

COS-D407. Scientific Modeling and Model Validation

Lecturer: Christina Bohk-Ewald

Week 7

University of Helsinki, Finland
26.10.2020–09.12.2020

Seventh week's class:

Scientific modeling & model validation in practice

- Q&A: recap of material of previous session
- Present and discuss your findings of previous lab session
- Toolbox for selecting suitable methods & for assessing model's performance with respect to explaining and predicting phenomena
- Course summary and writing the report

Seventh week's class in the lab:

Toolbox for selecting and assessing suitable methods.

- To be continued and completed: select and assess suitable model for predicting IFR_x starting from exponential model of Levin et al. (2020).
- Present and interactively discuss how to validate your research using, perhaps, new methods of this course.

Seventh week's class in the lab:

For seventh week's lab session, please prepare a brief description
of one of your research projects
(e.g., Bachelor or Master thesis)
and tell how you have evaluated your research findings so far
and how you would, perhaps, extend it.

Brief Q&A: recap material of previous session

- When it is about predicting the response of an outcome variable, what could possibly be wrong with fitting a model to *all* observations?
- Why could splitting *all* raw data into training data and testing data be useful when fitting a model?
- What is the procedure of the validation set approach?
- Bias-variance trade-off: what is meant by low bias?
- Bias-variance trade-off: what is meant by low variance?
- What is meant by overfitting?

→ Open questions?

Present your findings of previous lab session:

Six models (M1-M6) for predicting IFR by age. Which one of the six models is most suitable for predicting IFR_x based on:

- *all* raw data and R-squared?
- *all* raw data and MSE?
- training data and testing data adopting the validation set approach?

→ Open issues?

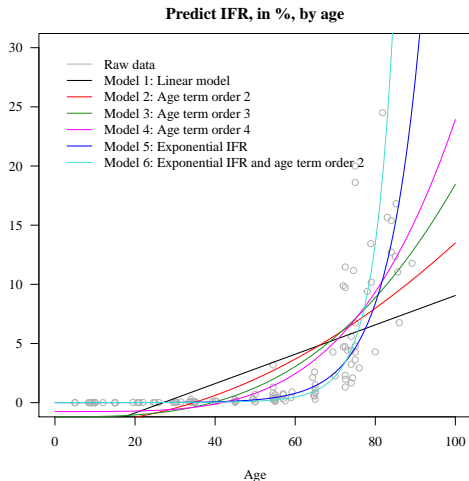
Toolbox for selecting and assessing methods

Application to select model

for predicting COVID-19-related infection fatality rates by age

based on data provided by Levin et al. (2020).

Model IFR estimates of Levin et al.



Based on *all* raw data:

- M5 has the largest R-squared
- M4 has the smallest MSE

Model IFR estimates of Levin et al.

It is not so much about finding the model
that fits best to all the observed data.

It is rather about finding the model
that predicts best IFR by age for data we do not know yet
(→ machine learning; generalization of underlying pattern).

⇒ Following this line of thinking, raw data should be split
into *training* data and *testing* data

Training data and testing data

Split raw data into *training* data and *testing* data using, e.g.,:

- Validation set approach
- k-fold cross validation
- ...

IFR estimates of Levin et al.

At first, we used the validation set approach
to select the best (of the six) models
for predicting the IFR by age.

Training data and testing data

Validation set approach:

- ➊ Randomly split all data into two parts: training data and testing data
- ➋ Fit models on training data to predict IFR by age
- ➌ Apply fitted models on testing data to predict IFR by age
- ➍ Calculate MSE between observed and predicted IFRs of testing data
- ➎ Select model with the smallest test MSE
- ➏ Could repeat entire procedure multiple times to get average test MSE

Validation set approach

Remember, it is about finding a suitable model with low bias and low variance that does not overfit training data.

Bias-variance trade-off and overfitting

- Low train MSE indicates that a model fits well the training data.
- Low test MSE indicates that a model is suitable to predict IFR by age with small or low bias for these particular training and testing data.
- Low mean test MSE indicates that a model is suitable to predict IFR by age with small or low bias across various testing data (when fitted to various training data).

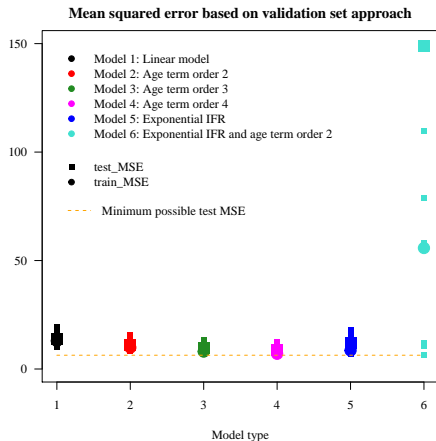
Bias-variance trade-off and overfitting

- Low variance indicates that a model's predictions of IFR by age would not change much even though the model was fitted to different training data.
- Low variance in test MSE indicates that a model's predictions of IFR by age are similarly accurate even though the model was fitted to different training data.

