Problem Set 4

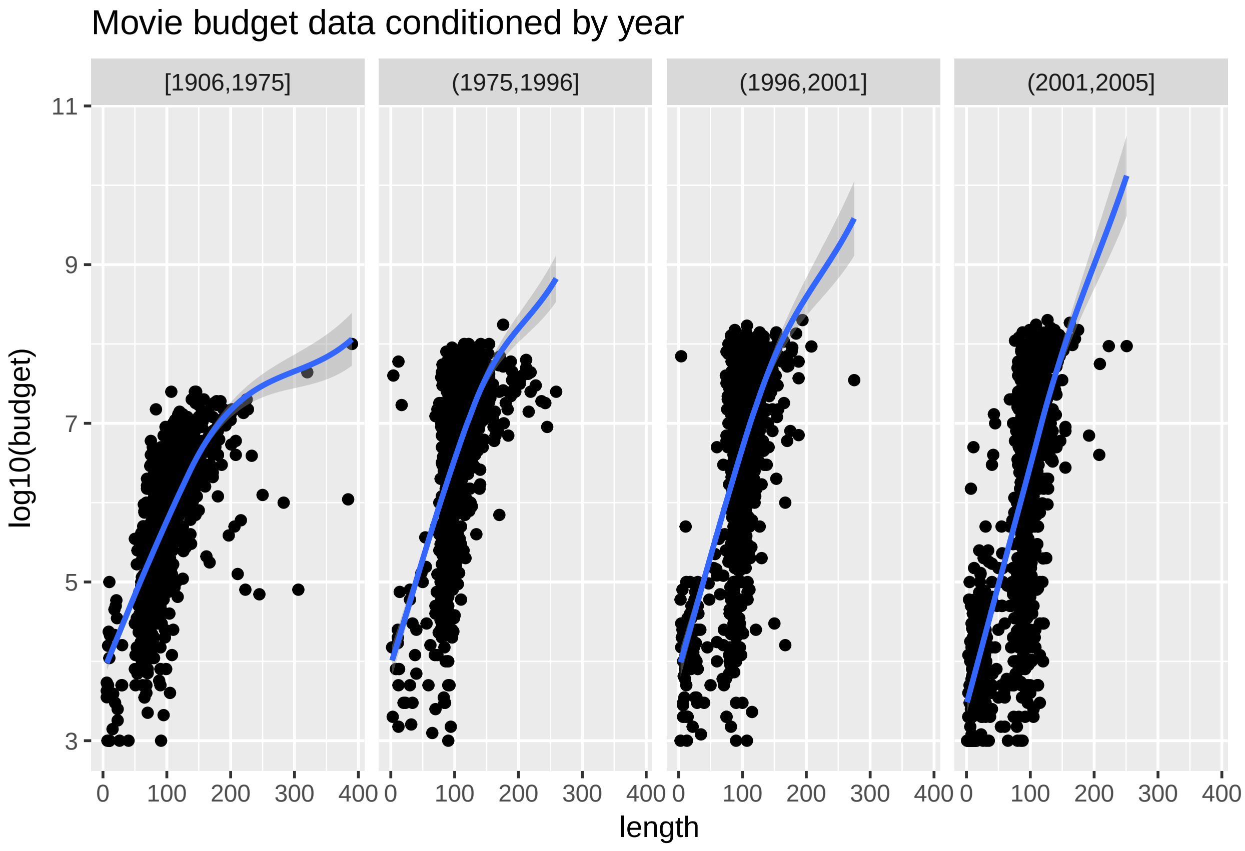
Ashwed Patil, Vatsal Jatakia, Akshay Naik, Saniya Ambavanekar (Team Hong Kong)

February 19, 2018

Introduction and Conditional Plot

The dataframe **movie_budgets** contains 4 measurements on 5183 movies. We are interested in studying log10(budget) as our response variable and year and length as the explanatory variables. To explore the trivariate relationship, we decided to create a conditional plot of budget and length faceted by year. We will cut the year variable into four categories, plot scatterplots of budget against length and add loess smoothers.

```
movie_budgets = read.table("movie_budgets.txt", header = TRUE)
ggplot(movie_budgets, aes(x = length, y = log10(budget))) + geom_point() +
    geom_smooth(method = "loess", span = 1, method.args = list(degree = 1)) +
    facet_grid(~cut_number(year, n = 4)) + labs(title = "Movie budget data conditioned by year")
```



There's a similar increasing relationship in each panel. In the first panel, the curve somewhat tends to bend down because of the presence of a few outliers. We can conclude that the trend here is monotonic and it seems that a loess fit would be adequate.

