Problem Set 4
By Sumeet Mishra

Discussed with Kimball Wu, Prahasan Gadugu

Q1:
1(a) Should you fit a linear or curved function for year?

We tried both linear and loess model(curved model) for data. I observed,there is a constant increase in the budgets in log scale over the years as described by the linear model. But the loess model described the data well by capturing the decrease in budgets in log scale around the year 1996 and after for which I would choose curved function over linear for year.

1(b) Should you fit a linear or curved function for length?

The linear model describes the data by ignoring the outliers when it comes to length of the movie. The loess model describes the outliers well. I would choose curved function over linear one for length of the movie.

1(c) Do you need Interaction between year and length?

I observed that,we did not get any new information even if we take interaction between length and year into consideration. So, I would not take interaction between them into consideration.

1(d) What span should you use in your loess smoother?

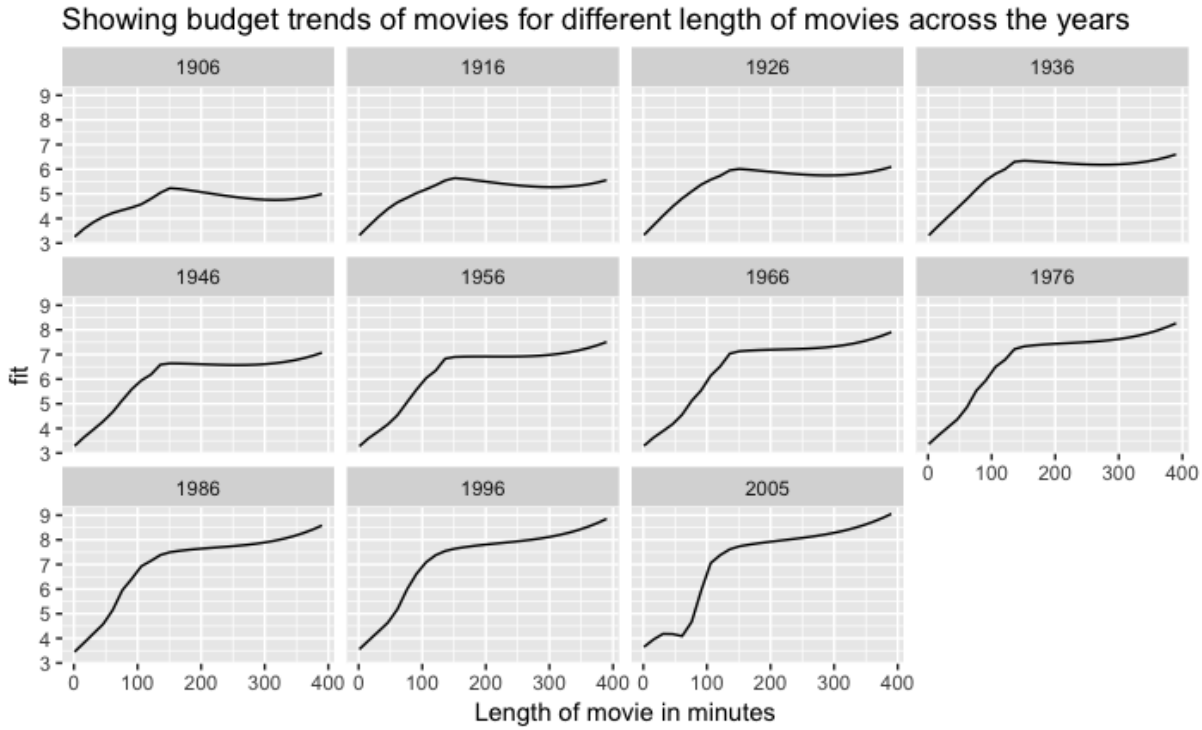
I have tested with span 0.25,0.5,1 and decided to use span 1 as -
1.Span 0.25 and 0.5 overfitting the data points.
2.The plots are smoother with span=1.

1(e) Should you fit using least squares or a robust fit?
I used 'Symmetric' family of loess model which is robust fit. The reason of choosing this was it captures the outliers well(when there was too many outliers).

