Model selection II

Patrick Breheny

February 22

The Swiss fertility data set

- Today we will explore how to calculate and apply model selection criteria in R and SAS
- We will look at a data set describing fertility in Switzerland in 1888
- Some background: fertility (number of children a woman gives birth to over the course of her life) is high in developing nations and low in developed countries
- In 1888, Switzerland was at the critical point in its development where it was undergoing the "demographic transition" in which its fertility was falling to levels seen in developed nations

The Swiss fertility data set (cont'd)

The following variables were collected (primarily from military records) for each of the 47 French-speaking provinces in Switzerland:

- `Fertility` (standardized)
- `Agriculture`: Percent of males involved in agriculture as an occupation
- `Examination`: Percent of draftees who received the highest mark on their army examination
- `Education`: Percent of draftees educated beyond primary school
- `Catholic`: Percent Catholic (as opposed to Protestant)
- `InfantMortality`: Percent of live births who live less than one year

Model selection criteria in R

- In R, R^2 and $\mathrm{R}^2_{\mathrm{adj}}$ are given by `summary`
- AIC, BIC, and C_p can all be obtained from the `extractAIC` function:

```
fit <- lm(Fertility~.,data=swiss)
extractAIC(fit)
extractAIC(fit,k=log(n)) ## BIC
extractAIC(fit,scale=sig2) ## Cp
```

  Note: there is also a function `AIC`, though be aware that the two functions do not return exactly the same number (`AIC` drops constant terms)
- There is no default function to calculate PRESS, but you can calculate it via:

```
sum((fit$resid/(1-hatvalues(fit)))^2)
```

Model selection criteria in SAS

- In SAS, unfortunately, none of these options are available in PROC GLM – you have to use PROC REG
- In PROC REG, you can get all the criteria with

  ```
  PROC REG DATA=swiss OUTEST=fits;
    MODEL Fertility = Agriculture Examination Education
     Catholic InfantMortality / ADJRSQ CP AIC BIC PRESS;
  RUN;
  ```

  although be aware that this estimate of CP is meaningless without an external estimate of $\hat{\sigma}^2$
- Note that PROC REG allows multiple model statements; all these models show up in fits, so you can select from among a list of models by, say, running PROC SORT on fits

Automatic model selection: Overview

- Given that we have an objective way of choosing between models, it is possible to automate the model selection process by fitting a number of models and ranking them according to some criterion
- There are two common strategies to searching through the set of all possible models:
  - Best subsets selection, an exhaustive search of all possible models
  - Stepwise selection, in which the search is simplified by taking it one step at a time

Best subsets selection in SAS

- We will start with best subsets regression and illustrate its use on the Swiss fertility data set
- PROC REG in SAS allows for best subsets regression based on C_p through the SELECTION option in a MODEL statement:

```
PROC REG DATA=swiss;
  MODEL Fertility = Agriculture Examination Education
           Catholic InfantMortality / SELECTION = CP;
RUN;
```

Best subsets selection in R

- In R, best subsets selection is available through the `leaps` package, which you will have to install:

  ```
  install.packages("leaps")
  require(leaps)
  ```

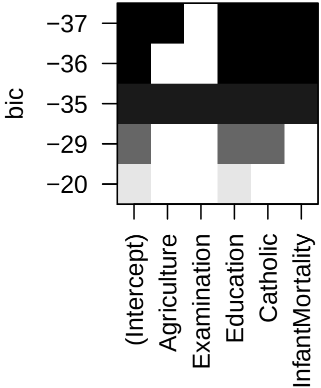
- Once installed, you can perform best subsets selection via:

  ```
  regs <- regsubsets(Fertility~.,data=swiss)
  ```

  where the `.` in a model formula means "all the variables not already in the formula"

Best subsets selection in R

These models can be plotted via `plot(regs)`:



Options in leaps

- By default, the package only returns the best model of each size (the best one-variable model, two-variable model, etc)
- This can be modified with the `nbest` option:

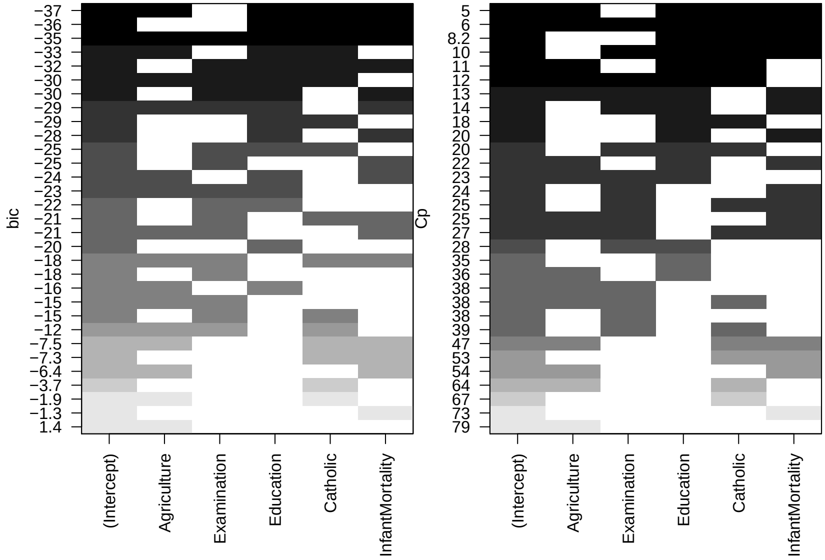
  ```
  regs <- regsubsets(Fertility~.,data=swiss,nbest=10)
  ```

- The `leaps` package calculates BIC and C_p by default; to rank by C_p, we can submit

  ```
  plot(regs,scale="Cp")
  ```

- `regsubsets` returns RSS and p, so in principle, one could calculate and sort by AIC as well

Best subsets



Interpreting the results

- Note that the best subsets approach selects all of the variables except `Examination`
- This is somewhat interesting, as `Examination` is highly correlated with `Fertility` (-0.65, $p = 7 \times 10^{-7}$)
- However, `Examination` is also highly correlated with `Education` and `Agriculture`, and seems to add little to the model if those two variables are already present

Stepwise selection

- Note that there are 2^p total subsets in a model with p candidate explanatory variables
- This number gets extremely large very quickly as p increases, to the point where, if p is above 40 or so, it is computationally infeasible to fit all these models
- Thus, *stepwise* approaches have been proposed, in which you find the best one-variable model, then find the best two-variable model that can be constructed by adding a variable to the best one-variable model, and so on
- In other words, we don't look at all two-variable models, only those that contain the best one-variable model

Stepwise selection (cont'd)

- Specifically, this approach is known as *forward* stepwise selection
- An alternative approach is to start with the full model and successively eliminate variables, which is known as *backward* stepwise selection
- There are other variants as well, capable of moving both forward and backward, allowing for a variable to be added, but then removed later if it no longer seems necessary

Stepwise selection in R

- In R, stepwise selection can be requested via the `method` option, as in

```
regs <- regsubsets(Fertility~.,data=swiss,method="forward")
```

- Other options for `method` include `''backward''` and `''seqrep''` for sequential replacement (an approach which considers both forward and backward steps)
- In this case, of course, stepwise selection is unnecessary, as we can perform the exhaustive search
- However, it is worth noting that the stepwise approach in this case selects the same model (this is not always the case)

Stepwise selection in SAS

- In SAS, stepwise selection can be requested via the `SELECTION` option in the model statement (*e.g.*, `SELECTION=FORWARD`, `SELECTION=BACKWARD`, `SELECTION=STEPWISE`)
- A subtle distinction between SAS and R is that SAS considers "stepwise" to be synonymous with choosing variables based on p-values
- In other words, it moves forward by adding the variable with the most significant p-value, and backwards by dropping the variable with the least significant p-value
- This approach is popular because, by construction, it leads to models with significant terms, but is on a weak foundation, as the construction of the model is not guided by any meaningful model selection criterion

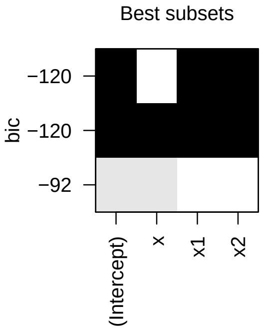
Greedy algorithms

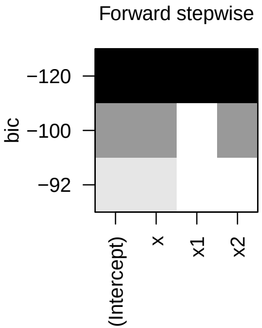
- In the computer science literature, stepwise approaches are known as *greedy algorithms*, in the sense that they operate according to grabbing the variable that will help most in the short term
- However, just as this is not always the best strategy in other areas of life, there is no guarantee that the stepwise approach will find the best model

Example

For example, consider:

$$X_1 \sim N(0,1)$$
$$X_2 \sim N(0,1)$$
$$X|X_1,X_2 \sim N(X_1+X_2,0.5)$$
$$Y|X_1,X_2,X \sim N(X_1+X_2,1)$$





Best subsets vs. stepwise approaches

- In low dimensions, forward stepwise algorithms and best subset approaches often agree
- Even if they do not (as in the previous example), simple tweaks such as sequential replacement usually fix the problem
- In higher dimensions, however, stepwise approaches only investigate an exceptionally small fraction of the possible models, and rarely find the optimal model