Variable Selection

Chapter 9 of LMWR2, Chapter 10 of ALR4

Subrata Paul

6/4/2020

Overview

Variable selection is intended to (objectively) find the best subset of predictors.

Reasons for this include:

- We want the simplest model that adequately explains the data.
- Unnecessary regressors will add noise to all model estimates.
    - Degrees of freedom are wasted.
    - A smaller model might achieve more precise estimates and predictions.
- Removing excess regressors aids in interpretation and helps to prevent problems with linearly dependent regressors.
- If the model is to be used for prediction, we can save time and/or money by not having to measure extra predictors (and improve our prediction!).

Two aspects to variable selection:

- The criterion used to compare models.
- The strategy used to search for the “optimal” model.

Selection Criteria (P-values)

P-values are a common criterion for selecting regressors to keep in our regression model.

This criterion keeps the regressors with the smallest p-values in the model, specifically the regressors with p-values less than some threshold, α_{crit}.

Selection Criteria (AIC)

Akaike's Information Criterion is a information-based criterion for variable selection.

$$AIC(\mathcal{M}) = -2\ell(\mathcal{M}) + 2p_{\mathcal{M}}$$

where,

- $\mathcal{M}$ is the model
- $\ell(\mathcal{M})$ is the log likelihood of the model using the MLE estimates of the parameters
- $p_{\mathcal{M}}$ is the number of regression coefficients in model $\mathcal{M}$

For linear regression models, $-2\ell(\mathcal{M}) = n\log(RSS/n) + c$, where c is a constant that depends only on the observed data and not on the model. c can be ignored when comparing between models (on same data).

Selection Criteria (BIC)

Bayesian Information Criterion is another information-based criterion for variable selection.

$$BIC(\mathcal{M}) = -2\ell(\mathcal{M}) + \log(n)p_{\mathcal{M}}$$

where,

- $\mathcal{M}$ is the model
- $\ell(\mathcal{M})$ is the log likelihood of the model using the MLE estimates of the parameters
- $p_{\mathcal{M}}$ is the number of regression coefficients in model $\mathcal{M}$

Smaller AIC or BIC is better

We favor models with smaller AIC or BIC.

The information criteria capture two aspects of model fit:

- The $-2\ell(\mathcal{M})$ measures how well the fitted model matches the observed data.
- The second component penalizes the model according to the number of parameters it includes (its complexity).
  - The more parameters, the larger the penalty.
- Models with more parameters will fit better (reducing the RSS), but will be penalized more for having additional parameters.
- AIC and BIC provide criteria for balancing model fit with model complexity.
- BIC tends to penalize complex models more heavily than AIC (anytime $\log(n) > 2$, i.e., n≥8), so it tends to suggest simpler models than the AIC criterion.

Selection Criteria (R^2)

R^2 never decreases as new regressors are added to the model.

- It is useless for comparing models with different numbers of regressors.

Adjusted R^2, R_a^2, is a better criterion for assessing model fit.
* The adjusted R^2 criterion penalizes for the number of parameters in the model.

For mdoel $\mathcal{M}$ with $p_{\mathcal{M}}$ regression coefficients,

$$R_a^2 = 1 - \frac{\frac{RSS_{\mathcal{M}}}{n-p_{\mathcal{M}}}}{\frac{TSS}{n-1}} = 1 - (\frac{n-1}{n-p_{\mathcal{M}}})(1-R^2) = 1 - \frac{\hat{\sigma}^2_{\mathcal{M}}}{\hat{\sigma}^2_{null}}$$

We favor models that produce larger R_a^2.

Selection Criteria (Mallow’s C_p)

Mallow’s C_p statistic is a criterion designed to quantify the predictive usefulness of a model.

At its core, Mallow’s C_p statistic is trying to estimate the standardized total mean square prediction error, given by

$$\frac{1}{\sigma^2}\sum_i E(\hat{y}_i - E(y_i))^2 = \frac{1}{\sigma^2}\sum_i MSE(\hat{y}_i)$$

For model $\mathcal{M}$ with $p_{\mathcal{M}}$ regression coefficients, this quantity is estimated by

$$C_{p_{\mathcal{M}}} = \frac{RSS_{\mathcal{M}}}{\hat{\sigma}^2_{\Omega}} + 2p_{\mathcal{M}} - n$$

Intuition behind Mallow’s C_p

- Total error in fitted value of ith observation

$$\hat{Y}_i - \mu_i$$

* Easy to show

$$E[(\hat{Y}_i - \mu_i)^2] = (E(\hat{Y}_i) - \mu_i)^2 + Var(\hat{Y}_i$$

* The total mean squared error

$$\sum_i E[(\hat{Y}_i - \mu_i)^2] = \sum_i (E(\hat{Y}_i) - \mu_i)^2 + \sum_i Var(\hat{Y}_i$$

- The criterion measures

$$\Gamma_p = \frac{1}{\sigma^2}\left[\sum_i (E(\hat{Y}_i) - \mu_i)^2 + \sum_i Var(\hat{Y}_i\right]$$

- Unbiased estimate of Γ_p is the Mallow’s C_p.

