


Chapter 4: Model Building and Effect Selection

Sections

4.1 Stepwise Selection Using Significance Level

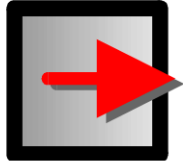
4.2 Information Criterion and Other Selection Options

4.3 All Possible Selection

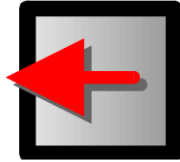


4.1 Stepwise Selection using Significance Level (SL)

Stepwise Selection Methods



FORWARD
SELECTION



BACKWARD
ELIMINATION



STEPWISE
SELECTION

Forward Selection

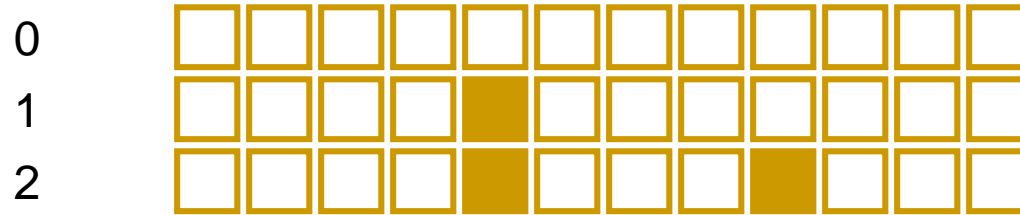
0



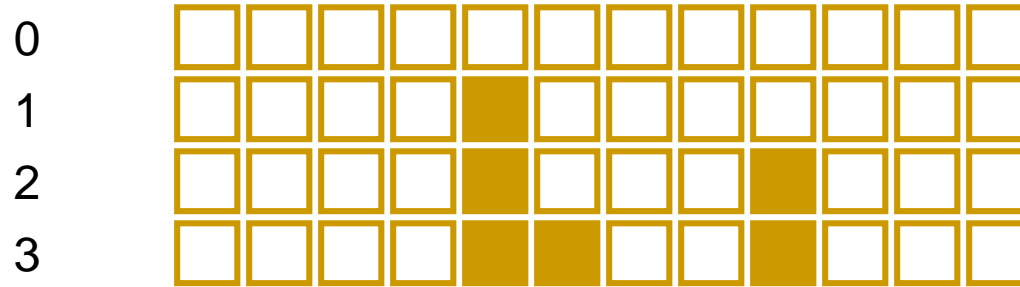
Forward Selection

0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

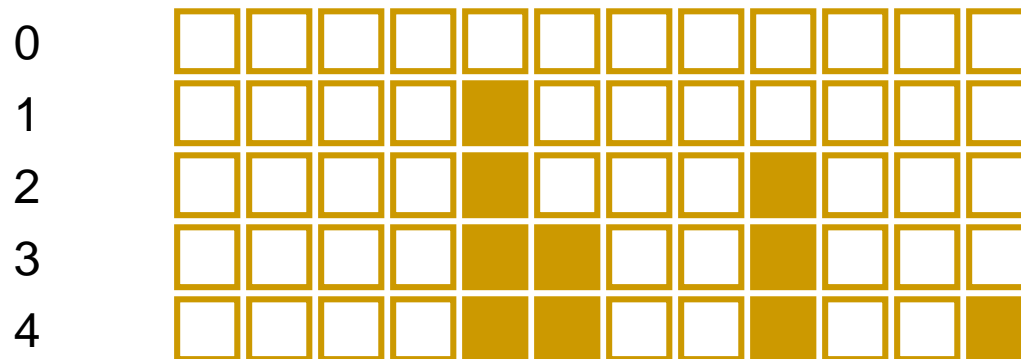
Forward Selection



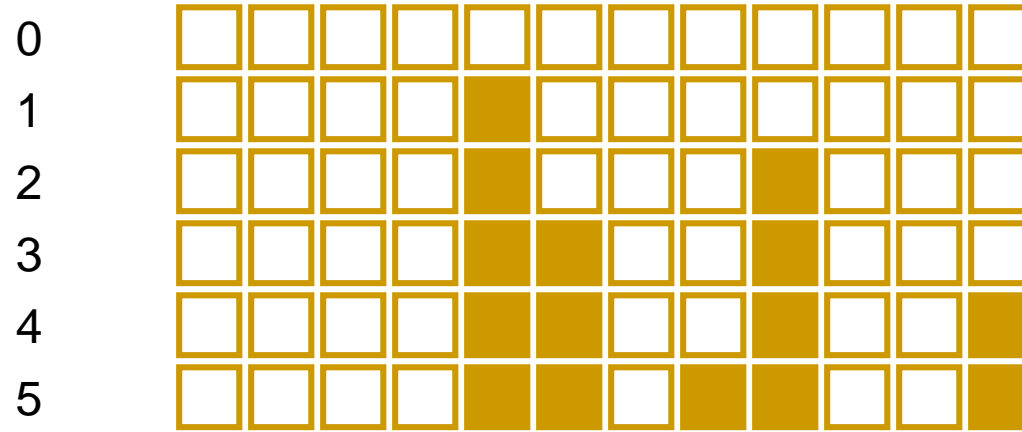
Forward Selection



