Contents

Chapter 6

Variable Screening Methods

In this chapter, we explain the importance of the deterministic portion of a linear model. We also present models with only numerical variables, only categorical variables, or both types of variables. Finally, we explain some basic procedures for building good linear models.

6.1 Introduction: Why Use a Variable Screening Method

One of the common problems in a multiple regression model is deciding which independent variables among a large number of independent variables of a data set should be included.

Model Building

Writing a model that will provide a good fit to a set of data and that will give good estimates of the mean value of y and good predictions of future values of y for given values of the independent variables. By entering larger number of independent variables into the model, we potentially have more terms. Hence, we need a larger smaple size to ensure for example the degrees of freedom for estimating σ^2 is not zero. Also interpreting the β parameters in this complex model will be difficult. To reduce a large list of potential predictors to a more manageable one, we consider two systematic methods known as **variable screening procedures** such as **stepwise regression** and **all-possible-regressions-selection**.

6.2 Stepwise Regression

Stepwise regression is one of the most widely used variable screening methods.

To run a stepwise regression:

1. Identify the dependent variable (response), y.
2. Identify set of potentially important independent variables, $x_1, x_2, ..., x_k$.

   Note: This set of variables could include both first-order and higher-order terms as well as interactions.
3. Enter the data into the computer software.

Computer Software Procedure

1. fitting all possible one-variable models of the form to the data, where x_i is i^{th} independent variable when $i = 1, 2, ..., k$.

   Then conducting individula t-test for each model:

$$H_0 : \beta_i = 0$$

$$H_a : \beta_i \neq 0$$

Finally, the independent variable (e.g. x_1) with the largest (absolute) t-value is selected the best one-variable predictor of y (or the one with the largest (absolute) Pearson product moment correlation, r with y).

2. Selecting the best two-variable model when one of them was selected in step 1. (e.g. x_1, x_2). The selection is based on conducting $k-1$ individual t-test and considering the independent variable (e.g. x_2) with the largest (absolute) t-value.

   Note: Before proceeding to Step 3, the stepwise routine will go back and check the t-value of $\hat{\beta}_1$ after $\hat{\beta}_2 x_2$ has been added to the model. If the t-value has become nonsignificant at some specified α level (say $\alpha = .05$), the variable x_1 is removed and a search is made for the independent variable with a α parameter that will yield the most significant t-value in the presence of $\hat{\beta}_2 x_2$.

3. Selecting a third independent variable to include in the model with x_1 and x_2 using the same criterion of selecting an independent variable with the largest (absolute) t-value. The program also recheck the t-values corresponding to the x_1 and x_2 coefficients, replacing the variables that yield nonsignificant t-values. This procedure is continued until no further independent variables can be found that yield significant t-values (at the specified α level) in the presence of the variables already in the model.

   Result of Stepwise Regression

   The model contains only those terms with t -values that are significant at the specified α level.

   Note: Do not conclude that the set of independent variables selected by stepwise procedure are important for predicting y or that the unimportant independent variables have been eliminated. Why?

Caution: Be cautious:

- when using the results of stepwise regression to make inferences about the relationship between $E(y)$ and the independent variables in the resulting first- order model.
- of having Type I or Type II errors.
- of entering only first-order and main effect terms.
- use stepwise regression when necessary (variable screening procedure).

Since RStudio uses other criteria to select variable(s), we explain the following common criteria:

- R^2-Adjusted

  Note: If the value of R^2-Adjusted increases very little by adding one predictor, generally the model with fewer number of independent variables is preferred.

- Akaike's Information Criterion (AIC) (based on balancing goodness of fit and a penalty for model complexity.)

$$AIC = nlog(\frac{SSE}{n}) + 2p$$

  When:

  p is the number of independent variables

  n is the number of the observartions

  Note: AIC is defined such that the smaller the value of AIC the better the model.

