AMA3602
Applied Linear Models
Department of Applied Mathematics

02/2024

Chapter 3-

I - Variable Selection in Regression Equations

References:

Chapter 10: Variable Selection and Model Building
Montgomery, D.C., Peck, E.A., & Vining, G.G. (2012) *Introduction to Linear Regression Analysis*. 5th Ed. Wiley.

Outline

Introduction
- Goals of model selection
- Criteria to compare models
- Model-building problem
- Consequences of model misspecification
- Criteria for evaluating equations

Computational Techniques for Variables Selection
- All possible regression
- Stepwise Regression Methods

Strategy for variable selection and model building

Introduction
- Goals of model selection
- Criteria to compare models
- Model-building problem
- Consequences of model misspecification
- Criteria for evaluating equations

Computational Techniques for Variables Selection

Strategy for variable selection and model building

Model selection: goals

- When we have many predictors (with many possible interactions), it can be difficult to find a good model.
- Which main effects do we include?
- Which interactions do we include?
- Model selection tries to "simplify" this task.
- This is an "unsolved" problem in statistics: there are no magic procedures to get you the "best model".
- In some sense, model selection is "data mining".
- Data miners / machine learners often work with very many predictors.

Model selection: strategies

- To "implement" this, we need:
  - ⋆ a criterion or benchmark to compare two models.
  - ⋆ a search strategy.
- With a limited number of predictors, it is possible to search all possible models.

Possible criteria

- R^2: not a good criterion. Always increase with model size $\rightarrow$ "optimum" is to take the biggest model.
- Adjusted R^2: better. It "penalized" bigger models.
- Mallow's C_p.
- Akaike's Information Criterion(AIC), Schwarz's BIC.

Model-building problem

- Previously our concern: functional specification correct or not; underlying assumption about the error term valid or not.
- We have employed the classical approach to regression model selection, which assumes that we have a very good idea of the basic form of the model and that we know all (or nearly all) of the regressors that should be used.
- In practice, particularly retrospective studies, variable selection problem: find an appropriate subset of regressors for the model when
  1. there is no clear-cut theory to determine the variables;
  2. there is a rather large pool of candidate variables;
  3. only a few are likely to be important.
- Good variable selection methods are very important in the presence of multicollinearity
  - ⋆ It help to justify the presence of these highly related regressors in the final model;
  - ⋆ It does not guarantee elimination of multicollinearity.

- Building a regression model that includes only a subset of the available regressors involves two conflicting objectives.
  1. The model includes as many regressors as possible so that the information content in these factors can influence the predicted value of y;
  2. The model includes as few regressors as possible because the variance of the prediction $\hat{y}$ increases as the number of regressors increases. Also the more regressors, the greater the costs of data collection and model maintenance.

  The process of finding a model that is a compromise between these two objectives is called selecting the **"best" regression equation**.
- None of the variable selection procedures are guaranteed to produce the best regression equation for a given data set.

Consequences of model misspecification

Consequences of **incorrect model specification**

- Assume that there are K candidate regressors $x_1, \ldots, x_K$ and $n \geq K+1$ observations on these regressors and the response y.

$$y_i = \beta_0 + \sum_{j=1}^{K} \beta_j x_{ij} + \varepsilon_i, \ i = 1, \ldots, n \quad \text{or} \quad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

- Let r be the number of regressors that are deleted form (1). Then the number of variables tht are retained is $p = K + 1 - r$, i.e. the subset model contains $p - 1 = K - r$ of the original regressors

$$\boldsymbol{y} = \boldsymbol{X}_p \boldsymbol{\beta}_p + \boldsymbol{X}_r \boldsymbol{\beta}_r + \boldsymbol{\varepsilon} \tag{2}$$

  - ⋆ The least-squares of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}^* = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$
  - ⋆ The estimate of the residual variance σ^2 is $\hat{\sigma}^{2*} = \frac{\boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}^{*\prime}\boldsymbol{X}'\boldsymbol{y}}{n-K-1} = \frac{\boldsymbol{y}'[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']\boldsymbol{y}}{n-K-1}$

- For the subset model

$$\boldsymbol{y} = \boldsymbol{X}_p \boldsymbol{\beta}_p + \boldsymbol{\varepsilon} \tag{3}$$

  - ⋆ The least-squares of $\boldsymbol{\beta}_p$ is $\hat{\boldsymbol{\beta}}_p = (\boldsymbol{X}_p'\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p'\boldsymbol{y}$
  - ⋆ The estimate of the residual variance σ^2 is $\hat{\sigma}^2 = \frac{\boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}_p'\boldsymbol{X}_p'\boldsymbol{y}}{n-p} = \frac{\boldsymbol{y}'[\boldsymbol{I} - \boldsymbol{X}_p(\boldsymbol{X}_p'\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p']\boldsymbol{y}}{n-p}$

