STAT 401A - Statistical Methods for Research Workers

Variable selection

Jarad Niemi (Dr. J)

Iowa State University

last updated: November 19, 2014

Why choose a subset of the explanatory variables?

Scenarios where you want to choose a subset of the explanatory variables:

1. Adjusting for a large set of explanatory variables
2. Fishing for explanation
3. Prediction

Reasons 1 and 3 have little to no interpretation of the resulting parameters and their significance. Yet, often, interpretation of all parameters is performed and importance is placed on the included explanatory variables. Great restraint should be exercised.

Model selection criteria

- Criteria for linear regression, i.e. the data are normal
  - R^2: always increases as parameters are added
  - Adjusted R^2: "generally favors models with too many variables"
  - F-test: statistical test for normal, nested models
  - Mallow's Cp: $(n-p)\hat{\sigma}^2/\hat{\sigma}^2_{full} + 2p - n$
- More general criteria
  - Akaike's information criterion (AIC): $n\log(\hat{\sigma}^2) + 2p$
  - Bayesian informaiton criterion (BIC): $n\log(\hat{\sigma}^2) + \log(n)p$
  - Cross validation

Approach

- If the models can be enumerated,

  choose a criterion and calculate it for all models

- If the models cannot be enumerated,
  1. choose a criterion and
  2. perform a stepwise variable selection procedure:
     - forward: start from null model and add explanatory variables
     - backward: start from full model and remove explanatory variables
     - stepwise: start from any model and use both forward and backward steps

Model enumeration in SAS

```
DATA case1201;
  INFILE 'case1201.csv' DSD FIRSTOBS=2;
  INPUT State $ SAT Takers Income Years Public Expend Rank;
  lTakers = log(Takers);

PROC REG;
  MODEL SAT = lTakers Rank Years Income Public Expend / SELECTION=Cp AIC SBC; /* SBC is our BIC */
  RUN; QUIT;
```

Model enumeration in SAS

Number in Model	C(p)	R-Square	AIC	SBC	Variables in Model
4	3.0834	0.8917	323.9013	333.46138	lTakers Rank Years Expend
3	4.6829	0.8827	325.9132	333.56128	lTakers Years Expend
5	5.0404	0.8918	325.8513	337.32341	lTakers Rank Years Income Expend
5	5.0760	0.8917	325.8927	337.36481	lTakers Rank Years Public Expend
4	5.7877	0.8849	326.9451	336.50518	lTakers Years Public Expend
4	6.1429	0.8840	327.3314	336.89156	lTakers Years Income Expend
6	7.0000	0.8919	327.8043	341.18850	lTakers Rank Years Income Public Expend
5	7.7048	0.8851	328.8545	340.32660	lTakers Years Income Public Expend
2	8.7126	0.8675	329.9871	335.72313	lTakers Expend
4	8.9207	0.8771	330.2544	339.81450	Rank Years Public Expend
3	9.2862	0.8711	330.6154	338.26348	Rank Years Expend
4	9.9138	0.8746	331.2592	340.81934	Rank Years Income Expend
3	10.0193	0.8693	331.3250	338.97312	lTakers Income Expend
3	10.3732	0.8684	331.6641	339.31220	lTakers Rank Expend
3	10.6389	0.8677	331.9171	339.56522	lTakers Public Expend
5	10.8060	0.8773	332.1370	343.60910	Rank Years Income Public Expend
4	11.5125	0.8705	332.8356	342.39567	lTakers Income Public Expend
4	11.6922	0.8701	333.0097	342.56980	lTakers Rank Public Expend
4	11.8600	0.8697	333.1717	342.73182	lTakers Rank Income Expend
5	12.4221	0.8733	333.7658	345.23796	lTakers Rank Income Public Expend
3	14.3572	0.8584	335.3298	342.97790	Rank Years Income
4	15.0584	0.8616	336.1641	345.72422	lTakers Rank Years Income
5	15.1323	0.8665	336.3836	347.85569	lTakers Rank Years Income Public
3	15.7468	0.8549	336.5477	344.19581	lTakers Rank Years
4	15.9756	0.8593	336.9901	346.55022	Rank Years Income Public
4	16.3018	0.8585	337.2806	346.84076	lTakers Years Income Public
2	16.8576	0.8471	337.1712	342.90724	Rank Years
4	16.9035	0.8570	337.8122	347.37232	lTakers Rank Years Public
3	18.2222	0.8487	338.6464	346.29445	lTakers Years Public

