COS-D407. Scientific Modeling and Model Validation

Lecturer: Christina Bohk-Ewald

Week 6

University of Helsinki, Finland
01.11.2021–15.12.2021

Sixth week's class:

Scientific modeling & model validation in practice

- Q&A: recap of material of previous session
- Present your findings of two previous lab sessions
- 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: Toolbox for selecting and assessing suitable methods.

- Select and assess suitable models for predicting IFR_x starting from the exponential model of Levin et al. (2020).

→ 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 sources are there for getting the numbers of COVID-19 infections in a country of interest?
- How does the demographic scaling model estimate the total numbers of COVID-19 infections in a country of interest?
- What are the core input parameter of the demographic scaling model?
- What different sources of IFR_x estimates do you know of?
- How to analyze the quality of the COVID-19 infection estimates of the demographic scaling model?

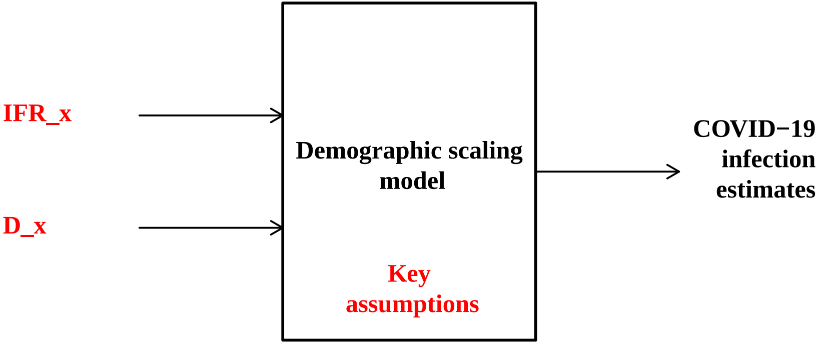
$\longrightarrow$ Open questions?

Present your findings of the two previous lab sessions:

- How large are the COVID-19 infection estimates for Finland produced with the demographic scaling model? (week 4)
- What are the two key assumptions of the demographic scaling model and how likely are they to hold in real-world applications? How might this impair the estimations? (week 4)
- How robust are the COVID-19 infection estimates for Finland with respect to taking infection fatality rates from different sources? (week 5)
- How plausible are the COVID-19 infection estimates, also considering the time (in)variance of input data? (week 5)

→ Open issues?

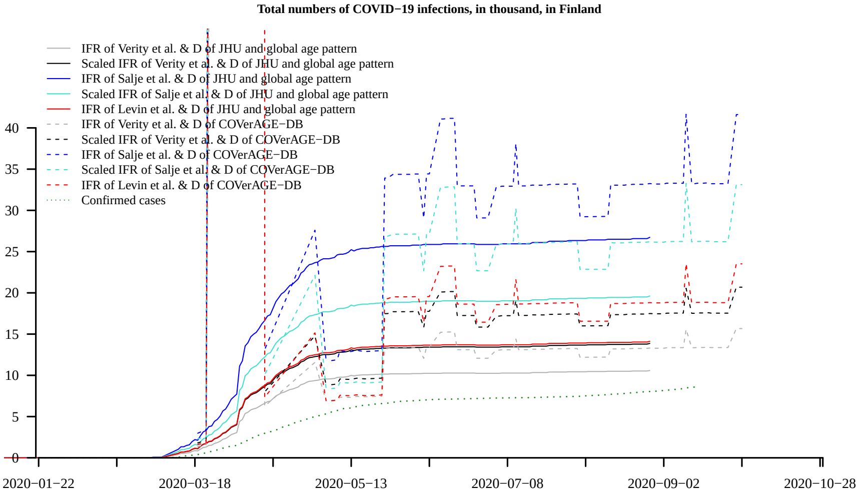
Some more thoughts on evaluating a model without true realizations



→ Think creatively and critically about the model's key assumptions, how likely they are to hold in real-world-applications, and if they might change over time.

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

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
- ...

⇒ 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 (→ input data)
- Programming issues
- Issues with software and hardware
- ...

