Chapter 9 Model Selection and Validation

Adapted from Timothy Hanson

Department of Statistics, University of South Carolina

Stat 704: Data Analysis I

Model building overview (pp.343-349)

Chapter 9: Model variable selection and validation

Book outlines four steps in data analysis

1. Data collection and preparation (acquiring and "cleaning")
2. Reduction of explanatory variables (for exploratory observational studies). Mass screening for "decent" predictors.
3. Model refinement and selection.
4. Model validation.

We usually obtain data after step 1, though this step has received much more attention from statisticians in recent years.

9.1 Model building overview

- Book has flowchart for model building process on p. 344.
- *Designed experiments* are typically easy; experimenter *manipulates* treatment variables during experiment (and expects them to be significant); experimenter may adjust for other variables.
- With *confirmatory* observational studies, the goal is to determine whether (or how) the response is related to one or more pre-specified explanatory variables. No need to weed them out.
- *Exploratory* observational studies are done when we have little previous knowledge of exactly which predictors are related to the response. Need to "weed out" good from useless predictors.
- We may have a list of *potentially* useful predictors; *variable selection* can help us "screen out" useless ones and build a good, predictive model.

Controlled experiments

- These include clinical trials, laboratory experiments on monkeys and pigs, etc., community-based intervention trials, etc.
- The experimenters control one or more variables that are related to the response. Often these variables are "treatment" and "control." Can ascribe causality if populations are the same except for the control variables.
- Sometimes other variables (not experimentally assigned) that may also affect the response are collected too, e.g. gender, weight, blood chemistry levels, viral load, whether other family members smoke, etc.
- When building the model the treatment is always included. Other variables are included as needed to reduce variability and zoom in on the treatment factors. Some of these variables may be useful and some not, so part of the model building process is weeding out "noise" variables.

Confirmatory observational studies

- Used to test a hypothesis built from other studies or a "hunch."
- Variables involved in the hypothesis (amount of fiber in diet) that affect the response (cholesterol) are measured along with other variables that can affect the outcome (age, exercise, gender, race, etc.) – nothing is controlled. Variables involved in the hypothesis are called *primary* variables; the others are called risk factors; epidemiologists like to "adjust" for "risk factors."
- Note that your book discusses Vitamin E and cancer on p. 345. Such studies has received serious scrutiny recently.
- Usually all variables are retained in the analysis; they were chosen ahead of time.

Observational studies

- When people are involved, often not possible to conduct controlled experiments.
- Example: maternal smoking affects infant birthweight. One would have to randomly allocate the treatments "smoking" and "non-smoking" to pregnant moms – ethical problems.
- Investigators consider *anything* that is easy to measure that might be related to the response. Many, many variables are considered, and models painstakingly built. One of my professors called this "data dredging."

Observational studies

- There's a problem here – one is sure to find *something* if they look hard enough. Often "signals" are there spuriously, and sometimes *in the wrong direction*.
- The number of variables to consider can be large; there can be high multicollinearity. Keeping too many predictors can make prediction *worse*.
- "*The identification of "good"...variables to be included in the...regression model and the determination of appropriate functional and interaction relations...constitute some of the most difficult problems in regression analysis.*"

Section 9.2: GPA 2008 example

- First steps often involve plots:
  - Plots to indicate correct functional form of predictors and/or response.
  - Plots to indicate possible interaction.
  - Exploration of correlation among predictors (maybe).
  - Often a first-order model is a good starting point.
- Once a reasonable set of potential predictors is identified, formal model selection begins.
- If the number of predictors is large, say $k \geq 10$, we can use (automated) stepwise procedures to reduce the number of variables (and models) under consideration.

9.3 Model selection (pp. 353-361)

Once we reduce the set of potential predictors to a reasonable number, we can examine all possible models and choose the "best" according to some criterion.

Say we have k predictors $x_1, \ldots, x_k$ and we want to find a good subset of predictors that predict the data well. There are several useful criteria to help choose a subset of predictors.

Adjusted-R^2, R^2_a

"Regular" R^2 measures how well the model predicts the data that built it. It is possible to have a model with $R^2 = 1$ (predicts perfectly the data that built it), but has *lousy out-of-sample prediction*. The adjusted R^2, denoted R^2_a, provide a "fix" to R^2 to provide a measure of how well the model will predict data not used to build the model. For a candidate model with $p-1$ predictors

$$R^2_a = 1 - \frac{n-1}{n-p}\frac{SSE_p}{SSTO} \left(= 1 - \frac{MSE_p}{s^2_y} \right)$$

.

- Equivalent to choosing the model with the *smallest* MSE_p.
- If irrelevant variables are added, R^2_a may decrease unlike "regular" R^2 (R^2_a can be negative!).
- R^2_a penalizes model for being too complex.
- Problem: R^2_a is greater for a "full" model whenever the F-statistic for comparing full to reduced is greater than 1. We usually want F-statistics to be a lot biggger than 1 before adding in new predictors $\implies$ *too liberal*.

AIC

Choose model with smallest Akaike Information Criterion (AIC). For normal error model,

$$AIC = nlog(SSE_p) - nlog(n) + 2p.$$

- $n\log(SSE_p) - n\log(n) = C - 2\log\left\{\mathcal{L}\left(\hat{\beta}, \hat{\sigma}^2\right)\right\}$ from the normal model where C is a constant.
- $2p$ is "penalty" term for adding predictors.
- Like R_a^2, AIC favors models with small SSE, but penalizes models with too many variables p.