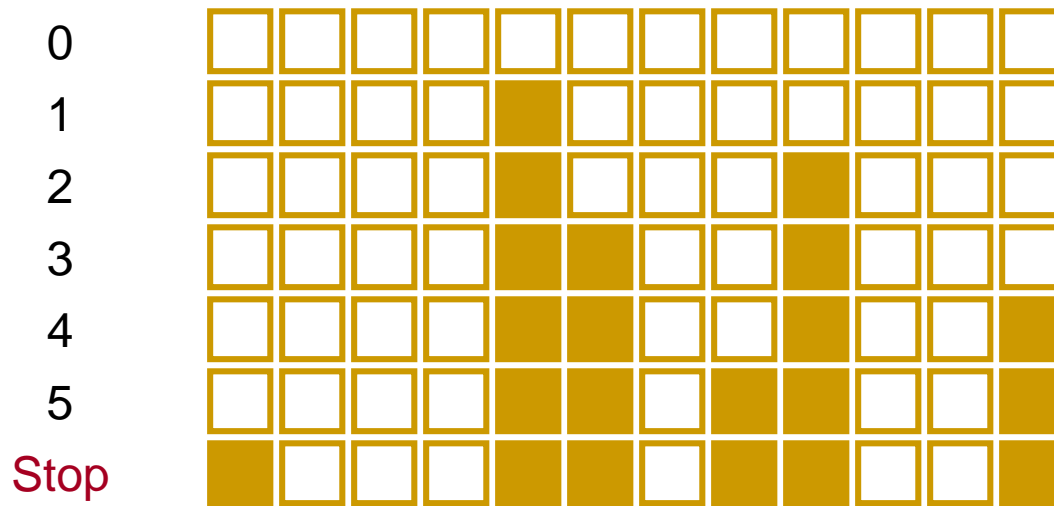
Forward Selection



Forward Selection



Forward Selection



Backward Elimination

0



Backward Elimination

0											
1											

Backward Elimination

0											
1											
2											

Backward Elimination

0											
1											
2											
3											

Backward Elimination

0											
1											
2											
3											
4											

Backward Elimination

0												
1												
2												
3												
4												
5												

Backward Elimination

0											
1											
2											
3											
4											
5											
6											

Backward Elimination

0											
1											
2											
3											
4											
5											
6											
Stop											