Toolbox for selecting and assessing methods

Application to select model

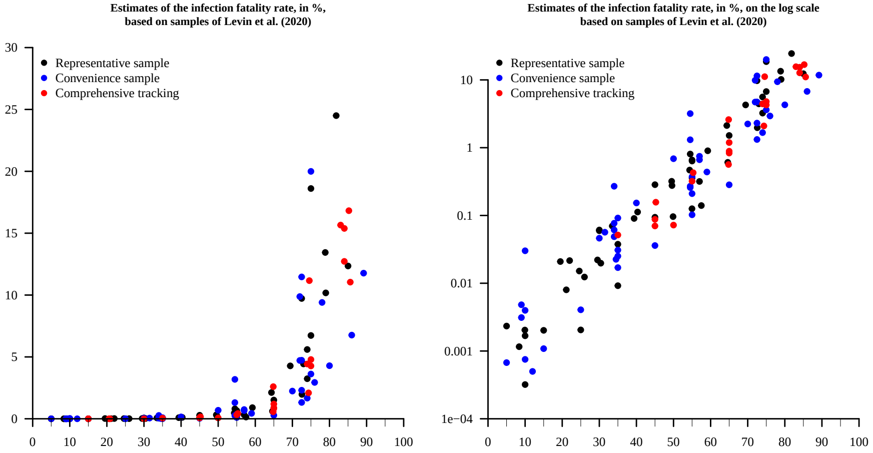
for predicting COVID-19-related infection fatality rates by age

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

IFR estimates of Levin et al. (2020) — just to remember from week 5

- Exponential relationship between the IFR (in %) and age:
  $\log \mathit{IFR} = -7.53 + 0.119 \times \mathit{age}$
- 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: $\log \mathit{IFR} = -7.345 + 0.118 \times \mathit{age}$
- Levin et al. (2020): $\log \mathit{IFR} = -7.53 + 0.119 \times \mathit{age}$

$\Rightarrow$ *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: $\log \mathit{IFR} = -7.345 + 0.118 \times \mathit{age}$
- Levin et al. (2020): $\log \mathit{IFR} = -7.53 + 0.119 \times \mathit{age}$

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

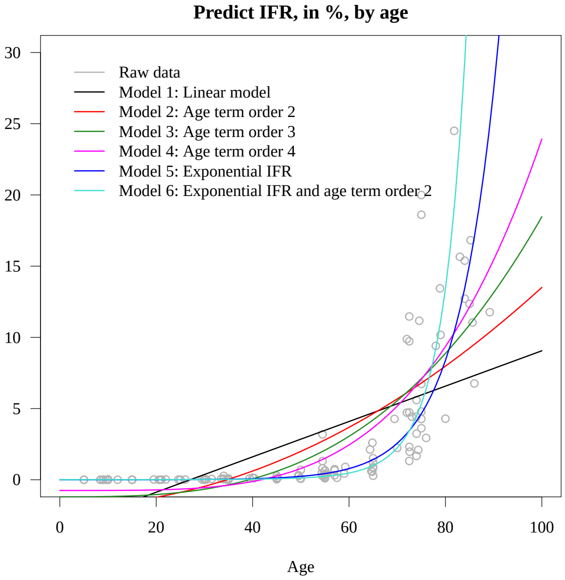
- Not all digits of IFR values in Excel spreadsheet?
- Rounding errors?
- Reporting error?
- Different model implementation in adopted software?
- ... → Try to be aware of these issues

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: $lm(\log IFR \sim age^2)$

$\longrightarrow$ 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: $< 2.2e - 16$.

→ M5 has the highest R-squared value

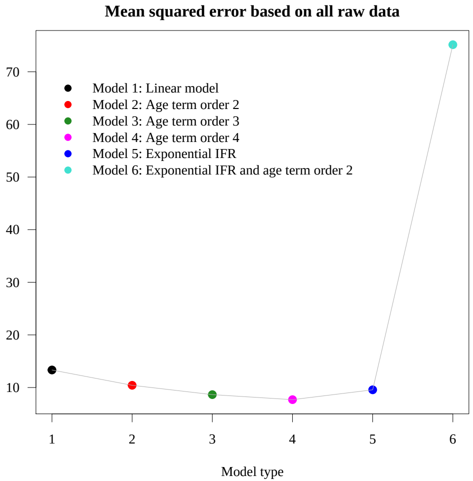
Model IFR estimates of Levin et al.

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 $\hat{y}$:

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

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**

⇒ 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 $\hat{y}$:

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

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
($\rightarrow$ 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
- ...

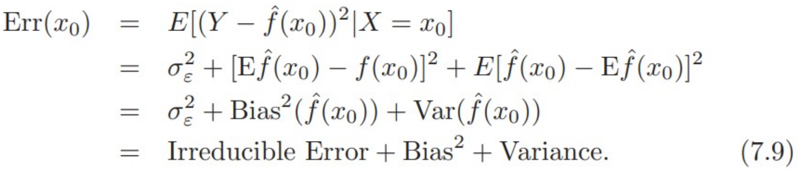
→ Raw data could even be split into: *training* data, *testing* data, and *validation* data.

