Lecture 15

Model Building: Selection Criteria

STAT 512

Spring 2011

Background Reading

KNNL: Chapter 9

Topic Overview

- Selecting and Refining a Regression Model
- Model Selection Criteria / Statistics
- Automated Search Procedures
- CDI Case Study

Overview of Model Building

Strategy employs four phases:

1. Data Collection
2. Reduction of Explanatory/Predictor Variables
3. Model Refinement / Selection
4. Model Validation

Data Collection Phase

- Studies can be ***controlled experiments*** or ***observational studies.*** It is important to know the difference, and know which of the two you are working with.

- Studies have different types of variables, and it is again important to understand this prior to analysis.

Controlled Experiments

- In a ***controlled experiment***, levels of explanatory variables are controlled; a combination of these levels (called a ***treatment***) is assigned to each exp. unit.
- If completely controlled, can get a ***balanced design*** – and hence no intercorrelation among predictors (the ideal situation).
- Sometimes also have uncontrolled variables which are often called ***covariates***

Controlled Experiments (2)

- Usually a small number of variables involved
- Correct model is often “obvious” and selection of explanatory variables unneeded.
- May need to develop the correct functional form for the model (interactions, non-linear terms, etc.)

Observational Studies

- In an observational study all variables are "observed" rather than "controlled" and hence multicollinearity is much more likely.
- Study can be ***exploratory*** – we want to narrow a large group of explanatory variables to find those that are important.
- Study could also be ***confirmatory*** – have specific hypotheses going in and collect exactly the data needed to test them.

Observational Studies (2)

- Exploratory studies require the most in terms of model selection.
- Start with a large number of potential explanatory variables (some will be intercorrelated) and reduce to a few appropriate subsets for final consideration.
- Many approaches, statistics to help us do this – we will learn what these are and how to use them.

Questions in Model Selection

- How many variables to use? Smaller sets are more convenient, but larger sets may explain more of the variation in the response.
- Given a subset size, which variables should we choose for the model? Some statistics will help us compare them.

Model Selection Criteria

The following statistics are some of the most commonly used in model selection:

- R^2 / SSE
- Adjusted R^2 / MSE
- Mallow's C_p Criterion
- AIC / SBC
- PRESS Statistic

R_p^2 (SSE_p) Criterion

- Subscript *p* corresponds to the number of parameters in the model.

$$R_p^2 = R^2 = 1 - \frac{SSE_p}{SST}$$

- Goal is maximization. However, as we already know, R^2 continues to go up as variables are added – eventually extra variables are just going to get in the way.

R_p^2 (SSE_p) Criterion (2)

- Useful comparing models with same number of variables. Among these, the model with the highest R^2 could be considered "best".

- Maximizing R_p^2 is equivalent to choosing the (p–1)-variable model with the smallest SSE.

Adjusted R^2 (MSE) Criterion

- Penalizes the R^2 value based on the number of variables in the model:

$$R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SSTO}$$

- End up subtracting off more if p gets larger, so if this is not counterbalanced by enough decrease in SSE, R_a^2 can decrease as variables are added.

Adjusted R^2 (MSE) Criterion (2)

- Can also write as

$$R_a^2 = 1 - \frac{MSE}{SSTO/\left(n-1\right)} = 1 - \frac{MSE}{\text{Constant}}$$

- Maximizing the adjusted R^2, or equivalently minimizing the MSE, is one way to determine a best model.
- Advantage over regular R^2 since can compare models of different size.

Mallows' C_p Criterion

- Basic idea: Compare predictive ability of subset models to that of full model
- Full model generally best for prediction; but if multicollinearity then parameter estimates not useful.
- Subset of full model that doesn't have as much multicollinearity will be better as long as there is no substantial "bias" in predicted values relative to full model (i.e. close to same predictive ability.)

Mallows' C_p Criterion (2)

- C_p considers ratio of SSE for $p - 1$ variable model to MSE for full model; then penalizes for the number of variables:

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

- A model is considered "good" if $C_p \leq p$. May consider choosing the smallest model for which this is true (to reduce intercorrelation).

