Cook's Distance Measures for Panel Data Models

David Vincent
dvincent@dveconometrics.co.uk

8 September 2022

2022 UK Stata Conference

Contents

1. Introduction
2. Cross Sectional Data
3. Panel Data Omit Row
4. Panel Data Omit Subject
5. `cooksd2`

Introduction

- When working with small samples, it is not uncommon to discover that regression estimates are very sensitive to the inclusion of just few datapoints
- These points are often referred to as influential observations as they have a disproportionate impact on the estimates
- Such anomalies may be caused by data recording errors, using an incorrect functional form or infrequently occurring events
- Identifying influential data points is an important step in empirical analysis as estimates that are being driven by a few observations casts doubt on the reliability of the analysis

Introduction

- The usual approach is to determine the impact of each observation when it is removed from the estimation sample
- A popular measure of influence is the distance statistic by Cook (1977), although the row-deletion formulas are designed for ordinary least squares and independent observations
- This presentation describes a new command `cooksd2`, that generates Cook's distance statistics for the fixed, random and between-effects regression estimators
- The updating formulas are based on Christensen et al. (1992) and extended to measure the influence of an entire subject following the methods described by Banerjee and Frees (1997)

Row Deletion

- Assume a standard linear regression $y_i = x_i'\beta + v_i$ where x_i is a K vector of explanatory variables and v_i is the error
- For N observations, the model in matrix form is $y = X\beta + v$ and the OLS estimates are $\widehat{\beta} = (X'X)^{-1}X'y$
- Letting r_i denote the ith residual, the OLS estimates $\widehat{\beta}_{(i)}$ when the ith row is deleted from y and X are:

$$\widehat{\beta}_{(i)} = \widehat{\beta} - \frac{(X'X)^{-1}x_i r_i}{1 - h_i} \tag{1}$$

This is based on the Woodbury matrix identify which is a numerically cheap way to compute the inverse $(X_{[i]}'X_{[i]})^{-1}$ where $X_{[i]}$ is the matrix without the ith row

Row Deletion

- The quantity h_i is the leverage and corresponds to the ith row and column of the hat matrix $H = X(X'X)^{-1}X'$
- As the fitted values $\widehat{y} = Hy$, the leverage of data point i is the contribution to $\widehat{y}_i$ that is made by y_i
- The leverage of a data point will be large when it has an extreme value for one or more of its regressors
- For the general case of K regressors, Cook (1977) provides an easily interpretable measure of the distance of $\widehat{\beta}_{(i)}$ from $\widehat{\beta}$ to assess the influence of each data point

Cook's Distance

- Assuming the errors $v \sim iidn(0, I\sigma^2)$, a $100(1-\alpha)\%$ confidence region for β is the set of vectors β^* that satisfy:

$$Pr\left(F_{K,N-K} \leq \frac{(\beta^* - \widehat{\beta})'(X'X)(\beta^* - \widehat{\beta})}{K\widehat{\sigma}^2}\right) = 1-\alpha$$

- The approach by Cook (1977) is to define the statistic:

$$D_i = \frac{(\widehat{\beta}_{(i)} - \widehat{\beta})'(X'X)(\widehat{\beta}_{(i)} - \widehat{\beta})}{K\widehat{\sigma}^2} \qquad (2)$$

which sets β^* to the fixed value $\widehat{\beta}_{(i)}$ and then compute the percentile of $F_{K,N-K}$ which corresponds to the value of D_i

