S670 Problem set 5

Erik Parker

February 23, 2017

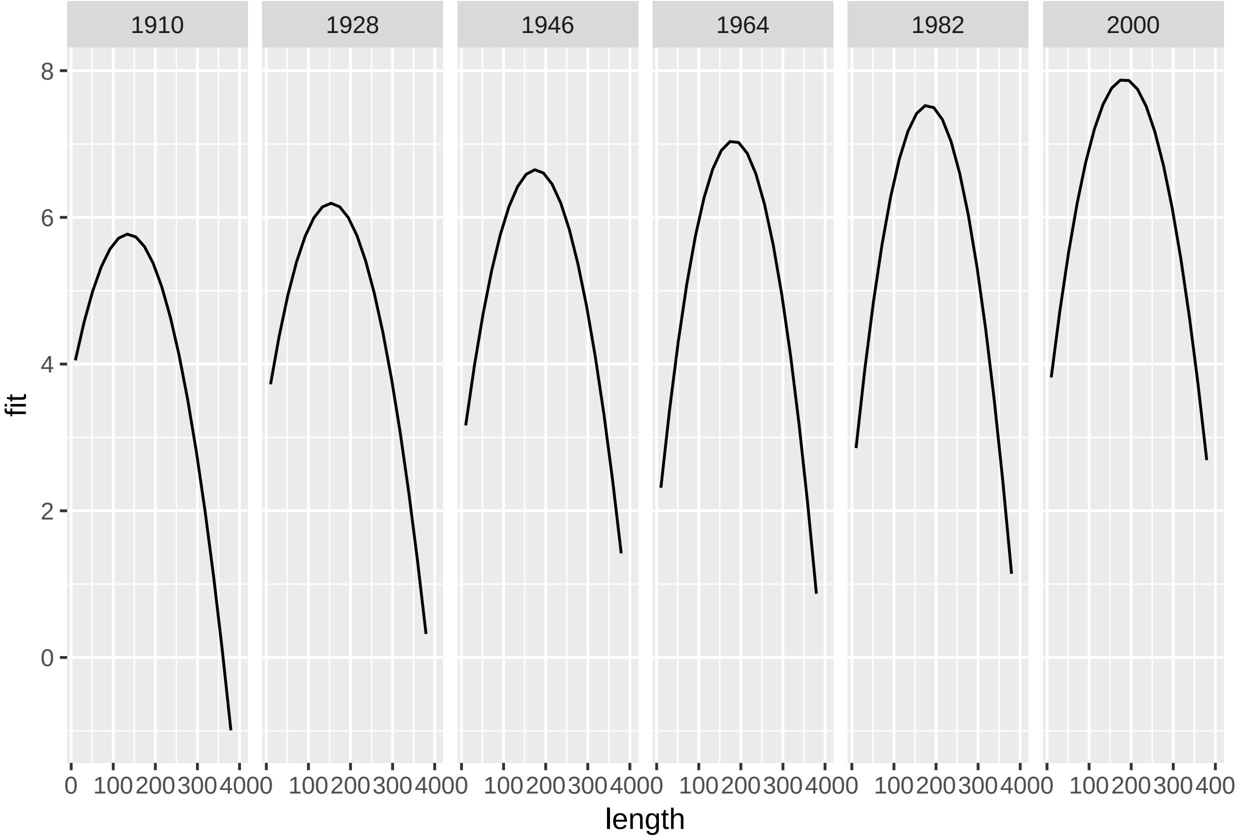
Model

```
model <- loess(log.budget ~ year * length, data = movie_budgets, degree = 2,
    family = "symmetric", drop.square = "year", parametric = "length", span = 0.3)
```