Mallows' C_p Criterion (3)

- Benefit to C_p is you can use it to select model size – getting a good model that contains as few variables as possible.
- Might also choose to pick the model with the minimum C_p, but it is more about C_p *relative to p*, and getting a smaller number of variables in the model while still having the same predictive ability.

AIC and SBC

- AIC is Akaike's Information Criterion

$$AIC_p = n\log\left(\frac{SSE_p}{n}\right) + 2p$$

- SBC is Schwarz' Bayesian Criterion

$$SBC_p = n\log\left(\frac{SSE_p}{n}\right) + p\log(n)$$

- Want to minimize these for "best model". They are the same except for their "penalty" terms.

*Remember, log refers to natural log in this class.

PRESS Statistic

- PRESS stands for Prediction Sum of Squares
- Measures how well fitted values predict the observed responses (SSE is a similar measure)
- Different from SSE in that, for each observation, model based on data *excluding that observation* is used for the prediction

$$PRESS_p = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_{i(i)} \right)^2$$

- $\hat{Y}_{i(i)}$ is prediction for i^{th} using all data except i^{th}

PRESS Statistic (2)

- Computation can be done without running n regression models; generally computer will do it for us, so not terribly important

- Models with a small PRESS statistic are considered good candidate models

Putting Things Together

- Consider goals of modeling (prediction? interpretation?)

- Look at multiple statistics, not just one. Generally they will all say similar things.

Physicians Case Study
(cdi_modelselect1.sas)

What is the best model for predicting the # of active physicians in a county? Possible predictors include:

1. X1 = Total Population
2. X2 = Total Personal Income
3. X3 = Land Area
4. X4 = Percent of Pop. Age 65 or older
5. X5 = Number of hospital beds
6. X6 = Total Serious Crimes
7. X7 = Percent of population HS Grads
8. X8 = Percent Unemployment

Initial Model

- We know there is multicollinearity.
- We take log of response, and remove LA and Chicago since these are special cases of their own. We expect our model to be valid only for small to large (not super-sized) counties.
- Adding a couple more variables for consideration.

Multicollinearity

	pop	inc	land	eld	bed	crm	hsg	unem
tot_pop		**0.97**	0.18	-0.02	**0.86**	**0.80**	0.04	-0.03
tot_income	**0.97**		0.10	-0.01	**0.81**	**0.72**	0.14	-0.09
land_area	0.18	0.10		0.01	0.03	0.10	-0.09	0.20
pop_eld	-0.02	-0.01	0.01		0.09	-0.03	-0.27	0.24
beds	**0.86**	**0.81**	0.03	0.09		**0.75**	-0.10	-0.02
crimes	**0.80**	**0.72**	0.10	-0.03	**0.75**		-0.09	0.03
hsgrad	0.04	0.14	-0.09	-0.27	-0.10	-0.09		-0.59
unemploy	-0.03	-0.09	0.20	0.24	-0.02	0.03	-0.59	

- Use only one of population, income variables. (checking one-var model indicates income is the better one)
- Perhaps some others have issues as well.

Full Model

- Everything in the model except total population which we've discarded.

Source	DF	SS	MS	F	P-value
Model	7	399	56.99	165.45	<.0001
Error	430	148	0.344		
Total	437	547			

R-Square 0.7292 Adj R-Sq 0.7248

Full Model (2)

Variable	DF	t Value	Pr > |t|	VIF
Intercept	1	7.20	<.0001	0
tot_income	1	7.62	<.0001	3.857
land_area	1	0.41	0.6856	1.075
pop_elderly	1	1.27	0.2030	1.129
beds	1	11.46	<.0001	4.106
crimes	1	-2.84	0.0047	2.602
hsgrad	1	4.10	<.0001	1.876
unemploy	1	-1.96	0.0511	1.650

- Seems multicollinearity may not be an issue at this point.

