7 Variable Selection

In the presentation so far we always assumed to know which variables (predictors) have to be included with a model, but in many application we face a data exploration problem, in which we try to determine the "best" model for the response variable choosing predictors from a given set.
We will have to define a "best" model, and then investigate different strategies for finding the "best" model. The different model selection strategies though will not necessarily result in the "best" model as we sill see.
Before discussing the process of choosing a model the effect of model misspecification shall be presented.

7.1 Model misspecification effect

Assume the true (complete) model describing the population is

$$Y = \beta_0 + \sum_{j=1}^{K} \beta_j x_j + \varepsilon$$

with K predictors. To estimate the model parameters the sample size n has to be at least $K+1$, because otherwise $X'X$ can not have full rank.
Therefore assume $n \geq K+1$, then the model for a sample of size n is written in matrix form as

$$\vec{Y} = X\vec{\beta} + \vec{\varepsilon}$$

or as

$$Y_i = \beta_0 + \sum_{j=1}^{K} \beta_j x_{ij} + \varepsilon_i, \quad 1 \leq i \leq n$$

We will study the effect of omitting r variables from the model. The complete model has the $K+1$ parameters and the incomplete model has $p := K+1-r$ parameters, when arranging the columns and components of $\vec{\beta}$ properly the complete model can be written as

$$\vec{Y} = X_p\vec{\beta}_p + X_r\vec{\beta}_r + \vec{\varepsilon}$$

where X_p is the design matrix including the the column for the intercept and the columns of X for the variables included with the incomplete model and X_r includes the columns of X for the variables not included with the incomplete model. $X = [X_p \mid X_r]$.
Assume that $\vec{\beta}$ is partitioned according to the partitioning of X into X_p and X_r. $\vec{\beta}' = [\vec{\beta}'_p \mid \vec{\beta}'_r]$.
Let $\hat{\beta}^*$ be the least squares estimator for the true model and $\hat{\sigma}^2_*$ the least squares estimator for σ^2 from the true model.
Recall

$$\hat{\beta}^* = (X'X)^{-1}X'\vec{Y}, \quad \hat{\sigma}^2_* = \frac{\vec{Y}'(I-H)\vec{Y}}{n-K-1}$$

The part of $\hat{\beta}^*$ relating to X_p and X_r are denoted by $\hat{\beta}^*_p$ and $\hat{\beta}^*_r$, respectively. Let $\hat{Y}^*$ denote the vector of fitted values.

Now consider the incomplete model missing all predictors in X_r.

$$\vec{Y} = X_p\vec{\beta}_p + \vec{\varepsilon}$$

$\hat{\beta}_p$ is the least squares estimator for $\vec{\beta}_p$ and $\hat{\sigma}^2$ the least squares estimator for σ^2 from the incomplete model. It is

$$\hat{\beta}_p = (X_p'X_p)^{-1}X_p'\vec{Y}, \quad \hat{\sigma}^2 = \frac{\vec{Y}'(I-H_p)\vec{Y}}{n-(K+1-r)} = \frac{\vec{Y}'(I-X_p(X_p'X_p)^{-1}X_p')\vec{Y}}{n-p}$$

and the fitted values are denoted as $\hat{Y}$.

The mean square error of an estimator $\hat{\psi}$ for a parameter vector $\vec{\psi}$ is defined as

$$MSE(\hat{\psi}) = E[||\hat{\psi}-\vec{\psi}||^2] = trace(Cov(\hat{\psi})) + ||Bias(\hat{\psi})||^2$$

Lemma 1.
It is

1. $E(\hat{\beta}_p) = \vec{\beta}_p + (X_p'X_p)^{-1}X_p'X_r\vec{\beta}_r = \vec{\beta}_p + A\vec{\beta}_r$
2. $Cov(\hat{\beta}_p) = \sigma^2(X_p'X_p)^{-1}$ and $Cov(\hat{\beta}^*) = \sigma^2(X'X)^{-1}$.
3. $Cov(\hat{\beta}_p^*) - Cov(\hat{\beta}_p)$ is positive semidefinite. (A is positive definite, if $\vec{x}'A\vec{x} \geq 0 \ \ \forall\ \vec{x}$).
4. $MSE(\hat{\beta}_p) = \sigma^2 trace((X_p'X_p)^{-1}) + \vec{\beta}_r'A'A\vec{\beta}_r$
5. $E(\hat{\sigma}^2) = \sigma^2 + \dfrac{\vec{\beta}_r'X_r'(I-H_p)X_r'\vec{\beta}_r}{n-p}$
6. For a prediction of the response, if the predictor vector is $\vec{x}_p$, we get $\hat{Y} = \vec{x}_p{}'\hat{\beta}_p$, then $E(\hat{Y}) = \vec{x}_p'\vec{\beta}_p + x_p'A\vec{\beta}_r$ and prediction Mean Square Error

$$MSE(\hat{Y}) = \sigma^2[1 + \vec{x}_p{}'(X_p'X_p)^{-1}\vec{x}_p] + (\vec{x}_p{}'A\vec{\beta}_r - \vec{x}_r{}'\vec{\beta}_r)^2.$$

