Six of one, half-dozen the other: in practice, many models fit the data equally well

W. John Boscardin, PhD

john.boscardin@ucsf.edu

Departments of Medicine and Epidemiology & Biostatistics

University of California, San Francisco

Acknowledgements

- UCSF Pepper Center (P30/AG044281) Statistics Core, Methodology Development Project
- Core Members/Collaborators: Yinghui Miao, MPH; Irena Cenzer, MA; Kate Kirby, MS
- UCSF Geriatrics Collaborators: Seth Landefeld (R01/AG029233), Ken Covinsky (R01/AG028481)
- SFVAMC Health Services Support: REAP Biostatistics Contract (REA/01-097)
- UCSF CTSI Methodology Grant: Boscardin (under UL1/RR024131)

The Source

STATISTICS IN MEDICINE, VOL. 15, 361–387 (1996)

TUTORIAL IN BIOSTATISTICS

MULTIVARIABLE PROGNOSTIC MODELS: ISSUES IN DEVELOPING MODELS, EVALUATING ASSUMPTIONS AND ADEQUACY, AND MEASURING AND REDUCING ERRORS

FRANK E. HARRELL Jr., KERRY L. LEE AND DANIEL B. MARK

Divisions of Biometry and Cardiology, Box 3363, Duke University Medical Center, Durham, North Carolina 27710, U.S.A.

Typical Setting for Prognostic Model Building

- Long-term survival data on adults age 70+ ($n \approx 1000$, e.g.).
- Have maybe $P = 50$ baseline, admission, discharge characteristics potentially predicting survival
- Goal: build a reasonably parsimonious ($p = 10$ or $p = 15$ predictors), clinically practical and sensible model that has good discrimination and calibration

Common Approach

- Many researchers in this area do following:
  - divide data set into training and validation halves
  - use stepwise selection to trim down set of all (or all bivariate significant) predictors
  - compare discrimination (e. g. Harrell's c-statistic) and calibration in training and validation sets
- First problem: cross-validation or bootstrapping are preferable to single splitting for assessing over-fitting
- Second problem: not ideal procedure for selecting predictors

Rewriting this approach

- Present researcher with long list of statistically similar models
- Researcher can choose model based on parsimony, practicality, sensibility
- Report/correct overfitting for the entire process of model selection using bootstrapping (or CV)

Barriers to this approach

- To bootstrap the process, need to algorithmize the (subjective) model selection
- Need software to do this easily
- Need evidence that this works well

Overfitting

- “Over-optimism” has two components
- First: whatever procedure was used to select a good model was almost certainly driven by data at hand
- Second: the coefficients for that model are optimized to provide the best fit to the data at hand
- Thus when we try to assess the model performance in a new data set, we will almost always have degradation in the model performance measure
- Problem in trying to assess this with a single split sample is that you can’t separate random variability from systematic overfitting

Bootstrapping Optimism

- Instead of split-sample validation, use bootstrapping to assess over-fitting
- Develop a prognostic model in original dataset using some model selection algorithm. Get c^{orig}
- Generate M bootstrapped datasets
- For each, develop a model using same procedure as in original. Look at its performance in the bootstrapped compared to original dataset
- Specifically, for $m = 1, \ldots, M$, use same model selection algorithm and get c_m^{boot} and c_m^{orig}
- Average amount by which c_m^{boot} exceeds c_m^{orig} measures over-optimism

Types of Bootstrapping

- Standard is to compare the c-statistics of this model in the bootstrapped and original data sets.
- Alternative is .632 bootstrapping: compare the c-statistics in the bootstrapped data set and the (approximately $36.8\% = 1/e$) original observations that did not make it into the bootstrapped data set.
- Optimism for .632 bootstrapping is a weighted average of the two ideas.

Stepwise Selection and Best Subsets

- Many sources have criticized stepwise model selection:
  - Standard errors of coefficients artificially small
  - Coefficient estimates biased away from zero
  - R^2 biased upward
  - Performs poorly in presence of multicollinearity
- Best subset selection usually viewed as even worse in all of these senses than stepwise
- Ronan Conroy: “I would no more let an automatic routine select my model than I would let some best-fit procedure pack my suitcase”.