LOESS Model

Our exploration lead to some suggestions for modeling the LOESS fit:

- Because all of the coplot lines have similar slopes, there's no obvious need for interaction between year and length
- Because there's a significance presence of outliers can affect the linear relationship, we prefer using the robust fit. After several trial and error methods, a span of 1 resulted in the best possible option for the fit.

R Code for LOESS Fit

```
movie.lo = loess(log10(budget) ~ length + year, span = 1, family = "symmetric",
    data = movie_budgets)
movie.lo.df = augment(movie.lo)

summary(movie.lo)
```

```
## Call:
## loess(formula = log10(budget) ~ length + year, data = movie_budgets,
##     span = 1, family = "symmetric")
##
## Number of Observations: 5183
## Equivalent Number of Parameters: 6.51
## Residual Scale Estimate: 0.7291
## Trace of smoother matrix: 7.03  (exact)
##
## Control settings:
##   span     :  1
##   degree   :  2
##   family   :  symmetric      iterations = 4
##   surface  :  interpolate      cell = 0.2
##   normalize:  TRUE
##  parametric:  FALSE FALSE
## drop.square:  FALSE FALSE
```

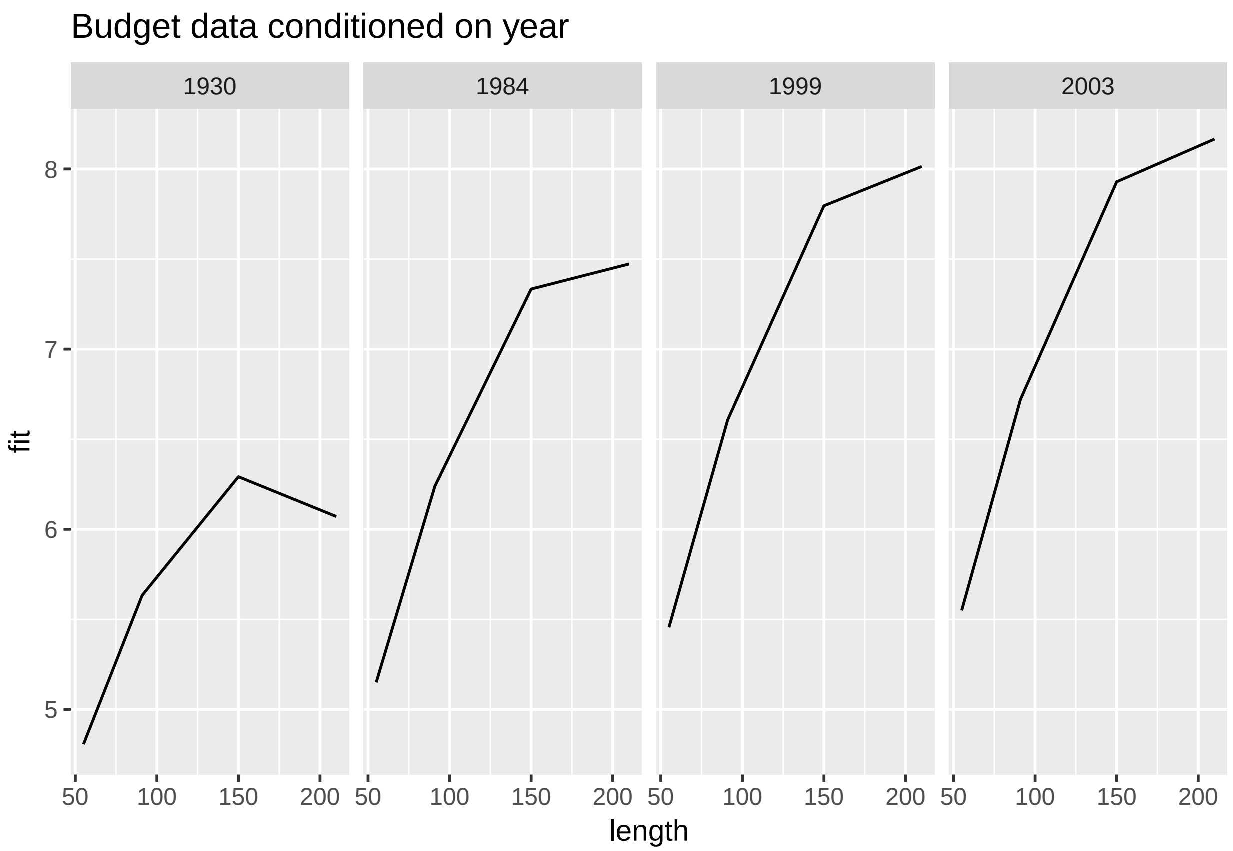
Visualizing the fitted model

