

Selecting the Best Regression Model

Applied Regression and Other Multivariable Methods
Sections 16-1 - 16-9

17

Overview

- Have response Y and predictors X_1, X_2, \dots, X_k
 - Goal: to determine the “best” subset of these predictors
- Reliability Model provides the best prediction for some new observation or set of observations. What variables end up in the model is of little consequence.
- Validity Model allows accurate quantification of the relationship between Y and certain X 's after controlling for other variables.
- With reliability, want an accurate yet parsimonious model. If model has too many variables, you can run into collinearity problems. Can also “overfit” the Y 's so model does not fit other data sets.
 - With validity, focus is on specific regression coefficients. Other variables in the model are for control purposes only. Again want “small” model to avoid potential collinearity.

17-1

Best in Terms of Reliability

- Five step process in determination
 1. Select maximum model
 2. Choose selection criterion
 3. Choose selection strategy
 4. Run strategy/analysis
 5. Evaluate reliability
1. Select Maximum Model
- The **maximum model** is the largest model to be considered at any point. All other models are created by deleting predictors from this maximum model.
- The **maximum model** should contain any possible transformations of the predictors (e.g., polynomial terms like X_j^2 or interactions) that would be considered. May need to center variables to avoid collinearity problems.
- Examples:** Given variables X_1 and X_2 , the maximum model could be
- $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$
- $$Y = \beta_0 + \beta_1 (X_1 - \bar{X}_1) + \beta_2 (X_2 - \bar{X}_2) + \beta_3 (X_1 - \bar{X}_1)^2 + \beta_4 (X_2 - \bar{X}_2)^2 + \beta_5 (X_1 - \bar{X}_1)(X_2 - \bar{X}_2)$$
- $$Y = \beta_0 + \beta_1 X_1 + \beta_2 \log(X_2) + \beta_3 X_1 \log(X_2)$$

17-2

1. Select Maximum Model

The total number of parameters, k , is limited by the size of data set used in the study

 - Since df error equals $n - k - 1$, need $n > k + 1$
 - Want decent sized df error so general rules are

$$n - k - 1 \geq 10$$

$$n \geq 5k$$

The **correct model** involves an unknown subset of $p \leq k$ predictors in the maximum model. The multiple squared correlation coefficient of this correct model, $\rho^2(Y|X_1, X_2, \dots, X_p)$, is no smaller than any other model with more predictors. Goal can then be defined as finding this correct model.
2. The selection criterion

Will consider several numeric summaries to compare models. All are closely related.

 - a. **The coefficient of determination**

$$R^2(Y|X_1, \dots, X_p) = 1 - (\text{SSE}(p)/\text{SSY})$$

Problem: Will always go up. Largest for maximum model
 - b. **Adjusted R^2**

$$R_a^2(Y|X_1, \dots, X_p) = 1 - \left(\frac{\text{SSE}(p)}{\text{SSY}} \times \frac{n-1}{n-p-1} \right)$$

Penalizes R^2 for the complexity of the model

17-3

2. The selection criterion

c. The mean square error

$$MSE(p) = \frac{SSE(p)}{n - p - 1}$$

SSE always goes down but MSE penalizes for complexity

d. The partial F test

$$F = \frac{\frac{\text{Model SS(Large)} - \text{Model SS(Small)}}{\text{Model df(Large)} - \text{Model df(Small)}}}{\text{Error SS(Large)}/\text{Error df(Large)}}$$

Can consider both variables-added-last (Type III) and variables-added-in-order tests (Type I).

e. Mallows' Cp

$$C_p = \frac{SSE(p)}{MSE(k)} - [n - 2(p + 1)]$$

Helps decide how many variables to have in the model.

Will approximately equal p + 1 when MSE(p) = MSE(k)

If key variables missing, Cp should be larger than p + 1. A small Cp is preferable. Lower bound of Cp is 2p - k + 1.

3. Searching strategy

Ideally want to look at all possible models. This is the only way to guarantee you find the "best" model. This is often infeasible since the number of models is 2^k - 1. The following strategies are commonly used.

a. Backwards Elimination

Fit the maximum model and determine partial F tests for all variables as if added last (Type III). Remove variable with highest P-value and refit the model. Continue to remove variables until all remaining variables are "significant".

b. Forward Selection

Start with most highly correlated variable and compute the partial F statistic for each variable as if it were added next. Select variable with highest F value (lowest P-value, highest partial r^2). Continue to add variables until remaining variables when added next are not "significant".

c. Stepwise Regression

This is a hybrid of the two procedures. Sometimes as additional variables are added, early fitted variables become unimportant. To adjust for this, after each step of the forward selection, the entire model is backward analyzed and variables are removed if not significant. Then an additional variable is added and the same backward process is implemented. This is repeated until nothing can be added or removed.

