Model Selection – Solutions

STAT-UB.0103 – Statistics for Business Control and Regression Models

Model Selection

1. Here are the results from fitting two models for Text. The first model using a single predictor variable, Social:

```
Analysis of Variance

Source         DF    Adj SS   Adj MS  F-Value  P-Value
Regression      1   1885519  1885519     3.39    0.072
Error          44  24439192   555436
Total          45  26324711

Model Summary

      S   R-sq  R-sq(adj)  R-sq(pred)
745.276  7.16%      5.05%       0.27%

Regression Equation

Text = 174 + 0.706 Social
```

The second model uses two predictor variables: Social and Audio.

```
Analysis of Variance

Source         DF    Adj SS   Adj MS  F-Value  P-Value
Regression      2   2563229  1281615     2.32    0.111
Error          43  23761482   552593
Total          45  26324711

Model Summary

      S   R-sq  R-sq(adj)  R-sq(pred)
743.366  9.74%      5.54%       0.00%

Regression Equation

Text = 77 + 0.712 Social + 0.992 Audio
```

(a) Which model has the highest value of R^2?

Solution: The model with 2 predictors.

(b) Compute the value of AIC for the first model.

Solution:

$$\begin{aligned}\text{AIC} &= n\log(\text{SSE}/n) + 2(k+1)\\ &= (46)\log(24439192/46) + 2(1+1)\\ &= 610.4206\end{aligned}$$

(c) Compute the value of AIC for the second model.

Solution:

$$\begin{aligned}\text{AIC} &= n\log(\text{SSE}/n) + 2(k+1)\\ &= (46)\log(23761482/46) + 2(2+1)\\ &= 611.127\end{aligned}$$

(d) According to AIC, which of these two models is preferable?

Solution: The first model has a smaller value of AIC, and is therefore preferable.

(e) According to R_a^2, which of these two models is preferable?

Solution: The second model has a larger value of R_a^2, and is therefor preferred over the first. Note that R_a^2 and AIC give different answers as to which model is "best."

Best Subsets Regression

2. Here is the output from using best subsets regression with response Text and predictor variables Video, Audio, Email, Social, and Mail:

```
Response is Text
                                                  S
                                            V A E o
                                            i u m c M
                                            d d a i a
             R-Sq    R-Sq  Mallows          e i i a i
Vars  R-Sq  (adj)  (pred)       Cp        S o o l l l
   1   7.2    5.1     0.3      0.9  745.28        X
   2   9.7    5.5     0.0      1.7  743.37    X   X
   3  13.0    6.8     0.0      2.2  738.48  X X   X
   4  13.3    4.9     0.0      4.0  745.98  X X   X X
   5  13.4    2.5     0.0      6.0  755.15  X X X X X
```

Use the output to answer the following questions:

(a) Of all candidate models with exactly 3 predictors, which fitted model has the smallest value of SSE?

Solution: The model with Video, Audio, and Social as predictors.

(b) Of all candidate models with up to 5 predictors, which fitted model has the smallest value of SSE?

Solution: The model with all 5 predictors (R^2 is highest here).

(c) Write an expression for SSE in terms of R^2 and SST.

Solution: Since $R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$, we have

$$\text{SSE} = \text{SST} \cdot (1 - R^2).$$

(d) Write an expression for AIC in terms of R^2, SST, n, and k.

Solution: Since

$$\text{SSE} = \text{SST} \cdot (1 - R^2)$$

and

$$\text{AIC} = n \log \frac{\text{SSE}}{n} + 2(k+1)$$

we have

$$\text{AIC} = n \log \frac{\text{SST}}{n} + n \log(1 - R^2) + 2(k+1)$$

(e) Use the answer from the previous part to find the candidate model with the smallest value of AIC.

Solution: Since SST and n are the same for all models, the model with the smallest value of AIC is the one with the smallest value of $n \log(1 - R^2) + 2(k+1)$. Ignoring the factor of $n \log(\text{SST}/n)$, we compute:

$$\begin{aligned}
\text{AIC}_1 &= 46 \log(1 - .072) + 2(1+1) = 0.56 \\
\text{AIC}_2 &= 46 \log(1 - .097) + 2(2+1) = 1.30 \\
\text{AIC}_3 &= 46 \log(1 - .130) + 2(3+1) = 1.59 \\
\text{AIC}_4 &= 46 \log(1 - .133) + 2(4+1) = 3.43 \\
\text{AIC}_5 &= 46 \log(1 - .134) + 2(5+1) = 5.38
\end{aligned}$$

The model with the smallest AIC is the one with a single predictor, Social.

(f) In this situation, does AIC agree with R_a^2?

Solution: No. The model with the highest value of R_a^2 is the model with 3 predictors (Video, Audio, and Social). This is a different model than the one chosen by AIC.