Math 455 Class 36 April 24

New homework is posted.

Yes, there will be a sample final exam.

Today: Chapter 10: Model Selection.

Q: What problem are we trying to solve?

A: We already know how to decide whether a single variable merits inclusion in a model: we have the t-test.

We can compare two nested models with the F-test.

What if we have to compare two models that are NOT nested?

The methods of this chapter do exactly that.

Four methods: Adjusted R^2, AIC, BIC and Mallows' C_p.

Most popular: AIC.

All these methods balance "goodness of fit" with a penalty for having many variables.

Every model gets a score which is improved by having a better fit but made worse by having more variables.

Warning: Not all scales are oriented in the same way.

Adjusted R^2: higher is better.

AIC, BIC, Mallows' C_p: lower is better.

What are all these scoring methods?

Adjusted R^2: we understand ordinary R^2

$0 \leq R^2 \leq 1$.

We can't use ordinary R^2 as a scoring method; because ordinary R^2 always prefers.

the model with more variables (if nested).

The ordinary R^2 has NO penalty for introducing extra variables.

Adjusted R^2 introduces a penalty for more variables in such a way that if you introduce a "noise" variable that has NO connection with the response Y, the expected increase in Adjusted R^2 is zero.

Pros and Cons of Adjusted R^2:

$0 \leq$ Adjusted $R^2 \leq 1$: easy to interpret.

Simple formulas:

$$\text{Adjusted } R^2 = R_a^2 = \frac{RSS/(n-p)}{TSS/(n-1)}$$

$$= 1 - \left(\frac{n-1}{n-p}\right)(1-R^2)$$ ← ordinary R^2

* these also aid interpretability.

Cons: Adjusted R^2 tends to be very lax about adding variables: selecting this way will likely include a lot of noise variables.

What is AIC?

AIC = Akaike Information Criterion.

Idea: Maximize the log-likelihood.

The observed log-likelihood is not an unbiased estimator of the true log-likelihood.

Correct for the bias, and maximize the corrected value.

This corresponds to MINIMIZING the AIC

$$AIC = -2L(\hat{\theta}) + 2p$$

$L(\hat{\theta})$: log-likelihood

p = # parameters.

Model selection Criterion: minimize AIC.

Why is this "most popular"?

It balances "laxness" and "strictness"

It tends to perform better for predictive tasks.

There is a connection to RSS: AIC can be expressed in terms of RSS and other things, e.g. p.

Why is this an issue?

AIC is a <u>relative</u> criterion.

We prefer models with lower AIC, but we don't try to interpret the absolute score.

Thus some R functions replace absolute AIC with another AIC which is a function of RSS and retains the

same relative AIC scores.

This means that the AIC from one R function (say AIC) is NOT the same as the AIC from another. (say extract AIC).

BIC: Bayesian Information Criterion.

similar to AIC, but harsher penalty:

$p(\log n)$ in place of $2p$.

$\uparrow$

$\log n > 2$ if $n > 8$.

More "strict" about including extra variables.

Smaller models result: easier to justify and interpet; but worse for prediction.

Mallows' C_p: $= \frac{RSS}{\hat{\sigma}^2} + 2p - n.$

As with AIC & BIC, lower is better.

One good feature: interpretability:

For a well-calibrated model, $C_p \approx p$.

That is Mallows' C_p tells you the number of parameters it "thinks" you are supposed to have.

Q: Why is model selection an issue?

A: Suppose that we have 30 potential explanatory variables. We now have 2^{30} potential models, corresponding to inclusion/exclusion of variables.

It might not be possible to look at all of them.

With a small number of variables, say 8, we could potentially look at all possibilities.

Function to look at all possibilities: regsubsets. in "leaps".

What if we can't do that?

There are "forward" and "backward" selection procedures:

Forward: start with the null model: no predictors.

Consider models $\beta_0 + \beta_i X_i + \varepsilon$.

include the X_i whose p-value is lowest and also below a threshold $\alpha_{crit.}$ (typically 0.15 - 0.20). ex

Say this is X_4. Then we consider models:

$\beta_0 + \beta_4 X_4 + \beta_i X_i + \varepsilon \quad (i \neq 4)$.

and pick the X_i whose p-value is lowest and below the $\alpha_{crit.}$.

Say this is X_2. Now we consider models

ex $\beta_0 + \beta_2 X_2 + \beta_4 X_4 + \beta_i X_i + \varepsilon \quad (i \neq 2, 4)$.

continue in the same way.

until there are no variables that meet the threshold α_{crit}.

We only have to look at:

$(p-1) + (p-2) + \cdots \to 1$ possibilities at worst.

and not 2^p.

There is a similar procedure for "backward selection" that starts with the full model, and eliminates variables with p-values over α_{crit}.

Also: back-and-forth "stepwise" procedures.