

Selecting statistical models: a discussion

J. Hambuckers

Chair of Statistics
University of Göttingen (Germany)

14 June 2018, Science Campus



GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

About me

- ▶ 2006-2009: B.Sc. Business Engineering (option: mathematical modelling) at University of Liège (ULg, Belgium).
- ▶ 2009-2011: M.Sc. Financial Engineering (option: Finance), ULg.
- ▶ 2011-2015: Ph.D. in Economics and Management Science (option: Financial Econometrics), ULg.
- ▶ **Since 2016**: Postdoc researcher, Chair of Statistics and Econometrics, University of Göttingen.

Mostly a **methodological** view of statistics, in the field of finance/economics. What follows relies on my research agenda and my collaborations in the field of ecology.

Agenda

- I. Reflecting on statistical models
- II. What is model selection ?
- III. A (non-exhaustive) list of approaches
- IV. Model selection vs p-value

What is a statistical model?

Phenomena in the real world are complex...

- ▶ Baboons are resting, then start moving. What does decide their order of departure?
- ▶ Farmers allocate lands to grow various cereals. What are the factors impacting their decision?
- ▶ Populations of woolly monkeys in South America move from one rainforest to another. Which factors do influence their behaviour?
- ▶ Frauds are routinely committed inside banks. How do the sizes of the frauds relate to internal context and the environment of the bank?

Answering to these questions is relevant for decision making, policy making, forecasts or simply a general understanding of the world.

What is a statistical model?

In an ideal world, and if our intuition about these processes is correct, one could **exactly predict** what will happen.

E.g. Woolly monkeys move because of rain, altitude, food constraints, interactions with humans, age of the monkey, number of animals, how many are pregnant, if a baby monkey died recently, etc...**in a deterministic way.**

$$\text{Presence} = f(\text{food}, \text{rain}, \dots, \text{etc.})$$

"Sadly", we don't live a perfect world. **Some event cannot be perfectly predicted.** Notion of randomness ?

Random: *a characteristic associated to a phenomenon so complex that it is impossible to predict an exact outcome.*

What is a statistical model?

How do we deal with these problems anyway ?

- ▶ Assume that what we don't know can be replaced by a stochastic/random process (e.g. error terms).
- ▶ Axiomatic reduction from the unknown to the random allows for confinement of complexity.
- ▶ *Ludwig Wittgenstein*: "We use probability only in default of certainty, when our knowledge of a fact is not complete."

A statistical model is, fundamentally, **an approximation of the reality**. Nevertheless, this approximation **reduces the complexity of the phenomenon to its core components**, allowing for educated guesses.

What is a statistical model?

Practically speaking, a statistical model is a set of equations describing the stochastic nature of a phenomenon (i.e. its distribution).

Usually, it allows for linking features of this distribution (mean, variance, probability of a given event) with *explanatory variables/predictors/covariates/causes*.

E.g.: $height = f(age) + \epsilon \sim N(0, \sigma^2)$

Age is not *causing* an increase in expected height!

For theoretical reasons, estimations rely on the assumption that models are true, i.e. are the processes that generated the observed data.

Physics vs biology vs economics?

What is model selection?

Suppose

$y_i = \text{Presence of woolly monkey}_t,$

$$y_i \sim f(\mu(x_i), \sigma^2),$$

$$\mu(x_i) = h(x_i).$$

1. What should be $f(\cdot)$?
2. What should be $h(\cdot)$?
3. What should go in x_i ?

Model selection will allow to make an educated guess regarding these choices, relying on a given objective function.

A (non-exhaustive) list of approaches

Notion of **likelihood**:

Suppose that we observe a sample $\mathbf{y} = \{y_1, \dots, y_n\}$.

Under my model being the truth, what is the joint probability to observe this sample?

Given that the observed data are (conditionally) independent, the joint probability is the product of the individual probabilities.

Taking the log:

$$\mathcal{L}(\mathbf{y}; \Theta) = \sum_i^n \log(f(y_i; x_i, \Theta)),$$

where Θ is the set of all parameters related to our model. In practice, Θ needs to be estimated. Simply achieved by maximizing $\mathcal{L}(\mathbf{y}; \Theta)$ w.r.t Θ .

A (non-exhaustive) list of approaches

The higher the likelihood, the more likely we judge our model, given a realization \mathbf{y} . The difference in likelihood between two models tells us about the relative distance with the true (unknown) model.

However, as for linear regression, if we increase the number of parameters, we can always increase the likelihood. We penalize for complexity:

$$AIC = -2 * \mathcal{L}(\mathbf{y}; \hat{\Theta}) + 2 * p$$

$$BIC = -2 * \mathcal{L}(\mathbf{y}; \hat{\Theta}) + \log(n) * p$$

$$IS = -2 * \mathcal{L}(\mathbf{y}; \hat{\Theta}) + k * p$$

BIC is more stringent on complexity than AIC.

A (non-exhaustive) list of approaches

When you have a small number of models, such approaches are sufficient.

Remark: \mathbf{y} is random, as well as \mathcal{L} , AIC , BIC , IS . One should account for the variability that will arise when repeating the experiment.

For nested model: likelihood ratio tests.

For non-nested model: Vuong's test (Vuong, 1989, *Econometrica*) and following.

