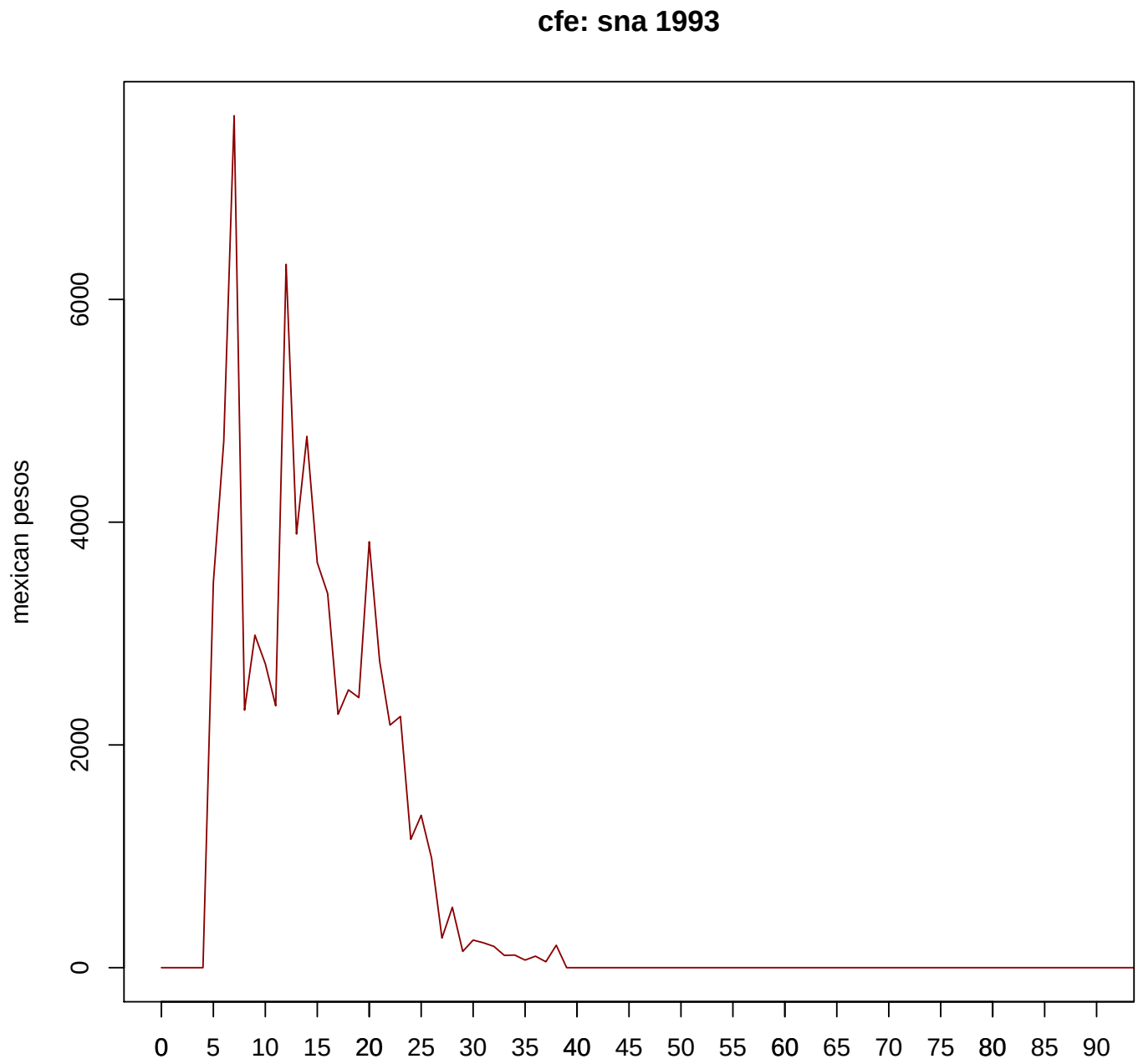
1. Smoothing

1. Smoothing

The per capita age profiles are noisy, particularly at ages with relatively few observations, and except as noted below should be smoothed. The following guidelines should be followed (NTA Manual):

- **The per capita education profile should not be smoothed.**

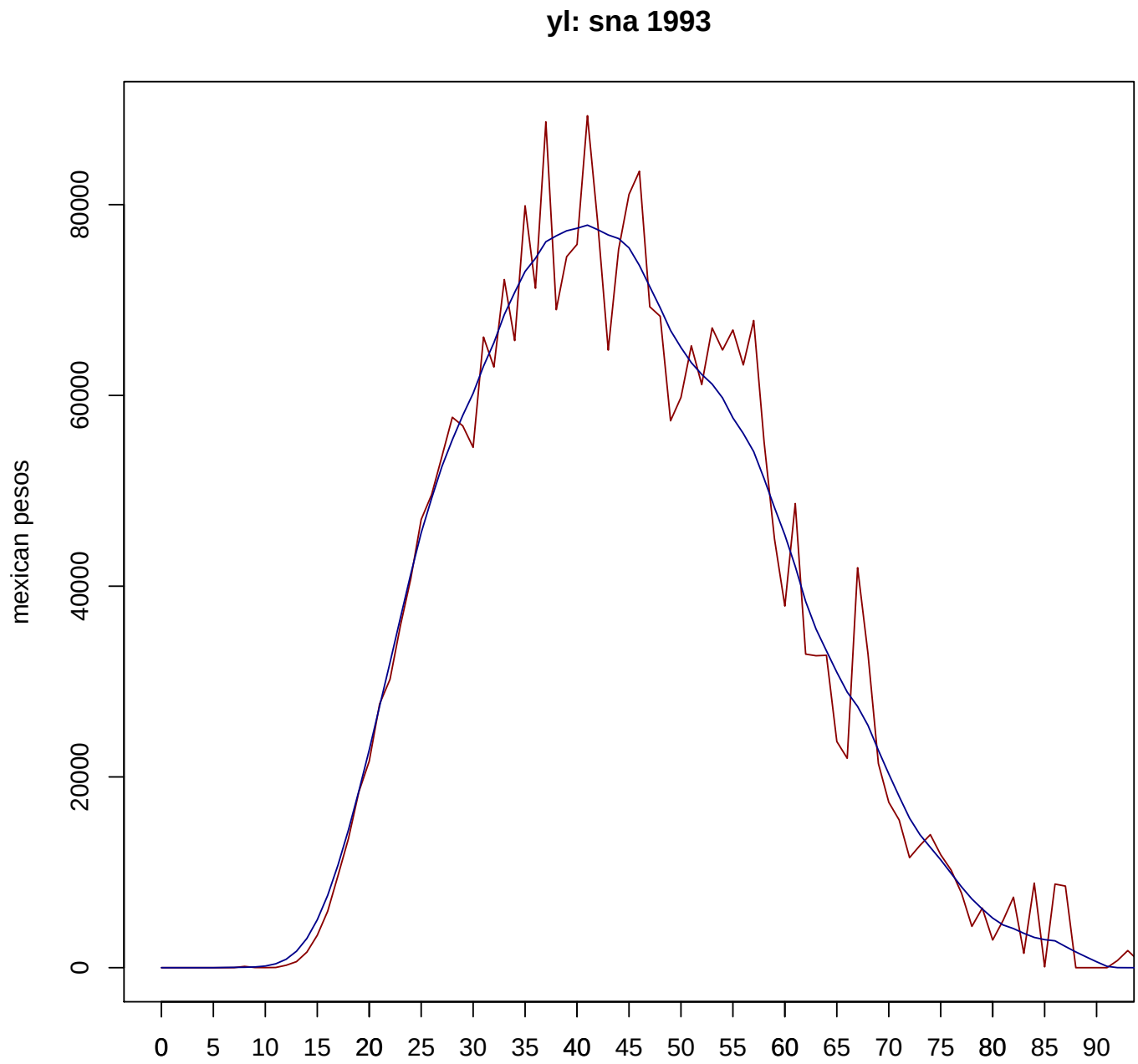
1. Smoothing: education age profile (Mexico 2004)



