STA302/1001: Methods of Data Analysis

Instructor: Fang Yao

Chapter 10: Variable Selection

Variable Selection

- also known as model selection
- goal: given a set of predictor variables $X_1, \ldots, X_p$, we want to identify the correct model that describes the behaviour of the response Y
- that is, we will ask questions like:
  – how many predictors should be included?
  – any interaction terms (e.g., $X_i X_j$)?
  – do we need higher order terms (e.g. X_i^2)?
- in real life there is never a “correct” model
- all we could do is to find the “best” model for the problem that we are trying to solve

Collinearity

basically : no independence
→ collinearity

- issue caused by redundant terms
- or known as multi-collinearity
- two terms are exactly collinear, if $c_1 \neq 0, c_2 \neq 0$,

$$c_1 X_1 + c_2 X_2 = c_0$$

c_0 is not necessarily 0
b/c we have intercept column.

for some constants c_0, c_1 and c_2 (this holds for for all observations in the data)

- in other words, given c_0, c_1, c_2 and X_1, we can determine X_2, and vice versa

Collinearity - Multi-terms etc

- this concept can be generalized to more than 2 terms
- and also to "approximately collinear"

$$\boxed{c_1 X_1 + c_2 X_2 + \cdots + c_p X_p \approx c_0}$$

(at least two c_j's are not 0)

- collinearity is measured by the square of sample correlation (r_{12}^2 for two terms, max of all r_{ij}^2 for multiple terms)
- $r_{ij}^2 = 1 \Rightarrow$ collinearity
- r_{ij}^2 close to 1 $\Rightarrow$ approximate collinearity

Collinearity - What is the Harm?

- what will happen if collinearity exists?
- inverse of $\mathbf{X}'\mathbf{X}$ does not exist
- so no fitting can be done
- one way to solve it is to drop some terms in fact R automatically ignores redundant terms
- what will happen if approximate collinearity exists?
- variance of $\hat{\beta}$ will be undesirably large (one consequence)

so it's not so bad to have a perfect collinearity

This is the difficulty.

Then the following
t-value (too small)
insignificance ....
all will be influenced.

Approximate Collinearity & Variance

- to see this, consider a 2–term model:

$$\mathrm{E}(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- it can be shown that

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{1 - r_{12}^2}\frac{1}{SX_jX_j}$$

where $SX_jX_j = \sum(x_{ij} - \bar{x}_j)^2$

- so we do not want collinearity, which can be achieved by dropping terms or by doing transformation

example: 7.1 in textbook

Fixing Collinearity

- mathematically: try to make $\mathbf{X}'\mathbf{X}$ as diagonal as possible
- statistically: want to minimize $\mathrm{Var}(\hat{\beta})$
- practically: try to remove terms that do not provide additional information (remove redundant terms)
- automatic methods for doing model selection
- first notice that least squares is a method for parameter estimation, but not for model selection
- it is because it always favors models with larger number of parameters

参数越多,模型越准确

Selection Criteria: Basic Idea

- instead of minimizing just the RSS, most model selection methods choose the best model as the one that minimizes

$$f_1(RSS) + f_2(p) \longrightarrow$$

it's like a penalty, more para, more penalty (increase the value you want to minimize)

where f_1 and f_2 are increasing functions and p is the number of parameters in the model

- large models: $RSS \downarrow$ and $p \uparrow$
- small models: $RSS \uparrow$ and $p \downarrow$
- goal: find a good balance between these two aspects

Four Common Criteria

- Akaike Information Criteria (AIC)

$$\underbrace{n\log\left(\frac{RSS}{n}\right)}_{f_1(RSS)}+\underbrace{2p}_{f_2(p)}$$

- Bayesian Information Criteria (BIC)

$$n\log\left(\frac{RSS}{n}\right)+\underbrace{p\log(n)}_{f_2(p)}$$

"penalize heavier"

- Mallows' C_p

$$\frac{RSS}{\hat{\sigma}^2}+2p-n$$

where $\hat{\sigma}^2$ is estimated with all terms

Four Common Criteria -con't

- cross-validation (CV)

$$\sum_{i=1}^{n}(y_i - \hat{y}_{i(i)})^2 = \sum_{i=1}^{n} \frac{\hat{e}_i^2}{(1-h_{ii})^2}$$

called "predictive residual sum of squares" (PRESS)

- easy to compute for linear models, no need to refit with leave-one-observation-out

like 3, 4, 5 ...

