SSSA
Small Sample Statistical Analysis

Day 3
Non-parametric Methods II

Dominik Duell (University of Essex)

September 28, 2016

1. Non-parametric regression slope estimator
   - Theil statistic
   - Theil-Sen median slope estimator
   - Confidence interval for the Theil statistic
2. Life and survival analysis
   - Kaplan-Meier estimator
   - Kolmogorov confidence bands for distribution functions
3. Local regression, density estimation, and other smoothing methods

Theil statistic: Basics

- Detect stochastic relationship between two variables – aka regression line
- Distribution free test of $H_0 : \beta = \beta_0$
- Assumptions
  - $Y_i = \alpha + \beta x_i + e_i$ for $i = 1, \ldots, n$
  - errors are a random sample from a continuous population that has median 0

Theil statistic: Procedure

- Compute n differences: $D_i = Y_i - \beta_0 x_i$
- Theil statistic $C = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} c(D_j - D_i)$

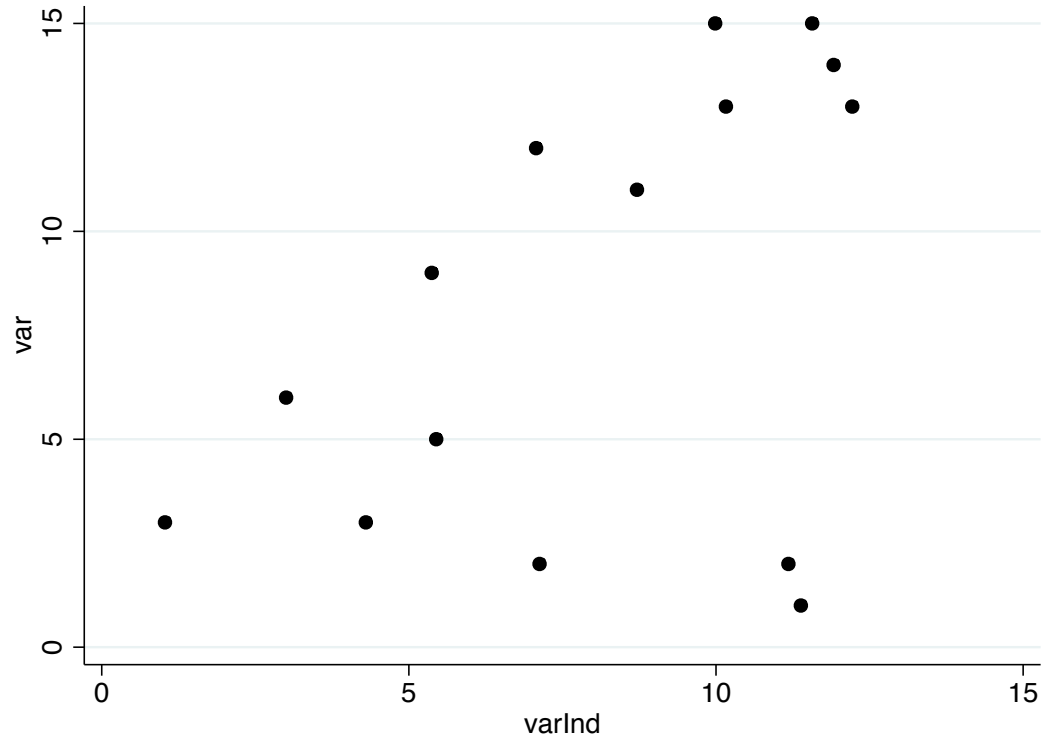
  where $c(d) = -1$ if D_j

$$c(d) = \begin{cases} -1 \text{ if } d < 0 \\ 0 \text{ if } d = 0 \\ 1 \text{ if } d > 0 \end{cases}$$

  with $1 \leq i < j \geq n$

- C will be large when many $D_j > D_i$
- Reject H_0 if $C \leq -k_\alpha$ for lower-tail test, $|C| \geq k_{\alpha/2}$ for two-sided test

Theil statistic: Example



Theil statistic: Example

```
.  reg var varInd;

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  1,    13) =    3.60
       Model |  85.2031241     1  85.2031241           Prob > F      =  0.0802
    Residual |  307.730209    13  23.6715546           R-squared     =  0.2168
-------------+------------------------------           Adj R-squared =  0.1566
       Total |  392.933333    14  28.0666667           Root MSE      =  4.8653

------------------------------------------------------------------------------
         var |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      varInd |   .6902718   .3638361     1.90   0.080    -.0957483    1.476292
       _cons |   2.722711   3.180751     0.86   0.408    -4.148884    9.594306
------------------------------------------------------------------------------
```

Theil statistic: Example

Say, $\beta_0 = 0$ then $D_i = Y_i$

	var	i	j	D	cD
1.	3	1	2	2	1
2.	5	1	3	0	.
3.	3	1	4	12	1
4.	15	1	5	9	1
5.	12	1	6	3	1
6.	6	1	7	11	1
7.	14	1	8	-1	-1
8.	2	1	9	6	1
9.	9	1	10	8	1
10.	11	1	11	12	1
11.	15	1	12	-2	-1
12.	1	1	13	10	1
13.	13	1	14	10	1
14.	13	1	15	-1	-1
15.	2	2	3	-2	-1
16.	.	2	4	10	1
17.	.	2	5	7	1
18.	.	2	6	1	1

Theil statistic: Example

- $C = \sum_{i=1}^{14} \sum_{j=i+1}^{15} c(D_j - D_i) = 52$
- We reject $H_0 : \beta_0 = 0$

	n						
x	4	5	8	9	12	13	16
0	.625	.592	.548	.540	.527	.524	.518
2	.375	.408	.452	.460	.473	.476	.482
4	.167	.242	.360	.381	.420	.429	.447
6	.042	.117	.274	.306	.369	.383	.412
8		.042	.199	.238	.319	.338	.378
10		.008	.138	.179	.273	.295	.345
12			.089	.130	.230	.255	.313
14			.054	.090	.190	.218	.282
16			.031	.060	.155	.184	.253
18			.016	.038	.125	.153	.225
20			.007	.022	.098	.126	.199
22			.002	.012	.076	.102	.175
24			.001	.006	.058	.082	.153
26			.000	.003	.043	.064	.133
28				.001	.031	.050	.114
30				.000	.022	.038	.097
32					.016	.029	.083
34					.010	.021	.070
36					.007	.015	.058
38					.004	.011	.048
40					.003	.007	.039
42					.002	.005	.032
44					.001	.003	.026
46					.000	.002	.021
48						.001	.016
50						.001	.013
52						.000	.010