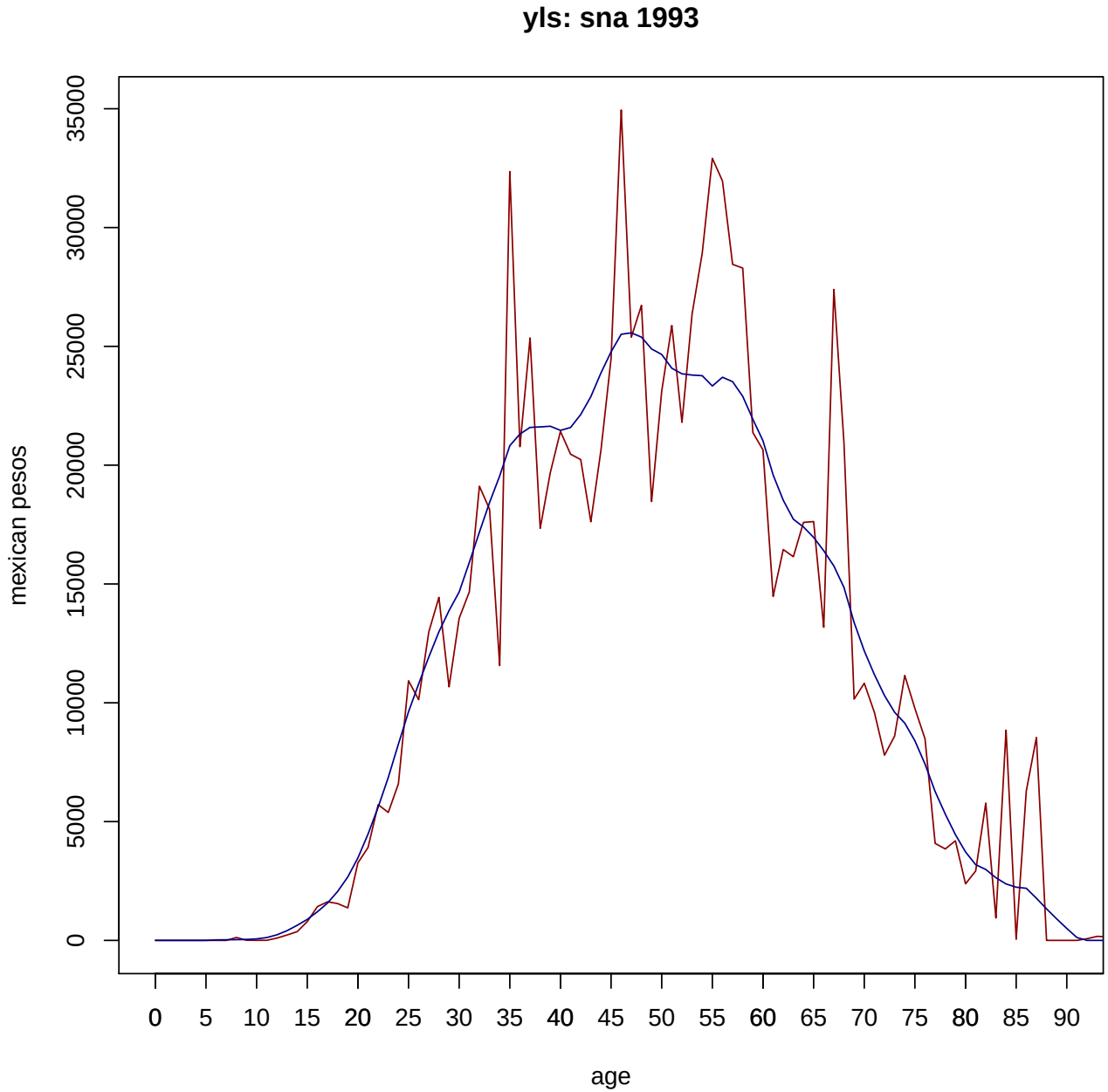
1. Smoothing...

- **Basic components should be smoothed, but not aggregations. For example, earnings and unincorporated income profiles should be smoothed, but the sum of the two should not be smoothed.**

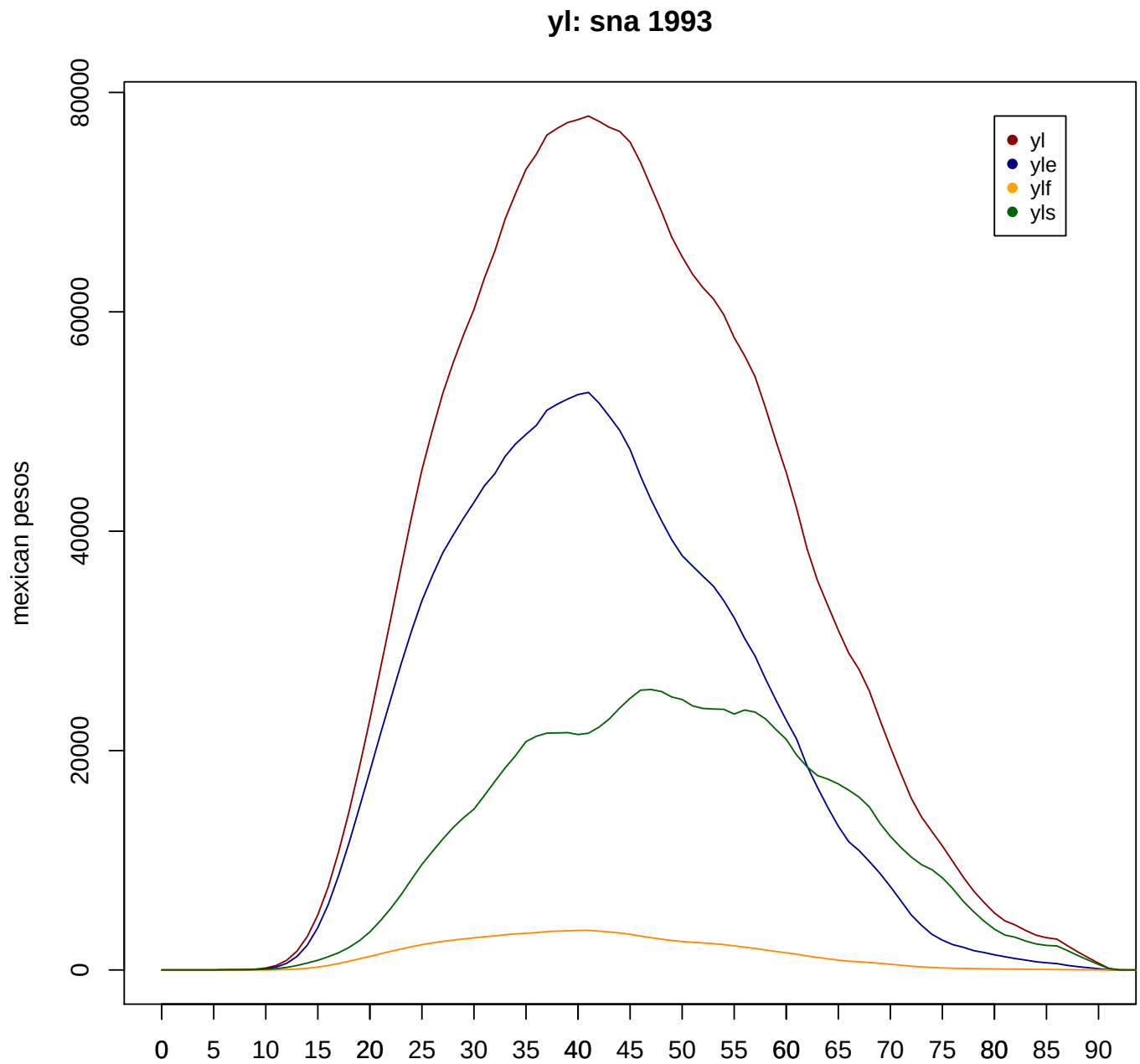
1. Smoothing: earnings (Mexico 2004)



