Stat763 HW10

1. Variables y, x_1, x_2, x_3, x_4, x_5 in Table B2 on page 575 are also stored in file B2.txt.

   (1) Find the best model $y = \beta_0 + \beta_1 x_i + \beta_2 x_j + \epsilon$ by the largest R^2 criterion and its R^2.

   The best model by the largest R^2 criterion is

   $$y = \beta_0 + \beta_1 x_3 + \beta_2 x_4 + \epsilon \text{ with } R^2 = 0.8589.$$

   SAS: `proc reg; model y=x1 x2 x3 x4 x5/selection=rsquare; run;`

   (2) Find the best model $y = \beta_1 x_i + \beta_2 x_j + \epsilon$ by the largest R^2 criterion and its R^2.

   The best model by the largest R^2 criterion is

   $$y = \beta_1 x_1 + \beta_2 x_2 + \epsilon \text{ with } R^2 = 0.9955.$$

   SAS: `proc reg; model y=x1 x2 x3 x4 x5/noint selection=rsquare; run;`

   (3) Find the best model $y = \beta_0 + \beta_1 x_i + \beta_2 x_j + \epsilon$ by forward selection.

   The best model by forward selection is

   $$y = \beta_0 + \beta_1 x_4 + \beta_2 x_3 + \epsilon$$

   SAS: `proc reg; model y=x1 x2 x3 x4 x5/selection=forward; run;`

   (4) Find the best model with intercept by the smallest Cp criterion and its Cp.

   The best model by the smallest Cp criterion is

   $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon \text{ with } Cp = 5.7661.$$

   SAS: `proc reg; model y=x1 x2 x3 x4 x5/selection=cp; run;`

   (5) Find the best model without intercept by the smallest Cp criterion and its Cp, AIC, SBC and BIC.

   The best model by the smallest Cp criterion is

   $$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon \text{ with } Cp = 3.0132$$

   $$\text{and } AIC = 133.8447, \ SBC = 139.3130, \ BIC = 137.4198.$$

   SAS: `proc reg; model y=y=x1 x2 x3 x4 x5/noint selection=cp AIC SBC BIC; run;`

2. Statistics table

```
    R-square    n=52
 ----------------------------
 1   0.2021      x3
 1   0.0857      x2
 1   0.0001      x1
 ---------------------------
 2   0.2596     x2 x3
 2   0.2216     x1 x3
 2   0.0859     x1 x2
-----------------------------
 3   0.2757     x1 x2 x3
```

is produced by

```
proc reg;
   model y=x1 x2 x3/selection=rsquare;
   run;
```

(1) Find C_p for the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_3 + \epsilon$.
Hint: Use notations such as SSE(1, x1, x3), MSE(1, x1, x2, x3), $R^2(1, x1, x2, x3)$.

$$
\begin{aligned}
C_p &= 2p - n + \frac{SSE_{(p)}}{MSE} = 2 \cdot 3 - 52 + \frac{SSE(1, x1, x3)}{SSE(1, x1, x2, x3)} \times (52 - 4) \\
&= -46 + \frac{SSTO - SSM(1, x1, x3)}{SSTO - SSM(1, x1, x2, x3)} \times 48 = -46 \frac{1 - R^2(1, x1, x3)}{1 - R^2(1, x1, x2, x3)} \times 48 \\
&= -46 + \frac{1 - 0.2216}{1 - 0.2757} \times 48 = 5.5853.
\end{aligned}
$$

(2) Find the first predictor selected by

```
proc reg;
    model y=x1 x2 x3/selection=Forward SLE=1.00;
    run;
```

and explain why.

The first predictor selected is x_3
In each of 3 models with 1 predictor, the test on the hypothesis that the coefficient of the predictor is zero is performed. The one that produces smallest p-value$<$ 1 is selected. But this test is the global F-test in ANOVA table. The smallest p-value criterion is equivalent to the largest R^2 criterion. The model with x_3 produced the largest $R^2 = 0.2021$. Thus x_3 is selected as the first predictor.