COS-D407. Scientific Modeling and Model Validation

Lecturer: Christina Bohk-Ewald

Week 6

University of Helsinki, Finland
26.10.2020–09.12.2020

Sixth week's class:

Scientific modeling & model validation in practice

- Q&A: recap of material of previous session
- Present your findings of previous lab session
- Validity & sensitivity of the demographic scaling model's COVID-19 infection estimates, continued and completed
- Toolbox for selecting suitable methods & for assessing model's performance with respect to explaining and predicting phenomena

Sixth week's class in the lab: Sensitivity of demographic scaling model's results & toolbox for selecting and assessing suitable methods.

- Analyze the sensitivity of the demographic scaling model's results for Finland with respect to IFR_x & D_x together.
- Select and assess suitable model for predicting IFR_x starting from exponential model of Levin et al. (2020).

$\rightarrow$ Present and discuss your findings in class at the beginning of the next session on Monday.

Seventh week's class in the lab: toolbox for selecting and assessing methods

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

- What different sources of D_x estimates do you know of?
- What are their pros and cons?
- How does the age profile of COVID-19-related death counts looks like?

$\rightarrow$ Open questions?

Present your findings of previous lab session:

- How large are the COVID-19 infection estimates for Finland based on different sources of the D_x most recently?
- Have you done the same analysis for another country? Do the results differ?
- Have you analyzed the combined impact of IFR_x & D_x on the demographic scaling model's results?

$\rightarrow$ Open issues?

Some more thoughts on this

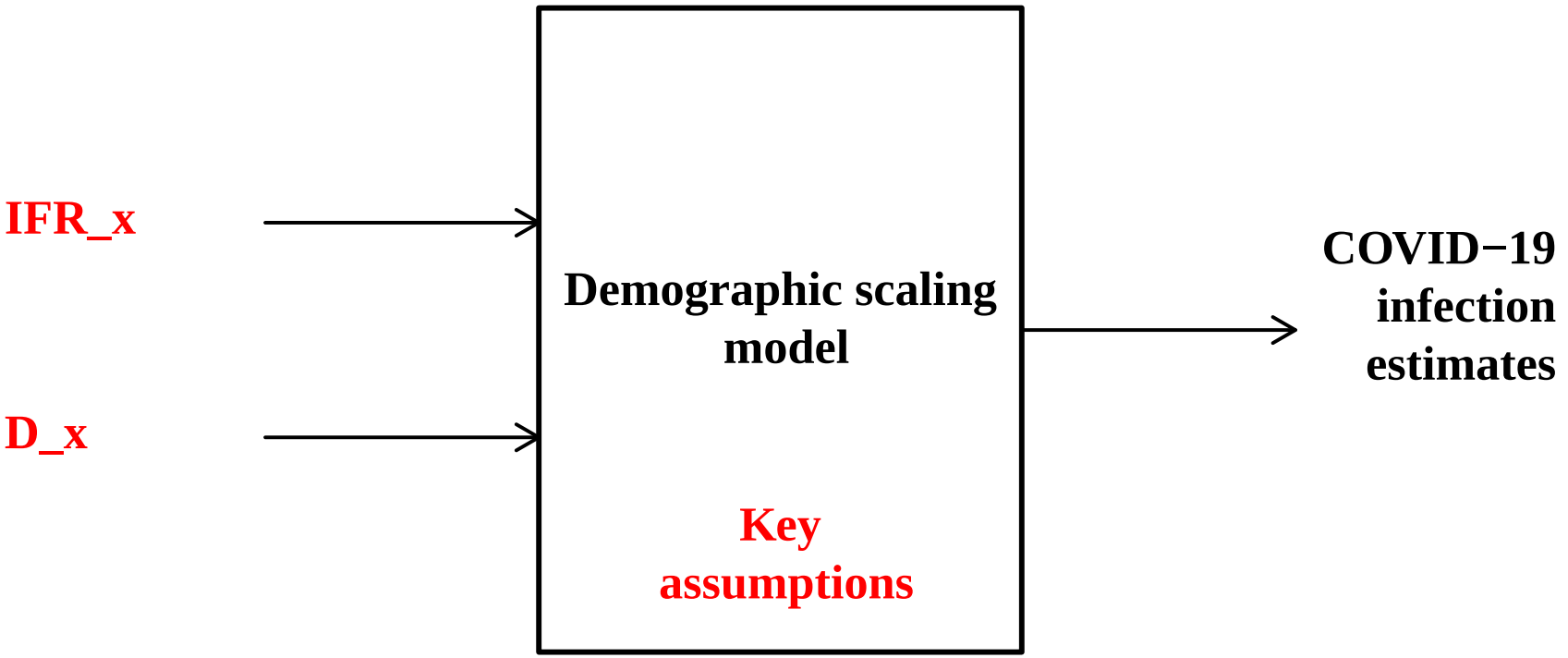
What do you think:

What impact is bigger: the one of IFR_x or D_x estimates?

Or the combined effect of IFR_x and D_x estimates?

$\Rightarrow$ How to test for this?

How sensitive are model infection estimates wrt IFR_x & D_x?



$\rightarrow$ Think creatively and critically about the combined impact of IFR_x and D_x on demographic scaling model's infection estimates.