4. The Analysis

Ideally, a residual analysis should be done for each model investigated. Strategies of selection depend on the model assumptions of constant variance, normality, and independence. Very important to check the residuals of final model.

5. Evaluating Reliability

Model chosen which best describes the data at hand. Using this same data set to assess reliability will overstate the accuracy of the model. Really interested in how this model generalizes to other data sets. Possible approaches around this are

a. Follow-up Study

Sometimes another data set can be made available and the model can be used to predict the response variable Y.

b. Split-Sample Analysis

This is similar to the jackknife residual idea. Randomly split the data set up into a training group to develop the model and a holdout group which will be used to assess the reliability of the model.

Sometimes have important strata/groups and will want to make sure each group is represented in the training and holdout groups. Can randomly assign subjects from each of these strata to the two groups (stratified sampling)

5. Evaluating Reliability

How big should the training sample and holdout sample be? The proportion should really be tailored to the problem at hand. If the dataset is quite large, a training sample of 200-500 observations may be at most 25% of the sample. If the dataset is quite small, this sort of sample may constitute up to 75% of the data. The key is to have a training sample that is representative of the population.

Let 1 represent the data in the training sample group and 2 be the data in the holdout group. In fitting the data, you obtain squared multiple correlation coefficient.

$$R^2(1) = R^2(Y_1|X_1, X_2, \dots, X_p) = r^2(Y_1, \hat{Y}_1)$$

Using the model to predict the holdout data set, Y2*, you obtain

$$R_*^2(2) = r^2(Y_2, \hat{Y}_2^*)$$

This is known as the cross-validation squared correlation. The difference between these two values R^2(1) - R_*^2(2) is known as the shrinkage on cross validation. This shrinkage is almost always positive. There are no hard rules as to what the shrinkage should be. Can usually consider a shrinkage of less than 20% to be pretty good.

Example

Consider the SBP example of Chapter 5. There are three independent variables but we will also consider all possible interactions. To avoid potential interactions, all variables will be centered. This results in some different model results than what is shown in the text.

The maximum model is

$$Y = \beta_0 + \beta_1 \text{AGE}^* + \beta_2 \text{QUET}^* + \beta_3 \text{SMK}^* + \beta_4 \text{AGE}^* \times \text{QUET}^* + \beta_5 \text{AGE}^* \times \text{SMK}^* + \beta_6 \text{QUET}^* \times \text{SMK}^* + \beta_7 \text{AGE}^* \times \text{SMK}^* \times \text{QUET}^*$$

We will use the selection=, cp, adjrsq, and mse options to get results like those in the book AND to perform forward, backward, and stepwise regression.

NOTE: It is common to include higher order interactions only when lower order terms are already in the model. While SAS will consider $2^7 - 1 = 127$ models, we will consider only a subset of them. You can use the include option to keep certain variables in all models considered.

Variable	Correlation			
	a	q	s	aq
a	1.0000	0.8028	-0.1395	0.1658
q	0.8028	1.0000	-0.0714	0.1379
s	-0.1395	-0.0714	1.0000	0.0347
aq	0.1658	0.1379	0.0347	1.0000
as	-0.1699	0.0345	0.0175	0.0169
qs	0.0353	0.2962	0.0091	-0.0038
aqs	-0.0755	-0.0484	0.6338	0.2740
sbp	0.7752	0.7420	0.2473	0.3183

Variable	Correlation			
	as	qs	aqs	sbp
a	-0.1699	0.0353	-0.0755	0.7752
q	0.0345	0.2962	-0.0484	0.7420
s	0.0175	0.0091	0.6338	0.2473
aq	0.0169	-0.0038	0.2740	0.3183
as	1.0000	0.8046	0.1441	-0.0321
qs	0.8046	1.0000	0.1194	0.1338
aqs	0.1441	0.1194	1.0000	0.2488
sbp	-0.0321	0.1338	0.2488	1.0000

There do not seem to be any troubling pairwise correlations among these centered variables and interactions.

```
data problem81;
  infile 'i:\www\datasets502\EX0502.DAT' firstobs=2 dlm='09'x;
  input person sbp quet age smk;
  aq=age*quet; as=age*smk; qs=quet*smk; aqs=age*quet*smk;

data centered;
  set problem81;
  q = quet - 3.4410938 ; a = age - 53.25 ;
  s = smk - 0.53125 ; aq=a*q;
  as = a*s; qs=q*s; aqs=a*q*s;

/* Results like those in the book */
/* Too much collinearity among these variables */
proc reg corr data=problem81;
  model sbp = age quet smk aq as qs aqs / selection=rsquare cp mse;

proc reg corr data=centered;
  model sbp = a q s aq as qs aqs / selection=rsquare cp mse adjrsq;
  model sbp = a q s aq as qs aqs / selection=forward;
  model sbp = a q s aq as qs aqs / selection=backward;
  model sbp = a q s aq as qs aqs / selection=stepwise;
run;
quit;
```

