MLR Model Selection

Author: Nicholas G Reich, Jeff Goldsmith

Today's Lecture

- Model selection vs. model checking
- Stepwise model selection
- Criterion-based approaches

Model checking and selection: KNN, Chapters 9 and 10.

Model selection vs. model checking

Assume $y|\mathbf{x} = f(\mathbf{x}) + \epsilon$

- model selection focuses on how you construct $f(\cdot)$;
- model checking asks whether the ϵ match the assumed form.

Why are you building a model in the first place?

Model selection: considerations

Things to keep in mind...

- **Why am I building a model?** Some common answers
  - Estimate an association
  - Test a particular hypothesis
  - Predict new values
- What predictors will I allow?
- What predictors are needed?
- What forms for $f(x)$ should I consider?

Different answers to these questions will yield different final models.

Model selection: realities

All models are wrong. Some are more useful than others.
- George Box

- If we are asking which is the "true" model, we will have a bad time
- In practice, issues with sample size, collinearity, and available predictors are real problems
- It is often possible to differentiate between better models and less-good models, though

Basic idea for model selection

A very general algorithm

- Specify a "class" of models
- Define a criterion to quantify the fit of each model in the class
- Select the model that optimizes the criterion you're using

Again, we're focusing on $f(x)$ in the model specification. Once you've selected a model, you should subject it to regression diagnostics – which might change or augment the class of models you specify or alter your criterion.

Classes of models

Some examples of classes of models

- Linear models including all subsets of $x_1, ..., x_p$
- Linear models including all subsets of $x_1, ..., x_p$ and their first order interactions
- All functions $f(x_1)$ such that $f''(x_1)$ is continuous
- Additive models of the form $f(\mathbf{x}) = f_1(x_1) + f_2(x_2) + f_3(x_3)...$ where $f_k''(x_k)$ is continuous

Popular criteria

- Adjusted R^2
- Residual mean square error
- Akaike Information Criterion (AIC)
- Bayes Information Criterion (BIC)
- Prediction RSS (PRESS)
- F- or t-tests (stepwise selection)

Adjusted R^2

- Recall:

$$R^2 = 1 - \frac{RSS}{TSS}$$

- Definition of adjusted R^2:

$$\begin{aligned} R_a^2 &= 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)} = 1 - \frac{\hat{\sigma}^2_{model}}{\hat{\sigma}^2_{null}} \\ &= 1 - \frac{n-1}{n-p-1}(1 - R^2) \end{aligned}$$

- Minimizing the standard error of prediction means minimizing $\hat{\sigma}^2_{model}$ which in turn means maximizing R_a^2
- Adding a predictor will not necessarily increase R_a^2 unless it has some predictive value

Residual Mean Square Error

Equivalent to Adjusted R^2...

$$RMSE = \frac{RSS}{n - p - 1}$$

Can choose either based on

- the model with minimum RMSE, or
- the model that has RMSE approximately equal to the MSE from the full model

Note: minimizing RMSE is equivalent to maximizing Adjusted R^2

Sidebar: Confusing notation about p

p can mean different things

- p can be the number of covariates you have in your model (not including your column of 1s and the intercept
- p can be the number of betas you estimate.

In these slides, p is the former: the number of covariates.

AIC

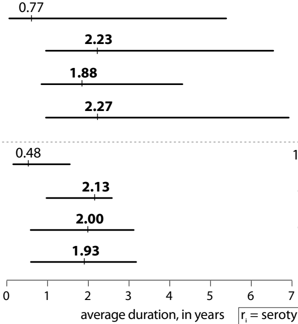
AIC ("An Information Criterion") measures goodness-of-fit through RSS (equivalently, log likelihood) and penalizes model size:

$$AIC = n \log(RSS/n) + 2(p+1)$$

- Small AIC's are better, but scores are not directly interpretable
- Penalty on model size tries to induce *parsimony*