Bias-variance trade-off and overfitting

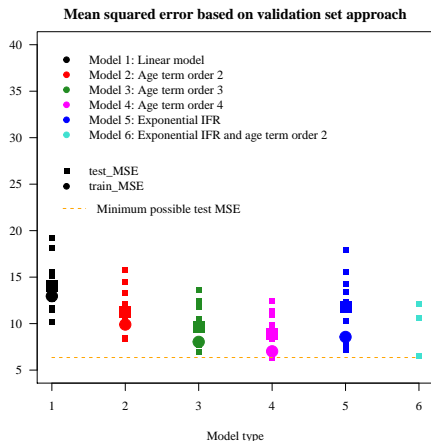
- Small gap between (mean) train MSE and (mean) test MSE can indicate that a model does not overfit the training data; i.e., that the model captures well the underlying or regular pattern but not the random noise.

IFR estimates of Levin et al.



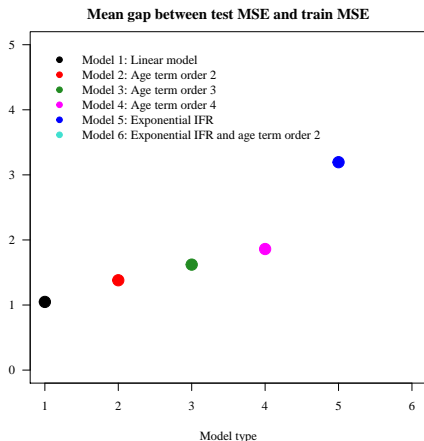
- Validation set approach applied 10 times
- M6 is not suitable to predict IFR by age

IFR estimates of Levin et al. — zoom in to better compare M1–M5



- Validation set approach applied 10 times
- Mean test MSE is consistently larger than mean train MSE
- Mean test MSE is smallest for M4 (low bias)
- M3-M5 are all close to minimum possible test MSE
- Test MSE varies stronger for M5 than for M4 (→ does M5 tend to overfit training data?)

IFR estimates of Levin et al. — zoom in to better compare M1–M5



- Gap between mean train MSE and mean test MSE tends to increase with model complexity (→ overfitting)
- Smallest gap for M1, largest gap for M5

Putting this together we can select a model based on...

Too few raw data (134) for validation set approach?

Let us try k-fold cross validation.

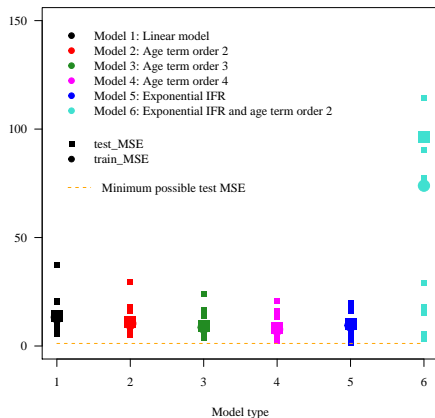
Training data and testing data

k-fold cross validation:

- 1 Systematically split all data into k parts
- 2 In each trial, hold out one part of all data to define testing data and use remaining data as training data
- 3 Fit models on training data to predict IFR by age
- 4 Apply fitted models on testing data to predict IFR by age
- 5 Calculate MSE between observed and predicted IFRs of testing data
- 6 Repeat this procedure until each part (of all k parts; step 1) has been hold out once and calculate average test MSE: $\frac{1}{k} \sum_{i=1}^k \text{testMSE}_i$
- 7 Select model with the smallest average test MSE

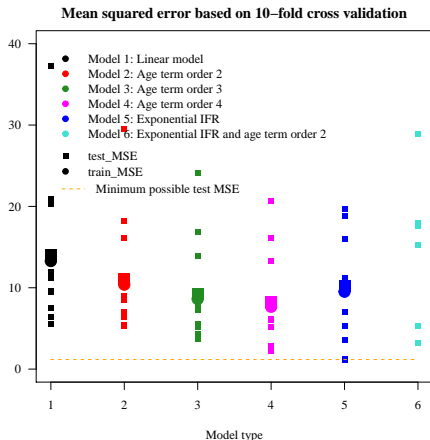
IFR estimates of Levin et al.