COVID-19 infection estimates for Finland

Week 4: Impact of IFR_x on COVID-19 infection estimates for Finland as of September 15, 2020:

IFR_x	I
Verity et al., original	10 589
Salje et al., original	26 735
Levin et al., original	14 122
Verity et al., scaled	13 868
Salje et al., scaled	19 618

$\rightarrow$ Using D_x from JHU CSSE & global age pattern

COVID-19 infection estimates for Finland

Week 5: Impact of D_x on COVID-19 infection estimates for Finland as of September 15, 2020:

D_x	I
COVerAGE-DB	17 494
JHU CSSE & global age pattern	13 868

$\rightarrow$ Using scaled IFR_x of Verity et al. to better match Finnish context regarding age structure, preconditions, and medical services

COVID-19 infection estimates for Finland

Week 5 & 6: Impact of IFR_x and D_x on COVID-19 infection estimates for Finland as of September 15, 2020:

IFR_x	JHU & global age pattern	COVerAGE-DB
Verity et al., original	10 589	13 306
Salje et al., original	26 735	33 250
Levin et al., original	14 122	18 816
Verity et al., scaled	13 868	17 494
Salje et al., scaled	19 618	26 190

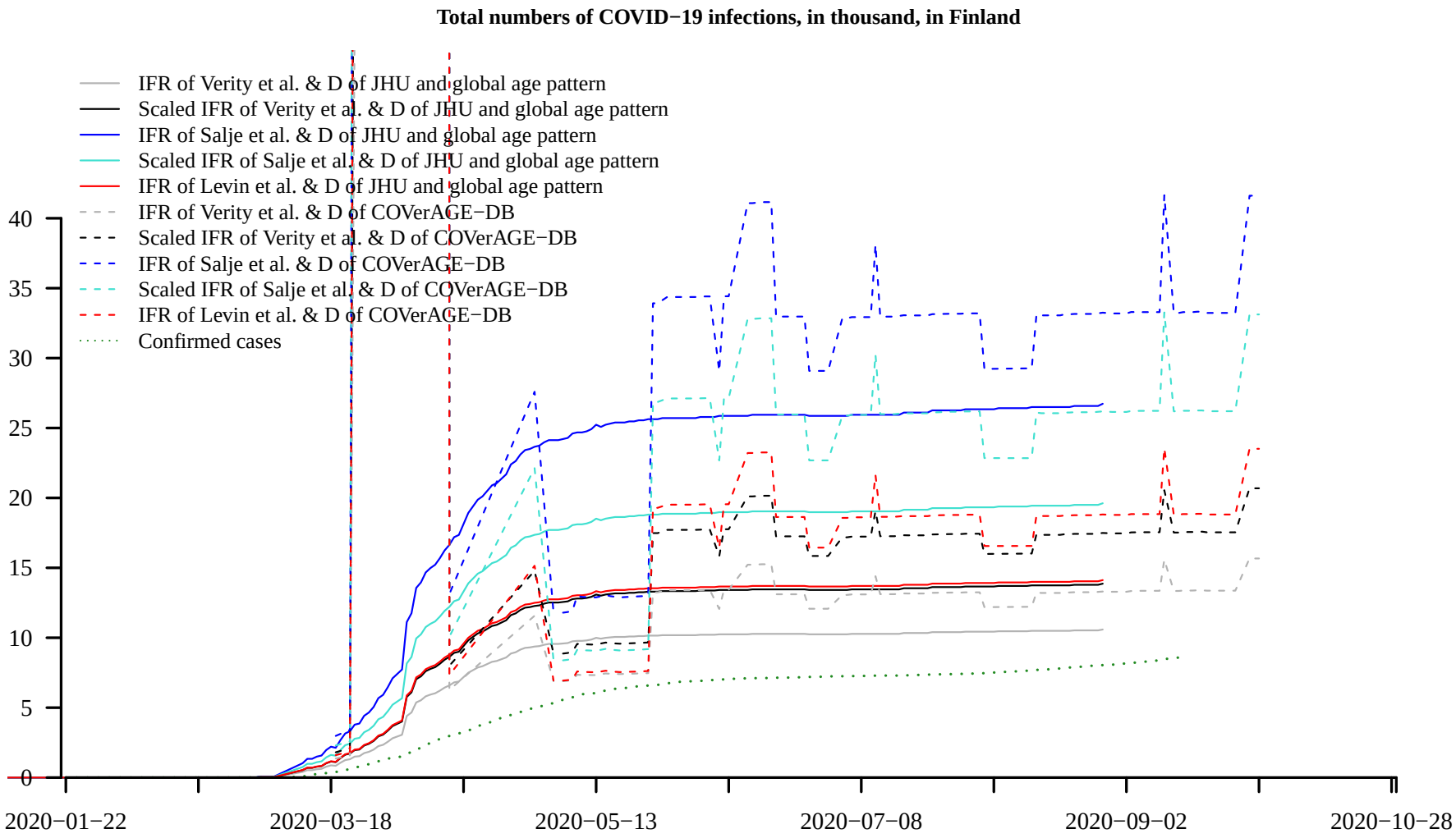
$\longrightarrow$ Largest difference, 22 661, between (i) original IFR_x of Verity et al. & D_x of JHU and (ii) original IFR_x of Salje et al. & D_x of COVerAGE-DB