1. Smoothing: unincorporated income (Mexico 2004)



1. Smoothing: labor income (Mexico 2004)



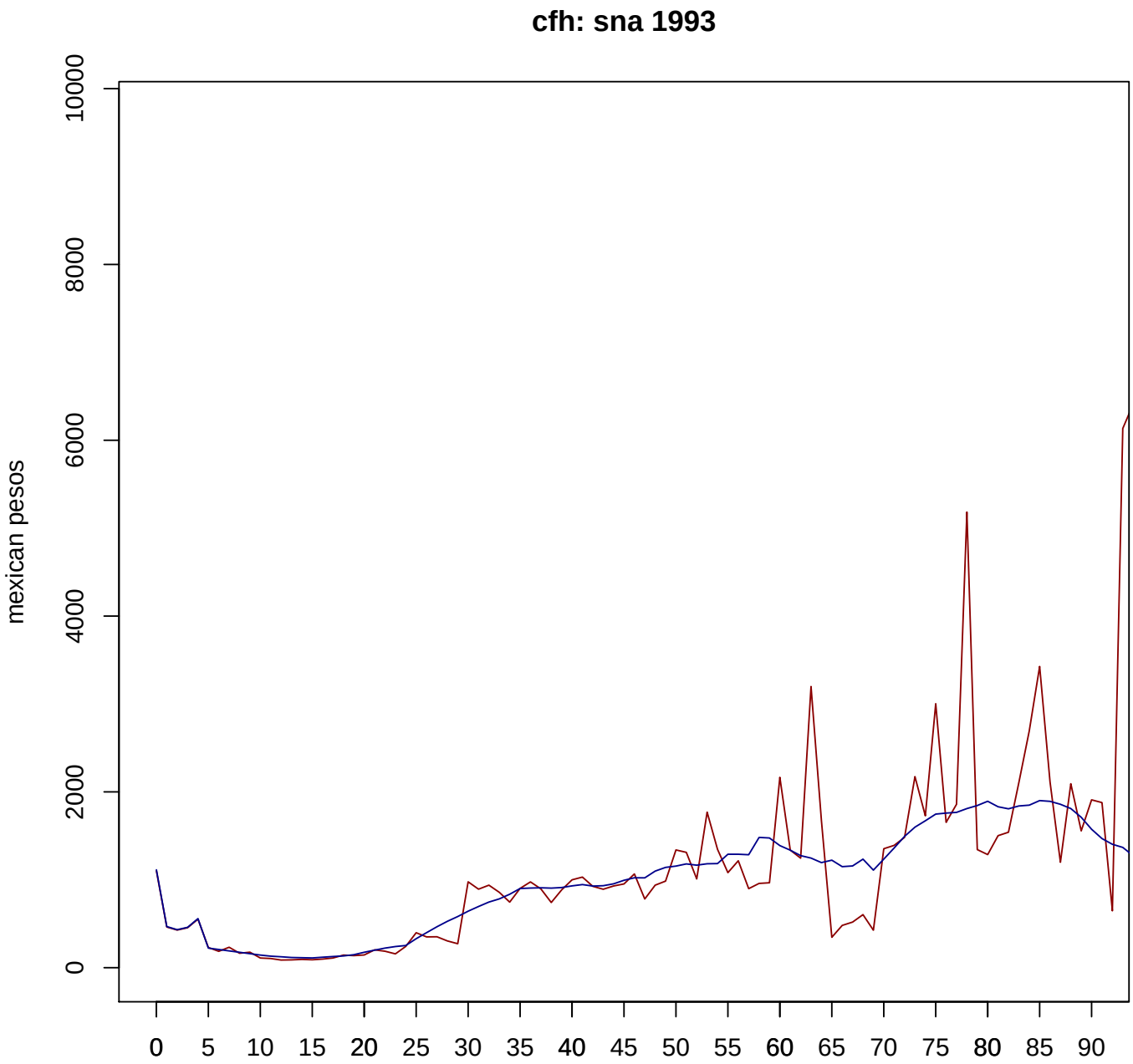
1. Smoothing...

- **The objective is to reduce sampling variance but not eliminate what may be "real" features of the data. For example, Public health spending may increase dramatically when individuals reach an age threshold, e.g., 65. This kind of feature of the data should not be smoothed away.**

1. Smoothing...

- **Due to unusual high health consumption by newborns, we tend not to smooth health consumption by age 0. This could be done by including estimated unsmoothed health consumption by newborns to the age profile of smoothed private health consumption by other age groups.**

1. Smoothing: private health consumption (Mexico 2004)



1. Smoothing...

- **Only adults (usually ages 15 and older) receive income, pay income taxes and make familial transfer outflows. Thus, when we smooth these age profiles, we begin smoothing from the adults, excluding those younger age group who do not earn income.**

2. Friedman's Super Smoother (supsmu)

2. Friedman's Super Smoother (supsmu)

There are a couple of steps to smoothing the per capita profile:

1. Create a spreadsheet, which contains unsmoothed age profile and the number of observations for each age.
2. Use Friedman's SuperSmoother (supsmu function in R) to smooth the per capita profile incorporating the number of observations.

The following is the R code to use the command "supsmu". Suppose "thyl.csv" is the file name (tab delimited excel file format), yl the unsmoothed variable name, and sample is the number of observations for each age in the data. The R programming code is:

$nta <- read.csv("thyl.csv", header = T)$ *Read in data. Work name is nta*

$test <- supsmu(nta\$age, nta\$yl, nta\$sample)$ *Smooth data. Work name is test*

$write.csv(test, "smoothed\_yl.csv")$ *Write out data using name "smoothed_yl"*

2. supsmu: R code

- supsmu(x, y, wt, span = "cv", periodic = FALSE, bass = 0)
  -Arguments:
  x: x values for smoothing
  y: y values for smoothing
  wt: case weights, by default all equal
  span: the fraction of the observations in the span of the running lines smoother, or "cv" to choose this by leave-one-out cross-validation.
  periodic: if TRUE, the x values are assumed to be in [0, 1] and of period 1.
  bass: controls the smoothness of the fitted curve. Values of up to 10 indicate increasing smoothness.

