Advanced Methods for Data Analysis (36-402/36-608)
Influence-like illness infection rate forecast
HiJae Kim

Introduction

In this case, Center for Disease Control (CDC) is concerned about ways to forecast the percentage of patients it sees with an influenza-like illness (ILI). I am given with weekly ILI(WILI) percentage records of 10 years and 29 weeks. This data set is composed of three components: week, season and WILI. WILI is the weekly value of the percentage of patients with an influenza-like illness across the US population. Week and season show the underlying week of each WILI value.

Figure 1 below shows the plot of WILI values (response) against the underlying week and this will show a general trend of the WILI value over the past 10 years and 29 weeks. The red line connects the average of each week's value.

In this report, I am going to make forecasts as CDC requested. In particular, I only have the first 29 weeks of the 11th year and I will predict the WILI for the remaining 23 weeks of the 11th year. Then, I will check the test error of my forecast and finally find the forecasted peak of season 11 and its variability.

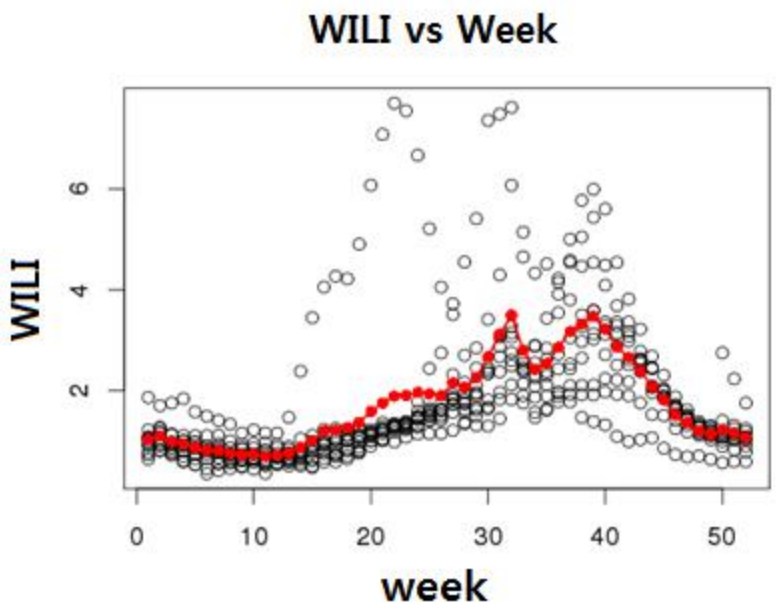

Figure1: plot of all WILI and week

Forecasting Season 11

To find out what the last curve going to look like over week 50, … 72 for season11, I first fit a nonparametric smoother(smooth spline) to the flu curves of the 10 past years and the first 29 weeks of the 11th year. I divided the data set into 10 seasons and fit a smoothing spline estimate to the WILI across all 10 separate seasons. I chose the degree of freedom with the internally set value using leave-one-out cross-validation.

Then, I find which of the first 29 weeks of the first 10 seasons resembles most of those of the 11th year. For this, I use mean squared error (average squared differences between each of first 29weeks of 11th season and those of each of first 10 seasons) my metric of comparison. The table below shows the mean squared errors across 10 seasons compared to 11th season.

Season	1	2	3	4	5	6	7	8	9	10
Error	0.94659	0.07561	0.01259	0.02399	0.02206	0.11702	8.39420	**0.01001**	0.03220	0.05621

Table1: mean squared errors across the first 29 weeks mean(wili[season==11][1:29]-flu.pred$y)^2 with degrees of freedom generated by cross-validation. Lowest value is bolded.

This table shows that season eighth has the lowest error with this model. I also noted that season 7 looks “very” different from season 11 and, in fact, from any other seasons.

Next step is refitting the smoothing spline estimate with degrees of freedom scaled 75% of the degrees of freedom from the original leave-one-out cross-validation. This method may seem ad hoc, but often leave-one-out cross-validation does not provide us with smooth enough fits, and this is a way of increasing the amount of achieved smoothness. As you can see from the below table3, the MSE of newly made smoothing spline estimates using 0.75 of the original degrees of freedom actually provides slightly less error.

Season	1	2	3	4	5	6	7	8	9	10
Error	0.99506	0.07298	0.01053	0.02262	0.02139	0.11780	8.37130	**0.00908**	0.02945	0.06121

*Table2: mean squared errors across the first 29 weeks with new degrees of freedom (0.75*d). Lowest value is bolded.*

Again, season8 has the lowest MSE. This means that season eight is the best-fitting season to predict the season 11, so I will predict weeks 50-72 of season 11 with the fitted values of season 8. Below two figures show the prediction and plots using the optimal season (8).