- The properties of the estimates $\hat{\boldsymbol{\beta}}_p$ and $\hat{\sigma}^2$
  - $E(\hat{\boldsymbol{\beta}}_p) = \boldsymbol{\beta}_p + (\boldsymbol{X}_p'\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p'\boldsymbol{X}_r\boldsymbol{\beta}_r = \boldsymbol{\beta}_p + \boldsymbol{A}\boldsymbol{\beta}_r$ where $\boldsymbol{A} = (\boldsymbol{X}_p'\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p'\boldsymbol{X}_r$;
  - $Var(\hat{\boldsymbol{\beta}}_p) = \sigma^2(\boldsymbol{X}_p'\boldsymbol{X}_p)^{-1}$ and $Var(\hat{\boldsymbol{\beta}}^*) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Also $Var(\hat{\boldsymbol{\beta}}^*) - Var(\hat{\boldsymbol{\beta}}_p)$ is positive semidefinite;
  - Since $\hat{\boldsymbol{\beta}}_p$ is a biased estimate of $\boldsymbol{\beta}_p$ and $\hat{\boldsymbol{\beta}}^*$ is not, it is more reasonable to compare the precision of the parameter estimates from the full and subset models in terms of means square error;
  - The parameter $\hat{\sigma}^{2*}$ from the full model is an unbiased estimate of σ^2. However, for the subset model $E(\hat{\sigma}^2) = \sigma^2 + \frac{\boldsymbol{\beta}_r'\boldsymbol{X}_r'[\boldsymbol{I} - \boldsymbol{X}_p(\boldsymbol{X}_p'\boldsymbol{X}_p)^{-1}\boldsymbol{X}_p']\boldsymbol{X}_r\boldsymbol{\beta}_r}{n-p}$. That is $\hat{\sigma}^2$ is generally biased upward as an estimate of σ^2;
  - Suppose we wish to predict the response at the point $\boldsymbol{x}' = [\boldsymbol{x}_p', \boldsymbol{x}_r']$. If we use the full model, the predicted value is $\hat{\boldsymbol{y}}^* = \boldsymbol{x}'\hat{\boldsymbol{\beta}}^*$, with mean $\boldsymbol{x}'\boldsymbol{\beta}$ and prediction variance $Var(\hat{\boldsymbol{y}}^*) = \sigma^2[1 + \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}]$

Motivation for variable selection

- Improve the precision of the parameter estimates of the retained variables by deleting variables from the model, even though some of the deleted variables are not negligible.
  - ⋆ This is also true for the variance of a predicted response;
  - ⋆ Deleting variables potentially introduces bias into the estimates of the coefficients of retained variables and the response.
  - ⋆ However, if the deleted variables have small effects, the MSE of the biased estimates will be less than the variance of the unbiased estimates.
  - ⋆ There is danger in retaining negligible variables, that is, variables with zero coefficients or coefficients less than their corresponding standard errors from the full model. This danger is that the variances of the estimates of the parameters and the predicted response are increased.

Criteria for evaluating subset regression models

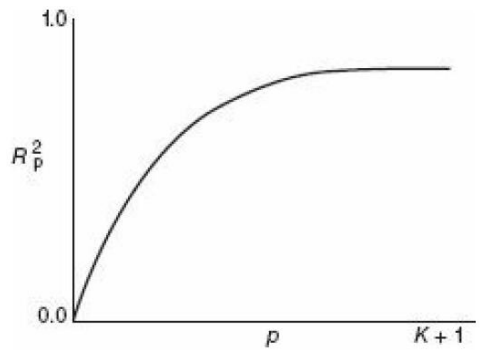
- Two key aspects of the variable selection:
  - ⋆ Generating the subset models;
  - ⋆ Deciding if one subset is better than another.

Computational methods for variable selection

- (**Coefficient of Multiple Determination** R^2) A measure of the adequacy of a regression model. Let R_p^2 denote the coefficient of multiple determination for a subset regression model with p terms, that is, $p-1$ regressors and an intercept term β_0

$$R_p^2 = \frac{SS_R(p)}{SS_T} = 1 - \frac{SS_{Res}(p)}{SS_T} \tag{4}$$

  - ⋆ There are $\binom{K}{p-1}$ values of R_p^2 for each value of p, and R_p^2 increases as p increases and is a maximum when $p = K+1$;
  - ⋆ The analyst uses this criterion by adding regressors to the model up to the point where an additional variable is not useful in that it provides only a small increase in R_p^2