Conclusions:

1. From 1. we find that $\hat{\beta}_p$ is not unbiased unless $\vec{\beta}_r = \vec{0}$ or X_p and X_r are orthogonal.
2. From 3. estimates from the complete model have a higher (or equal) variance than the equivalent estimates from the subset model. *Therefore removing variables from the model does not increase the variance of estimates for the remaining parameters. The estimates are in general biased but have smaller variance.*
3. From 4. we can derive that if a variable x_i from the true model is omitted from the model, but has a "small" slope β_i then the MSE for the slopes of remaining variables can be smaller for the incomplete model than for the complete true model.

4. From 5. $\hat{\sigma}^2$ is a biased estimator for σ^2 if $\vec{\beta}_r \neq \vec{0}$ and X_p and X_r are not orthogonal, in fact the estimate for σ^2 from the incomplete model is overestimating the true value.

5. From 6. $\hat{Y}$ is a biased estimate of Y if $\vec{\beta}_r \neq \vec{0}$ and X_p and X_r are not orthogonal. The MSE from the incomplete model can be smaller than for the complete model if the slope of the omitted variables is small enough.

Therefore, removing relevant variables from the model adds bias to the estimates, if the design is not orthogonal. On the other hand the variance in the estimates is smaller than for the complete model. Taking these two together, when omitting variables from the model with only small effects on the response the MSE for the estimates based on the reduces model can be smaller than for the complete model.

Adding variables of no relevance to the model, also adds bias, and increases variance, which is always bad.

7.2 Evaluating Models

The most commonly used criteria for model selection are the Adjusted Coefficient of Determination, R_a^2, and the Residual Sum of Squares, SS_{Res}:

1. The residual sum of squares measure for a model M, $SS_{Res}(M)$, gives the amount of variation, which can not be explained by the model. Therefore it is a reasonable goal to find the model with the smallest $SS_{Res}(M)$, but adding a variable to a given model never increases $SS_{Res}(M)$. Therefore the model with all K variables will always be "best" in terms of minimizing $SS_{Res}(M)$. Therefore conducting (extra sum of squares) tests if the addition of another variable significantly decreases $SS_{Res}(M)$ to decide if the variable should be included with the "best" model are conducted.

   I.e. an optimal model, is such that the addition of any of the not included variables is not significantly decreasing $SS_{Res}(M)$, AND the removal of a variable in the model results in a significant increase in the $SS_{Res}(M)$.

   Since the definition is not based on a global measure, several models might satisfy the above definition of an optimal model.

2. Recall that for a model M

$$R_a^2(M) = 1 - \frac{SS_{Res}(M)/n - p}{SS_T/(n-1)} = 1 - \frac{MS_{Res}(M)}{MS_T}$$

   SS_T and MS_T are the same for all models, since they measure the total variance in the response, regardless of the predictors.

   $R_a^2(M)$ is adjusted for the number of predictors in the model, p. It does not increase for every variable added to the model and therefore is a good measure of fit.

   The "best" model in regard to $R_a^2(M)$ is the model M with largest $R_a^2(M)$.

Alternative measures are Akaike's Information Criterion (AIC), and the Bayesian Information Criterion (BIC, from Schwartz).
Both are based on the likelihood, L, of the estimated model and a penalty for the number of slope parameters in the model.

$$AIC = -2ln(L) + 2p, \qquad BIC = -2ln(L) + pln(n)$$

We prefer model with a large like lihood, which means with small AIC and BIC.
Of these two measures BIC is used more often used in applied statistics.

If the number of variables is the same in two models the penalty for the two models is the same and AIC, BIC, and R^2_{adj} will prefer the same model.
The penalty for extra parameters is different and therefore the three measures can potentially lead to the recommendation of different models.

7.3 Techniques for Model Selection

To find the SS_{Res} optimal and R^2_a best models all possible models can be analyzed.

1. In order to find all the optimal models in regard to the residual sum of squares, for each model the conditions for optimality have to be checked. For K variables there are 2^K possible models to be considered, for each model K tests have to be conducted.
   For $K = 3$, there are 24 tests.
   For $K = 4$, there are 56 tests.
   For $K = 10$, there are 10240 tests.
   This seems unreasonable.

2. To find an R^2_a best model, at least all models have to be fit and evaluated, therefore 2^K models have to be analyzed. Even this is too much ($K = 10$, then $2^{10} = 1024$).

This might be doable for up to 4 variables, but beyond the amount of tests and analyses seem computationally unreasonable.
Stepwise procedures have been developed to help finding optimal models, all three procedures result in optimal models, but the optimal models might be all different.

For a model M and a variable v, the model $M + v$ is defined to be the model including all variables in M and v. Analogously define $M - v$.