$$Q_i = Pr(F_{K,N-K} \leq D_i)$$

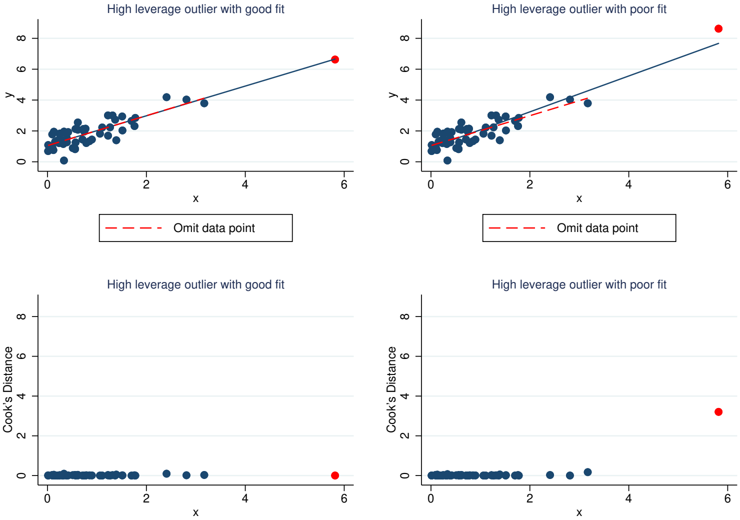
Cook's Distance

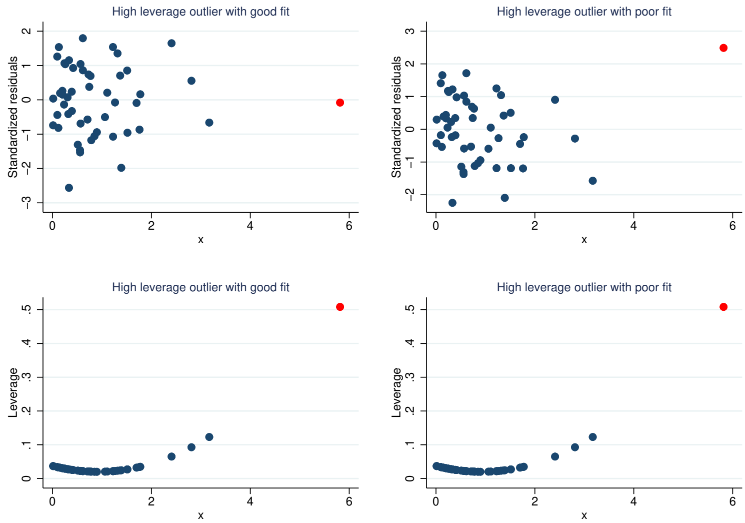
- If $Q_i = 0.5$, the removal of the ith data point moves the OLS estimates to the edge of a 50% confidence region for the unknown vector β based on $\widehat{\beta}$
- A popular cut-off is $D_i = 1$, as then $Pr(F_{K,N-K} \leq 1) \approx 0.5$ when $K > 5$. Other cut-offs such as $4/N$ are used, although it is better to look for large differences and not just whether they exceed suggested cut-offs
- Cook's distance can also be computed for subsets of the K parameters, but for an overall influence, it simplifies by substituting (1) into (2):

$$D_i = \left[\frac{r_i}{\widehat{\sigma}\sqrt{1 - h_i}} \right]^2 \frac{h_i}{K(1 - h_i)} \tag{3}$$

Leverage & Residual

- The influence of the *i*th data point is a combined measure of its (standardized) residual and its leverage. This is illustrated in the following figure for a outlier in the y and x-space
- In the left hand pane, the outlier is not influential as it follows the rest of the data. This is confirmed by the Cook's distance in the lower plot
- In the right hand pane the outlier is influential as it has a some residual. Removing this data point has a sizable impact on the estimates
- The second figure plots the residuals and leverage. The influential outlier cannot be detected from the residuals which illustrates why measures of influence also consider leverage


High leverage outlier with good fit
y
0
2
4
6
8
0
2
4
6
x
Omit data point
High leverage outlier with poor fit
y
0
2
4
6
8
0
2
4
6
x
Omit data point
High leverage outlier with good fit
Cook's Distance
0
2
4
6
8
0
2
4
6
x
High leverage outlier with poor fit
Cook's Distance
0
2
4
6
8
0
2
4
6
x


High leverage outlier with good fit
Standardized residuals
−3
−2
−1
0
1
2
0
2
4
6
x
High leverage outlier with poor fit
Standardized residuals
−2
−1
0
1
2
3
0
2
4
6
x
High leverage outlier with good fit
Leverage
0
.1
.2
.3
.4
.5
0
2
4
6
x
High leverage outlier with poor fit
Leverage
0
.1
.2
.3
.4
.5
0
2
4
6
x

Cook's Distance in Panel Settings

- Let y_{it} denote the outcome for individual $i = 1, .., N$ in period $t = 1, .., T_i$ and let x_{it} denote the K vector of regressors. The model to be estimated is:

$$y_{it} = x_{it}^{'}\beta + u_i + \epsilon_{it}$$

  where $u_i \sim iid(0, \sigma_u^2)$ and $\epsilon_{it} \sim iid(0, \sigma_\epsilon^2)$ are the individual specific and error components. Letting $v_{it} = u_i + \epsilon_{it}$ and stacking over T observations yields $y_i = X_i\beta + v_i$

- Transforming the model $y_i^* = W^{1/2}y_i$, $X_i^* = W^{1/2}X_i$ and $v_i^* = W^{1/2}v_i$, stacking over all N individuals and estimating by OLS yields different panel estimators:

$$\widehat{\beta} = (X^{*'}X^*)^{-1}X^{*'}y^* = \left(\sum_{i=1}^{N} X_i^{'}WX_i\right)^{-1} \sum_{i=1}^{N} X_i^{'}Wy_i$$

