McMaster University

Department of Mathematics and Statistics

STATISTICS 3A03: Applied Regression Analysis with SAS

Fall 2017

SAS Lab 12

Note that there is not an actual SAS lab this week but this material is provided for you to learn the SAS code related to the final week of lectures on your own.

This Lab will cover variable selection using SAS `PROC REG`.

All Possible Regressions

SAS will do all possible regressions for a given set of covariates and produce model selection criteria for each regression. The most common used method is to calculate the R^2 value for each model. SAS will then output a list of the models, sorted by model size and R^2 value. This allows us to easily select the model with the largest R^2 for a given number of covariates. We could then fit these models separately and compare them using other model selection criteria. Model selection is specified by using the `selection=` option in the `Model` statement. Here is the code to do this using the `CSData.txt` dataset.

```
PROC REG Data=S3A3.CSdata plots=none;
    Model GPA=Sex HSM HSS HSE SATM SATV / Selection=RSquare;
run;
```

Two other criteria are available for all possible regressions. They are `selection=ADJRSQ` to use the adjusted R^2 criteria which adjusts the regular R^2 for the number of parameters, and `selection=CP` to use Mallow's C_p criterion discussed in the lecture notes. For each of these methods, the models are sorted according to the selection criterion used. In the case of the adjusted R^2 they are in decreasing order and for Mallow's C_p they are in increasing order. For both of these methods, the regular R^2 as well as the selection criterion requested is output. Since the criteria are different, there is no reason that they will give the same "best" model. They do not in the case of the computer science dataset.

```
PROC REG Data=S3A3.CSdata plots=none;
    Model GPA=Sex HSM HSS HSE SATM SATV / Selection=CP;
run;

PROC REG Data=S3A3.CSdata plots=none;
    Model GPA=Sex HSM HSS HSE SATM SATV / Selection=AdjRsq;
run;
```

Stepwise Procedures

While the All Possible Regressions approach is quite fast in SAS for the datasets we are using in this course, it does become slow when the possible number of covariates is large. Also since a line of output is produced for every model, we may need to sift through many thousand such models. For this reason, most model selection is often done using a stepwise procedure

as described in class. Since these methods do not actually evaluate all possible models, they are generally faster but may not find the best model in general. SAS will allow us to do the 3 stepwise methods seen in class: Backward Elimination, Forward Selection and Stepwise Regression. See the class notes and textbook chapter 11 for details on these methods.

These are requested using the `Selection=` option in the `Model` statement. The relevant commands are `selection=backward`, `selection=forward` and `selection=stepwise` respectively. For backward elimination we need to give a value for the significance level at which variables remain in the model. Only variables with significance level greater than this threshold are considered for removal at any step. The significance level is specified using the `SLStay=` option in the `Model` statement. Similarly we need to specify a `SLEntry=` significance level for forward selection. Only variables whose significance level as an added variable in the model is less than this threshold are considered for entry into the model at any step. For stepwise regression we need to specify both of these levels. Usually both `SLStay` and `SLEntry` are set to the same value.

Here is code for applying these methods to the same dataset as before.

```
PROC REG Data=S3A3.CSData plots=none;
    Model GPA= Sex HSM HSS HSE SATM SATV
          / Selection=Forward SLEntry=0.05;
run;

PROC REG Data=S3A3.CSData plots=none;
    Model GPA= Sex HSM HSS HSE SATM SATV
          / Selection=Backward SLStay=0.05;
run;

PROC REG Data=S3A3.CSData plots=none;
    Model GPA= Sex HSM HSS HSE SATM SATV
          / Selection=Stepwise SLEntry=0.05 SLStay=0.05;
run;
```

In this case all three methods produce the same model but that does not happen in general. Note, however, that the model selected is not the same as the model with the best adjusted-R^2 or Mallows C_p.

Exercise: Show that forward selection and backward elimination give very different models for the simulated dataset in `Mydata2.txt`. Also check which models are ranked the "best" based on the adjusted-R^2 and Mallow's C_p criteria. Which model would you select for this dataset?

Exercise: The data for the example given in your textbook (Section 11.10) is in the file `Supervisor.txt`. Use model selection to find a good model for this dataset. Show that all stepwise methods give the same model but that model is not chosen if you use either the adjusted-R^2 or Mallow's C_p.