A Slightly Different View

- All of these things true (to some extent), but I think there is more important point
- Stepwise selection only shows one model and does not output comparisons to other potential models
- Best subsets regression gives a huge amount of useful information for comparing models, and in practice, a large number of models of reasonable parsimony are statistically nearly indistinguishable
- It is tremendously valuable to clinicians to view a lot of similarly performing prognostic models to choose ones that are most practically applied
- All the other criticisms can be addressed with bootstrapped over-optimism

Best Subsets Selection

- Computationally infeasible to fit all 2^P possible subset models
- But for each of $p = 1, 2, 3, ..., P - 1$ it is blazingly fast (using both branch and bound and properties of score test) to find the best (or best k) models according to score statistic
- This gives a list of $k(P - 1)$ models most of which are good in some sense
- Deficiency with Best Subsets: no CLASS variables allowed

Best Subsets in Proc Logistic

Regression Models Selected by Score Criterion

Number of Variables	Score Chi-Square	Variables Included in Model
1	87.0900	COMO1
1	57.3305	HOSP_2
1	49.7609	COMO5
1	49.5145	L_AGE2
1	36.5083	COMO7
1	36.4640	COMO6
1	35.0908	COMO3
1	34.6017	OPT2
1	34.2243	COMO4
1	32.6954	COMO8
2	118.1335	COMO1 HOSP_2
2	117.9059	COMO1 L_AGE2
2	114.7389	COMO1 COMO5
2	113.4627	COMO1 COMO4
2	105.3416	COMO1 COMO6
2	103.9949	COMO1 COMO7
2	103.9094	COMO1 COMO3
2	101.7466	COMO1 OPT2
2	98.2874	L_AGE2 HOSP_2
2	97.0521	COMO1 HC1
3	145.9586	COMO1 L_AGE2 HOSP_2
3	140.2519	COMO1 COMO4 HOSP_2
3	138.4447	COMO1 COMO4 COMO5
3	138.2988	COMO1 COMO4 L_AGE2
3	138.0759	COMO1 COMO5 L_AGE2
3	135.6308	COMO1 COMO6 L_AGE2
3	135.2440	COMO1 COMO5 HOSP_2
3	133.7228	COMO1 COMO3 HOSP_2
3	133.6824	COMO1 COMO3 L_AGE2
3	131.4789	COMO1 L_AGE2 OPT2

