UNIVERSITY OF MICHIGAN
INSTITUTE FOR SOCIAL RESEARCH
Social Science in the Public Interest
Program in Survey Methodology

Regression Diagnostics for Survey Data

Richard Valliant

Joint Program in Survey Methodology,

University of Maryland and

University of Michigan USA

Jianzhu Li (Westat), Dan Liao (JPSM)

JPSM The Joint Program in Survey Methodology

Introduction

- Topics—adaptations to survey data of ...

  Leverages

  DFBETAS

  DFFITS

  Cook's D

  Collinearity measures

  Forward search

- Comparisons to standard diagnostics

Linear Regression on Survey Data

- Weighted least squares estimates (fixed effects)

$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim \left(0, v_i \sigma^2\right)$ independent (no clustering but can be handled)

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \mathbf{W} \mathbf{V}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W} \mathbf{V}^{-1} \mathbf{Y}$$

If constant variance, $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}$

$\mathbf{W}$ = diagonal matrix of survey weights

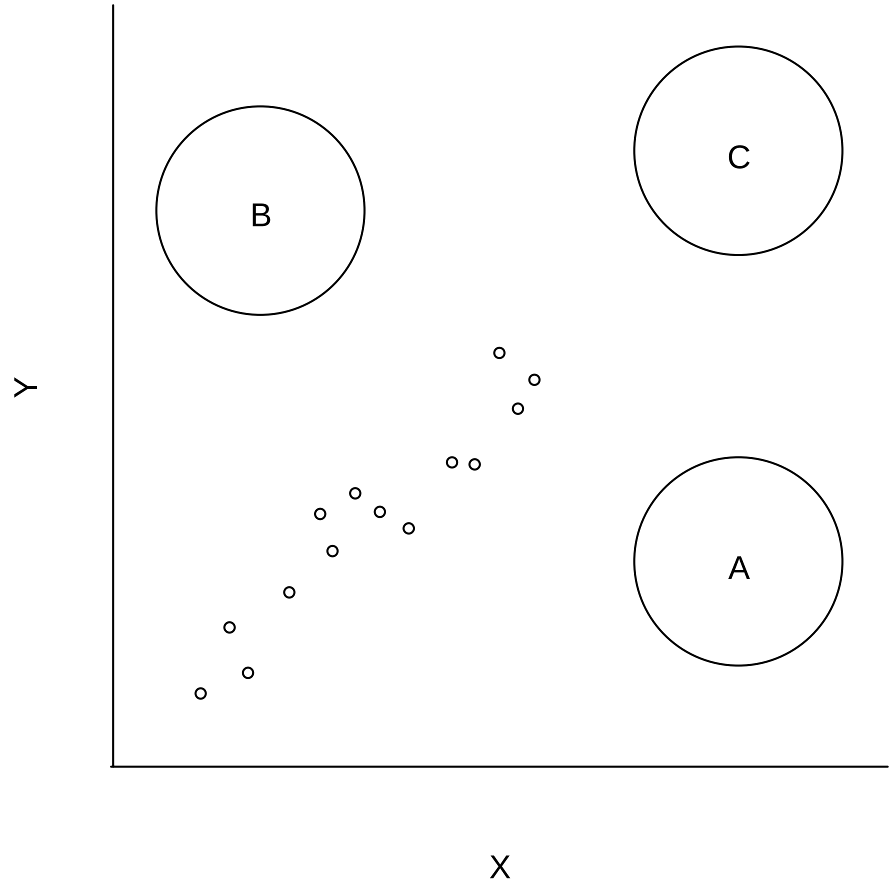
- $\hat{\boldsymbol{\beta}}$ can be interpreted as an estimate of

