Paper 1461-2014

A SAS® Macro to Diagnose Influential Subjects in Longitudinal Studies

Grant W. Schneider, The Ohio State University; Randall D. Tobias, SAS Institute Inc.

ABSTRACT

Influence analysis in statistical modeling looks for observations that unduly influence the fitted model. Cook's distance is a standard tool for influence analysis in regression. It works by measuring the difference in the fitted parameters as individual observations are deleted. You can apply the same idea to examining influence of groups of observations—for example, the multiple observations for subjects in longitudinal or clustered data—but you need to adapt it to the fact that different subjects can have different numbers of observations. Such an adaptation is discussed by Zhu, Ibrahim, and Cho (2012), who generalize the subject size factor as the so-called degree of perturbation, and correspondingly generalize Cook's distances as the scaled Cook's distance. This paper presents the %SCDMixed SAS® macro, which implements these ideas for analyzing influence in mixed models for longitudinal or clustered data. The macro calculates the degree of perturbation and scaled Cook's distance measures of Zhu et al. (2012) and presents the results with useful tabular and graphical summaries. The underlying theory is discussed, as well as some of the programming tricks useful for computing these influence measures efficiently. The macro is demonstrated using both simulated and real data to show how you can interpret its results for analyzing influence in your longitudinal modeling.

INTRODUCTION

Cook's distance (CD) (Cook 1977) is commonly used to detect influential individual or subsets of observations in linear regression for cross-sectional data. The basic idea is to measure how the predicted values from a regression model change when individual observations are dropped.

CD extends to mixed models for longitudinal data by considering the effect of dropping entire subjects or clusters of observations. However, deleting subsets with different numbers of observations introduces different degrees of perturbation to the model that is fitted to the data. Because the magnitude of CD is positively associated with the degree of perturbation, a large value of the CD for deleting a subset may be due either to a large number of observations or to the presence of influential observations in the subset. To address this ambiguity, Zhu, Ibrahim, and Cho (2012) introduce the dual notions of the scaled Cook's D (SCD) and the degree of perturbation for general maximum likelihood modeling.

The %SCDMixed macro computes SCD and perturbation for models with independent clusters of correlated observations. The macro builds on the existing INFLUENCE option in PROC MIXED, and in fact in order to use the %SCDMixed macro, you need to define an adjunct %SCDModel macro with a specific structure that uses PROC MIXED and the INFLUENCE option to fit your model. The %SCDMixed macro also incorporates matrix computations based on approximation formulae given in Zhu, Ibrahim, and Cho (2012). These calculations are performed using SAS/IML® software.

In the following sections, the concepts of scaled Cook's distance and degree of perturbation are reviewed and the computational details of the %SCDMixed macro are discussed. The arguments for the macro are documented, and finally both the method and the macro are demonstrated with a series of example, using both simulated and actual data.

COOK'S DISTANCE

To determine the degree of influence that the i^{th} data point has on the model, define Cook's D for the i^{th} subject to be

$$CD_i = \frac{(\hat{\beta}_{(-i)} - \hat{\beta})' X'X (\hat{\beta}_{(-i)} - \hat{\beta})}{pMSE},$$

where $\hat{\beta}_{(-i)}$ is the least squares estimate of β with the i^{th} subject deleted, $\hat{\beta}$ is the least squares estimate of β based on the full data, X is the design matrix, p is the number of parameters, and MSE is the mean squared error of the model based on the full data.

Although various guidelines exist for determining a "large" CD, these rules are arbitrary and are not based on distributional properties. Furthermore, in longitudinal studies where the subjects have different numbers of observations, larger subjects will have stochastically larger CD values than smaller subjects. Because of this, CD are not directly comparable for subjects with different degrees of perturbation.

DEGREE OF PERTURBATION

Degree of perturbation is formally defined in Zhu et al. (2012) to be the Kullback-Leibler distance between the fitted probability function and the probability function of a model for characterizing the deletion of the i^{th} subject. In the case of repeated observations with covariates $\mathbf{x}_i$ and covariance matrix $R_i(\boldsymbol{\alpha}) = \sigma_y^2 \mathbf{I}_{m_i} + \sigma_b^2 \mathbf{1}_{m_i}^{\otimes 2}$ for the i^{th} subject (where $\boldsymbol{\alpha} = (\sigma_b^2, \sigma_y^2)^T$), the degree of perturbation for the i^{th} subject is given as

$$P(i \mid M) = 0.5\text{tr}\{\mathbf{x}_i^T R_i(\hat{\alpha})^{-1} \mathbf{x}_i E_\beta[(\beta - \beta_*)(\beta - \beta_*)^T]\},$$

where $E_\beta[(\beta - \beta_*)(\beta - \beta_*)^T]$, the covariance matrix of β can be estimated by the results of the COVB option in the MIXED procedure.

SCALED COOK'S DISTANCE

SCD scales the raw CD values to account for the positive relationship between CD and perturbation. SCD is calculated for the i^{th} subject as

$$SCD_i = \frac{CD_i - \overline{CD}_i}{SD(CD_i)},$$

where $\overline{CD}_i$ and $SD(CD_i)$ are the sample mean and standard deviation of CD for the i^{th} subject based on one of two possible simulation methods.

TWO METHODS FOR SCD

The %SCDMixed macro provides two methods for calculating SCD in mixed model analyses of repeated measures studies:

- A method that calculates a SCD for a given group using the first order approximation described in Zhu et al. (2012), and
- A bootstrap approach that uses Monte Carlo simulation with repeated calls to the MIXED procedure to obtain SCD for many groups simultaneously

For the first method, Zhu et al. (2012) give a first-order approximation to the CD for deleting the i^{th} cluster and suggest that this be used for simulated data to efficiently produce scaling factors for the SCD. This approach is implemented as METHOD=FIRST ORDER in %SCDMIXED, using the SAS/IML matrix language to compute the approximation. The method applies this approximation with a given number of Normal random variables with mean 0 and covariance equal to the covariance matrix for subject *i* to estimate the distribution of the CD. The sample mean and standard deviation allow for the estimation of SCD. This process is repeated for each subject.

The second method implemented in %SCDMixed, specified by METHOD=BOOTSTRAP, is based on simulating the whole mixed model analysis. It begins by replicating the predicted values from %SCDModel (stored in outpm) a given number of times. Next, a given number of bootstrap samples are simulated for each subject based on the covariance matrix with that subject deleted. For each bootstrap replicate, %SCDModel is fit to the predicted values plus simulated errors and CD is stored. Once this process has been repeated for each of the replicates, the sample mean and standard deviation allow for the estimation of SCD.

For large Iteration values, the two methods produce equivalent results. Computationally, the First Order method tends to outperform the Bootstrap method when the number of subjects is fairly small or when Iteration is large. Due to the matrix calculations involved in the first order approximation employed by the First Order method, the Bootstrap method tends to perform better when there are large numbers of subjects and thus larger matrix operations. Regardless of the number of observations per subject, the Bootstrap method outperforms the First Order method with many subjects.

THE %SCDMIXED AND %SCDMODEL MACROS

The %SCDMixed macro requires you to specify the mixed model to be fit by defining the macro %SCDModel, which will act as a substitute for the usual call to PROC MIXED you would use in your analysis. As the macro relies on output from the MIXED procedure, there are options that you must include so that %SCDMixed can locate the required MIXED output. The %SCDModel macro you define must match the following template.

```
%macro SCDModel(Datam=, Effectm=, Outpm=, nSub=);
   proc mixed data=&Datam method=ml;
      class &Effectm << any additional classification variables >>;
      model << dependent variable >> = << fixed effects >> /
         influence(effect=&Effectm est) covb outpm=&Outpm;
      repeated /type=<< within-subject variance model >>
         sub=&Effectm r=1 to &nSub by 1;
%mend SCDModel;
```

Be sure *not* to include a "run;" statement at the end.

THE %SCDMODEL MACRO SYNTAX

The %SCDModel macro must be defined in terms of the parameters DataM, EffectM, OutpM, and nSub so that arguments may be passed from %SCDMixed to %SCDModel as needed. Also, the MODEL statement must contain the option

```
influence(effect=&Effectm est) covb outpm=&Outpm
```

in addition to any other statements you provide, as this output is required by the %SCDMixed macro.