Best Subsets in Proc Logistic (2)

```
  8    199.8464  COMO1 COMO3 COMO4 COMO6 COMO7 L_AGE2 HOSP_1 HOSP_2
  8    199.7390  COMO1 COMO3 COMO4 COMO6 COMO7 L_AGE1 L_AGE2 HOSP_2
  8    199.5003  COMO1 COMO3 COMO4 COMO6 L_AGE1 L_AGE2 HOSP_1 HOSP_2
  8    199.4539  COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 L_AGE2 HOSP_2
  8    198.7950  COMO1 COMO3 COMO4 COMO5 COMO6 L_AGE2 HOSP_1 HOSP_2
  8    198.3175  COMO1 COMO3 COMO4 COMO5 COMO6 L_AGE1 L_AGE2 HOSP_2
  8    197.9948  COMO1 COMO3 COMO4 COMO6 L_AGE2 RACERE3 HOSP_1 HOSP_2
  8    197.3978  COMO1 COMO3 COMO4 COMO6 COMO7 COMO8 L_AGE2 HOSP_2
  8    197.3055  COMO1 COMO3 COMO4 COMO6 COMO7 L_AGE2 RACERE3 HOSP_2
  8    197.2433  COMO1 COMO3 COMO4 COMO6 COMO8 L_AGE2 HOSP_1 HOSP_2

  9    204.5854  COMO1 COMO3 COMO4 COMO6 COMO7 L_AGE1 L_AGE2 HOSP_1 HOSP_2
  9    204.1594  COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 L_AGE1 L_AGE2 HOSP_2
  9    203.6430  COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 L_AGE2 HOSP_1 HOSP_2
  9    203.4108  COMO1 COMO3 COMO4 COMO6 COMO7 COMO8 L_AGE1 L_AGE2 HOSP_2
  9    203.2493  COMO1 COMO3 COMO4 COMO5 COMO6 L_AGE1 L_AGE2 HOSP_1 HOSP_2
  9    203.2171  COMO1 COMO3 COMO4 COMO6 COMO7 L_AGE2 RACERE3 HOSP_1 HOSP_2
  9    202.8507  COMO1 COMO3 COMO4 COMO6 COMO8 L_AGE1 L_AGE2 HOSP_1 HOSP_2
  9    202.5386  COMO1 COMO3 COMO4 COMO6 COMO7 COMO8 L_AGE2 HOSP_1 HOSP_2
  9    202.5024  COMO1 COMO3 COMO4 COMO6 L_AGE1 L_AGE2 RACERE3 HOSP_1 HOSP_2
  9    202.4004  COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 L_AGE2 HOSP_2

 10    207.9635  COMO1 COMO3 COMO4 COMO6 COMO7 COMO8 L_AGE1 L_AGE2 HOSP_1 HOSP_2
 10    207.9599  COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 L_AGE1 L_AGE2 HOSP_1 HOSP_2
 10    207.8042  COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 L_AGE1 L_AGE2 HOSP_2
 10    207.5607  COMO1 COMO3 COMO4 COMO6 COMO7 L_AGE1 L_AGE2 RACERE3 HOSP_1 HOSP_2
 10    206.6688  COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 L_AGE2 RACERE3 HOSP_1 HOSP_2
 10    206.6097  COMO1 COMO3 COMO4 COMO5 COMO6 COMO8 L_AGE1 L_AGE2 HOSP_1 HOSP_2
 10    206.4228  COMO1 COMO3 COMO4 COMO6 COMO7 L_AGE1 L_AGE2 HOSP_1 HOSP_2 HC_2
 10    206.3936  COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 L_AGE1 L_AGE2 RACERE3 HOSP_2
 10    206.3726  COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 L_AGE2 HOSP_1 HOSP_2
 10    206.3384  COMO1 COMO3 COMO4 COMO6 COMO7 L_AGE1 L_AGE2 HOSP_1 HOSP_2 OPT2
```

Using Best Subset to Select a Single Model

- To attempt to algorithmize the use of best subset, consider adding a predictor until the jump in the score statistic no longer exceeds 3.84 (which would be for nested models a test at $p = 0.05$)
- Alternatively can actually manually calculate AIC $= -2LLH + 2(p+1)$ and BIC $= -2LLH + \log(n)(p+1))$ in the best subset models (see Shtatland et al.)
  - even though score test and LLR test are asympotically equivalent in theory, the values of the test statistics can be quite different in practice
- This is a fairly greedy use of best subset – is there a price to be paid?

SAS Macro Description

- Regression Models: Logistic, Cox
- Selection Methods: Nested Score, Best AIC, Best BIC, All Bivariates, Stepwise on Bivariates, Regular Stepwise
- Bootstrapping: Standard, .632
- Class Variables Allowed for Three Best Subset Methods

SAS Macro Output Summary Table

Model-Summary [generated in original dataset]

