Chapter 8
Variable Selection

Terry Dielman

Applied Regression Analysis:

A Second Course in Business and Economic Statistics, fourth edition

8.1 Introduction

- Previously we discussed some tests (t-test and partial F) that helped us determine whether certain variables should be in the regression.

- Here we will look at several *variable selection strategies* that expand on this idea.

Why is This Important?

- If an important variable is omitted, the estimated regression coefficients can become biased (systematically too high or low).
- Their standard errors can become inflated, leading to imprecise intervals and poor power in hypothesis tests.

Strategies

- All possible regressions: computer procedures that briefly examine every possible combination of Xs and report summaries of fit ability.
- Selection algorithms: rules for deciding when to drop or add variables
  1. Backwards Elimination
  2. Forward Selection
  3. Stepwise Regression

Words of Caution

- None guarantee you get the right model because they do not check assumptions or search for omitted factors like curvature.
- None have the ability to use a researcher's knowledge about the business or economic situation being analyzed.

8.2 All Possible Regressions

- If there are k *x* variables to consider using, there are (2^k) possible subsets. For example, with only k=5, there are 32 regression equations.

- Obtaining these sounds like a ton of work but programs like SAS or Minitab have algorithms that can measure fit ability without really producing the equation.

Typical Output

- The program will usually give you a summary table.
- Each line on the table will tell you which variables were in the model, plus measures of fit ability.
- These measures include R^2, adjusted R^2, S_e and a new one, C_p

The C_p Statistic (Mallow's Cp)

p = k + 1 is the number of terms in the model, including the intercept.

SSE_p is the SSE of this model

MSE_F is the MSE in the "full model" (with all the variables)

$$C_p = \frac{SSE_p}{MSE_F} - (n - 2p)$$

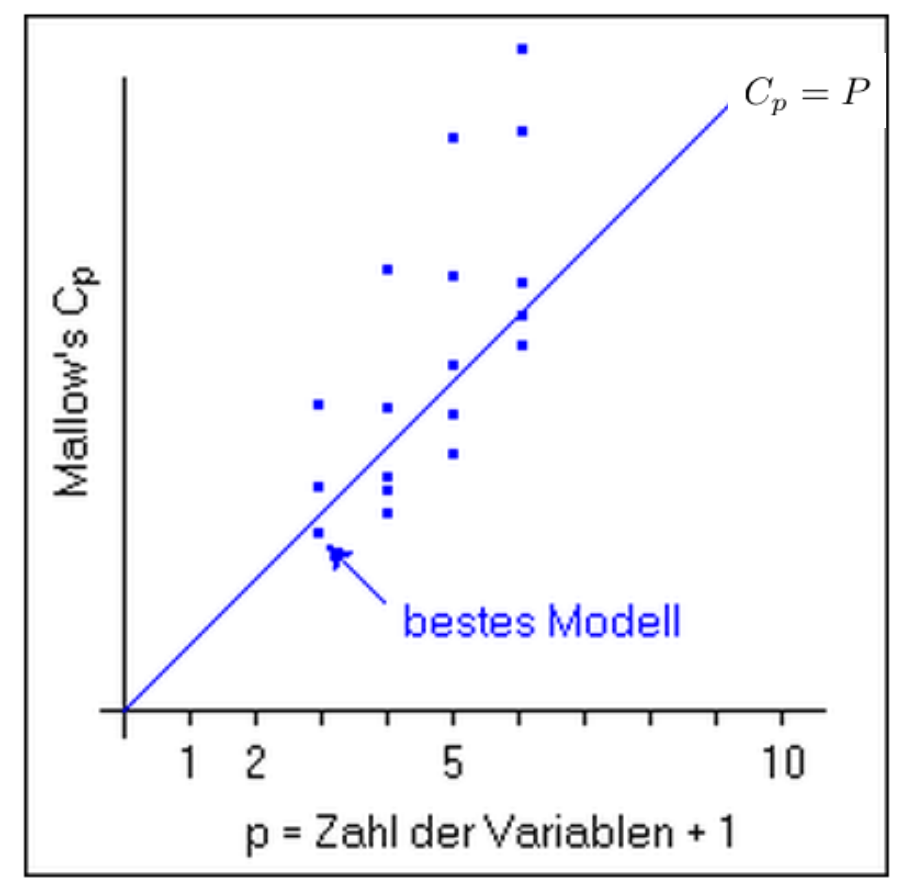
Using The C_p Statistic

Theory says that in a model with bias, C_p will be large.

