Multi-Model Inference

FW8051 Statistics for Ecologists

Department of Fisheries, Wildlife and Conservation Biology

Learning objectives

1. Gain an appreciation for challenges associated with selecting among competing models and performing multi-model inference.
2. Understand common approaches used to select a model (e.g., stepwise selection using p-values, AIC, Adjusted R^2).
3. Understand the implications of model selection for statistical inference.
4. Gain exposure to alternatives to traditional model selection, including full model inference (df spending), model averaging, and penalized likelihood/regularization techniques.
5. Be able to evaluate evaluate model performance using cross-validation and model stability using the bootstrap.
6. Be able to choose an appropriate modeling strategy, depending on the goal of the analysis (describe, predict, or infer).

Goals of multivariable regression modeling

- Describe: capture the main features of the data in a parsimonious way; which factors are related to the response variable and how?

Goals of multivariable regression modeling

- Describe: capture the main features of the data in a parsimonious way; which factors are related to the response variable and how?

- Explain: test and compare existing theories about which variables, or combination of variables, are causally related to the response variable. Regression coefficients should ideally capture how each variable affects the response variable.

Goals of multivariable regression modeling

- Describe: capture the main features of the data in a parsimonious way; which factors are related to the response variable and how?
- Explain: test and compare existing theories about which variables, or combination of variables, are causally related to the response variable. Regression coefficients should ideally capture how each variable affects the response variable.
- Predict: Use a set of sample data to make predictions about future data. It will be important to consider causality/confounding when making predictions for new populations (e.g., new spatial locations).

My Experience

Linear Models Mid-term

- Asked to develop a predictive model of Serum Albumin levels
- Given real data set (lots of predictors, missing data, etc)
- Given a weekend to complete the analysis and write up our report

My Experience

Linear Models Mid-term

- Asked to develop a predictive model of Serum Albumin levels
- Given real data set (lots of predictors, missing data, etc)
- Given a weekend to complete the analysis and write up our report

Goals were largely to demonstrate:

- Knowledge of linear models (diagnostics, model selection, etc)
- Ability to generate reliable inference

Strategy I took:

- $n = 136$ observations, split 80% (training data) and 20% (test data)

Strategy I took:

- $n = 136$ observations, split 80% (training data) and 20% (test data)
- Did lots of stuff w/ the training data
  - Grouped variables into similar categories (e.g., socio-economic status, stature [height, weight, etc], dietary intake variables)
  - Examined variables for collinearity and tried to pick the best one in a group using all possible regressions involving the group of variables
  - Fit a "best model", considered diagnostics (normality, constant variance, etc), then looked to see if any important variables were omitted

Strategy I took:

- $n = 136$ observations, split 80% (training data) and 20% (test data)
- Did lots of stuff w/ the training data
  - Grouped variables into similar categories (e.g., socio-economic status, stature [height, weight, etc], dietary intake variables)
  - Examined variables for collinearity and tried to pick the best one in a group using all possible regressions involving the group of variables
  - Fit a "best model", considered diagnostics (normality, constant variance, etc), then looked to see if any important variables were omitted
- Used the test data to evaluate predictive ability

Results

John Fieberg
Bios 163
Midterm

Final Model

Training Data Set:

NOTE: Due to missing values, only 136 observations can be used in this analysis.

Model: Albumin = $\beta 0 + \beta 1 * sciron + \beta 2 * scage + \beta 3 * scwt + \beta 4 * scwt^2 + \beta 5 * sczinc + E$

Variable Added	R^2 Model	Delta R^2	p
sciron	0.1121	0.1121	0.0001
scage	0.1840	0.0718	0.0004
scwt	0.1889	0.0050	0.7666
$scwt^2$	0.2327	0.0481	0.0036
sczinc	0.2865	0.0554	0.0022

Cross Validation Correlation R^2*(hold)=0.07643

Percent Relative Shrinkage = 100*(0.2865 - 0.0764)/0.2865 = 73%

Traditional approaches used to compare models

Nested and Non-nested Models

Nested = can get from one model to the other by setting one or more parameters = 0.