MODEL_TYPE	Variables in complete model	AIC	BIC	C Stat	Score	Optimism corrected c [Bootstrap]	Optimism corrected c [.632-Bootstrap]
Best AIC	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1036.2431	1094.2524	0.783498	202.4004	0.76557	0.76508
Best BIC	COMO1 COMO3 COMO4 COMO6 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1044.8632	1088.3702	0.769678	193.2394	0.75361	0.75515
Nested Score	COMO1 COMO3 COMO4 COMO6 COMO7 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1040.3808	1088.7219	0.774647	204.5854	0.75695	0.75953
All Biv Signif	COMO1 COMO2 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 HC1 HOSP_1-HOSP_2 L_AGE1-L_AGE2 OPT1-OPT2 RACERE1-RACERE3	1042.5715	1134.4196	0.787297	217.4752	0.77097	0.76934
Stepwise_Biv Signif	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1036.2431	1094.2524	0.783498	211.3465	0.76721	0.76448
Stepwise_Regular	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1036.2431	1094.2524	0.783498	202.4004	0.76683	0.76401

SAS Macro Output Best AIC

Best Model Generated in Original Dataset by BestSubset Procedure

Number in original model	Variables in original model	Number of variables in complete model	Variables in complete model	AIC with covariates in complete model	SC with covariates in complete model	HARRELL_C	Score Chi-Square
9	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 L_AGE2 HOSP_2	11	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1036.2431	1094.2524	0.783498	202.4004
11	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 L_AGE1 L_AGE2 HOSP_2 HC_2	12	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 HC_2 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1036.3489	1099.1923	0.784226	209.5317
12	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 L_AGE1 L_AGE2 HOSP_1 HOSP_2 HC1	12	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 HC1 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1037.0989	1099.9423	0.783046	212.4253
13	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 L_AGE1 L_AGE2 HOSP_1 HOSP_2 HC1 HC_2	13	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 HC1 HC_2 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1037.3285	1105.0060	0.783919	214.1122
13	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 L_AGE1 L_AGE2 HOSP_1 HOSP_2 HC_2 OPT2	14	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 HC_2 HOSP_1-HOSP_2 L_AGE1-L_AGE2 OPT1-OPT2	1037.8533	1110.3649	0.785783	215.1766
11	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 L_AGE1 L_AGE2 HOSP_2 OPT2	13	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 HOSP_1-HOSP_2 L_AGE1-L_AGE2 OPT1-OPT2	1037.9469	1105.6245	0.785793	210.1383
8	COMO1 COMO3 COMO4 COMO6 COMO7 COMO8 L_AGE2 HOSP_2	10	COMO1 COMO3 COMO4 COMO6 COMO7 COMO8 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1038.2663	1091.4415	0.779227	197.3978
11	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 L_AGE1 L_AGE2 HOSP_1 HOSP_2 HC_2	11	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 HC_2 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1038.2845	1096.2938	0.780103	209.8595
8	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 L_AGE2 HOSP_2	10	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1038.3059	1091.4811	0.779665	199.4539
11	COMO1 COMO3 COMO4 COMO6 COMO7 COMO8 L_AGE1 L_AGE2 HOSP_1 HOSP_2 HC_2	11	COMO1 COMO3 COMO4 COMO6 COMO7 COMO8 HC_2 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1038.4257	1096.4350	0.780661	209.7136