Cook's Distance in Panel Settings

- Unless otherwise stated, I assume a balanced panel to simplify the exposition, although the results generalize
- The fixed effects estimator uses deviations from the means of each individual. This corresponds to $W = Q_T$ where $Q_T = I_T - T^{-1}ee^{'}$ where e is a T vector of 1's.
- The random effects estimator applies a GLS transformation for an efficiency gain over OLS and corresponds to $W = \Omega^{-1}$ where $\Omega = E[v_i v_i^{'} \mid X_i]$
- Finally, the between estimator is a regression of the means, which corresponds to $W = M_T / T$ where $M_T = T^{-1}ee^{'}$

Cook's Distance in Panel Settings

- Despite estimation by OLS, Cook's formula in (3) cannot be applied, as the updating equation in (1) does not provide the correct estimates when a row is deleted from the data
- Instead this provides the OLS estimates when a row is removed from the transformed variables based on the full sample. But when a row is removed from the raw data, the other transformed values for that individual will also change
- Consider the fixed effects transformation $y^*_{it} = y_{it} - \bar{y}_i$. When this observation is removed from the data, the ij-th transformed value should now be $y_{ij} - \bar{y}_{i(t)}$
- Instead, the updating formula in (1) uses $y_{ij} - \bar{y}_i$ which continues to subtract the mean using all T observations

Cook's Distance in Panel Settings

- As the estimates are wrong, so too are the residuals and leverage values which Cook's distance in (3) uses

- This implies that the influence of each data point on the usual panel data estimators cannot be assessed from the full sample OLS estimates using the transformed data, after `xtdata`

- One approach is to simply re-fit the model without each data point and note that Cook's distance is equivalent to the F-statistic provided by `test` for testing the null $\beta = \widehat{\beta}_{(i)}$

- However this can be slow when there are a large number of observations or regressors. A more practical approach is to develop the appropriate leave-one-out formulas for each estimator that do not require refitting the entire model

Row Deletion

- Letting $y^*_{(it)}$ and $X^*_{(it)}$ denote the transformed data after dropping the it-th observation, the estimates become:

$$\widehat{\beta}_{(it)} = (X^{*'}_{(it)} X^*_{(it)})^{-1} X^{*'}_{(it)} y^*_{(it)} \tag{4}$$

- For the fixed and random effects estimators, the above cross products can be written as a subject i correction to the full sample values:

$$X^{*'}_{(it)} X^*_{(it)} = X^{*'} X^* - \frac{\widetilde{x}_{it} \widetilde{x}'_{it}}{s} \tag{5}$$

$$X^{*'}_{(it)} y^*_{(it)} = X^{*'} y^* - \frac{\widetilde{x}_{it} \widetilde{y}_{it}}{s} \tag{6}$$

where $\widetilde{x}_{it}$ and $\widetilde{y}_{it}$ are transformations that are estimator specific and where s is a scaling factor

Row Deletion

- This follows as dropping the *it*-th observation only impacts the contribution made by individual *i* hence for the regressors say, we can write:

$$X^{*'}_{(it)}X^{*}_{(it)} = \sum_{j\neq i}^{N} X_j^{*'}X_j^{*} + X^{*'}_{i(t)}X^{*}_{i(t)}$$

- After some algebra, it can be shown that for the fixed and random effects estimators:

$$X^{*'}_{i(t)}X^{*}_{i(t)} = X_i^{*'}X_i^{*} - \tilde{x}_{it}\tilde{x}'_{it}/s$$

- Finally, for OLS with no data transformation, the above is $X'_{i[t]}X_{i[t]} = X'_i X'_i - x_i x'_i$ where $X_{i[t]}$ is the matrix without row t

Row Deletion

- As (5) is a correction to the full sample result, the Woodbury matrix identity can be applied to find the inverse:

$$(X^{*\prime}_{(it)}X^{*}_{(it)})^{-1} = (X^{*\prime}X^{*})^{-1} + \frac{(X^{*\prime}X^{*})^{-1}\tilde{x}_{it}\tilde{x}'_{it}(X^{*\prime}X^{*})^{-1}}{s - \tilde{h}_{it}} \quad (7)$$

  where $\tilde{h}_{it}$ is the generalized leverage:

$$\tilde{h}_{it} = \tilde{x}'_{it}(X^{*\prime}X^{*})^{-1}\tilde{x}_{it}$$

- This provides the inverse of the corrected matrix by making a correction to the inverse of the original matrix and will be quicker than `invsym(A)`, especially when K is large

Row Deletion

- Substituting (6) and (7) into (4) provides an efficient updating formula without the *it*-th observation:

$$\widehat{\beta}_{(it)} = \widehat{\beta} - (X^{*\prime}X^{*})^{-1}\widetilde{x}_{it}\frac{(\widetilde{y}_{it} - \widetilde{x}'_{it}\widehat{\beta})}{s - \widetilde{h}_{it}} \tag{8}$$