Similarly, the REPEATED statement must include the option

```
r=1 to &nSub by 1
```

in addition to any others you provide.

In the examples that follow, three different specifications of the %SCDModel macro are demonstrated.

THE %SCDMIXED MACRO SYNTAX

The macro generates an output dataset which contains the number of observations, the raw CD, the degree of perturbation, the SCD, and the simulated P-value for the SCD for each subject. The macro also produces plots of several measures of influence for the subjects and a table of the influence measures for the most influential subjects.

The %SCDMixed macro has both required and optional input parameters. The input are displayed in Table 1 and their explanations follow.

Parameter	**Required?**
Data	Yes
Effect	Yes
DepVar	Yes
Fixed	Yes
Class	Yes
Iteration	No
OutTable	No
Method	No
PlotVar	No
PlotPercentile	No
TablePercentile	No
HighlightedSub	No

Table 1. Parameters for the %SCDMixed Macro

- **Data=**SAS-data-set

  specifies the dataset containing the repeated measures study to be analyzed. This is the dataset that you specify in the PROC MIXED statement's DATA= option for a mixed model analysis.

- **Effect=**variable

  specifies the variable that indexes the subjects in the study, which is specified in the MIXED procedure's REPEATED or RANDOM statement for the mixed model analysis.

- **DepVar=**variable

  specifies the dependent variable in the repeated measures study.

- **Fixed=**variables/options

  specifies the fixed effects and MODEL statement options using the MIXED procedure to analyze the study.

- **Class=**variable

  specifies the class variables in the mixed model analysis.

- **Iteration=**number

  specifies the number of iterations to be performed in the Monte Carlo simulation for computing the SCD. By default, Iteration is 100.

- **OutTable=**SAS-data-set

  specifies the name of the dataset in which to store the output of the macro. By default, the output is stored as scd.

- **Method=**First Order|Bootstrap

  specifies whether to use the First Order or Bootstrap method to approximate SCD. The First Order method is the default.

- **PlotVar=**variable

  specifies the horizontal axis variable for the within-subject plot of repeated measures. Typically this will be the time variable in a longitudinal study. By default, PlotVar is the order of the observations within a subject.

- **PlotPercentile=**number(s)

  specifies the percentiles at which the largest SCD and largest perturbation will be highlighted in the plots. If two numbers are given, the first will be the percentile for SCD and the second will be the percentile for perturbation. If one number is given, this number will be the percentile for both SCD and perturbation. By default, PlotPercentile is 98.

- **TablePercentile=**number(s)

  specifies the percentiles at which the largest SCD and largest perturbation will be included in the printed table. If two numbers are given, the first will be the percentile for SCD and the second will be the percentile for perturbation. If one number is given, this number will be the percentile for both SCD and perturbation. By default, TablePercentile is 90.

- **HighlightedSub=**number(s)

  specifies subjects which you would like to be highlighted on the plots. It may be necessary to increase

  PlotPercentile so that the highlighted subjects are more visible.

EXAMPLE 1 – SIMULATED DATA

This example uses simulated data to explore how to use the %SCDMixed macro to diagnose influential subjects. The following SAS statements simulate 50 subjects with an autoregressive covariance model at 200 time points per subject. To account for the fact that observations on a subject are often unequally spaced throughout time, keep a subset of 24 observations for each subject. This example will ply the %SCDMixed macro on this data and several variations, constructed to exhibit different kinds of influence.

```
data AR1;
   retain index 0 subject 1;
   do j=1 to 50;
      ytm1=rannor(12);
         do i=1 to 201;
            et=rannor(1);
            yt=0.7*ytm1+et;
            output;
            ytm1=yt;
            index+1;
         end;
         index=0;
         subject+1;
      end;
      keep yt index subject;
run;

data AR1;
   set AR1(where=(index in (0, 5, 10, 15, 20, 25, 35, 45, 55, 65, 75, 100, 120, 150,
   155, 160, 165, 170, 175, 180, 185, 190, 195, 200)));
run;
```

First, introduce noisy outliers to the data by scaling subjects 9, 16, 25, 36, and 49 by a factor of 5.

```
data ScaledAR1;
   set AR1;
   if subject in (9,16,25,36,49) then yt=5*yt;
run;
```

Next, introduce shifted outliers to the data by adding 5 to measurements from subjects 9, 16, 25, 36, and 49.

```
data ShiftedAR1;
   set AR1;
   if subject in (9,16,25,36,49) then yt=5+yt;
run;
```

The perturbation index computed by the %SCDMixed macro is most useful when subjects are not observed at the same points, and thus, perturbation varies from subject to subject. To investigate this situation, uniformly sample a “dropout” time on the observations for each subject.

```
data Cutoffs;
   call streaminit(1);
   do subject = 1 to 50;
      u = rand("Uniform");
      cutoff = ceil(200*u);
      output;
   end;
run;
```

The following dataset will be used to explore perturbation when there are no outliers.

```
data Longitudinal;
   merge Cutoffs(keep=subject cutoff) AR1;
   by subject;
   if index>cutoff then delete;
   drop cutoff;
run;
```

For the final simulated dataset, introduce outliers into the longitudinal data with dropouts. As before, scale subjects 9, 16, 25, 36, and 49 by a factor of 5.

```
data ScaledLongitudinal;
   merge Cutoffs(keep=subject cutoff) ScaledAR1;
   by subject;
   if index>cutoff then delete;
   drop cutoff;
run;
```

For all analyses of this simulated data, the following %SCDModel macro defines the mixed model to be fit. It is based on the true model used to generate the data.

```
%macro SCDModel(Datam=, Effectm=, Outpm=, nSub=);
   proc mixed data=&Datam method=ml;
      class &Effectm;
      model yt = index /
         influence(effect=&Effectm est)            /*Required*/
         covb                                      /*Required*/
         outpm=&Outpm;                             /*Required*/
      repeated /type=sp(exp)(index) sub=&Effectm
         r=1 to &nSub by 1;                        /*Required*/
%mend SCDModel;
```

The following invocation runs the %SCDMixed macro on the original dataset AR1, producing the results shown in Figure 1.

```
%SCDMixed(Data=AR1,
          Effect=subject,
          DepVar=yt,
          Fixed=index,
          Class=subject,
          PlotVar=index,
          PlotPercentile=90 100,
          TablePercentile=80 100);
```

