Some Ideas concerning Model Selection

John Kolassa

November 26, 2018
An Example I

- Example data from http://lib.stat.cmu.edu/datasets/NO2.dat
  - 500 observations regarding air quality in a city in Norway
  - Response: Log NO2 concentration
  - Explanatory variables
    - Log of cars per hour
    - Temperature at 2 meters (C)
    - Wind speed (m/s)
    - Difference between temperature (C) at 25 m and at 2 m
    - Wind direction (degrees)
    - Time of day (hour)
    - Number of day since 1 Oct 2001
Some Notation I

Standard Multiple Regression Model:

\[ E[Y_i] = x_i \beta \]

- \( Y_j \) independent
- \( Y_j \) have same variance (denoted by \( \sigma \))

\[ Y_j \sim \text{Gaussian} \quad (1) \]

Generalized Linear Model modifies (1)
- Postulate new distribution for \( Y_j \)
- Postulate link function indexing distribution by \( E[Y_j] \)
Some Definitions I

- **Random variable**: quantity observed on some or all of the various units of observation in a study.
- **Response variable**: Variable being explained.
- **Explanatory variables**: Variable used in explanation.
- **Factor**: Variable whose values consist in categories.
  - **Level**: One of the possible categories for a factor.
  - Factor gives rise to one fewer parameters than the number of levels.
  - One of the levels gets treated as baseline.
The Task I

- Determine transformations for $Y_j$, components of $x_j$
- Choose a stochastic structure for $Y_j$
- Choose which components of $x_j$ ought to be considered
- Check to see if choices made above are supported by the data.
Building models I

▶ Blindly-built regression model: add all seven covariates as linear predictors

▶ Smarter model will use mathematical and subject matter knowledge to build a better model.
  ▶ Response is always positive, and so taking log puts it on a scale that makes linear fits meaningful.
  ▶ Consider scale for explanatory variable that makes sense
    ▶ You might want to allow for a model in which NO2 is $\propto$ cars per hour,
    ▶ so logging cars per hour is also sensible.

▶ Hour and wind direction are cyclic.
  ▶ Model using sines and cosines.
Is this model too small?

▶ Consider adding terms that allow for a more flexible dependence of response on covariates.
  ▶ Maybe add polynomial terms in explanatory variables.
▶ Consider adding terms that allow for effect of one variable to depend on the level of another.
  ▶ Most standard way is to add interactions: products of covariates
▶ Adding all of these terms gives \( 7 + (7 \times 6)/2 = 28 \) terms total.
Principles for Reasonable submodels I

Some ideas about what makes for an incoherent model [McCullagh(2002)]

- If powers of a term appear in the model, shifts in the origin of the measurement scale can arbitrarily knock out lower terms.
  - Hence do not consider removing lower order terms in the presence of higher-order terms.
  - Similar issues apply to interaction terms.
- Removing parameters associated with some factors collapses that category with the baseline category.
  - Removing parameter associated with one level of a factor collapses the associated level into baseline.
    - Model selection becomes dependent baseline choice, which is usually arbitrary.
- Removing one sine-cosine pair members fixes start of cycle.
  - Remember the sine-of-difference and cosine-of-difference formulae from trig?
  - Unless the model is parameterized to explicitly have a meaningful null-hypothesis start of the cycle, these coefficients should only be evaluated as a pair.
Reducing the model I

Is your model too big?

- Search among submodels for the best model.
- This is a big job, since there are $2^{28}$ models to search through.
  - Previous principles reduce the number of models to consider.
  - Even so, cycling through all models is tough.
- People often use stepwise:
  - Start with an initial model.
  - Consider models with separate (groups of) parameters added or removed, one at a time.
  - Move to model with numerical criteria improved.
  - Gives a local, rather than guaranteed global, optimum.
Some Model Criteria I

What makes for a good model? [Kadane and Lazar(2004)]

- Akaike Information Criterion: Lower is better
  - Maximized Log likelihood

\[
\ell(\hat{\beta}, \hat{\sigma}) = -(n/2) \log(2\pi \hat{\sigma}^2) - \sum_{i=1}^{n} \frac{(Y_i - \hat{\beta} x_i)^2}{2\hat{\sigma}^2}
\]

\[
= -(n/2) \log(2\pi \hat{\sigma}^2) - n/2.
\]

- AIC is \(2p - 2\ell(\hat{\beta}, \hat{\sigma}) = 2p + n[\log(2\pi \hat{\sigma}^2) + 1]\)
  - \(p\) is number of parameters, incl. \(\sigma\)
  - Software reliably reports this only up to an additive constant.

- Bayesian Information Criterion: Lower is better
  - \(p \ln(n) - 2\ell(\hat{\beta}, \hat{\sigma})\)
  - Large-sample approximation to Bayesian result with prior on number of models.
    - As often happens, precise form of prior disappears in first-order approximation.
  - Penalty for larger model increases as \(n\) increases.
Some Less Popular Model Criteria I

- Mallows $C_p$
  - $C_p = \frac{SSE_p}{\hat{\sigma}^2} - n + 2p$ for $SSE_p$ the residual sum of squares using $p$ parameters (now not including $\hat{\sigma}$)
  - Best when $C_p \approx p$.

- $p$-values: Remove variables sequentially with large $p$-values
  - Threshold generally much larger than for significance.
  - Not recommended: Variable entry can depend on ordering of consideration.
Interpreting Models I

- **Interpretation:**
  - Model parameters measure effect of explanatory variable in light of all other variables in model.
  - Hence interpretation of parameter changes as other variables move in and out of the model.

- **Inference after selection problematic.**
  - Effect of variables in a best-fitting model will be exaggerated relative to a model selected a priori.
  - This exaggerated effect must be adjusted for when performing post-selection inference.
  - A solution: test and training set.
Joseph B. Kadane and Nicole A. Lazar. 
Methods and criteria for model selection. 
ISSN 01621459. 

Peter McCullagh. 
What is a statistical model? 
URL https://doi.org/10.1214/aos/1035844977.