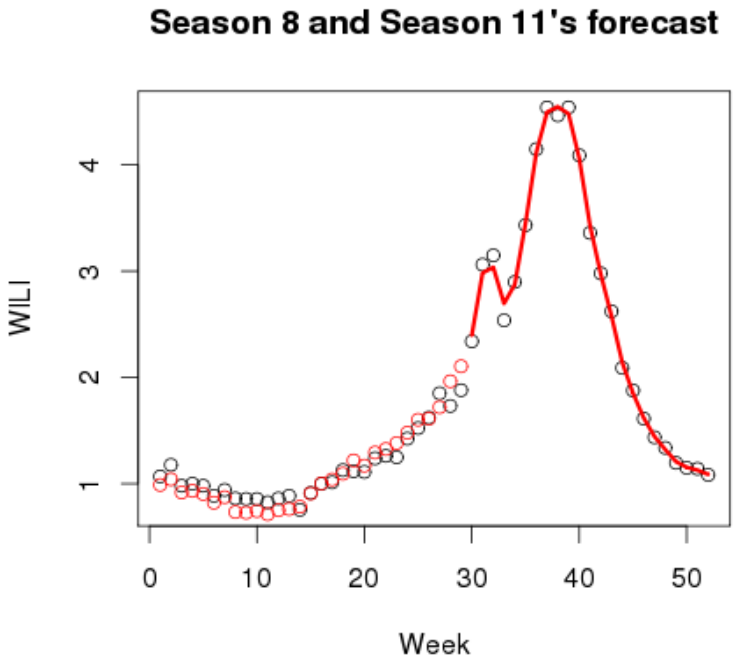

Figure2: Black points show the plot of season 8. Red points show the plot of season 11. The red line shows WILI prediction of season 11.

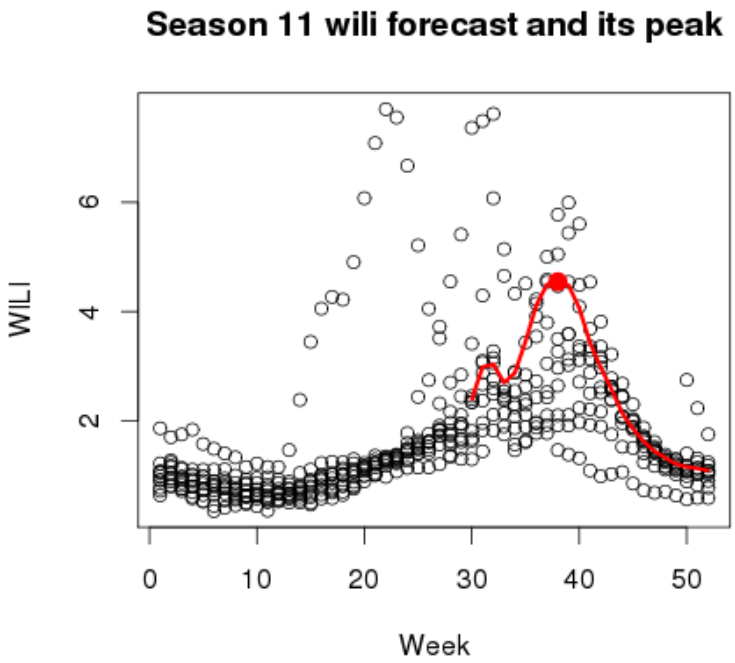

Figure3: WILI prediction of season 11. Red line shows the forecast and the fit to the overall plot. Red fijlled dot shows the maximum value of the curve (i.e. the point estimate of 11th season's peak WILI value)

Assessing forecast error

Now, I want to make an estimate of test error of my forecast to report to CDC how accurate my forecast is. However, because I have no observation over the week 30-52 in season 11, I cannot directly test my model. Thus, I am going to use leave-one-out cross validation over the first 10 seasons to assess the forecast error.

Essentially, I am using the same method I used to build a forecast model for season 11 to forecast the week 30-52 for each of the first 10 seasons. I can train the model with the first 1-29 just like I did for season 11, but this time I can test the model since I do have the test set, week 30-52 values, for the first 10 seasons. Below table shows the ten test errors.

Season	1	2	3	4	5	6	7	8	9	10
Fitted Season	10	4	8	5	9	2	10	3	5	8
Forecast Error	2.8834	1.8299	1.2468	1.2895	2.7728	2.8132	2.0520	1.0645	2.6158	2.7439

Table3: Fitted season shows which season's fitted line had the lowest MSE against each season. Forecast Error shows mean squared errors across the last 23 weeks' forecasts. Mean of all ten forecast errors is 2.13.

2.13, the mean of ten forecast errors, is the expected forecast error of my model as a result of this leave-one-out cross-validation,

Estimating forecast variability

Finally, CDC is particularly interested in the forecasted peak for the coming season—this is the maximum WILI value across the entire season. As shown in figure 3, the forecasted peak of the 11^{th} season is 4.546, which happens during on the 38^{th} week of season 11.

To provide the variability of this point estimate, bootstrap method is used to resample the WILI values of seasons 1 through 11, 500 times. This gives me 500 new data sets of resampled WILI values. By repeating the works I did above with those 500 new data set, I can see the distribution of 500 newly found forecasted peaks and find the standard deviation, which is the forecast variability.

Mean of the 500 newly found forecast peaks is 3.3654 and the standard deviation is 0.8508.

Summary

My analysis has covered forecasting the WILI value of week 30-52 of season 11. And I conclude that season 8 is the optimal year to make the forecast. I also found the forecasted peak during these weeks is 4.546 and discussed about variability of it. One concern I had while I was finding the optimal season is that simply using one year to forecast(though smooth splined) could be risky. I believe that there could be more comprehensive method that could represent better scope of our past data. Hence more samples would have been helpful.

Also, when using 0.75 as the multiplier to the degrees of freedom in my smooth spline, I could have formalized a better method in the future, called the "one standard error rule" instead of simply multiplying 0.75 to the original degrees of freedom.