Source: Hollander/Wolfe (1999), p.724

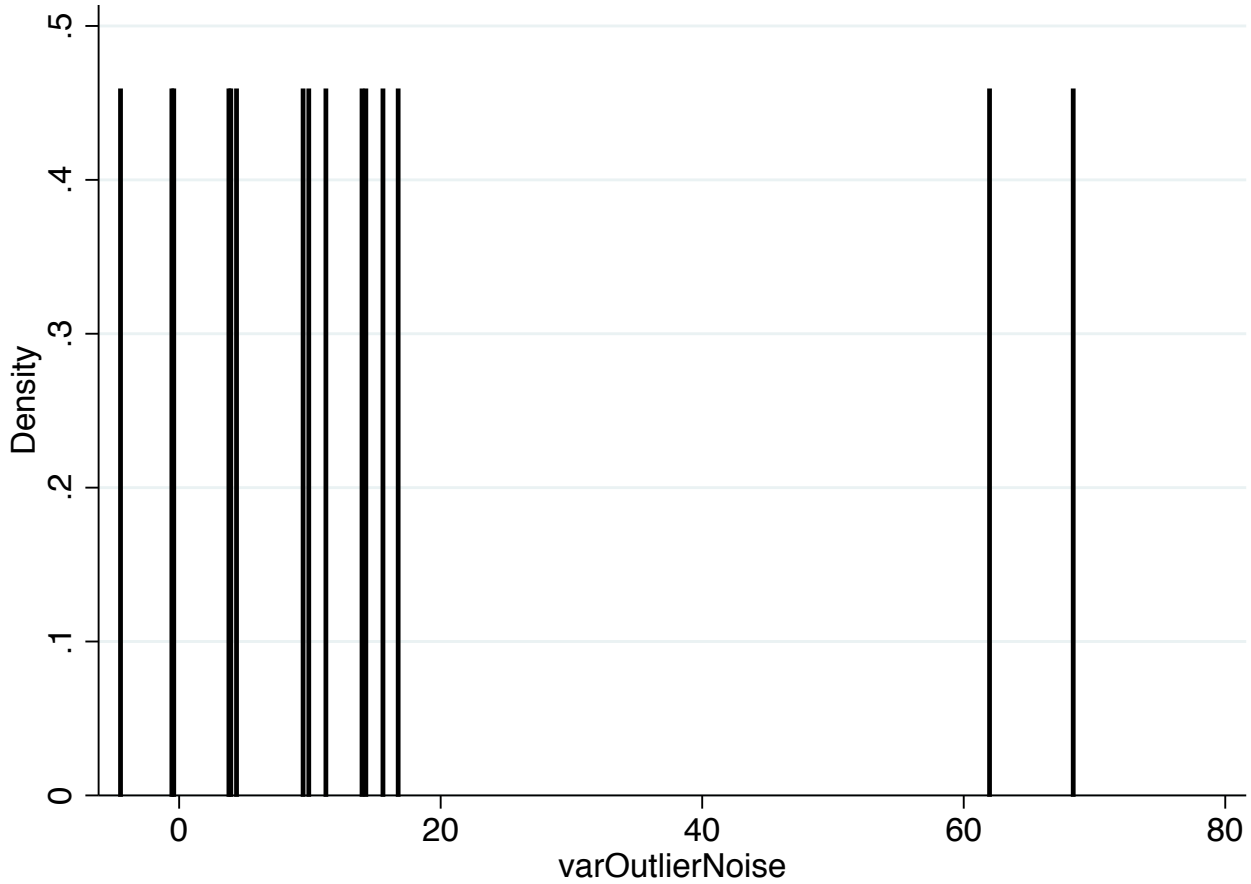
Theil-Sen: Basics

- Find point estimate for slope parameter of regression line
- Based on statistic introduced by Theil (1950) and extended by Sen (1968)
- Assumptions
  - $Y_i = \alpha + \beta x_i + e_i$ for $i = 1, \ldots, n$
  - errors are a random sample from a continuous population that has median 0
- Stratifying computation allows to control for confounding factors – moving into multivariate analysis

Theil-Sen: Procedure

- Compute $N = n(n-1)/2$ sample slope values $S_{ij} = (Y_j - Y_i)/(x_j - x_i)$ where $1 \leq i < j \leq n$
- $\hat{\beta} = median\{S_{ij}, 1 \leq i < j \leq n$
- Let $S^{(1)}, \ldots, S^{(N)}$ denote the ordered values of the S_{ij}
- $\hat{\beta} = S^{(N-1)/2+1}$ when N is odd
- $\hat{\beta} = [S^{N/2} + S^{N/2+1}]/2$ when N is even
- Sen extension defines statistic only from pairs with distinct x-coordinates

Theil-Sen: Example



Theil-Sen: Example

```
.  reg var varOutlierNoise, robust;

Linear regression                               Number of obs =      15
                                                F(  1,    13) =    4.68
                                                Prob > F      =  0.0498
                                                R-squared     =  0.1428
                                                Root MSE      =  5.0902

------------------------------------------------------------------------------
             |               Robust
         var |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
varOutlierNoise |   .0940331   .0434716     2.16   0.050     .0001184    .1879478
       _cons |   6.837187   1.770414     3.86   0.002      3.01244    10.66193
------------------------------------------------------------------------------
```

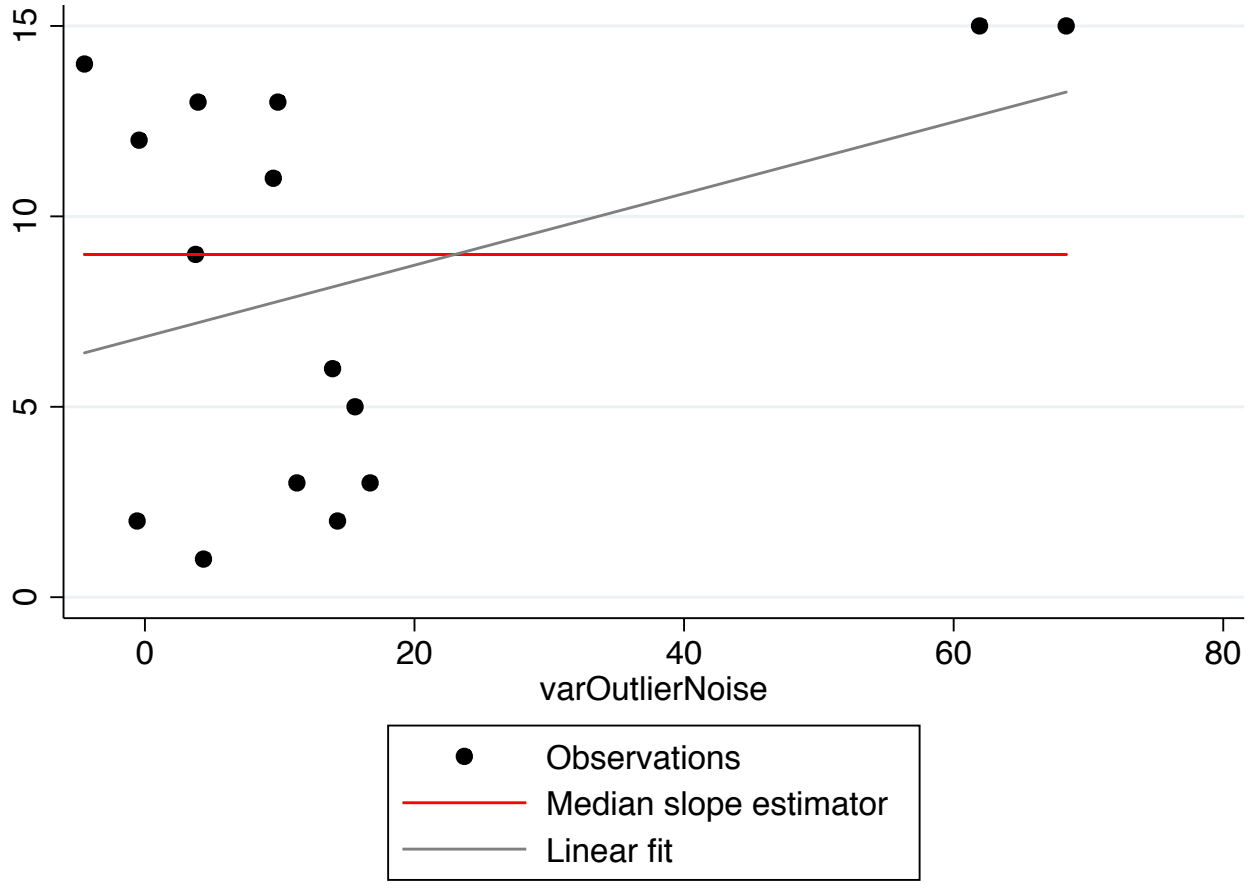
Theil-Sen: Example

```
. censlope var varOutlierNoise, ystar(residuals);
Outcome variable: var
Somers' D with variable: varOutlierNoise
Transformation: Untransformed
Valid observations: 15

Symmetric 95% CI
------------------------------------------------------------------------------
             |             Jackknife
varOutlier~e |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         var |   .0285714   .2612582     0.11   0.913    -.4834852    .5406281
------------------------------------------------------------------------------

95% CI(s) for percentile slope(s)
   Percent  Pctl_Slope     Minimum     Maximum
        50   4.777e-07  -.68684053   .21874658
```