(i) parameter in underlying model or of

(ii) "census fit" parameter

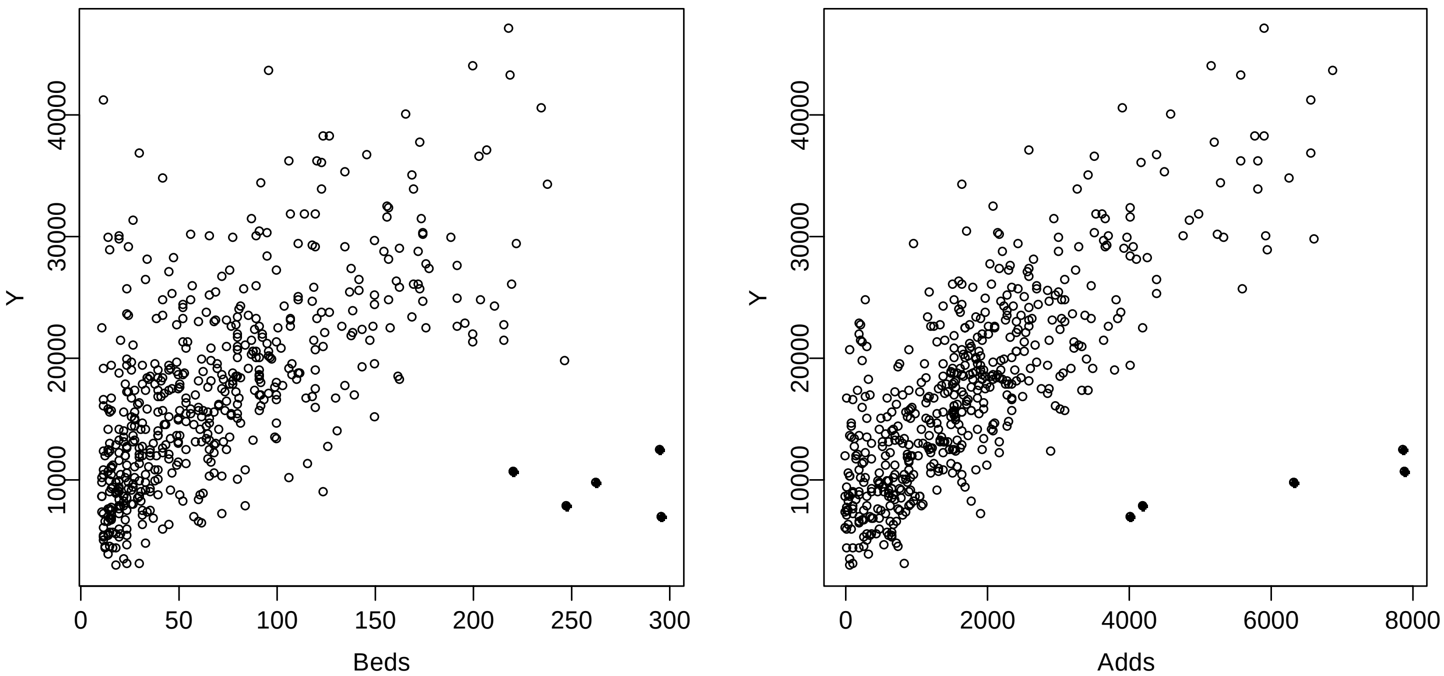
Reasons for Using Diagnostics

- Extreme points can affect regression parameter estimates, hypothesis tests, & confidence intervals
- Extremes can be due to
  - outlying X's or Y's (survey or non-survey data)
  - large weights (survey data)
  - interaction of weights with X's and Y's



A, B, and C are all influential. A, C may affect estimated slope.

C will not affect slope but may reduce SE of slope.



Generated data based on a survey of mental health organizations

The 5 points in the lower right may or may not be influential

depending on size of their survey weights.

Survey Weights

- Survey weights are intended to expand a sample to a finite population. They are <u>NOT</u> same as inverse-variance weights in usual WLS regression.
- Reasons for variation in size of weights due to sample design
  - Household surveys
    - Different sampling rates for demographic groups (e.g., to get equal sample sizes for groups)
  - Business/institution surveys
    - Varying sampling rates by type of business (retail, service, etc)
    - PPS sampling (probs $\propto$ no. of employees)

- More reasons for variation in size of weights
  - Differential follow-up for nonresponse, i.e., subsampling of neighborhoods at different rates for nonresponse conversion, callbacks
  - Low response rates followed by large nonresponse adjustments in some groups
  - Use of auxiliary data in estimation—poststratification by age, race, sex; general regression estimation using no. of employees, prior year expenditures, etc.