2. Alternative to supsmu...

The alternative smoothing method is "lowess" smoothing. The procedure is found to be unreliable because it does not incorporate sample weights. We recommend that it not be used. (see the NTA Manual for more detail about it if you feel more comfortable using the Stata rather than the R program, and would prefer to use the lowess smoothing method).

3. Variance estimation for age profiles

3. Variance estimation for age profiles

- Age profile estimation in NTA:

$$\bar{y}_a = \frac{y}{w} = \frac{\sum_a^{n_a} w_{ia} y_{ia}}{\sum_a^{n_a} w_{ia}} \tag{1}$$

where $\bar{y}_a$ is the mean value of variable y (e.g. education) for individual aged a, w_{ia} is the sampling weight for the individual i age a, n_a is the sampling size of individuals in the age group a.

- Survey design:
  a) Simple Random Sampling (SRS)
  b) Complex design survey (CDS): estratified multi-stage cluster
  * Survey variables in CDS: 1) strata, 2) primary sampling units, 3) weights

3. Variance estimation for age profiles

- Variance estimation for Simple Random Samples (SRS):

$$Var\left(\frac{y}{w}\right) = \frac{s^2}{n}$$

- Variance estimation for CDS: $Var\left(\frac{y}{w}\right) \neq \frac{Var(y)}{Var(w)}$

  * Taylor series linearization method (TSL): let's define $r = \frac{y}{w}$, then:

$$var(\bar{y}_a) = \frac{1}{w^2}[var(y) + r^2 \cdot var(w) - 2 \cdot r \cdot cov(y, w)] \quad (2)$$

where:

$var(y) = \sum_{h=1}^{H} \left(\frac{n_h}{n_h - 1}\right) \left[\sum_{\alpha=1}^{n_h} y_{h\alpha}^2 - \frac{y_h^2}{n_h}\right]$

$var(w) = \sum_{h=1}^{H} \left(\frac{n_h}{n_h - 1}\right) \left[\sum_{\alpha=1}^{n_h} w_{h\alpha}^2 - \frac{w_h^2}{n_h}\right]$

$cov(y, w) = \sum_{h=1}^{H} \left(\frac{n_h}{n_h - 1}\right) \left[\sum_{\alpha=1}^{n_h} y_{h\alpha} w_{h\alpha} - \frac{y_h w_h}{n_h}\right]$

where:

H : number of estrata

n_h: number of individuals in stratum h

3. Stata code for variance estimation

- SRS:
  mean yl [pw=factor], over(age)
  where:
  yl: NTA variable, i.e. labor income
  factor: sampling weight
  age: 'age' survey variable
- CDS:
  svyset psu [pw=factor], strata (stratum)
  svy: mean yl, over(age)
  where:
  psu: primary sampling unit survey variable
  stratum: strata survey variable

3. Stata output

yle

Over	Mean	Std. Err.	[95% Conf.	Interval]
0	0	0	.	.
1	0	0	.	.
2	0	0	.	.
3	0	0	.	.
...				
30	7133.63	256.329	6631.23	7636.03
31	8576.72	419.072	7755.34	9398.09
32	7959.72	347.977	7277.69	8641.75
33	9022.32	395.903	8246.35	9798.28
34	8751.68	374.232	8018.19	9485.17
35	8395.42	421.098	7570.07	9220.77
...				
86	490.310	463.267	-417.69	1398.31
87	9.375	9.375	-8.9999	27.7499
...				

