Thank you, Phu, i will no proceed with the presentation of our models.

First of all, Phu gave you some more information about the hospitalization and how different aspects, like the individual age groups or the individual states can impact the number of newly hospitalized people.

As you saw in qians presentation it is quite difficult to predict the hospitalization using infections as the dependent variable while keeping in mind that there a big differences over the several states.

In other words, we cannot assume, that the hospitalization can be reliably predicted in the individual states by including the infections, because it is much likely, that the number of infections is way higher in a given state then it would normally be due to tourism activities.

With these information’s in mind, we tried to first understand a little bit better how the hospitalization is affected by individual covariables using a Gam, secondly applying those understandings to a time series model.

First we started using a GAM to predict the hospitalization for hole germany, using the number of infections, which should be way closer to the true value than for the individual states.

For the Gam we used, that the interaction between Infections and Age groups has a nearly linear effect on the hospitalization. We also used, that the effect of time is not linear and that the effect of the time has the same structure for each year. So we can conclude that the use of a cyclic spline should be valid for either prediction and forecasting.

Cyclic Spline definition: Besides P-Splines, which are Penalized B-Splines, there are also other techniques to form one-dimensional smoothers.

For example there are cubic splines, which parameterize its values at its knots

“It is quite often appropriate for a model smooth function to be ‘cyclic’, meaning that the function has the same value and first few derivatives at its upper and lower boundaries.”

A

We could also have used P-Splines to predict the smoothing term over the hole course of the pandemic. Of course, this would yield a nearly perfect match.

But the problem by using P-Spline is, that they are quite good for explaining, but not for forecasting since after the last knot we do not have any estimated function left to calculate the new data.

In other words, we can describe the course of the hospitalization quite well using a timevariing covariable and using P-Splines. So each observation is nearly perfectly fitted. But we would fail at forecasting and describing unseen data in the future.

Since we are handling with countdata the poissonmodell seemed suitable, but by analyzing the mean and the variance in each age group for hole germany we can see, that the assumption for the poissonmodell, that mean and variance are nearly the same is not given. By using a quasipoissonmodell we can correct the estimation of the variance by estimating a dispersionparameter.

In the End we expect, that the hospitalization for hole Germany can be modelled the following way:

E(hospitalization\_i) = b\_0 + b\_Infection + b\_Agegroups + b\_Infection\*Agegroup + f(Kalenderwoche, Agegroup)

Where we expect, that hosp\_i|x\_i ~ Poi(lambda) with some Overdispersion.

This should give us some bigger confidence intervals but should also lead to more reliable estimations.

Here you can see the result of our model. As you can see most of the deviance is explained, and the scale parameter is estimated to be … which is way higher than 1.

Another way to identify the problem is by reducing it to the hospitalization. That way we receive an univariate timeseries which than can be analysed and modelled.

By following this path we do not model the hospitalization dependent on the infections anymore. We also do not need the distinction between several agegroups anymore, which is quite usefull since we can assume, that many people especially in the early ages are more likely to be infected (even more than once), just by going to school or work.

In order to understand a little bit better how we preceded from here we will define some words that we need to know in order to predict and forecast time series data.

What exactly is a time series. The term "univariate time series" refers to a time series that consists of single (scalar) observations recorded sequentially over equal time increments.

Here we use one week increments.

Normally univariate time series´s can be explained by their properties, like for example

Stationarity (which means, is there a clear trend visible)

Seasonality (which means the timeseries is it repetitive)

As we saw in Phus analysis the hospitalization is obviously not stationary, where this holds for each individual state as well. Here for example, you can see a stationary time series in contrast to the hospitalization over the course of time.

Next the time series

By normalizing the data, we get rid of these high numbers in contrast to low numbers.

In the end both models have their problems. The GAM uses more information which could lead to more precise predictions but it also shows a lot of heteroscedasity which can be explained by several factors. For example the missing distinction between the individual states, the inability to really account for infections which led to the hospitalization in each state or the fact, that is still possible, that people get infected multiple times, which we cannot account for, since all the data is anonymized.

On the other side we have our arima model using seasonality which only accounts for hospitalizations and is estimating the future trend, given the past observations