COVID-19 infection estimates for Finland

- COVID-19 infection estimates for Finland tend to be consistently larger when D_x are from COVerAGE-DB as opposed to being based on data of the JHU & global age pattern
- COVID-19 infection estimates for Finland tend to be smaller when IFR_x are based on data from Verity et al. as opposed to data from Levin et al. and Salje et al.
- COVID-19 infection estimates for Finland tend to be more similar when they are based on scaled IFRs and IFR of Levin as opposed to original IFRs of Verity et al. (China) and Salje et al. (France)

$\rightarrow$ In general, the COVID-19 infection estimates for Finland appear to be sensitive to IFR_x and D_x

How robust are model infection estimates wrt IFR_x & D_x?



Take-home message from evaluating the demographic scaling model

The two key assumptions may only partially hold at the moment
and the model's results appear to be sensitive
towards both input parameters IFR_x and D_x.

However, as soon as better input data will become available,
the demographic scaling model can account for them,
and its COVID-19 infection estimates
are likely to become more accurate.

Take-home message from evaluating the demographic scaling model

It is important to think critically and creatively about any model's limitations and their possible implication for the model's outcome.

It is also important to carefully and rigorously check the sensitivity of any model's results with respect to its input.

Otherwise, *you* cannot fully understand what a model is doing and assess how valuable its results could possibly be in order to explain or predict a particular phenomenon.

Take-home message from evaluating the demographic scaling model

It is also important to comprehensively document the scientific process conducted in order to generate the presented findings.

This can also entail publishing source code and data used in order to facilitate reproducibility of scientific work and to support scientific debate.

Otherwise, *other scholars* cannot fully understand what a model is doing and assess how valuable its results could possibly be in order to explain or predict a particular phenomenon.

Topic today

Toolbox

for selecting suitable methods

&

for assessing its performance

with respect to explaining and predicting phenomena

Toolbox for selecting and assessing methods

Model selection deals with selecting a suitable method for explaining or predicting a phenomenon.

Model assessment deals with evaluating how well a selected method explains or predicts a phenomenon.

Toolbox for selecting methods

Model selection deals with selecting a suitable method for explaining or predicting a phenomenon.

Tools and concepts related to this:

- Bias-variance trade-off
- Bet-on sparcity principle
- Occam's razor (or the law of parsimony)
- ...

Toolbox for assessing methods

Model assessment deals with evaluating how well a selected method explains or predicts a phenomenon.

Tools and concepts related to this:

- Validation set approach
- Cross-validation
- Bootstrap
- ...

$\Rightarrow$ Model selection and assessment: next to AIC, BIC, R-squared, and other common diagnostic test statistics

Toolbox for selecting and assessing methods

And not to forget general sources of error when selecting (or developing) and assessing methods:

- Model misspecification
- Data issues ($\rightarrow$ input data)
- Programming issues
- Issues with software and hardware
- ...

Toolbox for selecting and assessing methods

Application to select model

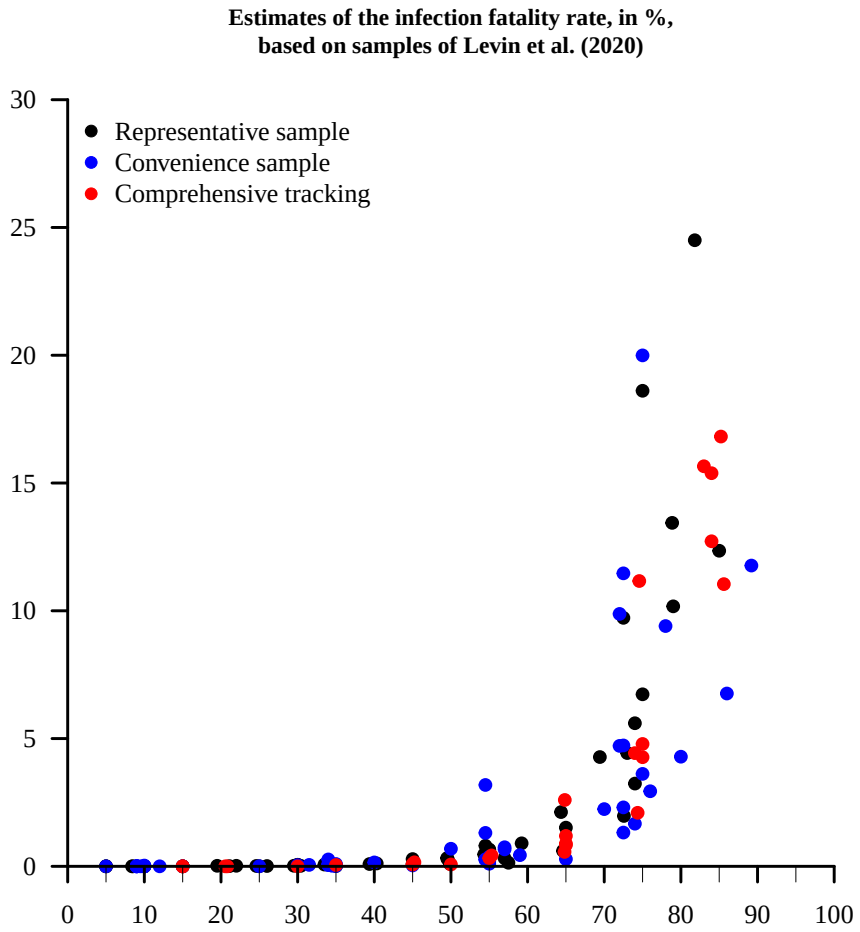
for predicting COVID-19-related infection fatality rates by age