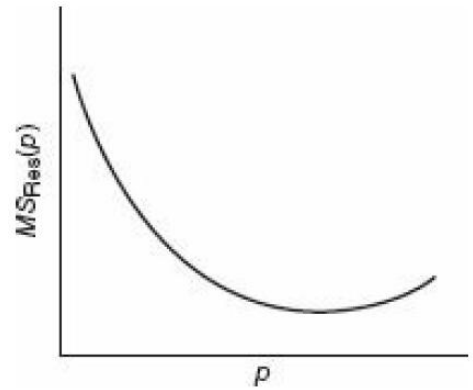


- Since we cannot find an "optimum" value of R^2 for a subset regression model, we must look for a "satisfactory" value.
  - ⋆
    $$R_0^2 = 1 - (1 - R_{K+1}^2)(1 + d_{\alpha,n,K}), \tag{5}$$
    where $d_{\alpha,n,K} = \frac{KF_{\alpha,K,n-K-1}}{n-K-1}$ and R_{K+1}^2 is the value of R^2 for the full model.
  - ⋆ Any subset of regressor variables producing an R^2 greater than R_0^2 is called an R^2-**adequate** (α) **subset**.
- (**Adjusted** R^2) The adjusted R^2 statistic, defined for a p-term equation as
$$R_{\text{Adj},p}^2 = 1 - \left(\frac{n-1}{n-p}\right)(1 - R_p^2) \tag{6}$$
  - ⋆ $R_{\text{Adj},p}^2$ statistic does not necessarily increase as additional regressors are introduced into the model.
  - ⋆ In fact, if s regressors are added to the model, $R_{\text{Adj},p+s}^2$ will exceed $R_{\text{Adj},p}^2$ if and only if the partial F statistic for testing the significance of the s additional regressors exceeds 1;
  - ⋆ Consequently, one criterion for selection of an optimum subset model is to choose the model that has a maximum $R_{\text{Adj},p}^2$.

- (**Residual mean square**) The residual mean square for a subset regression model

$$MS_{Res}(p) = \frac{SS_{Res}(p)}{n-p} \tag{7}$$

  - ⋆ The general behavior of $MS_{Res}(p)$ as p increases is illustrated in the following figure. $MS_{Res}(p)$ initially decreases, then stabilizes, and eventually may increase.



(Remark: the eventual increase in $MS_{Res}(p)$ occurs when the reduction in $SS_{Res}(p)$ from adding a regressor to the model is not sufficient to compensate for the loss of one degress of freedom in the denominator of (7).)

  - ⋆ The subset regression model that minimizes $MS_{Res}(p)$ will also maximize $R^2_{\text{Adj},p}$.

$$R^2_{\text{Adj},p} = 1 - \frac{n-1}{n-p}(1-R^2_p) = 1 - \frac{n-1}{n-p}\frac{SS_{Res}(p)}{SS_T} = 1 - \frac{MS_{Res}(p)}{SS_T/(n-1)}$$

Thus, the criteria minimum $MS_{Res}(p)$ and maximum adjusted R^2 are **equivalent**.

- (**Mallow's C_p Statistic**)

$$C_p(\mathcal{M}) = \frac{SS_{Res}(\mathcal{M})}{\widehat{\sigma}^2} - n + 2 \cdot p(\mathcal{M}) \tag{8}$$

  - ⋆ $\widehat{\sigma}^2 = SS_{Res}(F)/df_F$ is the "best" estimate of σ^2, we have (use the fullest model).
  - ⋆ $SS_{Res}(\mathcal{M}) = \|Y - \widehat{Y}_{\mathcal{M}}\|^2$ is the SS_{Res} of the model $\mathcal{M}$.
  - ⋆ $p(\mathcal{M})$ is the number of predictors in $\mathcal{M}$, or the degrees of freedom used up by the model.
  - ⋆ Based on an estimate of

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbb{E}((Y_i - \mathbb{E}(Y_i))^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathbb{E}((Y_i - \widehat{Y}_i)^2) + \mathrm{Var}(\widehat{Y}_i).$$

- (**AIC & BIC**)
  - ⋆ Mallow's C_p is (almost) a special case of Akaike Information Criterion(AIC)

$$AIC(\mathcal{M}) = -2\log L(\mathcal{M}) + 2 \cdot p(\mathcal{M}).$$

  - ⋆ $L(\mathcal{M})$ is the likelihood function of the parameters in model $\mathcal{M}$ evaluated at the MLE (Maximum Likelihood Estimators).
  - ⋆ Schwarz's Bayesian Information Criterion (BIC)

$$BIC(\mathcal{M}) = -2\log L(\mathcal{M}) + p(\mathcal{M}) \cdot \log n.$$

Search strategies

- "Best subset": search all possible models and take the one with highest R_a^2 or lowest C_p.
- Stepwise (forward, backward or both): useful when the number of predictors is large. Choose an initial model and be "greedy".
- "Greedy" means always take the biggest jump (up or down) in your selected criterion.

Implementations in *R*

- "Best subset": use the function *leaps*. Works only for multiple linear regression models.
- Stepwise: use the function *step*. Works for any model with Akaike Information Criterion (AIC). In multiple linear regression, AIC is (almost) a linear function of C_p .

Introduction

Computational Techniques for Variables Selection
- All possible regression
- Stepwise Regression Methods

Strategy for variable selection and model building

Selection for variables

- To find the subset of variables to use in the final equation, it is natural to consider fitting models with various combinations of the candidate regressors.
- Computational techniques for generating subset regression models
  - ⋆ All possible regression
  - ⋆ Stepwise regression methods
    - Forward selection;
    - Backward elimination;
    - Stepwise regression.

