Best practices in data analysis

Selecting a model

Stefanie Muff

Open Science course, Hjerkinn, November 2025

Introduction and overview

- “Best practices in data analysis” is a huge field.
- Today: some words on model selection.

Developing a model

In statistics courses, the correct models often “fall from heaven”: The model family and the terms in the model were almost always given.

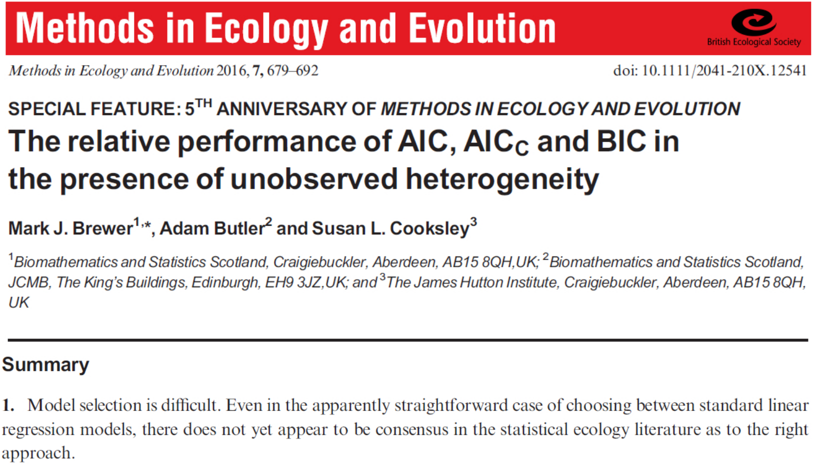
However, it is often not immediately obvious which terms are relevant to include in a model.

Importantly, the approach to find a model **heavily depends on the aim** for which the model is built.

The following distinction is important:

- The aim is to predict future values of y from known regressors.
- The aim is to explain y using known regressors. Ultimately, the aim is to *find causal relationships*.

→ Even among statisticians there is no real consensus about how, if, or when to select a model:



Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2016, **7,** 679–692 doi: 10.1111/2041-210X.12541

SPECIAL FEATURE: 5TH ANNIVERSARY OF *METHODS IN ECOLOGY AND EVOLUTION*

The relative performance of AIC, AIC$_C$ and BIC in the presence of unobserved heterogeneity

Mark J. Brewer[1,*], Adam Butler[2] and Susan L. Cooksley[3]

[1]Biomathematics and Statistics Scotland, Craigiebuckler, Aberdeen, AB15 8QH,UK; [2]Biomathematics and Statistics Scotland, JCMB, The King's Buildings, Edinburgh, EH9 3JZ,UK; and [3]The James Hutton Institute, Craigiebuckler, Aberdeen, AB15 8QH, UK

Summary

1. Model selection is difficult. Even in the apparently straightforward case of choosing between standard linear regression models, there does not yet appear to be consensus in the statistical ecology literature as to the right approach.

Note: The first sentence of a paper in *Methods in Ecology and Evolution* from 2016 is: “Model selection is difficult.’’

Why is finding a model so hard?

Note that a model is only an *approximation* to reality. The aim of a data analysis is to understand something about the real world thanks to *simplifications*.

Box (1979): **“All models are wrong, but some are useful.’ ’**

→ There is often not a “right” or a “wrong” model – but there are more and less useful ones.

→ Finding a model with good properties is sometimes an art...

Two examples

1. **Prognostic factors for body fat**

Research question: Which factors allow for precise estimation (prediction) of body fat?

Method: Study with 241 male participants. Measured variable were, among others, body fat (%), age, weight, body size, BMI, neck thickness and abdominal girth.

2. **Understanding the effect of mercury (Hg) in the soil**

Research question: Is the Hg level in the environment (soil) of people's homes associated to the Hg levels in their bodies (urin, hair)?

Method: Measurements of Hg concentrations on people's properties, as well as measurements and survey of children and their mothers living on these properties.

Predictive and explanatory models

Before we continue to discuss model/variable selection, we need to be clear about the scope of the model:

- Predictive models: These are models that aim to predict the outcome of future subjects.

  Example: In the bodyfat example the aim is to predict people's bodyfat from factors that are easy to measure (age, BMI, weight,..).

- Explanatory models: These are models that aim at understanding the (causal) relationship between covariates and the response.

  Example: The mercury study aims to understand if Hg-concentrations in the soil (covariable) influence the Hg-concentrations in humans (response).

→ The model selection strategy depends on this distinction.

Prediction vs explanation

> *When the aim is prediction, the best model is the one that best predicts the fate of a future subject. This is a well defined task and "objective" variable selection strategies to find the model which is best in this sense are potentially useful.*
>
> *However, when used for explanation the best model will depend on the scientific question being asked, and* ***automatic variable selection strategies have no place****.*

(Clayton and Hills 1993)

Model selection for predictive models

- Cross-validation (CV)
- AIC, BIC, DIC, ...
- If you want, even forward and backward selection is ok.
- Any (other) optimization procedure of some “cost function”.

