Problem Set 4

Raj Thakkar

February 20, 2019

Note: I worked with Bhavna Sinha and Varun Miranda from the class

Using loess or otherwise, fit a model to predict log10(budget) from year and length. For simplicity, do not transform year and length (even though a transformation of length would probably be sensible.)

Should you fit a linear or curved function for year?

Should you fit a linear or curved function for length?

Do you need an interaction between year and length?

What span should you use in your loess smoother?

Should you fit using least squares or a robust fit?

Some of these choices are clear-cut, while others will be a matter of preference. Either way, you must justify all your choices.

```r
library(tidyverse, quietly = TRUE)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.2
```

```
## -- Attaching packages ------------------------------------------------------------------------------
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.0.1     v dplyr   0.7.6
## v tidyr   0.8.2     v stringr 1.4.0
## v readr   1.1.1     v forcats 0.3.0
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
## Warning: package 'tibble' was built under R version 3.5.2
```

```
## Warning: package 'tidyr' was built under R version 3.5.2
```

```
## Warning: package 'stringr' was built under R version 3.5.2
```

```
## -- Conflicts ---------------------------------------------------------------------------------- tid
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

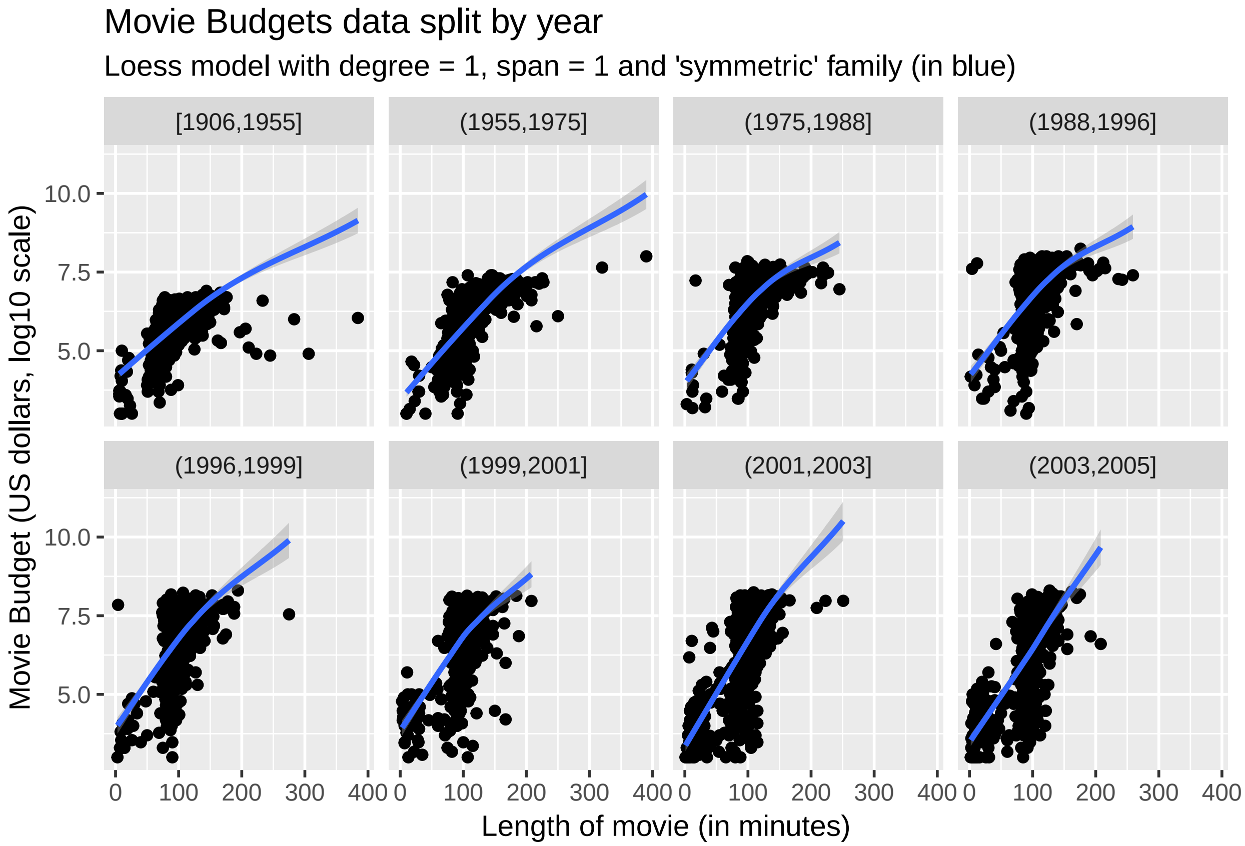
```r
movie_budgets <- read.csv("movie_budgets.txt", sep = "")
movie_budgets$log_budget <- log10(movie_budgets$budget)
```

1. We observe that log_budget increases with increase in length but the relationship is non monotonic as the log_budget variable becomes almost constant after length = 200 conditional on year. Because of this we will ft a curved model (loess) for length conditional on year. The non monotonic of relationship between log_budget and length is only because of the outliers which are movies with length > 200.
2. We observe that the shape of curve fitted for different cuts on years are almost similar except when there are outliers i.e. the movies with length > 200. Thus we don't need an interaction between year and length.

3. We have used a span of 1 for loess as we wanted the curve to not capture the trend in outliers i.e movies with length > 200.
4. Since, we have outliers in the dataset, we have used "symmetric" family in loess which is a robust fit version of loess.
5. Since correlation between log_budget and year is very low. We will use length to predict the log_budget conditional on year.

2. Draw ONE set of faceted plots to display the fit either condition on year or length, whichever seems to you to be more interesting. Choosing a sensible number of panels. Briefly describe what this set of plots shows you.

