COS-D407. Scientific Modeling and Model Validation

Lecturer: Christina Bohk-Ewald

Week 7

University of Helsinki, Finland
01.11.2021–15.12.2021

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 & assessing methods continued: cross validation
- Summary and course journal

Seventh week's class in the lab: Toolbox for selecting and assessing suitable methods.

- Self-study: select a suitable model for predicting IFR_x using k-fold cross validation.
- Present and interactively discuss how to validate your research using, perhaps, new methods & concepts 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?
- Bias-variance trade-off: where to place a suitable model?

$\rightarrow$ 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?

$\rightarrow$ Open issues?

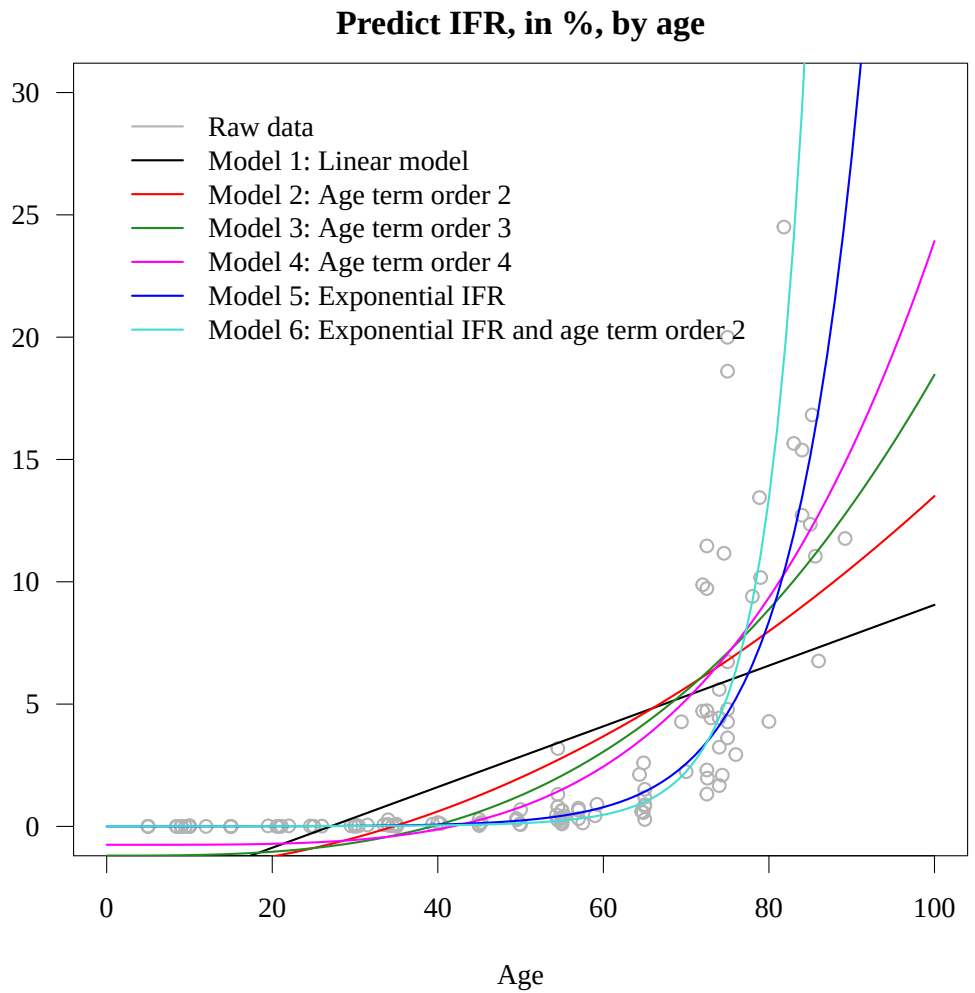
Toolbox for selecting and assessing methods

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 and M6 look suitable from eye-balling
- M5 has the largest R-squared value
- 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
($\longrightarrow$ machine learning; generalization of underlying pattern).