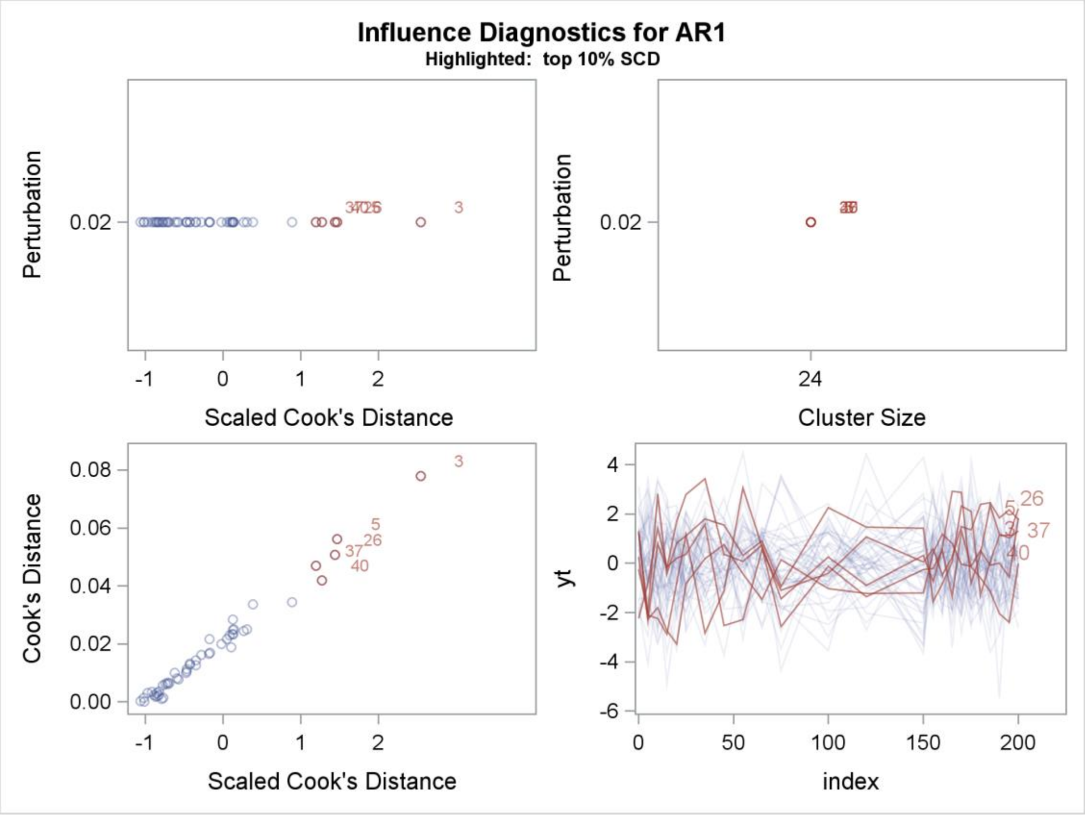


Figure 1. Diagnostic Plots from %SCDMixed Call with No Outliers

The four plots in Figure 1 indicate what the plots will look like when the true model has been fit, each subject is observed at the same time points, and there are no outliers. Since all of the subjects are observed at the same time points and the model is quite simple, all perturbation indices are the same, and thus the top two plots are not very useful. Note that this is the reason that the second arguments of PlotPercentile and TablePercentile are specified to be 100. Also, since SCD is CD adjusted for the scale of perturbation, there is a high degree of correlation in the bottom left plot, with only random variation due to simulation. There does not seem to be an obvious pattern for the highlighted subjects in the bottom right plot.

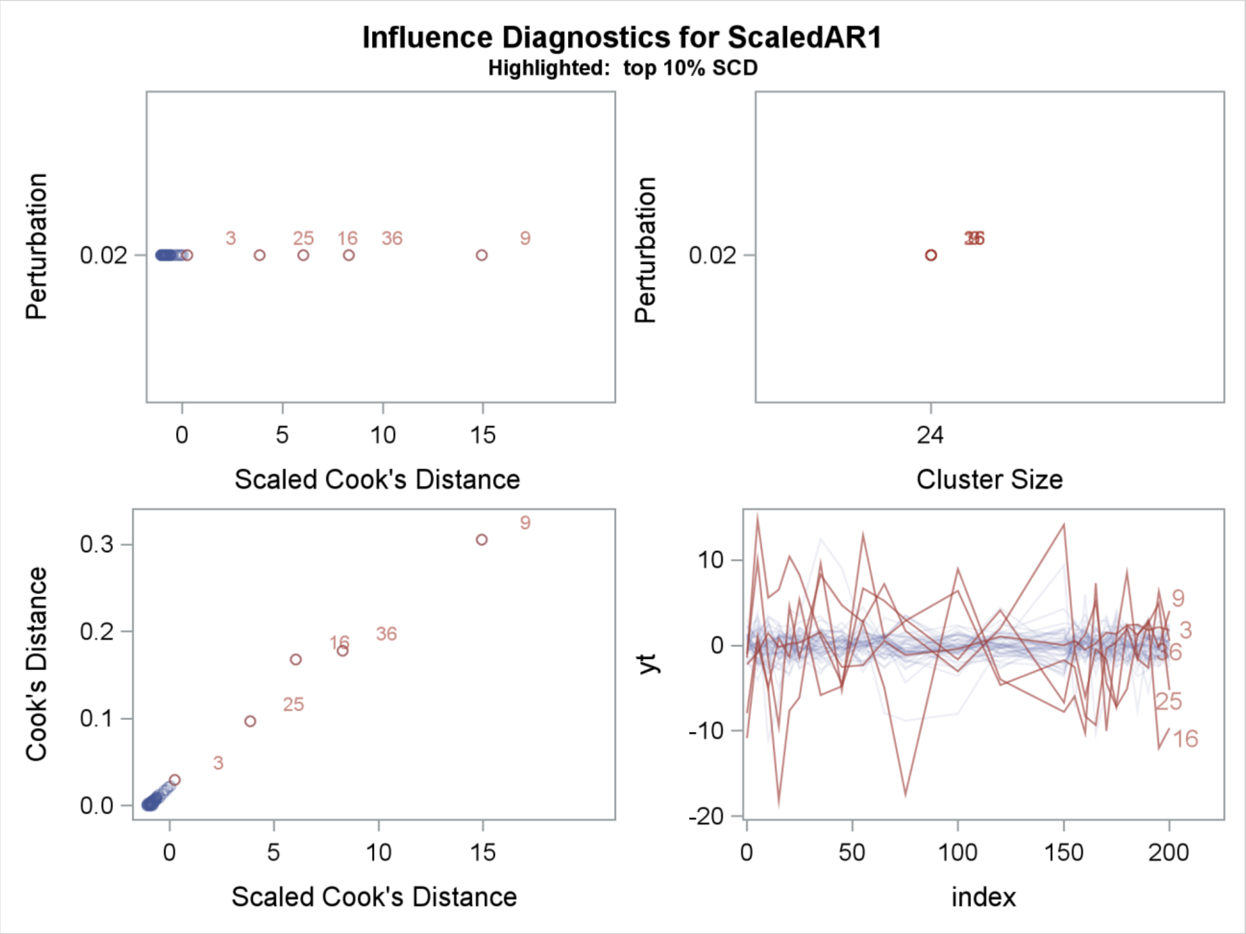


Figure 2. Diagnostic Plots from %SCDMixed Call with Scaled Outliers

For the dataset ScaledAR1 with several noisy outliers, the %SCDMixed analysis shows the outlying observations as distinctive, as you can see in Figure 2. Again, the top two plots provide very little information, as the perturbations are still the same across subjects. You can see that both CD and SCD were able to distinguish 4 out of the 5 outliers. Note the difference in scale for both CD and SCD once the outliers are introduced into the data. While the highlighted points in the lower right plot of Figure 1 appear to be the extreme values of an IID sample, those in Figure 2 appear to be from an entirely different distribution - a clear indication that these can be diagnosed as exceptional subjects, which by construction they indeed are.