R-Square Selection Method

Number in Model	R-Square	Adjusted R-Square	C(p)	MSE	Variables
1	0.6009	0.5876	18.7539	85.47795	a
1	0.5506	0.5356	24.6554	96.26743	q
1	0.0612	0.0299	81.9934	201.09569	s

2	0.7298	0.7112	5.6565	59.87188	a s
2	0.6412	0.6165	16.0324	79.49574	a q
2	0.6412	0.6165	16.0365	79.50353	q s

3	0.7609	0.7353	4.0075	54.86225	a q s
3	0.7405	0.7127	6.4033	59.55518	a s as
3	0.6776	0.6431	13.7682	73.98197	a q aq
3	0.6511	0.6137	16.8744	80.06647	q s qs

4	0.7901	0.7590	2.5924	49.95688	a q s aq
4	0.7639	0.7289	5.6625	56.19343	a q s as
4	0.7616	0.7262	5.9357	56.74834	a q s qs

5	0.7922	0.7523	4.3428	51.35172	a q s aq as
5	0.7908	0.7506	4.5061	51.69621	a q s aq qs
5	0.7650	0.7198	7.5370	58.08994	a q s as qs

6	0.7925	0.7427	6.3063	53.32556	a q s aq as qs

7	0.7952	0.7354	8.0000	54.84757	a q s aq as qs aqs

Adjusted R^2 , C_p , and MSE all tend to favor the model with variables a, q, s, and aq.

Forward Selection: Step 1

Variable a Entered: R-Square = 0.6009 and C(p) = 18.7539

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3861.63038	3861.63038	45.18	<.0001
Error	30	2564.33838	85.47795		
Corrected Total	31	6425.96875			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	144.53125	1.63438	668457	7820.23	<.0001
a	1.60450	0.23872	3861.63038	45.18	<.0001

Bounds on condition number: 1, 1

Forward Selection: Step 2

Variable s Entered: R-Square = 0.7298 and C(p) = 5.6565

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4689.68423	2344.84211	39.16	<.0001
Error	29	1736.28452	59.87188		
Corrected Total	31	6425.96875			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	144.53125	1.36784	668457	11164.8	<.0001
a	1.70916	0.20176	4296.58607	71.76	<.0001
s	10.29439	2.76811	828.05385	13.83	0.0009

Bounds on condition number: 1.0198, 4.0794

Forward Selection: Step 3

Variable q Entered: R-Square = 0.7609 and C(p) = 4.0075

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4889.82570	1629.94190	29.71	<.0001
Error	28	1536.14305	54.86225		
Corrected Total	31	6425.96875			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	144.53125	1.30937	668457	12184.3	<.0001
a	1.21271	0.32382	769.45920	14.03	0.0008
q	8.59245	4.49868	200.14147	3.65	0.0664
s	9.94557	2.65606	769.23345	14.02	0.0008

Bounds on condition number: 2.867, 20.152

Forward Selection: Step 4

Variable aq Entered: R-Square = 0.7901 and C(p) = 2.5924

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	5077.13311	1269.28328	25.41	<.0001
Error	27	1348.83564	49.95688		
Corrected Total	31	6425.96875			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	142.54885	1.61533	389044	7787.60	<.0001
a	1.15221	0.31058	687.56966	13.76	0.0009
q	8.55841	4.29289	198.55559	3.97	0.0564
s	9.65649	2.53893	722.65945	14.47	0.0007
aq	0.73724	0.38074	187.30741	3.75	0.0634

Bounds on condition number: 2.8963, 31.128

No other variable met the 0.5000 significance level for entry into the model.

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	a	1	0.6009	0.6009	18.7539	45.18	<.0001
2	s	2	0.1289	0.7298	5.6565	13.83	0.0009
3	q	3	0.0311	0.7609	4.0075	3.65	0.0664
4	aq	4	0.0291	0.7901	2.5924	3.75	0.0634

This procedure ended up with the same model that was decided on by looking at all the models. The alpha level of 0.50 is the default for forward selection. This is chosen to make sure important variables are not included. To change this level, the slentry= option can be used.

