Model Building and Variable Selection

Dr. Mutua Kilai

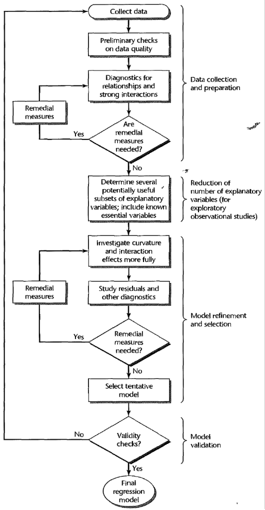
Department of Pure and Applied Sciences

2024-01-13

Overview of Model Building Process

- The strategy of building a regression model include the following phases:
    i. Data collection and preparation
    ii. Reduction of explanatory or predictor variables
    iii. Model refinement and selection
    iv. Model validation

Cont'd



Collect data
Preliminary checks on data quality
Diagnostics for relationships and strong interactions
Remedial measures
Yes
Are remedial measures needed?
No
Data collection and preparation
Determine several potentially useful subsets of explanatory variables; include known essential variables
Reduction of number of explanatory variables (for exploratory observational studies)
Investigate curvature and interaction effects more fully
Remedial measures
Study residuals and other diagnostics
Yes
Remedial measures needed?
No
Select tentative model
Model refinement and selection
No
Validity checks?
Yes
Model validation
Final regression model

Variable selection for prediction

- Variable selection means choosing among many variables which to include in a particular model, that is, to select appropriate variables from a complete list of variables by removing those that are irrelevant or redundant.
- The purpose of such selection is to determine a set of variables that will provide the best fit for the model so that accurate predictions can be made
- Variable Selection serves two purposes:
  i. It helps determine all of the variables that are related to the outcome, which makes the model complete and accurate
  ii. Second, it helps select a model with few variables by eliminating irrelevant variables that decrease the precision and increase the complexity of the model.

Variable Selection Steps

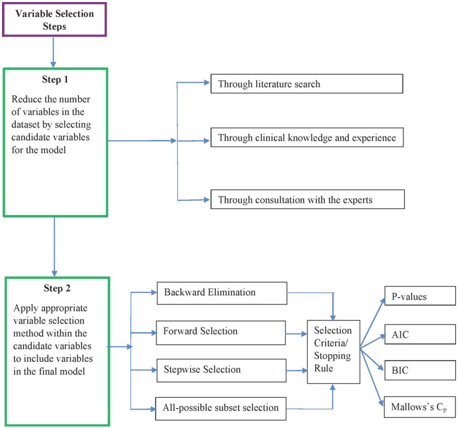


Figure 1: Variable Selection Steps

Variable Selection Procedures

- Chowdhury and Turin (2020) listed some variable selection methods:
  i. Backward Elimination
  ii. Forward selection
  iii. Stepwise selection
  iv. All possible subset selection

Backward Selection

- Backward elimination is the simplest of all variable selection methods.
- This method starts with a full model that considers all of the variables to be included in the model.
- Variables then are deleted one by one from the full model until all remaining variables are considered to have some significant contribution to the outcome.
- The variable with the smallest test statistic (a measure of the variable's contribution to the model) less than the cut-off value or with the highest p value greater than the cut-off value—the least significant variable—is deleted first.
- This process is repeated until every remaining variable is significant at the cut-off value.

Advantages and Disadvantages of Backward Elimination

- Backward elimination has the advantage to assess the joint predictive ability of variables as the process starts with all variables being included in the model.
- Backward elimination also removes the least important variables early on and leaves only the most important variables in the model.
- One disadvantage of the backward elimination method is that once a variable is eliminated from the model it is not re-entered again.