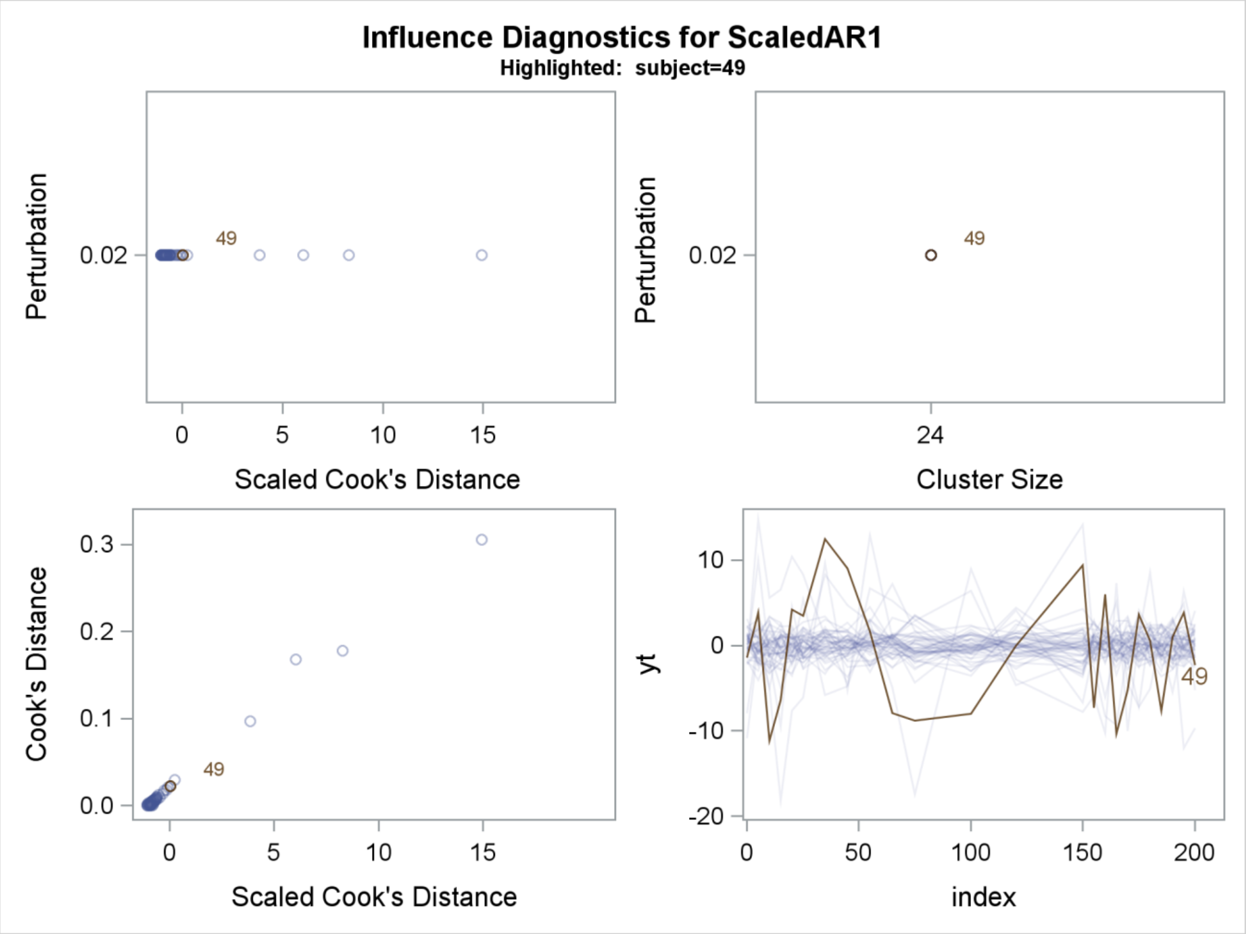


Figure 3. Diagnostic Plots from %SCDMixed Call with Scaled Outliers and Subject 49 Highlighted

To investigate the subject that was missed (49), you can rerun the %SCDMixed macro, now specifying PlotPercentile=100 (so that no subject is highlighted for its SCD or its perturbation) and HighlightedSub=49 options. The results are shown in Figure 3. Although Subject 49 seems to stand out somewhat in the bottom right plot here, it has a small CD and thus a small SCD in this setting. Also, it should be noted that since PlotPercentile was specified to be 90 in Figure 2, the 5 subjects (10 percent) with the largest SCD were highlighted. Subject 49 had the 6th largest SCD and would have been highlighted for PlotPerctile less than or equal to 88.

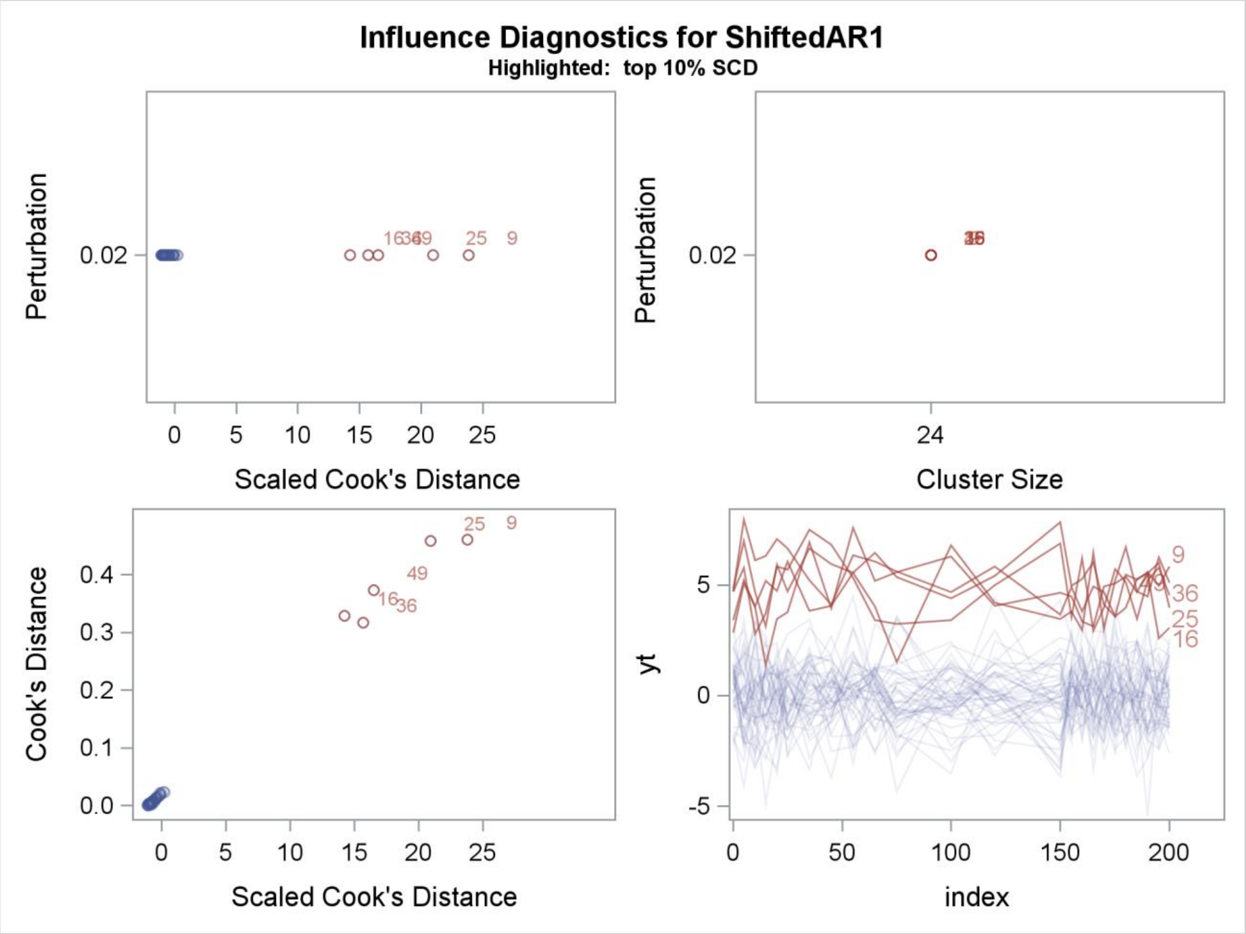


Figure 4. Diagnostic Plots from %SCDMixed Call with Shifted Outliers

For the ShiftedAR1 dataset with shifted outliers, Figure 4 shows that both CD and SCD were able to distinguish all of the 5 true outliers. It does not come as a surprise that CD and SCD perform better for shifted outliers than scaled outliers, as they are measuring the change in parameter estimates due to the deletion of a given subject. Again, note the difference in scale for both CD and SCD once the outliers are introduced into the data.