Examples

- 1999-2002 National Health & Nutrition Examination Survey (NHANES)

  Weight range for Mexican-Americans: 698 – 103,831 (148:1)

- 1998 Survey of Mental Health Organizations

  Weight range: 1 - 159

- 2002 Status of the Armed Forces Survey

  Weight range: 2.3 – 384 (168:1)

Hat Matrix and Leverages

(Li & Valliant, *Survey Methodology* 2009)

- Predicted values: $\hat{\mathbf{Y}} = \mathbf{HY}$

$$\mathbf{H} = \mathbf{XA}^{-1}\mathbf{X}^T\mathbf{W} \text{ with } \mathbf{A} = \mathbf{X}^T\mathbf{WX}$$

- Leverages on the diagonal of hat matrix are $h_{ii} = \mathbf{x}_i^T \mathbf{A}^{-1}\mathbf{x}_i w_i$

- When model has an intercept, leverage can be decomposed as

$$h_{ii} = \frac{1}{n}\frac{w_i}{\bar{w}}\left[1 + \hat{N}\left(\mathbf{x}_i - \bar{\mathbf{x}}_W\right)^T \mathbf{S}^{-1}\left(\mathbf{x}_i - \bar{\mathbf{x}}_W\right)\right],$$

$\mathbf{S}$ is a x-product matrix involving x's; $\bar{\mathbf{x}}_W$ wtd mean of x's

- A point has high leverage if its weight is >> average or $\mathbf{x}_i$ is toward edge of ellipsoid centered at $\bar{\mathbf{x}}_W$.

An Example

- 1998 Survey of Mental Health Organizations (SMHO). PPS sample
- Regress expenditures on no. of beds (BEDS), no. patients added during years (ADDS)

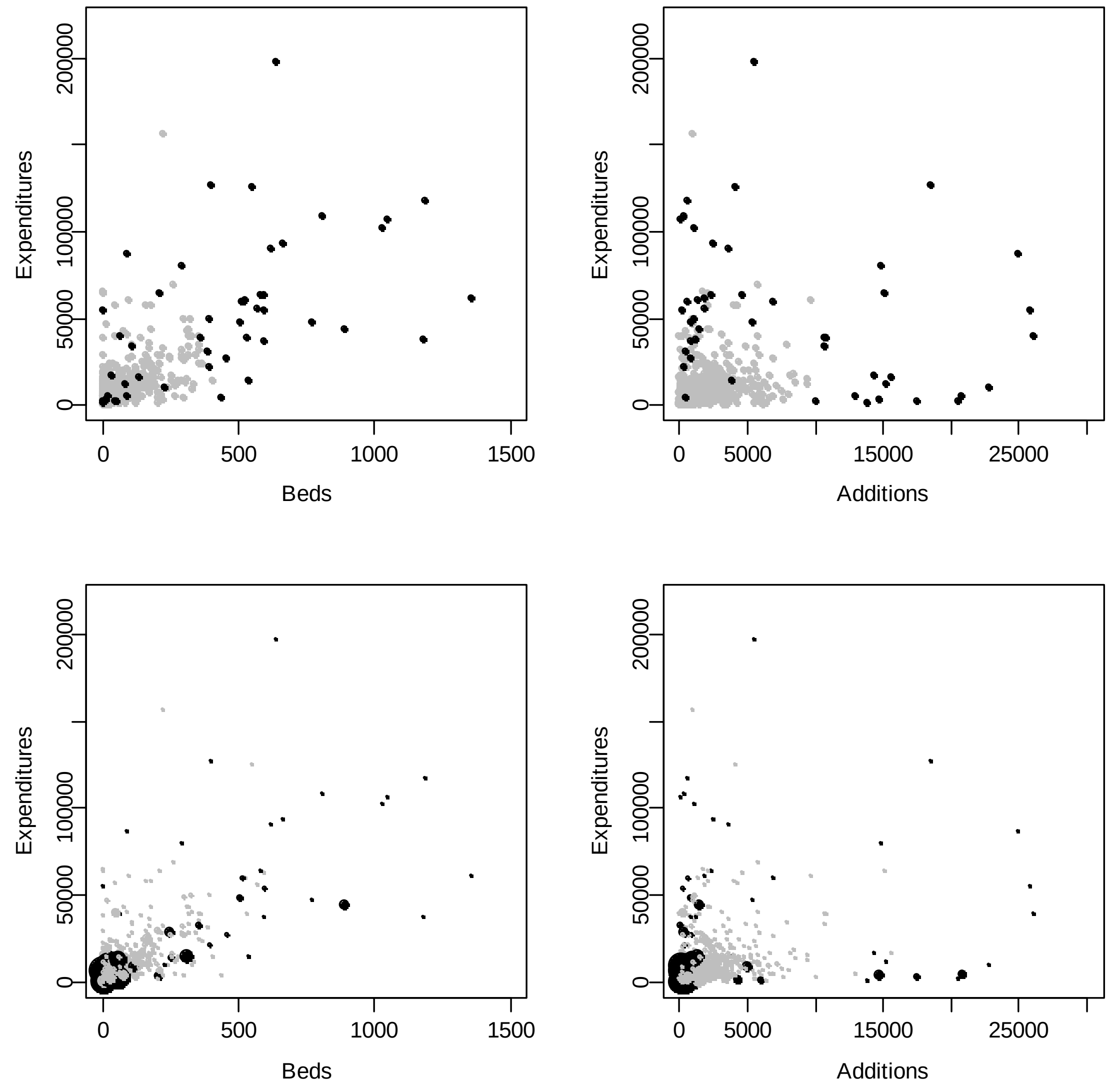
Quantiles of Variables in SMHO Regression.

