Chapter 4 Model Building and Effect Selection

4.1 All Possible Selection (Self-Study)

Objectives

- Explain the REG procedure options for all possible model selection.
- Describe model selection options and interpret output to evaluate the fit of several models.

47

Model Selection

Data set contains eight interval variables as potential predictors.

Possible Option #1:

Use a form of Stepwise Selection by hand or with assistance from SAS.

Possible Option #2:

Explore all possible models and determine “best.”

48

A process for selecting models might be to start with all the interval variables in the **STAT1.ameshousing3** data set and invoke some form of stepwise selection discussed in previous sections. This could be done by hand or with the assistance of SAS.

An alternative option would be to explore all possible models capable from the predictor variables provided and determine which is “best.” This method of all possible selection can be performed using PROC REG.

Model Selection Options

The SELECTION= option in the MODEL statement of PROC REG supports these model selection techniques:

Stepwise selection methods

- STEPWISE, FORWARD, or BACKWARD using significance level

All-possible regressions ranked using

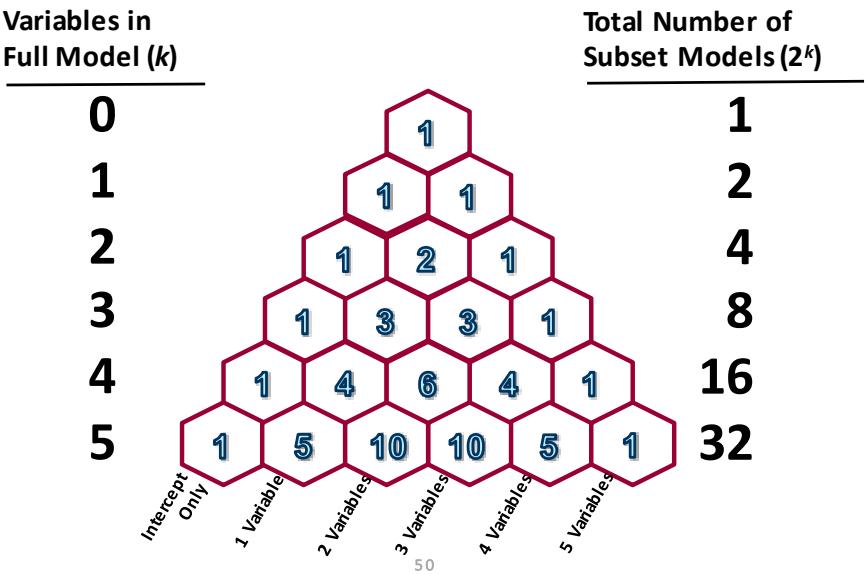
- RSQUARE, ADJRSQ, or CP

SELECTION=NONE is the default.

49

sas

RSQUARE, ADJRSQ, CP Selection Options



50

sas

In the **STAT1.ameshousing3** data set, there are eight possible independent variables. Therefore, there are 2^8=256 possible regression models. There are eight possible one-variable models, 28 possible two-variable models, 56 possible three-variable models, and so on.

You can choose to only look at the best n (as measured by the model R^2 for k=1, 2, 3, …, 7) by using the BEST= option on the model statement. The BEST= option only reduces the output. All regressions are still calculated.

If there were 20 possible independent variables, there would be more than 1,000,000 models.

Mallows' C_p

- Mallows' C_p is a simple indicator of effective variable selection within a model.
- Look for models with $C_p \leq p$, where p equals the number of parameters in the model, including the intercept.

Mallows recommends choosing the first (fewest variables) model where C_p approaches p.

51

Mallows' C_p (1973) is estimated by $C_p = p + \frac{(MSE_p - MSE_{full})(n-p)}{MSE_{full}}$

where

MSE_p is the mean squared error for the model with p parameters.

MSE_{full} is the mean squared error for the full model used to estimate the true residual variance.

n is the number of observations.

p is the number of parameters, including an intercept parameter, if estimated.

The choice of the best model based on C_p is debatable, as will be shown in the slide about Hocking's criterion. Many choose the model with the smallest C_p value. However, Mallows recommended that the best model will have a C_p value approximating p. The most parsimonious model that fits that criterion is generally considered to be a good choice, although subject-matter knowledge should also be a guide in the selection from among competing models.

