Model selection in a creative destruction setting

Raffaello Seri[1] and Massimo Rusconi[1]

[1]InsIDE Lab, DiEco, Università degli Studi dell'Insubria

2024 BTSCON, 8-10/10/2024

Outline

Outline

Introduction I

- The traditional way to verify the adequacy of a theory to the data has been to perform a single experiment in which the assumptions underlying the theory or the predictions of the theory are tested.
- While one experiment in itself is not sufficient to guarantee the acceptance or the rejection of a theory, attempts to replicate an already performed experiment have often been met with skepticism.
- Big team science has the potential to overcome this, making it possible to perform (several instances of) the same experiment in different countries and contexts and to considerably increase the sample size.
- This makes it possible to reliably test several related theories at the same time.

Introduction II

- As recently expressed, "initiatives to assess the robustness of findings [...] should aim to simultaneously test competing ideas operating in the same theoretical space" (Tierney et al., 2020, p. 291).
- This approach has been likened to the gales of creative destruction that were proposed by Schumpeter in innovation economics.
- The availability of statistical methods for the comparison of a collection of models incorporating these competing ideas is a prerequisite for this process of simultaneous verification of several theories.

Introduction III

- The aim of this presentation is to discuss these methods and make a case for information criteria as a general method to perform multi-criteria model selection in a creative destruction setting.

Caveats

- I will not discuss how models are created. Others did that, and better than I could ever do.
- I will make several simplifications that do not alter the general message.
- I will avoid mathematical details even when they simplify a lot the presentation.

Outline

Framework I

- A *(statistical) model* $\mathscr{M}$ is a mathematical description of the behaviour of one or more variables containing some unknown parameters.

A regression model - I

Consider the model defined as

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

This is the statistical model that we (implicitly) use when we estimate a simple linear regression.

- We suppose to have some data.

Framework II

A regression model - II

For the previous model, we observe the couples (x_i, y_i) for $i \in \{1, \dots, n\}$.

- We estimate the parameters.
- We have a *goodness-of-fit measure* Q_n expressing how well the model adapts to the data. Here n is the "size" of the data.

A regression model - III

For the model $y = \beta_0 + \beta_1 x + \varepsilon$, a goodness-of-fit measure is

$$Q_n = -\frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 .$$

Framework III

- We suppose to have several models that we index with $j = 1, 2, \ldots$, as in $\mathscr{M}_j$. We call $\{\mathscr{M}_1, \mathscr{M}_2, \ldots\}$ the *candidate set*. The goodness-of-fit measures are $Q_{n,1}, Q_{n,2}, \ldots$ or $Q_{n,j}$ for $j = 1, 2, \ldots$
- The objective is to look for a model that explains the data better than the other models of the candidate set, but in case of a tie we want the model with fewer parameters.
- Why the number of parameters? Because this is a measure of (absence of) *parsimony*. (Quote Occam's razor if you like.)

Framework IV

Elephants

Enrico Fermi told (to Freeman Dyson) that John von Neumann used to say that "with four parameters I can fit an elephant, and with five I can make him wiggle his trunk". (Lucien M. LeCam attributed the same sentence to Joseph Bertrand...) He meant that using too many parameters to describe a phenomenon makes the description prone to *overfitting*.

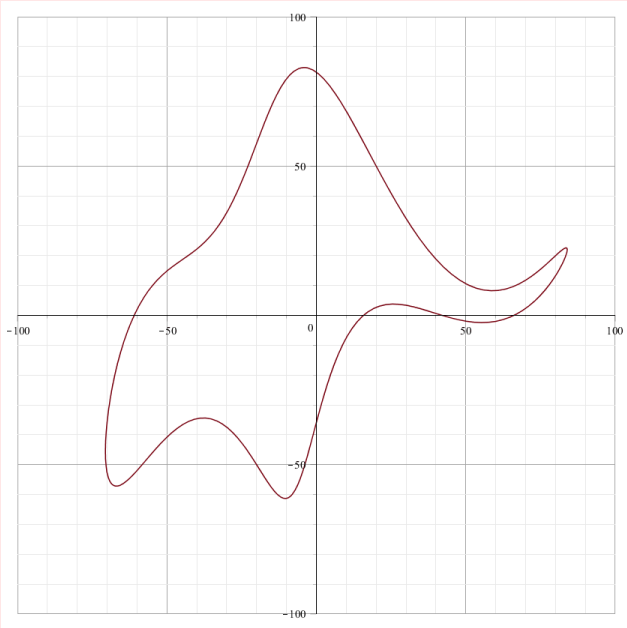

Figure: Fermi–Neumann elephant

Outline

Model Selection through Testing I

- A possible approach to model selection is through testing.
- Indeed, most restrictions of one model to another can be formulated as constraints on the parameters and testing is the most widely used way to verify the validity of parameter constraints.
- However, we need to grasp the relations between different models.