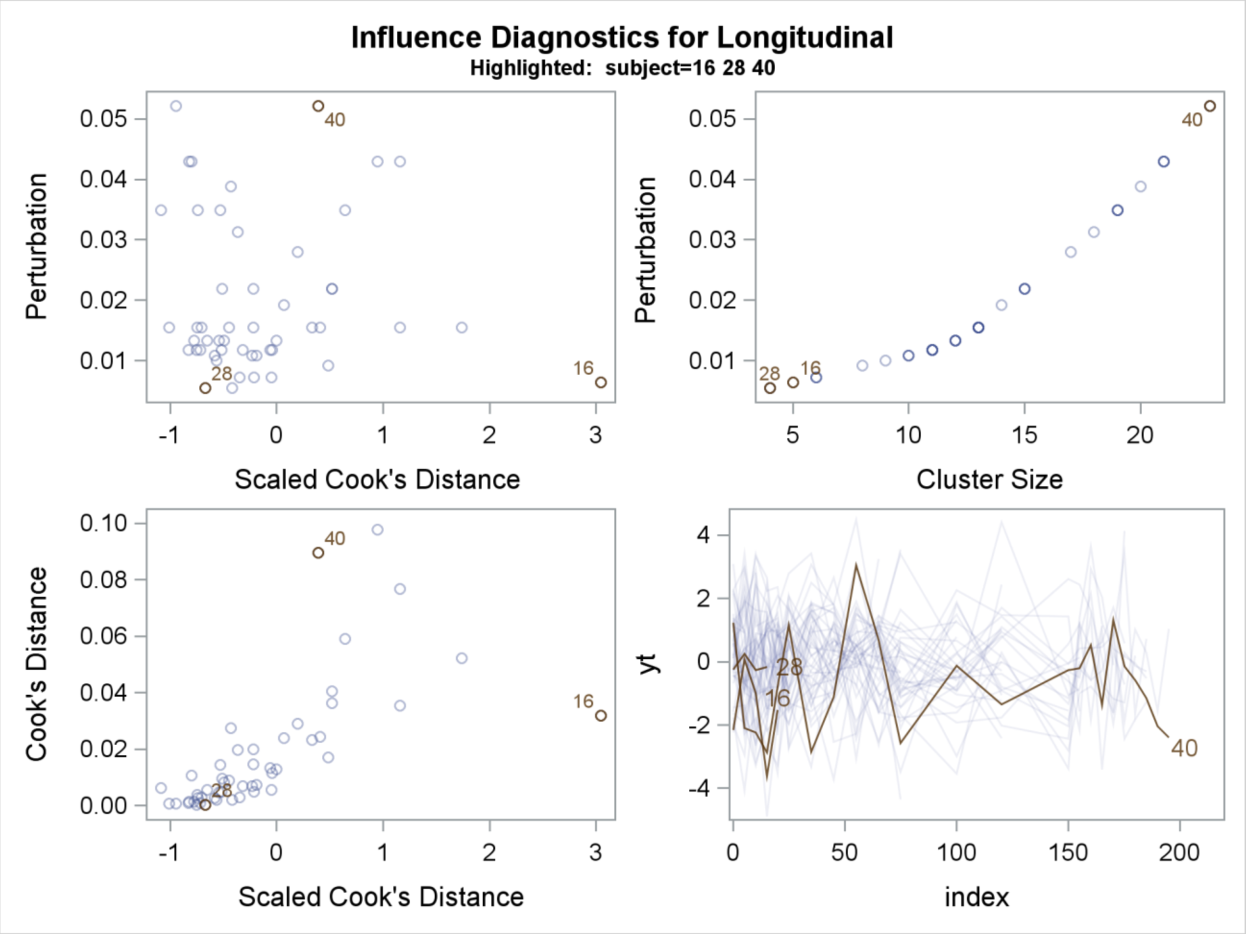


Figure 5. Diagnostic Plots from %SCDMixed Call for Longitudinal Data with No Outliers

Figure 5 is the results of running the %SCDMixed macro on the Longitudinal dataset with PlotPercentile=100 and HighlightedSub=16 28 40. You can see that subject 40, which is observed until time 195, has the second largest CD. However, once the adjustment is made for the fact that this subject causes such a large perturbation to the model, you can see that its SCD is not even in the top 10 largest values. Conversely, subject 16 has a relatively unremarkable CD, but the subject was only observed until time 20. Once the adjustment is made for the small perturbation to the model, you can see that the subject’s SCD is the largest. Subject 28, which is observed until time 15, is highlighted here to demonstrate that a small perturbation will not automatically cause a large scaled CD. While subject 16 makes a relatively large excursion away from zero, subject 28 does not. Subject 16 couples a small number of measurements with large variability, making it one of the most exceptionally influential observations.

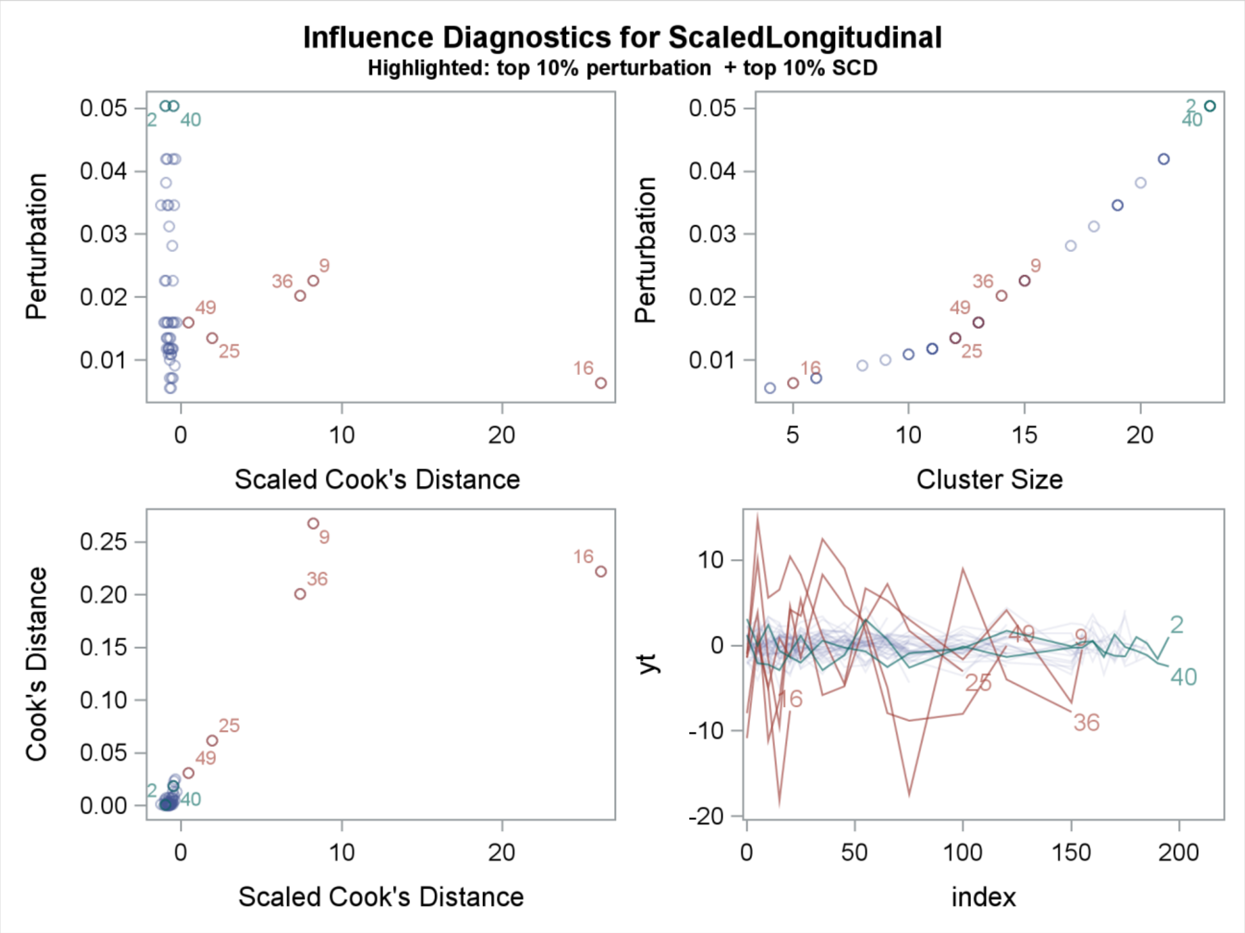


Figure 6. Diagnostic Plots from %SCDMixed Call for Longitudinal Data with Scaled Outliers

As in the scaled outlier case, the %SCDMixed macro is able to identify 4 out of the 5 outliers in the ScaledLongitudinal dataset using SCD. In Figure 6, you can see that while 16 is still the subject with the largest SCD, the value of SCD has increased by a factor of 10. Also note that the largest perturbations correspond exactly to the largest cluster sizes. This may not be the case when the model is misspecified or when the data are missing at random times rather than this longitudinal structure.

Along with the plots previously discussed, the %SCDMixed macro provides table output. In addition to the statistics in the plots, the table provides pseudo p-values for each subject’s SCD as a measure of how exceptional that value is. This pseudo p-value is the percentile for the observed SCD among the simulated SCD values.

Subject	No Outliers	Scaled Outliers	Shifted Outliers	Longitudinal	Long. with Outliers
3	0.04	0.27	0.47	0.43	0.76
5	0.07	0.36	0.38	0.04	0.26
9*	0.07	0	0	0.14	0
12	0.19	0.59	0.74	0.05	0.36
16*	0.27	0	0	0.01	0
25*	0.64	0.02	0	0.63	0.09
36*	0.38	0	0	0.45	0
40	0.05	0.51	0.36	0.23	0.54
49*	0.85	0.34	0	0.93	0.47

Table 2. Pseudo P-Values for Simulated Data

Table 2 displays the pseudo p-values for any subject determined have a significant SCD at the 0.05 level in at least one of the five scenarios considered above. In the scenario where there were no outliers, notice that two of the pseudo p-values are at or below the 0.05 level. It is important to keep multiple comparison issues in mind when interpreting pseudo p-values. As indicated visually, once outliers are introduced, they stand out much more than variation due to random chance. The actual outliers (denoted by an asterisk) introduced in these examples do indeed often have pseudo p-values of zero.