$\Rightarrow$ 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:

1. Randomly split all data into two parts: training data and testing data
2. Fit models on training data to predict IFR by age
3. Apply fitted models on testing data to predict IFR by age
4. Calculate MSE between observed and predicted IFRs of testing data
5. Select model with the smallest test MSE
6. Could repeat entire procedure multiple times to get average test MSE

Validation set approach

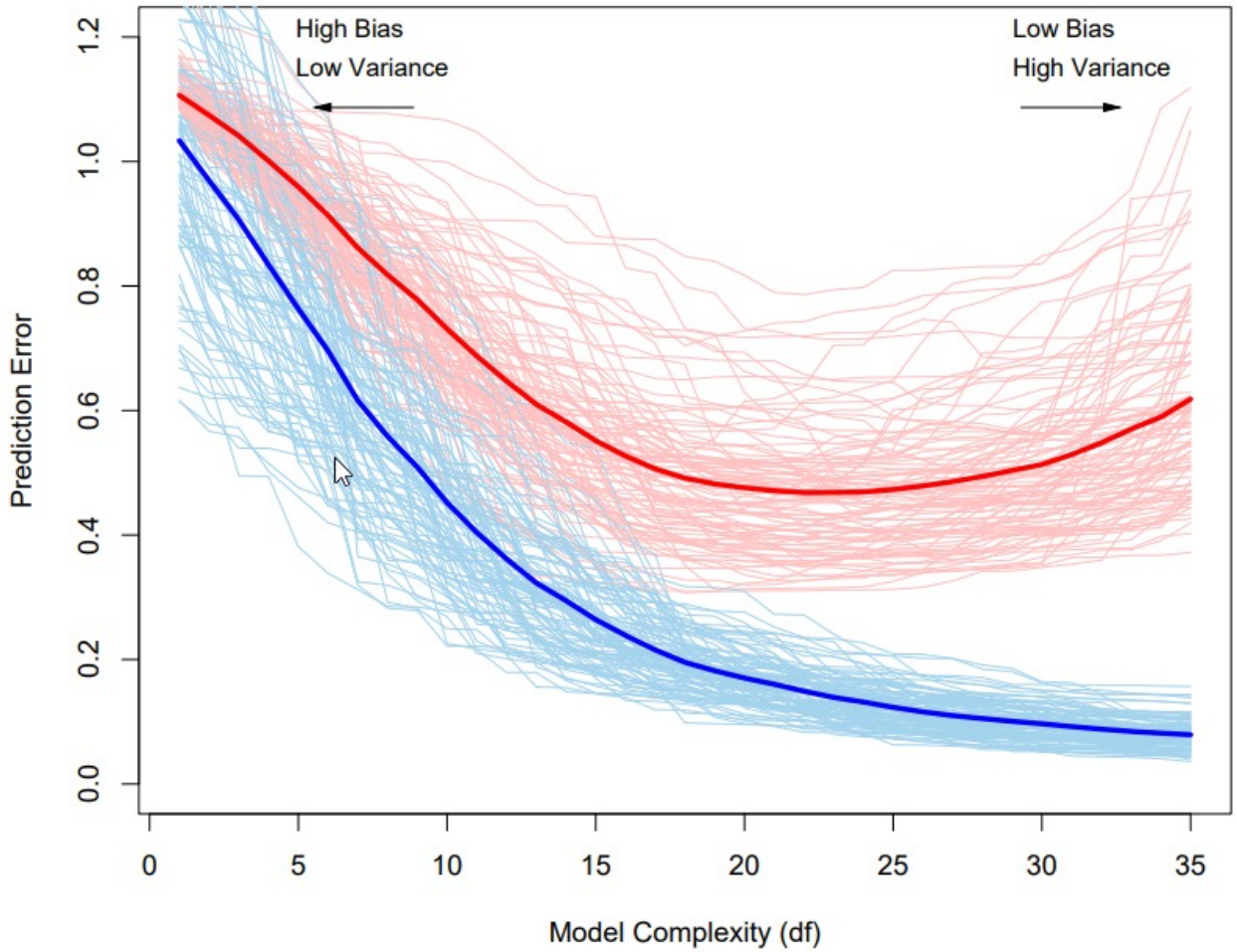
Remember, it is about finding a suitable model with comparably *low bias* and *low variance* that neither *underfits* nor *overfits* training data.