Other (1): bootstrap confidence interval. Resample the data, repeat the procedure, re-calculate difference in BIC.

Other (2): cross-validation. Split the data in K groups. Repeat the procedure setting each time one group aside. Average.

A (non-exhaustive) list of approaches

Remark 1: None of these approaches allow you to say that a model is good (i.e. is the true one). We only know if one model is closer than another to the true one.

Remark 2: So far we spoke about information criteria. One can imagine other criteria based on your objectives. Examples involve least squares, censored likelihood, predictive likelihood, prediction errors, classification rates, profit measures.

A (non-exhaustive) list of approaches

Common issue: the number of models to compare is potentially large.

E.g. I have k candidate explanatory variables. I want to compare all possible models based on these factors. How many models?

A (non-exhaustive) list of approaches

Common issue: the number of models to compare is potentially large.

E.g. I have k candidate explanatory variables. I want to compare all possible models based on these factors. How many models?

Answer: 2^k .

$k = 10 \rightarrow 1,024$

$k = 15 \rightarrow 32,768$

$k = 20 \rightarrow 1,048,576$

Forget about *testing*, due to the multiple comparison problem.

A common approach is to use stepwise methods. I would not recommend that as it leads often to suboptimal selection (Fan and Li, 2001, JASA).

A (non-exhaustive) list of approaches

Solutions ?

- ▶ Reduce k . Better a narrow set of well-justified candidate than ending up with ice-cream price predicting tomorrow's weather.
- ▶ Up to $k = 10$: all subset selections and deny uncertainty. If you want to do your analysis really well, apply bootstrap multiple comparison techniques (e.g. White, 2000, *Econometrica*).
- ▶ Beyond: use of shrinkage techniques (e.g. Tibshirani, 1996, *JRSSB*; Zou, 2006, *JASA*).

A (non-exhaustive) list of approaches

Terms that you might have heard: LASSO (least absolute shrinkage and selection operator), SCAD penalties.

General idea: Build a model with all candidate variables, but along the estimation, **set automatically some regression coefficients to zero.**

$$\text{LASSO: } \hat{\beta} = \arg \max_{\beta} \sum_i^n \log(f(y_i; x_i, \beta)) - \kappa \sum_{j=1}^k |\beta_j|$$

For κ appropriately chosen, we will obtain something like

$$\hat{\beta} = (1.22, 0, 0, 0.432, -0.987, 0, 0, 0, 0, 0, 0, 0, 1.655, 0)^T$$

Thus, we automatically select the relevant variable *without a direct comparison between models.*

A (non-exhaustive) list of approaches

Existing packages in R: `glmnet`, `glmlasso`, `glmmlasso`, `bamlss`.

In the Bayesian context, the same result can be achieved by choosing an appropriate prior distribution. Some techniques: Horse-shoe prior, spike-and-slab prior.

Advantage of Bayesian vs frequentist ? **Mostly inference (estimation and confidence intervals).**

Why not p-values instead?

Model selection vs p-value

- ▶ Model selection does not formally rely on one model being the truth.
- ▶ IS comparison: difference w.r.t the truth without knowing it.
- ▶ Wald tests: under the null hypothesis, the model **is** the true model.
- ▶ Rejecting Wald: the model has to be correctly specified but the regression coefficient is not equal to zero.
- ▶ Simultaneous confidence interval? Ok, but what about having a correctly specified model?
- ▶ Wald tests are highly sensitive to assumptions (see e.g. Hambuckers et al. (2018)).

Conclusion

Model selection is a hot topic in statistics right now. This is the consequence of the big data era and technological improvements.

Conclusion

Model selection is a hot topic in statistics right now. This is the consequence of the big data era and technological improvements.

Many challenges remain (random effect, beyond GLM, time series, collinearity issue...).

Conclusion

Model selection is a hot topic in statistics right now. This is the consequence of the big data era and technological improvements.

Many challenges remain (random effect, beyond GLM, time series, collinearity issue...).

A good practice consists in questioning the theoretical foundation of the set of models that are compared.

Conclusion

Model selection is a hot topic in statistics right now. This is the consequence of the big data era and technological improvements.

Many challenges remain (random effect, beyond GLM, time series, collinearity issue...).

A good practice consists in questioning the theoretical foundation of the set of models that are compared.

Inference cannot be ignored for nested models.

Conclusion

Model selection is a hot topic in statistics right now. This is the consequence of the big data era and technological improvements.

Many challenges remain (random effect, beyond GLM, time series, collinearity issue...).

A good practice consists in questioning the theoretical foundation of the set of models that are compared.

Inference cannot be ignored for nested models.

Shrinkage is the modern way to do model selection. In 10 years, every field will do it. Being among the firsts to understand these methods can give you a strong competitive advantage.

References I

- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.
- Hambuckers, J., Groll, A., and Kneib, T. (2018). Understanding the economic determinants of the severity of operational losses: A regularized generalized pareto regression approach. *Journal of Applied Econometrics* (forthcoming).
- Tibshirani, R. (1996). Regression Selection and Shrinkage via the LASSO. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 58(1):267–288.
- Vuong, Q. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57(2):307–333.
- White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5):1097–1126.

References II

Zou, H. (2006). The adaptive LASSO and its oracle properties.
Journal of the American Statistical Association,
101(476):1418–1429.