Example of AIC in practice

		average duration of cross protection	% protected (δ)	r_i*	r(t)*	loglik	df	ΔAIC
No cross-protection	N		0			-902.4	3	0
	N_a		0			-946.7	15	112.6
	N_b		0	•		**-882.9**	**19**	**-7.0**
	N_c		0		•	**-817.2**	**41**	**-94.4**
	N_d		0	•	•	**-781.1**	**119**	**-10.7**
Exponential (λ)	E_a	0.77	100			-943.3	16	107.7
	E_b	**2.23**	**100**	•		**-873.9**	**20**	**-23.1**
	E_c	**1.88**	**100**		•	**-810.0**	**42**	**-106.9**
	E_d	**2.27**	**100**	•	•	**-773.5**	**120**	**-23.9**
Fixed duration ($k \cdot \delta$)	F_a	0.48	100 (30 - 100)			-941.8	17	106.7
	F_b	**2.13**	**84 (39 - 100)**	•		**-870.5**	**21**	**-28.0**
	F_c	**2.00**	**80 (33 - 100)**		•	**-807.3**	**43**	**-110.2**
	F_d	**1.93**	**76 (32 - 100)**	•	•	**-771.0**	**121**	**-26.8**



r_i = serotype-specific transmission parameters included
r(t) = seasonal transmission parameters included
loglik = log likelihood for the given model
df = degrees of freedom of the model
ΔAIC = change in Akaike Information Criterion over null model

Reich et al. (2013) *Journal of the Royal Society Interface*

BIC

BIC ("Bayes Information Criterion") similarly measures goodness-of-fit through RSS (equivalently, log likelihood) and penalizes model size:

$$BIC = n \log(RSS/n) + (p+1)\log(n)$$

- Small BIC's are better, but scores are not directly interpretable
- AIC and BIC measure goodness-of-fit through RSS, but use different penalties for model size. They won't always give the same answer

Bonus link! Bolker on AIC vs. BIC

Example of BIC in practice

Step	Number of Predictors in Model	Breslow's Thickness	DCCD	Ulceration	Age	Nodal Status[a]	Localization	Gender	BIC
1	7	<0.0001	0.0068	0.0009	0.0051	0.0371	0.1380	0.8052	1,657.8
2	6	<0.0001	0.0069	0.0008	0.0050	0.0340	0.1035	—	1,650.9
3	5	<0.0001	0.0011	0.0008	0.0054	0.0475	—	—	1,646.6
4	4	<0.0001	<0.0001	0.0005	0.0127	—	—	—	1,643.6
5	3	<0.0001	<0.0001	0.0002	—	—	—	—	1,642.9
6	2	<0.0001	<0.0001	—	—	—	—	—	1,649.8
7	1	<0.0001	—	—	—	—	—	—	1,679.1

p-Values are for testing whether a hazard ratio equals 1; low BIC identifies best model.
[a]As determined by routine histopathology.
doi:10.1371/journal.pmed.1001604.t004

Vasantha and Venkatesan (2014) *PLoS ONE*

PRESS

Prediction residual sum of squares is the most clearly focused on prediction

$$PRESS = \sum (y_i - \mathbf{x}_i^T \hat{\beta}_{(-i)})^2$$

Looks computationally intensive, but for linear regression models this is equivalent to

$$PRESS = \sum \left(\frac{\hat{\epsilon}_i}{1 - h_{ii}} \right)^2$$

Example of model selection in practice

TABLE 2. Results of unrestricted longitudinal latent class analysis in the Medical Research Council 1946 National Survey of Health and Development (pooled sexes, $n = 3{,}272$)