Hocking's Criterion versus Mallows' C_p

Hocking (1976) suggests selecting a model based on the following:

- $C_p \leq p$ for prediction
- $C_p \leq 2p - p_{full} + 1$ for parameter estimation

52

Hocking suggested the use of the C_p statistic, but with alternative criteria, depending on the purpose of the analysis. His suggestion of ($C_p \leq 2p - p_{full} + 1$) is included in the REG procedure's calculations of criteria reference plots for best models.

All Possible Model Selection

Example: Invoke PROC REG to produce a regression of **SalePrice** on all the other interval variables in the **STAT1.ameshousing3** data set.

Note: Currently, **stepwise**, **forward**, and **backward** are the only three selection methods that can be chosen in the SAS Studio task. To perform model selection using a method other than these three, either manually edit the generated code or write the code directly.

```
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
         Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;

/*st104d03.sas*/ /*Part A*/
ods graphics on;
proc reg data=STAT1.ameshousing3 plots(only)=(rsquare adjrsq cp);
   ALLPOSS: model SalePrice=&interval
                  / selection=rsquare adjrsq cp;
   title "All Possible Model Selection for SalePrice";
run;
quit;
```

Selected MODEL statement options:

SELECTION= enables you to choose the different selection methods – RSQUARE, ADJRSQ, and CP. The first listed method is the one that determines the sorting order in the output.

Selected SELECTION= option methods:

RSQUARE tells PROC REG to use the model R-square to rank the model from best to worst for a given number of variables.

ADJRSQ prints the adjusted R-square for each model.

CP prints Mallows' C_p statistic for each model.