	Quantiles				
Variables	0%	25%	50%	75%	100%
Expenditure (1000's)	17	2,932	6,240	11,842	519,863
BEDS	0	6	36	93	2,405
ADDS	0	558	1,410	2,406	79,808
Weights	1	1.42	2.48	7.76	158.86

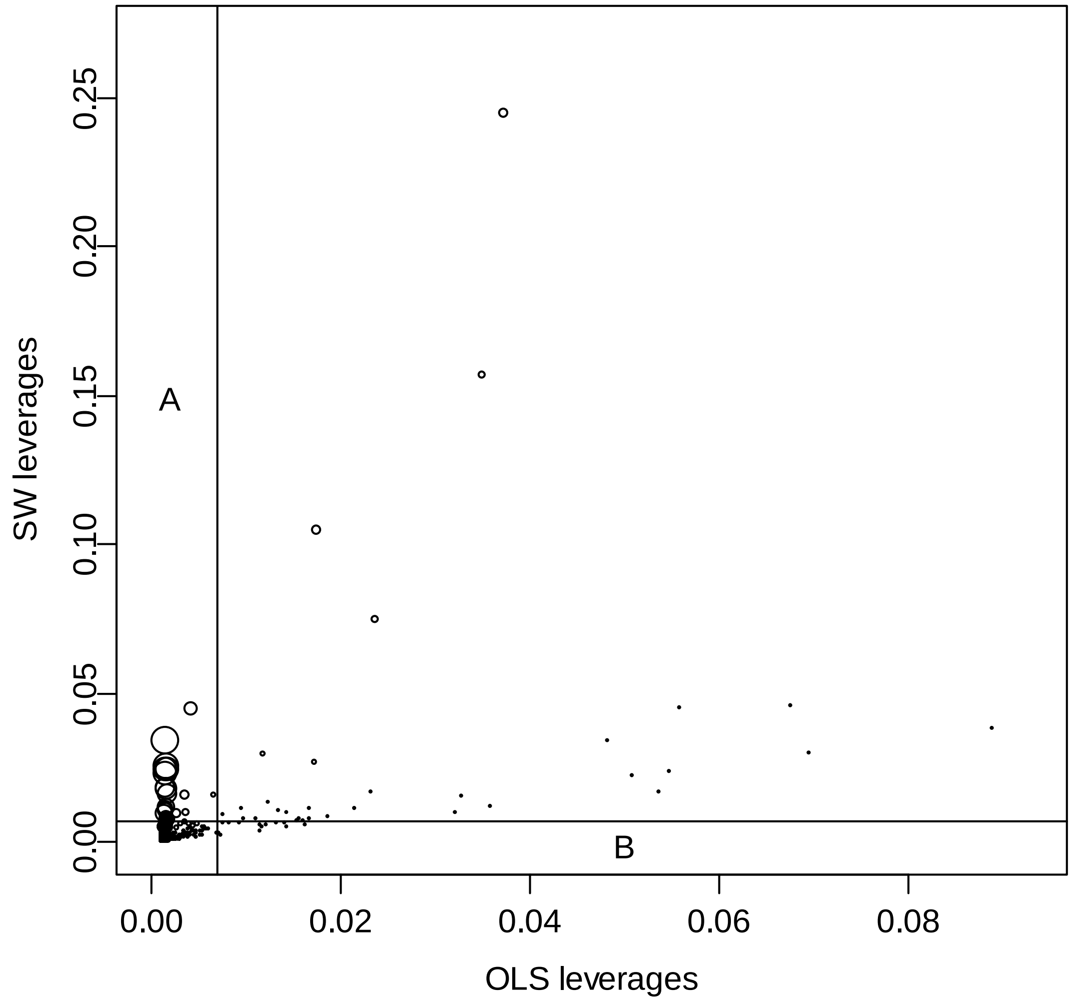
Scatterplots of expenditures versus beds and additions. High leverage points based on OLS (SW) are highlighted in top (bottom) row.