All possible regression

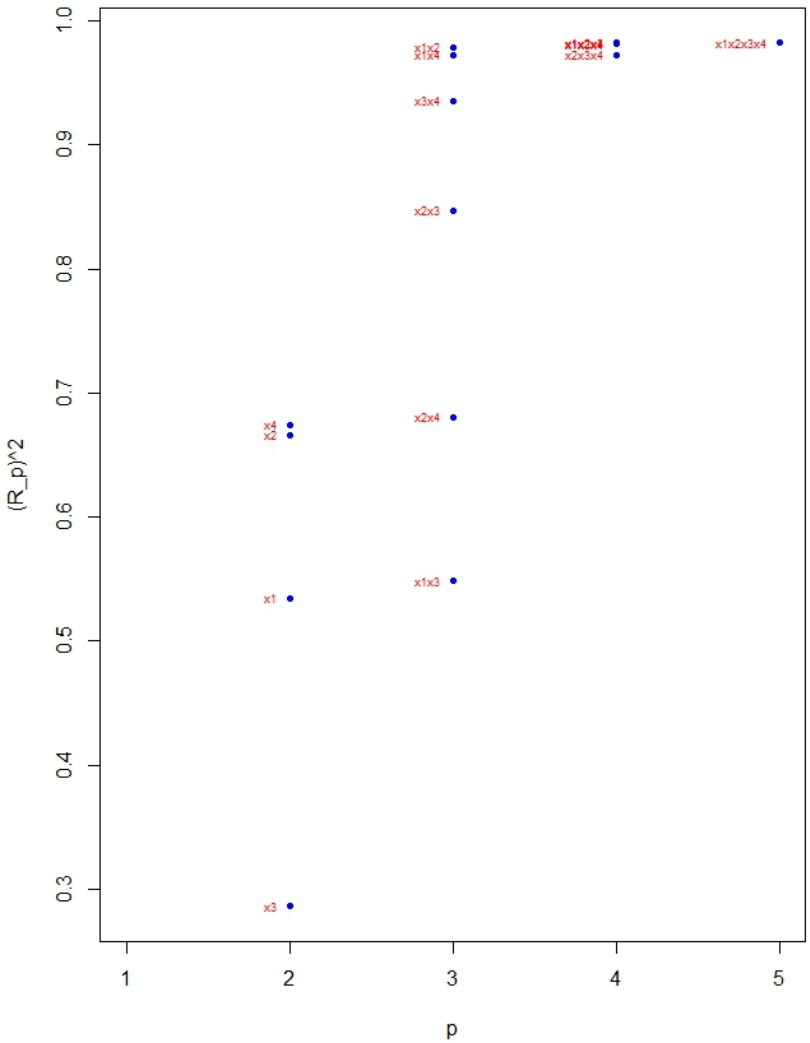
- This procedure requires that the analyst fit all the regression equations involving one candidate regressor, two candidate regressors, and so on.
- If there are K candidate regressors, there are 2^K total equations to be estimated and examined. Clearly the number of equations to be examined increases rapidly as the number of candidate regressors increases.
- **Example 1**: The Hald Cement Data

Number of Regressors in Model	p	Regressors in Model	$SS_{Res}(p)$	R_p^2	$R_{Adj,p}^2$	$MS_{Res}(p)$	C_p
None	1	None	2715.7635	0	0	226.3136	442.92
1	2	x_1	1265.6867	0.53395	0.49158	115.0624	202.55
1	2	x_2	906.3363	0.66627	0.63593	82.3942	142.49
1	2	x_3	1939.4005	0.28587	0.22095	176.3092	315.16
1	2	x_4	883.8669	0.67459	0.64495	80.3515	138.73
2	3	x_1x_2	57.9045	0.97868	0.97441	5.7904	2.68
2	3	x_1x_3	1227.0721	0.54817	0.45780	122.7073	198.10
2	3	x_1x_4	74.7621	0.97247	0.96697	7.4762	5.50
2	3	x_2x_3	415.4427	0.84703	0.81644	41.5443	62.44
2	3	x_2x_4	868.8801	0.68006	0.61607	86.8880	138.23
2	3	x_3x_4	175.7380	0.93529	0.92235	17.5738	22.37
3	4	$x_1x_2x_3$	48.1106	0.98228	0.97638	5.3456	3.04
3	4	$x_1x_2x_4$	47.9727	0.98234	0.97645	5.3303	3.02
3	4	$x_1x_3x_4$	50.8361	0.98128	0.97504	5.6485	3.50
3	4	$x_2x_3x_4$	73.8145	0.97282	0.96376	8.2017	7.34
4	5	$x_1x_2x_3x_4$	47.8636	0.98238	0.97356	5.9829	5.00

Variables in Model	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
x_1	81.479	1.869			
x_2	57.424		0.789		
x_3	110.203			−1.256	
x_4	117.568				−0.738
x_1x_2	52.577	1.468	0.662		
x_1x_3	72.349	2.312		0.494	
x_1x_4	103.097	1.440			−0.614
x_2x_3	72.075		0.731	−1.008	
x_2x_4	94.160		0.311		−0.457
x_3x_4	131.282			−1.200	−0.724
$x_1x_2x_3$	48.194	1.696	0.657	0.250	
$x_1x_2x_4$	71.648	1.452	0.416		−0.237
$x_2x_3x_4$	203.642		−0.923	−1.448	−1.557
$x_1x_3x_4$	111.684	1.052		−0.410	−0.643
$x_1x_2x_3x_4$	62.405	1.551	0.510	0.102	−0.144

- Clearly the least squares in estimate of an individual regression coefficient depends heavily on the other regressors in the model.
- The large changes in the regression coefficients observed in the Hald cement data are consistent with a serous problem with multicollinearity.