Mean squared error based on 10-fold cross validation



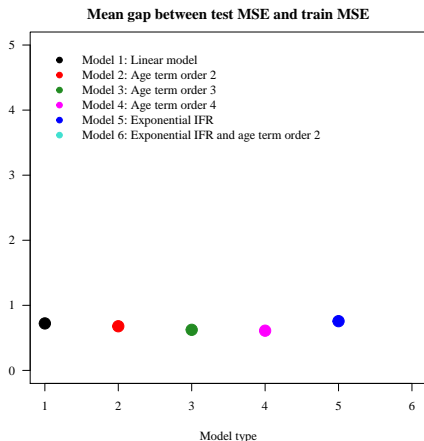
- 10-fold cross validation
- M6 is not suitable to predict IFR by age

IFR estimates of Levin et al. — zoom in to better compare M1–M5



- 10-fold cross validation
- Mean test MSE is consistently larger than mean train MSE
- Mean test MSE is smallest for M4 (low bias)
- M5 has minimum possible test MSE
- Test MSE varies similarly for M4 and M5

IFR estimates of Levin et al. — zoom in to better compare M1–M5



- Gap between mean train MSE and mean test MSE is smaller in 10-fold cross validation than in validation set approach for all models
- Gap appears to depend less on model complexity in 10-fold cross validation
- M4 has smallest gap

⇒ 10-fold cross validation appears to be more suitable than validation set approach for selecting most suitable model for predicting IFR by age

Predicting IFR by age...

So, what is the most suitable model of our six models, M1 through M6, for predicting IFR by age?

Selecting a suitable model

- Selecting a model based on *all* raw data does not reflect real-world applications, which are about making predictions facing new data
 - That is why we select a model based on training data and testing data
 - 10-fold cross validation provides enough data to first train and then test each model
 - Perhaps too few data points (134) for validation set approach
 - We look for a model with low bias and low variance that does not overfit training data
 - M4 appears to have the lowest mean test MSE, comparably low variance in test MSEs, and comparably small gap between mean train MSE and mean test MSE
- ⇒ **But does that mean that M4 is more suitable than M5, which has been chosen by Levin et al. (2020)?**

Selecting a suitable model

- Select a model with low bias and low variance that does not overfit training data
 - Applies most to M4, but also to, e.g., M5
 - Mechanistic, data-driven perspective
- Select a model with *theroretical explanation or meaning*
 - Considering that mortality is often modeled to increase exponentially with age, M5 might be more suitable for predicting IFR by age than M4 from this perspective?
- Select a simple model (over more complex models)
 - Occam's razor (or law of parsimony). If two models make predictions that are similarly accurate, select the simpler model (as it is more testable → scientific method ⇒ course weeks 1 & 2).
 - M4: age term order 4 versus M5: exponential function?
- ...

Selecting a suitable model

- Select a model with low bias and low variance that does not overfit training data
- Select a model with *theroretical explanation or meaning*
- Select a model that is simple (over more complex models)
- ...

→ Repeat and replicate this analysis in order to get rid of potential errors (e.g., due to computation, implementation, reporting) and to, perhaps, account for new data (→ representative of many possible cases)

⇒ There is no simple answer for selecting the most suitable model, but it is important to be aware of the various issues to consider when selecting a suitable model for the task at hand!

Selecting a suitable model

- Select a model with low bias and low variance that does not overfit training data
- Select a model with *theroretical explanation or meaning*
- Select a model that is simple (over more complex models)
- ...

→ Repeat and replicate this analysis

⇒ There is no simple answer for selecting the most suitable model, but it is important to be aware of the various issues to consider when selecting a suitable model for the task at hand!

⇒ Please do not just select a model because it is available somehow. There is more to think of and more to do here. → [Link to course material of previous weeks.](#)

Selecting a suitable model

Link to course material of course weeks 3, 4, and 5 in this context:

Please do not just select a model because it is available somehow. Try to also think critically and creatively about, e.g.,:

- a model's key assumptions and to what extent they may hold in the real world
- the quality of the data used to train and test a model (if applicable)
- the sensitivity of a model's results with respect to changes in input parameters
- ...

in order to fully understand the meaning and the quality of a model's results; what factors might impair a model's findings; and in what situations a model might be more suitable than in others.

→ Only then you will have good reasons for selecting a particular model to be suitable for a specific task at hand

Selecting a suitable model

But how to weigh all these different pieces of information in order to eventually make a reasonable decision on a suitable model?

Link to course material of course weeks 1 and 2:

Being aware of and considering all these different issues and pieces of information could lead to different conclusions when selecting a suitable method for a specific task at hand.

For example, the meaning and the importance of each of these issues and pieces of information (i) could be evaluated differently (→ Paul Feyerabend) and (ii) could change over time depending on, e.g., the current state of knowledge and scientific progress (→ Thomas S. Kuhn; paradigm).