Properties of Mallow's C_p

- For the model with all regressors (model Ω with p_Ω regression coefficients), $C_{p_\Omega} = p_\Omega$
- If a model with $p_\mathcal{M}$ regression coefficients fits the data well and has little or no bias, then $E(C_{p_\mathcal{M}}) \approx p_\mathcal{M}$
- When the C_p values for all possible regression models are plotted against p, those models with little bias will tend to fall near the $C_p = p$ line (p on the horizontal axis).
- Models considerably above the line are biased.
- Models bellow the line are considered unbiased and being below the line due to sampling error.
- We favor models with small $p_\mathcal{M}$ and $C_{p_\mathcal{M}}$ close to $p_\mathcal{M}$

selection Criteria (MSE)

The mean squared error (MSE) for prediction is simply the average of the squared deviations between the fitted values and the observed data, i.e., $\frac{1}{n}\sum(y_i - \hat{y}_i)^2$

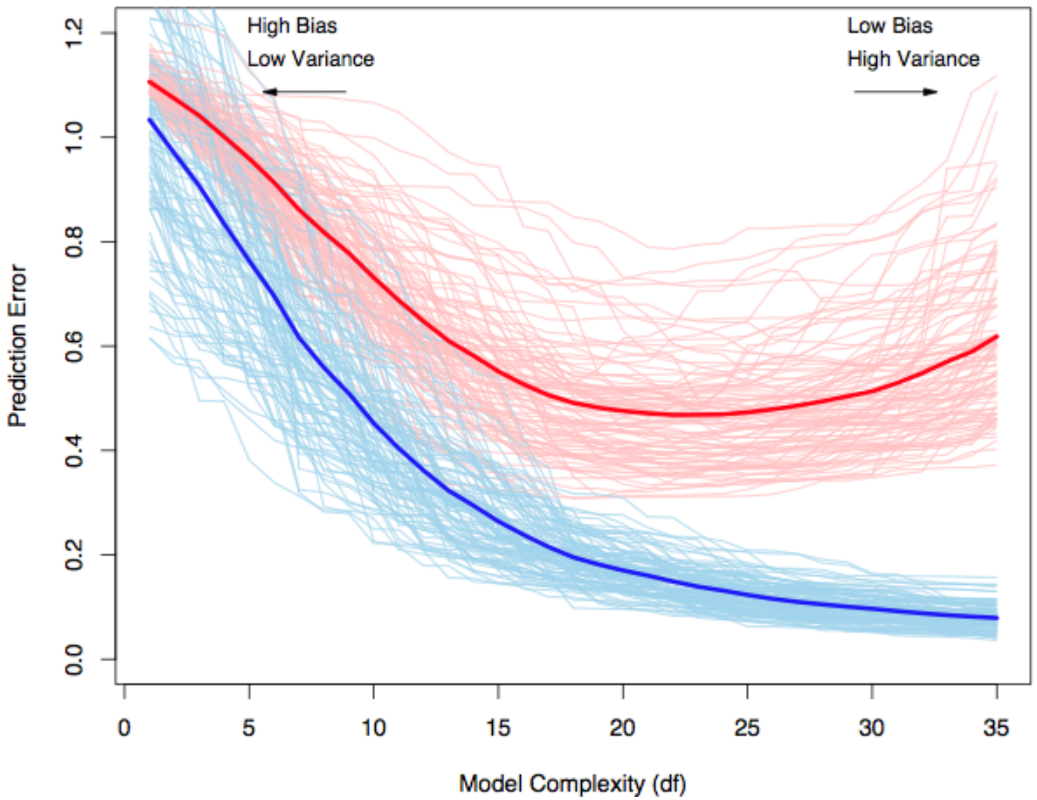
- We favor models with smaller mean squared error, but the search algorithm is very important, otherwise you just use the model with the most regressors.
- The RMSE (root mean squared error) is simply the square root of the MSE, and is sometimes used in place of the MSE.
    - The RMSE or MSE will produce identical variable selection results since they are 1-1 transformations of each other.

