

Assignment 5

- The data in `boston.dat` (on the course web site) are a subset of variables from the Boston House-Price Data.¹ Each line in `boston.dat` represents a community in the Boston area, and each column is a variable. The variables, in column order, are

```

CRIM      per capita crime rate by town
NOX       nitric oxides concentration (parts per 10 million)
AGE       proportion of owner-occupied units built prior to 1940
DIS       weighted distances to five Boston employment centres
PTRATIO   pupil-teacher ratio by town
LSTAT     % lower status of the population
MEDV     Median value of owner-occupied homes in $1000's
    
```

You will apply variable-selection techniques to find good linear regression models for the median home value `MEDV` as a function of the other variables.

To complete the parts that follow, use the SAS® script `boston.sas` or the R script `boston.R` from the course web site. Your personal data set will be a random subset of 30 communities from the full data set. Replace the word “seed” in `boston.sas` or `boston.R` with the number listed after your name below:

BAI LEI	841	LIU MINZHAO	473
BROWN STEPHEN V	627	LU CUIE	640
CHANG CHE-SHUN	716	LUO XUAN	396
CHEN OU	189	MA LU	206
DONOHUE MICHAEL	792	MALLICK PRANJAL	994
GAO HAIBING	712	MARCUS GABRIEL	113
GARG DIVYA	221	NAMKOONG YOUNG	703
GLUCK MATHEW R	796	NEAL DANIEL W	630
GORDON ROBERT F	976	PETTERSON SONIA	453
GUCI LEDIA	207	PRANO BRIJIDA A	424
HARIHARAN POOJA	326	SHAO ANQI	369
HUANG LEI	380	SINHA AMIT	931
KIM CHANMIN	544	THAYER LAURA K	354
KIRPICH ALEX	834	YE RONGZHONG	172
LEARY EMILY V	395	ZHOU ZHUO	446
LI KE	732	ZHU XIAOYU	180
LIN TONG	100	demo	656

¹Harrison, D. and Rubinfeld, D.L. “Hedonic prices and the demand for clean air”, *J. Environ. Economics & Management*, vol. 5, 81-102, 1978. Available from StatLib (<http://lib.stat.cmu.edu/datasets>).

- (a) Attach a printout of the results from the all-possible-regressions analysis. Determine the best variable subset of each size. Are all of your best subsets nested? (That is, does each best subset contain all smaller best subsets?)
- (b) Attach printouts of the stepwise regression, forward selection, and backward elimination results. Which variables appear in the final model after stepwise regression? After forward selection? After backward elimination?
2. The following table gives the R^2 , adjusted R^2 , C_p , AIC, MS(Res), and SBC values for the *best* subset models (according to R^2) of sizes 1 through 9 from a linear regression:

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	MSE	SBC
1	0.9099	0.9067	113.3924	-42.2117	0.22960	-39.4093
2	0.9161	0.9099	105.8522	-42.3386	0.22181	-38.1350
3	0.9718	0.9685	21.6610	-73.0369	0.07745	-67.4320
4	0.9808	0.9777	9.7381	-82.5577	0.05486	-75.5516
5	0.9845	0.9813	6.0179	-86.9665	0.04616	-78.5593
6	0.9860	0.9824	5.6609	-88.0652	0.04344	-78.2568
7	0.9865	0.9822	6.8614	-87.1935	0.04373	-75.9839
8	0.9870	0.9821	8.0581	-86.3715	0.04405	-73.7606
9	0.9871	0.9813	10.0000	-84.4585	0.04612	-70.4465

Plot each of these criteria versus subset size (except for C_p , which should be plotted versus p' , the number of mean-related terms). (You may produce the plots either by hand or using a computer.) For each criterion, except R^2 , determine which subset size is best.

3. Perform the following exercises from the textbook, Section 7.7. Show your work!
- (a) Write MS(Res) in terms of R^2 for a model with p independent variables. (Exercise 7.1)
- (b) Write C_p in terms of the R^2 value (that is, the R^2 value for the p -variable subset model). Separately, write C_p in terms of MS(Res)_p . (Exercise 7.2)
- (c) Exercise 7.3 (You may ignore any bias caused by variable selection.)
- (d) Exercise 7.4 (Assume, as usual, that the model has an intercept.)