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

Components of the expected test MSE



$$\begin{aligned}
\mathrm{Err}(x_0) &= E[(Y-\hat{f}(x_0))^2|X=x_0] \\
&= \sigma_\varepsilon^2 + [\mathrm{E}\hat{f}(x_0)-f(x_0)]^2 + E[\hat{f}(x_0)-\mathrm{E}\hat{f}(x_0)]^2 \\
&= \sigma_\varepsilon^2 + \mathrm{Bias}^2(\hat{f}(x_0)) + \mathrm{Var}(\hat{f}(x_0)) \\
&= \text{Irreducible Error} + \mathrm{Bias}^2 + \text{Variance}. \qquad (7.9)
\end{aligned}$$

Hastie et al. (2009; page 223; Equation 7.9; https://hastie.su.domains/Papers/ESLII.pdf**)**

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

Bias refers to the error that is introduced by approximating a real-life problem by a much simpler model. It is about the deviation from the truth.

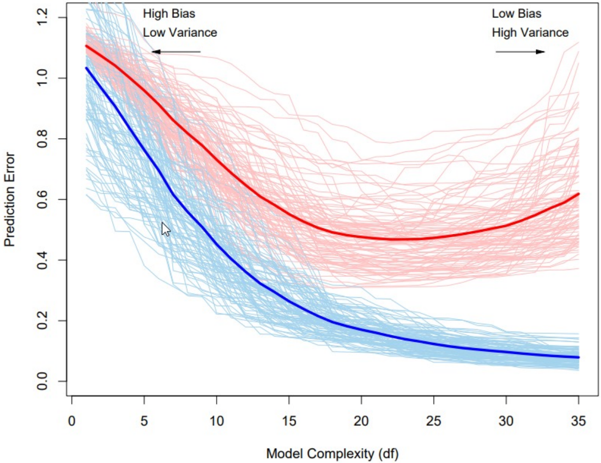
Variance refers to the amount by which the fitted model function (and its predictions) would change if different training data were used. It is about the dependence from training data.

A suitable model should have *low bias* and *low variance*. That is, it should approximate reality well and it should rather catch underlying pattern and not random noise in the data.

However, there is the *bias-variance trade-off*.

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`)

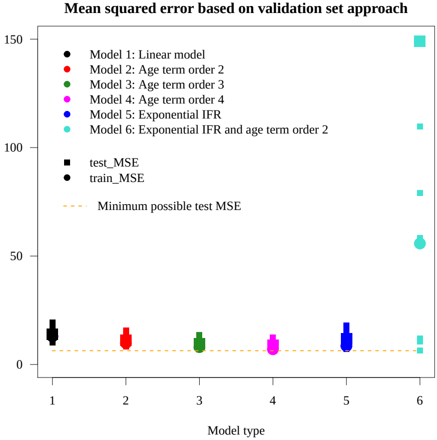
IFR estimates of Levin et al.

At first, we use the validation set approach

to select the best (of the six) models

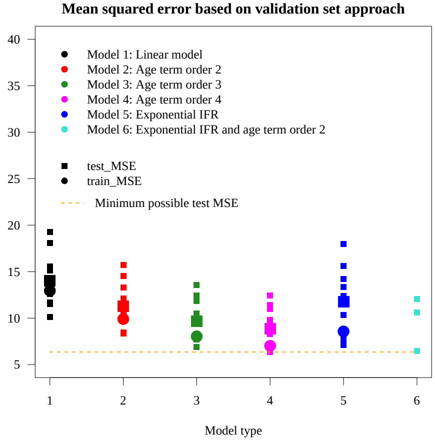
for predicting the IFR by age.

IFR estimates of Levin et al.



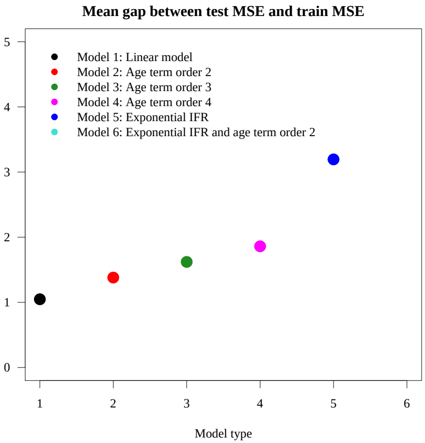
- Validation set approach applied 10 times
- M6 is not suitable to predict IFR by age

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

⇒ 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 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/2021-COS-D407-scientific-modeling-and-model-validation

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`

Thank you for your attention!

christina.bohk-ewald@helsinki.fi

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

- Select and assess suitable models for predicting IFR_x starting from the exponential model of Levin et al. (2020).

→ 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.