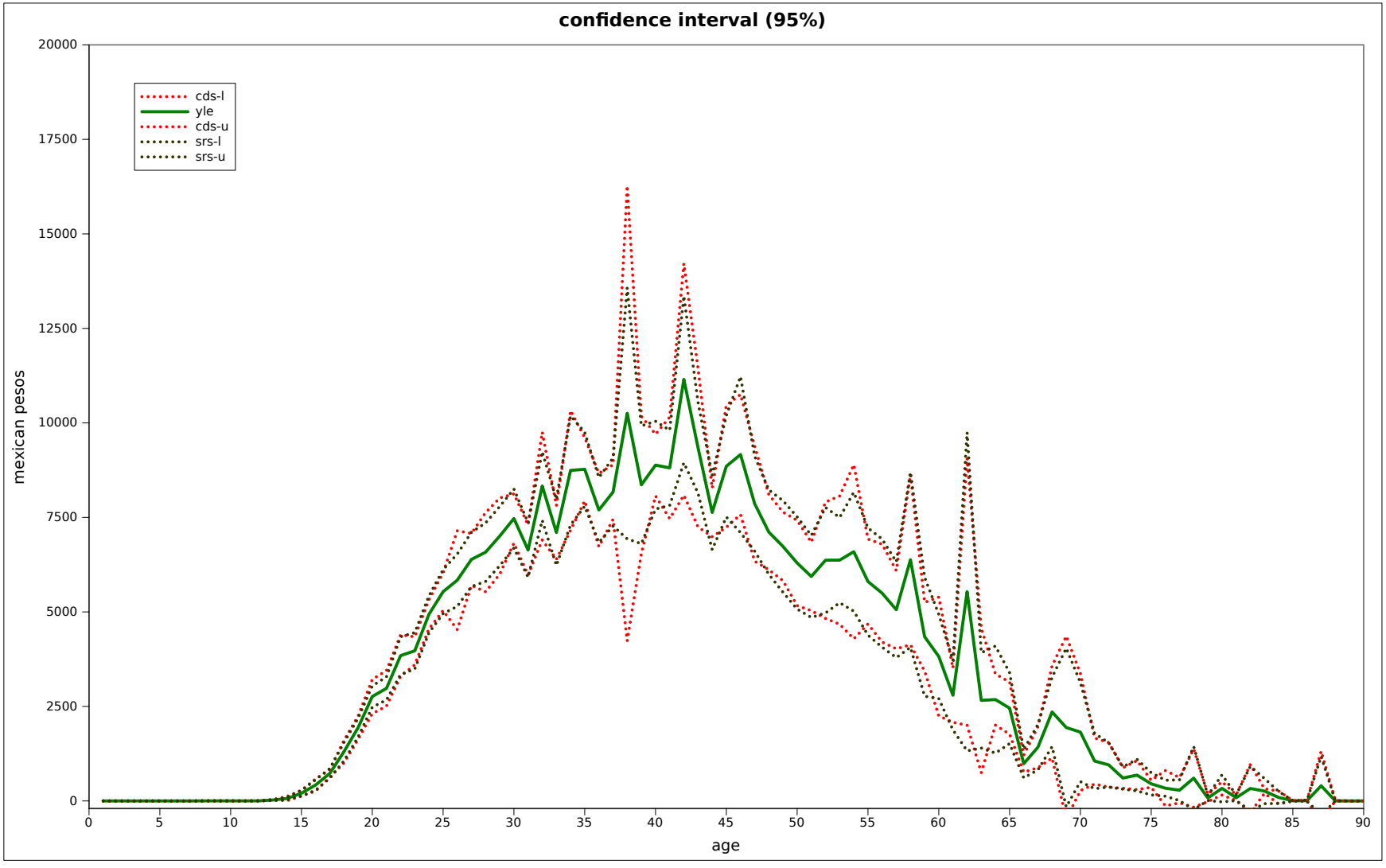
4. Confidence intervals

4. Stata output

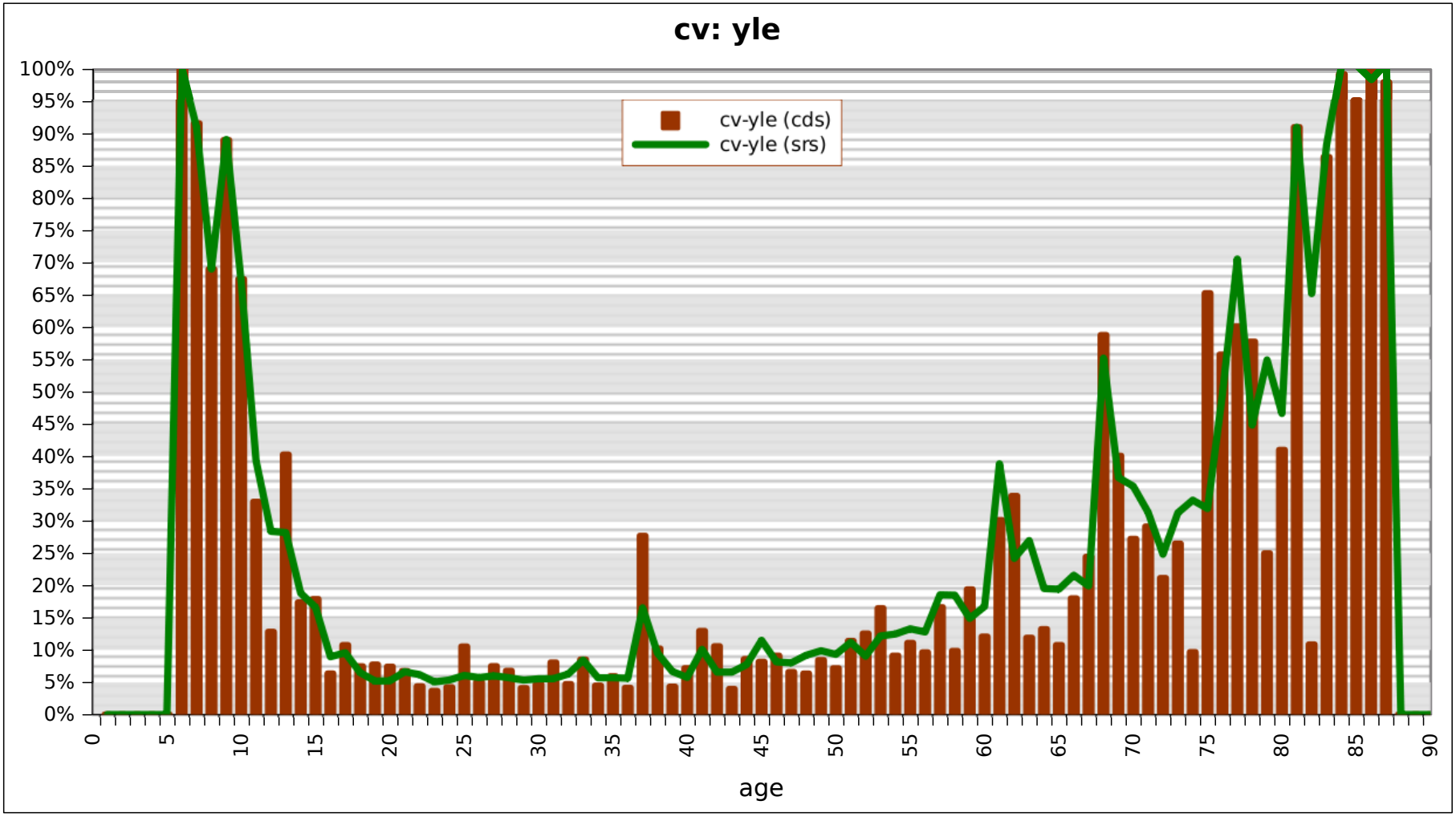
yle Over	Mean	Std. Err.	[95% Conf.	Interval]
...				
30	7133.63	256.329	6631.23	7636.03
31	8576.72	419.072	7755.34	9398.09

Mean: $\bar{y}_a$
Std. Err.: $se(\bar{y}_a)$
Conf. Interval: $\bar{y}_a \ +/- \ t_{df} * se(\bar{y}_a)$

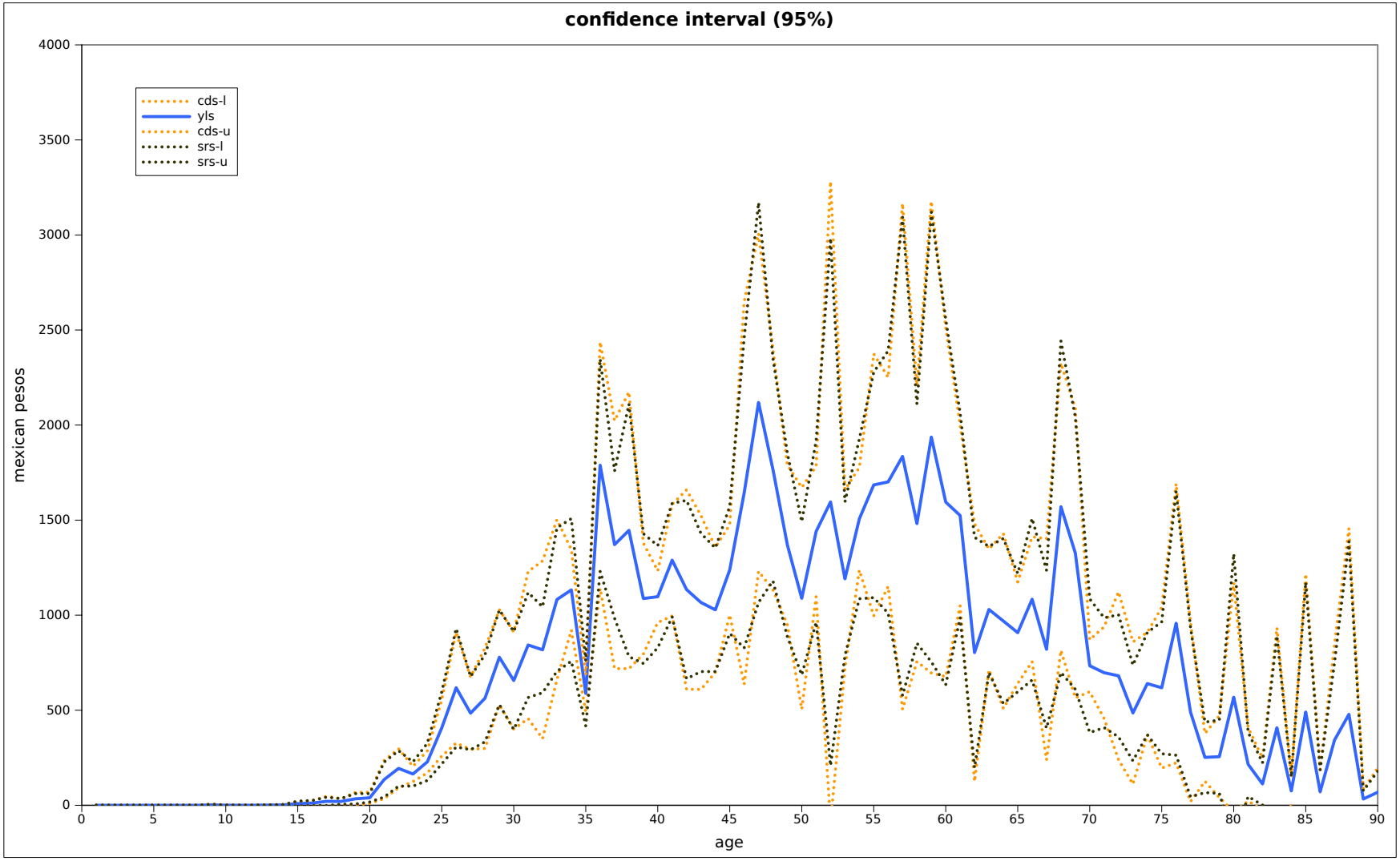
4. Example: YL: earnings (yle)