It also says that in a model with no bias, C_p should be equal to p.

It is thus recommended that we consider models with a small C_p and those with C_p near $p = k + 1$.

Mallow's C_p Statistic



Example 8.1 Meddicorp Revisited

n = 25 sales territories

y = Sales (in $1000s) in each territory

x_1 = Advertising ($100s) in the territory

x_2 = Bonuses paid (in $100s) in the territory

x_3 = Market share in the territory

x_4 = largest competitor's sales ($1000s)

x_5 = Region code (1 = S, 2 = W, 3 = MW)

We are not using region here because it should be converted to indicator variables which should be examined as a group.

Summary Results For All Possible Regressions

Variables in the Regression	R^2	R^2_{adj}	C_p	S_e
ADV	81.1	80.2	5.90	101.42
BONUS	32.3	29.3	75.19	191.76
COMPET	14.2	10.5	100.85	215.83
MKTSHR	0.1	0.0	120.97	232.97
ADV, BONUS	85.5	84.2	1.61	90.75
ADV, MKTSHR	81.2	79.5	7.66	103.23
ADV, COMPET	81.2	79.5	7.74	103.38
BONUS, COMPET	38.7	33.2	68.03	186.51
BONUS, MKTSHR	32.8	26.7	76.46	195.33
COMPET, MKTSHR	16.1	8.5	100.18	218.20
ADV, BONUS, MKTSHR	85.8	83.8	3.10	91.75
ADV, BONUS, COMPET	85.7	83.6	3.30	92.26
ADV, MKTSHR, COMPET	81.3	78.6	9.60	105.52
BONUS, MKTSHR, COMPET	40.9	32.5	66.90	187.48
ADV, BONUS, MKTSHR, COMPET	85.9	83.1	5.00	93.77

The Best Model?

- The two variable model with ADV and BONUS has the smallest C_p and highest adjusted R^2.
- The three variable models adding either MKTSHR or COMPET also have small C_p values but only modest increases in R^2.
- The two-variable model is probably the best.

Minitab Results

Vars	R-Sq	R-Sq(adj)	C-p	S	ADULT	BOOKS	MKTSPH	COMPUT
1	81.1	80.2	5.9	101.42	X			
1	32.3	29.3	75.2	191.76		X		
2	85.5	84.2	1.6	90.749	X	X		
2	81.2	79.5	7.7	103.23	X		X	
3	85.8	83.8	3.1	91.751	X	X	X	
3	85.7	83.6	3.3	92.255	X	X		X
4	85.9	83.1	5.0	93.770	X	X	X	X

By default, the Best Subsets procedure prints two models for each number of X variables. This can be increased up to 5.

Limitations

- With a large number of potential *x* variables, the all possible approach becomes unwieldy.
- Minitab can use up to 31 predictors, but warns that computational time can be long when as few as 15 are used.
- "Obviously good" predictors can be forced into the model, thus reducing search time, but this is not always what you want.

8.3 Other Variable Selection Techniques

- With a large number of potential *x* variables, it may be best to use one of the iterative selection methods.

- These will look at only the set of models that their rules will lead them too, so they may not yield a model as good as that returned by the all possible regressions approach.

8.3.1 Backwards Elimination

1. Start with all variables in the equation.
2. Examine the variables in the model for significance and identify the least significant one.
3. Remove this variable if it does not meet some minimum significance level.
4. Run a new regression and repeat until all remaining variables are significant.

No Search Routine Needed?

- Although most software packages have automatic procedures for backwards elimination, it is fairly easy to do interactively.
- Run a model, check its t-tests for significance, and identify the variable to drop.
- Run again with one less variable and repeat the steps.