Model Selection through Testing II

Regression models - I

Consider two models:

$$\begin{aligned} \mathscr{M}_1: &\quad y = \beta_0 + \varepsilon, \\ \mathscr{M}_2: &\quad y = \beta_0 + \beta_1 x + \varepsilon. \end{aligned}$$

If we set $\beta_1 \equiv 0$ in $\mathscr{M}_2$, we get $\mathscr{M}_1$. Therefore, $\mathscr{M}_1$ is a special case of (it is included in) $\mathscr{M}_2$. If $\rightarrow$ indicates inclusion, we have

$$\mathscr{M}_1 \rightarrow \mathscr{M}_2.$$

- How do we select a model through a test?

Model Selection through Testing III

Regression models - II

To choose between the two models, we test $H_0 : \beta_1 = 0$. Then,

- if H_0 is accepted, $\beta_1 = 0$ and $\mathscr{M}_1$ is as good as $\mathscr{M}_2$ ($\mathscr{M}_1 \simeq \mathscr{M}_2$); but $\mathscr{M}_1$ is preferable to $\mathscr{M}_2$ as it achieves the same explanatory power with fewer parameters;
- if H_0 is rejected, $\mathscr{M}_2$ is better than $\mathscr{M}_1$.

- We want to extend this to more than two models.
- The first problem with this approach is the *non-transitivity of testing*.

Model Selection through Testing IV

Regression models - III

Consider three models:

$$\begin{aligned}
\mathscr{M}_1: &\quad y = \beta_0 + \varepsilon, \\
\mathscr{M}_2: &\quad y = \beta_0 + \beta_1 x + \varepsilon, \\
\mathscr{M}_3: &\quad y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon.
\end{aligned}$$

We have

$$\begin{array}{ccccc}
 & & \mathscr{M}_2 & & \\
 & \nearrow & & \searrow & \\
\mathscr{M}_1 & & \longrightarrow & & \mathscr{M}_3
\end{array} .$$

Suppose a test accepts the restriction of $\mathscr{M}_3$ to $\mathscr{M}_2$ and of $\mathscr{M}_2$ to $\mathscr{M}_1$. Nothing guarantees that the restriction of $\mathscr{M}_3$ to $\mathscr{M}_1$ is accepted! We could get $\mathscr{M}_1 \simeq \mathscr{M}_2 \simeq \mathscr{M}_3$ but $\mathscr{M}_1 \not\simeq \mathscr{M}_3$!

Model Selection through Testing V

- The second problem is the *incomparability of models*.

Regression models - IV

Let us add

$$\mathscr{M}_4: \quad \ln y = \beta_0 + \beta_1 \ln x + \varepsilon.$$

Then

$$\begin{array}{ccccc} & & \mathscr{M}_2 & & \\ & \nearrow & & \searrow & \\ \mathscr{M}_1 & & \longrightarrow & & \mathscr{M}_3 \\ & & \mathscr{M}_4 & & \end{array} .$$

Model Selection through Testing VI

- Model selection through testing depends dramatically on the relations between the models, what we call the *topology* of the candidate set.
- There are methods to arrange some of these problems one by one, but hardly all of them together.
- Do we have an alternative?