EXAMPLE 2 – HEART RATE DATA

DATA STEP

The data used in this example is the Heart Rate data that was examined in Example 10.4 of Littell et al. (2006). Repeated measurements on the heart rates of patients were taken at five unequally spaced repeated time intervals: 1 minute, 5 minutes, 15 minutes, 30 minutes, and 1 hour. Each patient is subjected to one of three possible drug treatments, a standard drug, a test drug, and a placebo.

```
data HR;
   input patient drug$ basehr hr1 hr5 hr15 hr30 hr1h;
   array hra{5} hr1 hr5 hr15 hr30 hr1h;
   do i = 1 to 5;
      if (i = 1) then minute = 1/60;
      else if (i = 2) then minute = 5;
      else if (i = 3) then minute = 15;
      else if (i = 4) then minute = 30;
      else minute = 60;
      time = minute;
      hours = minute / 60;
      hours1 = hours;
      HR = hra{i};
      output;
   end;
   drop i hr1 hr5 hr15 hr30 hr1h;
   datalines;
   201 placebo 92 76 84 88 96 84
   202 test 54 58 60 60 60 64
      ... more lines ...
   222 test 88 88 98 98 96 88
   223 test 88 88 96 88 88 80
   224 placebo 88 78 84 64 68 64
   232 standard 78 72 72 78 80 68
   ;
```

%SCDMODEL SPECIFICATION

The %SCDModel macro below defines the same model fit in Littell et al. (2006). The options which are required by %SCDMixed are again denoted as such.

```
%macro SCDModel(Datam=, Effectm=, Outpm=, nSub=);
   proc mixed data=&Datam method=ml;
      class drug hours &Effectm;
      model hr = drug basehr / noint
         influence(effect=&Effectm est)                     /*Required*/
         covb                                               /*Required*/
         outpm=&Outpm;                                      /*Required*/
      repeated hours /type=sp(exp)(hours) sub=&Effectm
         r=1 to &nSub by 1;                                 /*Required*/
%mend;
```

%SCDMIXED CALL

Use the following %SCDMixed call to diagnose the influence of the various subjects on this analysis.

```
%SCDMixed(Data=HR,
          Effect=patient,
          DepVar=hr,
          Fixed=drug basehr /noint,
          Class=drug hours patient,
          PlotVar=hours,
          PlotPercentile=80,
          TablePercentile=75);
```

RESULTS AND DISCUSSION

The results are shown in Figure 7and Table 3. Note that the subjects are all observed at the same time points and thus all clusters have the same size. However, due to the fact that the subjects have different values in the design matrix *X*, perturbation is not uniquely determined by cluster size.

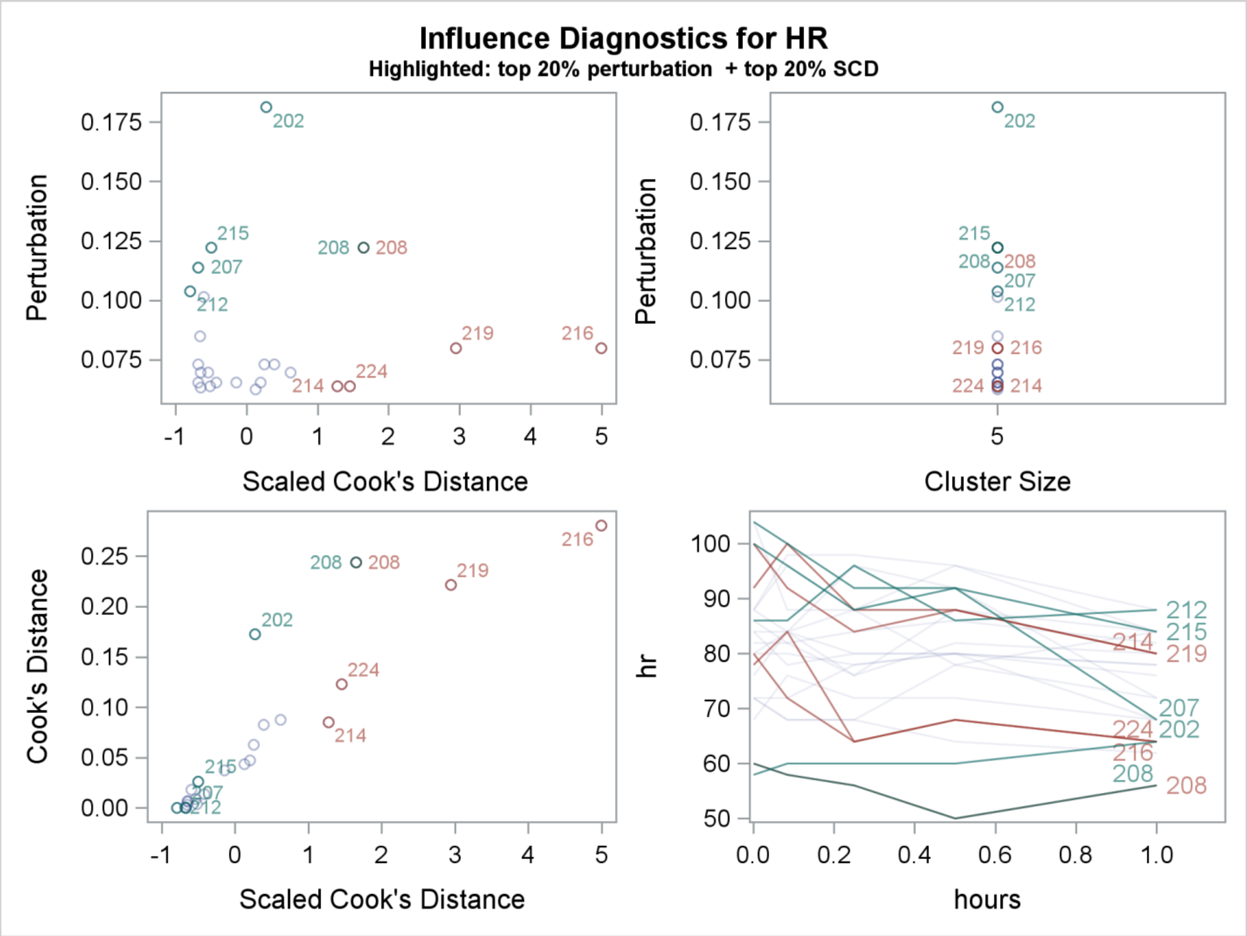
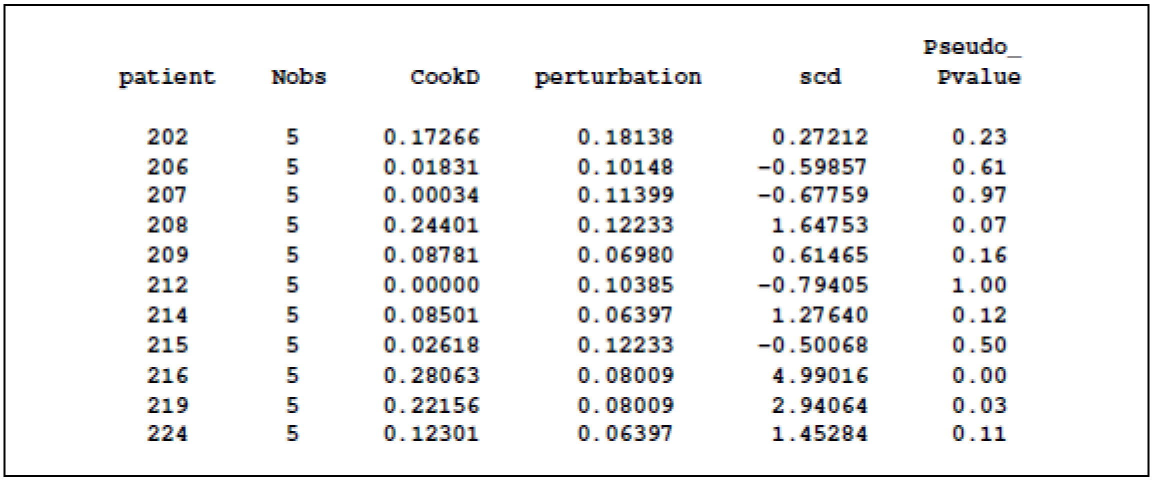