Bias-variance trade-off

- *Bias* is about the model-immanent deviation from the truth. It refers to the error that is introduced by approximating a real-life problem by a much simpler model.
- *Variance* is about the model dependence from training data. It refers to the amount by which the fitted model function (and its predictions) would change if different training data were used.
- A suitable model should have *low bias* and *low variance*. That is, it should approximate complex reality well and it should catch underlying patterns in the data (and not random noise) in order to produce accurate and robust predictions.

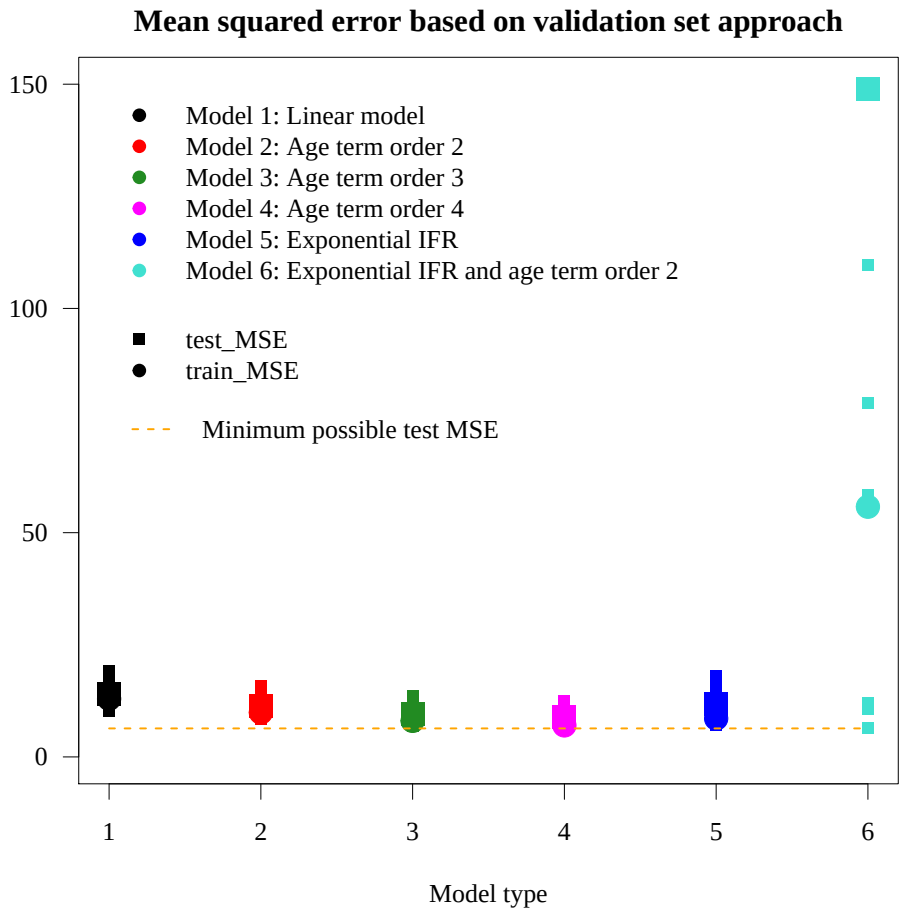
James et al. (2013; pages 29–36;
`https://hastie.su.domains/ISLR2/ISLRv2_website.pdf`**)**