Step 1 – All Variables

Regression Analysis: SALES versus ADV, BONUS, MKTSHR, COMPET

The regression equation is
SALES = - 594 + 2.51 ADV + 1.91 BONUS + 2.65 MKTSHR - 0.121 COMPET

Predictor	Coef	SE Coef	T	P
Constant	-593.5	259.2	-2.29	0.033
ADV	2.5131	0.3143	8.00	0.000
BONUS	1.9059	0.7424	2.57	0.018
MKTSHR	2.651	4.636	0.57	0.574
COMPET	-0.1207	0.3718	-0.32	0.749

S = 93.77 R-Sq = 85.9% R-Sq(adj) = 83.1%

Least Significant

Step 2 – COMPET Eliminated

Regression Analysis: SALES versus ADV, BONUS, MKTSHR

The regression equation is

SALES = - 621 + 2.47 ADV + 1.90 BONUS + 3.12 MKTSHR

Predictor	Coef	SE Coef	T	P
Constant	-620.6	240.1	-2.58	0.017
ADV	2.4698	0.2784	8.87	0.000
BONUS	1.9003	0.7262	2.62	0.016
MKTSHR	3.116	4.314	0.72	0.478

S = 91.75 R-Sq = 85.8% R-Sq(adj) = 83.8%

Step 3 – MKTSHR Eliminated

Regression Analysis: SALES versus ADV, BONUS

The regression equation is

SALES = - 516 + 2.47 ADV + 1.86 BONUS

Predictor	Coef	SE Coef	T	P
Constant	-516.4	189.9	-2.72	0.013
ADV	2.4732	0.2753	8.98	0.000
BONUS	1.8562	0.7157	2.59	0.017

S = 90.75 R-Sq = 85.5% R-Sq(adj) = 84.2%

8.3.2 Forward Selection

- At each stage, it looks at the *x* variables not in the current equation and tests to see if they will be significant if they are added.
- In the first stage, the *x* with the highest correlation with *y* is added.
- At later stages it is much harder to see how the next *x* is selected.

Minitab Output for Forward Selection

An option in the Stepwise procedure

```
Forward selection.  Alpha-to-Enter: 0.25

Response is SALES on 4 predictors, with N = 25

    Step          1        2
Constant     -157.3   -516.4

ADV            2.77     2.47
T-Value        9.92     8.98
P-Value       0.000    0.000

BONUS                   1.86
T-Value                 2.59
P-Value                0.017

S               101     90.7
R-Sq          81.06    85.49
R-Sq(adj)     80.24    84.18
C-p             5.9      1.6
```

Same Model as Backwards

- This data set is not too complex, so both procedures returned the same model.
- With larger data sets, particularly when the *x* variables are correlated among themselves, results can be different.

8.3.3 Stepwise Regression

- A limitation with the backwards procedure is that a variable that gets eliminated is never considered again.
- With forward selection, variables entering stay in, even if they lose significance.
- Stepwise regression corrects these flaws. A variable entering can later leave. A variable eliminated can later go back in.

Minitab Output for Stepwise Regression

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

Response is SALES on 4 predictors, with N = 25


    Step          1        2
Constant     -157.3   -516.4

ADV            2.77     2.47
T-Value        9.92     8.98
P-Value       0.000    0.000

BONUS                   1.86
T-Value                 2.59
P-Value                0.017

S               101     90.7
R-Sq          81.06    85.49
R-Sq(adj)     80.24    84.18
C-p             5.9      1.6
```

Selection Parameters

- For backwards elimination, the user specifies "Alpha to Remove", which is the maximum p-value a variable can have and stay in the equation.
- For forward selection, the user specifies "Alpha to Enter", which is the minimum p-vale a variable needs to enter the equation.
- Stepwise regression gets both.
- Often we use values like .15 or .20 because this encourages the procedures to look at models with more variables.

8.4 Which Procedure is Best?

- Unless there are too many *x* variables, the all possible models approach is favored because it looks at all combinations of variables.
- Of the other strategies, stepwise regression is probably best.
- If no search programs are available, backwards elimination can still provide a useful sifting of the data.

No Guarantees

- Because they do not check assumptions or examine the model residuals, there is no guarantee of returning the right model.
- Nonetheless, these can be effective tools filtering the data and identifying which variables to pay more attention to.

