Kevin H Shu (kshu)
36-402
6 March 2014

Midterm I

1 Introduction

Every week, physicians across the United States participating in a Center for Disease Control (CDC)-sponsored program submit reports to the CDC documenting the proportion of patients they treated with influenza-like illness (ILI); in turn, due to the varying composition of patients reporting ILIs, the CDC weights the ILI percentage measurements into weekly ILI (WILI) values that indicate how widespread the flu is in the U.S. during a particular week. The time of maximum occurrence of flu has implications for the utilisation of flu-fighting resources, such as vaccines in the autumn months. As averaged across 11 (weeks 21–49) or 10 (weeks 50–72) seasons Figure 1 displays

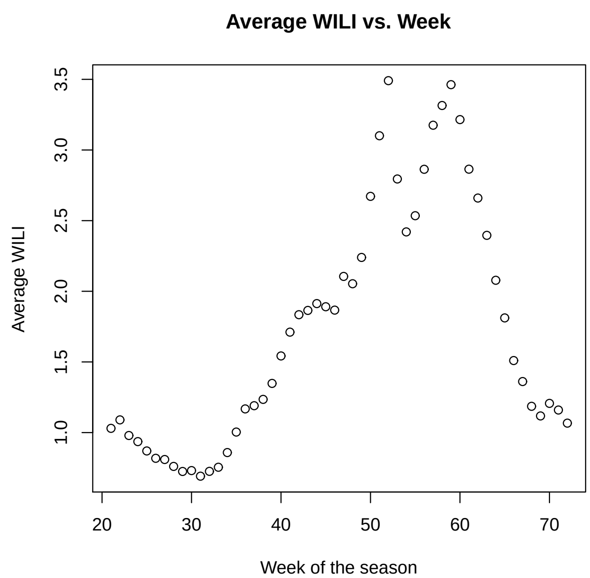


Figure 1: Scatterplot illustrating the relationship between average WILI and the week of the season

an evident trend: cases of flu reach their lowest in mid-summer before rising steadily in autumn and especially so in November, before decreasing sharply with warming temperatures. However, as alluded to previously, season 11 has data lacking for weeks 50–72, and it is the author's goal to provide a "forecast" for those 23 weeks and assess the error and variability of said forecast.

2 Forecast season 11

A "leave-one-out" cross-validation smooth-splining, where, for $i \in [10]$, the data from season i's weeks 21–49 were compared against the season 11 observed week 21–49 values, was used to select the number of degrees of freedom, and based on this model output, the author refit the smoothing spline estimate on $\frac{3}{4}$ the degrees of freedom. The season with the least mean squared error was chosen to be the eighth, with an error of 8.82×10^{-3}:

Season	Mean squared error
2003–04	1.0130
2004–05	0.0742
2005–06	0.0102
2006–07	0.0234
2007–08	0.0210
2008–09	0.1190
2009–10	8.3653
2010–11	0.0088
2011–12	0.0297
2012–13	0.0616

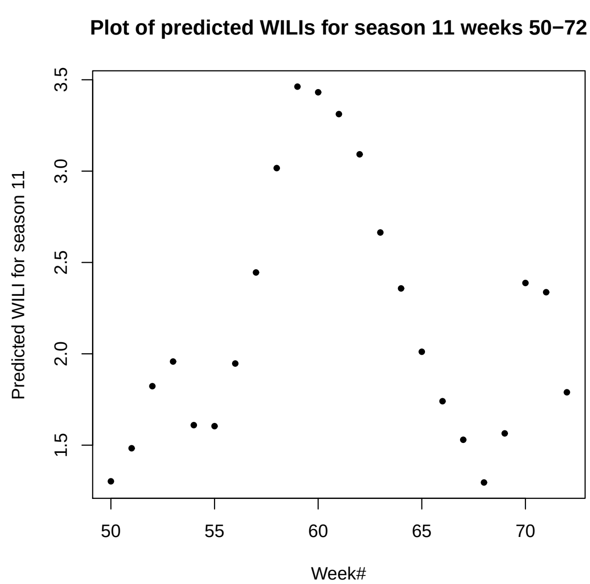


Figure 2: A forecast of WILI values for season 11 weeks 50–72 using smoothing spline model based off season 8