Bias-variance trade-off



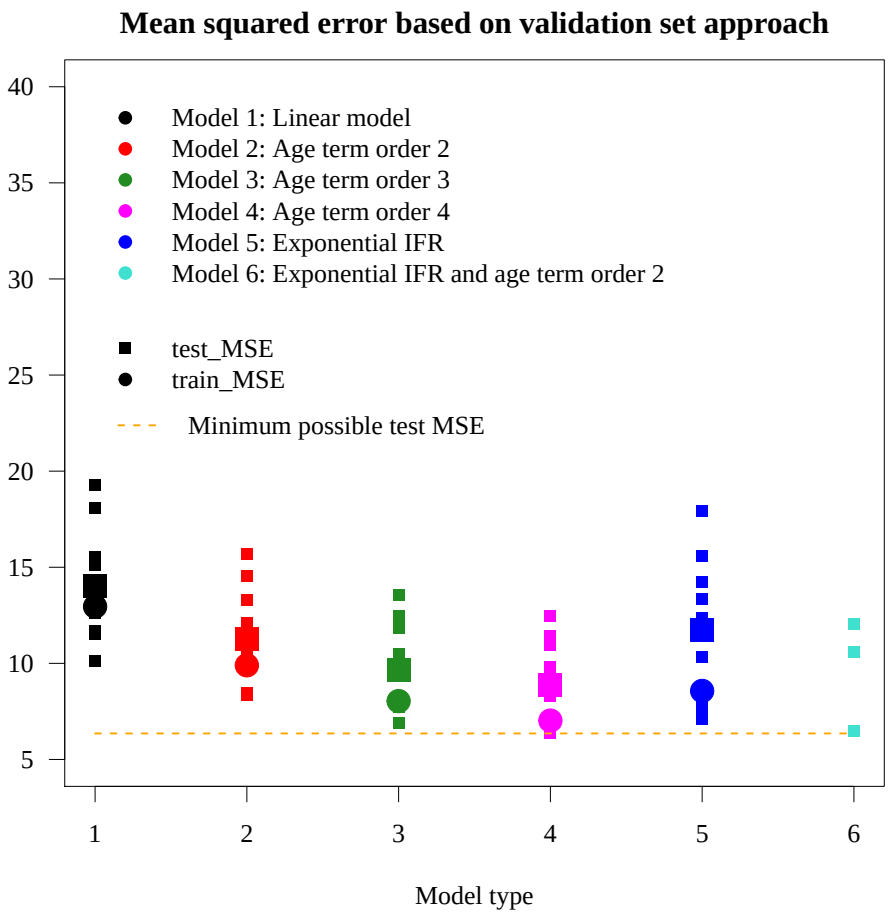
Hastie et al. (2009; page 220; Figure 7.1;
`https://hastie.su.domains/Papers/ESLII.pdf`**)**

IFR estimates of Levin et al.



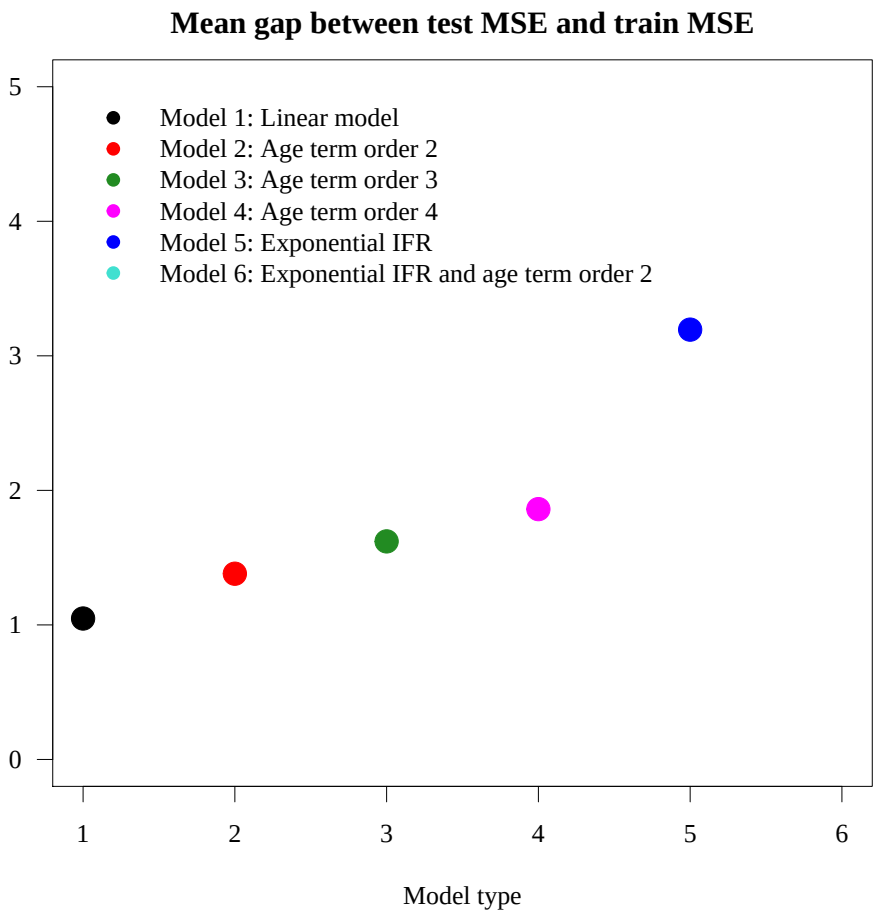
- Validation set approach applied 10 times
- M6 is not suitable to predict IFR by age
  - M6 strongly depends on training data so that its variance is too high