Figure 7. Diagnostic Plots from %SCDMixed Call



```
                                                                  Pseudo_
   patient     Nobs       CookD    perturbation        scd         Pvalue

     202         5      0.17266       0.18138       0.27212         0.23
     206         5      0.01831       0.10148      -0.59857         0.61
     207         5      0.00034       0.11399      -0.67759         0.97
     208         5      0.24401       0.12233       1.64753         0.07
     209         5      0.08781       0.06980       0.61465         0.16
     212         5      0.00000       0.10385      -0.79405         1.00
     214         5      0.08501       0.06397       1.27640         0.12
     215         5      0.02618       0.12233      -0.50068         0.50
     216         5      0.28063       0.08009       4.99016         0.00
     219         5      0.22156       0.08009       2.94064         0.03
     224         5      0.12301       0.06397       1.45284         0.11
```

Table 3. Output from %SCDMixed Call

Graphically, patients 202, 208, 216, and 219 in Figure 7 appear to stand out. To take a closer look at these three patients, specify that only these four subjects should be highlighted by re-running the %SCDMixed analysis with PlotPercentile=100 and HighlightedSub=202 208 216 219, which produces Figure 8.

```
%SCDMixed(Data=HR,
          Effect=patient,
          DepVar=hr,
          Fixed=drug basehr /noint,
          Class=drug hours patient,
          PlotVar=hours,
          PlotPercentile=100,
          TablePercentile=50,
          HighlightedSub=202 208 216 219);
```

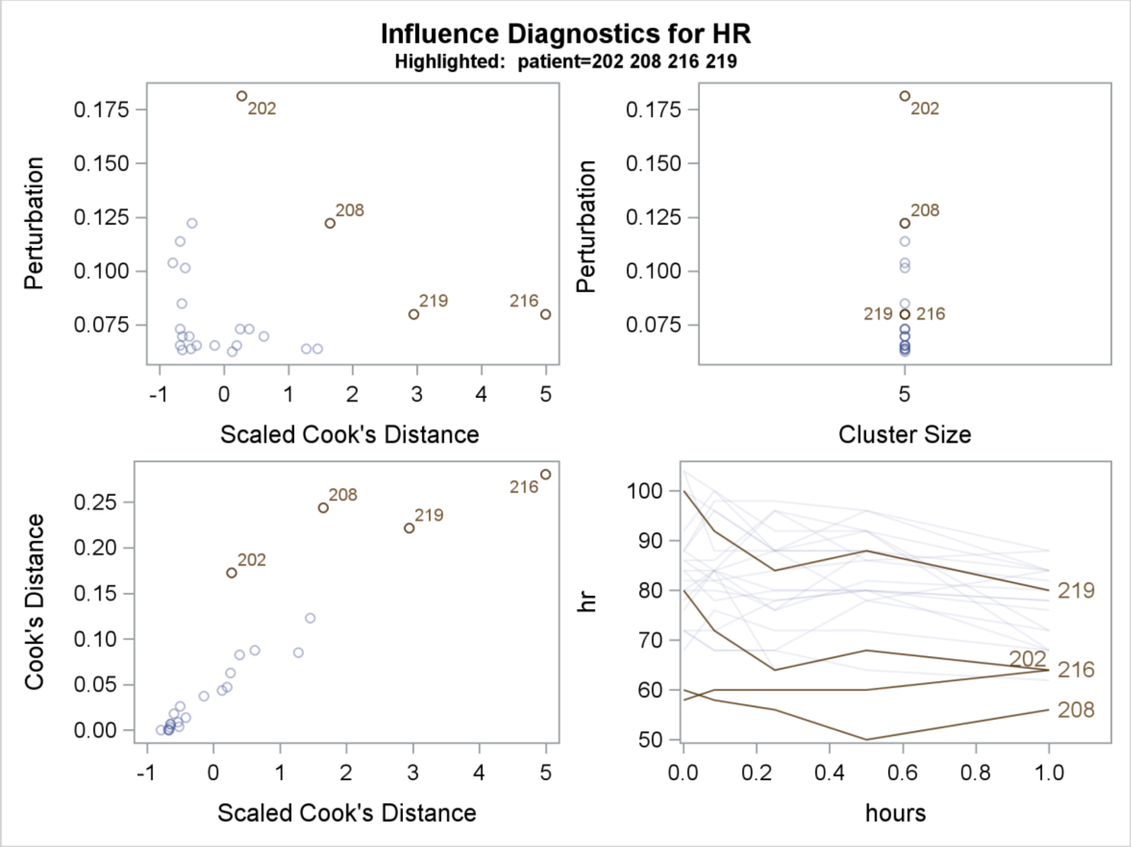


Figure 8. Diagnostic Plots from %SCDMixed Call

In Figure 8, Subjects 202 and 208 appear to be shifted downward in the lower right plot, which leads to these two subjects having fairly large Cook's distances. However, since these two have the largest perturbations, their scaled Cook's distances are relatively modest. Subjects 216 and 219, which are not clearly distinguishable in the lower right plot, have smaller perturbations and large Cook's distances. Therefore, they have the two largest scaled Cook's distances.

EXAMPLE 3 – YALE INFANT DATA

DATA STEP

The Yale infant growth data was examined in Zhu et al. (2012). The data were originally collected to study whether cocaine exposure during pregnancy leads to the maltreatment of infants after birth. The dataset contains observations of 298 children, with the number of observations per child varying from 2 to 30. The following SAS statements create the dataset Infants which contains the variable InfantID, to index the subjects, the variable Intercept, which is just a column of 1s, 9 specially constructed explanatory variables x1–x9, and a response variable Weight, which measures the weight (in kilograms) of the i^{th} subject on the j^{th} visit.

```
data Infants;
   input InfantID Gender GestAge nObs @@;
   do iObs = 1 to nObs;
      input VisitAge Weight @@;
      VisitSpl1 = max(0,VisitAge - 60);
      VisitSpl2 = max(0,VisitAge - 120);
      VisitSpl3 = max(0,VisitAge - 200);
      VisitSpl4 = max(0,VisitAge - 490);
      output;
   end;
cards;

   ... more lines ...

   298 0 2.0 5 0 1.210 45 2.400 78 3.000 78 3.000 109 3.800
;
```

%SCDMODEL SPECIFICATION

The following statements specify the %SCDModel macro to perform a repeated measures analysis with the MIXED procedure using a Compound Symmetry covariance structure. The CLASS statement declares the variable InfantID to be a classification variable. This is necessary because InfantID is also specified in the EFFECT= suboption of the MODEL statement, and the EFFECT= specification must contain only classification variables. The options which are required by %SCDMixed are denoted as such.

```
%macro SCDModel(Datam=, Effectm=, Outpm=, nSub=);
   proc mixed data=&Datam method=ml;
      class &Effectm;
      model Weight = VisitAge VisitSpl2 VisitSpl3 GestAge GestAge*VisitAge
                     GestAge*VisitSpl1 GestAge*VisitSpl4 Gender*VisitAge
                     Gender*VisitSpl2/
         influence(effect=&Effectm est)                          /*Required*/
         covb                                                    /*Required*/
         outpm=&Outpm;                                           /*Required*/
      repeated/type=cs subject=&Effectm
         r=1 to &nSub by 1;                                      /*Required*/
%mend;
```

%SCDMIXED CALL

Here is the appropriate %SCDMixed call for diagnosing influence in this data.

```
%SCDMixed(Data=Infants,
          Effect=InfantID,
          DepVar=weight,
          Fixed=VisitAge VisitSpl2 VisitSpl3 GestAge GestAge*VisitAge
                GestAge*VisitSpl1 GestAge*VisitSpl4 Gender*VisitAge
                Gender*VisitSpl2,
          Class=InfantID,
          PlotVar=VisitAge,
          TablePercentile=98);
```

RESULTS AND DISCUSSION

From Figure 9 and Table 4, subjects 269 and 274 appear to be the most noteworthy subjects, as they have the largest perturbation and scaled Cook's distance, respectively. You can see that subject 269 has the largest perturbation although its cluster size is relatively small. As noted in Zhu et al. (2012), this seems to be due to the fact that this infant was older than most others during visits. The modest perturbation and fairly large Cook's distance for subject 274 cause it to have the largest scaled Cook's distance.