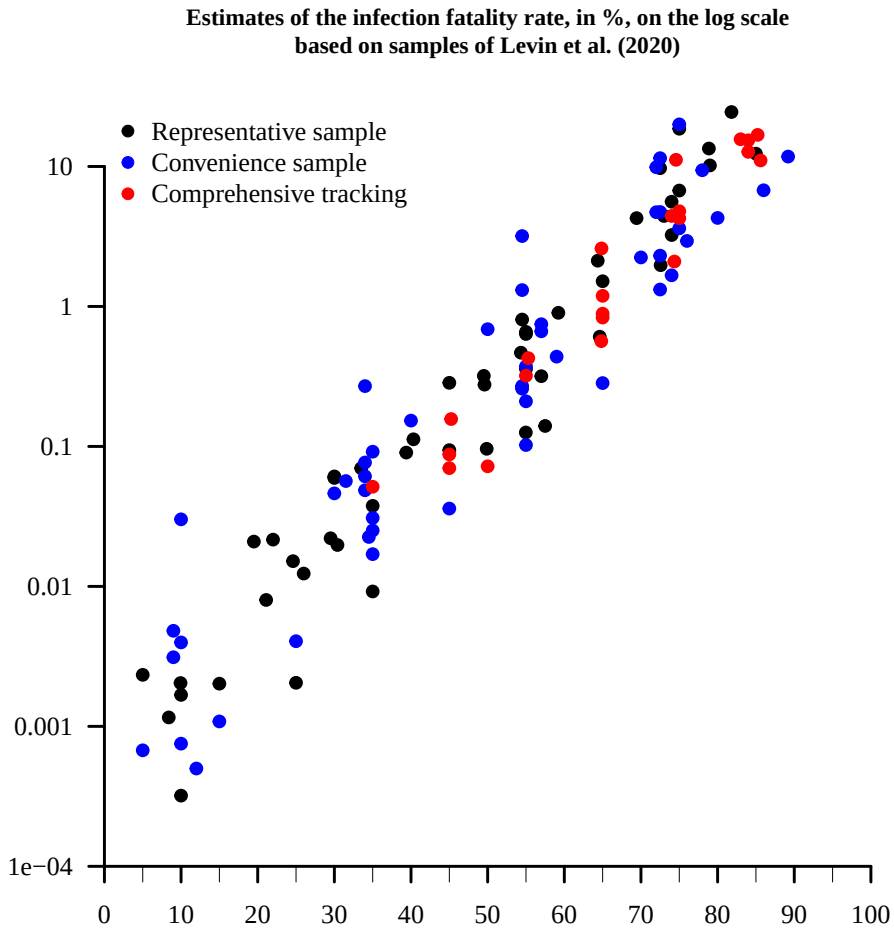
based on data provided by Levin et al. (2020).

IFR estimates of Levin et al. (2020) — just to remember from week 4

- Exponential relationship between the IFR (in %) and age: $\log IFR = -7.53 + 0.119 \times age$
- Based on data of 28 locations:
  - *Representative samples* (England, Ireland, Italy, Netherlands, Portugal, Spain, Geneva, Atlanta, Indiana, New York, Salt Lake City)
  - *Convenience samples* (Belgium, France, Sweden, Connecticut, Louisiana, Miami, Minneapolis, Missouri, Philadelphia, San Francisco, Seattle)
  - *Comprehensive tracing programs* (Australia, Iceland, Korea, Lithuania, New Zealand)
  - In total: 134 data points (IFR by age)

IFR estimates of Levin et al. — raw data





→ Source of data: Levin et al. (2020; excel spreadsheet)

Model IFR estimates of Levin et al.

Levin et al. (2020) introduce exponential model that is similar to model fitted in R:

- Model fitted in R: $\log IFR = -7.345 + 0.118 \times age$
- Levin et al. (2020): $\log IFR = -7.53 + 0.119 \times age$

$\Rightarrow$ *What do you think:*

Where could the small differences in coefficient estimates come from?

Model IFR estimates of Levin et al.

Levin et al. (2020) introduce exponential model that is similar to model fitted in R:

- Model fitted in R: $\log IFR = -7.345 + 0.118 \times age$
- Levin et al. (2020): $\log IFR = -7.53 + 0.119 \times age$

$\Rightarrow$ What do you think: where could small differences in coef come from?