- The above can also be used to obtain the error variance in the transformed model without the *it*-th observation. For the fixed effects method this is an estimate of $\sigma^2_{\epsilon(it)}$ given by:

$$\widehat{\sigma}^2_{\epsilon(it)} = \frac{NT - K}{NT - K - 1}\widehat{\sigma}^2_{\epsilon} - \frac{(\widetilde{y}_{it} - \widetilde{x}'_{it}\widehat{\beta})^2}{(s - \widetilde{h}_{it})(NT - K - 1)} \tag{9}$$

- Implementing these methods requires expressions for s, $\widetilde{y}_{it}$ and $\widetilde{x}_{it}$ which are presented below for each estimator

Fixed Effects Estimator

- For the fixed effects estimator, $\tilde{x}_{it}$ and $\tilde{y}_{it}$ are those that set $X'_{i[t]}Q_{T-1}X_{i[t]}$ and $X'_{i[t]}Q_{T-1}y_{i[t]}$ to their full sample values:
- These are $\tilde{x}_{it} = x_{it} - \bar{x}_{i(t)}$ and $\tilde{y}_{it} = y_{it} - \bar{y}_{i(t)}$ which are deviations from the means that exclude observation t and where the scaling factor is $T(T-1)^{-1}$
- These can be written in terms of the full sample fixed effects transformed variables x^*_{it} and y^*_{it} with redefined scaling factor:

$$\begin{aligned}\tilde{x}_{it} &= x_{it} - \bar{x}_i = x^*_{it}\\ \tilde{y}_{it} &= y_{it} - \bar{y}_i = y^*_{it}\\ s &= \frac{T-1}{T}\end{aligned}$$

Random Effects Estimator

- For the random effects estimator, these set $X'_{i[t]}\Omega^{-1}_{[T-1]}X_{i[t]}$ and $X'_{i[t]}\Omega^{-1}_{[T-1]}y_{i[t]}$ to their full sample values
- Using the approach in Christensen et al. (1992) who develop case deletion statistics in mixed models, it can be shown that:

$$\begin{aligned} \tilde{x}_{it} &= x_{it} - \bar{x}_{i(t)}\left[\frac{(T-1)\sigma_u^2}{(T-1)\sigma_u^2+\sigma_\epsilon^2}\right] \\ \tilde{y}_{it} &= y_{it} - \bar{y}_{i(t)}\left[\frac{(T-1)\sigma_u^2}{(T-1)\sigma_u^2+\sigma_\epsilon^2}\right] \end{aligned}$$

where the scaling factor is:

$$s = \frac{T\sigma_u^2+\sigma_\epsilon^2}{(T-1)\sigma_u^2+\sigma_\epsilon^2}$$

Random Effects Estimator

- Letting $r = \sigma_u^2/\sigma_\epsilon^2$, these expressions can be expressed in terms of the full sample transformations as follows:

$$\begin{aligned} \tilde{x}_{it} &= x_{it}^{*} - \lambda(1-\lambda)\bar{x}_i \\ \tilde{y}_{it} &= y_{it}^{*} - \lambda(1-\lambda)\bar{y}_i \end{aligned}$$

  where the scaling factor is now defined as:

$$s = \frac{(T-1)r+1}{Tr+1}$$

- The full sample GLS transformed variables are $x_{it}^{*} = x_{it} - \lambda\bar{x}_i$ and $y_{it}^{*} = y_{it} - \lambda\bar{y}_i$ where $\lambda = 1 - \sqrt{\frac{1}{Tr+1}}$

Between Effects Estimator

- An updating formula for the between-effects estimator can be derived in the same way, but involves matrix operations:

$$\widehat{\beta}_{(it)} = \widehat{\beta} - (X^{*'}X^{*})^{-1}\widetilde{X}_{it}[S - H_{it}]^{-1}(\widetilde{y}'_{it} - \widetilde{X}'_{it}\widehat{\beta}) \qquad (10)$$

  where $\widetilde{X}_{it}$ is a $K \times 2$ matrix and $\widetilde{y}_{it}$ is 1×2 vector:

$$\begin{aligned} \widetilde{X}_{it} &= \begin{bmatrix} \bar{x}_i & T\bar{x}_i - x_{it} \end{bmatrix} \\ \widetilde{y}_{it} &= \begin{bmatrix} \bar{y}_i & T\bar{y}_i - y_{it} \end{bmatrix} \end{aligned}$$

  and where:

$$H_{it} = \widetilde{X}'_{it}(X^{*'}X^{*})^{-1}\widetilde{X}_{it}$$

- Although this requires inverting $[S - H_{it}]^{-1}$, this is only a 2×2 matrix and is very quick to compute