To visualize the fitted surface, we want to plot a set of predictions for a grid of different values of year and length.

```
movie.grid = expand.grid(year = c(1930, 1984, 1999, 2003), length = c(55,
    91, 150, 210))
movie.predict = predict(movie.lo, newdata = movie.grid)
```

Suppose we want to see how the fit depends on length, conditioning on different values of year. We did this using a coplot.

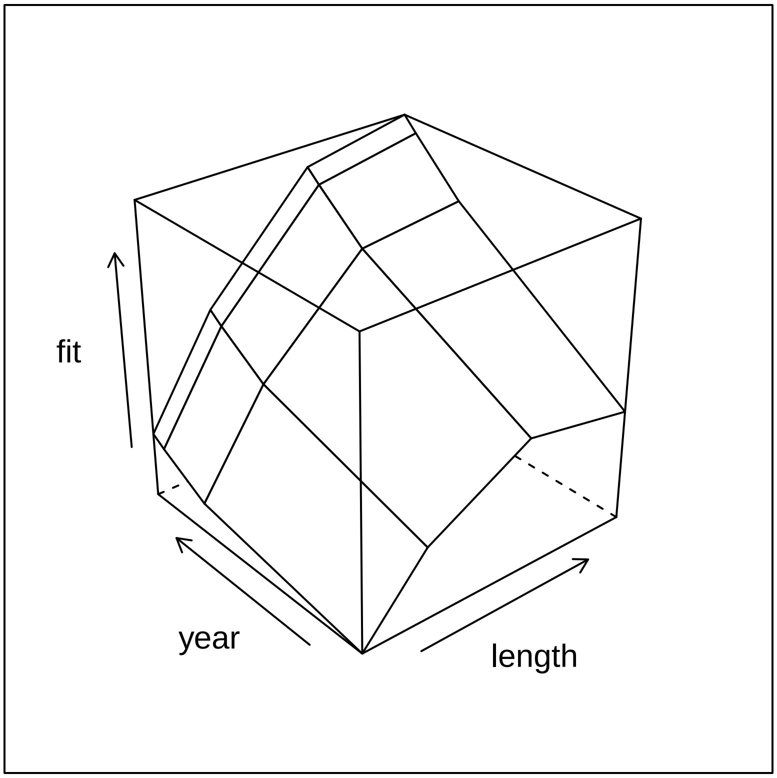
```
movie.plot.df = data.frame(movie.grid, fit = as.vector(movie.predict))

ggplot(movie.plot.df, aes(x = length, y = fit)) + geom_line() +
    facet_grid(~year) + labs(title = "Budget data conditioned on year")
```



Thus, the trend seems to similar for all the years. There seems to be an increasing relationship between budget and length until the movie length is 150 mins beyond which the curve flattens down. The bend in the curve in the first panel is probably due to outliers, but this isn't our concern as of now.

3D Plot

```r
library(lattice)
wireframe(fit ~ length + year, data = movie.plot.df)
```



We observe that the as length and year increases, the corresponding fitted budget also increases upto a certain point (around 150) after which it becomes almost constant.