Question 2

```{r}
library(ggplot2)
movie_budgets=read.table('movie_budgets.txt',header=TRUE)
```

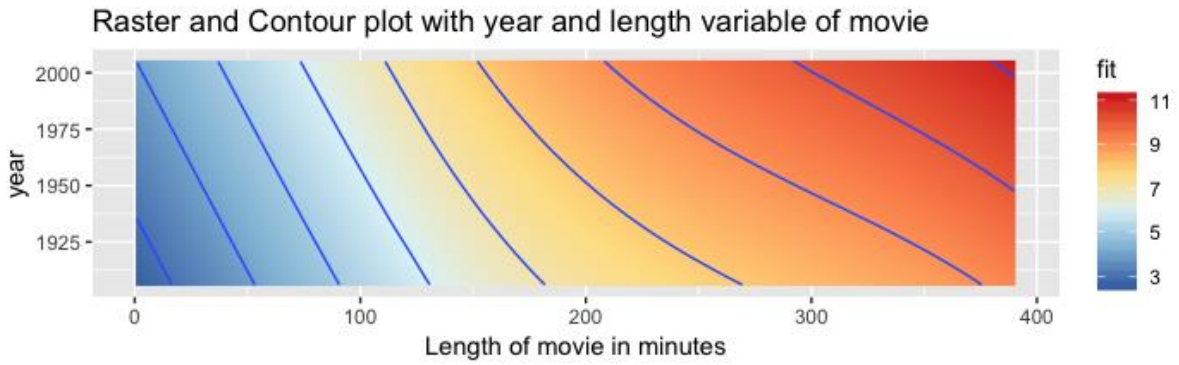
```{r}
movie_budgets1.grid = expand.grid(year=c(seq(1906,2005,10),2005),
length=c(seq(1,390,15),390))
movie_budgets1.lo = loess(log10(budget) ~ year*length, data = movie_budgets, span = 0.25,
                          family="symmetric", normalize=FALSE)
movie_budgets1.predict = predict(movie_budgets1.lo, newdata=movie_budgets1.grid)
movie_budgets1.plot.df = data.frame(movie_budgets1.grid,
fit=as.vector(movie_budgets1.predict))

ggplot(movie_budgets1.plot.df, aes(x=length, y=fit)) + geom_line() +facet_wrap(~year, ncol
= 4)+
  xlab('Length of movie in minutes')+
  ggtitle('Showing budget trends of movies for different length of movies across the years')
```



The plot was drawn fit ~ length faceted over years. The fit variable was generated by our movie_budget1 predict which was built over loess model of movie_budget1. For the predict model we have created a grid model consisting of 2 variables year and length. The year variable was observed between 1906 and 2005 with a interval of 10 years where as the length variable was observed between 1 and 390 with an interval of 15 minutes. I observed that, the fit variable increased initially a lot with increase of length but then steadily increased when length is more than 150.

Question 3

```{r}
movie_budgets.grid = expand.grid(year=seq(1906,2005,1), length=seq(1,390,1))
movie_budgets.lo = loess(log10(budget) ~ year*length, data = movie_budgets, span =
1,degree=1,
                  family="symmetric", normalize=FALSE)
movie_budgets.predict = predict(movie_budgets.lo, newdata=movie_budgets.grid)
movie_budgets.plot.df = data.frame(movie_budgets.grid,
fit=as.vector(movie_budgets.predict))
ggplot(movie_budgets.plot.df, aes(x=length, y=year, z=fit)) + geom_raster(aes(fill = fit))+
  coord_fixed() + scale_fill_distiller(palette="RdYlBu") +
  geom_contour()+xlab('Length of movie in minutes')+
  ggtitle('Raster and Contour plot with year and length variable of movie')
```



The plot was drawn length ~ year. The fit variable was generated by our movie_budget predict which was built over loess model of movie_budget. For the predict model we have created a grid model consisting of 2 variables year and length. The year variable was observed between 1906 and 2005 with a interval of 1 years where as the length variable was observed between 1 and 390 with an interval of 1 minutes. Apart from what we observed in the above plot(Q2) , I observed that the intervals between contours was more when the length of the movie increases more than 150 minutes, which means the length of movie does not make a huge contribution towards the budget if it is more than 150 minutes. The fit variable showed us how strong it could be when the length of the movie increased beyond 300 minutes towards the later years.