Training vs Test Error

Which one will be higher?

Bias-Variance tradeoff

Bias-Variance tradeoff



Courtesy : The Elements of Statistical Learning by Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie. Blue curves show the training errors on 100 samples of size 50. Red curves are the corresponding test set errors

Cross Validation

Cross-validation breaks the data into a training dataset and a test dataset to get a more accurate assessment of the predictive accuracy of a model.

- A model is fit to the training dataset and then the fitted model is used to predict the responses of the test dataset, from which an error criterion (e.g, the MSE) is calculated for the test dataset.
  - We favor the model the minimizes the MSE (or optimizes some other measure of prediction accuracy).

Cross Validation

There are many variations of how to choose the training and testing datasets for crossvalidation.

- Leave-one-out crossvalidation uses each observation (individually) as a test data set, using the other n-1 observations as the training data.
    - In principle, we must fit n models to find the mean squared error, though this can be done using only a single model if you scale things correctly..
- k-fold crossvalidation breaks the data into k unique sets.
    - For each set, the other k-1 sets are used as training data, and then the fitted model is used to predict the responses for the kth testing set.
    - We must fit k models to determine the mean squared error.
- There are other mechanisms for choosing the training and test datasets, but these are the most common.

Cross Validation

When using cross-validation as your selection criterion, we prefer the model that produces the lowest MSE or RMSE.

- You typically don’t do an exhaustive search or stepwise selection search.
- You often use one of the other selection criteria/search strategies to narrow down the possible models to a few final candidate models and then use cross-validation to make a final decision.

Search Strategies

Exhaustive Search

An exhaustive search looks at all possible models using all available regressors.

- This is not feasible unless the number of regressors is relatively small.
- If the number of regressors (including the intercept) is p_Ω, there are 2^{p_Ω} possible models.

Because of our error criteria, our search often simplifies to finding the model that minimizes $RSS_{\mathcal{M}}$ for each value of $p_{\mathcal{M}}$.
- This is the best subset searching strategy. - It’s really just a smart way to do an exhaustive search.

Stepwise

When the previous strategies may take too long, stepwise selection can be used to iteratively build models, choosing the next model as the one that maximizes or minimizes the criterion of interest.

- Backwards selection starts with the model having all regressors, then prunes the regressors one at a time until we can no longer improve the error criterion by removing a single regressor.
- Forward selection starts with the null model (only an intercept), and adds regressors one at a time until we can no longer improve the error criterion by adding a single regressor.
- “Both” (or commonly, stepwise selection) is similar to backward selection, except that we can add a regressor back into the model if it improves the error criterion.

Additional Notes on Model Selection

Stepwise selection can miss the optimal model because we do not consider all possible models due to the one-at-a-time nature of adding/removing regressors.

P-values should not be taken as very accurate in stepwise or best subset searches because we are bound to see small p-values due to chance alone.

Stepwise selection tends to produce simpler models that are not necessarily the best for prediction.

Model Hierarchy

We must respect hierarchy in models when it is naturally present.

- In polynomial models, x^2 is a higher order term than x.
- A lower order term should be retained if a higher order term is retained to increase the flexibility.
  - E.g., for the model $y = \beta_0 + \beta_2 x^2 + \epsilon$, the maximum/minimum value MUST occur at x=0.
  - For the model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$, the maximum/minimum value can occur anywhere along the real line (depending on what the data suggest).
  - Example: If we fit the model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ and ββ_1 is not significant, it would NOT make sense to remove x from the model but still keep x^2 .

Example

The U.S. Bureau of the Census collected data from the 50 states in the 1970s. Measured variables include:

- `Population`: population estimate as of July 1, 1975
- `Income`: per capita income (1974)
- `Illiteracy`: illiteracy (1970, percent of population)
- `Life.Exp`: life expectancy in years (1969–71)
- `Murder`: murder and non-negligent manslaughter rate per 100,000 population (1976)
- `HS Grad`: percent high-school graduates (1970)
- `Frost`: mean number of days with minimum temperature below freezing (1931–1960) in capital or large city
- `Area`: land area in square miles

