Example: Model Building Procedures in SAS

SENIC data set

These data were obtained as part of the Study on the Efficacy of Nosocomial Infection Control (SENIC) to determine whether infection surveillance and control programs have reduced the rates of nosocomial (hospital-acquired) infection in US hospitals.

This data set consists of a random sample of n=113 hospitals selected from the original N=338 hospitals surveyed. Each hospital is given an ID number, and is measured on 11 other variables. Definitions for the variables are given in the table below.

We will use model selection procedures to construct a model that estimates Infection Risk.

Variable	Description
idno	Identification number for hospital (1 – 113)
Stay	Average length of stay of all patients in the hospital (measured in days)
Age	Average age of patients (in years)
InfRisk	Infection Risk: Average estimated probability of acquiring infection in hospital (in percent)
CulRatio	Ratio of number of cultures performed to number of patients without signs or symptoms of hospital-acquired infection, times 100
XRay	Ratio of number of X-rays performed to number of patients without signs or symptoms of pneumonia, times 100
NumBeds	Average number of beds in hospital during study period
MedSch	Medical School Affiliation (1=Yes, 2=No)
Region	Geographic Region (1=Northeast, 2=North Central, 3=South, 4=West)
Census	Average number of patients in hospital per day during study period
Nurses	Average number of full-time equivalent registered and licensed practical nurses during study period (number of full-time + ½ number of part-time)
Services	Percent of 35 potential facilities and services that are provided by the hospital

An initial inspection of the variable descriptions reveals two qualitative variables: MedSchool and Region. Although their values are numeric, the numbers are simply codes for the categories. We should not treat these as numeric variables.

In SAS, the model selection procedures are implemented in Proc Reg, and this accepts only numeric variables. Therefore, MedSchool and Region can not be part of the automated model selection process unless we treat them as numeric. We don't want to ignore these variables, so we convert them to indicator variables and use the indicators in the model selection procedure. This approach is straightforward for MedSchool since there is only one indicator variable, but we must be careful with Region. There are three indicator variables for Region, and they must all be included or excluded in the the model. The automated procedure in SAS cannot enforce this restriction, so we may need to adjust the model manually.

To begin the analysis, we examine the correlation matrix. We are interested in the correlation between each potential predictor variable and the response variable (InfRisk). We are also interested in the correlation between potential predictors. For this part of the analysis, we omit the indicator variables.

Pearson Correlation Coefficients, N = 113									
	Stay	**Age**	**InfRisk**	**CulRatio**	**XRay**	**NumBeds**	**Census**	**Nurses**	**Services**
Stay	1.00000	0.18891	0.53344	0.32668	0.38248	0.40927	0.47389	0.34037	0.35554
Age	0.18891	1.00000	0.00109	-0.22585	-0.01885	-0.05882	-0.05477	-0.08294	-0.04045
InfRisk	0.53344	0.00109	1.00000	0.55916	0.45339	0.35977	0.38141	0.39398	0.41260
CulRatio	0.32668	-0.22585	0.55916	1.00000	0.42496	0.13972	0.14295	0.19890	0.18513
XRay	0.38248	-0.01885	0.45339	0.42496	1.00000	0.04582	0.06291	0.07738	0.11193
NumBeds	0.40927	-0.05882	0.35977	0.13972	0.04582	1.00000	0.98100	0.91550	0.79452
Census	0.47389	-0.05477	0.38141	0.14295	0.06291	0.98100	1.00000	0.90790	0.77806
Nurses	0.34037	-0.08294	0.39398	0.19890	0.07738	0.91550	0.90790	1.00000	0.78351
Services	0.35554	-0.04045	0.41260	0.18513	0.11193	0.79452	0.77806	0.78351	1.00000

Correlations between predictors and the response are shown in the third column. Except for Age, all of the correlations are between 0.3 and 0.6, so any of these variables could be effective predictors.

As an inital assessment of possible multicollinearity issues, we also examine the correlation between predictors. In general, correlations higher than 0.7 are reason for concern. We see several high

correlations, but the largest is 0.981 (between Census and NumBeds). Such a high correlation indicates that whatever information that is contained in Census is (almost) duplicated in NumBeds. Including both of these variables in the same model would create extreme multicollinearity problems, so we should use only one of these variables. Either one could be used, but we choose Census because its correlation with the response is slightly higher (0.381 vs. 0.360).