Stepwise Selection

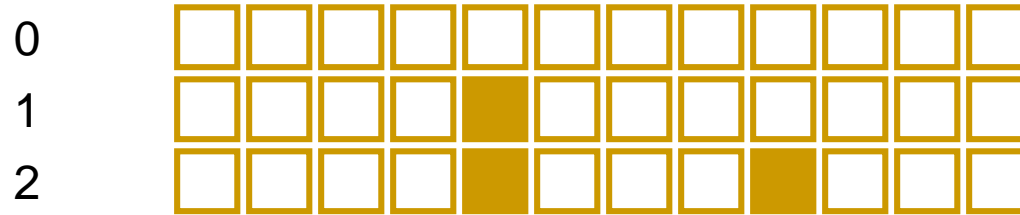
0



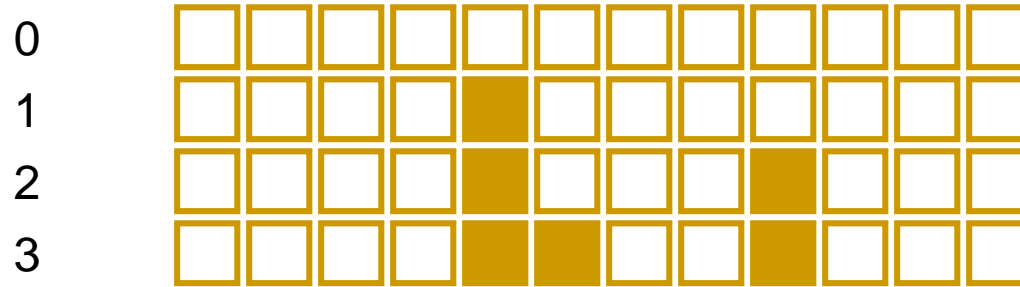
Stepwise Selection



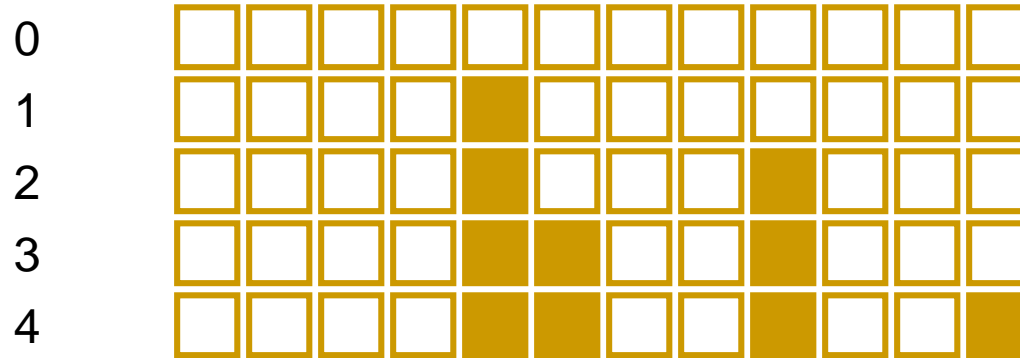
Stepwise Selection



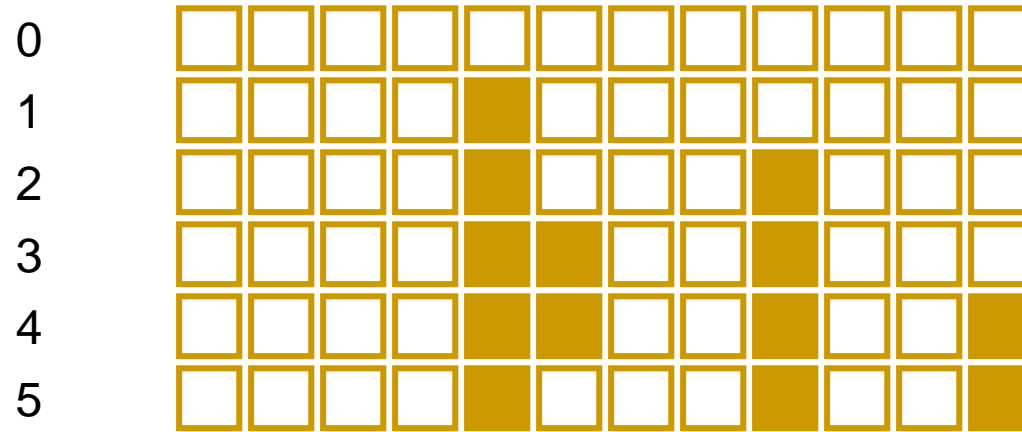
Stepwise Selection



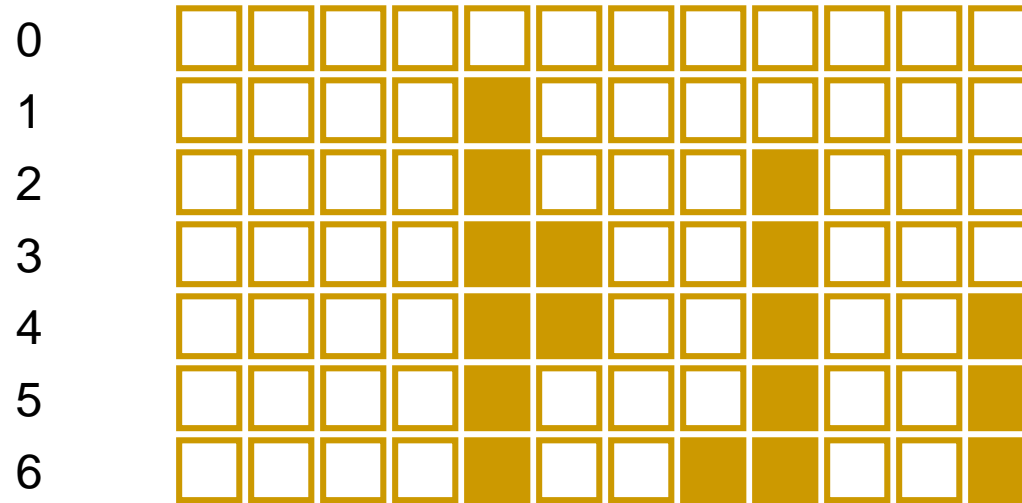
Stepwise Selection



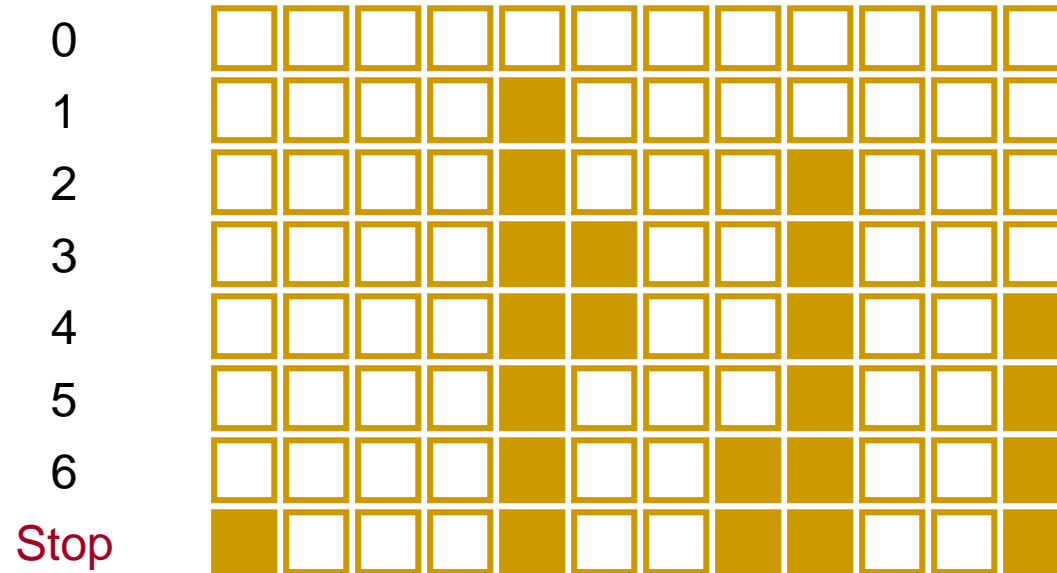
Stepwise Selection



Stepwise Selection



Stepwise Selection



PROC GLMSELECT

```
PROC GLMSELECT DATA=SAS-data-set <options>;  
    CLASS variables;  
    MODEL dependent(s)=regressor(s) </ options>;  
RUN;
```

Options within PROC GLMSELECT can assist with model selection.

Model Selection Options

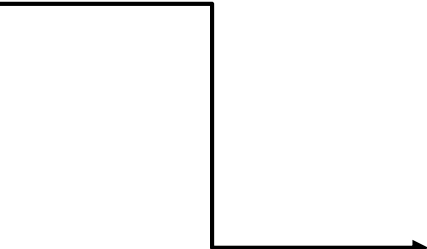
Options in the MODEL statement of PROC GLMSELECT support many model selection techniques and criteria.

- SELECTION=<*option*>
- CHOOSE=<*option*>
- SELECT=<*option*>
- STOP=<*option*>

Model Selection Options

Options in the MODEL statement of PROC GLMSELECT support many model selection techniques and criteria.