Backward Selection

```
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  7.0943e+01  1.7480e+00 40.5859 < 2.2e-16
## Population   5.1800e-05  2.9187e-05  1.7748   0.08318
## Income      -2.1804e-05  2.4443e-04 -0.0892   0.92934
## Illiteracy   3.3820e-02  3.6628e-01  0.0923   0.92687
## Murder      -3.0112e-01  4.6621e-02 -6.4590  8.68e-08
## HS.Grad      4.8929e-02  2.3323e-02  2.0979   0.04197
## Frost       -5.7350e-03  3.1432e-03 -1.8246   0.07519
## Area        -7.3832e-08  1.6682e-06 -0.0443   0.96491
##
## n = 50, p = 8, Residual SE = 0.74478, R-Squared = 0.74
```

- Higher murder rates decrease life expectancy!
- Many variables not significant.

Remove least significant predictor (Area):

```
lmod <- update(lmod, . ~ . - Area)
sumary(lmod)
```

```
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  7.0989e+01  1.3875e+00 51.1652 < 2.2e-16
## Population   5.1883e-05  2.8788e-05  1.8023   0.07852
## Income      -2.4440e-05  2.3429e-04 -0.1043   0.91740
## Illiteracy   2.8459e-02  3.4163e-01  0.0833   0.93400
## Murder      -3.0182e-01  4.3344e-02 -6.9634 1.454e-08
## HS.Grad      4.8472e-02  2.0667e-02  2.3454   0.02369
## Frost       -5.7758e-03  2.9702e-03 -1.9446   0.05839
##
## n = 50, p = 7, Residual SE = 0.73608, R-Squared = 0.74
```

Remove least significant predictor (Illiteracy)

```
lmod <- update(lmod, . ~ . - Illiteracy)
sumary(lmod)
```

```
##                 Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  7.1066e+01  1.0289e+00 69.0669 < 2.2e-16
## Population   5.1149e-05  2.7095e-05  1.8878   0.06566
## Income      -2.4771e-05  2.3160e-04 -0.1070   0.91531
## Murder      -3.0001e-01  3.7042e-02 -8.0992 2.907e-10
## HS.Grad      4.7758e-02  1.8591e-02  2.5689   0.01367
## Frost       -5.9099e-03  2.4678e-03 -2.3948   0.02095
##
## n = 50, p = 6, Residual SE = 0.72773, R-Squared = 0.74
```

Remove least significant predictor (Income)

```
lmod <- update(lmod, . ~ . - Income)
sumary(lmod)
```

```
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) 71.02712853  0.95285296 74.5415 < 2.2e-16
## Population   0.00005014  0.00002512  1.9960  0.052005
## Murder      -0.30014880  0.03660946 -8.1987 1.775e-10
## HS.Grad      0.04658225  0.01482706  3.1417  0.002968
## Frost       -0.00594329  0.00242087 -2.4550  0.018018
##
## n = 50, p = 5, Residual SE = 0.71969, R-Squared = 0.74
```

Remove `Population`?

Whether we should remove Population is a close call. We should probably keep it if it makes the model more interpretable.

```
lmod <- update(lmod, . ~ . - Population)
sumary(lmod)
```

```
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 71.0363788  0.9832622 72.2456 < 2.2e-16
## Murder      -0.2830652  0.0367313 -7.7064 8.039e-10
## HS.Grad      0.0499487  0.0152011  3.2859  0.001950
## Frost       -0.0069117  0.0024475 -2.8240  0.006988
##
## n = 50, p = 4, Residual SE = 0.74267, R-Squared = 0.71
```

All variables are now significant at $\alpha_{crit} = 0.05$

How much improvement

The R^2 for the full model is 0.736. Our final model has an R^2 of 0.713, which is only slightly lower.

- Removal of four predictors causes only a minor reduction in fit. This is NOT surprising.
- A better question might be: what would the effect of removing these variables be on a new independent sample?

Is it a good practice?

Eliminated variables may still be important.

- Replacing HS.Grad with Illiteracy …

```
sumary(lm(Life.Exp ~ Illiteracy + Murder + Frost, statedata))
```

```
##                Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 74.5567171  0.5842507 127.6108 < 2.2e-16
## Illiteracy  -0.6017607  0.2989270  -2.0131  0.049981
## Murder      -0.2800474  0.0433940  -6.4536 6.033e-08
## Frost       -0.0086910  0.0029595  -2.9367  0.005166
##
## n = 50, p = 4, Residual SE = 0.79112, R-Squared = 0.67
```

- Illiteracy does have some association with life expectancy
- High school graduation rate and illiteracy are likely correlated
- Impossible to know which is the important/causal variable
- Both could be important or both could be proxies for a third variable.