(Note: You could fit a model to all of the variables in the data set and use the variance inflation factors to assess multicollinearity issues. I did this, but I have not included the results in this report. The variance inflation factors for both Census and NumBeds were much larger than 10, indicating multicollinearity problems do exist when both of these variables are used.)

Fullest Possible Model

We are now ready to begin the model selection process. Our response is Infection Risk (InfRisk), and we consider the potential predictors Stay, Age, CulRatio, XRay, Census, Nurses, Services, and the indicator variables for Region and MedSchool. Using all of these variables, the R^2 is 0.5783 and adjusted R^2 is 0.5324. While this does not seem extraordinarily high, it is definitely reasonable given the fact that we are trying to model something as complicated as risk of infection. The parameter estimates for this model are shown in the table below. Note that Census has the largest variance inflation factor, but it is less than our customary cutoff of 10. In addition, the standard errors all look reasonable (none are extremely large), so we do not consider multicollinearity a serious issue. *At this stage of the analysis, we are not concerned with the actual parameter estimates or the results of their t-tests.* I am showing you this table solely for the variance inflation factors.

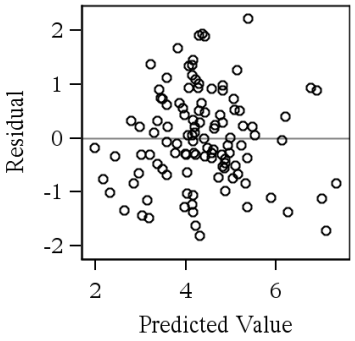
Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-0.67911	1.22155	-0.56	0.5795	0
Stay	1	0.25954	0.06918	3.75	0.0003	2.32921
Age	1	0.01069	0.02176	0.49	0.6242	1.25541
CulRatio	1	0.05293	0.01052	5.03	<.0001	1.54370
XRay	1	0.01161	0.00528	2.20	0.0303	1.39369
Census	1	-0.00005911	0.00169	-0.03	0.9722	9.03237
Nurses	1	0.00138	0.00167	0.83	0.4104	7.21885
Services	1	0.01756	0.00982	1.79	0.0769	2.97044
MedSch	1	-0.59913	0.31899	-1.88	0.0632	1.74784
Reg1	1	-1.06902	0.33465	-3.19	0.0019	2.80546
Reg2	1	-0.71900	0.29883	-2.41	0.0179	2.43620
Reg3	1	-0.76304	0.28955	-2.64	0.0097	2.48150

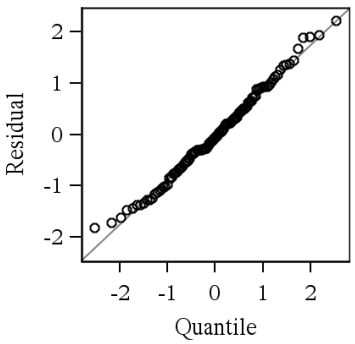
Next we generate alternatives to this model in an effort to get a simpler model that also fits the data well.

Foward Selection

The summary table for Forward Selection is shown below. Variables are listed in the order they were added to the model. Note that all three indicator variables for Region were added to the model. The model generated by this procedure contains 7 predictors (CulRatio, Stay, Services, Xray, Nurses, Region and MedSch). The (unadjusted) R^2 for this model is 0.5773.

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	CulRatio	1	0.3127	0.3127	55.6287	50.49	<.0001
2	Stay	2	0.1377	0.4504	24.6370	27.57	<.0001
3	Services	3	0.0430	0.4934	16.3384	9.25	0.0029
4	XRay	4	0.0227	0.5161	12.8940	5.07	0.0263
5	Reg1	5	0.0107	0.5268	12.3354	2.42	0.1231
6	MedSch	6	0.0097	0.5365	12.0206	2.21	0.1401
7	Nurses	7	0.0063	0.5428	12.5112	1.45	0.2317
8	Reg3	8	0.0077	0.5505	12.6567	1.79	0.1837
9	Reg2	9	0.0268	0.5773	8.2493	6.52	0.0121





The diagnostic plots for this model give no indication that the assumptions are violated.