Outline

Model Selection through Information Criteria I

- A solution is to penalize the goodness-of-fit measure $Q_{n,j}$ with a function of the number of parameters, say p_j for $\mathscr{M}_j$. We speak of *information criteria*.
- Here are some examples, of the form $Q_{n,j} - p_j f(n)$:
  - *Akaike Information Criterion (AIC)*

$$\mathrm{AIC}_j = Q_{n,j} - \frac{p_j}{n};$$

  - *Bayesian Information Criterion (BIC)*

$$\mathrm{BIC}_j = Q_{n,j} - \frac{p_j \ln n}{2n};$$

  - *Hannan–Quinn Information Criterion (HQIC)*

$$\mathrm{HQIC}_j = Q_{n,j} - \frac{p_j \ln \ln n}{n}.$$

Model Selection through Information Criteria II

Regression models - VI

How do they work? One computes them for all models in the candidate set:

$$\begin{aligned}
\mathscr{M}_1: &\quad y = \beta_0 + \varepsilon, &&\quad Q_{n,1} - f(n),\\
\mathscr{M}_2: &\quad y = \beta_0 + \beta_1 x + \varepsilon, &&\quad Q_{n,2} - 2f(n),\\
\mathscr{M}_3: &\quad y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon, &&\quad Q_{n,3} - 3f(n),\\
\mathscr{M}_4: &\quad \ln y = \beta_0 + \beta_1 \ln x + \varepsilon, &&\quad Q_{n,4} - 2f(n),
\end{aligned}$$

and selects the model with the highest value.

A simplification

They are often written in a different form (e.g., $\mathrm{AIC}_j = 2p_j - 2nQ_{n,j}$) but I stick to this formulation in the following.

Model Selection through Information Criteria III

- Let us start with some history.
- Akaike Hirotsugu (赤池 弘次) proposed an *information criterion* in (1973; 1974). It was later called *AIC*, Akaike Information Criterion.

"Nice try ;-)"

In 1974, Akaike suggested AIC as an acronym of "an Information Criterion", thus nudging the scientific community into replacing "an" with "Akaike". This opened a tradition of similar moves.

- In 1978, Gideon E. Schwarz proposed *BIC*, the *Bayesian Information Criterion*, a similar criterion with a different penalization (and properties), and no nudges.
- What is the rationale behind information criteria?

Model Selection through Information Criteria IV

- First of all, they can be obtained as approximations of quantities from information theory or from statistics. However, by borrowing and adapting the expression of Gouriéroux and Monfort (1995, Vol. 2, p. 308), "[t]he foundations [...] are not fully satisfactory." E.g., each information criterion has different origins and properties and the two are not necessarily linked.
- Second, when comparing two models, they can be seen as a (kind of) test.

- But: how general is an information criterion?

Generality of Information Criteria

- We would like to avoid the following negative phenomenon, called *rank reversal* and related to non-transitivity.
  - The researcher has a selection method (test, information criterion, coin flipping, etc.).
  - There are two models, $\mathscr{M}_1$ and $\mathscr{M}_2$. The candidate set is $\{\mathscr{M}_1, \mathscr{M}_2\}$.
  - On the basis of the data, the selection method chooses $\mathscr{M}_1$ over $\mathscr{M}_2$.
  - Suppose that a new model is introduced, $\mathscr{M}_3$. The candidate set is $\{\mathscr{M}_1, \mathscr{M}_2, \mathscr{M}_3\}$.
  - It could happen that, on the basis of the data, the selection method chooses $\mathscr{M}_2$ over $\mathscr{M}_1$ and $\mathscr{M}_3$.
  - Introducing a new model leads to a reversal of the order between $\mathscr{M}_1$ and $\mathscr{M}_2$.
- Absence of rank reversal is equivalent to selection through a (generalized) information criterion.

Generality of Information Criteria

- We would like to avoid the following negative phenomenon, called *rank reversal* and related to non-transitivity.
  - The researcher has a selection method (test, information criterion, coin flipping, etc.).
  - There are two models, $\mathscr{M}_1$ and $\mathscr{M}_2$. The candidate set is $\{\mathscr{M}_1, \mathscr{M}_2\}$.
  - On the basis of the data, the selection method chooses $\mathscr{M}_1$ over $\mathscr{M}_2$.
  - Suppose that a new model is introduced, $\mathscr{M}_3$. The candidate set is $\{\mathscr{M}_1, \mathscr{M}_2, \mathscr{M}_3\}$.
  - It could happen that, on the basis of the data, the selection method chooses $\mathscr{M}_2$ over $\mathscr{M}_1$ and $\mathscr{M}_3$.
  - Introducing a new model leads to a reversal of the order between $\mathscr{M}_1$ and $\mathscr{M}_2$.
- Absence of rank reversal is equivalent to selection through a (generalized) information criterion.