- Sleep $= \beta_0 + \beta_1 \text{Danger} + \beta_2 \text{LifeSpan}$
- Sleep $= \beta_0 + \beta_1 \text{Danger}$

Model comparisons

We can compare nested models using...

- p-values from hypothesis tests (nested models only), t-tests, F-tests, likelihood ratio tests

Model comparisons

We can compare nested models using...

- p-values from hypothesis tests (nested models only), t-tests, F-tests, likelihood ratio tests

And, nested or non-nested models using...

- Compare adjusted R^2 between competing models
- Compare model AIC values = -2*log-likelihood + 2p (nested and non-nested models; smaller is better)

Model comparisons

We can compare nested models using...

- p-values from hypothesis tests (nested models only), t-tests, F-tests, likelihood ratio tests

And, nested or non-nested models using...

- Compare adjusted R^2 between competing models
- Compare model AIC values = -2*log-likelihood + 2p (nested and non-nested models; smaller is better)

Problem 1: different (arbitrary) criteria often point to different models as "best."

Model comparisons

We can compare nested models using...

- p-values from hypothesis tests (nested models only), t-tests, F-tests, likelihood ratio tests

And, nested or non-nested models using...

- Compare adjusted R^2 between competing models
- Compare model AIC values = -2*log-likelihood + 2p (nested and non-nested models; smaller is better)

Problem 1: different (arbitrary) criteria often point to different models as "best."

Problem 2: the importance of a variable may depend on what else is in the model!

How do we choose a best model?

If we have several potential predictors, how can we decide on a best model?

How do we choose a best model?

If we have several potential predictors, how can we decide on a best model?

- We can compare a list of nested models in a structured way (adding variables or deleting variables using "stepwise selection" procedures)

How do we choose a best model?

If we have several potential predictors, how can we decide on a best model?

- We can compare a list of nested models in a structured way (adding variables or deleting variables using "stepwise selection" procedures)
- We can compare any list of models (nested or non-nested) using AIC

How do we choose a best model?

If we have several potential predictors, how can we decide on a best model?

- We can compare a list of nested models in a structured way (adding variables or deleting variables using "stepwise selection" procedures)
- We can compare any list of models (nested or non-nested) using AIC
- We can "think hard" about the problem and focus on a single most appropriate model (with others models considered as part of a "sensitivity analysis")

Traditional approaches used to compare models

Modeling Strategies

Stepwise Variable Selection

One method: Backward Elimination

- Start with a model with multiple predictors

Stepwise Variable Selection

One method: Backward Elimination

- Start with a model with multiple predictors
- Consider all possible models formed by dropping 1 of these predictors

Stepwise Variable Selection

One method: Backward Elimination

- Start with a model with multiple predictors
- Consider all possible models formed by dropping 1 of these predictors
- Keep the current model, or drop the "worst" predictor depending on:
  - p-values from the individual t-tests (drop the variable with the highest p-value, if > 0.05)
  - Adjusted R^2 values (higher values are better)
  - AIC (lower values are better)

Stepwise Variable Selection

One method: Backward Elimination

- Start with a model with multiple predictors
- Consider all possible models formed by dropping 1 of these predictors
- Keep the current model, or drop the "worst" predictor depending on:
  - p-values from the individual t-tests (drop the variable with the highest p-value, if > 0.05)
  - Adjusted R^2 values (higher values are better)
  - AIC (lower values are better)
- Rinse and repeat until you can no longer improve the model

The `stepAIC` function in the `MASS` library will do this for us.

