CHAPTER 10

Variable Selection and Model Building

Introduction

- **Model-Building Problem**
  - "Conflicting" goals in regression model building:
    - Want as many regressors as possible so that the "information content" in the variables will influence $\hat{y}$
    - Want as few regressors as necessary because the variance of $\hat{y}$ will increase as the number of regressors increases and more regressors can cost more money in data collection/model maintenance
  - Need to find a compromise that leads to the *best* regression equation

Introduction

- Notes on Variable selection techniques
  - None of the variable selection techniques can guarantee the best regression equation for the dataset of interest
  - The techniques may very well give different results
  - Complete reliance on the algorithm for results is to be avoided
  - Other valuable information such as experience with and knowledge of the data and problem should be utilized

Introduction

- **Consequences of Deleting Variables**
  - Improves the precision of the parameter estimates of retained variables
  - Improves the precision of the variance of the predicted response
  - Can induce bias into the estimates of coefficients and variance of predicted response

Criteria for Evaluating Subset Regression Models

- Coefficient of Multiple Determination
  - For a model with *p* terms:

$$R_p^2 = \frac{SS_{\mathrm{R}}(p)}{SS_{\mathrm{T}}} = 1 - \frac{SS_{\mathrm{Res}}(p)}{SS_{\mathrm{T}}}$$

  - Large values of R_p^2 are preferred, and adding terms will increase this value

R code

```
n=30
x=rnorm(n)
y=1+2*x+rnorm(n)
summary(lm(y~x))
x2=rnorm(n)
summary(lm(y~x+x2))
x3=rnorm(n)
summary(lm(y~x+x2+x3))
x4=rnorm(n)
summary(lm(y~x+x2+x3+x4))
```

Coefficient of Multiple Determination

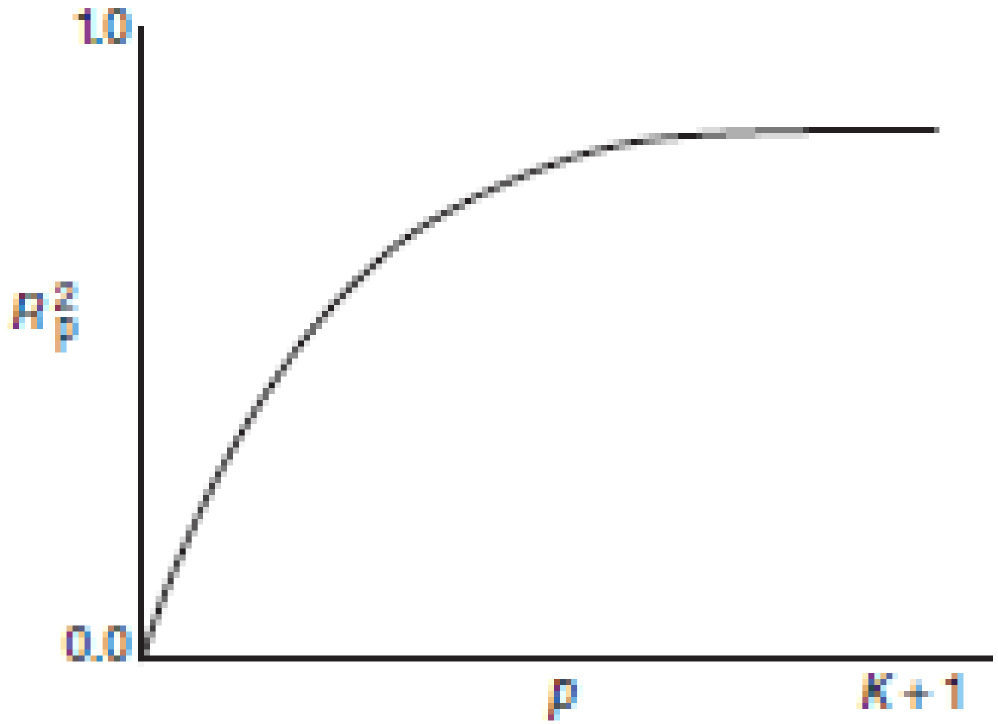

Figure 10.1 Plot of R_p^2 versus p.

