Model Selection Approaches

BIOS 6611

CU Anschutz

Week 14

Background

Purpose of the Model - I

Depending on our study design and purpose, we have various considerations for how we approach model selection.

For **epidemiologic study designs** (i.e., observational studies):

- One or a few exposure variables of interest; may include interactions
- Adjust for confounding
- May potentially collect a large number of variables without adequate scientific or historical information to inform which variables are important predictors and/or potential confounders (i.e., model selection is needed)

Purpose of the Model - II

Depending on our study design and purpose, we have various considerations for how we approach model selection.

For **clinical study designs** (i.e., randomized studies):

- Intervention effect is the variable of primary interest
- Little or no confounding is expected due to randomization; possible stratification in the study design
- Adjust for variables used in stratification/matching to increase precision; in general, these variables are identified during the design of the trial
- Model selection is usually not necessary or desired (models defined *a priori*)

Purpose of the Model - III

Depending on our study design and purpose, we have various considerations for how we approach model selection.

For **predictive modeling** (observational or randomized studies):

- Wish to identify a model that will best predict the outcome for future observations
- Model selection possible, but may or may not be necessary
- Use a hold-out sample (e.g., train/test sets) or external validation when interested in prediction

General Model Selection Considerations

Considerations for model selection procedures include:

- Goal of modeling: estimation, hypothesis testing, prediction, etc.
- Type I error inflation
- Adequate power for hypothesized effects, extent of confounding
- Form of the model: linear (BIOS 6611), generalized linear, nonlinear, etc.
- Functional form of the variables to suit model assumptions (e.g., linearity)
- Desire for a parsimonious model that avoids overfitting

Models can be compared by

- Numerical differences
- Statistically significant differences
- Scientifically meaningful differences

Methods for Model Selection

R^2 and Adjusted R^2

No matter how strong or weak an additional variable is, the SS_{Model} *never* decreases:

- If the new variable is a strong predictor, SS_{Model} significantly increases
- If the new variable is a poor predictor, SS_{Model} may change very little or stay the same

This implies the $R^2 = \frac{SS_{Model}}{SS_{Total}}$ will also never decrease with the addition of new variables.

Instead we use the *adjusted* R^2:

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

where p is the number of variables in the model (excluding the intercept). It can increase or decrease when a new variable is added.

In general, a larger value is desired, but no single benchmark exists for how much of an increase is meaningful.

F-tests and Partial F-tests

The overall F-test is meaningful to test if *any* variables in a given regression model are significantly associated with the outcome.

The partial F-test compares *nested* models to a given regression model:

$$F = \frac{[SS_{model}(full) - SS_{model}(reduced)]/k}{MS_{error}(full)} \sim F_{k,n-p-k-1}$$

where

- n be the number of observations
- p be the number of IVs in the *reduced model*
- k the number of IVs *removed* from the full model

If we wish to compare non-nested models, we cannot use this approach. For example, directly comparing a model with age & height to a model with age & weight as predictors.

Model Selection Criterion: AIC

Multiple proposed criteria have been proposed to evaluate the appropriateness of model fits. We discuss two classes here, noting many, many more exist.

The **Akaike Information Criterion (AIC)** is based on the goodness-of-fit and includes a penalty for increasing the number of parameters in the model:

$$AIC = 2k - 2 \times ln(L)$$

where k is the number of parameters in the model and $ln(L)$ is the log-likelihood for the model (specifically, L is the maximized likelihood function for the model).

For AIC,

- Models do NOT need to be nested (although they can be).
- The best model has the lowest AIC.
- An arbitrary rule of thumb, a difference of 2 is significant.

Model Selection Criterion: AICc

Like R^2 and the adjusted R^2, there is some criticism that AIC may favor models with more parameters based on its standard definition.

This can be especially true for smaller sample sizes, and may lead to overfitting.

A corrected version that has the same interpretation is

$$AICc = AIC + \frac{2k^2 + 2k}{n - k - 1}$$

Model Selection Criterion: BIC

The **Bayesian information criterion (BIC)**, also known as the *Schwarz Information Criterion (SIC)*, is similar to AIC in that it is based on the goodness-of-fit and includes a penalty for increasing the number of parameters in the model:

$$BIC = k \times ln(n) - 2 \times ln(L)$$

Like the AIC, the models do NOT need to be nested to be compared *and* the best model has the lowest BIC.

Some (arbitrary) rules for comparing to models[1] are:

ΔBIC	Evidence Against Higher BIC
0 to 2	Minimal
2 to 6	Positive
6 to 10	Strong
$>$10	Very Strong

[1] Kass, Robert E.; Raftery, Adrian E. (1995), "Bayes Factors", *Journal of the American Statistical Association*, 90 (430): 773–795.

Mallows' C_p

Mallows' C_p can be used to compare two models compared to a full reference model:

$$C_p = \frac{SSE_p(reduced)}{MSE_{p+k}(full)} - [n - 2(p+1)]$$

where the reduced model includes p variables and the full model includes $p + k$ variables.

