EDA - Assignment 5

Jivitesh Poojary

February 28, 2017

Answer 1

Q1. Use loess to 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.) You will have to make a number of modeling choices:

- 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.

ANSWER: - The LogBudget versus year is a curved plot in Loess, this shows that there is no direct linear relationship Logbudget and year

- The LogBudget versus length shows a almost linear relationship, however the length versus the Logbudget is a curved plot - this shows that there is no direct linear relationship Logbudget and year
- Because of the above two observations we will not be using a parametric model
- Facetted plots were drawn with conditioning on year and length. It was observed that the slopes varied a lot for Lodbudget versus year facetted with respect to length. Similary for Lodbudget versus length facetted with respect to year varied but the change was relatively small. Because of this we can say that there is a good chance that there is some interaction happening between year and length in the model.
- The span value was varied from 0.1 to 0.6, it was observed that for span value 0.10 to 0.25 we obtained a reasonably good fit without overfitting the data. The plots for residual vs length and residual vs year appears to be linear with slope 0 and centered around the residual value of 0. In our model we have used the span as 0.10.
- It appeared that there few outliers in the data when we modelled logbudget with length ,that may affect our model. Thus a robust fit seems to be a good option.

```
## Call:
## loess(formula = LogBudget ~ Length + Year + Length * Year, data = movie_budgets,
##     span = 0.1, family = "symmetric")
##
## Number of Observations: 5183
## Equivalent Number of Parameters: 60.68
## Residual Scale Estimate: 0.6484
## Trace of smoother matrix: 72.45  (exact)
##
## Control settings:
##   span     :  0.1
##   degree   :  2
##   family   :  symmetric      iterations = 4
```

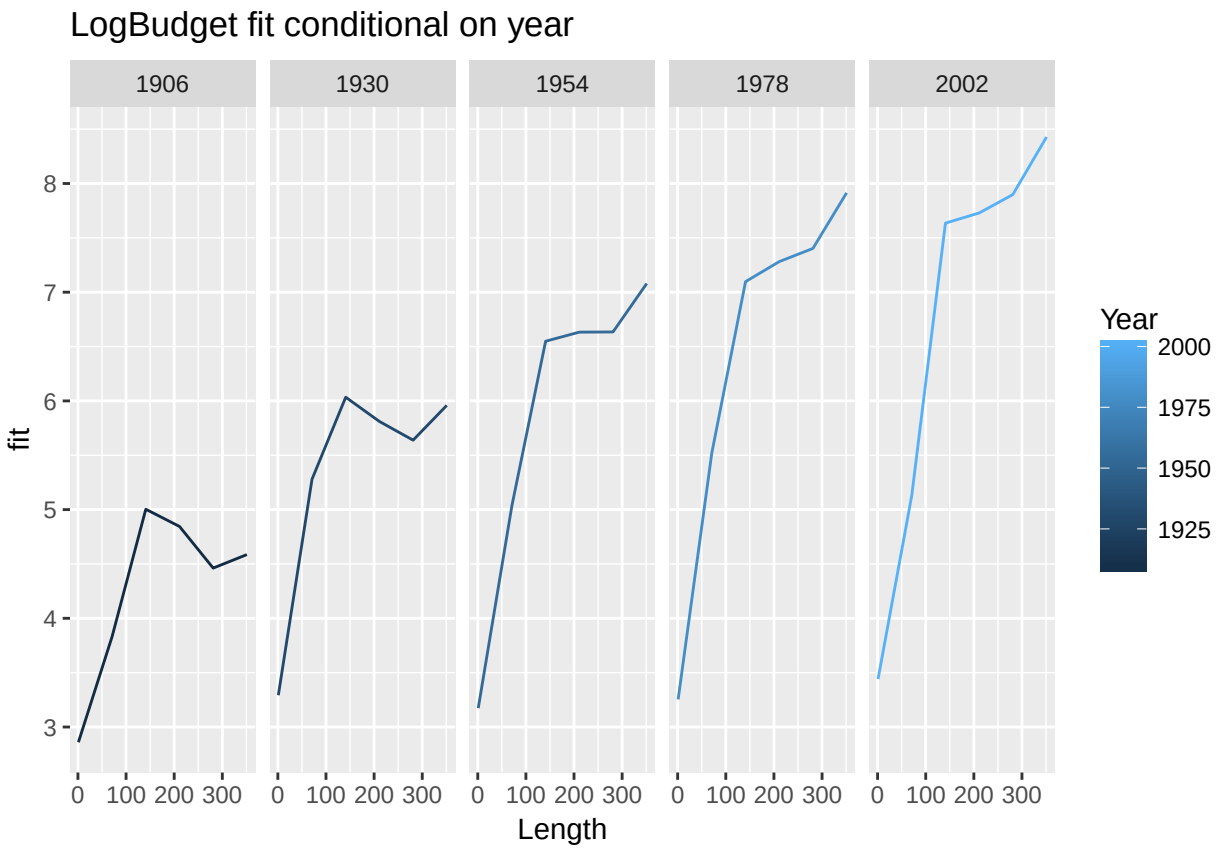
```
##   surface  :  interpolate      cell = 0.2
##   normalize:  TRUE
##  parametric:  FALSE FALSE
## drop.square:  FALSE FALSE
```

Answer 2

Q2. 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.

ANSWER: - The facetted fit plots with condition on year or length are different form each other. An show different patterns as the budget changes.