OLS

SW



Plot of survey weighted leverages versus OLS unweighted leverages.



Rule-of-thumb cutoff is $2\,p/n$

A = detected by SW only; B = detected by OLS only

OLS and SW parameter estimates of SMHO regression using all 875 sample cases.

Independent Variables	OLS Estimation			SW Estimation		
	Coefficient	SE	t	Coefficient	SE	t
Intercept	-1,201	**526**	-2.3	514	**1,158**	0.4
# of Beds	94	**3**	31	81	**13**	6.2
# of Additions	2.3	**0.13**	18	1.8	**0.8**	2.4

Deleting observations with leverages greater than $2p/n = 0.007$

Intercept	2,987	**490**	6	1,994	**354**	5.6
# of Beds	69	**4.4**	16	76	**6.7**	11.2
# of Additions	0.95	**0.20**	4.7	1.0	**0.20**	4.7

- After deleting high leverage points, SEs reduced, OLS and WLS estimates closer to each other.
- Significance of coefficients unchanged (except for intercept)

Variance Estimators

- Estimators of $\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right)$ are needed for several diagnostics

- Options are Binder sandwich (*ISR* 1983) or replication

  (jackknife, BRR, bootstrap)

  These are both design- and model-consistent.

- Purely model-based estimator useful for setting cutoffs

  $v_M\left(\hat{\boldsymbol{\beta}}\right)=\hat{\sigma}^2\mathbf{A}^{-1}\mathbf{X}^T\mathbf{W}^2\mathbf{X}\mathbf{A}^{-1}$ with $\hat{\sigma}^2=\sum_{i\in s} w_i e_i^2 \Big/ \left(\hat{N}-p\right)$

  $e_i = Y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$, $\hat{N}=\sum_{i\in s} w_i$

Standardized Residuals

- Standardizing so that residuals have (approximate) variance 1 makes interpretation easier.
- Use $e_i/\hat{\sigma}$
- Cutoff for large: 2 or 3 based on Gauss inequality

(No design-based, distribution theory for residuals, even asymptotically)

DFBETAS, DFFITS

(Li & Valliant 2009, submitted)

- Measure effect of single unit on each $\hat{\beta}_j$ separately

- $DFBETAS_{ij} = \dfrac{c_{ji}e_i \big/ (1-h_{ii})}{\sqrt{v(\hat{\beta}_j)}}$ with $c_{ij} = \left(\mathbf{A}^{-1}\mathbf{x}_i e_i w_i\right)_j$, i = unit, j = parm

  Based on $DFBETA_i = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i) = \mathbf{A}^{-1}\mathbf{x}_i e_i w_i \big/ (1-h_{ii})$

  Large if any of weight, residual, or leverage is large

  A lot to look at: np values

- Measure effect of unit *i* on prediction

  Multiply $DFBETA_i$ by $\mathbf{x}_i^T$ to get $DFFITS_i = \dfrac{h_{ii}e_i \big/ \left(1-h_{ii}\right)}{\sqrt{v\left(\hat{\beta}_j\right)}}$

- Heuristic cutoffs

  $DFBETAS_{ij} \quad z\big/\sqrt{n}$

  $DFFITS_i \quad z\sqrt{p/n}$, z = 2 or 3

  (Bonferroni adjustment to cutoffs can be used)