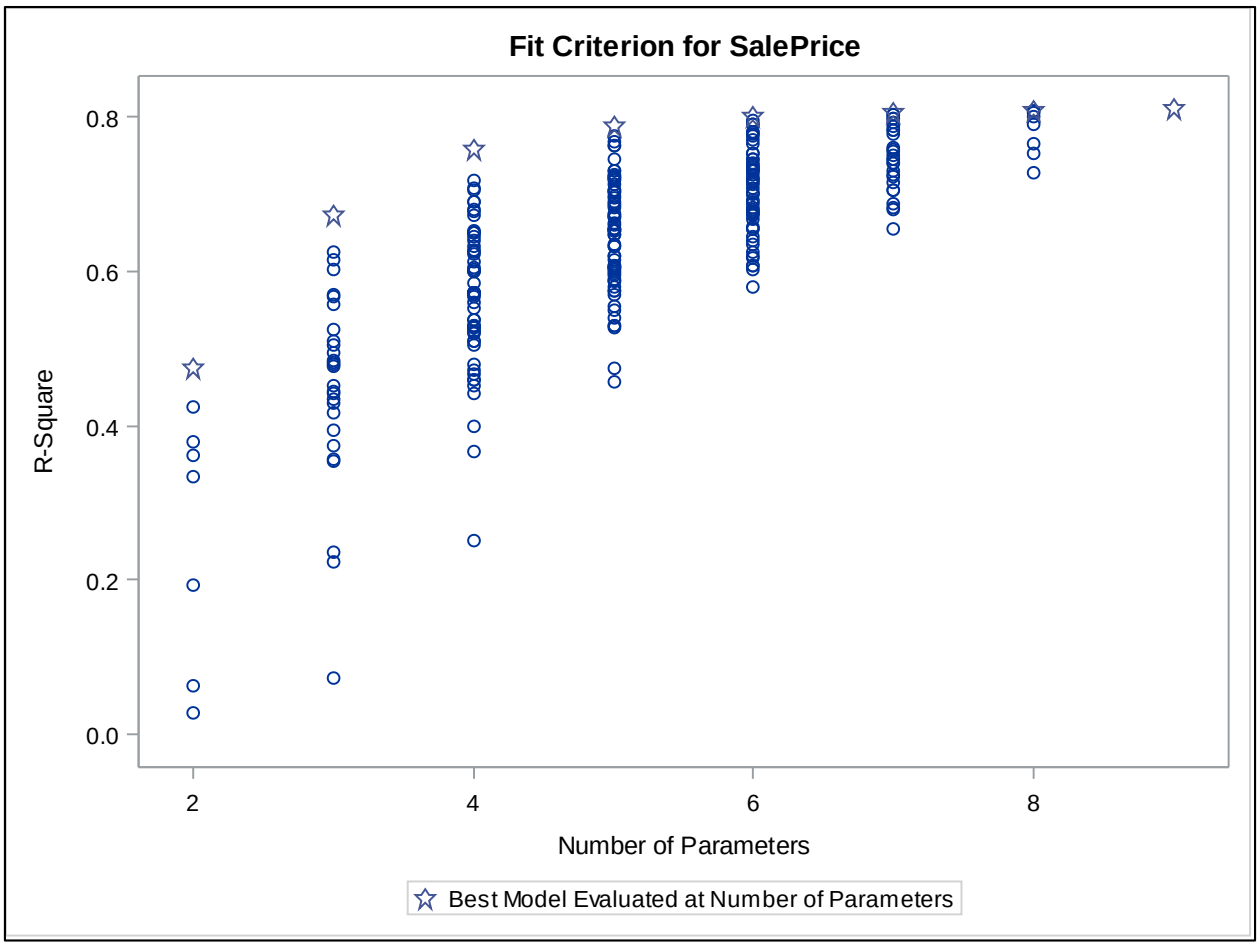
Partial PROC REG Output

Number of Observations Read	300
Number of Observations Used	300

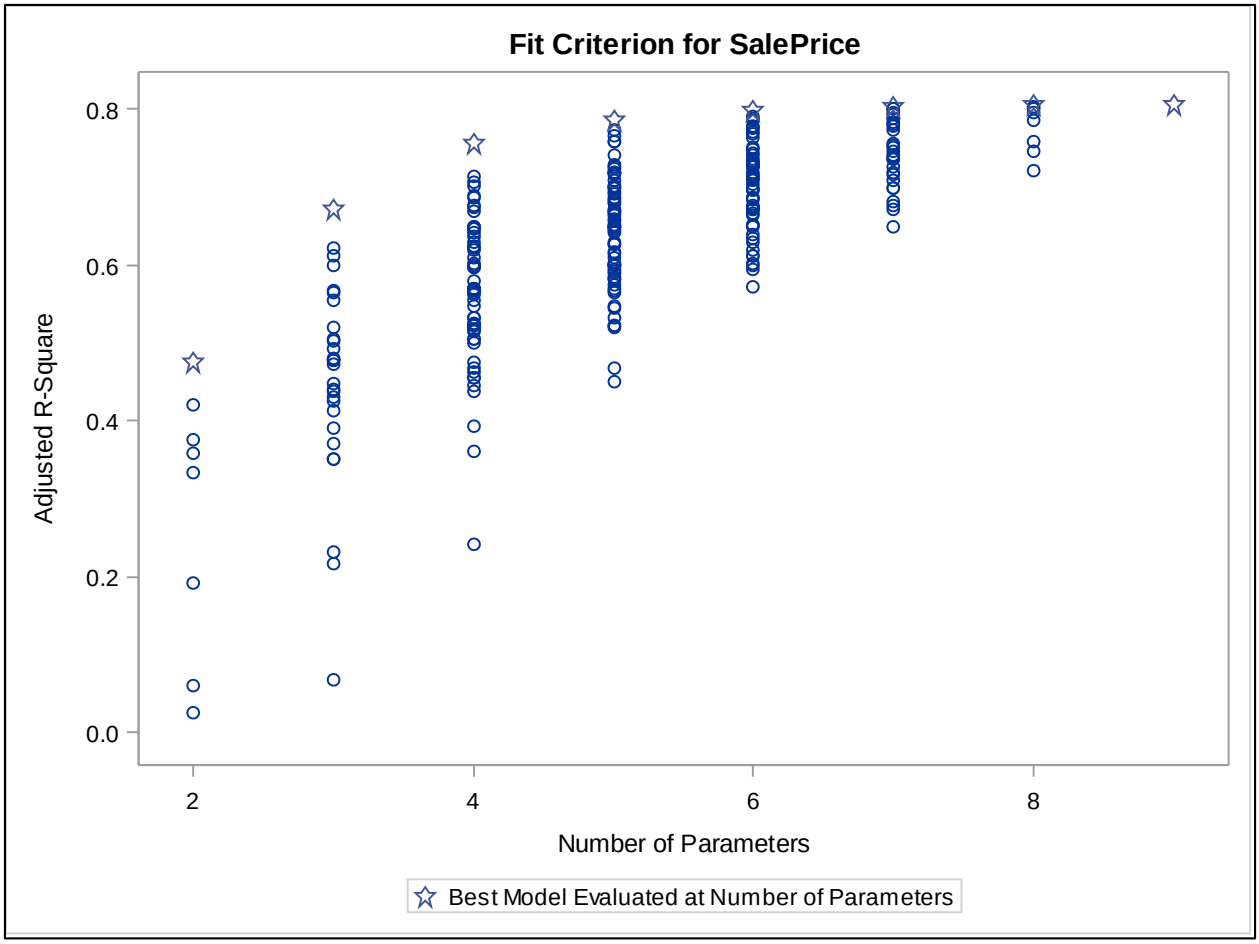
Model Index	Number in Model	R-Square	Adjusted R-Square	C(p)	Variables in Model
1	1	0.4755	0.4737	510.5367	Basement_Area
2	1	0.4231	0.4212	591.1052	Gr_Liv_Area
3	1	0.3787	0.3767	659.3108	Age_Sold
4	1	0.3605	0.3584	687.3455	Total_Bathroom
5	1	0.3351	0.3329	726.3533	Garage_Area
6	1	0.1935	0.1908	944.1662	Deck_Porch_Area
7	1	0.0642	0.0610	1143.019	Lot_Area
8	1	0.0275	0.0243	1199.378	Bedroom_AbvGr