- in practice: if the number of terms is not that large, we could fit all possible models (2^p of them), compute the criterion value using one of these four methods, and pick the one with the smallest value

of these different models (not of methods)

Four Common Criteria -con't

- if k is large, we can do

  1. forward selection (FS)

     Forward selection, which involves starting with no variables in the model, testing the addition of each variable using a chosen model comparison criterion, adding the variable(if any) that improves the model the most, and repeating this process until none improves the model.

  2. backward elimination (BE)

     Backward elimination, which involves starting with all candidate variables, testing the deletion of each variable using a chosen model comparison criterion, deleting the variable (if any) that improves the model most by being deleted, and repeating this process until no further improvement is possible.

  3. mixture of both (FS and BE at each step)

     Bidirectional elimination, a combination of the above, testing at each step for variables to be included or excluded.

STEPWISE REGRESSION (wiki)

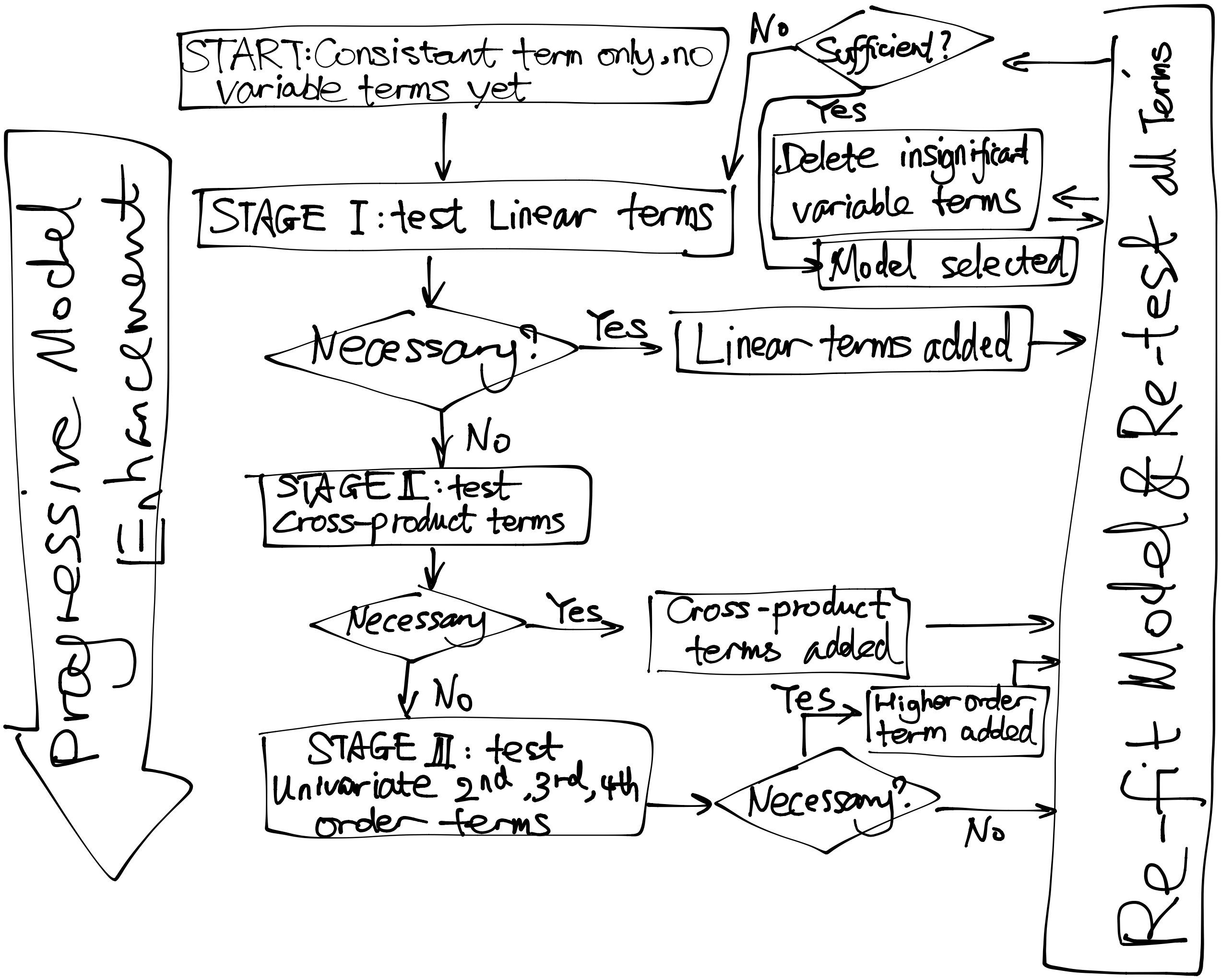
- two new phrases:

  1. underfitting: model too small, has bias and missing important predictors

  2. overfitting: model too big, has high variance and possibly wrong conclusions

- highway data example

  (bunch of variables)


START: Consistant term only, no variable terms yet
STAGE I: test Linear terms
Necessary?
Yes
Linear terms added
No
STAGE II: test Cross-product terms
Necessary
Yes
Cross-product terms added
No
STAGE III: test Univariate 2nd, 3rd, 4th order terms
Necessary?
Yes
Higher order term added
No
Re-fit Model & Re-test all Terms
Sufficient?
No
Yes
Delete insignificant variable terms
Model selected
Progressive Model Enhancement

Practical Model Building

- parsimony — strive for simplicity
- scope: under what conditions your model “works”?
- of course, parsimony and scope are related
- a famous quote (McCullagh & Nelder):

  “modeling in science remains, partly at least, an art”

- three principles from the same book:
  (i) all models are wrong, but some are useful
  (ii) do not fall in love with one particular model
  (iii) do diagnostic checking: it can tell you if anything went wrong

Other Principles and/or Hints

- the first step is not to look at the data, instead
  1. think about the process that generated the data
  2. think about the background behind
  3. bring “known” background knowledge into the model whenever possible
- main effects should not be excluded if interactions are to be included
- do not fully rely on automatic methods for finding a “correct model”: useful for initial screening
- final model may depend on other ground than purely statistical considerations, e.g., costs etc

A Quick Summary of STA302

- goal of modeling: “to find a good approximation of life”
- what you have learnt can be loosely grouped into 3 parts:

1. tools when you know which model you want to fit ch2-5
2. diagnostic checking (after term test)
3. correct/improve your fitted model

When you know which model you want to fit

- 3 assumptions of linear regression
- OLS, WLS

↙ some classic problems like the term test

- confidence intervals, tests, based on t, F distributions
- standard error calculations (include delta method)
- prediction (attach uncertainty) (how to count df correctly, like termtest)
- interpretation of your fitted model

fitted value interval v.s. prediction interval

Diagnostic Checking

- lack of fit test

Sulfer example.

- residual plots (3 components)

•tell what's going wrong with your model. (ex: Fuel consumption)

- leverage h_{ii}
- outlier tests
- Cook's distance

How to calculate these.

idea: outlier test
you cannot only test the most extreme points, have to adjust the whole data.
(that's why we need B-p value)

- Q-Q plots

Correcting/Improving your fitted model

- transformation (how many types?)
- adding/dropping terms, i.e., model selection
- ridge regression
- if you want to learn more...
  1. nonlinear regression
  2. generalized linear models
  3. nonparametric regression
  4. high-dimensional variable selection
  5. and many more...