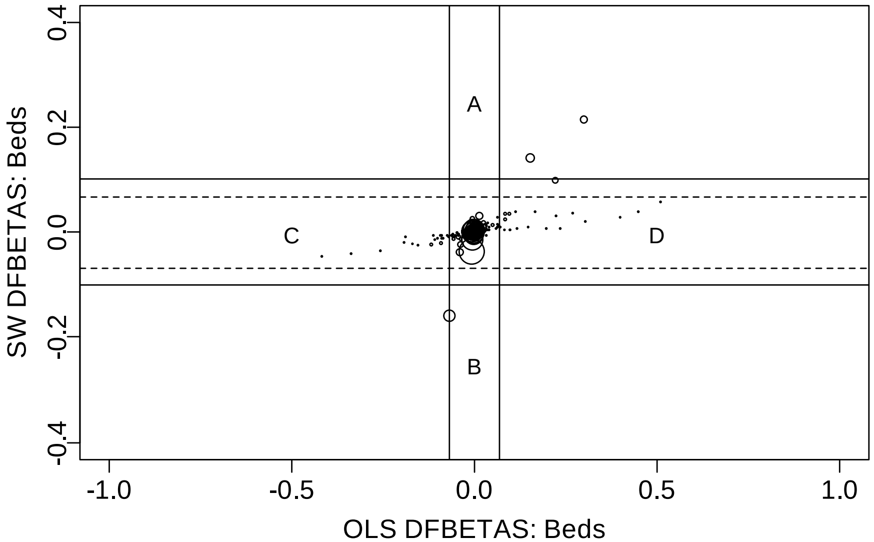
Extended Cook's D

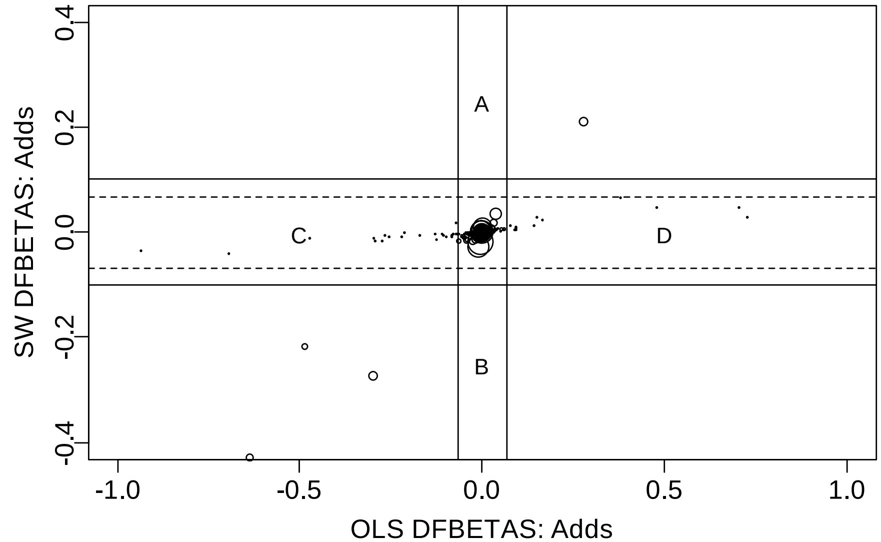
- Measures effect of single unit on vector estimate $\hat{\boldsymbol{\beta}}$
- $ED_i = \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i)\right)^T \left[v\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1} \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}(i)\right)$

  Compare to quantiles from $\chi^2(p)$ distribution. Influential units are ones that define a "large" ellipsoid centered at $\hat{\boldsymbol{\beta}}$.

- Per Atkinson (*JRSS-B* 1982), an alternative that detects more points is $MD_i = \sqrt{nED_i / p}$.
- Heuristic cutoff for MD_i is 2 or 3