Model Justifications

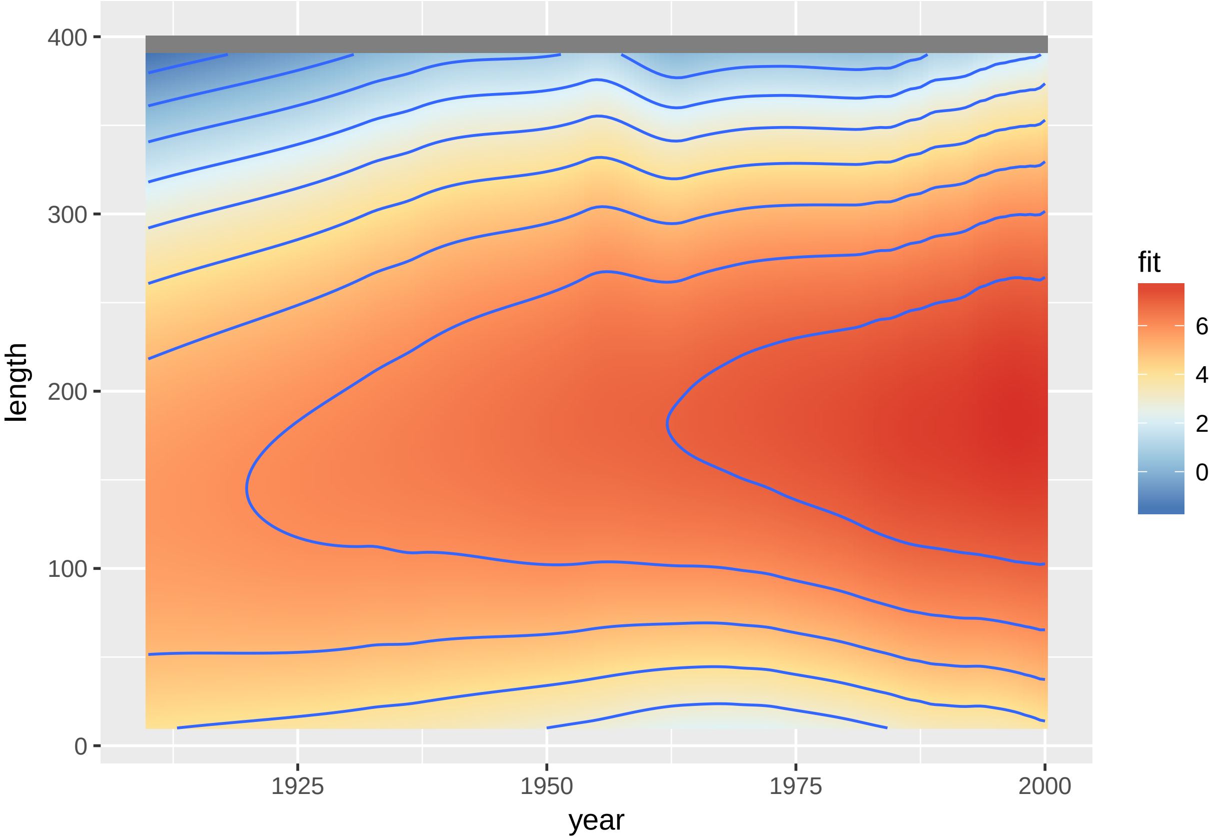
So, to get the best fitting model I possibly could I fit an interaction between year and length as the residual plots without the interaction showed a lot of weird, uneven curving, suggesting that an interaction was necessary to explain more of the variation. I used a robust symmetric fit as there seemed to be a significant number of outliers, and a robust fit lead to better behaved residuals. Interestingly, to get the best fit possible, I fit a parametric model to the length term, and I dropped the square term from the year variable. In my interpretation, doing these seperately leads to both length and year being fit with curved functions still - just non standard curved functions. I think fitting a global, parametric function for the length variable improved the fit because there were some locally weird outlying datapoints in the length by budget distribution when it was plotted, so fitting globally rather than locally in that term reduces the impact of those points. Honestly though, I'm not too sure what exactly the drop.squares argument did for the year variable, but specifying it did improve the fit of the residual plots while also increasing the amount of variation explained by the model in the residual-fit plot. Sticking with the running theme here, a span of .3 was chosen as it lead to the best outcome for the two types residual plots examined - but the choice also makes good sense based on the large amount of observations in this dataset.

Question 2



The above plot showing the predicted model fit of movie budget against movie length as faceted on year shows a pretty clear relationship of genearlly increasing movie budgets over time (across the year facets), and also a relationship of increasing budget with length, up until a certain point (between 150 and 180 minutes), after which budget decreases.

Question 3



The above contour plot doesn't seem to be providing any new information that the faceted fit plots above didn't already give us, but it is (in my opinion) a much more clear and better way to visualize the data. In this plot it is pretty clear that regardless of the year, there has always been a high budget sweet spot for a length between roughly two and three hours - among the movies represented in this dataset at least. Perhaps the movies represented here that are much longer or shorter than that are either too short to have much need for a large budget, or experimental artsy films not intended for general audiences (I mean, a 5+ hour long movie?!) which were not able to secure the financing required for a large budget production.