Generality of Information Criteria

- We would like to avoid the following negative phenomenon, called *rank reversal* and related to non-transitivity.
  - The researcher has a selection method (test, information criterion, coin flipping, etc.).
  - There are two models, $\mathscr{M}_1$ and $\mathscr{M}_2$. The candidate set is $\{\mathscr{M}_1, \mathscr{M}_2\}$.
  - On the basis of the data, the selection method chooses $\mathscr{M}_1$ over $\mathscr{M}_2$.
  - Suppose that a new model is introduced, $\mathscr{M}_3$. The candidate set is $\{\mathscr{M}_1, \mathscr{M}_2, \mathscr{M}_3\}$.
  - It could happen that, on the basis of the data, the selection method chooses $\mathscr{M}_2$ over $\mathscr{M}_1$ and $\mathscr{M}_3$.
  - Introducing a new model leads to a reversal of the order between $\mathscr{M}_1$ and $\mathscr{M}_2$.
- Absence of rank reversal is equivalent to selection through a (generalized) information criterion.

Generality of Information Criteria

- We would like to avoid the following negative phenomenon, called *rank reversal* and related to non-transitivity.
  - The researcher has a selection method (test, information criterion, coin flipping, etc.).
  - There are two models, $\mathscr{M}_1$ and $\mathscr{M}_2$. The candidate set is $\{\mathscr{M}_1, \mathscr{M}_2\}$.
  - On the basis of the data, the selection method chooses $\mathscr{M}_1$ over $\mathscr{M}_2$.
  - Suppose that a new model is introduced, $\mathscr{M}_3$. The candidate set is $\{\mathscr{M}_1, \mathscr{M}_2, \mathscr{M}_3\}$.
  - It could happen that, on the basis of the data, the selection method chooses $\mathscr{M}_2$ over $\mathscr{M}_1$ and $\mathscr{M}_3$.
  - Introducing a new model leads to a reversal of the order between $\mathscr{M}_1$ and $\mathscr{M}_2$.
- Absence of rank reversal is equivalent to selection through a (generalized) information criterion.

Generality of Information Criteria

- We would like to avoid the following negative phenomenon, called *rank reversal* and related to non-transitivity.
  - The researcher has a selection method (test, information criterion, coin flipping, etc.).
  - There are two models, $\mathscr{M}_1$ and $\mathscr{M}_2$. The candidate set is $\{\mathscr{M}_1, \mathscr{M}_2\}$.
  - On the basis of the data, the selection method chooses $\mathscr{M}_1$ over $\mathscr{M}_2$.
  - Suppose that a new model is introduced, $\mathscr{M}_3$. The candidate set is $\{\mathscr{M}_1, \mathscr{M}_2, \mathscr{M}_3\}$.
  - It could happen that, on the basis of the data, the selection method chooses $\mathscr{M}_2$ over $\mathscr{M}_1$ and $\mathscr{M}_3$.
  - Introducing a new model leads to a reversal of the order between $\mathscr{M}_1$ and $\mathscr{M}_2$.
- Absence of rank reversal is equivalent to selection through a (generalized) information criterion.

Generality of Information Criteria

- We would like to avoid the following negative phenomenon, called *rank reversal* and related to non-transitivity.
  - The researcher has a selection method (test, information criterion, coin flipping, etc.).
  - There are two models, $\mathscr{M}_1$ and $\mathscr{M}_2$. The candidate set is $\{\mathscr{M}_1, \mathscr{M}_2\}$.
  - On the basis of the data, the selection method chooses $\mathscr{M}_1$ over $\mathscr{M}_2$.
  - Suppose that a new model is introduced, $\mathscr{M}_3$. The candidate set is $\{\mathscr{M}_1, \mathscr{M}_2, \mathscr{M}_3\}$.
  - It could happen that, on the basis of the data, the selection method chooses $\mathscr{M}_2$ over $\mathscr{M}_1$ and $\mathscr{M}_3$.
  - Introducing a new model leads to a reversal of the order between $\mathscr{M}_1$ and $\mathscr{M}_2$.
- Absence of rank reversal is equivalent to selection through a (generalized) information criterion.