- Corrected AIC (AIC_C)

  Note: AIC_C can be used when the sample size is small, or when the number of parameters estimated is a moderate to large fraction of the sample size. Burnham and Anderson (2004) recommend that AIC_C be used instead of AIC unless $\frac{n}{K} > 40$. $K = p + 2$ (K is the number of estimated parameters (β and σ^2) in the fitted model.

$$AIC_c = AIC + \frac{2(p+2)(p+3)}{n-p-1}$$

- Bayesian Information Criterion (BIC) is defined such that the smaller the value of BIC the better the model.

  Note: BIC penalizes complex models more heavily than AIC, thus favoring simpler models than AIC. BIC is similar to AIC except that the factor 2 in the penalty term is replaced by $log(n)$.

- Mallows CP is defined such that the smaller the value of CP the better the model after taking into account the number of terms in the model.

$$CP = \frac{SSE_p}{MSE_k} + 2(p+1) - n$$

  when:

  - p is the number of independent variables selected from a set of k independent variables ($p < k$) (i.e. in the subset model).

- SSE_p is the sum of squares of residuals for the model with p independent variables.
- MSE_k is mean square error for the complete model with k variables.
- n is the sample size.

According to Hastie, Tibshirani and Freedman (2001, p. 208):

- Given a family of models, including the true model, the probability that BIC will select the correct model approaches one as the sample size $n \longrightarrow \infty$.
- AIC tends to choose models which are too complex as $n \longrightarrow \infty$.
- For finite samples, BIC often chooses models that are too simple, because of the heavy penalty on complexity.

A popular data analysis strategy which we shall adopt is to calculate R_a^2, AIC, AIC_C and BIC and compare the models which minimize AIC, AIC_C and BIC with the model that maximizes R_a^2.

Example 6.2.1 *Refer to Example 4.10 (p. 217 from the text book) and the multiple regression model for executive salary. Data set: EXECSAL2*

a) Use stepwise regression to decide which of the 10 variables should be included in the building of the final model for the natural log of executive salaries.

The dependent variable y is the natural logarithm of the executive salaries.

Consider the following independent variables in the executive salary example.

Note:

Table 6.1 Independent variables in the executive salary example

Independent Variable	Description
x_1	Experience (years)—quantitative
x_2	Education (years)—quantitative
x_3	Gender (1 if male, 0 if female)—qualitative
x_4	Number of employees supervised—quantitative
x_5	Corporate assets (millions of dollars)—quantitative
x_6	Board member (1 if yes, 0 if no)—qualitative
x_7	Age (years)—quantitative
x_8	Company profits (past 12 months, millions of dollars)—quantitative
x_9	Has international responsibility (1 if yes, 0 if no)—qualitative
x_{10}	Company's total sales (past 12 months, millions of dollars)—quantitative

The following partial output shows the final 5 independent variables, x_1, x_2, x_3, x_4, x_5, are selected.

The combination of backwrad and forward procedures using AIC criterion was used.

Other stepwise regression techniques are **forward selection** and **backward elimination**.

```
> step(lm(Y~.,data=EXECSAL2),direction="both")
Start:  AIC=-503.07
Y ~ id + X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10

       Df Sum of Sq     RSS     AIC
- X10   1   0.00061 0.51462 -504.95
- X7    1   0.00095 0.51496 -504.88
- id    1   0.00119 0.51520 -504.84
- X8    1   0.00158 0.51559 -504.76
- X6    1   0.00426 0.51827 -504.24
- X9    1   0.01001 0.52402 -503.14
<none>              0.51401 -503.07
- X5    1   0.08606 0.60007 -489.59
- X2    1   0.40323 0.91724 -447.16
- X4    1   0.63031 1.14432 -425.04
- X3    1   0.96695 1.48096 -399.25
- X1    1   1.43398 1.94799 -371.84


Step:  AIC=-511.93
Y ~ X1 + X2 + X3 + X4 + X5

       Df Sum of Sq    RSS     AIC
<none>              0.5304 -511.93
+ X9    1    0.0093 0.5211 -511.69
+ X6    1    0.0038 0.5267 -510.64
+ id    1    0.0014 0.5290 -510.20
+ X10   1    0.0004 0.5301 -509.99
+ X8    1    0.0002 0.5302 -509.97
+ X7    1    0.0000 0.5304 -509.93
- X5    1    0.0879 0.6183 -498.59
- X2    1    0.4289 0.9594 -454.67
- X4    1    0.6908 1.2212 -430.53
- X3    1    1.0656 1.5961 -403.76
- X1    1    3.9627 4.4932 -300.26

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5, data = EXECSAL2)

Coefficients:
(Intercept)           X1           X2           X3           X4           X5
  9.9619345    0.0272762    0.0290921    0.2246932    0.0005244    0.0019623
```