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 ($\rightarrow$ 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 ($\longrightarrow$ overfitting)
- Largest gap for M5

Let us try k-fold cross validation...

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

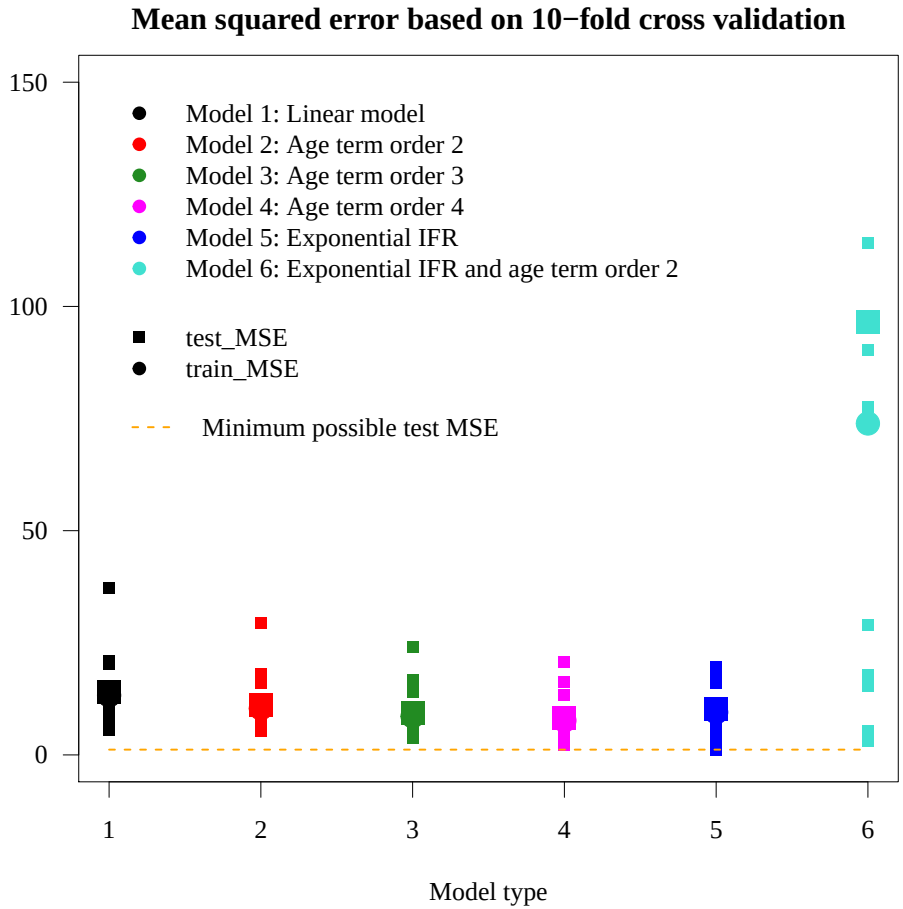
Let us try k-fold cross validation.

Training data and testing data

k-fold cross validation:

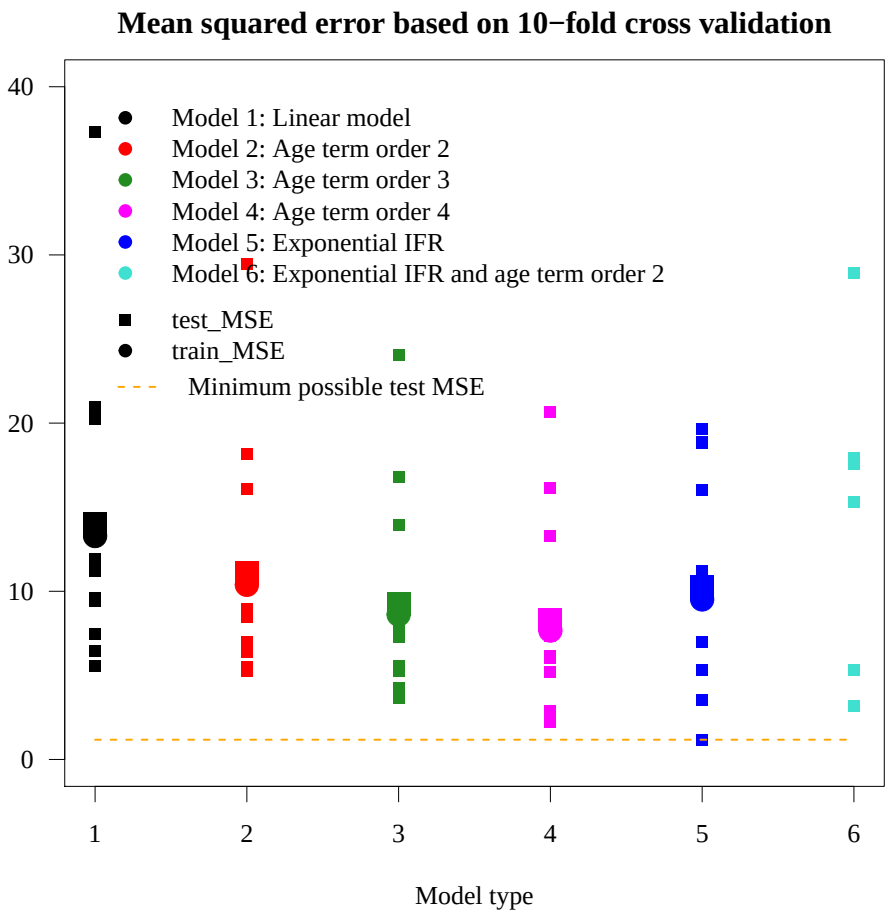
1. Systematically split all data into k parts
2. In each trial, hold out one part of all data to define testing data and use remaining data as training data
3. Fit models on training data to predict IFR by age
4. Apply fitted models on testing data to predict IFR by age
5. Calculate MSE between observed 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: $\frac{1}{k} \sum_{i=1}^{k} testMSE_i$
7. Select model with the smallest average test MSE

IFR estimates of Levin et al.



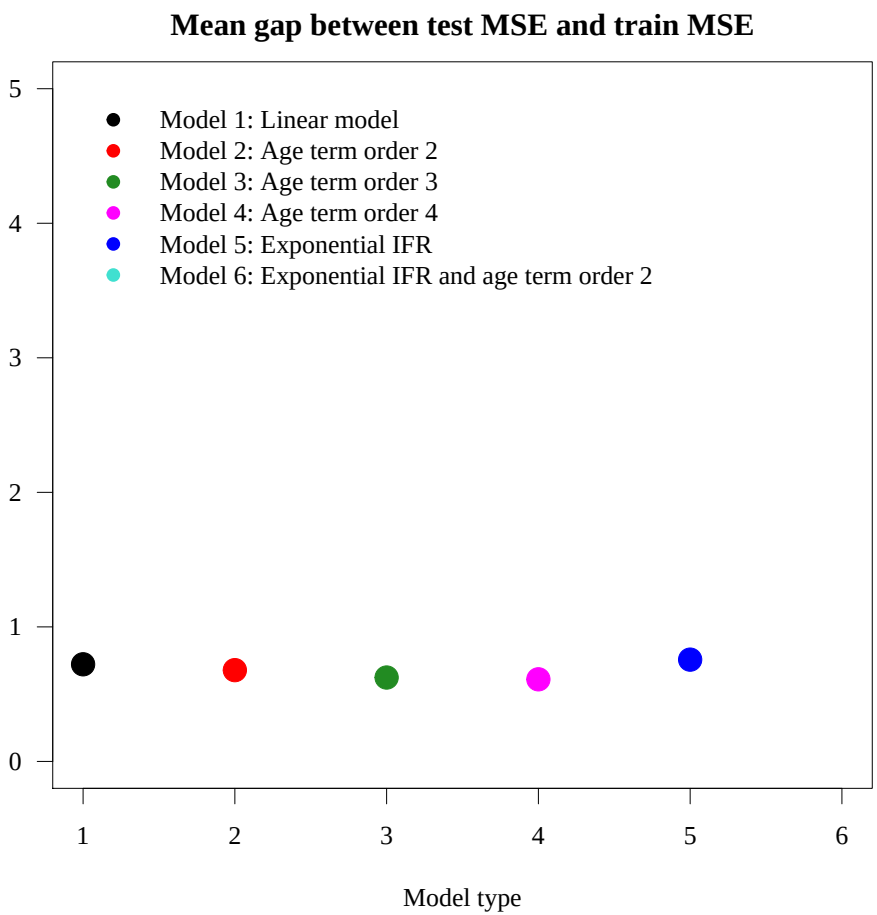
- 10-fold cross validation
- M6 is not suitable to predict IFR by age