Criteria for Evaluating Subset Regression Models

- Adjusted R^2
  - For a model with *p* terms:

$$R^2_{\text{Adj},p} = 1 - \left(\frac{n-1}{n-p}\right)(1 - R_p^2)$$

  - This value will not necessarily increase as additional terms are introduced into the model
  - Large values of adjusted R^2 are preferred

R code

```
n=30
x=rnorm(n)
y=1+2*x+rnorm(n)
summary(lm(y~x))
x2=rnorm(n)
summary(lm(y~x+x2))
x3=rnorm(n)
summary(lm(y~x+x2+x3))
x4=rnorm(n)
summary(lm(y~x+x2+x3+x4))
```

Criteria for Evaluating Subset Regression Models

- Residual Mean Square
  - The MS_{Res} for a subset regression model is

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

  - $MS_{Res}(p)$ increases as p increases, in general
  - The 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 degree of freedom.
  - Small values of $MS_{Res}(p)$ are preferred

R code

```
n=30
x=rnorm(n)
y=1+2*x+rnorm(n)
m1=lm(y~x)
MSRes1=sum(m1$residuals^2)/(n-2)
x2=rnorm(n)
m2=lm(y~x+x2)
MSRes2=sum(m2$residuals^2)/(n-3)
x3=rnorm(n)
m3=lm(y~x+x2+x3)
MSRes3=sum(m3$residuals^2)/(n-4)
x4=rnorm(n)
m4=lm(y~x+x2+x3+x4)
MSRes4=sum(m3$residuals^2)/(n-5)
MSRes1
MSRes2
MSRes3
MSRes4
summary(m1)
summary(m2)
summary(m3)
summary(m4)
```

Residual Mean Square

Same value of p that min $MS_{Res}(p)$ also max $R^2_{Adj,p}$

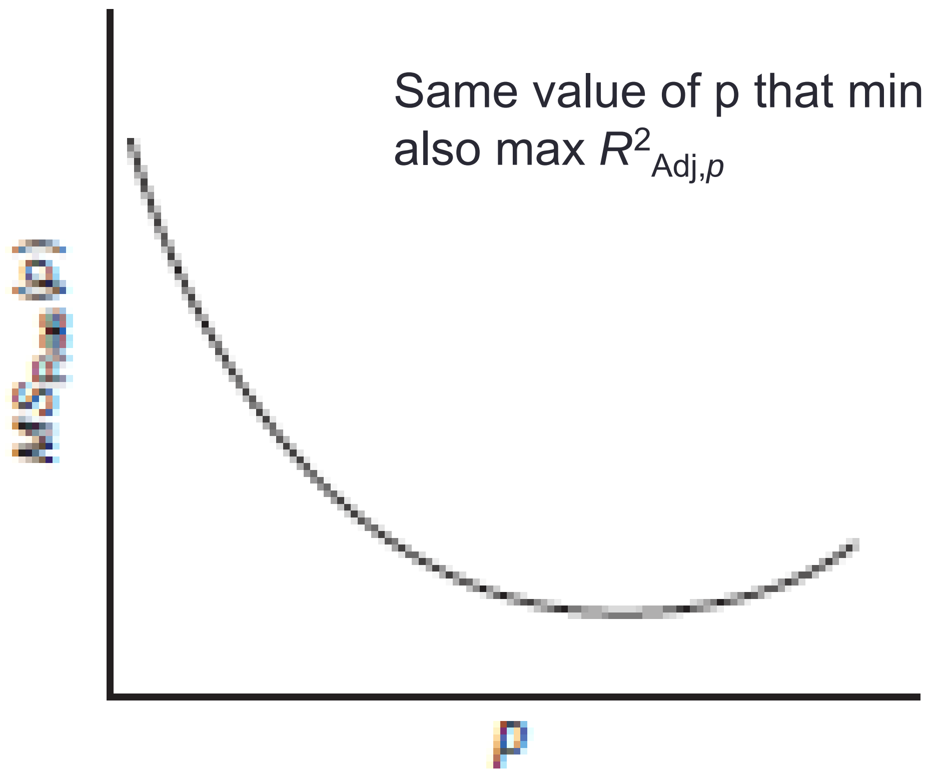

Figure 10.2 Plot of $MS_{Res}(p)$ versus p.