- **SELECTION=<option>**
- CHOOSE=<option>
- SELECT=<option>
- STOP=<option>

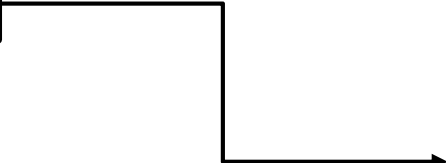


Specifies the method used in selecting the model (can choose: None, Forward, Backward, Stepwise, LAR, LASSO, Elasticnet (default is Stepwise))

Model Selection Options

Options in the MODEL statement of PROC GLMSELECT support many model selection techniques and criteria.

- SELECTION=<option>
- CHOOSE=<option>
- SELECT=<option>
- STOP=<option>

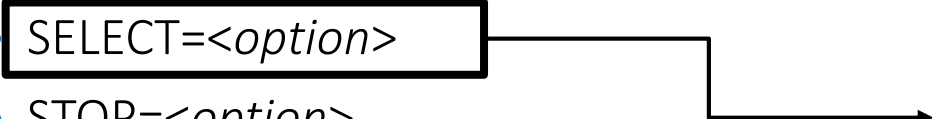


Specifies criterion for choosing the model (after all steps are done, this tells SAS to go back and look at this criterion in ALL steps and choose the best one (if this is not present, the last model is selected). Options are: ADJRSQ, AIC, AICC, BIC, CP, CV, PRESS, SBC, VALIDATE

Model Selection Options

Options in the MODEL statement of PROC GLMSELECT support many model selection techniques and criteria.

- SELECTION=<option>
- CHOOSE=<option>
- SELECT=<option>
- STOP=<option>



Specifies criterion on how variables enter (or exit) model. Options are: ADJRSQ, AIC, AICC, BIC, CP, CV, PRESS, RSQUARE, SBC, SL, and VALIDATE. Default is SBC.

Model Selection Options

Options in the MODEL statement of PROC GLMSELECT support many model selection techniques and criteria.

- SELECTION=<option>
- CHOOSE=<option>
- SELECT=<option>
- STOP=<option>

Specifies when to stop the selection process. If this is NOT specified but SELECT is, then the criterion in SELECT will be used. Options are: ADJRSQ, AIC, AICC, BIC, CP, CV, PRESS, SBC, SL, and VALIDATE. If neither STOP nor SELECT is specified, then SBC is the default.

Predictive Model Building

```
ods graphics on;  
proc glmselect data=ameshousing3_train plots=all;  
  FORWARD: model SalePrice=&interval/selection=forward  
  details=steps select=SL slentry=0.05;  
  title "Forward Model Selection for SalePrice-SL 0.05";  
  store out=glmameshousing3forward;  
run;
```

Predictive Model Building

```
ods graphics on;  
proc glmselect data=ameshousing3_train plots=all;  
  BACKWARD: model SalePrice=&interval/selection=backward  
  details=steps select=SL slstay=0.05;  
    title "Backward Model Selection for SalePrice-SL 0.05";  
    store out=glmameshousing3back;  
run;
```

Predictive Model Building

```
ods graphics on;  
proc glmselect data=ameshousing3_train plots=all;  
  STEPWISE: model SalePrice=&interval/selection=stepwise  
  details=steps select=SL slstay=0.05 slentry=0.05;  
  title "Stepwise Model Selection for SalePrice-SL 0.05";  
  store out=glmameshousing3step;  
run;
```

Are p -values and Parameter Estimates Correct?

Automated model selection results in the following:

- biases in parameter estimates, predictions, and standard errors
- incorrect calculation of degrees of freedom
- p -values that tend to err on the side of overestimating significance (increasing Type I Error probability)

Conservative Significance Levels

Sample Size				
Evidence	30	50	100	1000
Weak	.076	.053	.032	.009
Fair	.028	.019	.010	.003
Strong	.005	.003	.001	.0003
Very Strong	.001	.0005	.0001	.00004

4.01 Poll – Top Hat

The STEPWISE, BACKWARD, and FORWARD strategies result in the same final model if the same significance levels are used in all three.