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

$\Rightarrow$ 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

  $\longrightarrow$ 10-fold cross validation provides enough data to first train and then test each model

  $\longrightarrow$ 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

  $\longrightarrow$ 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

  $\Rightarrow$ **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

  $\longrightarrow$ Applies most to M4, but also to, e.g., M5

  $\longrightarrow$ Mechanistic, data-driven perspective

- Select a model with *theroretical explanation or meaning*

  $\longrightarrow$ 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)

  $\longrightarrow$ 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 $\longrightarrow$ scientific method $\Rightarrow$ course weeks 1 & 2).

  $\longrightarrow$ 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)
- ...

$\rightarrow$ 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 ($\rightarrow$ representative of many possible cases)

$\Rightarrow$ 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)
- ...

$\rightarrow$ Repeat and replicate this analysis

$\Rightarrow$ 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!

$\Rightarrow$ Please do not just select a model because it is available somehow. There is more to think of and more to do here. $\rightarrow$ Link to course material of previous weeks.

Selecting a suitable model

Link to course material of previous weeks 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.

$\rightarrow$ 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 ($\rightarrow$ Paul Feyerabend) and (ii) could change over time depending on, e.g., the current state of knowledge and scientific progress ($\rightarrow$ Thomas S. Kuhn; paradigm).

$\Rightarrow$ 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.

Recommended learning material for today's class

- **Hastie T, Friedman J, Tibshirani R**
  The Elements of Statistical Learning.
  Springer, New York, 2009
  DOI: https://doi.org/10.1007/978-0-387-84858-7
  `https://hastie.su.domains/Papers/ESLII.pdf`
- **James G, Witten D, Hastie T, Tibshirani R**
  An Introduction to Statistical Learning with Applications in R.
  Springer Science+Business Media New York 2013
  `https://hastie.su.domains/ISLR2/ISLRv2_website.pdf`
- **Levin et al. (2020)**
  Assessing the Age Specificity of Infection Fatality Rates for COVID-19: Systematic Review, Meta-Analysis, and Public Policy Implications.
  medRxiv 2020; published online July 24
  `https://doi.org/10.1101/2020.07.23.20160895`

Course learning materials

Course learning materials on GitHub:

`https://github.com/christina-bohk-ewald/2021-COS-D407-scientific-modeling-and-model-validation`

Course content 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.

- Week 3: Introduction to human mortality forecasting and how to systematically evaluate the performance of forecast methods using cross validation and *ex post* forecast errors. Guest lecturer: Ricarda Duerst.

- Weeks 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 trade-off) and to assess their performance (e.g., cross-validation).

The course journal

- Should show your understanding of and reflections on each of the core topics by course week: what have you learned about, e.g., purpose of science, scientific method, and selecting and assessing models regarding their suitability for explaining / predicting a phenomenon?
- If you like, you can include highlight-figures of the hands-on exercises in an appendix.
- The course journal should be approximately 1 400 words long; approximately 200 words per course week.
- Should be submitted as pdf, but can be written in any software such as Word, LaTeX, or `R Markdown`.
- Is to be submitted until December 16, 2021 by email to `christina.bohk-ewald@helsinki.fi`.

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.

- Self-study: select a suitable model for predicting IFR_x using k-fold cross validation.
- Present and interactively discuss how to validate your research using, perhaps, new methods & concepts 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.