Between Effects Estimator

- The scaling matrix S takes different values depending on which version of the estimator is being used
- The unweighted estimator uses the sample means $X^{*'}X^{*} = \sum \bar{x}_i\bar{x}_i'$ and is invoked by typing the command `xtreg,be`. For this method, the scaling matrix is:

$$S = \begin{bmatrix} 1 & 0 \\ 0 & -(T-1)^{-2} \end{bmatrix}$$

- For the T_i-weighted variant, invoked by typing `xtreg,be wls`, $X^{*'}X^{*} = \sum T_i\bar{x}_i\bar{x}_i'$, the and scaling matrix is:

$$S = \begin{bmatrix} T^{-1} & 0 \\ 0 & -(T-1) \end{bmatrix}$$

Between Effects Estimator

- After some algebra an updating formula for the estimate of variance in the between model is:

$$\widehat{\sigma}^2_{b(it)} = \widehat{\sigma}^2_b - \frac{(\widetilde{y}'_{it} - \widetilde{X}'_{it}\widehat{\beta})'[S - H_{it}]^{-1}(\widetilde{y}'_{it} - \widetilde{X}'_{it}\widehat{\beta})}{N-K} \tag{11}$$

- As illustrated in the following slides, the above is typically used to obtain an updating formula for the variance of the individual specific effects

Updating Formula for Variance Components

- For the random effects estimator, the variance parameters σ^2_ϵ and σ^2_u impact $(X^{*'}X^*)^{-1}$ and must be estimated
- The literature sets these to their full sample estimates, but this means the leave-one-out formula will be incorrect, as the variance estimates also change when an observation is deleted
- An updating formula for $\widehat{\sigma}^2_{\epsilon(it)}$ and $\widehat{\sigma}^2_{b(it)}$ is already provided by the fixed and between effects methods, hence $\widehat{\sigma}^2_{u(it)}$ is then:

$$\sigma^2_{u(it)} = \max\left\{\widehat{\sigma}^2_{b(it)} - \frac{\widehat{\sigma}^2_{\epsilon(it)}}{T^*_{(it)}}, 0\right\} \tag{12}$$

where $T^*_{(it)}$ is the harmonic mean. Thus to obtain estimates that are the same as re-running `xtreg,re` requires an efficient way to update $(X^{*'}X^*)^{-1}$ when the variance terms change

Updating Formula for Variance Components

- To simply the exposition I assume a balanced panel. For the unbalanced case, an approximation is obtained using the average number of periods. First recall that:

$$(X^{*'}X^{*})^{-1} = \left[\sum_{i=1}^{N} X_i'\Omega^{-1}X_i\right]^{-1} \qquad (13)$$

  and ignoring scaling factors note that:

$$\Omega^{-1} = [Q + \psi M] \qquad (14)$$

- Letting $r = (\sigma_u^2/\sigma_\epsilon^2)$, then:

$$\psi(r) = \frac{1}{Tr + 1}$$

Updating Formula for Variance Components

- Substituting (14) into (13) yields:

$$(X^{*'}X^{*})^{-1} = [\underbrace{\sum_{i=1}^{N} X_i Q X_i}_{A} + \psi \underbrace{\sum_{i=1}^{N} X_i' M X_i}_{B}]^{-1} = B^{-1}(AB^{-1} + I\psi)^{-1}$$

- By the eigendecomposition, $AB^{-1} = C \Lambda C^{-1}$, where C are the eigenvectors and Λ are the eigenvectors, then:

$$B^{-1}(AB^{-1} + I\psi) = B^{-1}(C\Lambda C^{-1} + I\psi)^{-1} = B^{-1}C(\Lambda + I\psi)^{-1}C^{-1}$$

- As Λ is diagonal, the above provides a quick method for updating $(X^{*'}X^{*})^{-1}$ as ψ and the variance estimates change

Generating Cook's Distance Statistics

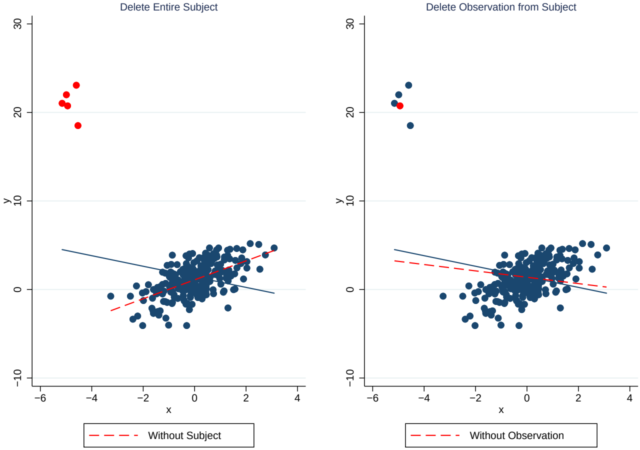
- The full set of NT leave-one-out estimates $\widehat{\beta}_{(it)}$, can be plugged into the underlying definition of the Cook's distance formula in (2) to assess the influence of each observation