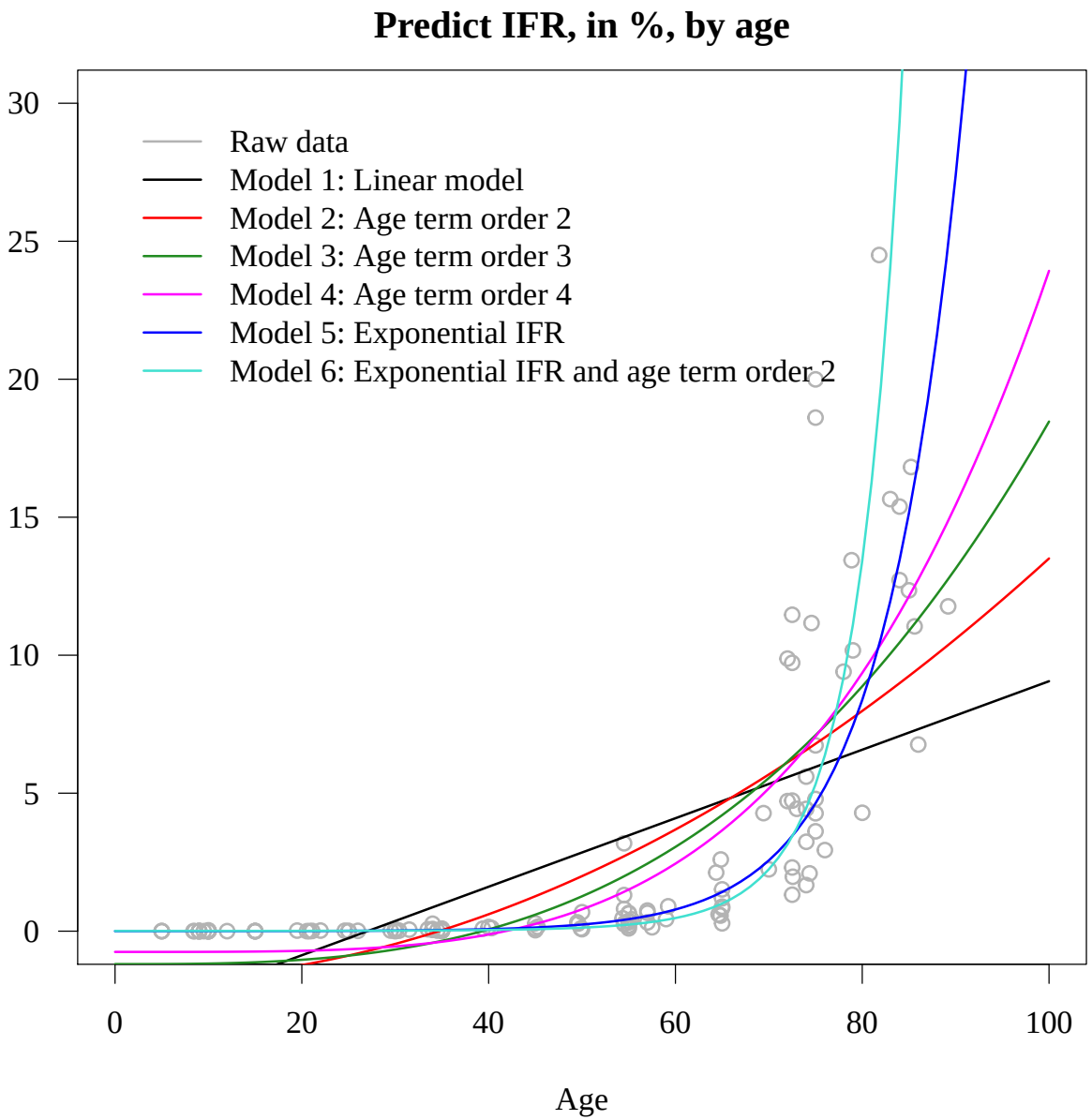
- Not all digits of IFR values in Excel spreadsheet?
- Rounding errors?
- Reporting error?
- Different model implementation in adopted software?
- ... $\rightarrow$ Try to be aware of these issues

Toolbox for selecting and assessing methods

Let us go back on track:

Is the exponential model introduced by Levin et al. (2020) the most suitable one for predicting IFR by age?

Model IFR estimates of Levin et al.



Model IFR estimates of Levin et al.

Fit different models to these raw IFR_x estimates provided by Levin et al.:

1. M1: $lm(IFR \sim age)$
2. M2: $lm(IFR \sim age^2)$
3. M3: $lm(IFR \sim age^3)$
4. M4: $lm(IFR \sim age^4)$
5. M5: $lm(\log IFR \sim age)$
6. M6: $lm(\log IFR \sim age^2)$

$\rightarrow$ Levin et al. (2020) introduce exponential model that is similar to M5

Model IFR estimates of Levin et al.

What do you think:

What model fits best raw data?

Which model would you select for predicting IFR_x?

Model IFR estimates of Levin et al.