Theil-Sen: Example



Small sample issues

- Works for asymmetric population distributions (of Y values given different X values)
- Very robust to outliers

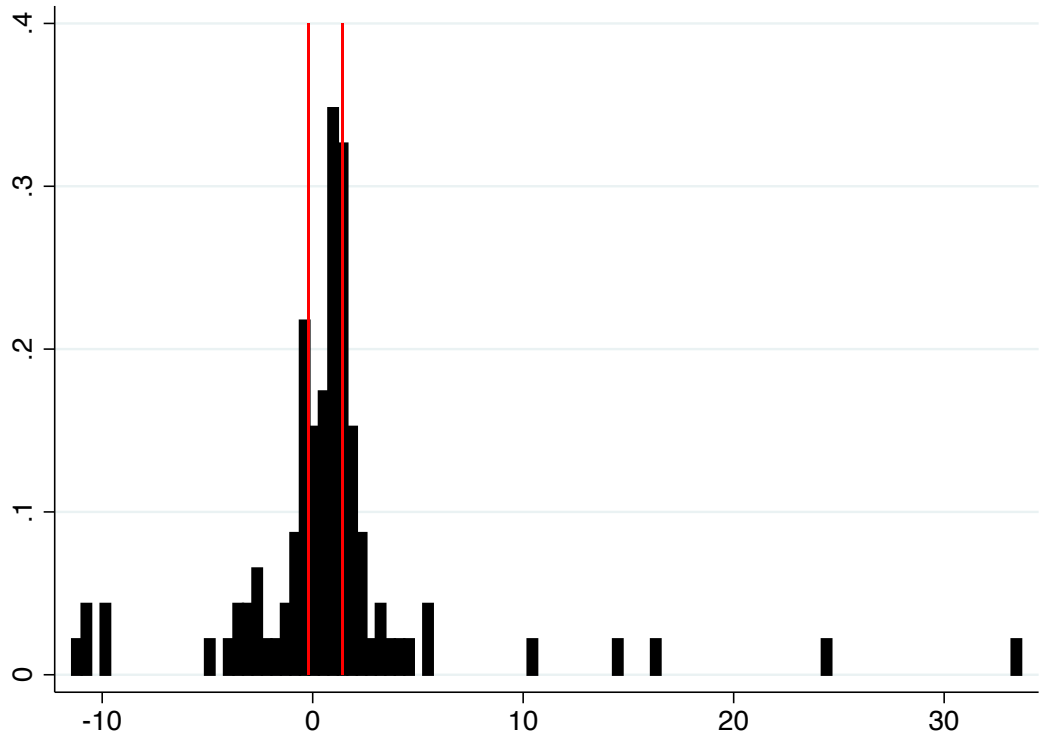
CI for Theil statistic: Procedure

- Compute $N = n(n-1)/2$ sample slope values $S_{ij} = (Y_j - Y_i)/(x_j - x_i)$ where $1 \leq i < j \leq n$
- Let $S^{(1)}, \ldots, S^{(N)}$ denote the ordered values of the S_{ij}
- Define $C_{\alpha/2} - 2$
  - $l = \frac{N - C_\alpha}{2}$
  - $u = \frac{N + C_\alpha}{2}$
    where $N = n(n-1)/2$
- Then,
  - $\beta_L = S^{(l)}$
  - $\beta_U = S^{(u+1)}$

CI for Theil statistic: Example

- $N = 15(14)/2$
- Say $\alpha = .05$, then $C_\alpha = 38$
- $l = (105 - 38)/2 = 33.5$
- $u = (105 + 38)/2 = 71.5$
- $\beta_l = S^{(33)}$
- $\beta_u = S^{(72)}$

CI for Theil statistic: Example



CI for Theil statistic: Example

	S
20.	-1.205393
21.	-.9693461
22.	-.8964559
23.	-.6737521
24.	-.6089196
25.	-.5966101
26.	-.5231877
27.	-.5186577
28.	-.4891254
29.	-.4092018
30.	-.3535548
31.	-.282382
32.	-.2350623
33.	-.1931933
34.	-.1639845
35.	-.1453534
36.	-.098527
37.	0
38.	0
39.	0
40.	0

	S
60.	1.028955
61.	1.041113
62.	1.050898
63.	1.094997
64.	1.13898
65.	1.181162
66.	1.262835
67.	1.265488
68.	1.288334
69.	1.30007
70.	1.339673
71.	1.380865
72.	1.381765
73.	1.391028
74.	1.40269
75.	1.425216
76.	1.444377
77.	1.474134
78.	1.489615
79.	1.521573
80.	1.633902

CI for Theil statistic: Example

- $N = 15(14)/2$
- Say $\alpha = .05$, then $C_\alpha = 38$
- $l = (105 - 38)/2 = 33.5$
- $u = (105 + 38)/2 = 71.5$
- $\beta_l = S^{(33)} = -.1632$
- $\beta_u = S^{(72)} = 1.391$