- For the fixed effects estimator, these can be compared to the percentiles of the F-distribution with K and $N(T-1)-K$ degrees of freedom to determine what confidence level for β corresponds to the distance $\widehat{\beta}_{(it)} - \widehat{\beta}$

- For the random effects estimator, $D_{it} \times K$ can be compared to the percentiles of the chi-square distribution with K degrees of freedom. All results generalize to the case of unbalanced panels by replacing T with the T_i, for $i = 1, .., N$

Subject Deletion

- Leave-one-out methods may not be useful in detecting a cluster of observations that do not fit the overall pattern



Subject Deletion

- The approach taken by Banerjee and Frees (1997) is to measure the overall impact of subject i on the estimates
- Letting $y^*_{(i)}$ and $X^*_{(i)}$ denote the $N-1$ stacked matrices without individual i, the relevant estimators are:

$$\widehat{\beta}_{(i)} = (X^{*\prime}_{(i)} X^*_{(i)})^{-1} X^{*\prime}_{(i)} y^*_{(i)} \tag{15}$$

where for all three estimators, the cross products are:

$$\begin{aligned} X^{*\prime}_{(i)} X^*_{(i)} &= X^{*\prime} X^* - X^{*\prime}_i X_i && (16) \\ X^{*\prime}_{(i)} y^*_{(i)} &= X^{*\prime} y^* - X^{*\prime}_i y_i && (17) \end{aligned}$$

Subject Deletion

- Applying the Woodbury Matrix identity, the inverse of (16):

$$(X_{(i)}^{*'} X_{(i)}^{*})^{-1} = (X^{*'} X^{*})^{-1} + (X^{*'} X^{*})^{-1} X_i^{*'} [I - H_i^{*}]^{-1} X_i^{*} (X^{*'} X^{*})^{-1}$$

- Then substituting (17) and the above into (15) provides an efficient updating formula without the *i*-th subject:

$$\widehat{\beta}_{(i)} = \widehat{\beta} - (X^{*'} X^{*})^{-1} X_i^{*'} [I - H_i^{*}]^{-1} (y_i^{*} - X_i^{*} \widehat{\beta}) \qquad (18)$$

  where the matrix:

$$H_i^{*} = X_i^{*} (X_i^{*'} X_i^{*})^{-1} X_i^{*'}$$

- Although this requires N inversions, H_i^{*} is a $T \times T$ matrix, and T is typically small in panel applications

Subject Deletion

- Although not discussed in Banerjee and Frees (1997), we can use (18) to obtain the updated error variance without individual i
- Letting $r_i^* = y_i^* - X_i^* \widehat{\beta}$ denote the residuals, for the fixed effects estimator, the updated variance estimate $\widehat{\sigma}^2_{\epsilon(i)}$ is:

$$\widehat{\sigma}^2_{\epsilon(i)} = \frac{NT-K}{NT-K-T}\widehat{\sigma}^2_{\epsilon} - \frac{r_i^{*'}[I-H_i^*]^{-1}r_i^*}{NT-K-T} \qquad (19)$$

- For the between-effects estimator (18) and the updating formula for $\widehat{\sigma}^2_{b(i)}$ simplify to the usual cross sectional formulas:

$$\widehat{\sigma}^2_{b(i)} = \frac{N-K}{N-K-1}\widehat{\sigma}^2_{b} - \frac{(\bar{y}_i - \bar{x}_i'\widehat{\beta})^2}{(1-h_i)(N-K-1)} \qquad (20)$$

cooksd2

Cook's distance after regress and xtreg:

cooksd2 *newvar* [, cvars(*varlist*) parms(*newvar*) panel(*varname*) noconstant]

---

cvars(*varlist*) computes Cook's distance using *varlist* including the constant. The default uses all variables in the regression

parms(*newvar*) adds the jackknifed regression coefficients to the dataset. These take the variable names with prefix *newvar*

panel(*varname*) evaluates the influence of the entire subject after xtreg. After regress, the group variable varname is required

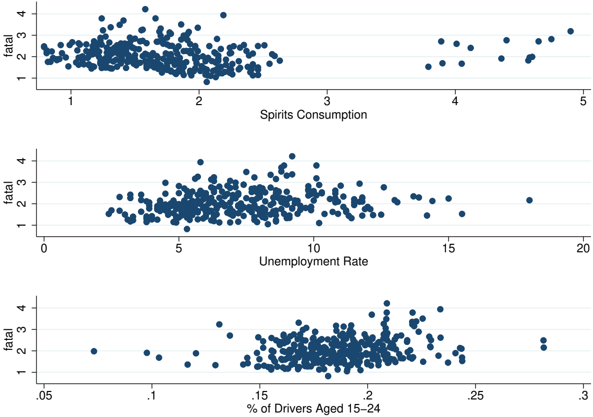
noconstant excludes the regression constant in the Cook's distance calculation. Helpful when the constant is unimportant