Example 2

Use best subset selection to minimize the AIC criterion.

- The regsubsets function in the leaps package can be used to do this.
- For each number of regression coefficients p, it finds the model that minimizes the RSS.
    - For each value of p, the model that minimizes the RSS will have the smallest AIC, BIC, R_a^2, and Mallow’s C_p.
- NOTE: By default, regsubsets only goes up to p=9. You have to set nvmax = j, where j is the number of regressors you want to consider.

Example 2

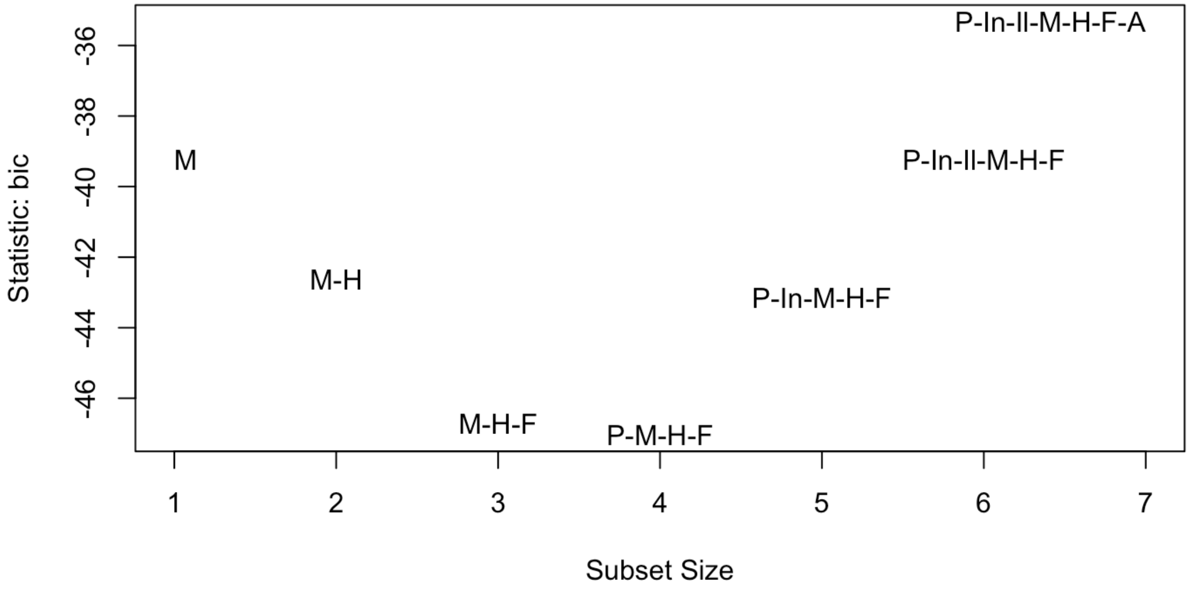
```
b <- leaps::regsubsets(Life.Exp ~ ., data = statedata)
rs <- summary(b)
rs$which
```

```
##   (Intercept) Population Income Illiteracy Murder HS.Grad Frost  Area
## 1        TRUE      FALSE  FALSE      FALSE   TRUE   FALSE FALSE FALSE
## 2        TRUE      FALSE  FALSE      FALSE   TRUE    TRUE FALSE FALSE
## 3        TRUE      FALSE  FALSE      FALSE   TRUE    TRUE  TRUE FALSE
## 4        TRUE       TRUE  FALSE      FALSE   TRUE    TRUE  TRUE FALSE
## 5        TRUE       TRUE   TRUE      FALSE   TRUE    TRUE  TRUE FALSE
## 6        TRUE       TRUE   TRUE       TRUE   TRUE    TRUE  TRUE FALSE
## 7        TRUE       TRUE   TRUE       TRUE   TRUE    TRUE  TRUE  TRUE
```