Forward selection

1. Let $i = 0$. The initial model is M_0 including only the intercept.

2. Until no variable v can be added to M_i to significantly improve $SS_{Res}(M_i)$.

   Use the extra sum of squares test for each variable v not in M_i, if its addition significantly decreases $SS_{Res}(M_i)$.

   If such a variable can be found add variable v with largest F-statistic $F_0(v)$.

   $i \leftarrow i + 1$.

Backward elimination

1. Let $i = 0$. The initial model is M_0 including all K variables.
2. Until no variable v can be eliminated from M_i which does not significantly improve $SS_Res(M_i - v)$.

   Use the extra sum of squares test for each variable v in M_i to test if its addition to $M_i - v$ significantly decreases $SS_Res(M_i - v)$.

   If such a variable can be found eliminate variable v with smallest F-statistic $F_0(v)$.

   $i \leftarrow i + 1$.

Stepwise Regression
For stepwise regression forward selection and backward elimination are combined.
Before considering to enter another variable, all variables entered so far are reassessed. If it is determined that a variable should be eliminated after others have been added, this is done at this time.

In contrast to the other procedures the number of variables in the model can go up AND down during this procedure, whereas in the the other procedures, either the model was strongly increasing or decreasing.

Example 7.1.
We will use the data on academic success in Californian Elementary schools to illustrate the three procedures.

The variables in the data set:
api00 - academic performance of the school,
acs-k3 - average class size in kindergarten through 3rd grade,
acs-46 - average class size in grades 4-6
meals - percentage of students receiving free meals,
full - percentage of teachers who have full teaching credentials,
growth - growth 1999 to 2000
ell - English language learners
hsg - at least one parent is a Highschool graduate
some-col - percent of children with at least one parent with some college education
enroll - enrollment

To do automatic model selection with R, use library "leaps". The function "regsubsets" in this library can be used to find optimal models depending on the number of variables in the model.

```
library(leaps)

nona<-na.omit(elemapi.data)
attach(nona)
```

```
#for finding best model with up to 6 variables
l<-regsubsets(x=cbind(growth,meals,ell,acs_k3,acs_46,some_col,full),y=api00,
method="exhaustive", intercept=TRUE, nvmax=6)
summary(l)

Subset selection object
7 Variables  (and intercept)

1 subsets of each size up to 6
Selection Algorithm: exhaustive
         growth meals ell acs_k3 acs_46 some_col full
1  ( 1 ) " "    "*"   " " " "    " "    " "      " "
2  ( 1 ) " "    "*"   " " " "    " "    " "      "*"
3  ( 1 ) " "    "*"   "*" " "    " "    " "      "*"
4  ( 1 ) "*"    "*"   "*" " "    " "    " "      "*"
5  ( 1 ) "*"    "*"   "*" " "    "*"    " "      "*"
6  ( 1 ) "*"    "*"   "*" " "    "*"    "*"      "*"

f<-regsubsets(x=cbind(growth,meals,ell,acs_k3,acs_46,some_col,full),y=api00,
method="forward",intercept=TRUE, nvmax=4)
summary(f)

Subset selection object
7 Variables  (and intercept)

1 subsets of each size up to 6
Selection Algorithm: forward
         growth meals ell acs_k3 acs_46 some_col full
1  ( 1 ) " "    "*"   " " " "    " "    " "      " "
2  ( 1 ) " "    "*"   " " " "    " "    " "      "*"
3  ( 1 ) " "    "*"   "*" " "    " "    " "      "*"
4  ( 1 ) "*"    "*"   "*" " "    " "    " "      "*"
5  ( 1 ) "*"    "*"   "*" " "    "*"    " "      "*"
6  ( 1 ) "*"    "*"   "*" " "    "*"    "*"      "*"

b<-regsubsets(x=cbind(growth,meals,ell,acs_k3,acs_46,some_col,full),y=api00,
method="backward",intercept=TRUE, nvmax=6)
summary(b)

Subset selection object
7 Variables  (and intercept)

1 subsets of each size up to 6
Selection Algorithm: backward
```

```
          growth meals ell acs_k3 acs_46 some_col full
1  ( 1 ) " "    "*"   " " " "    " "    " "      " "
2  ( 1 ) " "    "*"   " " " "    " "    " "      "*"
3  ( 1 ) " "    "*"   "*" " "    " "    " "      "*"
4  ( 1 ) "*"    "*"   "*" " "    " "    " "      "*"
5  ( 1 ) "*"    "*"   "*" " "    "*"    " "      "*"
6  ( 1 ) "*"    "*"   "*" " "    "*"    "*"      "*"
```

In the first output it shows the best models for 1, 2, up to 6 variables in the model. When the number of variables is fixed, all criteria lead to the same model, only when comparing models with different number of variables decision might deviate.
The second output, shows the best model for 1, 2, 3, etc. variables choosing the model using the forward selection procedure.
The last output, shows the best model for 1, 2, 3, etc. variables choosing the model using the backward elimination procedure.
Surprisingly all methods result in the same models for all the different number of variables. This is not necessary.