```
proc  reg  data=a ;
model  sales = ADV  BONUS
                MKTSHR  COMPET / selection = ?    ;
```

```
selection = rsquare   adjrsq   aic   cp
            forward  backward  stepwise  ;
```

$$AIC = n \cdot \ln\left(\frac{SSE}{n}\right) + 2p$$

```
selection = cp   best=4  ;
```

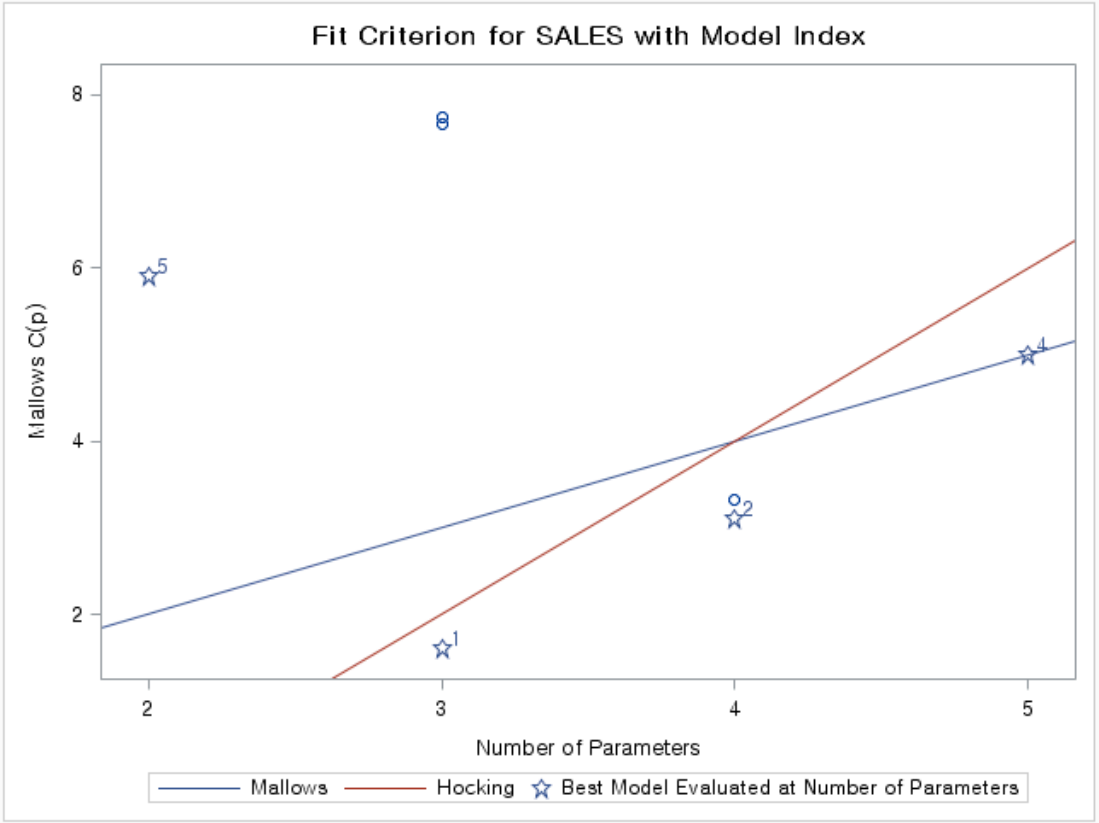
Number in Model	C(p)	R-Square	Variables in Model
2	1.6052	0.8549	ADV BONUS
3	3.1054	0.8585	ADV BONUS MKTSHR
3	3.3270	0.8569	ADV BONUS COMPET
4	5.0000	0.8592	ADV BONUS MKTSHR COMPET
1	5.9046	0.8106	ADV

Mallows' Cp in SAS

Mallows (1973) says Cp ~ p and small

Hocking (1976) says Cp <= 2p – (k+1)

Model Index	Number in Model	C(p)	R-Square	Variables in Model
1	2	1.6052	0.8549	ADV BONUS
2	3	3.1054	0.8585	ADV BONUS MKTSHR
3	3	3.3270	0.8569	ADV BONUS COMPET
4	4	5.0000	0.8592	ADV BONUS MKTSHR COMPET
5	1	5.9046	0.8106	ADV
6	2	7.6612	0.8123	ADV MKTSHR
7	2	7.7404	0.8117	ADV COMPET

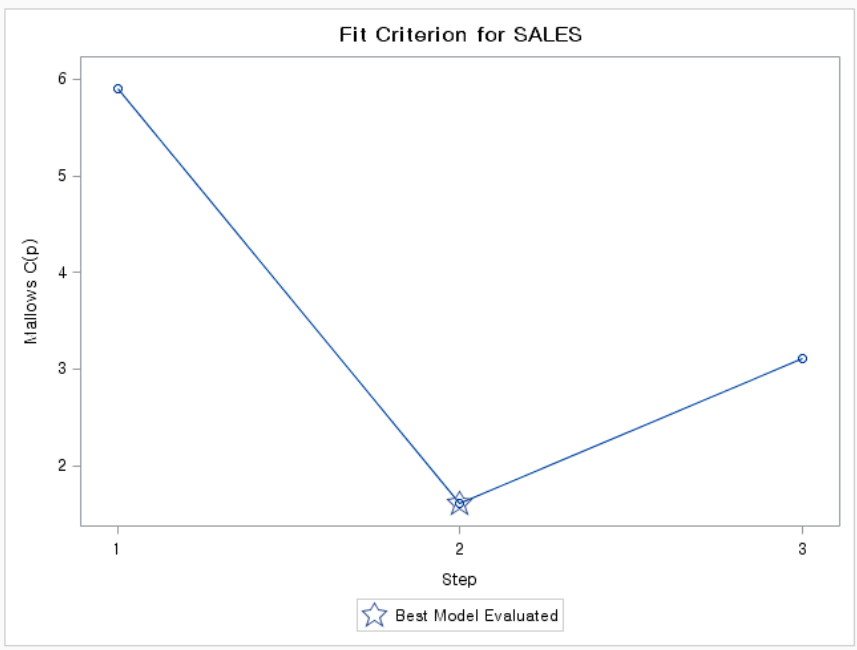


Forward Selection in SAS

selection = forward ;

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	ADV	1	0.8106	0.8106	5.9046	98.43	<.0001
2	BONUS	2	0.0443	0.8549	1.6052	6.73	0.0166
3	MKTSHR	3	0.0035	0.8585	3.1054	0.52	0.4780

Partial F-test with alpha=0.5



SLE = 0.5
(significance level for entry)

Backward Elimination in SAS

selection = backward ;

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.8592 and C(p) = 5.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	1073119	268280	30.51	<.0001