Example 2

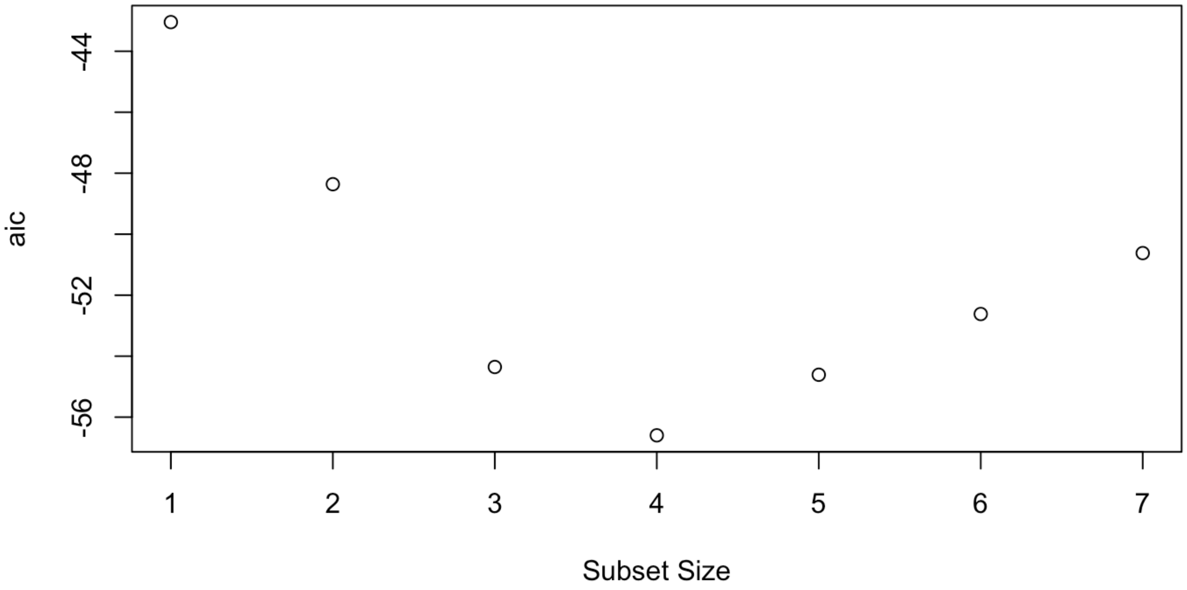
Plot of several selection criteria versus the best subsets are available via the subsets function in the car package.

```
car::subsets(b, legend = F)
```



Example 2

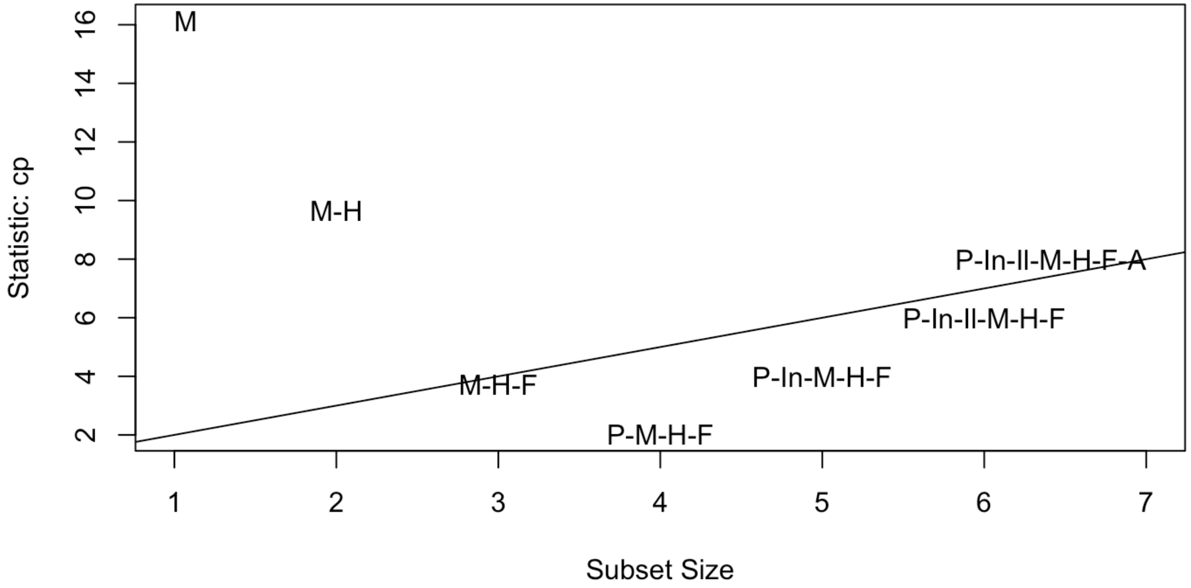
```
p = 2:8; pp = p-1
aic <- rs$bic + p *(2 - log(nrow(state.x77)))
plot(aic ~ pp, xlab = 'Subset Size')
```



The model with $p = 5$ is best. This model includes the predictors: population, murder, high school graduation rate, and frost.