Model Index	Number in Model	R-Square	Adjusted R-Square	C(p)	Variables in Model
9	2	0.6725	0.6703	209.5529	Gr_Liv_Area Age_Sold
10	2	0.6249	0.6224	282.7594	Gr_Liv_Area Basement_Area
11	2	0.6148	0.6122	298.3135	Basement_Area Age_Sold
12	2	0.6027	0.6000	316.9559	Basement_Area Garage_Area
13	2	0.5708	0.5679	365.9609	Gr_Liv_Area Garage_Area
14	2	0.5680	0.5651	370.2769	Basement_Area Total_Bathroom

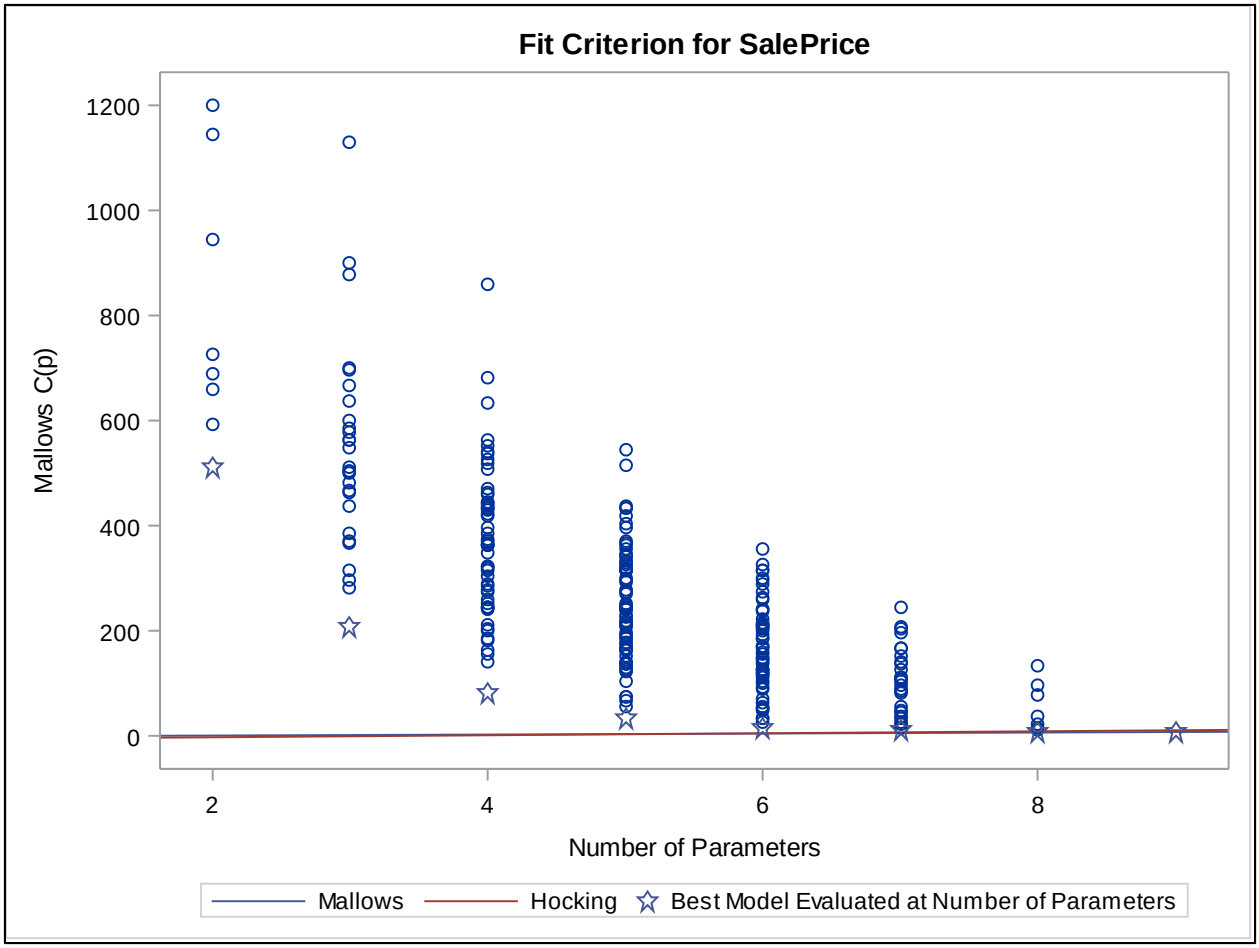
There are many models to compare. It would be unwieldy to try to determine the best model by viewing the output tables. Therefore, it is advisable to look at the ODS plots.