SBC (or BIC)

Models with smaller Schwarz Bayesian Criterion (SBC) are estimated to predict better. SBC is also known as *Bayesian Information Criterion*:

$$BIC = n\log(SSE_p) - n\log(n) + p\log(n).$$

- BIC is similar to AIC, but for $n \geq 8$, the BIC "penalty term" is more severe.
- Chooses model that "best predicts" the observed data according to asymptotic criteria.

Mallow's C_p

Let F be the full model with all k predictors and R be a reduced model with *p-1* predictors to be compared to the full model. Mallow's C_p is

$$C_p = \frac{SSE(R)}{MSE(F)} - n + 2p.$$

- Measures the bias in the reduced regression model relative to the full model.
- The full model is chosen to provide an unbiased estimate $\hat{\sigma}^2 = MSE(x_1, \ldots, x_k)$. Predictors must be in "correct form" and important interactions included.
- If a reduced model is unbiased, $E(\hat{Y}_i) = \mu_i$, then $E(C_p) = p$ (pp. 357-359).
- For the full model, $C_p \equiv k + 1$.
- If $C_p \approx p$ then the reduced model predicts as well as the full model. If $C_p < p$, then the reduced model is estimated to predict *better* than the full model.

Which criteria to use?

R_a^2, AIC, BIC, and C_p may give different "best" models, or they may agree. The ultimate goal is to find model(s) that balance(s):

- A good fit to the data.
- Low bias.
- Parsimony.

All else being equal, the simpler model is often easier to interpret and work with. Christensen (1996) recommends C_p and notes the similarity between C_p and AIC.

Two methods for "automatically" picking variables

- Two automated methods for variable selection are **best subsets** and **stepwise** procedures.
- Best subsets simply finds the models that are best according to some statistic, e.g., smallest C_p of a given size. Only `proc reg` does this automatically, but does not enforce hierarchical model building; grouping of coded categorical variables is ignored as well.
- Stepwise procedures add and/or subtract variables one at a time according to prespecified inclusions/exclusion criteria. Useful when you have a very large number of variables (e.g., $k > 30$). Both `proc reg` and `proc glmselect` incorporate stepwise methods.

Best subsets for Fall 2008 GPA data

Refer to the text example for a best subsets regression with continuous variables and interactions. The Fall 2008 data set includes a response (GPA), two continuous predictors (Verbal SAT and Math SAT) and numerous categorical predictors. We recoded the categorical predictors with a modest number of categories (Class, Race, Gender, Enrollment Status, Registration Status) and included each main effect as a set.

```
data fall08; set fall08;
proc reg data=fall08;
model cltotgpa=satv satm class2 class3 class4 housing raceaa raceo raceu
 genderf enrollft regn rego/selection=cp best=10;
run;
```

9.4 automated variable search (pp. 361–368)

Forward stepwise regression (pp. 364–365)
We start with k potential predictors $x_1, \ldots, x_k$. We add and delete predictors one at a time until all predictors are significant at some preset level. Let α_e be the significance level for adding variables, and α_r be significance level for removing them.

Note: We should choose $\alpha_e < \alpha_r$; in book exmaple, $\alpha_e = 0.1$ & $\alpha_r = 0.15$.

Forward stepwise regression

1. Regress Y on x_1 only, Y on x_2 only, up to Y on x_k only. In each case, look at the p-value for testing the slope is zero. Pick the x variable with the smallest p-value to include in the the model.
2. Fit all possible 2-predictor models (in general j-predictor models) than include the initially chosen x, along with each remaining x variable in turn. Pick new x variable with smallest p-value for testing slope equal to zero in model that already has first one chosen, as long as p-value $< \alpha_e$. Maybe nothing is added.
3. Remove the x variable with the *largest* p-value as long as p-value $> \alpha_r$. Maybe nothing is removed.
4. Repeat steps (2)-(3) until no x variables can be added or removed.

proc glmselect

- *Forward selection* and *backward elimination* are similar procedures; see p. 368. I suggest stepwise of the three.
- `proc glmselect` implements automated variable selection methods for regression models.
- Does stepwise, backwards, and forwards procedures as well as least angle regression (LAR) and lasso. Flom and Casell (2007) recommend either of these last two over all traditional stepwise approaches & note they both perform about the same.
- The syntax is the same as `proc glm`, and you can include class variables, interactions, etc.

proc glmselect

- The `hier=single` option buildes hierarchical models. To do stepwise as in your textbook, include `select=sl`. You can also use any of AIC, BIC, C_p, or R_a^2 rather than p-value cutoffs for model selection.
- `proc glmselect` will stop when you cannot add or remove any predictors, but the "best" model may have been found in an earlier iteration. Using `choose=cp`, for example, gives the model with the lowest C_p as the final model, regardless where the procedure stops.
- `include=`p includes the first p variables listed in the model statement in every model. Why might this be necessary?

proc glmselect

Fall 2008 enrollment data: Stepwise selection, choosing hierarchical model with smallest C_p during stepwise procedure.

```
proc glmselect;
class class race regstatus gender enroll housing
model cltotgpa=satv satm class housing race gender enroll regstat
  satv*race satv*gender satv*enroll satm*race satm*gender satm*enroll/select=sl
  sle=0.1 sls=0.15 hier=single;
run;
```

The most interesting part of this analysis was actually the EDA that identified a previously-overlooked problem with the Registration Status variable.

Stepwise procedures vs. best subsets

- Forward selection, backward elimination, and stepwise procedures are designed for very large numbers of variables.
- Best subsets works well when the number of potential variables is smaller. You can identify best subsets in `proc reg`, but SAS will not weed out non-hierarchical models.
- Choose `proc glmselect` for "large p" problems and choose `proc reg` for smaller numbers of predictors, e.g., $k < 30$.