3 Assessing forecast error

Now, instead of comparing smoothing spline fits based off weeks 21–49 in seasons 1 thru 10, for each season i, the author evaluated the models based on weeks 50–72 in the other nine seasons for which week 50–72 was available, executing the procedure described in the previous section 10 times. The mean of the minimum errors was chosen as an estimate of e_{test}, which was found to be 0.5081:

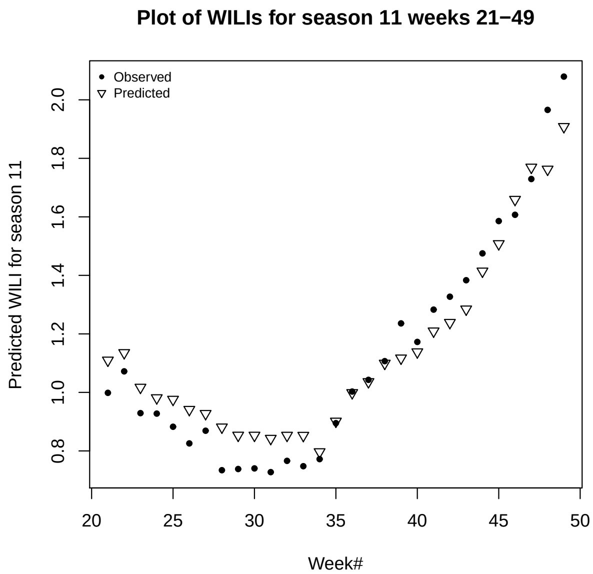
Season	Min CV error
2003–04	2.6090
2004–05	0.1319
2005–06	0.1001
2006–07	0.0977
2007–08	0.1917
2008–09	0.4095
2009–10	0.1501
2010–11	0.1481
2011–12	0.1594
2012–13	1.0836

4 Estimating forecast variability

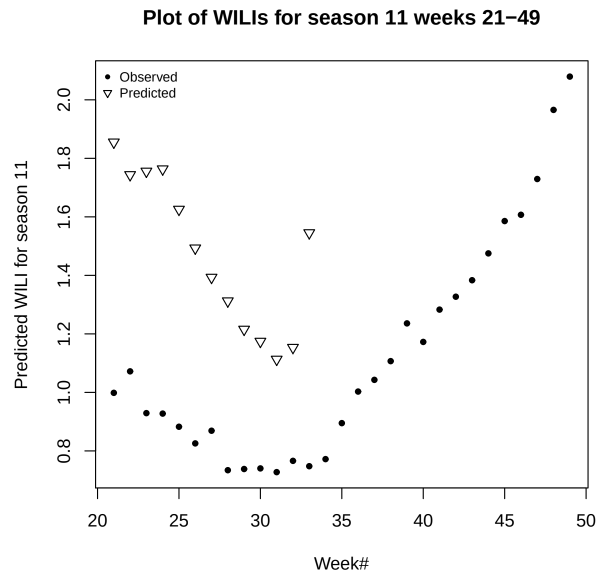
To estimate the forecast variability, for each of the seasons, a smooth spline model was fit to the observed WILI values and the model's residuals, as compared to the respective season's observed values, were randomly rearranged using bootstrapping. Next, the forecasting procedure (using smoothing spline based on weeks 21–49 of seasons 1 to 10 to predict the same weeks in season 11) adopted in Section 2 was repeated. This overall was executed 1000 times to reduce noise, and one run produced a standard variation of 0.6238. Hence, the estimate $\tilde{y}^{(11)}_{\max} \pm \delta = 4.5482 \pm 0.6238$

```
B=1000
peakVec=vector(mode="numeric",length=B)
for(b in 1:B)......
> sd(peakVec)
[1] 0.6237685
```

5 Discussion



(a) Predicted vs. Observed for the best-performing season's (2010–11) model



(b) Predicted vs. Observed for the worst-performing season's (2009–10) model

By inspection, Figure 3 shows that both the best and worst seasons had a significant and consistent bias in the summer/early autumn months, suggesting deficiencies with the smooth-splining. Also, note the error predicted by 2009–10 is particularly large, which may be due in part to the autumn and winter of that season being particularly cool and wet.