State Traffic Fatality Dataset

```
. use  http://www.stata-press.com/data/imeus/traffic,clear

. describe  fatal spircons unrate yngdrv

              storage  display    value
variable name   type   format     label      variable label
-----------------------------------------------------------------------------------------------------

fatal           float  %9.0g
spircons        float  %9.0g                 Spirits Consumption
unrate          float  %9.0g                 Unemployment Rate
yngdrv          float  %9.0g                 % of Drivers Aged 15-24
```

```
. xtsum  fatal spircons unrate yngdrv

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
fatal    overall |  2.040444   .5701938     .82121    4.21784 |     N =     336
         between |             .5461407   1.110077   3.653197 |     n =      48
         within  |             .1794253    1.45556   2.962664 |     T =       7
                 |                                            |
spircons overall |   1.75369   .6835745        .79        4.9 |     N =     336
         between |             .6734649   .8614286   4.388572 |     n =      48
         within  |              .147792   1.255119   2.265119 |     T =       7
                 |                                            |
unrate   overall |  7.346726   2.533405        2.4         18 |     N =     336
         between |             1.953377        4.1       13.2 |     n =      48
         within  |             1.634257   4.046726   12.14673 |     T =       7
                 |                                            |
yngdrv   overall |  .1859299   .0248736    .073137    .281625 |     N =     336
         between |              .017161   .1375446    .222699 |     n =      48
         within  |             .0181513   .1215223   .2513753 |     T =       7
```

State Traffic Fatality Dataset

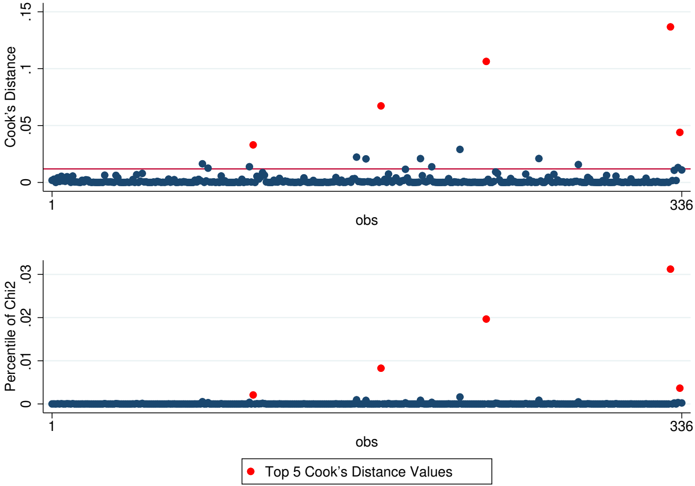


Random Effects Regression

```
. xtreg  fatal spircons unrate yngdrv, re

Random-effects GLS regression                   Number of obs     =        336
Group variable: state                           Number of groups  =         48

R-sq:                                           Obs per group:
     within  = 0.2423                                         min =          7
     between = 0.0269                                         avg =        7.0
     overall = 0.0103                                         max =          7

                                                Wald chi2(3)      =      65.98
corr(u_i, X)   = 0 (assumed)                    Prob > chi2       =     0.0000

------------------------------------------------------------------------------
       fatal |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    spircons |   .2539986   .0732514     3.47   0.001     .1104284    .3975687
      unrate |  -.0558281   .0072446    -7.71   0.000    -.0700271    -.041629
      yngdrv |   1.984222   .7457939     2.66   0.008     .5224924    3.445951
       _cons |   1.636236   .1359906    12.03   0.000       1.3697    1.902773
-------------+----------------------------------------------------------------
     sigma_u |  .49947472
     sigma_e |  .16643841
         rho |  .90005747   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Cook’s Distance for Each Observation

```
. cooksd2 cdre, parms(re_)
Cooks-distance using: spircons unrate yngdrv _cons
```

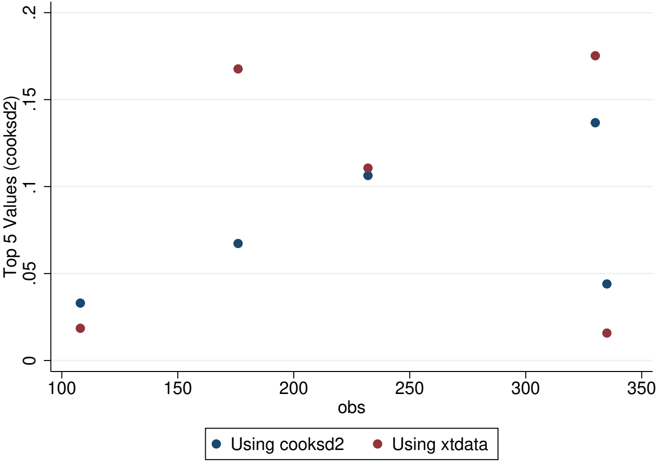


Largest 5 Cook's Distance Values