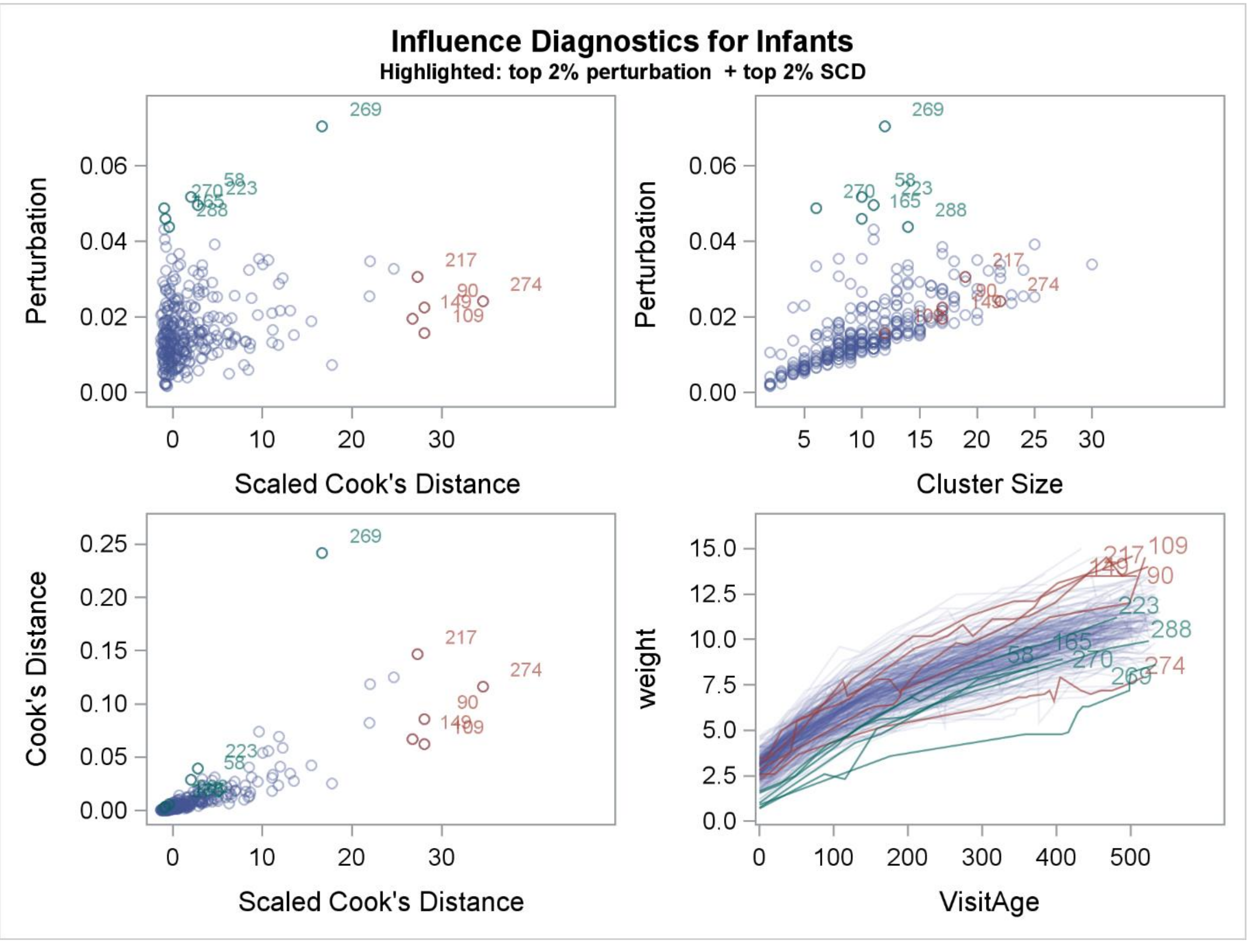
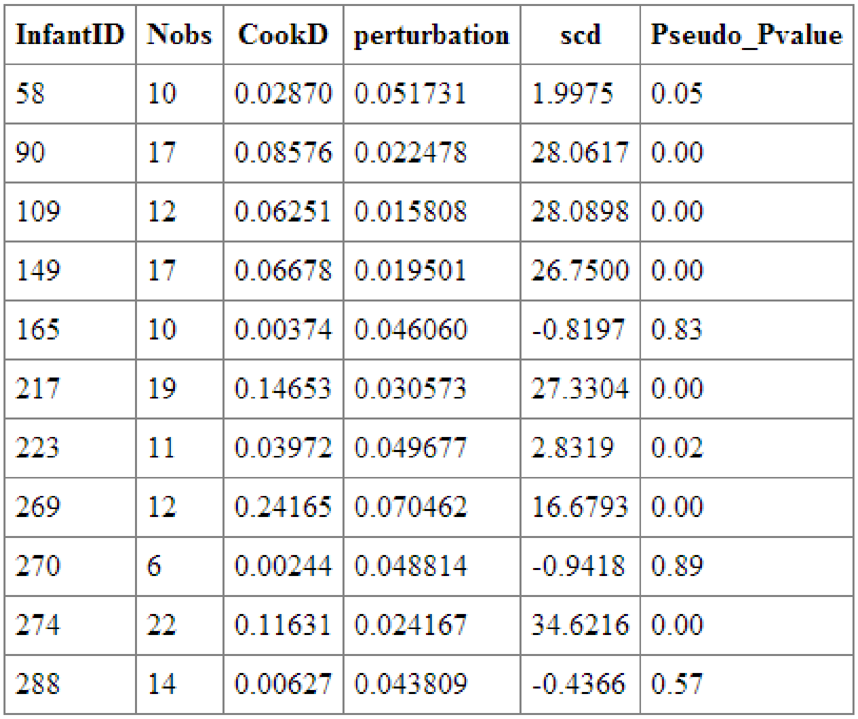


Figure 9. Diagnostic Plots from %SCDMixed Call



InfantID	Nobs	CookD	perturbation	scd	Pseudo_Pvalue
58	10	0.02870	0.051731	1.9975	0.05
90	17	0.08576	0.022478	28.0617	0.00
109	12	0.06251	0.015808	28.0898	0.00
149	17	0.06678	0.019501	26.7500	0.00
165	10	0.00374	0.046060	-0.8197	0.83
217	19	0.14653	0.030573	27.3304	0.00
223	11	0.03972	0.049677	2.8319	0.02
269	12	0.24165	0.070462	16.6793	0.00
270	6	0.00244	0.048814	-0.9418	0.89
274	22	0.11631	0.024167	34.6216	0.00
288	14	0.00627	0.043809	-0.4366	0.57

Table 4. Output from %SCDMixed Call

CONCLUSION

You can use the %SCDMixed macro to obtain useful influence diagnostics for longitudinal data in addition to the standard diagnostics output by the MIXED procedure. The macro provides the flexibility to fit a variety of longitudinal models and to produce custom graphics and tables that are tailored to your particular situation.

REFERENCES

- Cook, R. D. (1977), “Detection of Influential Observations in Linear Regression,” Technometrics, 19, 15–18.
- Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., and Schabenberger, O. (2006), SAS for Mixed Models, 2nd Edition, Cary, NC: SAS Institute Inc.
- Zhu, H., Ibrahim, J. G., and Cho, H. (2012), “Perturbation and Scaled Cook’s Distance,” Annals of Statistics, 40, 785–811.

ACKNOWLEDGMENTS

Thank you to Ja-An Lin and Allen McDowell for their early work on developing this macro. Also, thanks to everyone in the Advanced Analytics Department at SAS Institute for their help in developing this work and an overall great experience last summer.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Grant Schneider
The Ohio State University
1958 Neil Avenue, 404 Cockins Hall
Columbus, OH 43210
(614) 292-1567
schneider.393@osu.edu

Randy Tobias
SAS Institute Inc.
Research Drive
Cary, NC 27513
Randy.Tobias@sas.com
www.sas.com