Error	20	175855	8792.75990		
Corrected Total	24	1248974			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-593.53745	259.19585	46107	5.24	0.0330
ADV	2.51314	0.31428	562260	63.95	<.0001
BONUS	1.90595	0.74239	57955	6.59	0.0184
MKTSHR	2.65101	4.63566	2875.57485	0.33	0.5738
COMPET	-0.12073	0.37181	927.06773	0.11	0.7488

Bounds on condition number: 1.4799, 20.838

Backward Elimination: Step 1

Variable COMPET Removed: R-Square = 0.8585 and C(p) = 3.1054

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	COMPET	3	0.0007	0.8585	3.1054	0.11	0.7488
2	MKTSHR	2	0.0035	0.8549	1.6052	0.52	0.4780

Partial F-test with alpha=0.1

SLS = 0.1
(significance level for stay)

Stepwise Regression in SAS

selection = stepwise ;

Stepwise Selection: Step 1

Variable ADV Entered: R-Square = 0.8106 and C(p) = 5.9046

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1012408	1012408	98.43	<.0001
Error	23	236566	10285		
Corrected Total	24	1248974			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-157.33011	145.19120	12077	1.17	0.2898
ADV	2.77212	0.27941	1012408	98.43	<.0001

Bounds on condition number: 1, 1

Stepwise Selection: Step 2

Variable BONUS Entered: R-Square = 0.8549 and C(p) = 1.6052

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	ADV		1	0.8106	0.8106	5.9046	98.43	<.0001
2	BONUS		2	0.0443	0.8549	1.6052	6.73	0.0166