Fit different models to these raw IFR_x estimates provided by Levin et al.:

1. M1: $IFR = -3.355 + 0.124 \times age$.
   R-squared: 0.41; p-value: $< 2.2e-16$.
2. M2: $IFR = -1.843 + 0.0015 \times age^2$.
   R-squared: 0.539; p-value: $< 2.2e-16$.
3. M3: $IFR = -1.194 + 0.000019 \times age^3$.
   R-squared: 0.6175; p-value: $< 2.2e-16$.
4. M4: $IFR = -0.751 + 0.0000002 \times age^4$.
   R-squared: 0.6597; p-value: $< 2.2e-16$.
5. M5: $\log IFR = -0.7345 + 0.118 \times age$.
   R-squared: 0.9167; p-value: $< 2.2e-16$.
6. M6: $\log IFR = -5.039 + 0.0012 \times age^2$.
   R-squared: 0.8681; p-value: $< 2.2e-16$.

$\rightarrow$ M5 has the lowest R-squared value

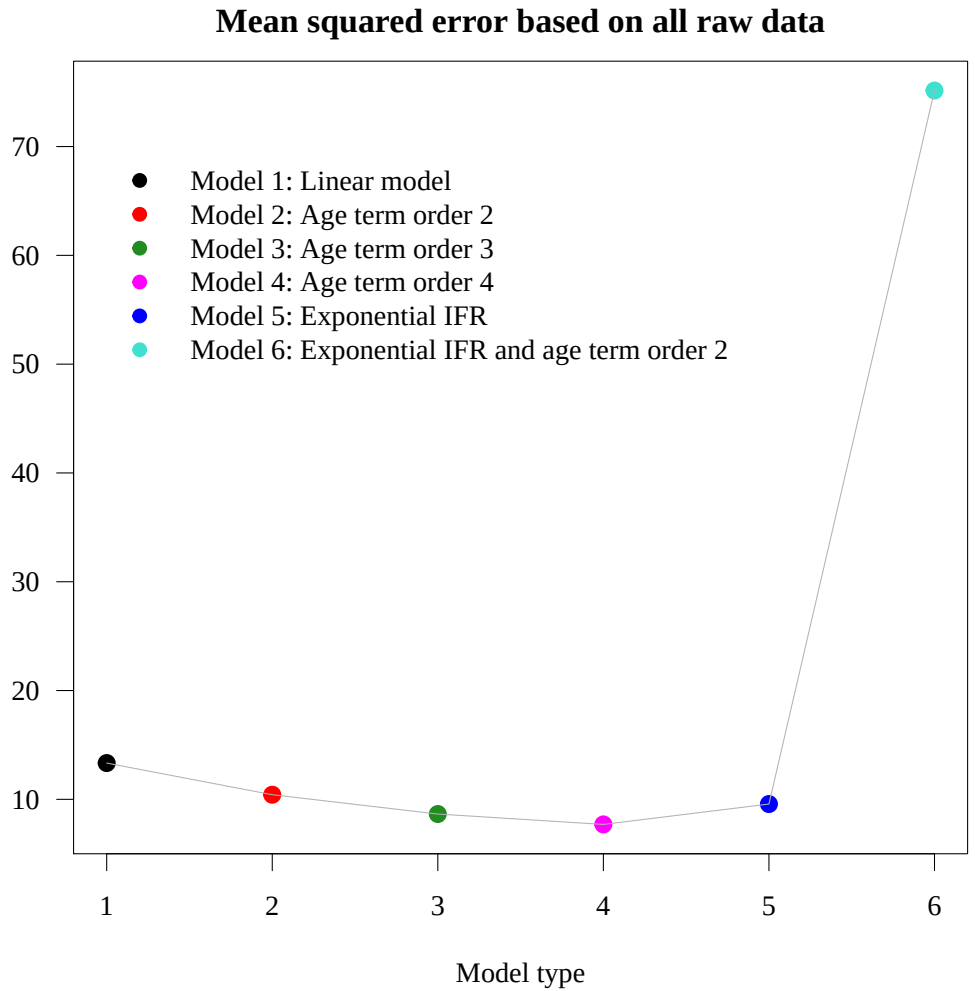
Model IFR estimates of Levin et al.

Another way for assessing how well the models M1 through M6 fit the raw IFR_x estimates is to calcuate and compare the mean squared error (MSE) between all n observed data y and their predicted values $\hat{y}$:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The smaller the MSE, the better does a model fit raw IFR_x estimates.

Model IFR estimates of Levin et al.



MSE in decreasing order:

- Model 6: 75.1
- Model 1: 13.3
- Model 2: 10.4
- *Model 5: 9.6*
- Model 3: 8.6
- **Model 4: 7.7**

$\Rightarrow$ Which model would you choose to predict IFR by age based on MSE?

Model IFR estimates of Levin et al.

Another way for assessing how well the models M1 through M6 fit the raw IFR_x estimates is to calcuate and compare the mean squared error (MSE) between all n observed data y and their predicted values $\hat{y}$:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The smaller the MSE, the better does a model fit raw IFR_x estimates.

But what does all this say about the predictive power with respect to IFR_x of each of these models?

Model IFR estimates of Levin et al.

What do you think:

What could possibly be wrong with

fitting different models to *all* raw data

and then selecting the one with the smallest mean squared error

(or, e.g., the largest R-squared value)?

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
- ...

$\rightarrow$ Raw data could even be split into: *training* data, *testing* data, and *validation* data.

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

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

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

The *bias-variance trade-off* describes the balance of two fundamental features of any statistical model.