```
. gsort -cdre
. format fatal cdre cdre_pr_chi2 re_*  %7.0g
. list obs state year cdre re_* in 1/5, abbreviate(13) noobs

  +--------------------------------------------------------------------------------------------------------------+
  | obs   state   year     cdre   re_b_spircons   re_b_unrate   re_b_yngdrv   re_b_cons   re_sigma_u   re_sigma_e |
  |--------------------------------------------------------------------------------------------------------------|
  | 330      WY   1982   .13672          .24102       -.05176        1.5969      1.6994       .49641       .16468 |
  | 232      OK   1982   .10637          .23609       -.05191        1.7116       1.687       .49795       .16123 |
  | 176      NV   1982   .06729          .22068       -.05653        2.1157      1.6739       .49973       .16554 |
  | 335      WY   1987   .04403          .25306        -.0536        1.7231      1.6714       .50207       .16516 |
  | 108      LA   1984   .03303          .24666       -.05748        2.2448      1.6136       .49726       .16638 |
  +--------------------------------------------------------------------------------------------------------------+
```

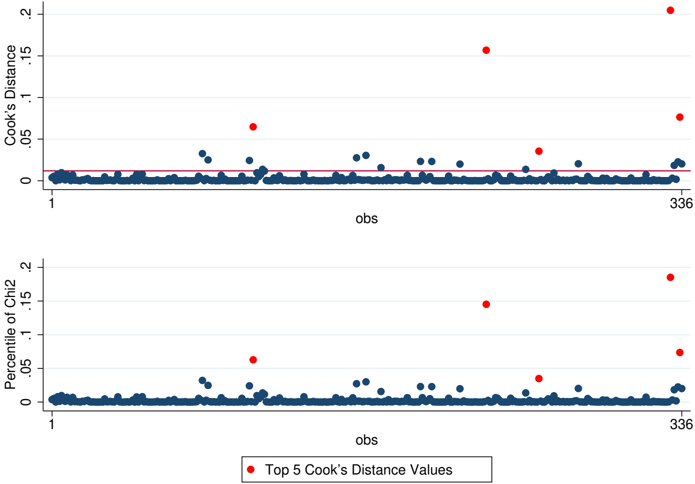
Cook's Distance after xtdata vs cooksd2

```
. local r=e(sigma_u)/e(sigma_e)
. xtdata year fatal spircons unrate yngdrv, ratio(`r´) clear
(theta=0.8750)
. qui reg fatal spircons unrate yngdrv constant, noconstant
. predict cdxtdata, cooksd
```



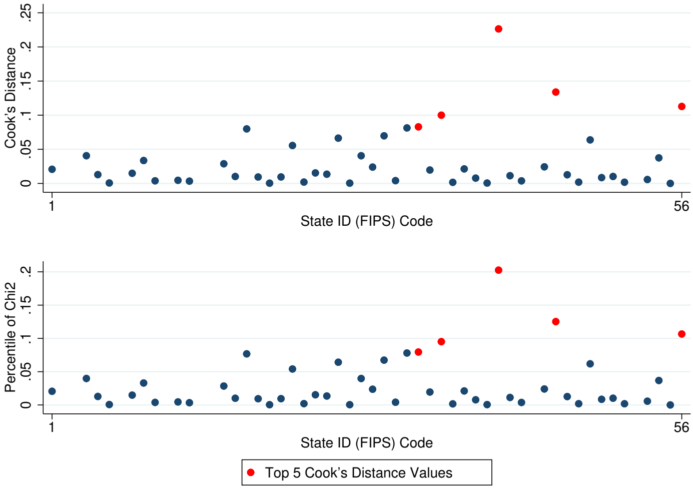
Cook’s Distance for Specific Variables

```
. cooksd2 cdre2, cvars(unrate yngdrv) nocons
Cooks-distance using: unrate yngdrv
```



Cooks's Distance for each State

```
. cooksd2 cdre3, cvars(unrate yngdrv) nocons panel
Cooks-distance using: unrate yngdrv
```



References

Banerjee, M., and E. Frees. 1997. Influence diagnostics for linear longitudinal models. *Journal of the American Statistical Association* 92(439): 999–1005.

Christensen, R., L. Pearson, and W. Johnson. 1992. Case-deletion diagnostics for mixed models. *Technometrics* 34(1): 38–45.

Cook, R. 1977. Detection of influential observation in linear regression. *Technometrics* 19(1): 15–18.