BIC stepwise model selection in SAS

```
DATA case1202;
  INFILE 'case1202.csv' DSD FIRSTOBS=2;
  INPUT Bsal Sal77 Sex $ Senior Age Educ Exper;
  lBsal = log(Bsal);

PROC GLMSELECT DATA=case1202;
  CLASS Sex(REF='Female');
  MODEL lBsal = Sex Senior Age Educ Exper / SELECTION=stepwise CHOOSE=SBC; /* SBC is our BIC */
  RUN;
```

BIC stepwise model selection in SAS

The GLMSELECT Procedure

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	Number Parms In	SBC
0	Intercept		1	1	-377.0828
1	Sex		2	2	-405.0656
2	Senior		3	3	-418.1975
3	Educ		4	4	-424.5180
4	Age		5	5	-427.3876*

* Optimal Value Of Criterion

Selection stopped at a local minimum of the SBC criterion.

Stop Details

Candidate For	Effect	Candidate SBC		Compare SBC
Entry	Exper	-423.0281	>	-427.3876
Removal	Age	-424.5180	>	-427.3876

BIC stepwise model selection in SAS

The GLMSELECT Procedure
Selected Model

The selected model, based on SBC, is the model at Step 4.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	4	0.80014	0.20003	23.92
Error	88	0.73591	0.00836	
Corrected Total	92	1.53604		

Root MSE	0.09145
Dependent Mean	8.58961
R-Square	0.5209
Adj R-Sq	0.4991
AIC	-345.05056
AICC	-344.07381
SBC	-427.38756

Parameter Estimates

Parameter		DF	Estimate	Standard Error	t Value
Intercept		1	8.572205	0.108342	79.12
Sex	Male	1	0.143944	0.021605	6.66
Sex	Female	0	0	.	.
Senior		1	-0.004100	0.000947	-4.33
Age		1	0.000195	0.000072259	2.70
Educ		1	0.017005	0.004474	3.80

BIC stepwise model selection in R

```
m = step(lm(log(Bsal)~Sex+Senior+Age+Educ+Exper, case1202), direction="both", k=log(nrow(case1202)))


Start:  AIC=-423.03
log(Bsal) ~ Sex + Senior + Age + Educ + Exper

         Df Sum of Sq     RSS     AIC
- Exper   1  0.001369 0.73591 -427.39
- Age     1  0.011414 0.74595 -426.13
<none>                0.73454 -423.03
- Educ    1  0.120093 0.85463 -413.48
- Senior  1  0.157972 0.89251 -409.44
- Sex     1  0.306421 1.04096 -395.14

Step:  AIC=-427.39
log(Bsal) ~ Sex + Senior + Age + Educ

         Df Sum of Sq     RSS     AIC
<none>                0.73591 -427.39
- Age     1   0.06097 0.79687 -424.52
+ Exper   1   0.00137 0.73454 -423.03
- Educ    1   0.12083 0.85674 -417.78
- Senior  1   0.15665 0.89256 -413.97
- Sex     1   0.37122 1.10713 -393.94
```

```
summary(m)
```

```

Call:
lm(formula = log(Bsal) ~ Sex + Senior + Age + Educ, data = case1202)

Residuals:
      Min        1Q    Median        3Q       Max
-0.223534 -0.065771 -0.005761  0.059315  0.209990

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.572e+00  1.083e-01  79.122  < 2e-16 ***
SexMale      1.439e-01  2.160e-02   6.663 2.25e-09 ***
Senior      -4.100e-03  9.472e-04  -4.328 3.96e-05 ***
Age          1.951e-04  7.226e-05   2.700 0.008313 **
Educ         1.700e-02  4.474e-03   3.801 0.000265 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09145 on 88 degrees of freedom
Multiple R-squared:  0.5209, Adjusted R-squared:  0.4991
F-statistic: 23.92 on 4 and 88 DF,  p-value: 2.076e-13
```

```
exp(confint(m))
```

```
                  2.5 %       97.5 %
(Intercept) 4259.4632883 6551.9087447
SexMale        1.1062868    1.2054807
Senior         0.9940356    0.9977851
Age            1.0000515    1.0003388
Educ           1.0081476    1.0262331
```

Healthy skepticism

Data simulated from the following model:

$$Y_i \overset{ind}{\sim} N(\mu_i, 1)$$

where

$$\begin{aligned} \mu_i &= 10X_{i,1} + 10X_{i,2} + 10X_{i,3} \\ &+ X_{i,4} + X_{i,5} + X_{i,6} \\ &+ 0.1X_{i,7} + 0.1X_{i,8} + 0.1X_{i,9} \end{aligned}$$

where $X_{i,j} \overset{iid}{\sim} N(0,1)$ for $i = 1, \ldots, 200$ and $j = 1, \ldots, 100$.

