McMaster University

Department of Mathematics and Statistics

STATISTICS 3A03: Applied Regression Analysis with SAS

Fall 2017

SAS Lab 7. October 31 – November 3, 2017

Influence Measures

In this week's lab we shall learn how to use SAS to get some of the measures of influence of individual points that we are currently examining in class. The main measures we are examining are:

1. The Leverages, p_{ii}.
2. Cook's Distances, C_i.
3. Welsch & Kuh measure DFITS$_i$.
4. The Hadi measure H_i.

The first three of these are available directly in SAS. The fourth needs to be calculates using saved output from the regression model. As with the fitted values and the three types of residuals, there are keywords for the leverages (`H.`), Cook's Distances (`Cookd.`) and the Welsch and Kuh measure (`DFFITS.`). All of these can also be stored in an output matrix using the same names without the period at the end. For plots of the measures against the observation number we get the observation number using the keyword `Obs.` which can be included in any `Plot` statement in `PROC REG`.

Here is a demonstration using the `Examination.txt` data used above.

```
PROC REG Data=S3A3.Exams plots=none;
  Model F=P1 P2;
  Plot H.*Obs.; /* The Leverages against the Observation number. */
  Plot Cookd.*Obs.; /* Cook's Distance against the Observation number. */
  Plot DFFITS.*Obs.; /* DFITS against Observation number */
  Output Out=ExamsOut
         Predicted=Fitted
         Residual=Resid_raw
         Student=Resid_student
         H=Leverage
         Cookd=Cook_Dist;
run;
```

```
 quit;
```

The Hadi Influence Measure

To construct the Hadi measure we need to get the Potential and Residual functions and their sum to get the Hadi measure. Since we would like to plot this measure against the observation number, we also need to include a column giving the observation number in our dataset. Here is the code to do that.

```
PROC REG Data=S3A3.Exam plots=none;
    Model F=P1 P2;
    Title "Examination Data Model";
    * We will now save the ANOVA table in a new dataset called exam_anova;
    ODS output anova=exam_anova;
    * Create another output dataframe containing the usual values used in diagnostics;
    Output Out=Exam_Out
        Predicted=Fitted
        Residual=Resid_raw
        Student=Resid_student
        H=Leverage
        Cookd=Cook_Dist;
run;

Data exam_anova;
    set exam_anova;
    * We now want to save the SSE in a scalar variable which we will call sse;
    If source='Error' then call symput ('sse', ss);
    * Similarly we extract and save the model degrees of freedom and save it as p;
    If source='Model' then call symput ('p', df);
run;

* When referring to the variables defined above we need to preced their name;
* with an ampersand (&) to tell SAS they are scalar variables rather than;
* a dataset or column of a dataset;

Data Exam_out;
    set Exam_out;
    * The value _N_ gives the row number currently being examined in a data step;
    id=_N_;
    * Now we calculate the values of the di;
    d=resid_raw/sqrt(&sse);
    * The potential and residual functions are then found using the definitions;
    potential=Leverage/(1-Leverage);
```

```
    residual_fun=(&p+1)*d**2/((1-Leverage)*(1-d**2));
    Hadi=potential+residual_fun;
run;
```

We have now created a new dataset which includes the potential and residual functions as well as the Hadi measure. It also contains a column (called `id`) with the observation numbers and all of the output saved from `PROC REG` above. Using this we can plot the Hadi measure (and any of the other influence measures) against the observation number.

```
PROC GPLOT Data=Exam_out;
    Plot Hadi*id;
    Plot Cook_Dist*id;
    Plot Leverage*id;
run;
```

We can also construct the Potential-Residual Plot from our calculations above.

```
PROC GPLOT Data=Exam_Out;
    Plot potential*residual_fun;
run;
```

The Residual Plus Component Plot

The final plot that we saw in class is the Residual plus Component plot. Once again this is not directly available in SAS but can be created from using the following code (and the earlier stored output)

```
PROC REG Data=S3A3.Exam;
    Model F=P1 P2;
    * This time we need the parameter estimates table;
    ods output ParameterEstimates=exam_ests;
run;

DATA exam_ests;
    set exam_ests;
    * Extract and save the estimate of beta1;
    If Variable='P1' then call symput ('Beta1', Estimate);
    * Extract and save the estimate of beta2;
    If Variable='P2' then call symput ('Beta2', Estimate);
run;

DATA Exam_Out;
    Set Exam_Out;
    * Calculate the residual plus component values;
```

```
    res_comp1=resid_raw+P1*&Beta1;
    res_comp2=resid_raw+P2*&Beta2;
run;

PROC GPLOT Data=Exam_Out;
    Plot res_comp1*P1;
    Plot res_comp2*P2;
run;
```

Exercises

1. For the data in `CSData.txt` consider a model with response `GPA` and three predictor variables `HSM`, `HSS` and `HSE`. construct the plots of the Hadi measure, the potential-residual plot and the residual plus component plots.

2. For the New York Rivers data which can be found in `NYRivers.txt`, consider the model given in Equation (4.18) in your textbook which has response variable the mean nitrogen content (`Nitrogen`) and covariate the percentage of land area in commercial or industrial usage (`ComIndl`). Construct the three plots of the influence measures against index as given in Figure 4.7 of your textbook. Also construct the Potential-Residual Plot in Figure 4.8.