Kaplan-Meier estimator: Basics

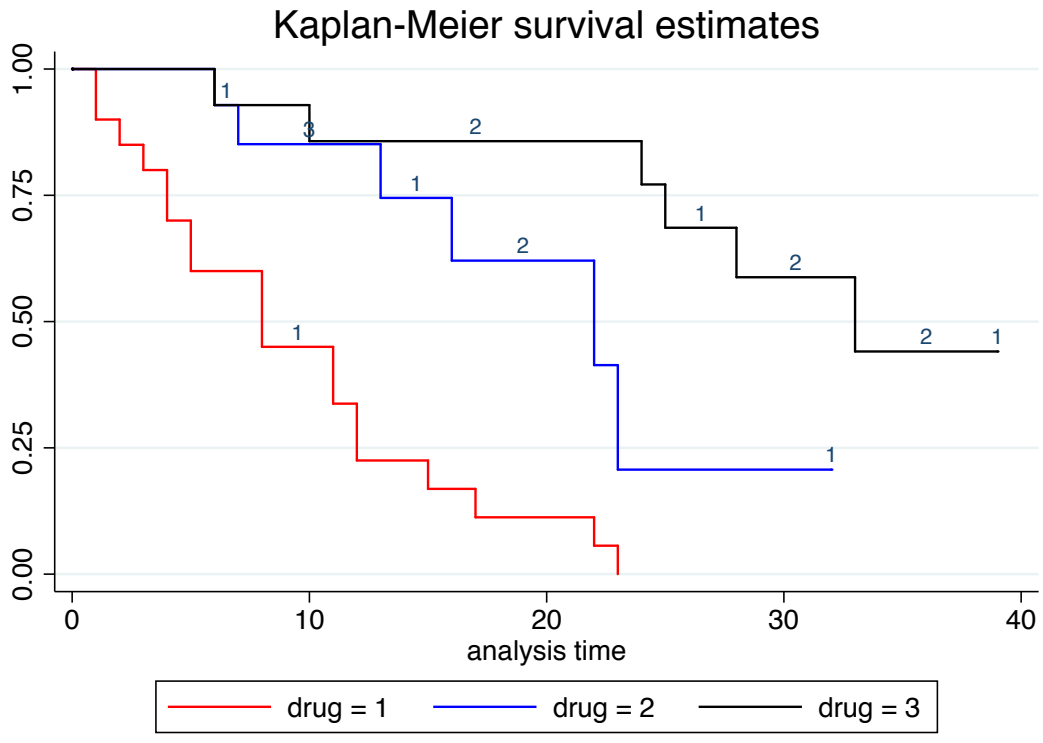
- Survival analysis with censored data
- Estimate of the probability of surviving in a given length of time
- Time is considered in many small intervals
- Time observations $T_1, \ldots, T_n$ with corresponding censoring observations $C_1, \ldots, C_n$
- Assumptions:
  - Time (censoring) observations are independent with continuous life (censoring) distribution F (G)
  - For each $i = 1 \ldots n$ $X_i = min\{T_i, C_i\}$ is observed
  - Time and censoring observations are independent
- Randomly right-censored model

Kaplan-Meier estimator: Procedure

- Let
  - $t_1, \ldots, t_k$ be the ordered failure times
  - n_i the number of cases at risk at t_i
  - d_i the number of failures at time t_i
- Then the Kaplan-Meier estimator of the survival function at time x is
  $\overline{F}_{KM} = \prod_{t_i \leq x} \left(1 - \frac{d_i}{n_i}\right)$
- and the estimator of the distribution function at time x is $1 - \overline{F}_{KM}(x)$
- Estimate of probability of death at t_i as number of death at t_i divided by the number of cases at risk n_i – take $1-$ estimator and get probability of surviving until t_i
- Kaplan-Meier (Product limit) estimator is den taking the product of conditional probabilities of not dying from zero until time of interest

Kaplan-Meier estimator: Example

- Take Stata dataset on cancer-treatment (`cancer.dta`)



Kaplan-Meier estimator: Small sample issues

- Assumption of constant survival probability within each interval probably not appropriate given that with smaller sample size the time intervals grow longer (intervals are defined by the survival times)
- If the number of cases at risk in or the number of cases survived to the beginning of that interval is small, variance estimate will underestimate actual variance

  Variance of Kaplan-Meier estimator of surviving past time x is

  $$\hat{var}(x) = [\overline{FM}_{KM}(x)]^2 \sum_{t_i \leq x} \frac{d_i}{n_i(n_i - d_i)}$$

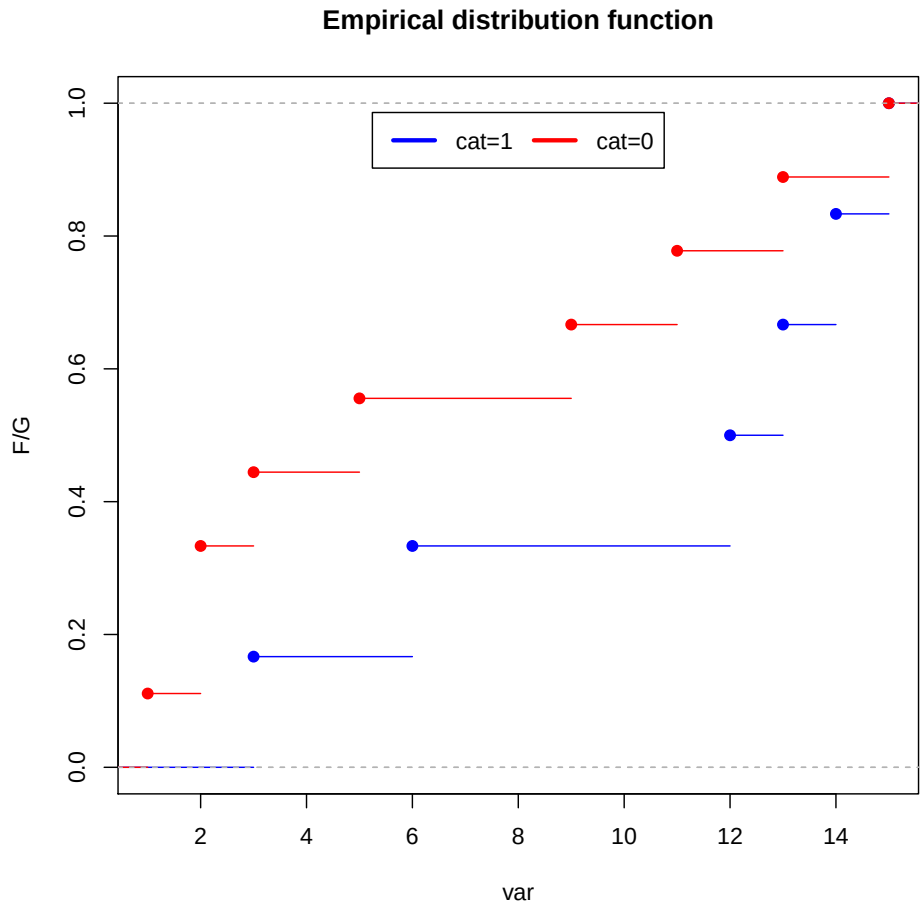
Kolmogorov confidence band: Basics

- Assumptions:
  - observations are random samples from underlying continous population and iid
  - distributed according to a continuous distribution function F

