Predicting number of patients in hospital care because of COVID-19 after two weeks in Finland

Introduction

COVID-19 is a disease which is caused by SARS-CoV-2-virus (1). The virus has spread globally, and coronavirus epidemic was declared as pandemic on March 11, 2020 by the World Health Organization (WHO) (1). This coronavirus causes respiratory infections and in small percentage the virus causes severe infection (1). In Finland 1,2% of confirmed coronavirus infections cases have died (situation 12.3.2021) and globally the estimated mortality is 2.2% (situation 12.3.2021) (2, 3). The numbers are estimations since all cases are not confirmed or reported.

Different restrictions are being made to prevent the spread of the coronavirus and thus secure the capacity of hospital care. The incubation period (time from infection to the onset of initial symptoms) is estimated to be 1-14 days and most common symptoms appear 4-5 days after infection (4). The number of COVID-19-patients in hospitals is reflected in the infection rates in delay. Number of infections are difficult to predict as different restrictions are being made and also transformed coronaviruses change the situation quickly. Especially new more contagious virus variant first detected in Britain has spread rapidly in Finland and increased the number of infections (2).

Vaccinations against the coronavirus have been started in the end of year 2020 (5). Vaccination coverage for the first dose is 9,7% and for the second dose is 1,5% (situation 12.3.2021) (5). Vaccinations are given first to health care personnel and after that for risk groups, starting with the oldest and those with predisposing underlying diseases (1). The progress of vaccinations is expected to reduce number of patients in hospital as people with highest risk for severe disease are vaccinated.

Problem formulation

Purpose of to create model for predicting number of patients in intensive care because of coronavirus after two weeks. This is done by using machine learning methods. As several factors affect the situation (including restrictions given by government) the situation is extremely difficult to predict. Aim is to create a rough model that gives an indicative estimate of the course of the disease situation. Since the patients admitted to hospital come with a delay after diagnosing the disease, I am estimating number of patients in hospital after two weeks. As vaccinations are given first for risk groups, also progress of vaccination should reduce the proportion of patients in hospitalcare. So here the aim is to do prediction based on confirmed infection cases and number of vaccinations given. Estimation is done by using numbers from the beginning of 2021 as the new coronavirus variant from Britain has become more common from the beginning of the year.

In this machine learning problem, the datapoints are individual days. Features for each day are estimated vaccination coverage (proportion of people received at least first vaccination) and number of new confirmed infection cases. The label is number of people in hospital care after 14 days. The number of confirmed infection cases daily 2021 can be found from THL's open data API (6). Number of vaccinations

given weekly 2021 can be found from THL's open data API (7). The vaccinations given daily are estimated as if number of given vaccinations that week is distributed evenly for each day of the week. Individual daily vaccinations are not used since those have changed greatly afterwards and reliable daily numbers are not achievable. Number of patients in hospital care are listed daily in THL (2).

Data is gathered from January 1, 2021 to February 28, 2021 so it includes 59 datapoints. The number of people in hospital care after 14 days (label) is gathered from January 14 to March 14. The data is split to two datasets: the first dataset consisting of 40 of datapoints (2/3) and second one consisting of 19 datapoints (1/3). Datapoints are split randomly to these two datasets. First dataset (40 datapoints) is used for training, validation and selecting of best suitable model. The second dataset (19 datapoints) is used for evaluation of the validity of selected model.

Method

The analysis if done by RStudio Version 1.4.1103 (10). The first inspection of the datapoints does not give clear picture of the relation between the features (vaccinations given and number of new cases) and the label (number of people in hospital after 2 weeks) of datapoints. Since the relation is not quite clear, I am testing both multiple linear regression and polynomial regression. Linear regression is based on the linear hypothesis space h(x) between feature space X and label space Y (8). Polynomial regression is based on nonlinear hypothesis space between feature space X and label space Y with different r degrees (8). Since we have now two features, we are using multiple linear regression, where features "vaccinations given" and "number of new infections" make up feature space X.

We need to interpret polynomial regression as a combination of a feature map (transformation) and linear regression (8). In both tests the quality of a predictor is measured by squared error ross. With the training data we compute predictor that minimize the average squared error loss (y-h(x))^2, where y is label and h(x) is predictor. Smaller the difference between label y and predictor h(x) is, the smaller the average squared error loss is.

Multiple linear regression can be considered as same than polynomial regression with degree r=1. For the multiple polynomial regression, we are testing different degrees r (>= 2). We are testing maximum degree 2, 3, 4 and 5. The resulting Squared error lost for training data for different models are listed in table 1.