Simulated model

```
# Simulated model
set.seed(1)
p = 100
n = 200
b = c(10,10,10,1,1,1,.1,.1,.1, rep(0,91))
x = matrix(rnorm(n*p), n, p)
y = rnorm(n,x%*%b)
d = data.frame(y=y,x=x)
mod = lm(y~.,d)
summary(mod)
mod.aic = step(mod)
mod.bic = step(mod, k=log(n))
```

```
> summary(mod.aic)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.18492    0.06404   2.888 0.004395 **
x.1         10.10298    0.06939 145.601  < 2e-16 ***
x.2         10.04751    0.06394 157.142  < 2e-16 ***
x.3         10.04937    0.06018 167.000  < 2e-16 ***
x.4          0.94539    0.05740  16.469  < 2e-16 ***
x.5          0.95183    0.05752  16.549  < 2e-16 ***
x.6          1.06018    0.06335  16.735  < 2e-16 ***
x.9          0.27968    0.05936   4.712 5.15e-06 ***
x.16        -0.24460    0.05935  -4.121 5.92e-05 ***
x.18        -0.14809    0.06648  -2.228 0.027241 *
x.19         0.13453    0.06275   2.144 0.033493 *
x.21         0.10957    0.06849   1.600 0.111505
x.22         0.08906    0.06248   1.425 0.155893
x.27         0.19548    0.06842   2.857 0.004819 **

... 31,32,34,35,38,40,44,45,49 are included ...

x.50        -0.13274    0.06931  -1.915 0.057178 .
x.61         0.10487    0.06581   1.594 0.112922
x.68         0.14039    0.06764   2.076 0.039471 *
x.72         0.08631    0.06472   1.334 0.184134
x.78        -0.10080    0.06324  -1.594 0.112849
x.81         0.12723    0.06201   2.052 0.041749 *
x.84         0.23409    0.06506   3.598 0.000422 ***
x.86         0.10954    0.06351   1.725 0.086446 .
x.90        -0.15650    0.06607  -2.369 0.018993 *
x.93         0.09983    0.05896   1.693 0.092263 .

Residual standard error: 0.8417 on 167 degrees of freedom
Multiple R-squared: 0.9981,  Adjusted R-squared: 0.9977
F-statistic:  2745 on 32 and 167 DF,  p-value: < 2.2e-16
```

```
> summary(mod.bic)

Call:
lm(formula = y ~ x.1 + x.2 + x.3 + x.4 + x.5 + x.6 + x.9 + x.16 +
    x.27 + x.84, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5419 -0.5243  0.1222  0.6292  2.5151

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.14420    0.06673   2.161 0.031967 *
x.1         10.03241    0.07132 140.673  < 2e-16 ***
x.2         10.00679    0.06484 154.324  < 2e-16 ***
x.3         10.05523    0.06155 163.378  < 2e-16 ***
x.4          0.99144    0.06031  16.438  < 2e-16 ***
x.5          0.98504    0.06144  16.033  < 2e-16 ***
x.6          1.05357    0.06607  15.946  < 2e-16 ***
x.9          0.20230    0.06038   3.351 0.000974 ***
x.16        -0.15225    0.06108  -2.493 0.013543 *
x.27         0.18068    0.07120   2.538 0.011966 *
x.84         0.17341    0.06718   2.581 0.010598 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.9184 on 189 degrees of freedom
Multiple R-squared: 0.9974,Adjusted R-squared: 0.9973
F-statistic:  7373 on 10 and 189 DF,  p-value: < 2.2e-16
```

Alternatives to variable selection

- Model averaging
  - Bayesian model averaging
  - AIC model averaging
  - BIC model averaging
- Keep all variables, but shrink the coefficients toward zero
  - Lasso
  - Ridge regression
  - Elastic net