Stepwise Variable Selection

One method: Backward Elimination

- Start with a model with multiple predictors
- Consider all possible models formed by dropping 1 of these predictors
- Keep the current model, or drop the "worst" predictor depending on:
  - p-values from the individual t-tests (drop the variable with the highest p-value, if > 0.05)
  - Adjusted R^2 values (higher values are better)
  - AIC (lower values are better)
- Rinse and repeat until you can no longer improve the model

The `stepAIC` function in the `MASS` library will do this for us.

Can also do "forward selection" (or fit all possible models "all subsets")

Change-in-estimate criterion

Heinze et al. (2017) suggest using "augmented backwards elimination"

Eliminate variables using significant testing, but only drop a variable if it does not lead to large changes in other regression coefficients.

Available in `abe` package (I have not used, but am intrigued. . .)

Data-Driven Inference

Problem 4: p-values, confidence intervals, etc assume the model has been pre-specified. Measures of fit, after model selection, will be overly optimistic.

Data-Driven Inference

Problem 4: p-values, confidence intervals, etc assume the model has been pre-specified. Measures of fit, after model selection, will be overly optimistic.

Problem 5: no guarantee that the 'best-fit' model is actually the most appropriate model for answering your question.

Stepwise selection

Harrel et al. 2001. Regression Modeling Strategies:

1. R^2 values are biased high.
2. The ordinary F and χ^2 test statistics do not have the claimed distribution
3. SEs of regression coefficients will be biased low and confidence intervals will be falsely narrow.
4. p-values will be too small and do not have the proper meaning (due to multiple comparison problems).
5. Regression coefficients will be biased high in absolute magnitude.
6. Rather than solve problems caused by collinearity, variable selection is made arbitrary by collinearity.
7. It allows us not to think about the problem.

Stepwise Selection

Copas and Long," the choice of the variables to be included depends on estimated regression coefficients rather than their true values, and so X_j is more likely to be included if its regression coefficient is over-estimated than if its regression coefficient is underestimated."

Stepwise Selection

Copas and Long," the choice of the variables to be included depends on estimated regression coefficients rather than their true values, and so X_j is more likely to be included if its regression coefficient is over-estimated than if its regression coefficient is underestimated."

Stepwise methods often select noise variables rather than ones that are truly important.

Stepwise Selection

Copas and Long," the choice of the variables to be included depends on estimated regression coefficients rather than their true values, and so X_j is more likely to be included if its regression coefficient is over-estimated than if its regression coefficient is underestimated."

Stepwise methods often select noise variables rather than ones that are truly important.

Problems get worse as you consider more candidate predictors and as these predictors become more highly correlated.

Regression Modeling Strategies (Frank Harrell's book)

Why do variable selection?

Regression Modeling Strategies (Frank Harrell's book)

Why do variable selection?

- Desire to find a 'concise' model
- Fear of collinearity influencing results
- False belief that it is not legitimate to include 'insignificant' regression coefficients.
- May increase precision by dropping unimportant variables.

df spending

Often sensible to determine how many 'degrees of freedom' you can spend, spend them, and then don't look back.

- limit model df (number of parameters) to $\leq n/10$ or $\leq n/20$, where n is your effective sample size (Harrell 2001. Regression Modeling Strategies).
- said another way, require 10-20 "events" per variable
- fit a "full model" without further simplification.

For an example, see:

Giudice, J., J. Fieberg, and M. Lenarz. 2012. Spending degrees of freedom in a poor economy: a case study of building a sightability model for Moose in northeastern Minnesota. Journal of Wildlife Management 76:75-87.

df spending

Often sensible to determine how many 'degrees of freedom' you can spend, spend them, and then don't look back.

- limit model df (number of parameters) to $\leq n/10$ or $\leq n/20$, where n is your effective sample size (Harrell 2001. Regression Modeling Strategies).
- said another way, require 10-20 "events" per variable
- fit a "full model" without further simplification.

For an example, see:

Giudice, J., J. Fieberg, and M. Lenarz. 2012. Spending degrees of freedom in a poor economy: a case study of building a sightability model for Moose in northeastern Minnesota. Journal of Wildlife Management 76:75-87.