For selecting best model, I am using k-fold Cross Validation with k=5. Testing dataset 1 (n=40) is split evenly into k=5 subsets (8). I repeat learning 5-times and each time I use one subset as validation set and k-1 subsets (=4) as training set (8). From computed validations errors I will compute average validation error. The resulting validation errors are listed in table 1.

	r=1 (linear)	r=2	r=3	r=4	r=5
Training data	847.81	570,63	238,98	188,92	70.0197
Validation data	888.43	650,48	592.64	1352.96	2378.57

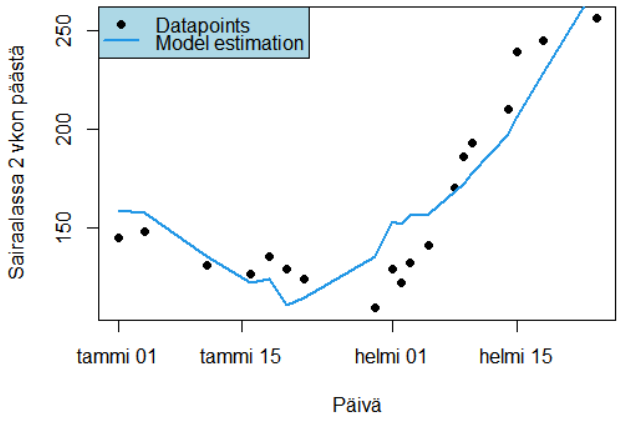
***Table 1**: Training and validation errors for linear regression model (r=1) and for polynomial regression models with different degrees of r.*

Results

Based on the result we select polynomial regression model with r=3 since it gave the smallest validation error. For r values 4 and 5 the regression model seemed to overfit the training data since the validation error is much larger than testing error. For the maximum degree r=3, the validation error is larger which indicates that also this model overfit but still this model is considered best suitable for our data.

The Dataset 2 is used for testing the validity of our model. The average loss for the Dataset 2 (test set) gives mean squared error loss 323,086 meaning that error in test set is smaller than validation error (592.64). Visual presentation of selected model with test data (Dataset 2) is presented in Picture 1 where dots present actual values of test data and line represent prediction given my selected model.

***Picture 2**. Dots represent actual values of test data and blue line represent estimation given by polynomial regression model with r=3*



Conclusion

For estimating the number of patients in hospital after two weeks because of COVID-19 infection, I created linear regression model and polynomial regression model with degrees r=2, 3, 4 and 5 and used number of new cases and estimated number of vaccinations given as features. By using k-fold Cross Validation with k=5 I chose polynomial regression model with r=3 as best model for estimating this data. The training error for this model was 238,98 and validation error was 592,64. The validation error was clearly larger than training error indicating that our model is overfitting.

The model was tested with Dataset 2 and resulting test error was 323,086, which is smaller than validation error, so model seem to work well in test data. The test data is also presented in Picture 1 and visual inspection gives impressions of overfitting.

It should be noted that when using multiple regression, different features should ideally be independent. In this case the features aren't completely independent since amount of given vaccinations affects number of new COVID-19 cases. In my model I am expecting that number of vaccinations alone would affect number of people in hospital after two week, since it is expected that people who has been vaccinated, are less probably having the severe form of the disease and thus getting to hospital care.

In the end it should be taken into consideration that COVID-19 disease is extremely difficult to predict and there are lots of factors which might affect both number of infections and number of people in hospital. There are different virus variants reported all over the world and these variants have shown that mutations might change situation quickly. It should be also taken into account that actual infection number is larger than confirmed infections. Some people are asymptomatic and some just do not go for the test even with symptoms. For example, some people do not want to go for test because of strict quarantine regulations in case positive test result. There has been also uncertainty in test results and both false positives and false negatives have been reported.

Even though the model seems to work quite well in my data, it should be taken into consideration that this dataset was relatively small (n=59) and vaccination coverage is still quite small in Finland. It is expected that vaccinations would reduce the number of people getting to hospital after COVID-19 infection, but in our model, it cannot be seen that clearly. It is expected that vaccination rate will increase in coming months and this is expected to also reduce number of patients in hospital. I will test this model in next couple weeks and possibly create new model after few weeks since larger dataset probably will give more accurate model.

References

1. Uusi koronavirus (COVID-19). *Duodecim - Terveyskirjasto*. Referenced 12.3.2021
2. Situation update on coronavirus. *Finland institute for health and welfare.* Referenced 12.3.2021.
3. WHO Coronavirus (COVId-19). Dashboard. *World Health Organization.* Referenced 12.3.2021.
4. Koronaviruksen tarttuminen ja itämisaika. *Finland institute for health and welfare.* Referenced 12.3.2021.
5. Koronarokotukset Suomessa. *Finland institute for health and welfare.* Referenced 12.3.2021.
6. COVID-19 cases in the infectious diseases registry. *Finland institute for health and welfare.* THL's open data API.
7. Vaccinations over time in Hospital Care District per age group. *Finland institute for health and welfare.* THL's open data API.
8. Jung, A. Machine Learning. The Basics. 2021. mlbook.cs.aalto.fi
9. Venttola, E. Tässä koottuna kaikki, mitä rokotteiden aikataulusta tiedetään nyt – Suomeen voi tulla pian parissa viikossa yhtä paljon rokotteita kuin tähän saakka yhteensä. *Aamulehti.* Referenced 19.3.2021.
10. The R Project for Statistical Computing. R project. Referenced 19.3.2021.