The R-square plot compares all models based on their R-square values. As noted earlier, adding variables to a model always increases R-square, and therefore the full model is always best. Therefore, you can only use the R-square value to compare models of equal numbers of parameters.



The adjusted R-square does not have the problem that the R-square has. You can compare models of different sizes. In this case, it is difficult to see which model has the higher adjusted R-square, the starred model for seven parameters or eight parameters.



The line $C_p=p$ is plotted to help you identify models that satisfy the criterion $C_p \leq p$ for prediction. The lower line is plotted to help identify which models satisfy Hocking's criterion $C_p \leq 2p - p_{full} + 1$ for parameter estimation.

Use the graph and review the output to select a relatively short list of models that satisfy the criterion appropriate for your objective. It is often the case that the best model is difficult to see because of the range of C_p values at the high end. These models are clearly not the best and therefore you can focus on the models near the bottom of the range of C_p.

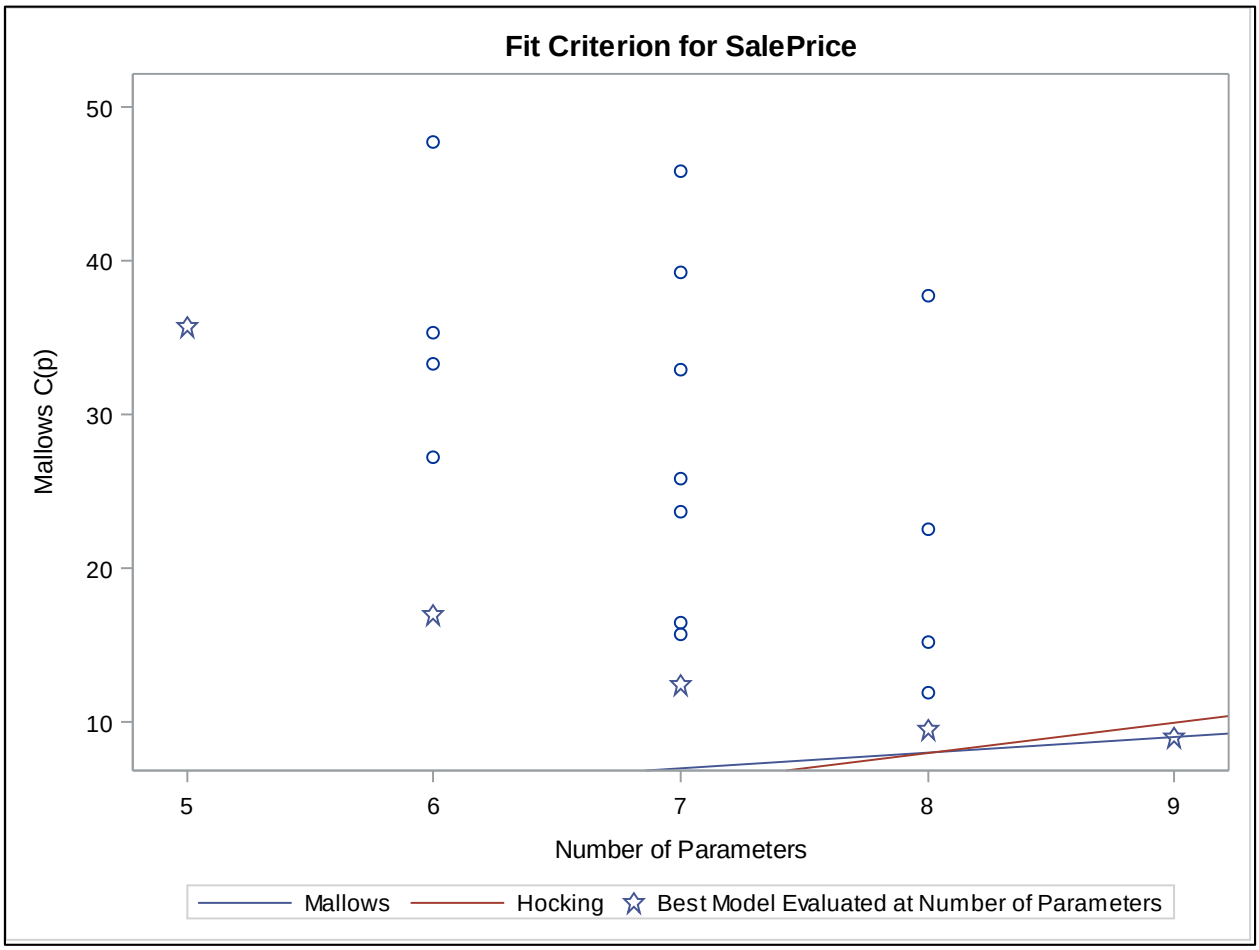
```
/*st104d03.sas*/ /*Part B*/
proc reg data=STAT1.ameshousing3 plots(only)=(cp);
     ALLPOSS: model SalePrice=&interval / selection=cp rsquare adjrsq
best=20;
     title "Best Models Using All Possible Selection for SalePrice";
run;
quit;
```