Kolmogorov confidence band: Procedure

- Task: Find random functions $l(x)$ and $u(x)$ such that $Prob[l(x) \leq F(x) \leq u(x), \forall x] \geq 1 - \alpha$
- Recall empirical distribution function $F_n(x)$

Kolmogorov confidence band: Procedure



Kolmogorov confidence band: Procedure

- Task: Find random functions $l(x)$ and $u(x)$ such that $Prob[l(x) \leq F(x) \leq u(x), \forall x] \geq 1 - \alpha$
- Recall empirical distribution function $F_n(x)$ and define Kolmogorov's statistic

  $D = sup_{-\infty<x<\infty}\{|F_n(x) - F(x)|\}$
- Chose d_α that satisfies

  $P_F\left(sup_{-\infty<x<\infty}\{|F_n(x) - F(x)|\} < d_\alpha\right) = 1 - \alpha$

Kolmogorov confidence band: Procedure

- Define

$$l(x) = \begin{cases} F_n(x) - d_\alpha \text{ if } F_n(x) - d_\alpha \geq 0 \\ 0 \qquad\qquad\quad \text{if } F_n(x) - d_\alpha < 0 \end{cases}$$

$$u(x) = \begin{cases} F_n(x) + d_\alpha \text{ if } F_n(x) + d_\alpha \leq 1 \\ 0 \qquad\qquad\quad \text{if } F_n(x) + d_\alpha > 1 \end{cases}$$

Kolmogorov confidence band: Example

- Recall example from Kolmogorov-Smirnov test

	var	F	G	D_FG	maxD_FG	M	N	J
1.	1	.1111111	0	.1111111	.4444444	9	6	8
2.	2	.3333333	0	.3333333	.4444444	9	6	8
3.	2	.3333333	0	.3333333	.4444444	9	6	8
4.	3	.4444444	.1666667	.2777778	.4444444	9	6	8
5.	3	.4444444	.1666667	.2777778	.4444444	9	6	8
6.	5	.5555556	.1666667	.3888889	.4444444	9	6	8
7.	6	.5555556	.3333333	.2222222	.4444444	9	6	8
8.	9	.6666667	.3333333	.3333333	.4444444	9	6	8
9.	11	.7777778	.3333333	.4444444	.4444444	9	6	8
10.	12	.7777778	.5	.2777778	.4444444	9	6	8
11.	13	.8888889	.6666667	.2222222	.4444444	9	6	8
12.	13	.8888889	.6666667	.2222222	.4444444	9	6	8
13.	14	.8888889	.8333333	.0555556	.4444444	9	6	8
14.	15	1	1	0	.4444444	9	6	8
15.	15	1	1	0	.4444444	9	6	8

Kolmogorov confidence band: Example

n	$1-\alpha$: .80	.90	.95	.98	.99
1	.900	.950	.975	.990	.995
2	.684	.776	.842	.900	.929
3	.565	.636	.708	.785	.829
4	.493	.565	.624	.689	.734
5	.447	.509	.563	.627	.669
6	.410	.468	.519	.577	.617
7	.381	.436	.483	.538	.576
8	.358	.410	.454	.507	.542
9	.339	.387	.430	.480	.513
10	.323	.369	.409	.457	.489
11	.308	.352	.391	.437	.468
12	.296	.338	.375	.419	.449
13	.285	.325	.361	.404	.432
14	.275	.314	.349	.390	.418
15	.266	.304	.338	.377	.404
16	.258	.295	.327	.366	.392
17	.250	.286	.318	.355	.381
18	.244	.279	.309	.346	.371
19	.237	.271	.301	.337	.361
20	.232	.265	.294	.329	.352
21	.226	.259	.287	.321	.344
22	.221	.253	.281	.314	.337
23	.216	.247	.275	.307	.330
24	.212	.242	.269	.301	.323
25	.208	.238	.264	.295	.317
26	.204	.233			