R_p^2 criterion



- From examining this display it is clear that after two regressors are in the model, there is little to be gained in terms of R^2 by introducing additional variables;
- Both of the two regressor models (x_1, x_2) and (x_1, x_4) have essentially the same R^2 values, and in terms of this criterion, it would make little difference which model is selected as the final regression equation;
- If we take $\alpha = 0.05$,

$$R_0^2 = 1 - (1 - R_5^2)\left(1 + \frac{4F_{0.04,4,8}}{8}\right)$$

$$= 1 - (1 - 0.98238)\left[1 + \frac{4(3.84)}{8}\right]$$

$$= 0.94855$$

Therefore, any subset regression model for which $R_p^2 > R_0^2 = 0.94855$ is adequate (0.05), that is, its R^2 is not significantly different from R_{K+1}^2.

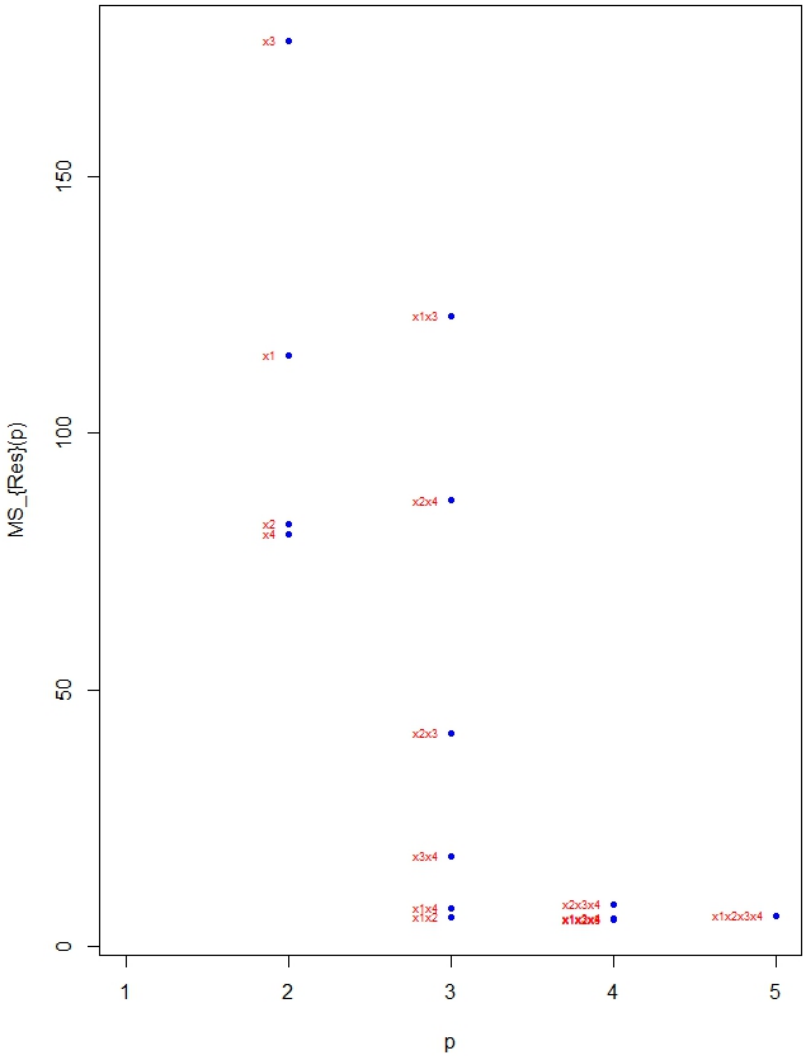
Simple correlations

```
> X<-scale(cbind(Ex1$x1, Ex1$x2, Ex1$x3, Ex1$x4, Ex1$y))
> cor(X)
            [,1]       [,2]       [,3]       [,4]       [,5]
[1,]  1.0000000  0.2285795 -0.8241338 -0.2454451  0.7307175
[2,]  0.2285795  1.0000000 -0.1392424 -0.9729550  0.8162526
[3,] -0.8241338 -0.1392424  1.0000000  0.0295370 -0.5346707
[4,] -0.2454451 -0.9729550  0.0295370  1.0000000 -0.8213050
[5,]  0.7307175  0.8162526 -0.5346707 -0.8213050  1.0000000
```