The forward selection is nearly identical to the stepwise procedure previously outlined. The only difference is that the forward selection technique provides no option for rechecking the t-values corresponding to the x's that have entered the model in an earlier step. Thus, stepwise regression is **preferred** to forward selection in practice.

The backward elimination fits a model containing terms for all potential independent variables. That is, for k independent variables. The variable with the smallest t (or F) statistic for testing $H_0 : \beta_i = 0$ is identified and dropped from the model if the t-value is less than some specified critical value. This process is repeated until no further nonsignificant independent variables can be found.

Advantage of Backward Elimination

Backward elimination method can be an advantage when at least one of the candidate independent variables is a qualitative variable at three or more levels (requiring at least two dummy variables), since the backward procedure tests the contribution of each dummy variable after the others have been entered into the model.

Disadvantage of Backward Elimination

The real disadvantage of using the backward elimination technique is that you need a sufficiently large number of data points to fit the initial model in Step 1.

The following partial RSudio outputs show the first and last step of the backward elimination technique for Example 6.2.1.

```
> step(lm(Y~.,data=EXECSAL2),direction="backward")
Start:  AIC=-503.07
Y ~ id + X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10

       Df Sum of Sq     RSS     AIC
- X10   1   0.00061 0.51462 -504.95
- X7    1   0.00095 0.51496 -504.88
- id    1   0.00119 0.51520 -504.84
- X8    1   0.00158 0.51559 -504.76
- X6    1   0.00426 0.51827 -504.24
- X9    1   0.01001 0.52402 -503.14
<none>              0.51401 -503.07
- X5    1   0.08606 0.60007 -489.59
- X2    1   0.40323 0.91724 -447.16
- X4    1   0.63031 1.14432 -425.04
- X3    1   0.96695 1.48096 -399.25
- X1    1   1.43398 1.94799 -371.84


Step:  AIC=-511.93
Y ~ X1 + X2 + X3 + X4 + X5

       Df Sum of Sq    RSS     AIC
<none>              0.5304 -511.93
- X5    1    0.0879 0.6183 -498.59
- X2    1    0.4289 0.9594 -454.67
- X4    1    0.6908 1.2212 -430.53
- X3    1    1.0656 1.5961 -403.76
- X1    1    3.9627 4.4932 -300.26

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5, data = EXECSAL2)

Coefficients:
(Intercept)           X1           X2           X3           X4           X5
  9.9619345    0.0272762    0.0290921    0.2246932    0.0005244    0.0019623
```

6.3 All-Possible-Regressions Selection Procedure

All-possible-regressions selection procedures consider all possible regression models given the set of potentially important predictors. These techniques use different criteria for selecting the "best" subset of variables such as:

R^2 Criterion

Considering the model that includes all k independent variables has the largest R^2, the objective of the R^2 criterion is to find a subset model (i.e., a model containing a subset of the k independent variables) so that adding more variables to the model will yield only small increases in R^2. Hence, the best model based byR^2 criterion does not necessarily have the largest R^2.

$$R^2 = 1 - \frac{SSE}{SST}$$

Adjusted R^2 or MSE Criterion

Recall from chapter 4, in multiple regression the adjusted multiple coefficient of determination, R^2, is often reported to adjust the sample size and the number of parameters in the model.

$$R_a^2 = 1 - (n-1)[\frac{SSE}{SST}]$$

Note: R_a^2 increases only if MSE decreases since SST is constant.