Selected SELECTION= option methods:

BEST=*n* limits the output to only the best *n* models.

Model Index	Number in Model	C(p)	R-Square	Adjusted R-Square	Variables in Model
1	8	9.0000	0.8108	0.8056	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom
2	7	9.4754	0.8091	0.8046	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr
3	7	11.8765	0.8076	0.8030	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Age_Sold Bedroom_AbvGr Total_Bathroom
4	6	12.4745	0.8059	0.8019	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Age_Sold Bedroom_AbvGr
5	7	15.1956	0.8054	0.8008	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Total_Bathroom
6	6	15.7530	0.8038	0.7997	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Age_Sold Total_Bathroom
7	6	16.4459	0.8033	0.7993	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold
8	5	17.0005	0.8017	0.7983	Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Age_Sold
9	7	22.5339	0.8007	0.7959	Gr_Liv_Area Basement_Area Garage_Area Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom
10	6	23.7403	0.7986	0.7944	Gr_Liv_Area Basement_Area Garage_Area Lot_Area Age_Sold Bedroom_AbvGr
11	6	25.8313	0.7972	0.7931	Gr_Liv_Area Basement_Area Garage_Area Age_Sold Bedroom_AbvGr Total_Bathroom
12	5	27.1943	0.7950	0.7915	Gr_Liv_Area Basement_Area Garage_Area Age_Sold Bedroom_AbvGr
13	6	32.9173	0.7926	0.7884	Gr_Liv_Area Basement_Area Garage_Area Lot_Area Age_Sold Total_Bathroom
14	5	33.3028	0.7911	0.7875	Gr_Liv_Area Basement_Area Garage_Area Age_Sold Total_Bathroom
15	5	35.3618	0.7897	0.7861	Gr_Liv_Area Basement_Area Garage_Area Lot_Area Age_Sold
16	4	35.7387	0.7882	0.7853	Gr_Liv_Area Basement_Area Garage_Area Age_Sold
17	7	37.7677	0.7907	0.7857	Gr_Liv_Area Basement_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom
18	6	39.3019	0.7885	0.7841	Gr_Liv_Area Basement_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr
19	6	45.8708	0.7842	0.7798	Gr_Liv_Area Basement_Area Deck_Porch_Area Age_Sold Bedroom_AbvGr Total_Bathroom
20	5	47.7363	0.7817	0.7780	Gr_Liv_Area Basement_Area Deck_Porch_Area Age_Sold Bedroom_AbvGr

Investigate the plot of Mallows' C(p).



In this example, the number of parameters in the full model, p_{full}, equals 9 (eight variables plus the intercept).