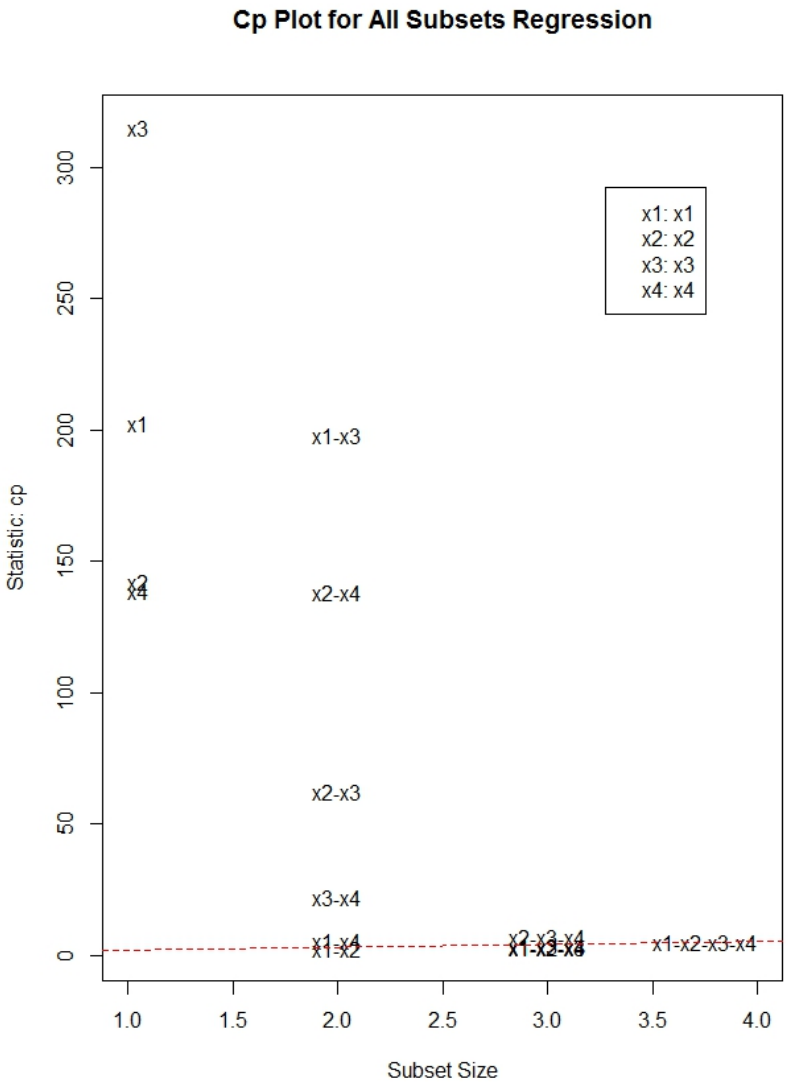
It is instructive to examine the pairwise correlation between x_i and x_j and between x_i and y.

- The pairs of regressor (x_1, x_3) and (x_2, x_4) are highly correlated;
- Consequently, adding further regressors when x_1 and x_2 or when x_1 and x_4 are already in the model will be of little use since the information content in the excluded regressors is essentially present in the regressors that are in the model.
- This correlative structure is partially responsible for the large changes in the regression coefficients.

$MS_{Res}(p)$ vs. p



- The minimum residual mean square model is (x_1, x_2, x_4), with $MS_{Res}(4) = 5.3303$ ($R^2_{\text{Adj}} = 0.97645$);
- As expected, the model that minimizes $MS_{Res}(p)$ also maximizes the adjusted R^2;
- However, two of the other three-regressor models [(x_1, x_2, x_3) and (x_1, x_3, x_4)] and the two-regressor models [(x_1, x_2) and (x_1, x_4)] have comparable values of the residual mean square.
  1. If either (x_1, x_2) or (x_1, x_4) is in the model, there is little reduction in residual mean square by adding further regressors.
  2. The subset model (x_1, x_2) may be more appropriate than (x_1, x_4) because it has a smaller value of the residual mean square.



$C_p(p)$ vs. p

- Suppose we take $\hat{\sigma}^2 = 5.9829$ (MS_{Res} from the full model), then

$$C_3 = \frac{SS_{Res}(3)}{\hat{\sigma}^2} - n + 2p$$
$$= \frac{74.7621}{5.9829} - 13 + 2(3) = 5.50$$

- From examination of this plot we find that there are four models that could be acceptable: (x_1, x_2), (x_1, x_2, x_3), (x_1, x_2, x_4), and (x_1, x_3, x_4);
- Without considering additional factors such as technical information about the regressors or the costs of data collection, it may be appropriate to choose the simplest model (x_1, x_2) as the final model because it has the smallest C_p.

- This example has illustrated the computational procedure associated with model building with all possible regressions.
- Note that there is no clear-cut choice of the best regression equation.
- Very often we find that different criteria suggest different equations. For example, the minimum C_p equation is (x_1, x_2) and the minimum MS_{Res} equation is (x_1, x_2, x_4).
- All "final" candidate models should be subjected to the usual tests for adequacy, including investigation of leverage points, influence, and multicollinearity.