- The below plots are of logbudget fitted values with condition on Year. The plot is facetted into 5 subplots to better represent the variation.
- We observe that for the fitted values increase as there year facet increases. Also the shift of the curve is upwards
- Similarly we can also see that there is general pattern in the curves, the fitted values for each year block increase around 150.
- After 150 we see that the curve decreases for earlier years while for latter years the curve increase with a small slope
- There is another inflection point at around 280 where the starts increasing again this time with a larger slope



Answer 3

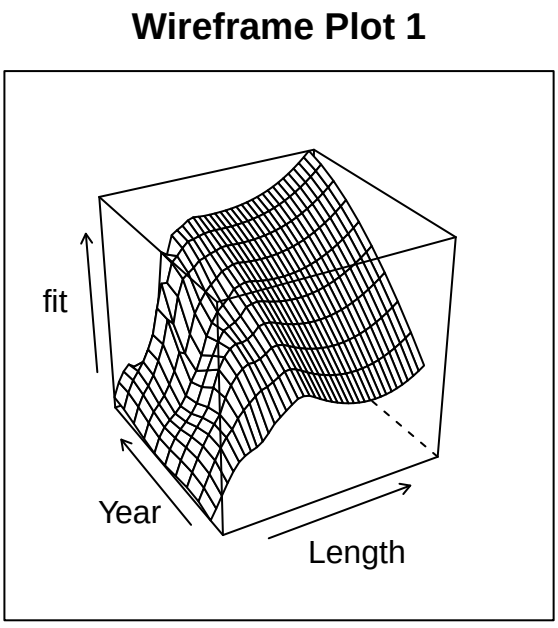
Q3. Draw EITHER a contour plot or a wireframe plot (whichever you prefer) to further display your fit. Briefly describe what, if anything, this plot shows you that your plot for question 2 didn't.

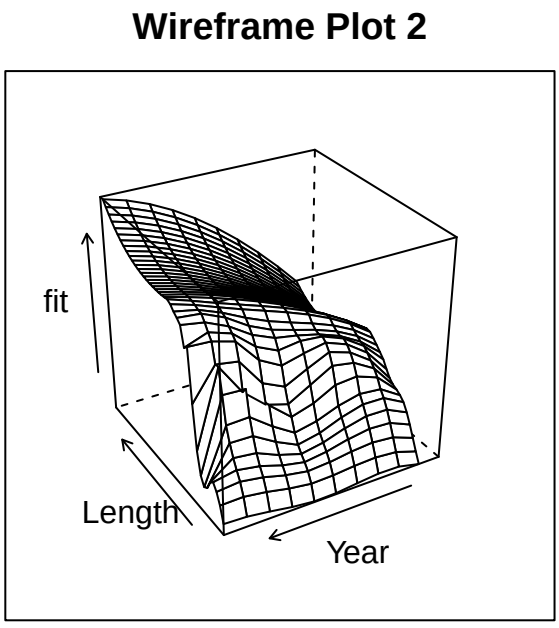
ANSWER: - Both the contour plots and the wireframes were drawn in this case.

- The contour plot gives us an idea about the density of the distribution. However it does not convey too much information as the density changes can be observed without the contour plot overlay.
- However the wireframe plots drawn from different positions give us a better understanding of the curve that is obtained from our fitted model.
- We see that the data varies with both length and year. The distriution is not linear for any single variable.
- It starts from the smallest values and the maximum value is reached at almost the maximum point in the curve.
- This makes us think that with each passing year the budget increase for the same length of the movie.
- The curve increases first then plateau for some time and then increases slowly to reach the maximum value.
- Another important but misleading observation was when we look at the wireframe plot 3 below, we see only a small variation with respect to change in year. This perception is because of the angle of view and the idea is disproved when we observe the curve from different angles.

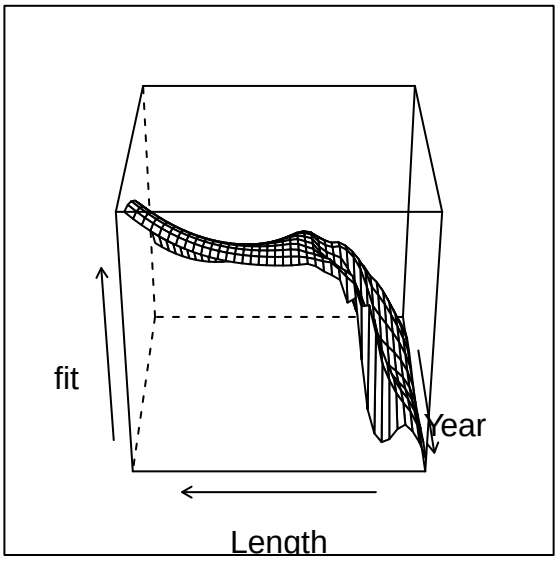
```
## Range of Years: 1906 2005
```

```
## Range of movie lengths: 1 390
```

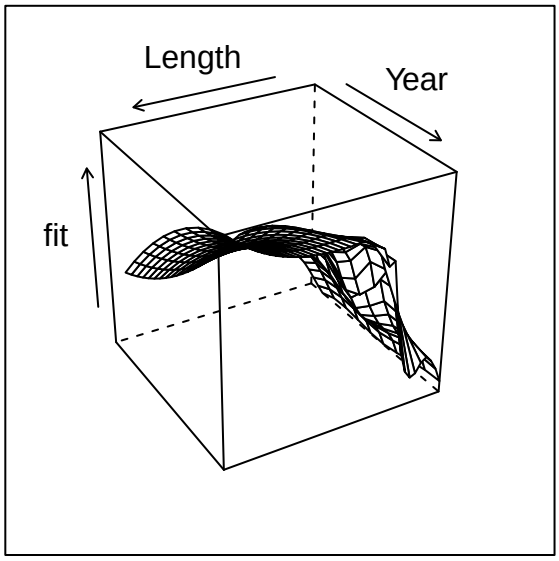




Wireframe Plot 3



Wireframe Plot 4



Wireframe Plot 5