SAS Macro Output Diminishing Score

Best Model Generated in Original Dataset by BestSubset Procedure

Number in original model	Variables in original model	Number of variables in complete model	Variables in complete model	AIC with covariates in complete model	SC with covariates in complete model	HARRELL_C	Score Chi-Square
14	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 L_AGE1 L_AGE2 RACERE3 HOSP_1 HOSP_2 HC1 HC_2	16	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 HC1 HC_2 HOSP_1-HOSP_2 L_AGE1-L_AGE2 RACERE1-RACERE3	1040.0451	1122.2249	0.785049	216.1907
10	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 L_AGE2 RACERE3 HOSP_1 HOSP_2	13	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 HOSP_1-HOSP_2 L_AGE1-L_AGE2 RACERE1-RACERE3	1040.0699	1107.7474	0.782391	206.6688
11	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 L_AGE1 L_AGE2 HOSP_1 HOSP_2 OPT2	12	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 HOSP_1-HOSP_2 L_AGE1-L_AGE2 OPT1-OPT2	1040.1927	1103.0361	0.781968	209.5846
10	COMO1 COMO3 COMO4 COMO5 COMO6 COMO8 L_AGE1 L_AGE2 HOSP_1 HOSP_2	10	COMO1 COMO3 COMO4 COMO5 COMO6 COMO8 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1040.3218	1093.4970	0.779583	206.6097
14	COMO1 COMO2 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 L_AGE1 L_AGE2 RACERE3 HOSP_1 HOSP_2 HC_2	16	COMO1 COMO2 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 HC_2 HOSP_1-HOSP_2 L_AGE1-L_AGE2 RACERE1-RACERE3	1040.3243	1122.5042	0.787051	215.9163
12	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 L_AGE1 L_AGE2 RACERE3 HOSP_2 OPT2	16	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 HOSP_1-HOSP_2 L_AGE1-L_AGE2 OPT1-OPT2 RACERE1-RACERE3	1040.3691	1122.5490	0.787059	212.0418
12	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 L_AGE1 L_AGE2 RACERE3 HOSP_1 HOSP_2 HC_2	14	COMO1 COMO3 COMO4 COMO5 COMO6 COMO7 HC_2 HOSP_1-HOSP_2 L_AGE1-L_AGE2 RACERE1-RACERE3	1040.3788	1112.8904	0.782350	212.3667
9	COMO1 COMO3 COMO4 COMO6 COMO7 L_AGE1 L_AGE2 HOSP_1 HOSP_2	9	COMO1 COMO3 COMO4 COMO6 COMO7 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1040.3808	1088.7219	0.774647	204.5854

SAS Macro Output Best BIC

Best Model Generated in Original Dataset by BestSubset Procedure

Number in original model	Variables in original model	Number of variables in complete model	Variables in complete model	AIC with covariates in complete model	SC with covariates in complete model	HARRELL_C	Score Chi-Square
17	COMO1 COMO2 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 L_AGE1 L_AGE2 GENDER2 RACERE3 HOSP_1 HOSP_2 HC1 HC_2 OPT2	20	COMO1 COMO2 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 GENDER2 HC1 HC_2 HOSP_1-HOSP_2 L_AGE1-L_AGE2 OPT1-OPT2 RACERE1-RACERE3	1044.6636	1146.1799	0.787823	218.8817
7	COMO1 COMO3 COMO4 COMO6 L_AGE1 L_AGE2 HOSP_2	8	COMO1 COMO3 COMO4 COMO6 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1044.8632	1088.3702	0.769678	193.2394
7	COMO1 COMO3 COMO4 COMO5 COMO7 L_AGE2 HOSP_2	9	COMO1 COMO3 COMO4 COMO5 COMO7 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1046.0410	1094.3821	0.770703	191.7362
7	COMO1 COMO3 COMO4 COMO6 L_AGE2 HOSP_2 OPT2	10	COMO1 COMO3 COMO4 COMO6 HOSP_1-HOSP_2 L_AGE1-L_AGE2 OPT1-OPT2	1046.3252	1099.5004	0.772362	190.2202
7	COMO1 COMO3 COMO4 COMO6 L_AGE2 RACERE3 HOSP_2	11	COMO1 COMO3 COMO4 COMO6 HOSP_1-HOSP_2 L_AGE1-L_AGE2 RACERE1-RACERE3	1046.3301	1104.3394	0.773701	190.4831
19	COMO1 COMO2 COMO3 COMO4 COMO6 COMO7 COMO8 L_AGE1 L_AGE2 GENDER2 RACERE1 RACERE2 RACERE3 HOSP_1 HOSP_2 HC1 HC_2 OPT1 OPT2	19	COMO1 COMO2 COMO3 COMO4 COMO6 COMO7 COMO8 GENDER2 HC1 HC_2 HOSP_1-HOSP_2 L_AGE1-L_AGE2 OPT1-OPT2 RACERE1-RACERE3	1046.4574	1143.1396	0.784669	216.1812
6	COMO1 COMO3 COMO4 COMO7 L_AGE2 HOSP_2	8	COMO1 COMO3 COMO4 COMO7 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1050.3070	1093.8140	0.766123	184.4826
6	COMO1 COMO3 COMO4 COMO5 L_AGE2 HOSP_2	8	COMO1 COMO3 COMO4 COMO5 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1050.9743	1094.4813	0.766105	184.9045