A suitable model has *low bias*, which indicates small test MSE, and *low variance*, which indicates similar IFR predictions of the same model when fitted to various training data.

This is also related to the concept of *overfitting*: a model with small training MSE and large test MSE is said to be likely to overfit training data and *vice versa*.

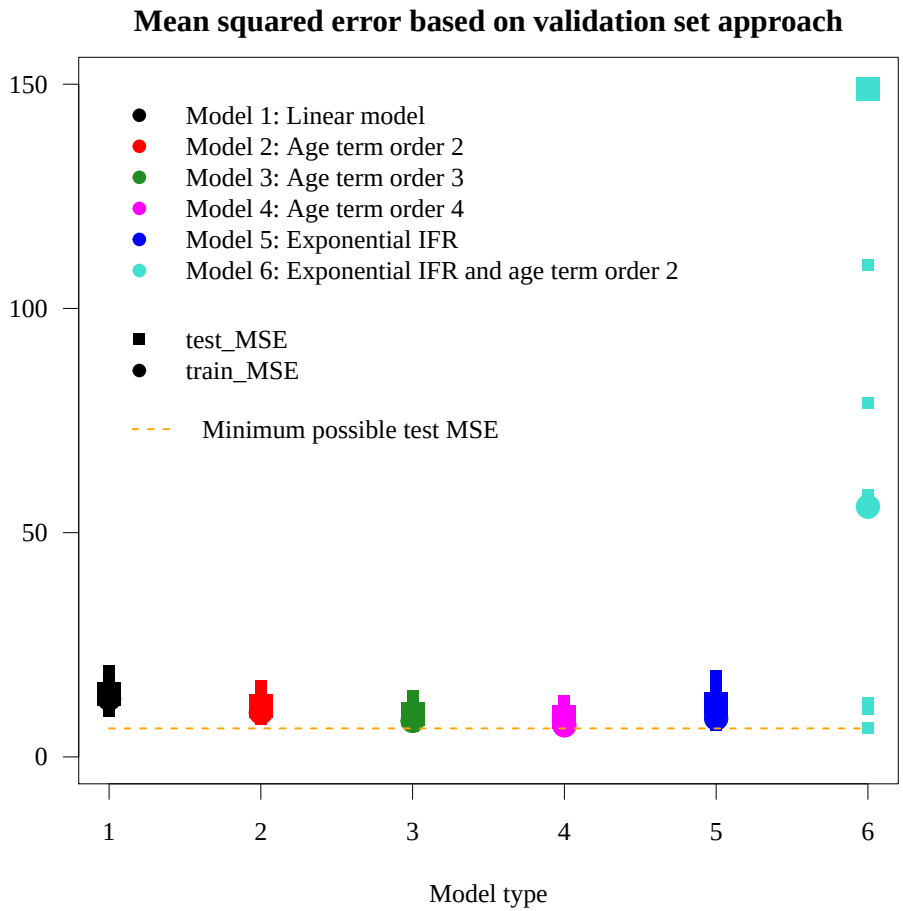
IFR estimates of Levin et al.

At first, we use the validation set approach

to select the best (of the six) models

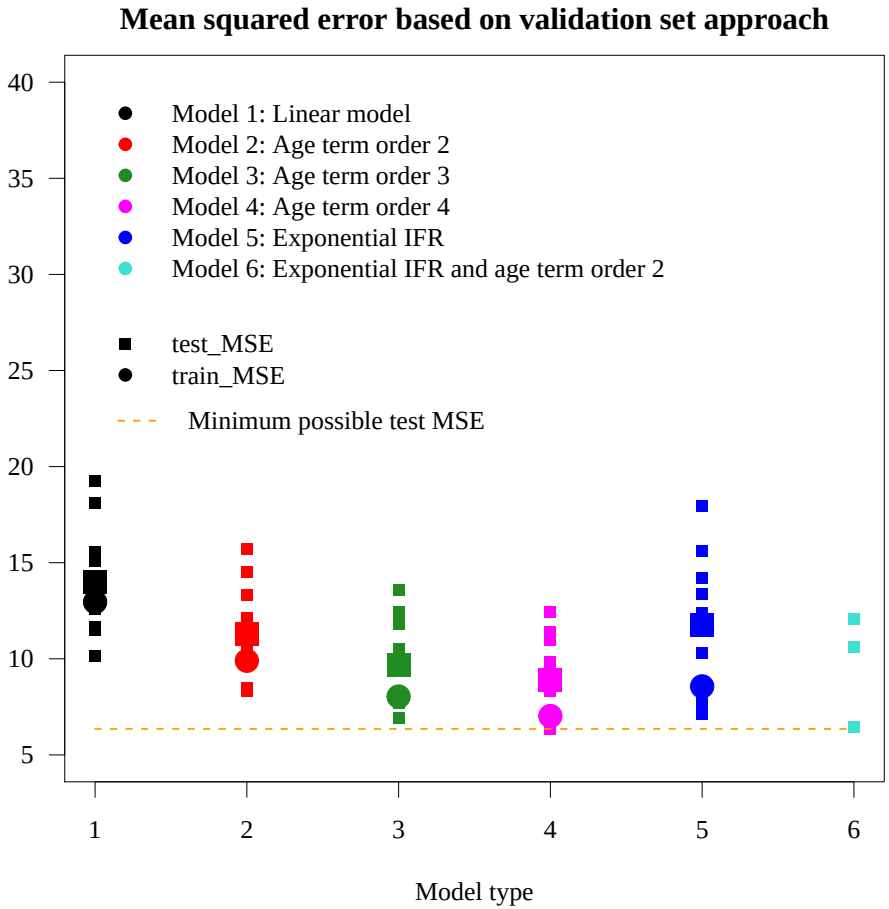
for predicting the IFR by age.