Observation i	$\hat{y} = 52.58 + 1.468x_1 + 0.662x_2$[a]			$\hat{y} = 71.65 + 1.452x_1 + 0.416x_2 - 0.237x_4$[b]		
	e_i	h_{ii}	$[e_i/(1-h_{ii})]^2$	e_i	h_{ii}	$[e_i/(1-h_{ii})]^2$
1	−1.5740	0.25119	4.4184	0.0617	0.52058	0.0166
2	−1.0491	0.26189	2.0202	1.4327	0.27670	3.9235
3	−1.5147	0.11890	2.9553	−1.8910	0.13315	4.7588
4	−1.6585	0.24225	4.7905	−1.8016	0.24431	5.6837
5	−1.3925	0.08362	2.3091	0.2562	0.35733	0.1589
6	4.0475	0.11512	20.9221	3.8982	0.11737	19.5061
7	−1.3031	0.36180	4.1627	−1.4287	0.36341	5.0369
8	−2.0754	0.24119	7.4806	−3.0919	0.34522	22.2977
9	1.8245	0.17195	4.9404	1.2818	0.20881	2.6247
10	1.3625	0.55002	9.1683	0.3539	0.65244	1.0368
11	3.2643	0.18402	16.0037	2.0977	0.32105	9.5458
12	0.8628	0.19666	1.1535	1.0556	0.20040	1.7428
13	−2.8934	0.21420	13.5579	−2.2247	0.25923	9.0194
		PRESS $x_1, x_2 = 93.8827$			PRESS $x_1, x_2, x_4 = 85.3516$	

[a] $R^2_{\text{Prediction}} = 0.9654$, $\text{VIF}_1 = 1.05$, $\text{VIF}_2 = 1.06$.

[b] $R^2_{\text{Prediction}} = 0.9684$, $\text{VIF}_1 = 1.07$, $\text{VIF}_2 = 18.78$, $\text{VIF}_4 = 18.94$.

- ⋆ This table examines the two models (x_1, x_2) and (x_1, x_2, x_4) with respect to PRESS and their variance inflation factors (VIFs).
- ⋆ Both models have very similar values of PRESS (roughly twice the residual sum of squares for the minimum MS_{Res} equation), and the R^2 for prediction computed from PRESS is similar for both models.
- ⋆ However, x_2 and x_4 are highly multicollinear, as evidenced by the larger variance inflation factors in (x_1, x_2, x_4).
- ⋆ Since both models have equivalent PRESS statistics, we would recommend the model with (x_1, x_2) based on the lack of multicollinearity in this model.

Selection of variables: stepwise-type procedures

- Cases: when there are a large number of potential explanatory variables ($\#=q$); not involve computing of all possible equations (2^q).
- Common feature: the variables are introduced or deleted from the equation one at a time; exam only a subset of all possible equations (evaluate at most $q+1$ equations)
- Procedure Categories:
  - ⋆ **Forward selection** procedure
    - goes through the full set of variables and provides with q possible equations.
  - ⋆ **Backward elimination** procedure
    - involves fitting at most q regression equations
  - ⋆ **Stepwise regression** (a modification of the FS procedure)
    - a number of possible combinations of the afore two procedures.

Forward selection procedure

- Starts with an equation containing only a constant term but no regressors.
- Step 1: The first variable x_1 included in the equation: the one which has the highest simple correlation (with y).
  - ⋆ Retain it when β_1 is significantly different from zero;
  - ⋆ Search for a second variable.
- Step 2: The second variable is the one which has the highest correlation with y, after y has been adjusted for the effect of the first variable.
  - ⋆ i.e. the variable has the highest simple correlation coefficient with the residuals from step 1;
  - ⋆ Retain x_2 when β_2 is significantly different from zero;
  - ⋆ Search for a second variable.
- ...
- Terminate the procedure: insignificant β_q.
  - ⋆ Judged by the standard t-statistic computed from the latest equation;
  - ⋆ Mostly by a low t cutoff value for testing the coefficient of the newly entered variable.

Backward elimination procedure

- Starts with the full equation and successively drops one variable at a time; The variables are dropped on the basis of their contribution to the reduction of SS_{res}.
- Step 1: The first variable x_1 deleted from the equation: the one which contributes the smallest to the reduction of SS_{res}.
  - ⋆ i.e. the variable has the smallest t ratio (the ratio of the regression coefficient to the standard error of the coefficient).
  - ⋆ If all the t ratios are significant, all q regressors will be retained in the equation.
  - ⋆ If there are one or more variables with insignificant t ratios, drop the variables with the smallest insignificant t ratios
- Step 2: The equation with the remaining $q-1$ variables is fitted and the t ratios for the new regression coefficients are examined.
- ...
- Terminate the procedure: when all the t ratios are significant or all but one variable has been deleted.
  - ⋆ Judged by the standard t-statistic computed from the latest equation.
  - ⋆ Mostly by a high t cutoff value so that the procedure runs through the whole set of variables.