Criteria for Evaluating Subset Regression Models

- Mallow's C_p Statistic

$$C_p = \frac{SS_{\text{Res}}(p)}{\hat{\sigma}^2} - n + 2p$$

  - C_p is a measure of variance in the fitted values and $(\text{bias})^2$
  - $C_p \gg p$, then significant bias
  - Small C_p values near p are desirable
  - Negative values of C_p could result because the MSE for the full model overestimates the true σ^2

R code

- library(leaps)
- n=100000
- x=rnorm(n)
- y=1+2*x+rnorm(n)
- m1=lm(y~x)
- x2=rnorm(n)
- m2=lm(y~x+x2)
- x3=rnorm(n)
- m3=lm(y~x+x2+x3)
- x4=rnorm(n)
- m4=lm(y~x+x2+x3+x4)
- leaps(x=cbind(x,x2,x3,x4), y=y, method="Cp")

Mallow's C_p Statistic

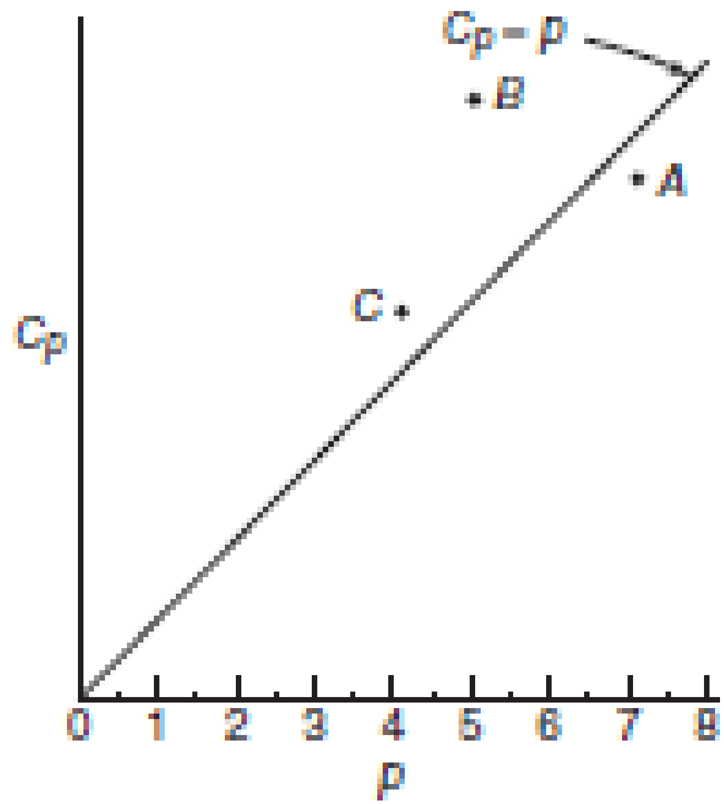

Figure 10.3 A C_p plot.

Criteria for Evaluating Subset Regression Models

- AIC and BIC Criteria
  - AIC: Akaike's Information Criterion
  - BIC: Schwarz' Bayesian Criterion
  - $\mathrm{AIC} = n \ln\left(SS_{\mathrm{Res}}/n\right) + 2p$
  - $\mathrm{BIC}_{\mathrm{Sch}} = n \ln\left(SS_{\mathrm{Res}}/n\right) + p\ln(n)$
  - Search for models with small values
  - Commonly used for more complicated modeling situations

R code

- library(leaps)
- n=1000000
- x=rnorm(n)
- y=1+2*x+rnorm(n)
- m1=lm(y~x)
- x2=rnorm(n)
- m2=lm(y~x+x2)
- x3=rnorm(n)
- m3=lm(y~x+x2+x3)
- x4=rnorm(n)
- m4=lm(y~x+x2+x3+x4)
- leaps(x=cbind(x,x2,x3,x4), y=y, method="Cp")

- extractAIC(m1)
- extractAIC(m1, k = log(n))

Criteria for Evaluating Subset Regression Models

- Uses of Regression and Model Evaluation Criteria
  - Regression equations may be used to make predictions
  - Minimizing the MSE for prediction may be an important criterion
  - The PRESS statistic can be used for comparisons of candidate models

$$PRESS_p = \sum_{i=1}^{n} \left(y_i - \hat{y}_{(i)} \right)^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2$$

  - Small values of PRESS are preferred

Computational Techniques for Variable Selection

- **All Possible Regressions**
  - Assume the intercept term is in all equations
  - If there are K regressors, we would investigate 2^K possible regression equations
  - Use the criteria discussed to determine some candidate models
  - Complete regression analysis on candidate models

Example 10.1 Hald Cement Data

Observation i	y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

Example 10.1 Hald Cement Data

TABLE 10.1 Summary of All Possible Regressions for 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

Example 10.1 Hald Cement Data

TABLE 10.2 Least-Squares Estimates for All Possible Regressions (Hald Cement Data)

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

Example 10.1 Hald Cement Data

TABLE 10.3 Matrix of Simple Correlations for Hald's Data in Example 10.1

	x_1	x_2	x_3	x_4	y
x_1	1.0				
x_2	0.229	1.0			
x_3	–0.824	–0.139	1.0		
x_4	–0.245	–0.973	0.030	1.0	
y	0.731	0.816	–0.535	–0.821	1.0