- ☐ True
- ☐ False



4.2 Information Criterion and Other Selection Options

Information Criteria

In SAS, the below calculations begin with $n\log(SSE/n)$ and then place a penalty on this quantity

- Akaike's information criterion (AIC)..penalty is $2p + n + 2$
- Corrected Akaike's information criterion (AICC)...penalty is $[n(n+p)/(n-p-2)]$
- Sawa Bayesian information criterion (BIC)....penalty is $2(p+2)q-2q^2$
- Schwarz Bayesian information criterion (SBC).... $p\log(n)$

Smaller is better.

Where $q = \frac{n\hat{\sigma}}{SSE}$

Adjusted R Square / Mallows' Cp

- Adjusted R Square allows proper comparison between models with different parameter counts.

$$R_{ADJ}^2 = 1 - \frac{(n-i)(1-R^2)}{n-p}$$

- Mallows' C_p is a simple indicator of effective variable selection within a model.

Where $i=1$ if there is an intercept and 0 otherwise

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 .

SAS Code

```
proc glmselect data=ameshousing3_train plots=all;  
  STEPWISEAIC: model  
  SalePrice=&interval/selection=stepwise  
  details=steps select=SBC;  
  title "Stepwise model for SalePrice - AIC";  
  store out=glmameshousing3AIC;  
run;
```



4.3 All Possible Selection

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.”

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.

RSQUARE, ADJRSQ, CP Selection Options

Variables in
Full Model (k)

Total Number of
Subset Models (2^k)

0

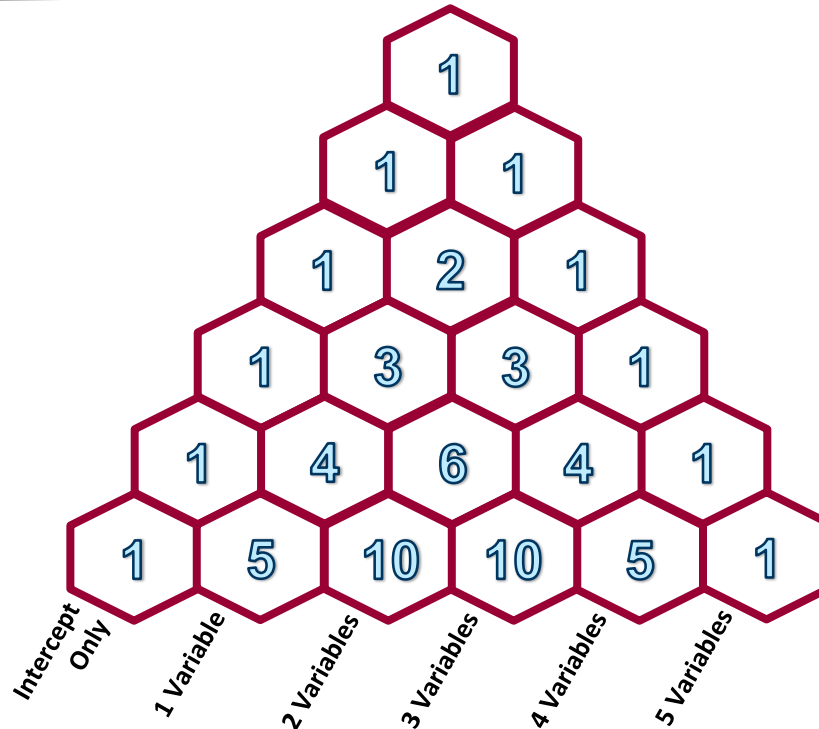
1

2

3

4

5



1

2

4

8

16

32

Hocking's Criterion versus Mallows' C_p

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

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

All-Possible Model Selection – R^2 , Adj R^2 , C_p

```
proc reg data=ameshousing3_train plots=all;  
  ALLPOSS: model SalePrice=&interval/selection=rsquare adjrsq cp;  
  title "All possible model selection for SalePrice";  
run;  
quit;
```

All-Regression Model Selection – C_p

```
proc reg data=ameshousing3_train plots=cp;  
  ALLPOSS: model SalePrice=&interval/selection=cp best=1;  
  title "All possible model selection for SalePrice";  
run;  
quit;
```

4.02 Multiple Choice Poll – Top Hat

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