Subset Models

```
proc reg data=cdi outest=fits;
   model lphys = tot_income land_area pop_elderly beds
   crimes hsgrad unemploy
   /cli selection=rsquare adjrsq cp aic sbc b best=3;
    run;
```

- Asks for the 3 best models for each possible number of variables
- "Best" in terms of R^2.
- The 'b' option gives parameter estimates
- Adds other selection statistics to the output.

Output - (Note: See SAS for full output)

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	---Parameter Estimates-- Intercept	tot_income
1	0.6265	0.6256	159.2450	-329.9566	-321.79219	5.41290	.
1	0.6261	0.6252	159.8538	-329.5074	-321.34293	5.37448	0.00010515
1	0.3323	0.3308	626.4402	-75.5522	-67.38778	5.78045	.
2	0.6916	0.6902	57.8087	-411.8741	-399.62744	5.31641	0.00005798
2	0.6896	0.6882	60.9543	-409.0702	-396.82355	2.26671	.
2	0.6551	0.6535	115.7024	-362.9420	-350.69532	5.94910	.
3	0.7201	0.7181	14.6031	-452.2869	-435.95806	3.06757	0.00004268
3	0.7108	0.7088	29.3215	-438.0219	-421.69307	5.76518	0.00005402
3	0.7018	0.6997	43.6551	-424.5627	-408.23383	5.29441	0.00006453

# in Model	R-Square	land_area	pop_elderly	beds	crimes	hsgrad
1	0.6265	.	.	0.00053367	.	.
1	0.6261	.	.	.	.	.
1	0.3323	.	.	.	0.00001437	.
2	0.6916	.	.	0.00029505	.	.
2	0.6896	.	.	0.00055059	.	0.04026
2	0.6551	.	.	0.00053155	.	.
3	0.7201	.	.	0.00037024	.	0.02910
3	0.7108	.	.	0.00030961	.	.
3	0.7018	.	.	0.00034847	-0.00000396	.

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	SBC	---Parameter Estimates--- Intercept	tot_income
4	0.7261	0.7236	6.9248	-459.9268	-439.51573	3.19707	0.00004880
4	0.7222	0.7197	13.1139	-453.7379	-433.32682	3.72411	0.00004419
4	0.7215	0.7189	14.3462	-452.5160	-432.10488	2.81694	0.00004316
5	0.7281	0.7250	5.7704	-461.1151	-436.62182	3.81939	0.00005013
5	0.7268	0.7237	7.8354	-459.0253	-434.53200	3.01427	0.00004882
5	0.7262	0.7230	8.9160	-457.9357	-433.44241	3.19085	0.00004871
6	0.7291	0.7254	6.1642	-460.7476	-432.17202	3.64978	0.00005026
6	0.7282	0.7244	7.6256	-459.2620	-430.68647	3.81529	0.00004979
6	0.7268	0.7230	9.8266	-457.0342	-428.45867	3.00805	0.00004873
7	0.7292	0.7248	8.0000	-458.9147	-426.25699	3.64436	0.00004991

# in Model	R-Square	land_area	pop_elderly	beds	crimes	hsgrad
4	0.7261	.	.	0.00040709	-0.00000309	0.02720
4	0.7222	.	.	0.00036078	.	0.02306
4	0.7215	.	0.01099	0.00036644	.	0.03063
5	0.7281	.	.	0.00039746	-0.00000304	0.02149
5	0.7268	.	0.00773	0.00040253	-0.00000293	0.02837
5	0.7262	0.00000175	.	0.00040758	-0.00000310	0.02726
6	0.7291	.	0.00945	0.00039104	-0.00000284	0.02242
6	0.7282	0.00000716	.	0.00039913	-0.00000307	0.02153
6	0.7268	0.00000174	0.00773	0.00040301	-0.00000294	0.02843
7	0.7292	0.00000763	0.00951	0.00039278	-0.00000287	0.02248

Best Model?

- C_p suggests 5-6 variables
- Minimal AIC for 5-variable model excluding pop_eld and land_area.
- Maximum Adjusted R^2 for 6-variable model excluding land area.

Upcoming in Lecture 16...

- Model Building: Automated Procedures
- Continuing Physicians Dataset Analysis