The smallest model that falls under the Hocking line has p=9, the full model. This model also has a Cp value that is equal to p exactly, falling directly on Mallows line. From this information, your full model appears to be a potential model for prediction and variable explanation. This result is likely to change if additional continuous predictors are included in the analysis.

If multiple models, sharing the same number of parameters, fall below these lines, there are several options that can be used to make a decision. First, the analyst can appeal to a subject matter expert who could potentially provide previous experiences that could "break the tie." Secondly, other fit statistics could be used as a comparison between the models. Perhaps one of the models has a higher adjusted R-square value. Thirdly, the models in question could be compared using a hold-out data set, especially when the focus is prediction.

End of Demonstration

4.02 Multiple Choice Poll

Which value tends to increase (can never decrease) as you add predictor variables to your regression model?

a. R square

b. Adjusted R square

c. Mallows' C_p

d. Both a and b

e. *F* statistic

f. All of the above

Exercises

1. Using All-Regression Techniques

Use the **STAT1.BodyFat2** data set to identify a set of "best" models.

a. With the SELECTION=CP option, use an all-possible regression technique to identify a set of candidate models that predict **PctBodyFat2** as a function of the variables **Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist**.

Hint: Select only the best 60 models based on C_p to compare.

End of Exercises

4.2 Solutions

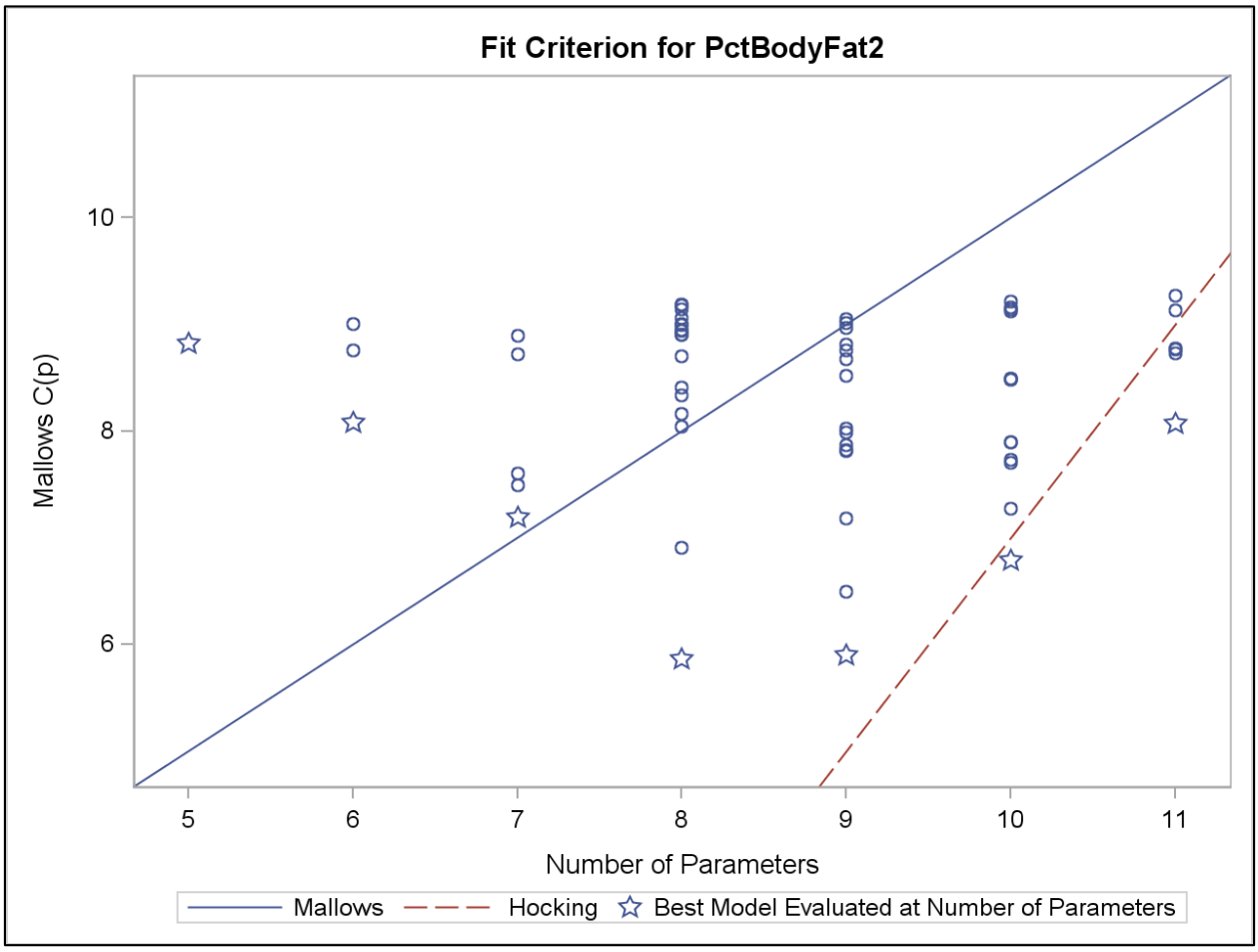
Solutions to Exercises

1. Using All-Regression Techniques

a. With the SELECTION=CP option, use an all-possible regression technique to identify a set of candidate models that predict **PctBodyFat2** as a function of the variables **Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist**. Hint: Select only the best 60 models based on C_p to compare.

```
/*st104s03.sas*/  /*Part A*/
ods graphics / imagemap=on;

proc reg data=STAT1.BodyFat2 plots(only)=(cp);
   model PctBodyFat2=Age Weight Height
         Neck Chest Abdomen Hip Thigh
         Knee Ankle Biceps Forearm Wrist
         / selection=cp best=60;
   title "Using Mallows Cp for Model Selection";
run;
quit;
```