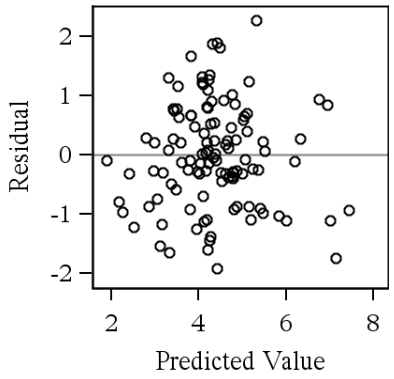
In summary, this is a reasonable model.

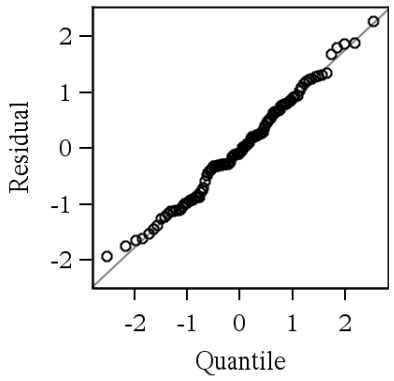
Backward Elimination

The summary table for Backward Elimination is shown below. This table is much shorter than the previous one, but it is giving the variables that are *removed* from the model. We must compare this list to the original list of potential predictors (Stay, Age, CulRatio, XRay, Census, Nurses, Services, and the indicator variables for Region and MedSchool) to determine which variables remain in the model.

This model contains 6 predictors (Stay, CulRatio, XRay, Services, Region and MedSchool), with $R^2 = 0.5712$. SAS reports that there are 8 predictor variables in the model because it is counting the indicator variables.

Summary of Backward Elimination							
Step	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	Census	10	0.0000	0.5783	10.0012	0.00	0.9722
2	Age	9	0.0010	0.5773	8.2493	0.25	0.6178
3	Nurses	8	0.0061	0.5712	7.7015	1.48	0.2270





The diagnostic plots do not suggest any violations of the model assumptions.

In summary, this is another reasonable model.

Stepwise

The summary for Stepwise is given in the table below. Note that, once entered, no variables were removed from the model. For a different dataset, it is possible for variables to enter, then exit and later re-enter the model. Among the three indicator variables for Region, this procedure has included only Reg1 in its final model. We either include or exclude all the indicators for a cateorical variable, so we choose to include all three of the indicators for Region. With this modification, the Stepwise model is the same as the Backward Elmination 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	CulRatio		1	0.3127	0.3127	55.6287	50.49	<.0001
2	Stay		2	0.1377	0.4504	24.6370	27.57	<.0001
3	Services		3	0.0430	0.4934	16.3384	9.25	0.0029
4	XRay		4	0.0227	0.5161	12.8940	5.07	0.0263
5	Reg1		5	0.0107	0.5268	12.3354	2.42	0.1231
6	MedSch		6	0.0097	0.5365	12.0206	2.21	0.1401

All three of the preceding methods use F tests to compare models and make decisions regarding which variables to include or exclude. The next set of methods use different model comparison criteria to make this decision.

Criterion: Adjusted R2

This procedure generates numerous models and ranks them according to adjusted R^2. The SAS output, shown in the table below, lists the models in decreasing order of adjusted R^2. Thus the first model in the list is the "best" according to this criterion, and the second model listed is the second best, etc. The top model contains all the predictors except Age and Census, so this is the same as the model generated by Forward Selection.

Number in Model	Adjusted R-Square	R-Square	Variables in Model
9	0.5403	0.5773	Stay CulRatio XRay Nurses Services MedSch Reg1 Reg2 Reg3
8	0.5382	0.5712	Stay CulRatio XRay Services MedSch Reg1 Reg2 Reg3
9	0.5371	0.5743	Stay CulRatio XRay Census Services MedSch Reg1 Reg2 Reg3
10	0.5370	0.5783	Stay Age CulRatio XRay Nurses Services MedSch Reg1 Reg2 Reg3
10	0.5359	0.5773	Stay CulRatio XRay Census Nurses Services MedSch Reg1 Reg2 Reg3
9	0.5346	0.5720	Stay Age CulRatio XRay Services MedSch Reg1 Reg2 Reg3

Criterion: Mallow's Cp

The output is similar to that of adjusted R2, but the ranking criterion is Mallow's Cp. The "best" model (in the top row) is the same as that generated by Backward Elimination and Stepwise.