SAS Macro Output Variable Selection AIC

Best Model Generated in Original Dataset by BestSubset Procedure

Number in original model	Variables in original model	Number of variables in complete model	Variables in complete model	AIC with covariates in complete model	SC with covariates in complete model	HARRELL_C	Score Chi-Square
17	COMO1 COMO2 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 L_AGE1 L_AGE2 GENDER2 RACERE3 HOSP_1 HOSP_2 HC1 HC_2 OPT2	20	COMO1 COMO2 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 GENDER2 HC1 HC_2 HOSP_1-HOSP_2 L_AGE1-L_AGE2 OPT1-OPT2 RACERE1-RACERE3	1044.6636	1146.1799	0.787823	218.8817
7	COMO1 COMO3 COMO4 COMO6 L_AGE1 L_AGE2 HOSP_2	8	COMO1 COMO3 COMO4 COMO6 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1044.8632	1088.3702	0.769678	193.2394
7	COMO1 COMO3 COMO4 COMO5 COMO7 L_AGE2 HOSP_2	9	COMO1 COMO3 COMO4 COMO5 COMO7 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1046.0410	1094.3821	0.770703	191.7362
7	COMO1 COMO3 COMO4 COMO6 L_AGE2 HOSP_2 OPT2	10	COMO1 COMO3 COMO4 COMO6 HOSP_1-HOSP_2 L_AGE1-L_AGE2 OPT1-OPT2	1046.3252	1099.5004	0.772362	190.2202
7	COMO1 COMO3 COMO4 COMO6 L_AGE2 RACERE3 HOSP_2	11	COMO1 COMO3 COMO4 COMO6 HOSP_1-HOSP_2 L_AGE1-L_AGE2 RACERE1-RACERE3	1046.3301	1104.3394	0.773701	190.4831
19	COMO1 COMO2 COMO3 COMO4 COMO6 COMO7 COMO8 L_AGE1 L_AGE2 GENDER2 RACERE1 RACERE2 RACERE3 HOSP_1 HOSP_2 HC1 HC_2 OPT1 OPT2	19	COMO1 COMO2 COMO3 COMO4 COMO6 COMO7 COMO8 GENDER2 HC1 HC_2 HOSP_1-HOSP_2 L_AGE1-L_AGE2 OPT1-OPT2 RACERE1-RACERE3	1046.4574	1143.1396	0.784669	216.1812
6	COMO1 COMO3 COMO4 COMO7 L_AGE2 HOSP_2	8	COMO1 COMO3 COMO4 COMO7 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1050.3070	1093.8140	0.766123	184.4826
6	COMO1 COMO3 COMO4 COMO5 L_AGE2 HOSP_2	8	COMO1 COMO3 COMO4 COMO5 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1050.9743	1094.4813	0.766105	184.9045