The plot indicates that the best model according to Mallows' criterion is an eight-parameter (seven variables plus an intercept) model. The best model according to Hocking's criterion has 10 parameters (including the intercept).

A partial table of the 60 models, their C(p) values, and the numbers of variables in the models is displayed.

Model Index	Number in Model	C(p)	R-Square	Variables in Model
1	7	5.8653	0.7445	Age Weight Neck Abdomen Thigh Forearm Wrist
2	8	5.8986	0.7466	Age Weight Neck Abdomen Hip Thigh Forearm Wrist
3	8	6.4929	0.7459	Age Weight Neck Abdomen Thigh Biceps Forearm Wrist
4	9	6.7834	0.7477	Age Weight Neck Abdomen Hip Thigh Biceps Forearm Wrist
5	7	6.9017	0.7434	Age Weight Neck Abdomen Biceps Forearm Wrist
6	8	7.1778	0.7452	Age Weight Neck Abdomen Thigh Ankle Forearm Wrist
7	6	7.1860	0.7410	Age Weight Abdomen Thigh Forearm Wrist
8	9	7.2729	0.7472	Age Weight Neck Abdomen Hip Thigh Ankle Forearm Wrist
9	6	7.4937	0.7406	Age Weight Neck Abdomen Forearm Wrist
10	6	7.6018	0.7405	Weight Neck Abdomen Biceps Forearm Wrist
11	9	7.7067	0.7468	Age Weight Neck Abdomen Thigh Ankle Biceps Forearm Wrist
12	9	7.7282	0.7467	Age Weight Height Neck Abdomen Hip Thigh Forearm Wrist
13	8	7.8146	0.7445	Age Weight Height Neck Abdomen Thigh Forearm Wrist
14	8	7.8246	0.7445	Age Weight Neck Chest Abdomen Thigh Forearm Wrist
15	8	7.8651	0.7445	Age Weight Neck Abdomen Thigh Knee Forearm Wrist
16	9	7.8966	0.7466	Age Weight Neck Abdomen Hip Thigh Knee Forearm Wrist
17	9	7.8986	0.7466	Age Weight Neck Chest Abdomen Hip Thigh Forearm Wrist
18	8	7.9907	0.7443	Age Weight Neck Abdomen Ankle Biceps Forearm Wrist

Note: Number in Model does not include the intercept in this table.

The best MALLOWS model is either the eight-parameter models, number 1 (includes the variables Age, Weight, Neck, Abdomen, Thigh, Forearm, and Wrist) or number 5 (includes the variables Age, Weight, Neck, Abdomen, Biceps, Forearm, and Wrist).

The best HOCKING model is number 4. It includes Hip, along with the variables in the best MALLOWS models listed above.

End of Solutions

Solutions to Student Activities (Polls/Quizzes)

4.02 Multiple Choice Poll – Correct Answer

Which value tends to increase (can never decrease) as you add predictor variables to your regression model?

a. R square
b. Adjusted R square
c. Mallows' C_p
d. Both a and b
e. *F* statistic
f. All of the above

55