Now we do backward selection

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.7952 and C(p) = 8.0000

Variable	Estimate	Error	Type II SS	F Value	Pr > F
Intercept	142.87015	1.74528	367546	6701.23	<.0001
a	1.18626	0.34978	630.86115	11.50	0.0024
q	8.55567	5.12851	152.64523	2.78	0.1083
s	8.43264	3.51479	315.70836	5.76	0.0246
aq	0.64309	0.42490	125.63711	2.29	0.1432
as	0.28967	0.69724	9.46675	0.17	0.6815
qs	-2.41874	10.53141	2.89309	0.05	0.8203
aqs	0.48106	0.86928	16.79731	0.31	0.5851

Bounds on condition number: 3.6974, 132.68

Backward Elimination: Step 1

Variable qs Removed: R-Square = 0.7947 and C(p) = 6.0527

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	142.82340	1.70021	372370	7056.55	<.0001
a	1.19786	0.33950	656.93174	12.45	0.0016
q	8.09586	4.63127	161.25275	3.06	0.0927
s	8.46257	3.44519	318.39161	6.03	0.0213
aq	0.65362	0.41434	131.31270	2.49	0.1273
as	0.16085	0.40626	8.27238	0.16	0.6955
aqs	0.46703	0.85054	15.91015	0.30	0.5878

Bounds on condition number: 3.2763, 74.729

Backward Elimination: Step 2

Variable as Removed: R-Square = 0.7934 and C(p) = 4.2036

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	142.75075	1.66265	376375	7371.53	<.0001
a	1.15208	0.31398	687.40652	13.46	0.0011
q	8.65036	4.34227	202.62721	3.97	0.0570
s	8.26280	3.35232	310.18987	6.08	0.0206
aq	0.65076	0.40751	130.20588	2.55	0.1224
aqs	0.53092	0.82144	21.32858	0.42	0.5237

Bounds on condition number: 2.8963, 52.555

Backward Elimination: Step 3
 Variable aqs Removed: R-Square = 0.7901 and C(p) = 2.5924

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	142.54885	1.61533	389044	7787.60	<.0001
a	1.15221	0.31058	687.56966	13.76	0.0009
q	8.55841	4.29289	198.55559	3.97	0.0564
s	9.65649	2.53893	722.65945	14.47	0.0007
aq	0.73724	0.38074	187.30741	3.75	0.0634

Bounds on condition number: 2.8963, 31.128

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	qs	6	0.0005	0.7947	6.0527	0.05	0.8203
2	as	5	0.0013	0.7934	4.2036	0.16	0.6955
3	aqs	4	0.0033	0.7901	2.5924	0.42	0.5237

This analysis also ended up with the same model. The default of 0.10 can be changed using the slentry= option.

Stepwise Selection: Step 1
 Variable a Entered: R-Square = 0.6009 and C(p) = 18.7539

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	144.53125	1.63438	668457	7820.23	<.0001
a	1.60450	0.23872	3861.63038	45.18	<.0001

Bounds on condition number: 1, 1

Stepwise Selection: Step 2
 Variable s Entered: R-Square = 0.7298 and C(p) = 5.6565

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	144.53125	1.36784	668457	11164.8	<.0001
a	1.70916	0.20176	4296.58607	71.76	<.0001
s	10.29439	2.76811	828.05385	13.83	0.0009

Bounds on condition number: 1.0198, 4.0794

Stepwise Selection: Step 3
 Variable q Entered: R-Square = 0.7609 and C(p) = 4.0075

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	144.53125	1.30937	668457	12184.3	<.0001
a	1.21271	0.32382	769.45920	14.03	0.0008
q	8.59245	4.49868	200.14147	3.65	0.0664
s	9.94557	2.65606	769.23345	14.02	0.0008

Bounds on condition number: 2.867, 20.152

Stepwise Selection: Step 4
 Variable aq Entered: R-Square = 0.7901 and C(p) = 2.5924

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	142.54885	1.61533	389044	7787.60	<.0001
a	1.15221	0.31058	687.56966	13.76	0.0009
q	8.55841	4.29289	198.55559	3.97	0.0564
s	9.65649	2.53893	722.65945	14.47	0.0007
aq	0.73724	0.38074	187.30741	3.75	0.0634

Bounds on condition number: 2.8963, 31.128

All variables left in the model are significant at the 0.1500 level.
 No other variable met the 0.1500 significance level for entry into the model.

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	a		1	0.6009	0.6009	18.7539	45.18	<.0001
2	s		2	0.1289	0.7298	5.6565	13.83	0.0009
3	q		3	0.0311	0.7609	4.0075	3.65	0.0664
4	aq		4	0.0291	0.7901	2.5924	3.75	0.0634

No variables are removed so this is the same as forward selection