SAS Macro Output Variable Selection BIC

Best Model Generated in Original Dataset by BestSubset Procedure

Number in original model	Variables in original model	Number of variables in complete model	Variables in complete model	AIC with covariates in complete model	SC with covariates in complete model	HARRELL_C	Score Chi-Square
17	COMO1 COMO2 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 L_AGE1 L_AGE2 GENDER2 RACERE3 HOSP_1 HOSP_2 HC1 HC_2 OPT2	20	COMO1 COMO2 COMO3 COMO4 COMO5 COMO6 COMO7 COMO8 GENDER2 HC1 HC_2 HOSP_1-HOSP_2 L_AGE1-L_AGE2 OPT1-OPT2 RACERE1-RACERE3	1044.6636	1146.1799	0.787823	218.8817
7	COMO1 COMO3 COMO4 COMO6 L_AGE1 L_AGE2 HOSP_2	8	COMO1 COMO3 COMO4 COMO6 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1044.8632	1088.3702	0.769678	193.2394
7	COMO1 COMO3 COMO4 COMO5 COMO7 L_AGE2 HOSP_2	9	COMO1 COMO3 COMO4 COMO5 COMO7 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1046.0410	1094.3821	0.770703	191.7362
7	COMO1 COMO3 COMO4 COMO6 L_AGE2 HOSP_2 OPT2	10	COMO1 COMO3 COMO4 COMO6 HOSP_1-HOSP_2 L_AGE1-L_AGE2 OPT1-OPT2	1046.3252	1099.5004	0.772362	190.2202
7	COMO1 COMO3 COMO4 COMO6 L_AGE2 RACERE3 HOSP_2	11	COMO1 COMO3 COMO4 COMO6 HOSP_1-HOSP_2 L_AGE1-L_AGE2 RACERE1-RACERE3	1046.3301	1104.3394	0.773701	190.4831
19	COMO1 COMO2 COMO3 COMO4 COMO6 COMO7 COMO8 L_AGE1 L_AGE2 GENDER2 RACERE1 RACERE2 RACERE3 HOSP_1 HOSP_2 HC1 HC_2 OPT1 OPT2	19	COMO1 COMO2 COMO3 COMO4 COMO6 COMO7 COMO8 GENDER2 HC1 HC_2 HOSP_1-HOSP_2 L_AGE1-L_AGE2 OPT1-OPT2 RACERE1-RACERE3	1046.4574	1143.1396	0.784669	216.1812
6	COMO1 COMO3 COMO4 COMO7 L_AGE2 HOSP_2	8	COMO1 COMO3 COMO4 COMO7 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1050.3070	1093.8140	0.766123	184.4826
6	COMO1 COMO3 COMO4 COMO5 L_AGE2 HOSP_2	8	COMO1 COMO3 COMO4 COMO5 HOSP_1-HOSP_2 L_AGE1-L_AGE2	1050.9743	1094.4813	0.766105	184.9045

SAS Macro Output Optimism

Method	Apparent c-Statistic	Optimism Selection Only	Optimism Selection and Estimation
Best AIC	0.783	0.011	0.018
Dimin Score	0.770	0.013	0.018
Best BIC	0.775	0.013	0.016
Bivariate All	0.787	0.007	0.016
Bivariate Step	0.783	0.012	0.016
Stepwise	0.783	0.012	0.017

Over-optimism Results

- Harrell’s c is about 0.78 for all selection procedures in original data
- Optimism is similar for all selection procedures
- Total over optimism due to variable selection and coefficient estimation is less than 0.02 of which a bit more than 0.01 is due to selection

Summary

- Best subset selection by AIC, BIC, or Diminishing score does not result in additional overfitting compared to Stepwise selection in a wide range of settings we have investigated
- Key reason: in this setting, best models perform similarly to each other – there is simply no room for latching on to artifacts in the data
- Results would be likely different with a greedier regression technique (e.g. regression trees) or with very unevenly distributed regressors and their interactions
- The output from best subsets is of great interest to clinical colleagues

References

- Harrell FE, Lee KL, Mark DB (1996). Tutorial in Biostatistics: Multivariable prognostic models. *Stat Med*, **15**, 361–387.
- King (2003). Running a best-subsets logistic regression: an alternative to stepwise methods. *Educ Psych Meas*, **63**, 392–403.
- Shtatland ES, Kleinman K, Cain EM (2003). Stepwise methods in using SAS Proc Logistic and SAS Enterpise Miner for prediction. *SUGI* **29**, 258-28.
- Cenzer IS, Miao Y, Kirby K, Boscardin WJ (2012). Estimating Harrell's optimism using bootstrap samples. *Proceedings of the Western Users of Sas Software Conference*, **74-12**.
- Miao Y, Cenzer IS, Kirby K, Boscardin WJ (2012). Estimating Harrell's optimism on predictive indices using bootstrap samples. *SUGI Proceedings*, **504-2013**.