Number in Model	C(p)	R-Square	Variables in Model
8	7.7015	0.5712	Stay CulRatio XRay Services MedSch Reg1 Reg2 Reg3
9	8.2493	0.5773	Stay CulRatio XRay Nurses Services MedSch Reg1 Reg2 Reg3
7	8.6301	0.5590	Stay CulRatio XRay Services Reg1 Reg2 Reg3
9	8.9610	0.5743	Stay CulRatio XRay Census Services MedSch Reg1 Reg2 Reg3
9	9.5227	0.5720	Stay Age CulRatio XRay Services MedSch Reg1 Reg2 Reg3
8	9.6892	0.5629	Stay CulRatio XRay Nurses MedSch Reg1 Reg2 Reg3

Summary

All of the model selection procedures generated similar models, each containing Stay, CulRatio, Xray, Services, MedSch and Region as useful predictors and excluded potential predictors Age and Census. Themodels differ with regard to the predictor Nurses. Both Forward Selection and adjusted R^2 judged Nurses to be a useful predictor, while Backward Elimination, Stepwise and Mallow's C_p did not.

At this stage of the analysis, we might want to re-consider some of our early choices. In particular, we decided to exclude NumBeds as a potential predictor since it is so highly correlated with Census. However, none of the model selection methods generated a model that contains Census. It is doubtful that the generated models will change if we re-performing the model selection process with NumBeds instead of Census. But if we want to be thorough we might consider this alternative.

Now that we have two candidate models, we can investigate their properties a little more closely. We should examine the diagnostic plots and parameter estimates for these two models and decide whether we should move forward with the model that contains Nurses or the one that does not. As part of this decision, we should consider how other researchers have modeled these quantities in the past, and how well we can justify the decision to either include or exclude Nurses. To facilitate a direct comparison between two or more models, we can tell SAS to save the summary information about each model and print it together as one table. (This table is very wide, so it looks like two tables in this document.)

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	Stay	CulRatio	XRay	Nurses	Services
1	WithNurses	PARMS	InfRisk	0.90912	-0.14941	0.26708	0.051606	0.011606	.001304167	0.017633
2	WithoutNurses	PARMS	InfRisk	0.91120	-0.32966	0.27469	0.052336	0.011326	.	0.025459

Obs	MedSch	Reg1	Reg2	Reg3	InfRisk	_IN_	_P_	_EDF_	_RSQ_	_ADJRSQ_	_AIC_	_SBC_
1	-0.61838	-1.07183	-0.74377	-0.77185	-1	9	10	103	0.57727	0.54034	-12.0045	15.2693
2	-0.50274	-1.10697	-0.76674	-0.75937	-1	8	9	104	0.57121	0.53823	-12.3954	12.1511

This table provides the parameters estimates as well as the model selection criteria. There are no drastic differences between these two models. AIC and SBC are slightly better (smaller) for the model without Nurses, but the adjusted R^2 is better (larger) for the model with Nurses. For my own personal opinion, I would choose to exclude Nurses. This is primarily because I prefer to use the smallest model that fits the data well. In other words, I choose to exclude a predictor unless there is a good reason to include it.