Increases the chance that the model will fit future data nearly as well as the current data set (versus *overfitting* current data)

Effective sample size

Table A.1. Limiting sample sizes for various response variables[a].

Type of response variable	Limiting sample size, m[b]
Continuous	n (total sample size)
Binary	$\min(n_0, n_1)$
Ordinal (k categories)	$n - 1/n^2 \sum_{i=1}^{k} n_i^3$
Failure (survival) time	Number of failures

[a] © 2001 Springer-Verlag New York, Inc. Reproduced with permission from Springer Science + Business Media: Regression modeling strategies, Chapter 4: multivariable modeling strategies, 2001, page 61, Frank E. Harrell Jr., Table 4.1.

How to choose predictors

- Subject matter knowledge
- Cost/feasibility of data collection
- Relevance to your research questions
- Potential to be a confounding variable

How to choose predictors

- Subject matter knowledge
- Cost/feasibility of data collection
- Relevance to your research questions
- Potential to be a confounding variable

- Degree of missingness
- Correlation with one or more other variables
- Degree of variability (may want to combine rare categories)

Boostrapping to Evaluate Model stability

How likely are you to end up with the same model if you collect another data set of the same size and apply the same model selection algorithm.

- can report bootstrap inclusion frequencies (how often variables are selected in final models)
- may want to also report "competitive models" (others that are frequently chosen)
- don't trust a single model unless it is almost always chosen

Boostrapping to Evaluate Model stability

How likely are you to end up with the same model if you collect another data set of the same size and apply the same model selection algorithm.

- can report bootstrap inclusion frequencies (how often variables are selected in final models)
- may want to also report "competitive models" (others that are frequently chosen)
- don't trust a single model unless it is almost always chosen

Can also use the bootstrap to get a more honest measure of fit.

See: Figure 1, Fieberg and Johnson (2015)

1. Estimate regression parameters using the full data set.
2. Form bootstrapped test and bootstrapped training data sets by resampling the original data set with replacement
3. Fit the full model to the bootstrapped training data set
4. Use the model from step 3 to form predicted responses for the bootstrapped test data set.
5. Fit a linear regression model using the test data responses (as the response variable) and the predicted values from step 4 as the only explanatory variable
6. Repeat steps 2-5 many times. The average slope in step 5 can be used to proportionally reduce, or shrink, the regression parameters from step 1 towards 0; if no overfitting has occurred, the average slope in step 5 will be 1.

Regression Modeling Strategies (rms)

R package: `rms` has functions for implementing bootstrap-based model validation/calibration:

- linear regression models (using `ols` rather than `lm`)
- generalized least squares (using `Gls` rather than `gls`)
- logistic regression models (using `lrm` rather than `glm`)
- cox proportional hazards models (using `cpm` rather than `cph`)

and a few others...but not mixed models.

Regression Modeling Strategies (rms)

R package: `rms` has functions for implementing bootstrap-based model validation/calibration:

- linear regression models (using `ols` rather than `lm`)
- generalized least squares (using `Gls` rather than `gls`)
- logistic regression models (using `lrm` rather than `glm`)
- cox proportional hazards models (using `cpm` rather than `cph`)

and a few others…but not mixed models.

See in-class example for a demonstration and paper by Heinze et al. (2017) on Canvas for further discussion.

AIC and model-averaging

Rather than choose a *best* model, another approach is to average predictions among "competitive" models or models with roughly equal "support".

AIC and model-averaging

Rather than choose a *best* model, another approach is to average predictions among "competitive" models or models with roughly equal "support".

Steps:

1. Start by writing down K biologically plausible models.

AIC and model-averaging

Rather than choose a *best* model, another approach is to average predictions among "competitive" models or models with roughly equal "support".

Steps:

1. Start by writing down K biologically plausible models.
2. Fit these models and calculate AIC (trades-offs goodness-of-fit and model complexity)

AIC and model-averaging

Rather than choose a *best* model, another approach is to average predictions among "competitive" models or models with roughly equal "support".