	Three classes (LLCA*-3)	Four classes (LLCA-4)	Five classes (LLCA-5)
Sequential model comparisons ($T + 1$ classes vs. T classes)	3 vs. 2	4 vs. 3	5 vs. 4
Log-likelihood value for model with $T + 1$ classes	–3,243.605	–3,211.173	–3,201.380
Log-likelihood value for model with T classes	–3,344.440	–3,243.605	–3,211.173
–2 difference in log-likelihood	201.669	64.863	19.587
Difference in no. of parameters ($T + 1$ classes vs. T classes)	7	8	8
Lo-Mendell-Rubin adjusted LRT* value	198.171	63.877	19.289
Lo-Mendell-Rubin adjusted LRT p value	<0.0001	<0.0001	0.0322
Bootstrap LRT p value	<0.01	<0.01	>0.50
Chi-square goodness-of-fit tests			
Degrees of freedom	43	36	29
LRT χ^2	123.588	58.725	39.138
p value	<0.0001	0.0098	0.0990
Bootstrap p value†	<0.01	0.02	0.11
Pearson χ^2	132.431	49.416	35.966
p value	<0.0001	0.0674	0.1746
Bootstrap p value†	<0.01	0.10	0.40
Information criterion‡			
Akaike's Information Criterion	6,527.210	6,476.347	*6,470.760*
Bayesian Information Criterion	6,649.073	*6,640.862*	6,677.927
Sample-size-adjusted Bayesian Information Criterion	6,585.524	*6,555.071*	6,569.894
Entropy	0.856	0.913	0.897
Condition number§	$0.120E^{-03}$	$0.783E^{-03}$	$0.379E^{-03}$

* LLCA, longitudinal latent class analysis; LRT, likelihood ratio test.
† Bootstrap p values were based on 200 resamples.
‡ Minimum values are shown in italic type.
§ Condition number = ratio of the largest eigenvalue to the smallest eigenvalue for the Fisher information matrix. Small values less than $10E^{-09}$ indicate problems with model identification.

Croudace et al (2003) *Amer J Epidemiology*

Model building is an art

Putting this all together requires

- knowledge of the process generating the data
- detailed data exploration
- checking assumptions
- careful model building
- patience patience patience

Sequential methods: PROCEED WITH CAUTION

Stepwise selection methods are dangerous if you want accurate inferences

- There are many potential models – usually exhausting the model space is difficult or infeasible
- Stepwise methods don't consider all possibilities
- One paper* showed that stepwise analyses produced models that...
    - represented noise 20-75% of the time
    - contained $<$50% of actual predictors
    - correlation btw predictors $\longrightarrow$ including more predictors
    - number of predictors correlated with number of noise predictors included

* Derksen and Keselman (1992) *British J Math Stat Psych*

Sequential methods: "forward selection"

- Start with "baseline" (usually intercept-only) model
- For every possible model that adds one term, evaluate the criterion you've settled on
- Choose the one with the best "score" (lowest AIC, smallest p-value)
- For every possible model that adds one term to the current model, evaluate your criterion
- Repeat until either adding a new term doesn't improve the model or all variables are included

Sequential methods: "backward selection/elimination"

- Start with every term in the model
- Consider all models with one predictor removed
- Remove the term that leads to the biggest score improvement
- Repeat until removing additional terms doesn't improve your model

MORE concerns with sequential methods

- It's common to treat the final model as if it were the only model ever considered – to base all interpretation on this model and to assume the inference is accurate
- This doesn't really reflect the true model building procedure, and can misrepresent what actually happened
- Inference is difficult in this case; it's hard to write down a statistical framework for the entire procedure
- Predictions can be made from the final model, but uncertainty around predictions will be understated
- P-values, CIs, etc will be incorrect

Variable selection in polynomial models

A quick note about polynomials. If you fit a model of the form

$$y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon_i$$

and find the quadratic term is significant but the linear term is not...

- You should still keep the linear term in the model
- Otherwise, your model is sensitive to centering – shifting x will change your model
- Using orthogonal polynomials helps with this

Sample size can limit the number of predictors

p (total number of βs) should be $< \frac{m}{15}$, where

Type of Response Variable	Limiting sample size m
Continuous	n (total sample size)
Binary	$min(n_1, n_2)$
Ordinal (k categories)	$n - \frac{1}{n^2} \sum_{i=1}^{k} n_i^3$
Failure (survival) time	number of failures

Table adapted from Harrel (2012) notes from "Regression Modeling Strategies" workshop.

A more modern approach: shrinkage/penalization

Penalized regression

- adds an explicit penalty to the least squares criterion
- keeps regression coefficients from being too large, or can shrink coefficients to zero
- Keywords for methods: LASSO, Ridge Regression
- More in Methods 3!

Much of modern statistics is devoted to figuring out what to do when $p \geq n$.