Sunny Park

Regression Analysis

December 2, 2012

We have learned several techniques for variable selection in regression analysis, such as All Possible Regression using Cp statistic, Adj-R square, or squared root of MSE and other algorithms such as Forward Selection, Backward Elimination, and Stepwise Regression. In this Lab class, we are going to learn about AIC, BIC criteria and some real example.

(1) AIC (Akaike Information Criterion)

$$AIC = n \ln(\frac{SSE}{n}) + 2p$$

n = no. of observations

p = k+1 = no. of parameters for the trial model (= p*)

SSE = Sum of Square Error for the trial model

>> "The smaller is the better"

(2) BIC (Bayesian Information Criterion)

$$BIC = n \ln(\frac{SSE}{n}) + 2(p+2)q - 2q^2$$

n = no. of observations

p = k+1 = no. of parameters for the trial model (= p*)

q= n MSE(full) / SSE

SSE = Sum of Square Error for the trial model

>> "The smaller is the better"

(3) SAS with Real Example (fuel consumption data)

<US Fuel Consumption Data (fuelcon6.xls 사용)>

-State : State Name

-POP : the population of the state

-AREA : the area of the state (in square miles)

-DRIVERS : the ratio of licensed drivers to private and commercial motor vehicles

(등록차량 대비 면허소지자 비율)

-HWYMILES : the number of federally funded miles of highway

(연방정부 기금으로 만든 고속도로 총 길이)

-GASTAX: the tax on a gallon of gas (휘발유 1갤론당 부여된 지방세)

-INCOME : the average personal income (주민 1인당 평균 소득)

1) 각 독립변수 값의 변화에 따른 Fuel Consumption 의 변화를 그래프로 살펴보아라. (생략)

2) All Possible Regression 을 하되 $R^2, Adj-R^2, C_p$, AIC 를 이용하라.

```
>> Proc Reg;    Id State
      Model  Fuelcon = P A D H G I  / selection = rsquare adjrsq cp aic ;  run;
```

3) 위 2)번 수행에서 Population과 GASTAX 변수를 꼭 회귀식에 포함시키면서 AIC 를 이용해 변수를 선택해 보아라.

```
>> Proc Reg;  ID State ;
        Model   Fuelcon = P G A D H I / include = 2  selection = AIC  ; run;
```

4) Forward Selection을 수행하되 SLE(Significance Level for Entry) 를 0.15 로 사용하라.

```
>> Proc Reg;  ID State;
        Model  Fuelcon =  P G A D H I / selection =forward  SLE=0.15 ; run;
```

5) Backward Elimination을 수행하되 SLS(Significance Level for Stay) 를 0.20 로 사용하라.

```
>>  Proc Reg;  ID State;
        Model  Fuelcon =  P G A D H I / selection =backward  SLS=0.20 ; run;
```

6) Stepwise Regression을 수행하되 SLE=0.15, SLS=0.20 을 사용하라.

```
>>  Proc Reg;  ID State;
        Model  Fuelcon =  P G A D H I / selection =stepwise  SLE=0.15 SLS=0.20 ; run;
```

7) 첨부된 cpplot macro 를 수행하여 Cp 통계량 그래프와 AIC 그래프를 그려보고 최종모형을 선정하여라.

```
>> %cpplot( data= fuel ,
        yvar= fuelcon ,  xvar = POP AREA DRIVERS HWYMILES GASTAX INCOME ,
        gplot = CP  AIC,
        plotchar =  P A D H G I  , cpmax=20 );
```