STA 5207 Assignment 3

Due Friday September 24

The data set prostate is in the faraway package. The data set has 97 observations and 8 predictors. A study on 97 men with prostate cancer who were due to receive a radical prostatectomy. The data set has the following columns (without headings)

Lcavol: log(cancer volume)

Lweight: log(prostate weight)

Age: age

Lbph: log(benign prostatic hyperplasia amount)

Svi: seminal vesicle invasion

Lcp: log(capsular penetration)

Gleason: Gleason score

pgg45: percentage Gleason scores 4 or 5

lpsa: log(prostate specific antigen)

It is also posted as prostate.txt in SAS Studio. The first column in the file prostate.txt contains observation numbers. The observation numbers must be read into SAS as a variable, but should be used in any of the procedures.

1. (60 points, 10 points each) Use lpsa as the response and the other variables as predictors. Implement the following variable selection methods to determine the best model. Create a table listing each method and the subset of variables chosen by the method.

Subsets Regression Summary

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | R-Square | Adj.  R-Square | Pred  R-Square | C(p) | AIC | SBIC | SBC | MSEP | FPE | HSP | APC |
| 1 | 0.5394 | 0.5346 | 0.5179 | 24.3946 | 232.9080 | -43.1213 | 240.6321 | 60.1553 | 0.6329 | 0.0066 | 0.4800 |
| 2 | 0.5859 | 0.5771 | 0.5534 | 14.5415 | 224.5837 | -51.1891 | 234.8825 | 54.6631 | 0.5809 | 0.0061 | 0.4405 |
| 3 | 0.6264 | 0.6144 | 0.5869 | 6.2169 | 216.5979 | -58.5271 | 229.4714 | 49.8518 | 0.5350 | 0.0056 | 0.4057 |
| 4 | 0.6366 | 0.6208 | 0.5898 | 5.6264 | 215.9223 | -58.8841 | 231.3705 | 49.0284 | 0.5313 | 0.0056 | 0.4029 |
| 5 | 0.6441 | 0.6245 | 0.5882 | 5.7150 | 215.8997 | -58.5526 | 233.9227 | 48.5502 | 0.5312 | 0.0056 | 0.4028 |
| 6 | 0.6493 | 0.6259 | 0.5856 | 6.4020 | 216.4854 | -57.6133 | 237.0831 | 48.3850 | 0.5345 | 0.0056 | 0.4053 |
| 7 | 0.6544 | 0.6273 | 0.5835 | 7.0822 | 217.0428 | -56.6319 | 240.2152 | 48.2125 | 0.5376 | 0.0056 | 0.4077 |
| 8 | 0.6548 | 0.6234 | 0.576 | 9.0000 | 218.9522 | -54.5019 | 244.6993 | 48.7211 | 0.5484 | 0.0058 | 0.4159 |

1. The largest Adjusted R2 value is 0.6273, corresponding to a model with lcavol, lweight, age, lbph, svi, lcp, and pgg45.
2. The smallest AIC value is 215.8997, corresponding to a model with lcavol, lweight, age, lbph, and svi.
3. The smallest SBC value 231.3705, corresponding to a model with lcavol, lweight, lbph, and svi.
4. Forward Selection (In R, use penter=.50 and in SAS use SLE = .50)
   1. lcavol, lweight, svi, lbph, age, pgg45, and lcp are chosen by forward selection.
5. Backward Elimination (In R use prem=.10 and in SAS use SLS = .10)
   1. 5 variables were removed, so lcavol, lweight, and svi were selected by backward elimination.
6. Stepwise Selection (In R use pent=.15 and prem=.15 and in SAS use SLE = .15 and SLS = .15)
   1. The predictors lcavol, lweight, svi, and lbph were chosen by stepwise selection.
7. (40 points) For each of the candidate models chosen in Question 1, calculate the PRESS statistic. Which model is preferred based on this criteria?
   1. The PRESS of the forward selection is 53.27402.
   2. The PRESS of the backward elimination 52.84734.
   3. The PRESS of the stepwise selection is 52.47784.
   4. Based on this criteria, we would prefer the model offered by stepwise selection.

No output or code needs to be submitted.