Source: Hollander/Wolfe (1999), p.740

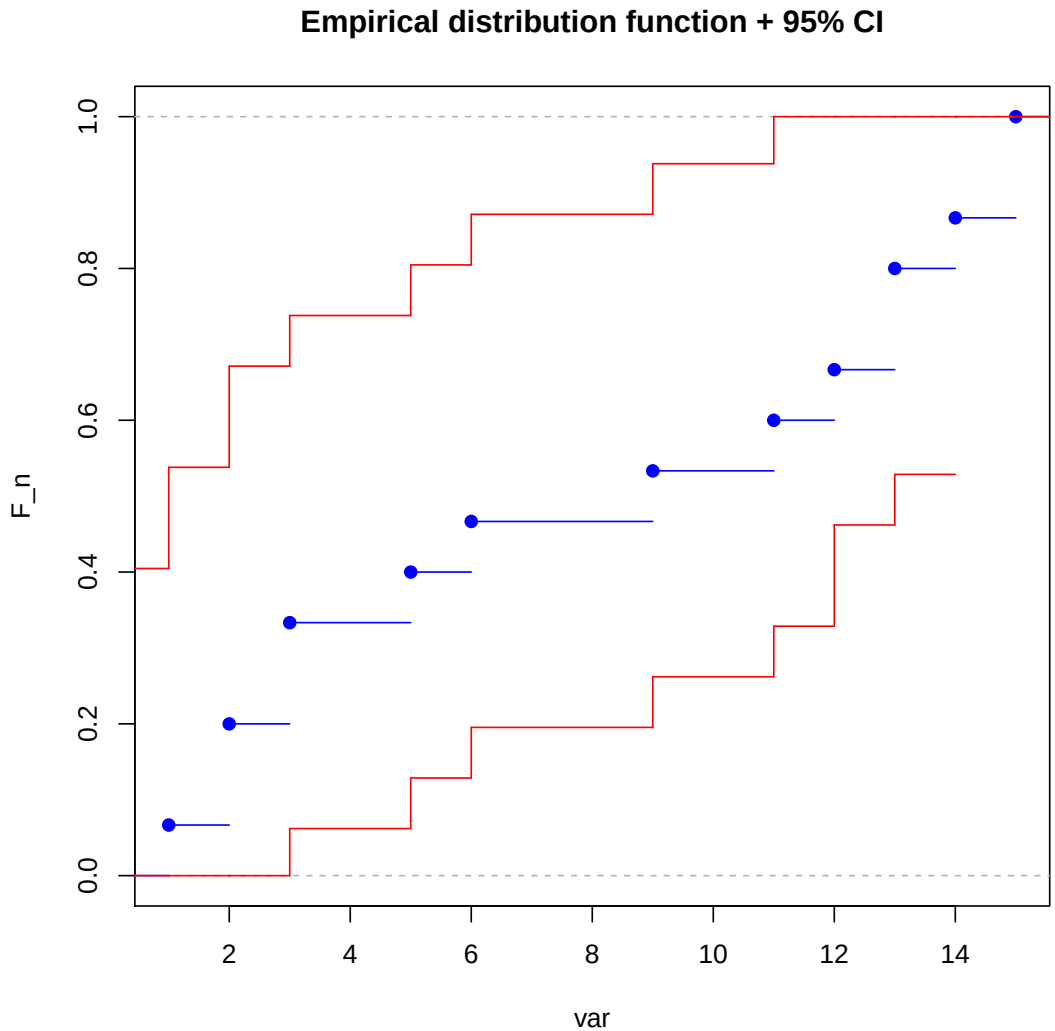
Kolmogorov confidence band: Example

- When n $= 15$, $d_{.05} = .338$

	var	F_n	l_x	u_x
1.	1	.0666667	0	.4046667
2.	2	.2	0	.538
3.	2	.2	0	.538
4.	3	.3333333	0	.6713333
5.	3	.3333333	0	.6713333
6.	5	.4	.062	.738
7.	6	.4666667	.1286667	.8046666
8.	9	.5333334	.1953334	.8713334
9.	11	.6	.262	.938
10.	12	.6666667	.3286667	1
11.	13	.8	.462	1
12.	13	.8	.462	1
13.	14	.8666667	.5286667	1
14.	15	1	.662	1
15.	15	1	.662	1

- How serve is the violation of the continuity of F assumption with that many ties?

Kolmogorov confidence band: Example



Parametric, semi-parametric, non-parametric models

- Why this section: let's review more tools for regression analysis that tackle peculiarities in the data – we will judge their small sample properties later and think about the fact that our sample may be too small to say something meaningful about the sampling (let alone the population) distribution
- Parametric:
  - model features finite-dimensional but unknown parameter
  - estimators are efficient if model is correct but inconsistent if it is not
  - Converge at rate $\sqrt{N}$

Parametric, semi-parametric, non-parametric models

- Semi-parametric:
  - model features finite-dimensional parameter plus infinite dimensional "nuisance" parameter
  - more generality than non-parametric
  - easier to compute
  - more efficient than non-parametric equivalents
- Non-parametric:
  - infinite-dimensional and unknown parameter
  - Convergence usually much slower than $\sqrt{N}$