```
# Cuts on Years
gg = ggplot(movie_budgets, aes(x = length, y = log_budget)) + geom_point() +
  geom_smooth(method = "loess", method.args = list(degree = 1, family = "symmetric"),
    span = 1) + facet_wrap(~cut_number(year, n = 8), ncol = 4)
gg + labs(title = "Movie Budgets data split by year", subtitle = "Loess model with degree = 1, span = 1
  xlab("Length of movie (in minutes)") + ylab("Movie Budget (US dollars, log10 scale)")
```



Observations from the faceted plots of the fit of Movie Budget in log10 scale conditional on year:

1. The curves are similar for all the years except [1906,1955] and (1955,1975] which have outliers (movies with length > 200)
2. The span of 1 gives us an ideal fit which is not affected a lot by the outliers (movies with length > 200)
3. The relationship between movie budget in log10 scale and length is not monotonic. There is a point of diminishing return when length = 200 as the budget on log10 scale decreases for length > 200

3. Draw a raster-and-contour plot (or other 3D plot of your choice) to further display your fit. Briefly describe what, if anything, this plot shows you that your plot for question 2 didn't.

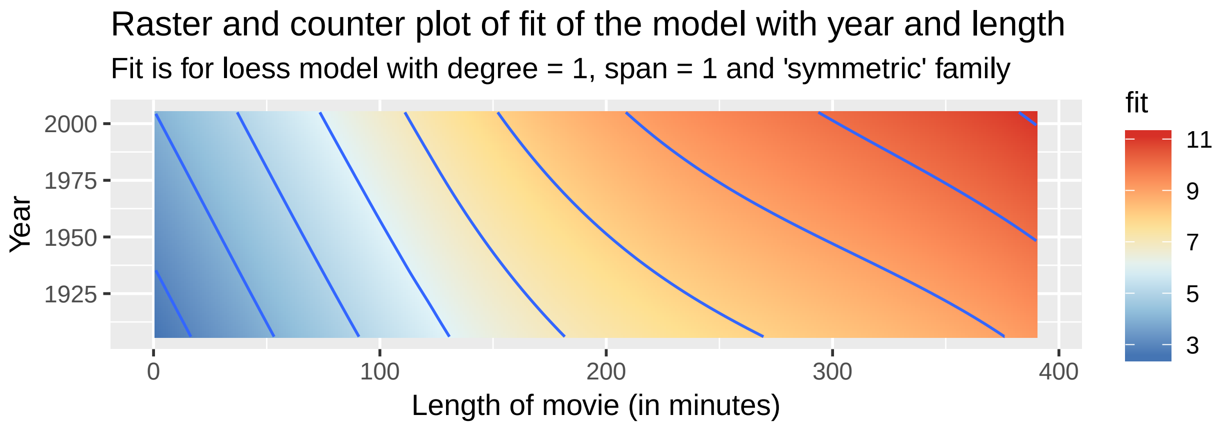
```
budget.lo = loess(log_budget ~ length + year, data = movie_budgets, degree = 1,
    family = "symmetric", span = 1)

movies.grid = expand.grid(year = seq(1906, 2005, 1), length = seq(1, 390, 1))
movies.predict = predict(budget.lo, newdata = movies.grid)




# Dataframe
movies.plot.df = data.frame(movies.grid, fit = as.vector(movies.predict))

ggplot(movies.plot.df, aes(x = length, y = year, z = fit)) + geom_raster(aes(fill = fit)) +
    coord_fixed() + scale_fill_distiller(palette = "RdYlBu") + geom_contour() +
    labs(title = "Raster and counter plot of fit of the model with year and length",
        subtitle = "Fit is for loess model with degree = 1, span = 1 and 'symmetric' family") +
    xlab("Length of movie (in minutes)") + ylab("Year")
```



Raster and counter plot of fit of the model with year and length

Fit is for loess model with degree = 1, span = 1 and 'symmetric' family

We see that model captures the fact that the budget on log10 scale of a movie doesn't increase a lot when the length increases beyond 200. We can see that the value is same for a movie released in 2000 when length of movie increases from 300 to 390 and the value of fit is high i.e. around 11.