SMHO Data: Regress expenditures on BEDS, ADDS

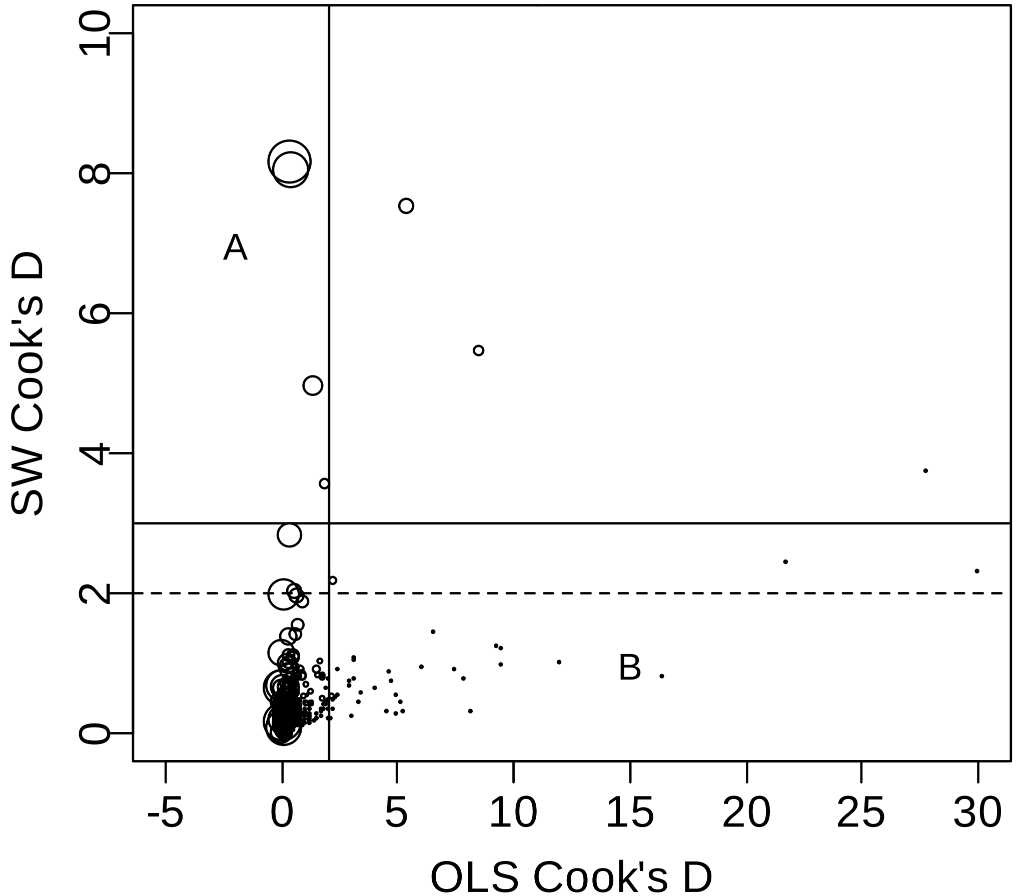




C & D are cases identified by OLS but not by SW

These are all cases with small weights.

OLS flags 57 cases; SW 9.



A = cases identified by SW only; B = OLS only

OLS flags 44; *MD* flags 10

OLS and SW Parameter Estimates after Deleting Observations with Large Modified Cook's Distance.

Independent Variables	OLS Estimation			SW Estimation		
	Coefficient	SE	*t*	Coefficient	SE	*t*
Intercept	-1,201	**526**	-2.3	514	**1,158**	0.4
# of Beds	94	**3**	31	81	**13**	6.2
# of Additions	2.3	**0.13**	18	1.8	**0.8**	2.4
No. units deleted	44			10		
Independent Variables	Coefficient	SE	*t*	Coefficient	SE	*t*
Intercept	1660	**335**	4.9	932	**345**	2.7
# of Beds	81	**2.4**	33	83	**5.7**	14.5
# of Additions	1.2	**0.12**	9.7	1.4	**0.3**	5.4

Forward Search

(Atkinson & Riani book 2000), Li & Valliant 2009, draft)

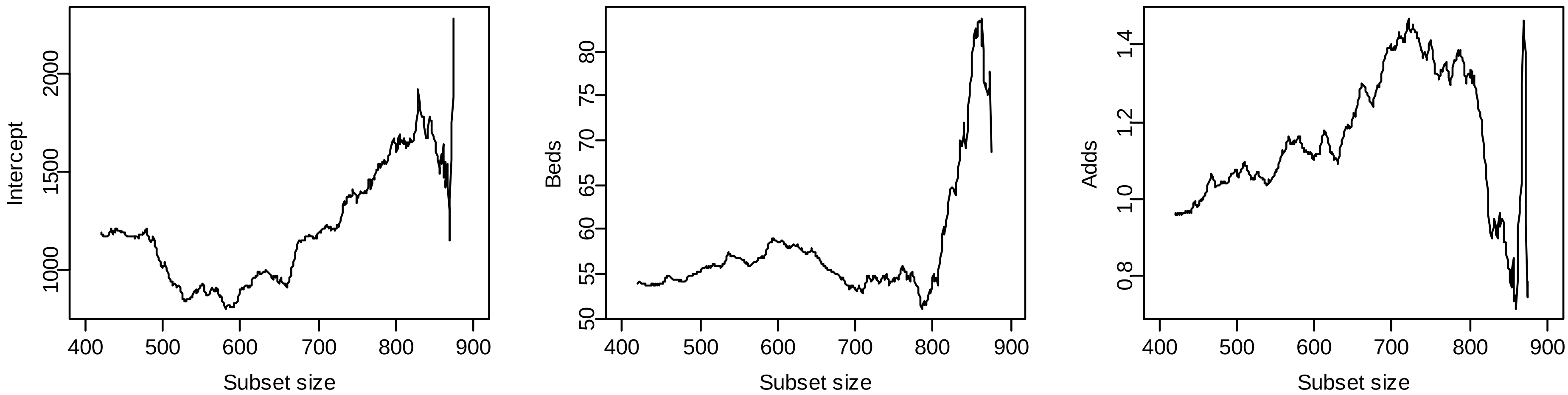
- One outlier can mask effect of another
- Identify **groups** of influential observations to avoid masking effect