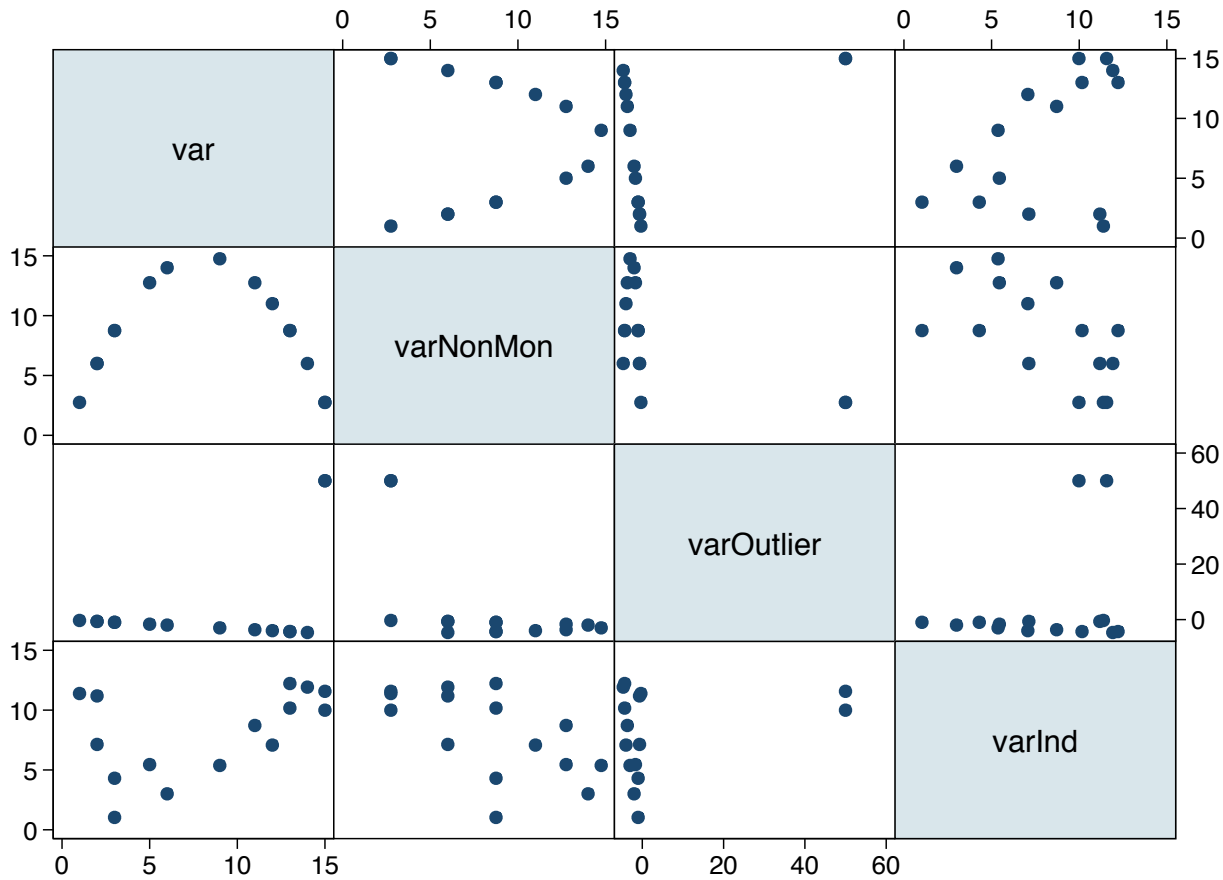
Explore the data, again

Take a look at bi- and multi-variate relationships

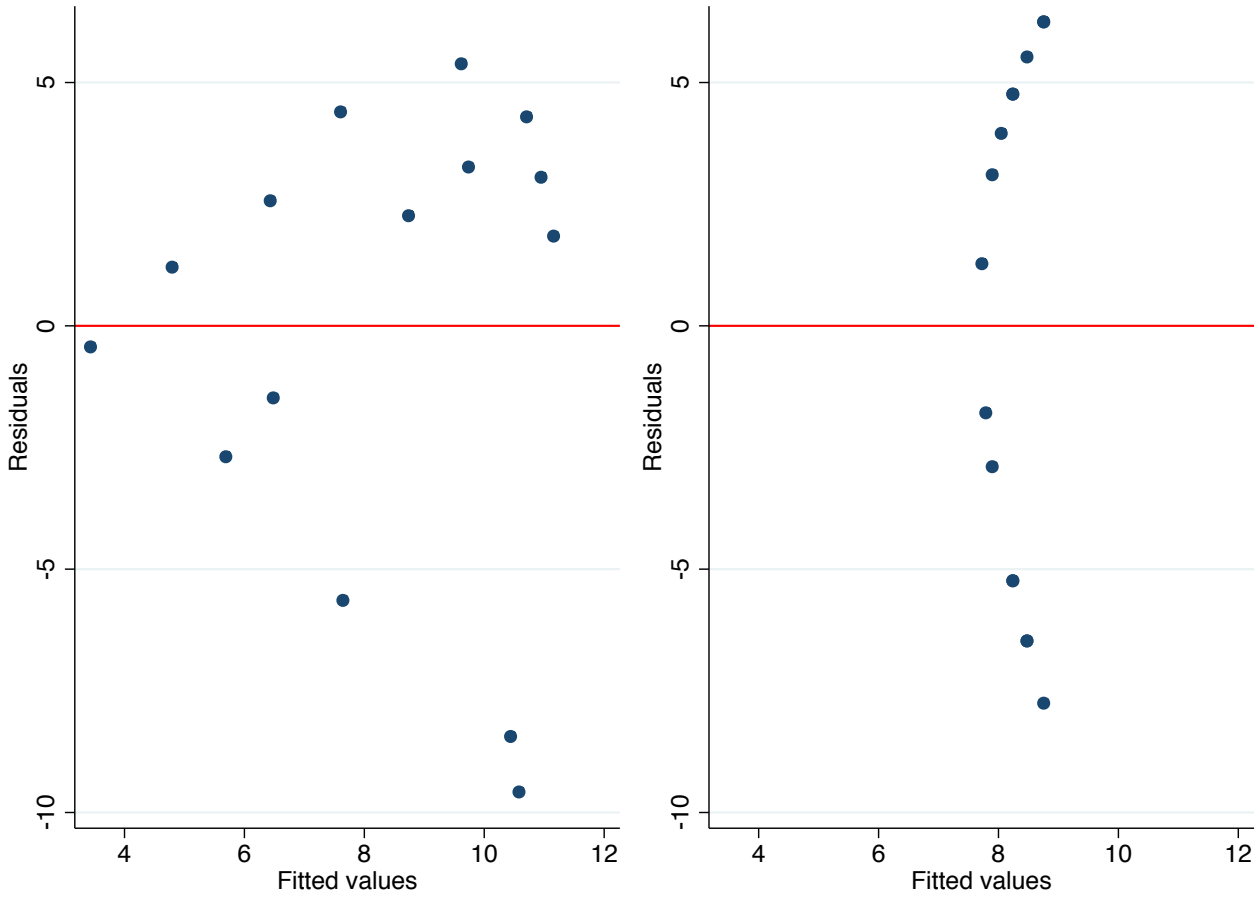
- Pairwise scatter plots
- 3D plots – yeah, not with *N*15
- Regression diagnostic plots: e.g., residual vs fitted plot

Sure, some more observations would be desirable for meaningful visualisation

Explore the data, again



Explore the data, again



Adjustments to standard parametric models

- Robust standard errors to tackle
  - heteroskedasticity
  - outliers
  - autocorrelation
  - clustering
  - etc
- Robust regression – iteratively re-weighted least squares, giving more weight to well behaved observations (judged by Crook's D)

Look at evaluations of small sample performance!

Semiparametric regression

- Partial least square regression:
  - Parametric estimation of only one part of a model, non-parametric lowess fitting of non-linear part
  - e.g. `plreg` in stata, `plsdepot`-package in R
- Semi-parametric regression:
  - Parametric estimation of a model except one variable that follows some non-linear relationship with the dependent variable
  - Non-parametric estimator could be a kernel smoother
  - e.g., `semipar` in Stata and package `SemiPar`-package in R

Again, look at evaluations of small sample performance

Non-parametric regression

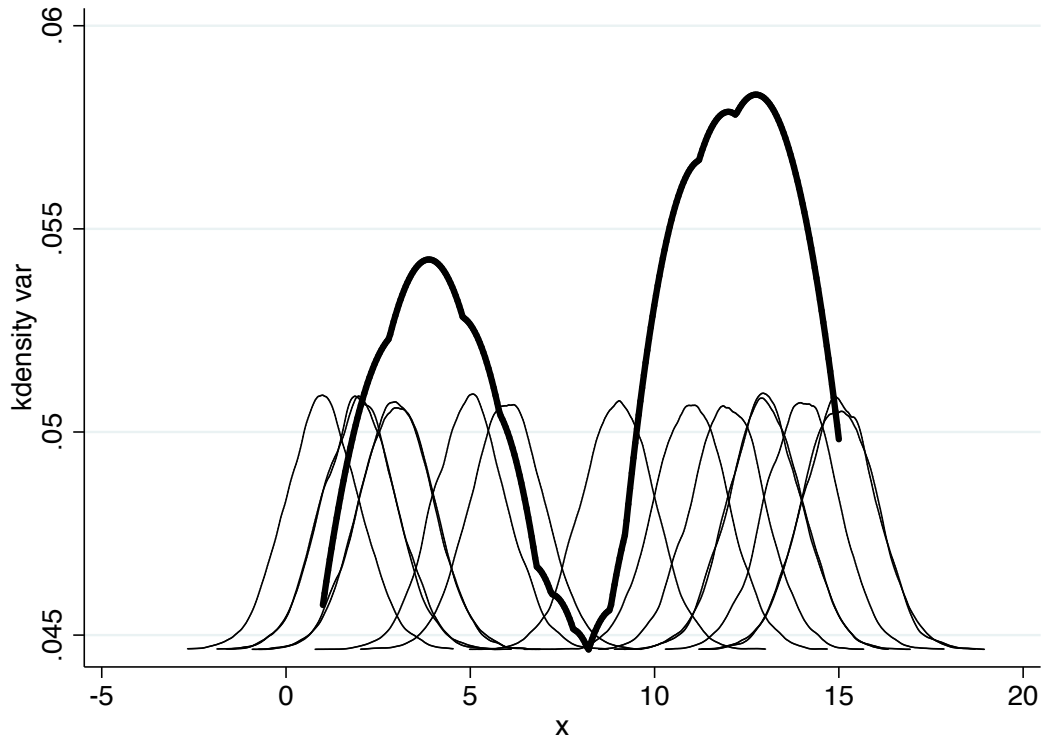
- Exact regression:
  - Distribution of test statistic based on permutations
  - Provides exact confidence intervals
  - e.g. `exlogistic` and `expoisson` in Stata and `elrm`-package in R
- Quantile regression
  - Estimates conditional median (or other quantile) – in contrast to conditional mean as in least square regression
  - Small samples may run into problems when estimating more extreme quantiles

Kernel density estimation: Basics

- What's a kernel: type of pdf that must be even
- What's kernel density estimation:
  - non-parametric method of estimating pdf of a continuous RV
  - adds kernel functions created around each value with that value at the centre then dividing that sum by the number of values
  - bandwith determines the width of each individual kernel function
- Assumptions:
  - Continous variable
  - Kernel is symmetric
  - Non-negative, real-valued
  - Definite integral over its support must equal 1
- No assumptions about the distribution of the data – that's why we care