C_p Criterion

$$CP = \frac{SSE_p}{MSE_k} + 2(p+1) - n$$

Note: C_p criterion focuses on minimizing total mean square error and the regression bias by choosing a model with the smallest C_p value.

The C_p criterion selects as the best model the subset model with:

1. A small value of $C_p \equiv$ A small total mean square error.

2. A value of C_p near $p+1 \equiv$ Slight or no bias exists in the subset regression model.

Note: A model is said to be unbiased if $E(\hat{y}) = E(y) \longrightarrow E(C_p) \approx p+1$.

PRESS Criterion

$$PRESS = \sum_{1=1}^{n} [y_i - \hat{y}_{(i)}]^2$$

where: $\hat{y}_{(i)}$ is the predicted value for the i^{th} observation obtained when the regression model is fit with the data point for the i^{th} observation omitted (or deleted) from the sample.

Note: A model with a small $PRESS$ is dsired. Apply the all-possible- regressions selection procedure to find the most important independent variables.

Example 6.3.1 *Refer to Example 6.2.1. Apply the all-possible- regressions selection procedure to find the most important independent variables.*

a) How many possible subset first-order models can we have?

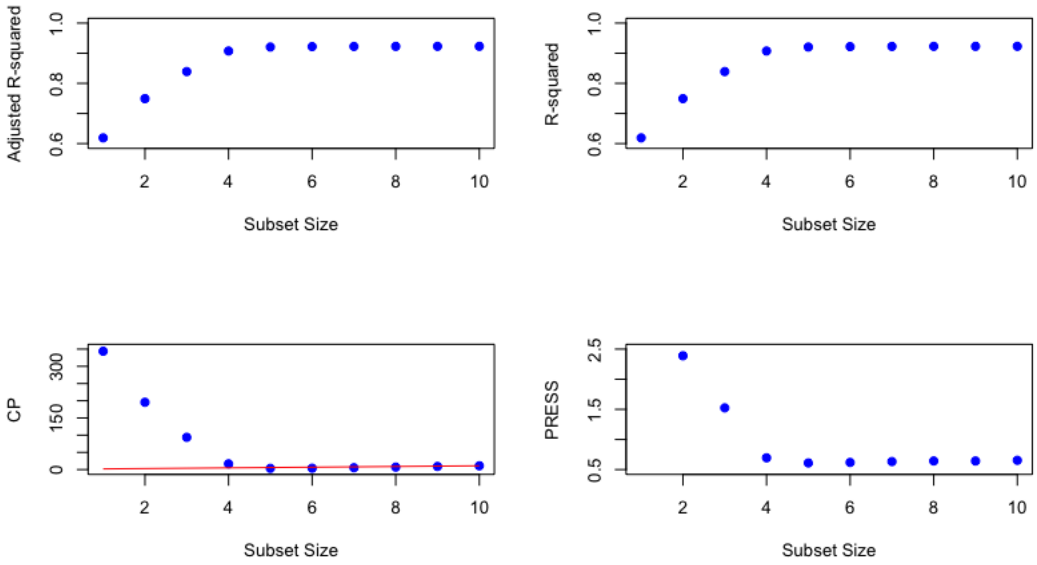
b) Show the possible models only for the "best" model for each value of p.

```
> Model=regsubsets(Y~X1+X2+X3+X4+X5+X6+X7+X8+X9+X10,data=EXECSAL2, nvmax = 10)
> summary(Model)
Subset selection object
Call: regsubsets.formula(Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 +
    X9 + X10, data = EXECSAL2, nvmax = 10)
10 Variables  (and intercept)
    Forced in Forced out
X1      FALSE      FALSE
X2      FALSE      FALSE
X3      FALSE      FALSE
X4      FALSE      FALSE
X5      FALSE      FALSE
X6      FALSE      FALSE
X7      FALSE      FALSE
X8      FALSE      FALSE
X9      FALSE      FALSE
X10     FALSE      FALSE
1 subsets of each size up to 10
Selection Algorithm: exhaustive
          X1  X2  X3  X4  X5  X6  X7  X8  X9  X10
1  ( 1 )  "*" " " " " " " " " " " " " " " " " " "
2  ( 1 )  "*" " " "*" " " " " " " " " " " " " " "
3  ( 1 )  "*" " " "*" "*" " " " " " " " " " " " "
4  ( 1 )  "*" "*" "*" "*" " " " " " " " " " " " "
5  ( 1 )  "*" "*" "*" "*" "*" " " " " " " " " " "
6  ( 1 )  "*" "*" "*" "*" "*" " " " " " " "*" " "
7  ( 1 )  "*" "*" "*" "*" "*" "*" " " " " "*" " "
8  ( 1 )  "*" "*" "*" "*" "*" "*" " " "*" "*" " "
9  ( 1 )  "*" "*" "*" "*" "*" "*" "*" "*" "*" " "
10  ( 1 ) "*" "*" "*" "*" "*" "*" "*" "*" "*" "*"
```