Generality of Information Criteria

- We would like to avoid the following negative phenomenon, called *rank reversal* and related to non-transitivity.
  - The researcher has a selection method (test, information criterion, coin flipping, etc.).
  - There are two models, $\mathscr{M}_1$ and $\mathscr{M}_2$. The candidate set is $\{\mathscr{M}_1, \mathscr{M}_2\}$.
  - On the basis of the data, the selection method chooses $\mathscr{M}_1$ over $\mathscr{M}_2$.
  - Suppose that a new model is introduced, $\mathscr{M}_3$. The candidate set is $\{\mathscr{M}_1, \mathscr{M}_2, \mathscr{M}_3\}$.
  - It could happen that, on the basis of the data, the selection method chooses $\mathscr{M}_2$ over $\mathscr{M}_1$ and $\mathscr{M}_3$.
  - Introducing a new model leads to a reversal of the order between $\mathscr{M}_1$ and $\mathscr{M}_2$.
- Absence of rank reversal is equivalent to selection through a (generalized) information criterion.

Generality of Information Criteria

- We would like to avoid the following negative phenomenon, called *rank reversal* and related to non-transitivity.
  - The researcher has a selection method (test, information criterion, coin flipping, etc.).
  - There are two models, $\mathscr{M}_1$ and $\mathscr{M}_2$. The candidate set is $\{\mathscr{M}_1, \mathscr{M}_2\}$.
  - On the basis of the data, the selection method chooses $\mathscr{M}_1$ over $\mathscr{M}_2$.
  - Suppose that a new model is introduced, $\mathscr{M}_3$. The candidate set is $\{\mathscr{M}_1, \mathscr{M}_2, \mathscr{M}_3\}$.
  - It could happen that, on the basis of the data, the selection method chooses $\mathscr{M}_2$ over $\mathscr{M}_1$ and $\mathscr{M}_3$.
  - Introducing a new model leads to a reversal of the order between $\mathscr{M}_1$ and $\mathscr{M}_2$.
- Absence of rank reversal is equivalent to selection through a (generalized) information criterion.

Objectives of Information Criteria I

- Different information criteria have different properties.
- We can interpret these differences in light of the distinction between explanation and prediction.
  - The objective of *explanatory modeling* is "the use of statistical models for testing causal explanations" (Shmueli, 2010, p. 290).
  - *Predictive modeling* is "the process of applying a statistical model [...] to data for the purpose of predicting new or future observations" (Shmueli, 2010, p. 291).
- BIC and HQIC select with high probability the model that explains the data better than any other model of a candidate set, but in case of a tie they select the model with fewer parameters.
  - This property is called *consistency* of the selection procedure.
  - These criteria are meant for explanation.

Objectives of Information Criteria II

- AIC selects with high probability the model with the smallest prediction error.
  - This model has more parameters (is less parsimonious) than the one selected by BIC and HQIC. It is therefore prone to overfitting.
  - This property is called *conservativeness* of the selection procedure.
  - This criterion is fit for prediction.
  - As a consequence, it has some constraints: e.g., it should be used with caution when comparing models with different dependent variables (e.g., y and $\ln y$) because their predictions are not comparable.

Outline

Conclusions

- When selecting among a large candidate set of models, tests are not always the optimal tool because they have difficulties with the topology of the models.
- Information criteria solve this problem.
- However, they have other drawbacks: they do not control for errors (like tests do), there is no unified theory behind them, their generality is unclear, and they are often misused.
- In this presentation we have provided an answer to some of these questions: we have shown that information criteria are very general and we have discussed their applicability.
- The other questions will be answered in forthcoming papers and presentations.

For Further Reading I

H. Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. In B.N. Petrov and F. Csaki, editors, *Proceedings of the 2nd International Symposium on Information Theory*, pages 267–281. Akademiai Kiado, Budapest, 1973.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, December 1974. doi: 10.1109/TAC.1974.1100705.

C. Gouriéroux and A. Monfort. *Statistics and Econometric Models*. Cambridge University Press, 1995.

G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, March 1978. doi: 10.1214/aos/1176344136.

G. Shmueli. To Explain or to Predict? *Statistical Science*, 25(3): 289 – 310, 2010. doi: 10.1214/10-STS330.

For Further Reading II

W. Tierney, J. H. Hardy, C. R. Ebersole, K. Leavitt, D. Viganola, E. G. Clemente, M. Gordon, A. Dreber, M. Johannesson, T. Pfeiffer, Hiring Decisions Forecasting Collaboration, and E. L. Uhlmann. Creative destruction in science. *Organizational Behavior and Human Decision Processes*, 161:291–309, 2020. doi: 10.1016/j.obhdp.2020.07.002.