Variable selection

STAT 401 - Statistical Methods for Research Workers

Jarad Niemi

Iowa State University

November 1, 2013

Why 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

AIC stepwise model selection in R

```
> step(lm(sat~log(takers)+income+years+public+expend+rank,d), direction="both")
Start:  AIC=327.8
sat ~ log(takers) + income + years + public + expend + rank

              Df Sum of Sq   RSS    AIC
- public       1      25.0 26610 325.85
- income       1      47.0 26632 325.89
<none>                     26585 327.80
- rank         1    1672.2 28257 328.85
- log(takers)  1    3589.6 30175 332.14
- years        1    4588.8 31174 333.77
- expend       1    6264.4 32850 336.38

Step:  AIC=325.85
sat ~ log(takers) + income + years + expend + rank

              Df Sum of Sq   RSS    AIC
- income       1      26.6 26637 323.90
<none>                     26610 325.85
- rank         1    1918.1 28528 327.33
+ public       1      25.0 26585 327.80
- log(takers)  1    4249.6 30860 331.26
- years        1    5452.8 32063 333.17
- expend       1    7430.3 34040 336.16
```

AIC stepwise model selection in R

```
Step:  AIC=323.9
sat ~ log(takers) + years + expend + rank

              Df Sum of Sq   RSS    AIC
<none>                     26637 323.90
+ income       1      26.6 26610 325.85
+ public       1       4.6 26632 325.89
- rank         1    2225.4 28862 325.91
- log(takers)  1    5071.4 31708 330.62
- years        1    5743.5 32380 331.66
- expend       1    9065.8 35703 336.55

Call:
lm(formula = sat ~ log(takers) + years + expend + rank, data = d)

Coefficients:
(Intercept)  log(takers)        years       expend         rank
    388.426      -38.015       17.857        2.423        4.004
```

Healthy skepticism

Data simulated from the following model:

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

where

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

where $X_{i,j} \stackrel{iid}{\sim} N(0,1)$ 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
```

Cross validation

1. Randomly split the data into:
   - training
   - testing
2. Use stepwise selection to find a model using the training data
3. Fit that model again on the testing data to obtain the final model

Approaches that improve on this basic idea:

- Leave-one-out cross-validation
- k-fold cross-validation

Cross validation

```
testing.indices = sample(n,n*.25)
training        = d[setdiff(1:200,testing.indices),]
testing         = d[testing.indices,]
mod             = lm(y~., training)
mod.training    = step(mod, k=log(nrow(training)))
keep            = as.numeric(gsub("[^0-9]","",names(mod.training$coefficients)[-1])
mod.testing     = step(lm(y~., testing[,c(1,1+keep)]), k=log(nrow(testing)))
```

Cross validation

```
> summary(mod.testing)

Call:
lm(formula = y ~ x.1 + x.2 + x.3 + x.4 + x.5 + x.6 + x.16 + x.64,
    data = testing[, c(1, 1 + keep)])

Residuals:
    Min      1Q  Median      3Q     Max
-1.8349 -0.5965  0.1962  0.6256  1.6548

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2833     0.1301   2.177   0.0353 *
x.1           9.9088     0.1417  69.941  < 2e-16 ***
x.2           9.8353     0.1319  74.552  < 2e-16 ***
x.3          10.0542     0.1132  88.838  < 2e-16 ***
x.4           0.8640     0.1138   7.591 2.45e-09 ***
x.5           0.9291     0.1372   6.773 3.45e-08 ***
x.6           1.1560     0.1461   7.915 8.70e-10 ***
x.16         -0.2889     0.1141  -2.532   0.0153 *
x.64          0.3453     0.1277   2.705   0.0099 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.8621 on 41 degrees of freedom
Multiple R-squared:  0.9975,Adjusted R-squared:  0.997
F-statistic:  2024 on 8 and 41 DF,  p-value: < 2.2e-16
```

Alternatives to variable selection

- Keep all models and calculate their posterior probability

$$p(M_j|D) = p(M_j)\frac{e^{-BIC_j}}{SUM}$$

where

$$SUM = \sum_{j=1}^{J} e^{-BIC_j}.$$

- Keep all variables, but shrink them toward zero
    - Lasso
    - Ridge regression
    - Elastic net