Example

- We use the `evals.csv` The data are gathered from end of semester student evaluations for 463 courses taught by a sample of 94 professors from the University of Texas at Austin. In addition, six students rate the professors' physical appearance. .
- There are 23 variables for each listing.
- **Our goal will be to build a model that predicts the score from the remaining variables.**
- The following packages are used `olsrr, glmnet`

Data

```
library(olsrr)
library(glmnet)
data = read.csv("evals.csv")
```

- First we run a regular linear regression using lm(), with score as the dependent variable:

```
model <- lm(score ~., data = data)
```

- And then we use ols_step_backward_aic()

Results

```r
library(olsrr)
library(glmnet)
data <-  read.csv("evals.csv")
model <- lm(score ~., data = data)
backward <- ols_step_backward_aic(model)
print(backward)
```

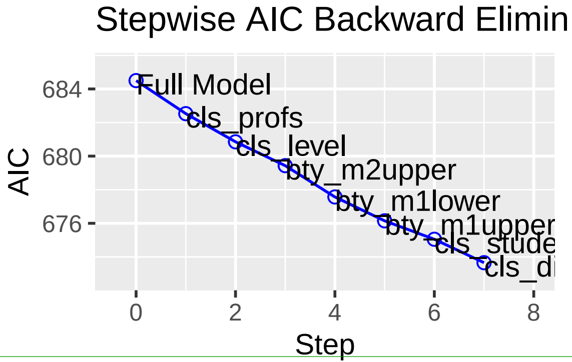
Cont'd

```
##
##
##                        Backward Elimination Summary
## --------------------------------------------------------
## Variable            AIC       RSS       Sum Sq     R-Sq
## --------------------------------------------------------
## Full Model        684.496   107.651    29.004    0.21224
## cls_profs         682.529   107.658    28.996    0.21219
## cls_level         680.852   107.733    28.921    0.21164
## bty_m2upper       679.433   107.869    28.786    0.21065
## bty_m1lower       677.570   107.901    28.754    0.21041
## bty_m1upper       676.148   108.035    28.619    0.20943
## cls_students      675.051   108.246    28.408    0.20788
## cls_did_eval      673.656   108.388    28.266    0.20685
## --------------------------------------------------------
```

Cont'd

- In the table above, we can see the names of 7 variables that are eliminated from the model.

```
plot(backward)
```



Forward Selection

- The forward selection method of variable selection is the reverse of the backward elimination method.
- The method starts with no variables in the model then adds variables to the model one by one until any variable not included in the model can add any significant contribution to the outcome of the model.
- At each step, each variable excluded from the model is tested for inclusion in the model. If an excluded variable is added to the model, the test statistic or p value is calculated.
- This process continues until no remaining variable is significant at the cut-off level when added to the model.
- In forward selection, if a variable is added to the model, it remains there.

Advantages and Disadvantages

- One advantage of forward selection is that it starts with smaller models.
- Also, this procedure is less susceptible to collinearity (very high intercorrelations or interassociations among independent variables).
- Inclusion of a new variable may make an existing variable in the model non-significant; however, the existing variable cannot be deleted from the model.

Example

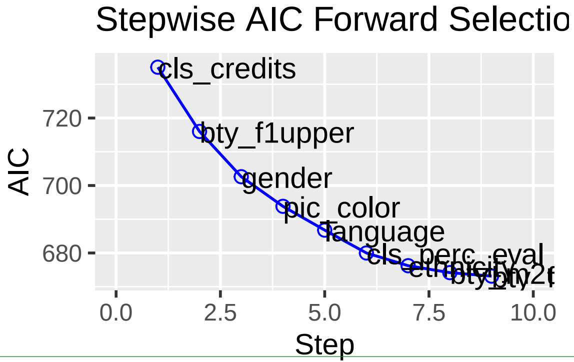
```r
library(olsrr)
library(glmnet)
data <-  read.csv("evals.csv")
model <- lm(score ~., data = data)
forward <- ols_step_forward_aic(model)
print(forward)
```

```
##
##                               Selection Summary
## ----------------------------------------------------------------
## Variable               AIC        Sum Sq        RSS         R-Sq
## ----------------------------------------------------------------
## cls_credits          735.077       5.742      130.913      0.04202
## bty_f1upper          716.019      11.563      125.092      0.08461
## gender               702.609      15.657      120.997      0.11458
```

Cont'd

- In the table above, we can see the names of 9 variables that are included in the model.

```
plot(forward)
```



Stepwise selection

- Stepwise selection methods are a widely used variable selection technique, particularly in medical applications.
- This method is a combination of forward and backward selection procedures that allows moving in both directions, adding and removing variables at different steps.
- The process can start with both a backward elimination and forward selection approach
- The stepwise selection method is perhaps the most widely used method of variable selection.
- One reason is that it is easy to apply in statistical software. This method allows researchers to examine models with different combinations of variables that otherwise may be overlooked.

