Demo: Looking for Influential Observations Using PROC GLMSELECT and PROC REG

Filename: **st105d02.sas**

In this demonstration, we look for influential observations in the ameshousing3 data set. First, we select a model by stepwise selection, PROC GLMSELECT. We then use PROC REG to generate influence statistics and plots for the selected model and save the plot data to temporary output data sets. We'll reference these data sets in the second part of the demonstration. All of the code for these demonstrations needs to be run in the same SAS session.

PROC GLMSELECT DATA=*SAS-data-set <options>***;**
CLASS *variable(s)***;**
<label:> **MODEL** *dependent = <effects> < / options>***;**
RUN;

PROC REG DATA=*SAS-data-set <options>***;**
MODEL *dependents = <regressors> < / options>***;**
RUN;

1. Open program st105d02.sas.

```
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
         Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;

/*st105d02.sas*/ /*Part A*/
ods select none;
proc glmselect data=STAT1.ameshousing3 plots=all;
        STEPWISE: model SalePrice = &interval / selection=stepwise details=steps select=SL slentry=(
        title "Stepwise Model Selection for SalePrice - SL 0.05";
run;
quit;
ods select all;

ods graphics on;
ods output RSTUDENTBYPREDICTED=Rstud
           COOKSDPLOT=Cook
           DFFITSPLOT=Dffits
           DFBETASPANEL=Dfbs;
proc reg data=STAT1.ameshousing3
         plots(only label)=
              (RSTUDENTBYPREDICTED
               COOKSD
               DFFITS
               DFBETAS);
    SigLimit: model SalePrice = &_GLSIND;
    title 'SigLimit Model - Plots of Diagnostic Statistics';
run;
quit;


/*st105d02.sas*/  /*Part B*/
title;
proc print data=Rstud;
run;

proc print data=Cook;
run;

proc print data=Dffits;
run;

proc print data=Dfbs;
run;

data Dfbs01;
        set Dfbs (obs=300);
run;

data Dfbs02;
```

```
        set Dfbs (firstobs=301);
run;

data Dfbs2;
        update Dfbs01 Dfbs02;
        by Observation;
run;


data influential;
/*  Merge datasets from above.*/
    merge Rstud
          Cook
          Dffits
                  Dfbs2;
    by observation;

/*  Flag observations that have exceeded at least one cutpoint;*/
    if (ABS(Rstudent)>3) or (Cooksdlabel ne ' ') or Dffitsout then flag=1;
    array dfbetas{*} _dfbetasout: ;
    do i=2 to dim(dfbetas);
         if dfbetas{i} then flag=1;
    end;

/*  Set to missing values of influence statistics for those*/
/*  that have not exceeded cutpoints;*/
    if ABS(Rstudent)<=3 then RStudent=.;
    if Cooksdlabel eq ' ' then CooksD=.;

/*  Subset only observations that have been flagged.*/
    if flag=1;
    drop i flag;
run;

title;
proc print data=influential;
    id observation;
    var Rstudent CooksD Dffitsout _dfbetasout:;
run;
```

In Part A, the PROC GLMSELECT step uses the stepwise selection method to automatically select a model. The specified selection criterion is significance level, and both the entry and stay criteria are 0.05.

In addition to the output that this step produces, it also automatically creates the macro variable _GLSIND, which stores a list of effects selected by PROC GLMSELECT. You can then reference the list as &_GLSIND in subsequent statements.

In this case, we want to create the list of effects in the _GLSIND macro variable, but we don't need to see the PROC GLMSELECT output. So, before the step, we add the statement ODS SELECT NONE, which suppresses the output, and we add ODS SELECT ALL at the end of the step to make sure that we get the output from the next step we run.

2. Submit Part A, beginning with the %LET statement, up to and including the ODS SELECT ALL statetment.

3. Check the SAS log.

   The log shows that the step processed. PROC GLMSELECT automatically saves the list of the chosen model effects as the _GLSIND macro variable. We could see the values of this macro variable in the log by submitting %put &_glsind;, but we'll see the model effects in the PROC REG output.

4. Let's look at our PROC REG step in Part A. The plots (only label)= option generates only the specified plots, and labels extreme observations in the plot. If we were to include an ID statement, SAS would use the value of the ID variable as the label. In this case, the extreme observations will be labeled with the observation numbers. In our MODEL statement, we specify SalePrice as the response variable, and for the predictor variables, we reference the macro variable to specify the list of effects. Notice that we've included a label, SigLimit, which is short for Significance Limit, in front of the MODEL statement. Later, we'll output this model into new data sets, which will include this SigLimit model label. Above the PROC REG step we include an ODS OUTPUT statement. This statement, along with the PLOTS= option, writes the data from the influence plots into separate output data sets. Notice that some of the plot objects that we reference here have slightly different names than the ones that we use in PROC REG. For example, to reference the data that creates the COOKSD plot, we reference COOKSDPLOT rather than just COOKSD.

5. Submit the rest of the Part A code, beginning with ODS GRAPHICS On.

6. Review the output.

   Here's the final model that was selected by PROC GLMSELECT. This is the same model as the previous stepwise selection demonstration in Lesson 4. Here, we'll focus on only the influence statistics.

   In the Diagnostic Plots, the R Student by Predicted plot shows 16 observations beyond two standard errors from the mean of 0, and they're identified with their observation numbers. Remember that RStudent residuals are assumed to be normally distributed and therefore, you expect approximately 5% of values to be beyond two standard errors from the mean. The fact that you have 16 beyond two standard errors is no cause for concern, because 5% of 300 is 15 expected observations. Observation 123 is the largest outlier, and it's well separated from the other points. We might want to recheck this observation.

   Next is the Cook's D plot, which is a needle plot. Cook's D is a measure of the simultaneous change in all parameter estimates when an observation is deleted.

The horizontal line shows the Cook's D cutoff boundary. The plot labels the 21 influential points that are above the cutoff.

Let's look at the DFFITS plot. Recall that DFFITS measures the impact that an observation has on the predicted value. This plot flags several observations as influential points based on DFFITS. At this point, it might be helpful to see which parameters these observations might influence the most using DFBETAS information. The DFBETAS plot is a panel plot, which contains one plot for each parameter. In this case, SAS created two panels. Each plot labels the points that are potentially influencing the parameter that's associated with each of the predictor variables.

Detection of outliers or influential observations with plots is convenient for relatively small data sets, but for larger data sets, it can be difficult to discern one observation from another. One method for extracting only the influential observations is to write the output of the ODS plots into data sets and then subset the influential observations. We'll do this in the next demonstration.

Close