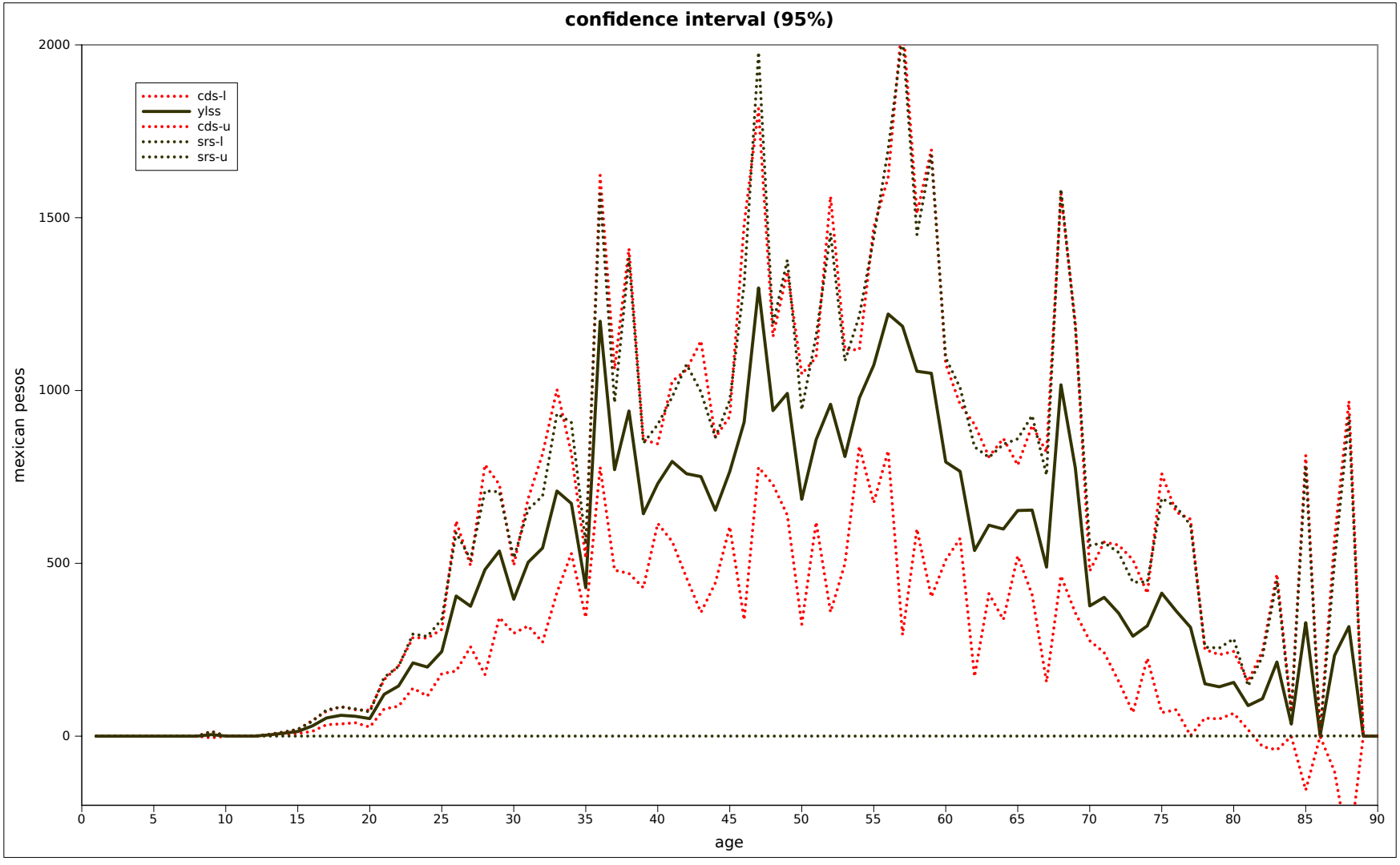
4. Coefficient of variation: $ce(\bar{y}_a) = se(\bar{y}_a)/\bar{y}_a$



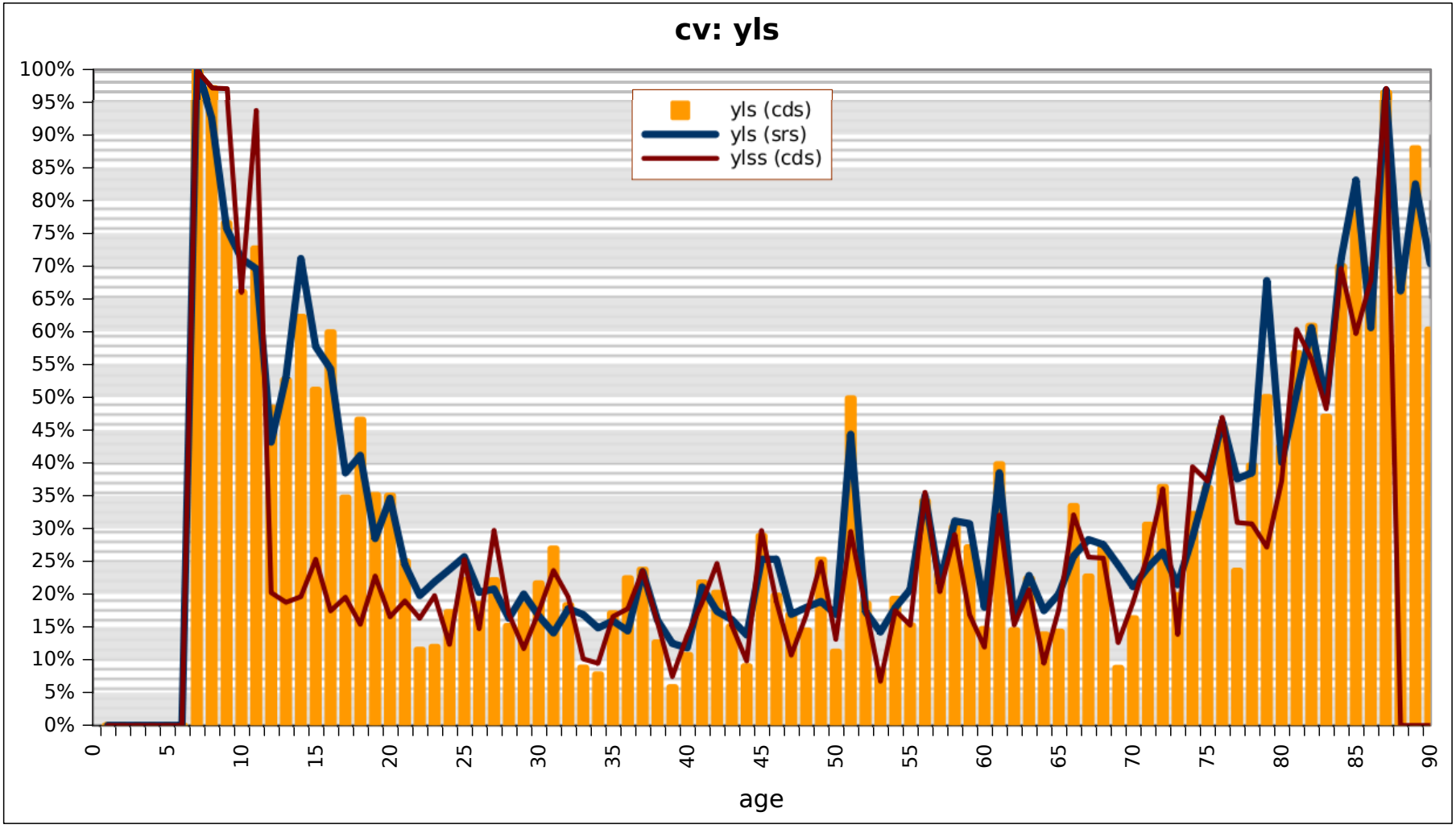
4. Example: YL: entrepreneurial income (yls)