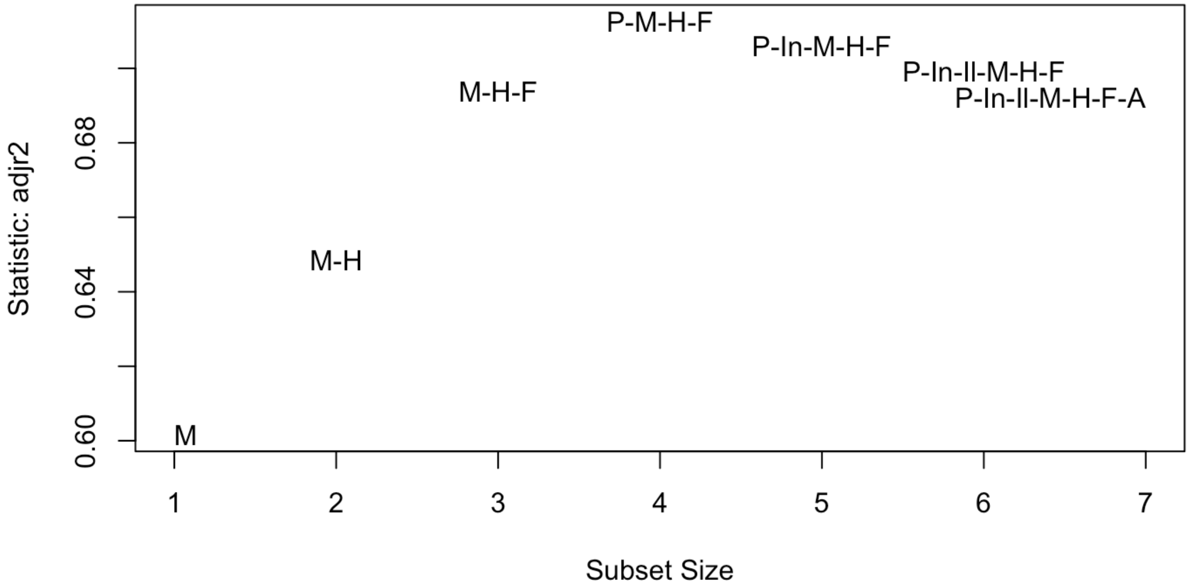
Example 2

```
car::subsets(b, statistic = "cp", legend = FALSE)
abline(1, 1) # corresponds to 45 degree line offset by 1 unit vertically
```



Example 2

```r
car::subsets(b, statistic = "adjr2", legend = FALSE)
```



Stepwise with AIC

```
lmod = lm(Life.Exp ~ ., data = statedata)
step(lmod, direction = 'both')
```

```
## Start:  AIC=-22.18
## Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad + 
##     Frost + Area
## 
##              Df Sum of Sq    RSS     AIC
## - Area        1    0.0011 23.298 -24.182
## - Income      1    0.0044 23.302 -24.175
## - Illiteracy  1    0.0047 23.302 -24.174
## <none>                    23.297 -22.185
## - Population  1    1.7472 25.044 -20.569
## - Frost       1    1.8466 25.144 -20.371
## - HS.Grad     1    2.4413 25.738 -19.202
## - Murder      1   23.1411 46.438  10.305
## 
## Step:  AIC=-24.18
## Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad + 
##     Frost
## 
##              Df Sum of Sq    RSS     AIC
```

Example (Cross Validation)

Comparison of full model to model with Population, Murder, HS.Grad, and Frost predictors using the RMSE criterion and both 10-fold crossvalidation and leave-one-out crossvalidation.

```r
library(caret)
# define training/test (control) data
cv_10fold = trainControl(method = "cv", number = 10) # 10-fold crossvalidation train/test data
cv_loo = trainControl(method = "LOOCV") # leave-one-out crossvalidation train/test data
cv_loo_slow = trainControl(method = "cv", number = 50) # loo crossvalidation train/test data

# train the full model
f1 = Life.Exp ~ . # formula for full model
# formula for reduced model with p = 5
f2 = Life.Exp ~ Population + Murder + HS.Grad + Frost
# formula for reduced model 2 with p = 4
f3 = Life.Exp ~ Murder + HS.Grad + Frost
```