Mallows' C_p attempts to balance the statistical trade-off between bias and variance:

- The full model is assumed to be unbiased and includes all relevant predictors (a somewhat strong assumption)
- Subset models with small values of C_p have a small estimated total variation in their predicted responses
- C_p "penalizes" the addition of variables to the model (helps to prevent overfitting and consider parsimonious models)

Mallows' C_p

- When comparing two models that have the same value for p, the model with the lower C_p is considered the better model
- C_p can be used to determine how many variables are in the "best" model since it will achieve a value of approximately $p+1$ if MSE(reduced) is roughly equal to MSE(full)
- If important predictors are omitted from the full model, $C_p > p+1$
- When C_p is near $p+1$, the bias is small and if it is much greater than $p+1$ there is substantial bias and important predictors may be omitted from the model
- For linear regression it has been shown that the standard AIC and C_p are equivalent.
- C_p is related to the partial F statistic (F_p below) by:

$$C_p = kF_p + (p - k + 1)$$

All Possible Subsets and an Example

All Possible Subsets

Assume you have identified the potential variables in to include in your model (we'll discuss variable selection considerations in another lecture).

One way to identify the "best" model is through **all possible subsets** regression, which fits *all* possible subsets of variables from the model. The most promising models can then be selected based on any of the previously discussed measures.

All Possible Subsets Considerations

1. Different measures can arrive at different "best" models (i.e., they aren't all equivalent measures).
2. All possible subsets regression can quickly become computationally prohibitive. With p total variables of interest, there are 2^p possible models.
3. If you have a hierarchical model (e.g., interaction or polynomial terms), some subsets will exclude main effect terms or lower order polynomials and may not make sense.

Example

Let's examine using our FEV dataset to identify the "best" model for predicting FEV from the available predictors of age, height, sex, and smoking status. For simplicity, we will assume there are no interactions, a need for polynomial terms, etc.

Given that we have 4 predictors, there are a total of $2^4 = 16$ possible models.

Example - R Code

```
dat <- read.csv('FEV_rosner.csv')

# create indicator variables with 1/0
dat$smoker <- as.numeric(dat$smoke == 'smoker')
dat$male <- as.numeric(dat$sex == 'male')

library(leaps) # package to implement all subset regression
asr <- regsubsets(fev ~ age + height + male + smoker, data=dat,
                  method='exhaustive', nbest=8)

# create object to summarize results
asr_summary <- cbind(
  summary(asr)$which, # variables in each model
  rsq = summary(asr)$rsq, # R^2
  adj_rsq = summary(asr)$adjr2, # adjusted R^2
  BIC = summary(asr)$bic, # BIC
  Cp = summary(asr)$cp) # Mallows' Cp
```

Example - SAS Code

```
data dat;
    set dat;
    if smoke='nonsmoker' then csmoke=0; /* smoker=1; nonsmoker=0 */
    else if smoke='smoker' then csmoke=1;

    if sex='male' then male=1; /* male=1; female=0 */
    else if sex='female' then male=0;
run;

PROC REG DATA=dat;
  MODEL fev = age height male csmoke / SELECTION=rsquare adjrsq bic cp;
RUN;
```

Example

```
round(asr_summary, 3) # round results to 3 places
```

```
##   (Intercept) age height male smoker   rsq adj_rsq      BIC       Cp
## 1           1   0      1    0      0 0.754   0.753 -903.311   61.702
## 1           1   1      0    0      0 0.572   0.572 -542.391  585.863
## 1           1   0      0    0      1 0.060   0.059  -27.663 2065.066
## 1           1   0      0    1      0 0.043   0.042  -16.077 2113.592
## 2           1   1      1    0      0 0.766   0.766 -931.584   26.867
## 2           1   0      1    1      0 0.759   0.758 -910.452   49.030
## 2           1   0      1    0      1 0.754   0.753 -896.840   63.689
## 2           1   1      0    1      0 0.607   0.606 -591.339  487.431
## 2           1   1      0    0      1 0.577   0.575 -542.604  575.275
## 2           1   0      0    1      1 0.112   0.109  -58.269 1917.375
## 3           1   1      1    1      0 0.775   0.774 -948.480    5.168
## 3           1   1      1    0      1 0.768   0.767 -928.486   25.382
## 3           1   0      1    1      1 0.759   0.758 -904.310   50.666
## 3           1   1      0    1      1 0.609   0.608 -588.768  482.660
## 4           1   1      1    1      1 0.775   0.774 -944.178    5.000
```

Closing Thoughts

Model selection is ultimately an art:

- There is no single "best" method to arrive at the "optimal" model
- However, there are several wrong models due to high collinearity, models that are not hierarchical, etc.

Model selection should be first and foremost be driven by your knowledge of the subject area and by your hypotheses.

Different selection strategies can lead to different sets of variables being included in the final model.

A good analysis should point out that there are different possible models when more than one "adequate" model is detected in an analysis.

For prediction modeling, we should utilize either an external validation set, split our data into training and testing subsets, or utilize cross validation approaches.