Example 10.1 Hald Cement Data

TABLE 10.4 Comparisons of Two Models for Hald's Cement Data

Observation i	$\hat{y} = 52.58 + 1.468x_1 + 0.662x_2$[a] e_i	h_{ii}	$[e_i/(1-h_{ii})]^2$	$\hat{y} = 71.65 + 1.452x_1 + 0.416x_2 - 0.237x_4$[b] 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] $R^2_{\text{Prediction}} = 0.9684$, $\text{VIF}_1 = 1.07$, $\text{VIF}_2 = 18.78$, $\text{VIF}_4 = 18.94$.

Example 10.1 Hald Cement Data

Best Subsets Regression: y versus x1, x2, x3, x4

Response is y

Vars	R-Sq	R-Sq(adj)	C-p	S	x 1	x 2	x 3	x 4
1	67.5	64.5	138.7	8.9639				X
1	66.6	63.6	142.5	9.0771		X		
1	53.4	49.2	202.5	10.727	X			
1	28.6	22.1	315.2	13.278			X	
2	97.9	97.4	2.7	2.4063	X	X		
2	97.2	96.7	5.5	2.7343	X			X
2	93.5	92.2	22.4	4.1921			X	X
2	84.7	81.6	62.4	6.4455		X	X	
2	68.0	61.6	138.2	9.3214		X		X
3	98.2	97.6	3.0	2.3087	X	X		X
3	98.2	97.6	3.0	2.3121	X	X	X	
3	98.1	97.5	3.5	2.3766	X		X	X
3	97.3	96.4	7.3	2.8638		X	X	X
4	98.2	97.4	5.0	2.4460	X	X	X	X

Figure 10.7 Computer output (Minitab) for Furnival and Wilson all-possible-regression algorithm.

Computational Techniques for Variable Selection

- **Stepwise Regression Methods**
  1. Forward Selection
  2. Backward Elimination
  3. Stepwise Regression (combination of forward and backward)

Stepwise Regression Methods

- Forward Selection
  - Procedure is based on the idea that no variables are in the model originally, but are added one at a time
  - The first regressor selected to be entered into the model is the one with the highest correlation with the response
  - If the F statistic corresponding to the model containing this variable is significant (larger than some predetermined value, F_{in}), then that regressor is left in the model
  - The second regressor examined is the one with the largest partial correlation with the response. If the *F*-statistic corresponding to the addition of this variable is significant, the regressor is retained
  - This process continues until all regressors are examined

Forward Selection

Stepwise Regression: y versus x1, x2, x3, x4

Forward selection. Alpha-to-enter: 0.25

Response is y on 4 predictors, with N = 13

Step	1	2	3
Constant	117.57	103.10	71.65
x4	-0.738	-0.614	-0.237
T-Value	-4.77	-12.62	-1.37
P-Value	0.001	0.000	0.205
x1		1.44	1.45
T-value		10.40	12.41
P-Value		0.000	0.000
x2			0.42
T-Value			2.24
R-Value			0.052
S	8.96	2.73	2.31
R-Sq	67.45	97.25	98.23
R-Sq(adj)	64.50	96.70	97.64
Mallows C-p	138.7	5.5	3.0

Figure 10.8 Forward selection results from Minitab for the Hald cement data.

Stepwise Regression Methods

- **Backward Elimination**
  - Procedure is based on the idea that all variables are in the model originally and examined one at a time and removed if not significant
  - The partial F statistic is calculated for each variable *as if it were the last one added to the model*
  - The regressor with the smallest F statistic is examined first and will be removed if this value is less than some predetermined value F_{out}
  - If this regressor is removed, then the model is refit with the remaining regressor variables and the partial F statistics calculated again
  - The regressor with the smallest partial F statistic will be removed if that value is less than F_{out}.
  - The process continues until all regressors are examined

Backward Elimination

Stepwise Regression: y versus x1, x2, x3, x4

Backward elimination. Alpha-to-Remove: 0.1

Response is y on 4 predictors, with N = 13

Step	1	2	3
Constant	62.41	71.65	52.58
x1	1.55	1.45	1.47
T-Value	2.08	12.41	12.10
P-Value	0.071	0.000	0.000
x2	0.510	0.416	0.662
T-Value	0.70	2.24	14.44
P-Value	0.501	0.052	0.000
x3	0.10		
T-Value	0.14		
P-Value	0.896		
x4	-0.14	-0.24	
T-Value	-0.20	-1.37	
P-Value	0.844	0.205	
S	2.45	2.31	2.41
R-Sq	98.24	98.23	97.87
R-Sq(adj)	97.36	97.64	97.44
Mallows C-p	5.0	3.0	2.7

Figure 10.9 Backward selection results from Minitab for the Hald cement data.

