Math 455 Class 34 April 23

Midterm: Monday April 28, in class 80 min.

~~Eco~~: Due date for practice midterm will be postponed to Sunday April 27.

Today: more about Ch. 10, Variable Selection.

Q: We already know how to compare models with an F-test. Why do we need all these techniques in Ch. 10?

A: The F-test can only be used to compare nested models. If there is a variable in model A that is NOT in model B, and vice-versa, then A & B are not comparable via F-test.

Some remarks from 10.1:

in general, domain-specific considerations can pre-empt variable selection.

- if we are considering a model in powers of X, generally we want to include all powers up to some degree.

Many reasons for this, but one simple one as an example:

if we have $Y \sim \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$

If we eliminate the X-term, then we are saying that the effect of ~~reducing~~ negative X is the same as that of ~~increasing~~ positive X.

We don't want to do this unless there is some prior expectation that this is so.

Other reasons: changes of scale:

$X \mapsto X + a$ $\qquad (X+a)^2 = X^2 + 2aX + a^2$

and the X term reappears.

Summary: there are situations where we should not eliminate variables even if the algo says it "wants" to.

We talked about 4 "metrics" for variable selection.

adjusted R^2, AIC, BIC, Mallow C_p.

$$R_a^2 = 1 - \frac{RSS/(n-p)}{TSS/(n-1)} = \underbrace{1 - \left(\frac{n-1}{n-p}\right)(1-R^2)}_{\star}$$

$$= 1 - \frac{\hat{\sigma}^2_{model}}{\hat{\sigma}^2_{null}}$$

p = # of parameters.

$(1-R^2)$ = amount of "unexplained variance".

ideally, we want p to be small.

and R^2_{adj} to be large. (close to 1)

If we increase p we increase R^2 so,

decrease $1-R^2$.

But this decreases the denom. in $\frac{n-1}{n-p}$.

Thus decreasing R^2_{adj}.

Formula $\star$ above has a trade-off.

All these ~~Forward~~ metrics (AIC, BIC, C_p)

have the same type of trade-off.

R^2_{adj} : easiest to understand.

AIC : preferred for prediction problems

BIC : prefers more parsimonious models than AIC
(larger penalty term)
Justified by Bayesian Statistics.

Mallow C_p : Warning : opposite scale.
that is high score = bad
whereas with adj. R^2, high score = good.

Note: R^2_{adj} is between 0 and 1.
this is NOT true for the other scores.

AIC = Akaike Information Criterion.

Defn : $-2L(\hat{\theta}) + 2p$
where $L(\hat{\theta})$ = log-likelihood.

$\hat{\theta}$ represents all of our parameter estimates
$\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_p$.

This AIC will be computed for you by R.

Convenient Fact: for linear regression models,

$$-2L(\hat{\theta}) = \underline{n \log(RSS/n)} + \text{const.}$$

So, this boils down to comparing models using the underlined term + penalty $2p$.

So some but not all R functions report AIC without the constant above.

This means AIC reported by one R function might not agree with AIC reported by another.

So: comparing models with AIC, make sure you use the same R function to compute AIC.

BIC: Bayesian Information Criterion.

Same but penalty term is $p \log n$.

Typically we have enough data that $\log n > 2$.

So penalty is larger, so smaller models are preferred.

Mallow C_p : $\frac{RSS}{\frac{1}{\sigma^2}} + 2p - n$.