Model selection for explanatory models?

Model selection may lead to biased parameter estimates, thus do not draw (biological, medical,..) conclusions from models that were optimized for prediction, for example by AIC/AICc/BIC minimization!

See, e.g., (Freedman 1983), (Copas 1983), (Berk et al. 2013).

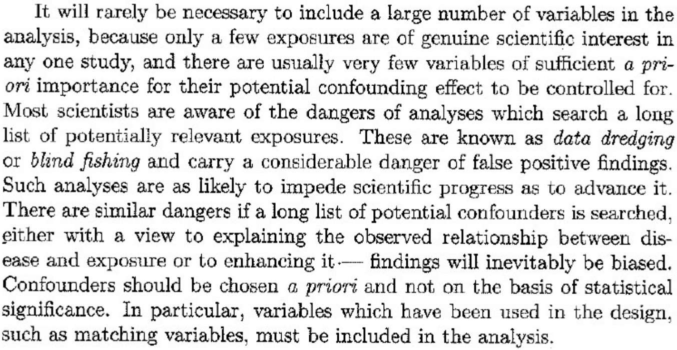
"Explanation" means that you will want to interpret the regression coefficients, 95% CIs and p-values. It is then often assumed that some sort of causality ($x \rightarrow y$) exists.

In such a situation, you should formulate a confirmatory model:

- Start with a **clear hypothesis.**
- Select your covariates according to **a priori knowledge.**
- Ideally, formulate **only one** or a few model(s) **before you start analysing your data**.

You might consider *pre-registration*.

Confirmatory models have a long tradition medicine. In fact, the main conclusions in a study are only allowed to be drawn from the main model (which needs to be specified even before data are collected):



It will rarely be necessary to include a large number of variables in the analysis, because only a few exposures are of genuine scientific interest in any one study, and there are usually very few variables of sufficient *a priori* importance for their potential confounding effect to be controlled for. Most scientists are aware of the dangers of analyses which search a long list of potentially relevant exposures. These are known as *data dredging* or *blind fishing* and carry a considerable danger of false positive findings. Such analyses are as likely to impede scientific progress as to advance it. There are similar dangers if a long list of potential confounders is searched, either with a view to explaining the observed relationship between disease and exposure or to enhancing it — findings will inevitably be biased. Confounders should be chosen *a priori* and not on the basis of statistical significance. In particular, variables which have been used in the design, such as matching variables, must be included in the analysis.

(Clayton and Hills 1993)

Model selection bias – coded example

https://htmlpreview.github.io/?https://github.com/stefaniemuff/statlearning/blob/master/OpenScience/bestPracticeAnalysis/Rcode_exercise/ModelSelectionBias.html

Aim of the example:
To illustrate how model selection purely based on AIC can lead to biased parameters and overestimated effects.

Confirmatory vs. exploratory

Any *additional analyses* that you potentially do with your data have the character of *exploratory models*.

→ Two sorts of explanatory models/analyses:

Confirmatory:

- Clear hypothesis and a priori selection of regressors for the response.
- No variable selection!
- Allowed to interpret the results and draw quantitative conclusions.

Confirmatory vs. exploratory

Any *additional analyses* that you potentially do with your data have the character of *exploratory models*.

→ Two sorts of explanatory models/analyses:

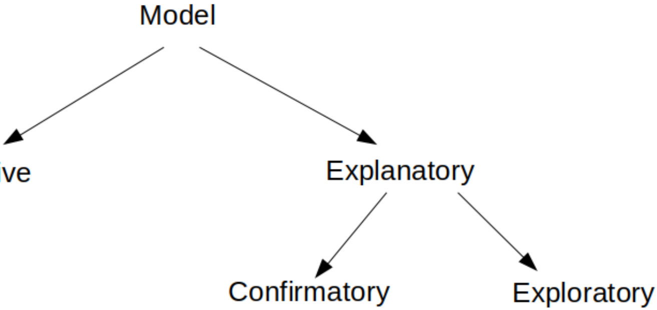
Confirmatory:

- Clear hypothesis and a priori selection of regressors for the response.
- No variable selection!
- Allowed to interpret the results and draw quantitative conclusions.

Exploratory:

- Build whatever model you want, but the results should only be used to generate new hypotheses, a.k.a. "speculations".
- Clearly label the results as "exploratory".

Summary



Model

Predictive

Explanatory

Confirmatory

Exploratory

References

Berk, Richard, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. 2013. "Valid post-selection inference." *The Annals of Statistics* 41 (2): 802–37. https://doi.org/10.1214/12-AOS1077.

Clayton, D., and M. Hills. 1993. *Statistical Models in Epidemiology*. Oxford: Oxford University Press.

Copas, J. B. 1983. "Regression, Prediction and Shrinkage." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 45: 311–54.

Freedman, D. A. 1983. "A Note on Screening Regression Equations." *The American Statistician* 37: 152–55.