⇒ There is no easy choice of selecting a suitable method; this choice will always require you to think critically and creatively and to test rigorously.

What you have learned today about selecting and assessing a suitable model

- Describe the idea to split observations into training data and testing data.
- Explain and compare validation set approach and k-fold cross validation.
- Describe the idea behind looking for a model with low bias and low variance that does not overfit training data.
- Explain what criteria can be considered in order to select a suitable model.

Course learning materials

Course learning materials on GitHub:

<https://github.com/christina-bohk-ewald/2020-COS-D407-scientific-modeling-and-model-validation>

Course summary by week

- Weeks 1 & 2: Introduction to science and the scientific method, and to the role of scientific modeling & model validation within the scientific process from a broad (scientific) perspective.
- Weeks 3, 4 & 5: Introduction to a statistical model for estimating COVID-19 infections and to strategies for assessing its outcome (even though *true* values are not available to compare the outcome to).
- Weeks 6 & 7: Introduction to a toolbox of classical concepts and tools for selecting statistical models (e.g., bias-variance tradeoff) and to assess their performance (e.g., cross-validation).

What is expected of you

Active participation in class to deeply understand principles and practices in scientific modeling and model validation.

Feedback, assessment, and grading

...are in alignment with key learning contents and learning activities:

- Formative feedback during course to discover and close learning gaps
- Summative assessment to generate course grade
 - ▶ Actively participating in class, and presenting and interactively discussing findings of your and other course participants' hands-on exercises. (40%)
 - ▶ Report that covers core topics of this course: scientific method, scientific modeling, and model validation. Report should briefly summarize the hands-on exercises and put them into a broader context; it should be approximately 3000 words long (not counting references, figures, tables, and R-code). You are supposed to write your report during this course, however, it will be due on December 18 / 21, 2020. (60%). → more information in week 7

The course report

- Should show your understanding of and reflections on each of the core course topics: purpose of science, scientific method, and selecting and assessing models regarding their suitability for explaining / predicting a phenomenon.
- Should show your familiarity with applying the introduced concepts and tools using real-world data in R (e.g., hands-on exercises).
- Should be approximately 1500–3000 words long (not counting references, figures, tables, and R-code).
- Could be written in form of a paper, summary, or course journal—that is up to you.
- Should be submitted as pdf, but can be written in any software such as Word, L^AT_EX, or R Markdown.
- Is to be submitted until December 18, 2020 (for students who are completing their studies on Dec 18 or according to the old degree requirements) or December 21, 2020 (all other students), by email to `christina.bohk-ewald@helsinki.fi`.

Some words on R Markdown

- Website: <https://rmarkdown.rstudio.com/>
- Allows you to combine your analysis in R with narrative text and to make it fully reproducible.
- For example, the hands-on exercises for this course have been written with R Markdown.
- You would need to generate a *Rmd* file (→ special syntax) in R. Running this *Rmd* file in R with the function `render()` would require you first to load the R-packages *rmarkdown* and *knitr* and having installed *pandoc* on your computer.
- There are plenty tutorials freely available online and you can just have a look at them if you wish.

⇒ There is no obligation or bonus for you using R Markdown for writing your course report; it is just to make you aware of this tool as it may become useful for you at some time.

Thank you for your time and attention!

`christina.bohk-ewald@helsinki.fi`

Seventh week's class in the lab:

Toolbox for selecting and assessing suitable methods.

- To be continued and completed: select and assess suitable model for predicting IFR_x starting from exponential model of Levin et al. (2020).
- Present and interactively discuss how to validate your research using, perhaps, new methods of this course.

Seventh week's class in the lab:

For seventh week's lab session, please prepare a brief description
of one of your research projects
(e.g., Bachelor or Master thesis)
and tell how you have evaluated your research findings so far
and how you would, perhaps, extend it.

Upcoming courses of Center for Social Data Science

Overview of courses in 2020/2021 can be found at
<https://www.helsinki.fi/en/faculty-of-social-sciences/centre-for-social-data-science-csds#section-70055>.

For example,

- Applied Research Project (Christina Bohk-Ewald, period 4)
- Perspectives to Digitalizing Society (Krista Lagus, period 4)
- Survey Methodology (Maria Valaste, period 3)
- Exploratory Methods of Multivariate Data Analysis (Kimmo Vehkalahti, period 3)
- Introduction to Atlas (Digital Constructive and Experimental Methods for Social Sciences) (Matti Nelimarkka, period 3)
- Social Data Science PhD Seminar I+II (CSDS, all 4 periods)

⇒ You are welcome to attend; also to our open CSDS Coffee Friday, e.g., if you want to talk about a research problem :-)