Continue

```
# the train function needs:
# formula - to formula for the model you want to fit,
# data - the data frame where the variables are located
# trControl - the training/testing data sets created using the trainControl function
# method - the type of model you want to fit.  There are a lot of choices.  We simply need "lm
modela = train(f1, data = statedata, trControl = cv_10fold,
               method = "lm")
modelb = train(f2, data = statedata, trControl = cv_10fold,
               method = "lm")
```

Continue

```
# compare mse (rmse) for the two models using 10-fold cv
print(modela$results) # full, 10-fold
```

```
##   intercept      RMSE  Rsquared       MAE    RMSESD RsquaredSD     MAESD
## 1      TRUE 0.8559381 0.5932933 0.7133981 0.2849117  0.3111669 0.2432284
```

```
print(modelb$results) # reduced, 10-fold
```

```
##   intercept     RMSE  Rsquared       MAE    RMSESD RsquaredSD    MAESD
## 1      TRUE 0.730817 0.7332559 0.6239986 0.2149312  0.2036376 0.201853
```

The smaller model is preferred (since it has smaller RMSE) using 10-fold crossvalidation.

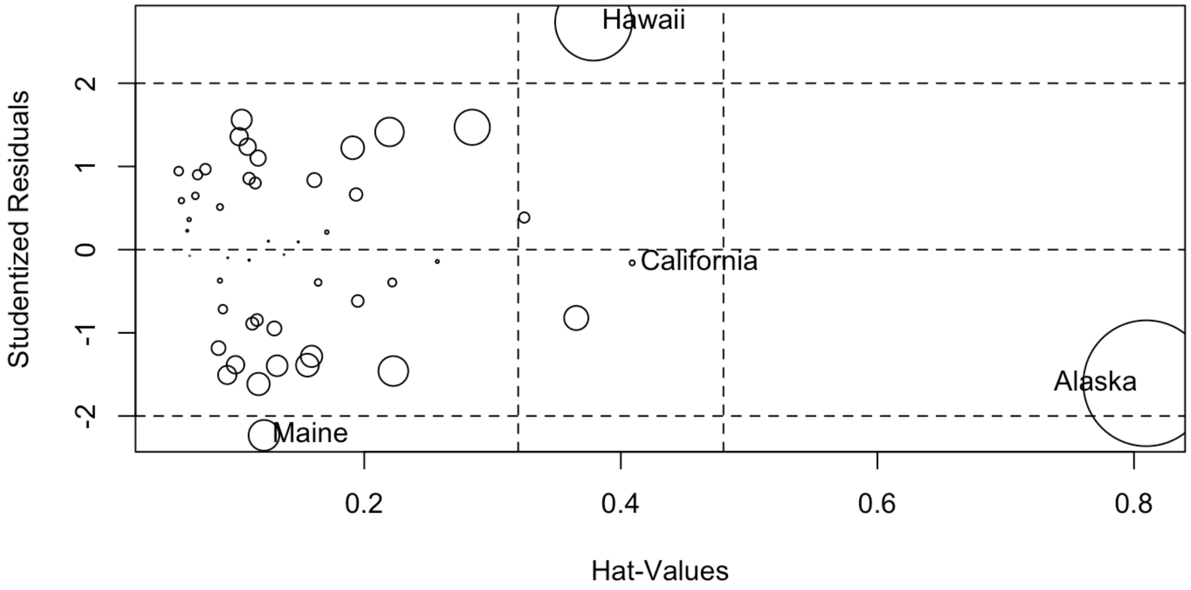
```
modelc = train(f1, data = statedata, trControl = cv_loo,
               method = "lm")
modeld = train(f2, data = statedata, trControl = cv_loo,
               method = "lm")
# compare mse (rmse) for the two models using 10-fold cv
print(modelc$results) # full 2, LOO
```

```
##   intercept      RMSE  Rsquared       MAE
## 1      TRUE 0.9090885 0.5469535 0.7196334
```

Influence of outliers

Variable selection can be affected by outliers and transformations.

Alaska is a high leverage point. What's the effect if we remove it?



```
##                  StudRes       Hat      CookD
```

Removing Alaska

```
b <- regsubsets(Life.Exp ~., data = statedata, subset = (state.abb!="AK"))
rs <- summary(b)
rs$which[which.max(rs$adjr), ]
```

```
## (Intercept)  Population      Income  Illiteracy      Murder     HS.Grad
##        TRUE        TRUE       FALSE       FALSE        TRUE        TRUE
##       Frost        Area
##        TRUE        TRUE
```

We now choose a 5 regressor model using R_a^2, whereas we chose 4 before.