- Method
  - Fit a robust regression (e.g., least median of squares) to subsample of full sample
  - Choose subsample that minimizes $median\left(e_{OLS,i}^{2}\right)$
  - Subsample $m = p$
  - Find $m+1$ cases with smallest squared residuals
  - Track $\hat{\sigma}^2$
  - Look for point at which $\hat{\sigma}^2$ makes abrupt change. All cases after that are called outliers.

    (No abrupt changes $\Rightarrow$ no outliers)
- Adaptations made for survey data

SMHO Data again

Plots of Parameter Estimates from Forward Search



83 points identified as influential; 20 never identified by single-case deletion methods (DFBETAS, DFFITS, modified Cook, etc)

Method may have promise but more work needed.

Collinearity

- Collinearity is worrisome for both numerical and statistical reasons.
- Estimates of slopes can be **numerically unstable**, i.e., small changes in the X's or the Y's can produce large changes in estimates.
- Correlation among predictors can lead to slope estimates with **large variances**.
- When X's are strongly correlated, R^2 can be large while the individual slope estimates are not statistically significant.
- Even if slope estimates are significant, they may have **opposite sign** of what is expected.

- Variance inflation factor (VIF)

  Measure of how much $\mathrm{var}\left(\hat{\beta}_j\right)$ is inflated compared to what it would be if x's were orthogonal.

$$\mathrm{Var}_M\left(\hat{\beta}_k\right) = \underbrace{\frac{1}{1-R_k^2}}_{VIF} \frac{\sigma^2}{\sum_{i \in s} x_{ik}^2}$$

R_k^2 is the R-square from regressing $\dot{\mathbf{x}}_k$ on the other x's.

$\dot{\mathbf{x}}_k$ = column k of $\mathbf{X}$

- For survey weighted regression estimator, if $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \left(\mathbf{0}, \mathbf{V}\right)$

$$\mathrm{Var}_M\left(\hat{\beta}_k\right) = \underbrace{\frac{\zeta_k \eta_k}{1 - R^2_{SW(k)}}}_{VIF} \left(\mathrm{Var}\ \text{if}\ \dot{\mathbf{x}}_k \perp \text{others}\right)$$

$R^2_{SW(k)}$ = R-square from SW of regression of $\dot{\mathbf{x}}_k$ on other *x*'s

$$\zeta_k = \frac{\mathbf{e}^T_{(k)} \mathbf{WVWe}_{(k)}}{\mathbf{e}^T_{(k)} \mathbf{We}_{(k)}}, \qquad \eta_k = \frac{\dot{\mathbf{x}}^T_k \mathbf{W} \dot{\mathbf{x}}_k}{\dot{\mathbf{x}}^T_k \mathbf{WVW} \dot{\mathbf{x}}_k},$$

$\mathbf{e}_{(k)}$ = vector of residuals from regressing $\dot{\mathbf{x}}_k$ on other *x*'s

- Approaches to estimation
  - Purely model-based
  - Think of census value of $\dfrac{\zeta_k \eta_k}{1 - R^2_{SW(k)}}$; fill in design-based estimates of each component.
- Variance decomposition using SVD: use to identify pairs of x's that are collinear (ala Belsley, Kuh, Welsch 1980)
- Work is in progress on this

Conclusion

- Different points can be influential in OLS and SW regression. Specialized diagnostics needed for survey data (assuming survey weighted LS used).
  - If you adopt OLS regression, use OLS diagnostics; if you adopt SW regression, use SW diagnostics.
- Little formal distribution theory available
- Packages do not currently include diagnostics for survey regressions

- Implications of dropping points based on diagnostics
  - "Core" model being fitted: one that fits for the portion of population that excludes influential points
  - Idea of estimating census parameter is lost

- What if mechanical procedure used that automatically drops points?
  - SE's too small, CI's cover at less than nominal rate, hypothesis tests reject too often
  - Similar to problems known for stepwise regression (Zhang *BMKA* 1992, Hurvich & Tsai *TAS* 1990)
- Collinearity has similar effects on survey estimators as in regular regression
  - Same inference problems may exist as above if automatic procedure used.