Cont'd

- Build regression model from a set of candidate predictor variables by entering and removing predictors based on p values, in a stepwise manner until there is no variable left to enter or remove any more.

```r
# stepwise regression
library(olsrr)
library(glmnet)
data <-  read.csv("evals.csv")
model <- lm(score ~., data = data)
forward <- ols_step_forward_aic(model)
ols_step_both_p(model)
```

All Possible Subset selection

- In all possible subset selection, every possible combination of variables is checked to determine the best subset of variables for the prediction model.
- With this procedure, all one-variable, two-variable, three-variable models, and so on, are built to determine which one is the best according to some specific criteria.
- If there are K variables, then there are 2K possible models that can be built.
- The R function regsubsets() [leaps package] can be used to identify different best models of different sizes.
- You need to specify the option nvmax, which represents the maximum number of predictors to incorporate in the model.

Example

```r
library(leaps)
data <-  read.csv("evals.csv")
models <- regsubsets(score ~., data = data, nvmax = 5)
summary(models)
```

Criteria for Model selection

- From any set of $p-1$ predictors 2^p-1 alternative models can be constructed.
- Model selection procedures have been developed to identify a small group of regression models that are good according to specified criterion.
- We will assume that the number of observations exceeds the maximum number of potential parameters.

$$n > P$$

Cont'd

- We will focus on the following criterion for model selection:

    i. R_p^2

    ii. $R_{\alpha,p}^2$

    iii. Mallows C_p criterion

    iv. AIC

    v. SBC_p

    vi. $PRESS_p$

R_p^2

- The R^2 criterion calls for the use of the coefficient of multiple determination in order to identify several good subsets of X.
- The subset with the highest R^2 is chosen.
- The
$$R^2 = 1 - \frac{SSE}{SST}$$
- $0 \leq R^2 \leq 1$

$R^2_{\alpha,p}$

- Since R^2 does not take into account of the number of parameters in the regression model and since $max(R^2)$ can never decrease as p increases, the adjusted multiple coefficient of determination is used.

$$R^2_{\alpha,p} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SST}$$

- This coefficient takes the number of parameters in the regression model into account through the degrees of freedom.
- $R^2_{\alpha,p}$ increases if and only if MSE_p decreases since $\frac{SST}{n-1}$ is fixed for the given Y observations.

Mallows C_p criterion

- This coefficient takes the number of parameters in the regression model into account through the degrees of freedom.
- The total mean squared error for all n fitted values $\hat{Y}_i$ is the sum of the n individual mean squared errors:

$$\sum_{i=1}^{n}[(E\{\hat{Y}_i\}-\mu_i)^2+\sigma^2\{\hat{Y}_i\}] = \sum_{i=1}^{n}(E\{\hat{Y}_i\}-\mu_i)^2+\sum_{i=1}^{n}\sigma^2\{\hat{Y}_i\} \quad (1)$$

- The criterion measure, denoted by Γ_p, is simply the total mean squared error divided by σ^2 the true error variance

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

Cont'd

- The estimator of Γ_p is C_p

$$C_p = \frac{SSE}{MSE} - (n - 2p)$$

- In using the C_p criterion, we seek to identify subsets of X variables for which (I) the C_p value is small and (2) the C_p value is near p.

AIC and *SBC* criteria

- Two popular alternatives that also provides penalties for adding predictors are *AIC* and *SBC*.
- $$AIC = n \ln SSE - n \ln n + 2p$$
- $$SBC = n \ln SSE - n \ln n + [\ln n]p$$
- Models with smallest AIC and SBC are considered.

$PRESS_p$ criterion

- The $PRESS_p$ (prediction sum of squares) criterion is a measure of how well the use of the fitted values for a subset model can predict the observed responses.
- The $PRESS_p$ criterion is the sum of the squared prediction errors over all the n cases

$$PRESS_p = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)$$

- Models with small $PRESS_p$ values are considered good candidate models.

Model Validation

- Model validation usually involves checking a candidate model against independent data.
- Three basic ways of validating a regression model are:
  i. Collection of new data to check the model and its predictive ability.
  ii. Comparison of results with theoretical expectations, earlier empirical results, and simulation results.
  iii. Use of a holdout sample to check the model and its predictive ability .

Thank You!