Stepwise Regression Methods

- **Stepwise Regression**
  - This procedure is a modification of forward selection
  - The contribution of each regressor variable that is put into the model is reassessed by way of its partial F statistic
  - A regressor that makes it into the model, may also be removed it if is found to be insignificant with the addition of other variables to the model
  - If the partial F-statistic is less than F_{out}, the variable will be removed.
  - Stepwise requires both an F_{in} value and F_{out} value

Stepwise Regression

Stepwise Regression: y versus x1, x2, x3, x4

Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15

Response is y on 4 predictors, with N = 13

Step	1	2	3	4
Constant	117.57	103.10	71.65	52.58
x4	-0.738	-0.614	-0.237	
T-Value	-4.77	-12.62	-1.37	
P-Value	0.001	0.000	0.205	
x1		1.44	1.45	1.47
T-Value		10.40	12.41	12.10
P-Value		0.000	0.000	0.000
x2			0.416	0.662
T-Value			2.24	14.44
P-Value			0.052	0.000
S	8.96	2.73	2.31	2.41
R-Sq	67.45	97.25	98.23	97.87
R-Sq(adj)	64.50	96.70	97.64	97.44
Mallows C-p	138.7	5.5	3.0	2.7

Figure 10.10 Stepwise selection results from Minitab for the Hald cement data.

Computational Techniques for Variable Selection

- **Stepwise Regression Methods**
  - No one model may be the "best"
  - The three stepwise techniques could result in different models
  - Inexperienced analysts may use the final model simply because the procedure spit it out.

Strategy for Regression Model Building

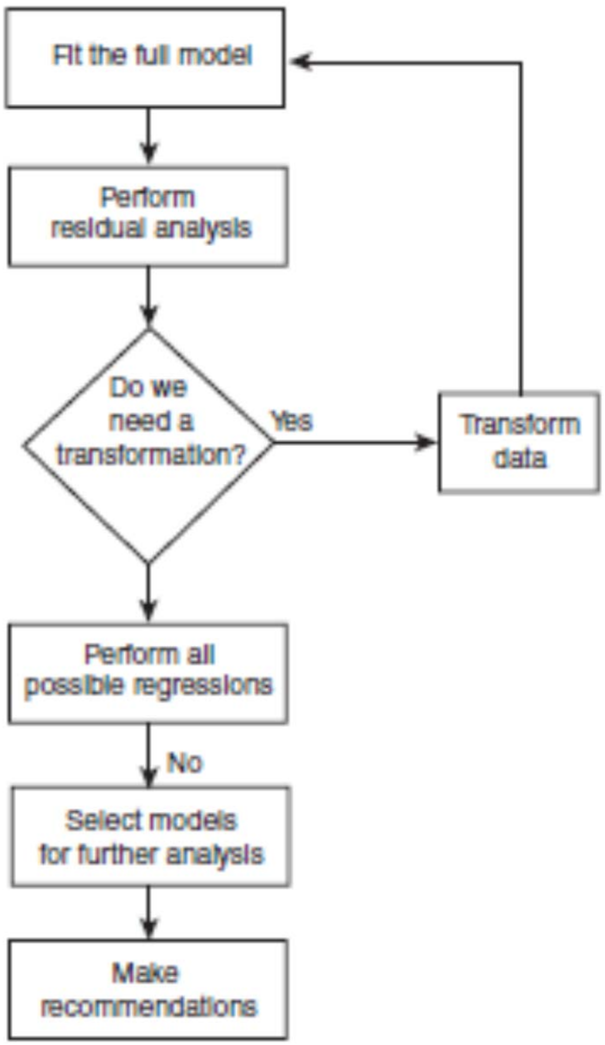

Figure 10.11 Flowchart of the model-building process.

R code

- library(MASS)
- asphalt <- read.csv("data-table-10-5 (Asphalt).csv",h=T)
- temp <- lm(y~x1+x2+x3+x4+x5+x6,data=asphalt)
- step <- stepAIC(temp, direction="both")
- step$anova
- library(leaps)
- temp <- regsubsets(y~x1+x2+x3+x4+x5+x6,data=asphalt,nbest=10)
- summary(temp)
- plot(temp)