IFR estimates of Levin et al.



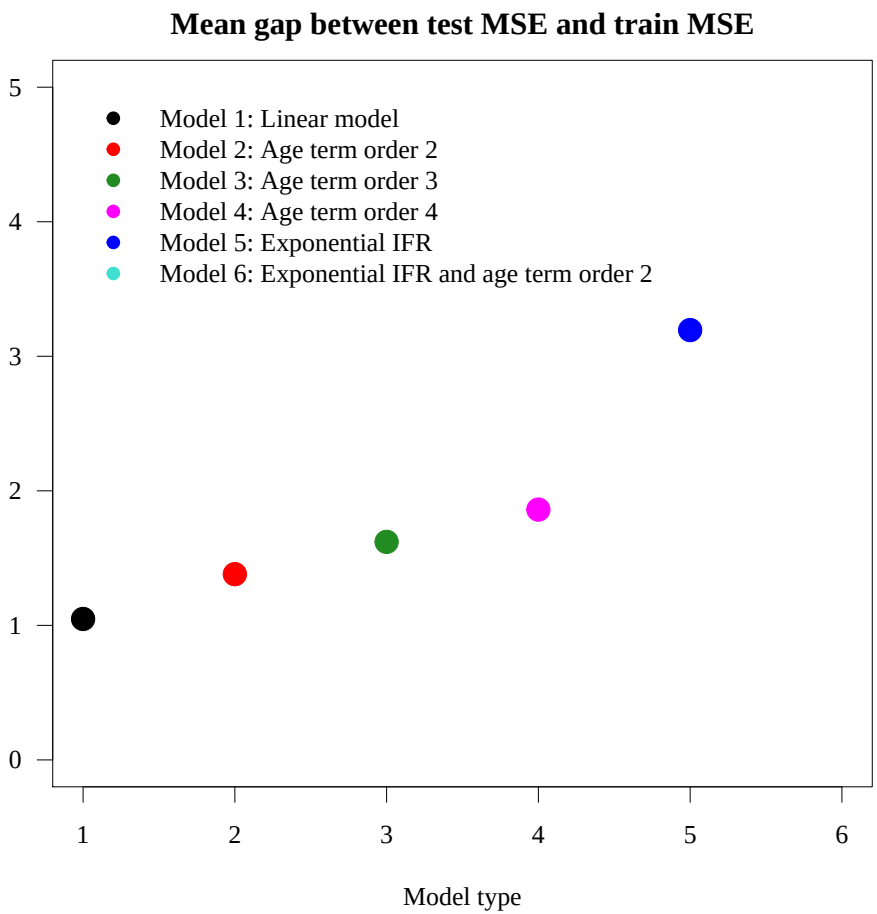
- Validation set approach applied 10 times
- Test MSE is consistently larger than train MSE
- Test MSE is smallest for M4 (low bias)
- M3-M5 are all close to minimum possible test MSE on average

IFR estimates of Levin et al.



- Validation set approach applied 10 times
- Test MSE is consistently larger than train MSE
- Test MSE is smallest for M4 (low bias)
- M3-M5 are all close to minimum possible test MSE on average

IFR estimates of Levin et al.



- Gap between average train MSE and average test MSE tends to increase with model complexity ($\longrightarrow$ variance of test MSE; overfitting)

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

IFR estimates of Levin et al.

We continue using k-fold cross validation **next week**

to select the best (of the six) models

for predicting the IFR by age.

What you have learned today about assessing the demographic scaling model

- Describe the impact of D_x and IFR_x on the COVID-19 infection estimates for Finland (and in other countries).
- Describe the idea for splitting observations into training data and testing data.
- Explain validation set approach.
- Describe the idea behind the bias-variance trade-off: what low bias and low variance mean.

Course learning materials

Course learning materials on GitHub:

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

Recommended learning material for today's class

- **Gareth J, Witten D, Hastie T, Tibshirani R**
  An Introduction to Statistical Learning with Applications in R.
  Springer Science+Business Media New York 2013
- **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`

Thank you for your attention!

christina.bohk-ewald@helsinki.fi

Sixth week's class in the lab: Sensitivity of demographic scaling model's results & toolbox for selecting and assessing suitable methods.

- Analyze the sensitivity of the demographic scaling model's results for Finland with respect to IFR_x & D_x together.
- Select and assess suitable model for predicting IFR_x starting from exponential model of Levin et al. (2020).

$\rightarrow$ Present and discuss your findings in class at the beginning of the next session on Monday.

Seventh week's class in the lab: toolbox for selecting and assessing methods

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.