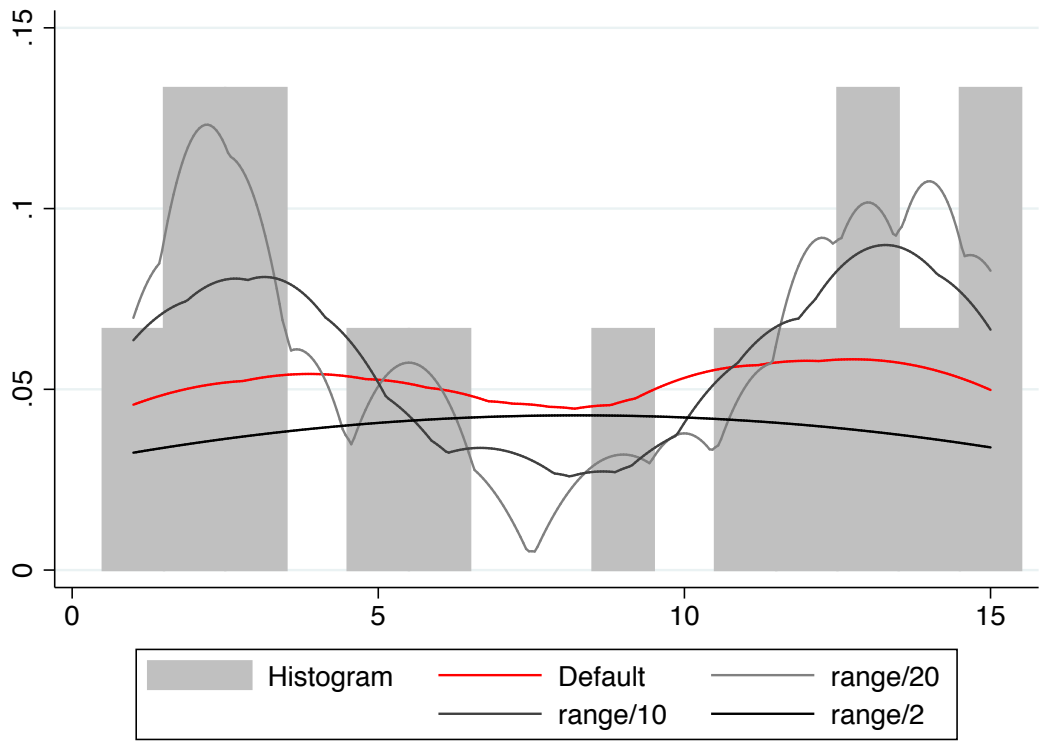
Kernel density estimation: Example

Figure: `kdensity varRdm` with overlaid normal distributions at each value



Kernel density estimation: Small sample issues

- What is density estimation good for in small samples?
- Exploratory data analysis only?
- Approximate discrete distribution with reasonable bandwidth
- Extract (continuous) approximation to extrapolate

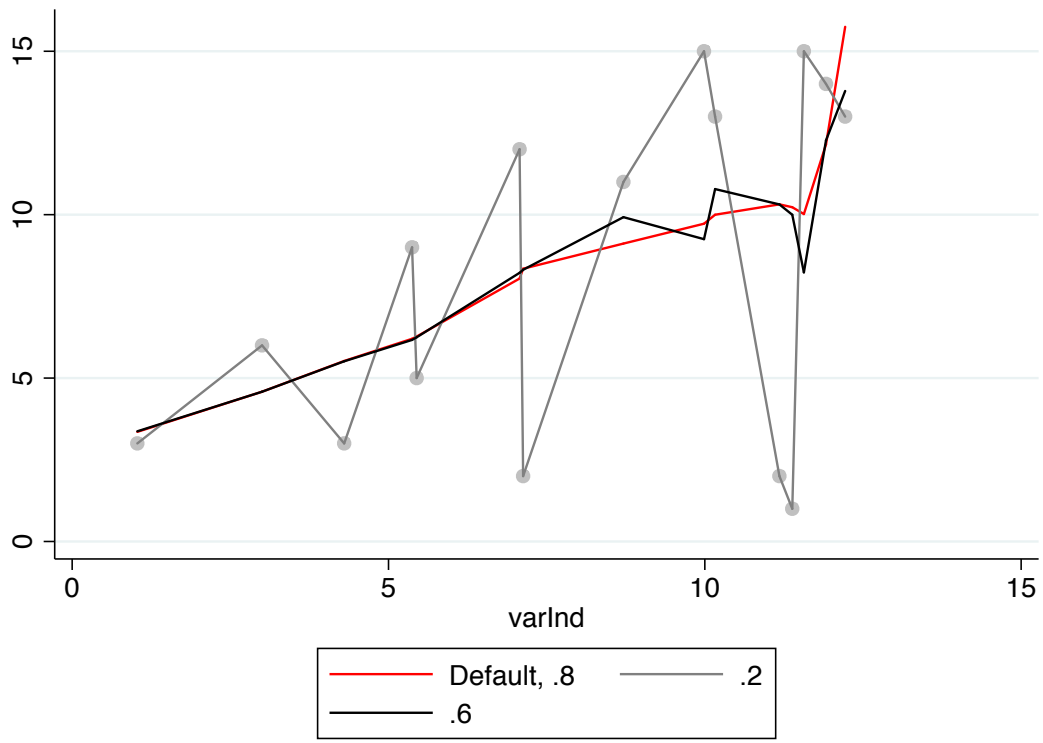


Kernel density estimation: Small sample issues

- Sufficient continous nature of data necessary
- Which bandwith to use for small samples? A large bandwidth that results in a large standard deviation covering more of the neighboring values
- Using the bandwith that minimizes the integrated squared error is only optimal if the data were normally distributed
  - Mostly Epanechnikov kernel as default – minimizes integrated squared error
  - All standard automated bandwith selection seem to fail in small samples (Heidenreich et al 2013)
  - eye-balling remains an option
- Strictly speaking, 2 observations are enough but for meaningful estimates of a well-behaved density in 1 dimension you need 4 (Silverman, 1986)

Local smoothing: Small sample issues

- What is density estimation good for in small samples?
- Exploratory data analysis only?
- Approximate discrete distribution with reasonable bandwidth
- Extract (continuous) approximation to extrapolate



References

Non-parametric regression slope estimator

- Hollander, Wolfe, and Chicken (2013): Nonparametric statistical methods, chapter 9.2, John Wiley & Sons
- Newson (2012): `censlope`

Semi-parametric regression analysis

- Yatchew (2003): Semiparametric Regression for the Applied Econometrician, Cambridge University Press
- Lokshin (2006): Difference-based semiparametric estimation of partial linear regression models, Stata Journal 6(3), pp. 377-83
- Verardi (2013): Semiparametric regression in Stata (slides)

Kernel density estimation

- Heidenreich, Schindler, Sperlich (2013): Bandwidth selection for kernel density estimation: a review of fully automatic selectors, Advances in Statistical Analysis 97(4), pp. 403-33