c) Summarize the "best subset" models along with their criteria.

```
> cbind(SUM$which,round(cbind(Rsq,AdRsq,CP,BIC,RSS,AIC, PRESS, MSE),4))
   (Intercept) X1 X2 X3 X4 X5 X6 X7 X8 X9 X10    Rsq  AdRsq       CP       BIC    RSS       AIC  PRESS    MSE
1            1  1  0  0  0  0  0  0  0  0   0 0.6190 0.6151 343.8566  -87.2799 2.5462  -92.4902 2.6639 0.0260
2            1  1  0  1  0  0  0  0  0  0   0 0.7492 0.7440 195.5192 -124.4974 1.6760 -132.3129 1.7880 0.0173
3            1  1  0  1  1  0  0  0  0  0   0 0.8391 0.8341  93.7538 -164.2722 1.0753 -174.6929 1.1712 0.0112
4            1  1  1  1  1  0  0  0  0  0   0 0.9075 0.9036  16.8128 -215.0014 0.6183 -228.0273 0.6956 0.0065
5            1  1  1  1  1  1  0  0  0  0   0 0.9206 0.9164   3.6279 -225.7305 0.5304 -241.3615 0.6102 0.0056
6            1  1  1  1  1  1  0  0  0  1   0 0.9220 0.9170   4.0235 -222.8918 0.5211 -241.1280 0.6104 0.0056
7            1  1  1  1  1  1  1  0  0  1   0 0.9225 0.9166   5.4499 -218.9258 0.5178 -239.7672 0.6198 0.0056
8            1  1  1  1  1  1  1  0  1  1   0 0.9227 0.9159   7.1956 -214.6054 0.5163 -238.0519 0.6287 0.0057
9            1  1  1  1  1  1  1  1  1  1   0 0.9228 0.9151   9.1091 -210.0972 0.5158 -236.1489 0.6427 0.0057
10           1  1  1  1  1  1  1  1  1  1   1 0.9229 0.9142  11.0000 -205.6146 0.5152 -234.2714 0.6542 0.0058
>
```

d) Plot the criteria against the number of variables, p.



6.4 Caveats

- Stepwise regression and the all-possible-regressions selection procedure are useful **variable screening methods** which might be different from **model-building methods**.
- Ba aware of the high probability of making at least one Type I error or at least one Type II error when applying these variable screening methods which means including at least one unimportant independent variable or leaving out at least one important independent variable in the final model.
- Typically higher-order terms or interactions in the list of potential predictors for stepwise regression are not included.
- Main terms might not be included by the stepwise and best subsets procedures when the interaction terms exist in the model.
- Using common sense or intuition when applying stepwise regression.

Acknowledgement

The core content of the slides are from the textbook of this course;

A Second Course in Statistics: Regression Analysis (7th Edition)

by

Mendenhall, William and Sincich, Terry; Pearson Education.

A Modern Approach to Regression with R

by

Simon J. Sheather