Stepwise method

- Essentially a forward selection procedure + proviso that at each stage the possibility of deleting a variable is considered.
  - ⋆ A variable that entered in the earlier stages of selection may be eliminated at later stages.
  - ⋆ The calculation made for inclusion and deletion of variables are the same as the afore two methods.
  - ⋆ Requires two cutoff values, one for entering and one for removing variables.
    - ▸ Frequently we choose $t_{IN} > t_{OUT}$, making it relatively more difficult to add a regressor than to delete one.
  - ⋆ Often different levels of significance are assumed for inclusion and exclusion of variables from the equation.
  - ⋆ Caution: the order in which the variables enter or leave the equation should not be interpreted as reflecting the relative importance of the variables. (intercorrelation affects!)

Comparison and comments

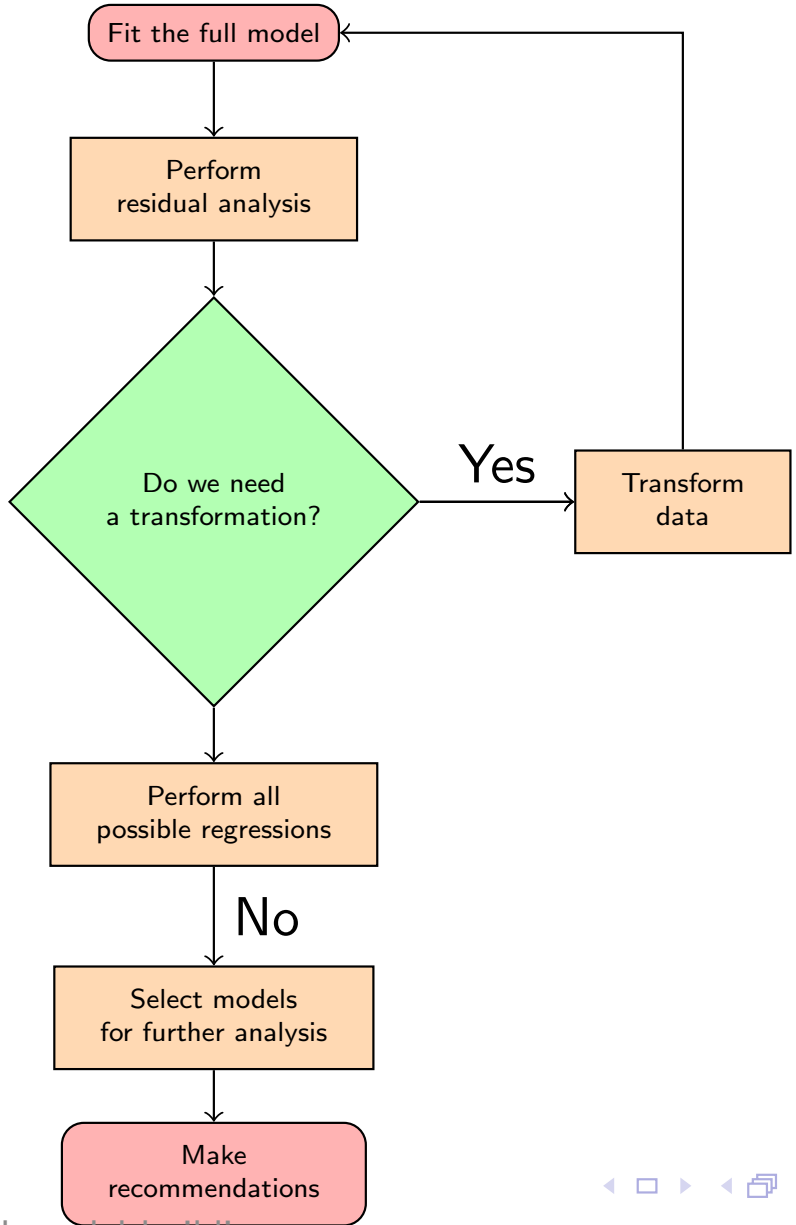
- FS and BE work in the opposite direction.
- A stopping rule: In FS, stop if minimum t ratio is less than 1; In BE, stop if minimum t ratio is greater than 1.
- Partial F-statistics is alternative of t-statistics since $t^2_{\alpha/2,\nu} = F_{\alpha,1,\nu}$.
- BE is particularly favored by analysts who like to see the effect of including all the candidate regressors, so that nothing obvious will be missed.
- All three work out nearly the same selection of variables with noncollinear data. But they do not necessarily lead to the same choice of the final model.
- FE: once a regressor has been added, it can not be removed at a later step.
- BE is better able to handle multicollinearity than FS because it is often less adversely affected by the correlative structure of the regressors than is FS.
- NONE generally guarantees that the best subset regression model of any size will be identified.

Introduction

Computational Techniques for Variables Selection

Strategy for variable selection and model building

Strategy for variable selection and model building



Variable selection for high-dim data

When high-dimensional or ultrahigh dimensional data is appeared in regression, i.e. $p >> n$, ??

- LASSO
- SCAD
- Elastic-net
- MCP
- ...
- SIS