Math 455 Class 33 April 22

Chapter 10 : Variable Selection.

Background : assume we have the "form" of the model already : we don't need to transform $Y \mapsto \log(Y)$ or $Y \mapsto Y^2$

don't need to add X^2 term, etc.

We are only considering

$$Y \sim \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

We are thinking about which X_i to include.

(One possible answer is "all of them")

Why might we NOT want to include all of them?

Tradeoffs we are making :

"Big" (many variables) models :

can be more accurate, account for more factors

"Small" (few variables) models :

more interpretable, less danger of "overfitting," tend to be more stable.

Where you want to be in this tradeoff depends on your application.

Examples: ① You are building a model which will be applied by a computer, second-to-second, to maximize ad revenue.

Here a "big" model can be appropriate.

- complexity doesn't matter to the computer.
- maximizing revenue is the main goal.
- don't care about explaining.

② We are trying to write something explaining public policy. The model has to be "interpretable" so that it is convincing to policy-makers.

A "small" model is more appropriate.

We want to make a yes/no case and don't care so much whether the result is perfectly accurate.

In Ch. 10, Variable Selection, there are lots of choices possible. Your choices will depend on your desired application:

One obvious method: "Best Subset."

This just means that we have some metric for "best" and we look at all 2^p subsets of $\{X_1, \ldots, X_p\}$ and we choose the one that is best according to our metric.

What is a metric? Many choices:

AIC, BIC (Explain later)

R^2, Adjusted-R^2

Mallow C_p (Explain)

Example question: We apply "best subset" variable selection and our metric is R^2. Which subset of $\{X_1, \ldots, X_p\}$ is most likely to be selected?

A: The whole set $\{X_1, \ldots, X_p\}$.

Because if we add a variable, we can only increase R^2. R^2 never goes down when we add a variable.

Notice that this method does not address the trade-offs we discussed.

Another problem with "Best Subset": we need to consider 2^p subsets. If $p = 30$, this means ~ 1 billion models.

Most methods of variable selection will have a penalty for more variables or some method for rejection of variables.

Method 1: "Backward elimination".

Start with the model including ALL variables.

Choose some "α_{crit}" - threshold p-value, maybe 0.05., maybe 0.15 ?

Are there any variables with p-value $> \alpha_{crit}$?

Now repeat the procedure.

This again does NOT consider all 2^p models.

We consider AT MOST

$p + (p-1) + (p-2) + \cdots$ potential models.

Method 3: "Stepwise Regression": at each stage consider <u>both</u> adding 1 variable and dropping 1 variable.

(Many possible methods for doing this.)

10.3: Criterion-based procedures.

Introduce a "score" for each model:

more points (better) for accuracy

lose points (worse) for having more variables.

Examples: Adjusted-R^2, AIC, BIC, C_p.

If so, throw away the variable with highest p-value, run a new regression, repeat.

If not, done.

Notice that this does NOT consider every possible subset!

This considers at most $p+1$ models.

This can be done "by hand" but there are R functions to do it for you.

Method 2: "Forward Selection"

Start with the "null model" 0 explanatory vars.

Start by adding 1 variable to the model.

We have p choices $X_1, \ldots, X_p$.

If none of them have p-value $< \alpha_{crit}$, stop (stick with the model we have).

If some do, add the variable with lowest p-value. Say it is X_3

Now we have a new model $Y = \beta_0 + \beta_3 X_3 + \varepsilon$

There are $p-1$ variables we could add.