SAS Code

The full SAS code for this example is in the file ModelSelection.SENIC.sas.
The file also includes the dataset.

```
data senic;
input idno Stay Age InfRisk CulRatio XRay NumBeds MedSch
      Region Census Nurses Services;

* Region is 1,2,3, or 4 -- these are categories -- ;
* Proc Reg does not allow categorical variables, so create dummy variables;
* Also change 'MedSch' to 0 if it is a 2;
*   (ie, replace MedSchool with indicator variable);

if MedSch eq 2 then MedSch = 0;
if Region eq 1 then do; Reg1=1; Reg2=0; Reg3=0; output; end;
if Region eq 2 then do; Reg1=0; Reg2=1; Reg3=0; output; end;
if Region eq 3 then do; Reg1=0; Reg2=0; Reg3=1; output; end;
if Region eq 4 then do; Reg1=0; Reg2=0; Reg3=0; output; end;

datalines;
  1   7.13  55.7  4.1   9.0   39.6  279  2  4  207  241  60.0
  2   8.82  58.2  1.6   3.8   51.7   80  2  2   51   52  40.0
  3   8.34  56.9  2.7   8.1   74.0  107  2  3   82   54  20.0
   . . . more data lines here . . .
112  17.94  56.2  5.9  26.4   91.8  835  1  1  791  407  62.9
113   9.41  59.5  3.1  20.6   91.7   29  2  3   20   22  22.9
;

* always print the data the first time you read it into SAS;
proc print data=senic;run;

/* First do some preliminaries to get a 'feel' for the data */
* cross-classify the two categorical predictors;
proc tabulate data=senic;
 class MedSch Region;
 table MedSch*Region;
 run;

* correlations for numeric variables;
proc corr data=senic noprob nosimple;
  var Stay Age InfRisk CulRatio XRay NumBeds Census Nurses Services;
  run;
* Census and NumBeds are highly corr'd -- disregard NumBeds ;
```

```
ods graphics on;
proc reg data=senic;
 fullmodel: model InfRisk = Stay Age CulRatio XRay Census Nurses Services
                            MedSch Reg1 Reg2 Reg3
             / vif;

 forward:   model InfRisk = Stay Age CulRatio XRay Census Nurses Services
                            MedSch Reg1 Reg2 Reg3
              / selection = forward  details=summary;

 backward:  model InfRisk = Stay Age CulRatio XRay Census Nurses Services
                            MedSch Reg1 Reg2 Reg3
              / selection = backward details=summary;

 stepwise:  model InfRisk = Stay Age CulRatio XRay Census Nurses Services
                            MedSch Reg1 Reg2 Reg3
              / selection = stepwise details=summary;

 AdjR2:     model InfRisk = Stay Age CulRatio XRay Census Nurses Services
                            MedSch Reg1 Reg2 Reg3
              / selection = adjrsq details=summary best=6;

 MallowsCp: model InfRisk = Stay Age CulRatio XRay Census Nurses Services
                            MedSch Reg1 Reg2 Reg3
             / selection = cp  details=summary best=6;
 run;


proc reg data=senic outest=modelinfo;
 WithNurses:    model InfRisk = Stay CulRatio XRay Nurses Services
                                MedSch Reg1 Reg2 Reg3
                  / aic sbc adjrsq;
 WithoutNurses: model InfRisk = Stay CulRatio XRay Services
                                MedSch Reg1 Reg2 Reg3
                 / aic sbc adjrsq;
 run;
proc print data=modelinfo;  run;


ods graphics off;
```

Explanation of the Code

Proc Reg allows multiple model statements. By default, they will be labeled "Model 1", "Model 2", etc., in the output. We can provide a more descriptive name -- this goes before the word 'model' and must end with a colon.

All of these models are given the same set of potential predictors, the differences occur in the options (after the slash).

For the full model, the only option we specified is to generate variance inflation factors. For the remaining models the 'selection' option invokes one of the model building procedures.

By default, SAS prints a large amount of information about each step of a model building procedure. Unless you are interested in analyzing the procedure itself, I recommend that you use the 'details=summary' option, which greatly reduces the amount of printed output.

For the Adjusted R^2 and Mallow's Cp criteria, SAS will print every possible model (or at least a very large number of them), in order from "best" to "worst". We are usually only interested in the "best", so the "best=6" option prints only the 6 best models (according to the chosen criteria).

After we have examined the output from all the models in the first instance of Proc Reg, we narrow our choices of predictors to two models: one with and one without the predictor Nurses. The second instance of Proc Reg takes a closer look at these two models. The `outest` option creates a temporary dataset we name 'modelinfo' that will store information about the models that are in the two later `model` statements. The options `aic, sbc` and `adjrsq` at the end of each model statement will put these summaries in the temporary dataset. Finally, `proc print` prints this temporary dataset. This provides a condensed format by which we can compare these models.