4. YL: imputed self-employed income (ylss)



4. YL: coefficient of variation (yls)



4. Confidence intervals for smoothed profiles: supsmu

- $(x_1, y_1)...(x_n, y_n)$:

$$y_i = s(x_i) + r_i, i = 1...n \tag{3}$$

- Smoothed value at point x_i:

$$s(x_i) = \frac{1}{J} \sum_{i-J/2}^{i+J/2} y_i$$

- Expected squared error at point x_i, under $E(r_i) = 0$, $Var(r_i) = \sigma^2$:

$$e^2(x_i \| J) = \left(f(x_i) - \frac{1}{J} \sum_{i-J/2}^{i+J/2} f(x_i) \right)^2 + \frac{1}{J}\sigma^2 \tag{4}$$

4. supsmu: NTA framework

- $(a, \bar{y}_a)...(a, \bar{y}_a)$:

$$\bar{y}_a = s(\bar{y}_a) + r_a, a = 0...\omega \tag{5}$$

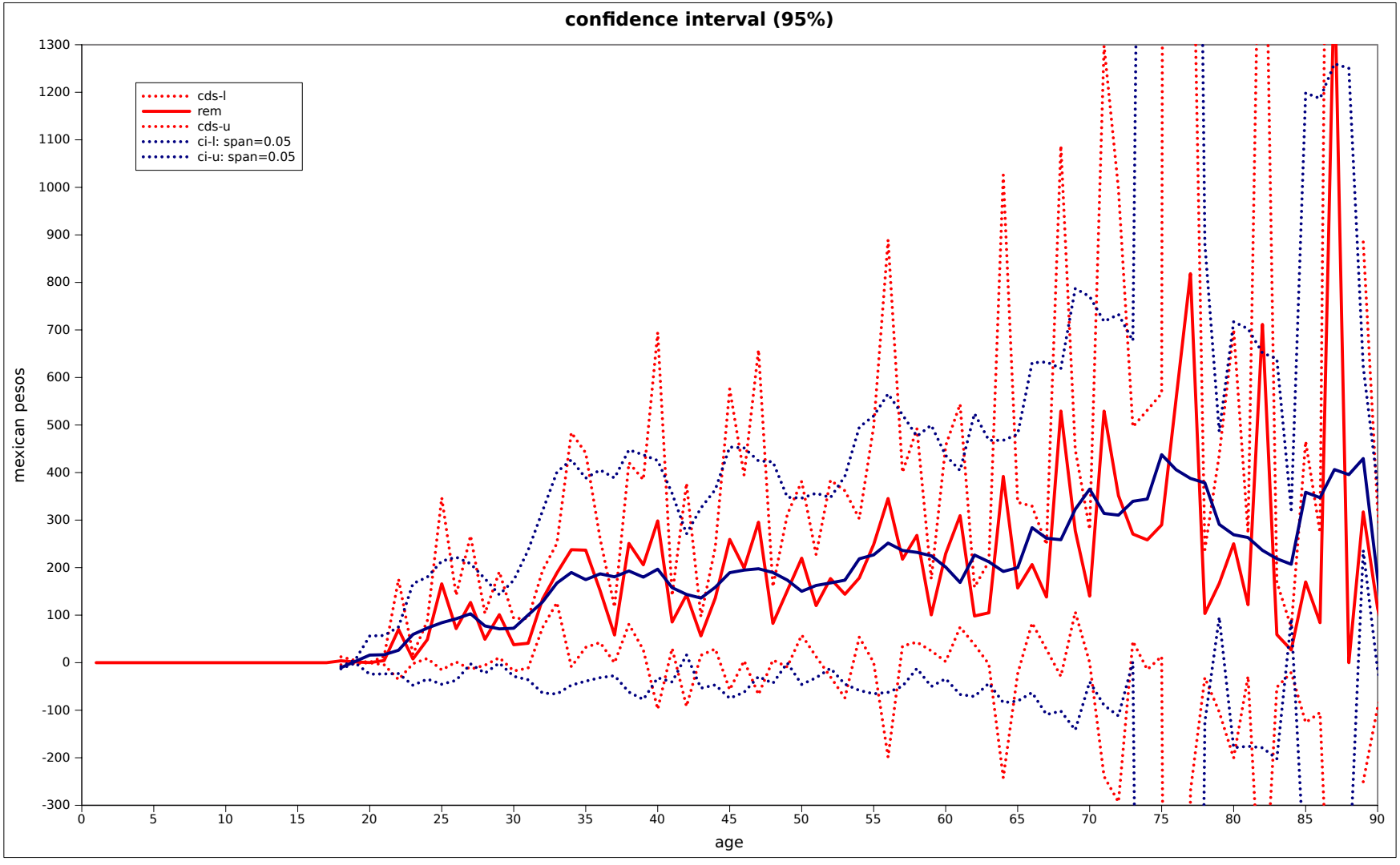
- Smoothed value at age a:

$$s(\bar{y}_a) = \frac{1}{J} \sum_{i-J/2}^{i+J/2} \bar{y}_a$$

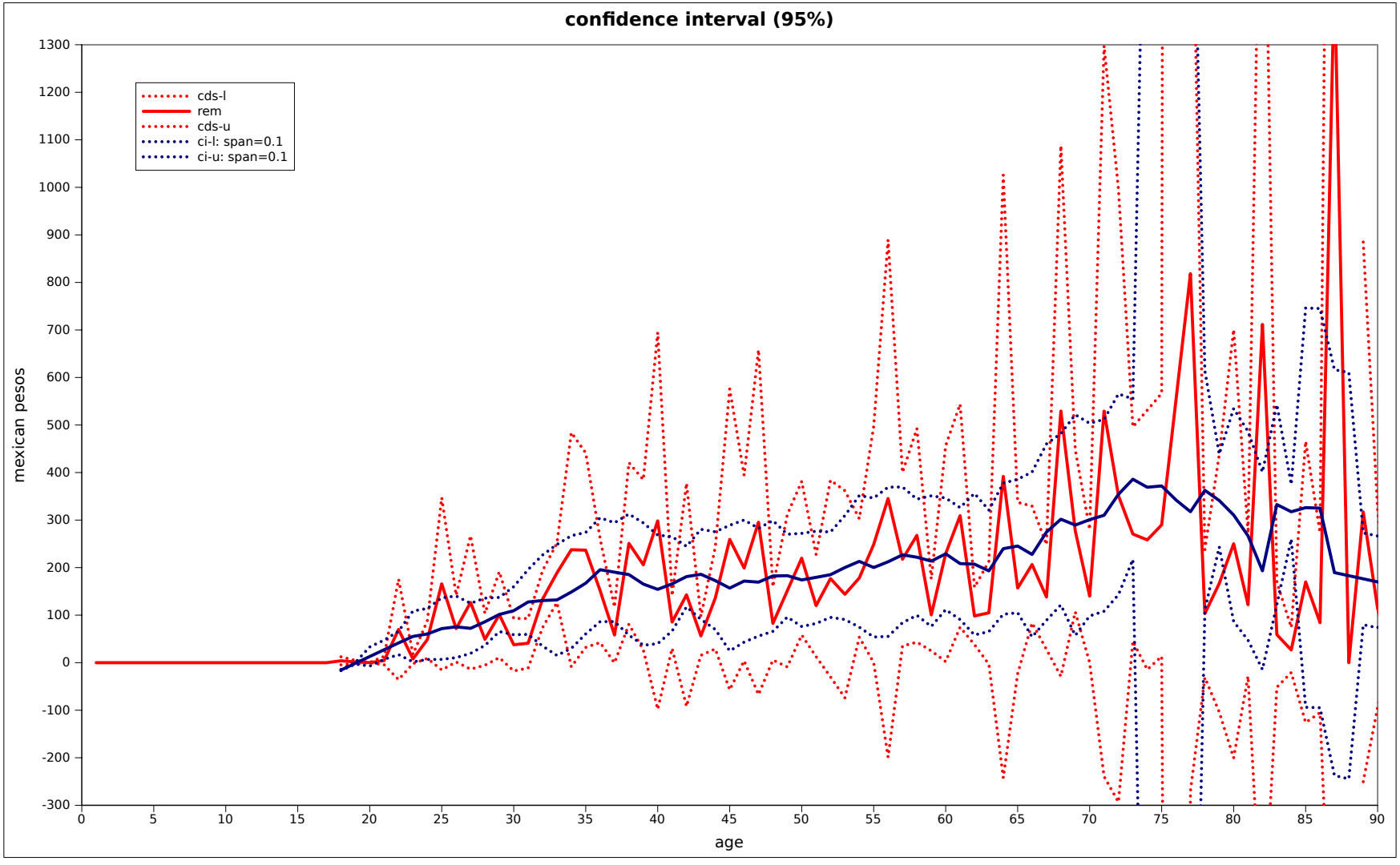
- Expected squared error at age a, under $E(r_a) = 0$, $Var(r_a) = \sigma_i^2 = Var_{cds}(\bar{y}_a)$:

$$e^2(a\|J) = \left(\bar{y}_a - \frac{1}{J} \sum_{a-J/2}^{a+J/2} \bar{y}_a \right)^2 + \frac{1}{J^2} \sum_{a-J/2}^{a+J/2} Var_{cds}(\bar{y}_a) \tag{6}$$

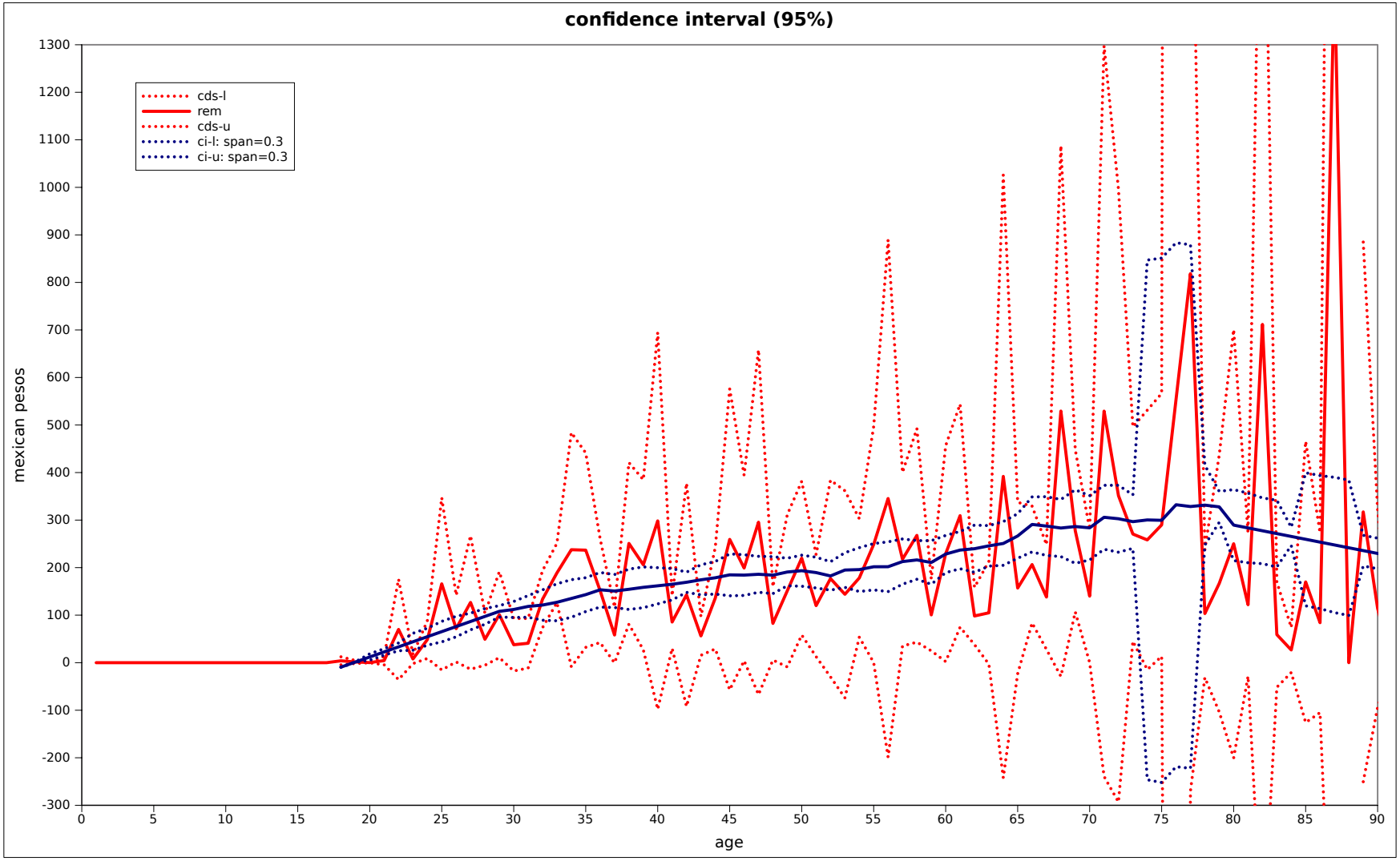
4. Example-supsmu: remittances (span=0.05)



4. Example-supsmu: remittances (span=0.1)



4. Example-supsmu: remittances (span=0.3)



4. Remarks

- This approach is valid only if you choose a single span selection.
- Do not use it if you select the cross validation option "cv" or if you specify the "bass" option in supsmu
- The software to do that is comming soon.....