Steps:

1. Start by writing down K biologically plausible models.
2. Fit these models and calculate AIC (trades-offs goodness-of-fit and model complexity)
3. Compute model weights, using the AIC values, reflecting "relative plausibility" of the different models.

$$w_i = \frac{\exp(-\Delta AIC_i)}{\sum_{k=1}^{K} \exp(-\Delta AIC_k)}$$

where $\Delta AIC_i = min_k(AIC_k) - AIC_i$ (difference in AIC between the "best" model and model i)

AIC and model-averaging

Rather than choose a *best* model, another approach is to average predictions among "competitive" models or models with roughly equal "support".

Steps:

1. Start by writing down K biologically plausible models.
2. Fit these models and calculate AIC (trades-offs goodness-of-fit and model complexity)
3. Compute model weights, using the AIC values, reflecting "relative plausibility" of the different models.

$$w_i = \frac{\exp(-\Delta AIC_i)}{\sum_{k=1}^{K} \exp(-\Delta AIC_k)}$$

where $\Delta AIC_i = min_k(AIC_k) - AIC_i$ (difference in AIC between the "best" model and model i)

4. Calculate weighted predictions and SEs that reflect model uncertainty and sampling uncertainty.

Model Averaging

Use AIC weights to calculate a weighted average prediction:

$$\hat{\theta}_{avg} = \sum_{k=1}^{K} w_k \hat{\theta}_k \ .$$

Model Averaging

Use AIC weights to calculate a weighted average prediction:

$$\hat{\theta}_{avg} = \sum_{k=1}^{K} w_k \hat{\theta}_k \ .$$

Calculate a standard error that accounts for model uncertainty and sampling uncertainty:

$$\widehat{SE}_{avg} = \sum_{k=1}^{K} w_k \sqrt{SE^2(\hat{\theta}_k) + (\hat{\theta}_k - \hat{\theta}_{avg})^2}$$

Model Averaging

Use AIC weights to calculate a weighted average prediction:

$$\hat{\theta}_{avg} = \sum_{k=1}^{K} w_k \hat{\theta}_k \ .$$

Calculate a standard error that accounts for model uncertainty and sampling uncertainty:

$$\widehat{SE}_{avg} = \sum_{k=1}^{K} w_k \sqrt{SE^2(\hat{\theta}_k) + (\hat{\theta}_k - \hat{\theta}_{avg})^2}$$

Typically, 95% CIs are formed using $\hat{\theta}_{avg} \pm 1.96\widehat{SE}_{avg}$, assuming that $\hat{\theta}_{avg}$ is normally distributed.

References

Burnham, Kenneth P., and David R. Anderson. Model selection and multimodel inference: a practical information-theoretic approach. Springer, 2002.

Kenneth P Burnham

FOLLOW

professor of statistics, Colorado State University
Verified email at lamar.colostate.edu

wildlife conservation biology statistics model selection capture-recapture

TITLE	CITED BY	YEAR
Practical use of the information-theoretic approach KP Burnham, DR Anderson Model selection and inference, 75-117	50113	1998

References

Burnham, Kenneth P., and David R. Anderson. Model selection and multimodel inference: a practical information-theoretic approach. Springer, 2002.

Kenneth P Burnham

FOLLOW

professor of statistics, Colorado State University
Verified email at lamar.colostate.edu

wildlife conservation biology statistics model selection capture-recapture

TITLE	CITED BY	YEAR
Practical use of the information-theoretic approach KP Burnham, DR Anderson Model selection and inference, 75-117	50113	1998

For a counter-point, see (optional reading on Canvas):

Cade, B. S. (2015). Model averaging and muddled multimodel inferences. Ecology, 96(9), 2370-2382.

And, Mark Brewer's talk on the Cult of AIC:

https://www.youtube.com/watch?v=lEDpZmq5rBw