Problem Set 4

Khusaal Giri Vaishnavi Srinivasan Ayesha Bhimdiwala Harsh Mehta

2/15/2018

```
options(scipen = 999)
library(ggplot2)
library(tidyr)
library(GGally)
library(MASS)
library(broom)
library(gridExtra)
library(plot3D)
```

Section 01:

Log of Budget and Year

By experimenting with both the linear as well as the loess model, we came to the conclusion that we should fit a loess curve as it tends to better depict the trend of the relationship between Year and Budget.

Also by fitting a loess model and visualizing the residual plot based on the explanatory variable Year, we see that the loess curve just wiggles around the zero line which to an extent says that the loess is a good model to fit.

```
###Budget and Year
  movies = read.table("movie_budgets.txt",header = TRUE)
  movies$logbudget <- log10(movies$budget)
  movies = subset(movies, movies$length<=180)
  gg.lo=ggplot(movies,aes(x=year,y=logbudget))+geom_smooth(span=0.5, method.args = list(degree = 1))+ge

  movies.loyear =loess(logbudget~year ,data = movies, span = 0.5)
  movies.loyear.df =augment(movies.loyear)
  movies.ggloyear=ggplot(movies.loyear.df,aes(x = year, y = .resid))+ geom_point()+ geom_smooth(method=
```

```
cat("The variance explained by the model is", var(movies.loyear.df$.fitted)/var(movies$logbudget)*100,"
```

```
## The variance explained by the model is 14.28752 %
```

Log of Budget and Length

By experimenting with both the linear as well as the loess model, we came to the conclusion that we should fit a loess curve as it tends to better depict the trend of the relationship between Length and Budget.

Also by fitting a loess model and visualizing the residual plot based on the explanatory variable, Length (in minutes), we see that the loess curve just wiggles around the zero line which to an extent says that the loess is a good model to fit whereas for the linear model the residual it tends to overfit the residuals of the linear model fit.

```
##Budget and length
  gg.lo=ggplot(movies,aes(x=length,y=logbudget))+geom_smooth()+geom_point()
```

```
  movies.lolen =loess(logbudget~length ,data = movies, span = 0.5)
  movies.lolen.df =augment(movies.lolen)
  movies.gglolen=ggplot(movies.lolen.df,aes(x = length, y = .resid))+ geom_point()+ geom_smooth()+geom_
```

```
cat("The variance explained by the model is", var(movies.lolen.df$.fitted)/var(movies$logbudget)*100,"%
```

```
## The variance explained by the model is 50.2043 %
```

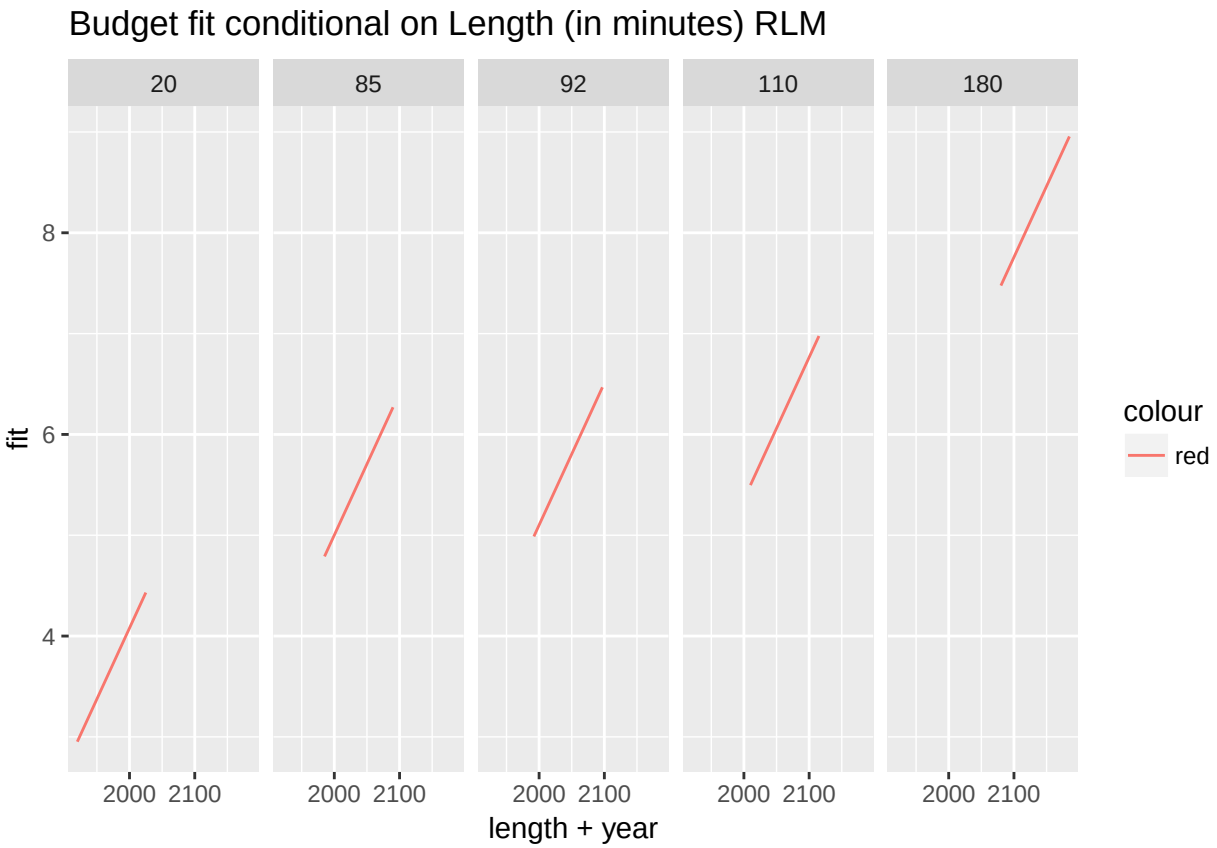
The R code for the loess models are:-

```
movies.loyear =loess(logbudget~year ,data = movies, span = 0.5)

movies.lolen =loess(logbudget~length ,data = movies, span = 0.5)
```

Since nearly all of the Hollywood movies have a length of fewer than 180 minutes (3 hours), we have filtered the data based on this condition (i.e. removed all movie lengths greater than 180 mins). Based on the co-plots and residual fit plots, we found no obvious need for interaction between the terms. While fitting using loess model, we have used a span of 0.5 as it provides a reasonably smoother curve. We will be using least square for model fit because gross outliers have been taken care of and the model performs similarly to robust linear model with outliers.

Section 02: Faceted Plot

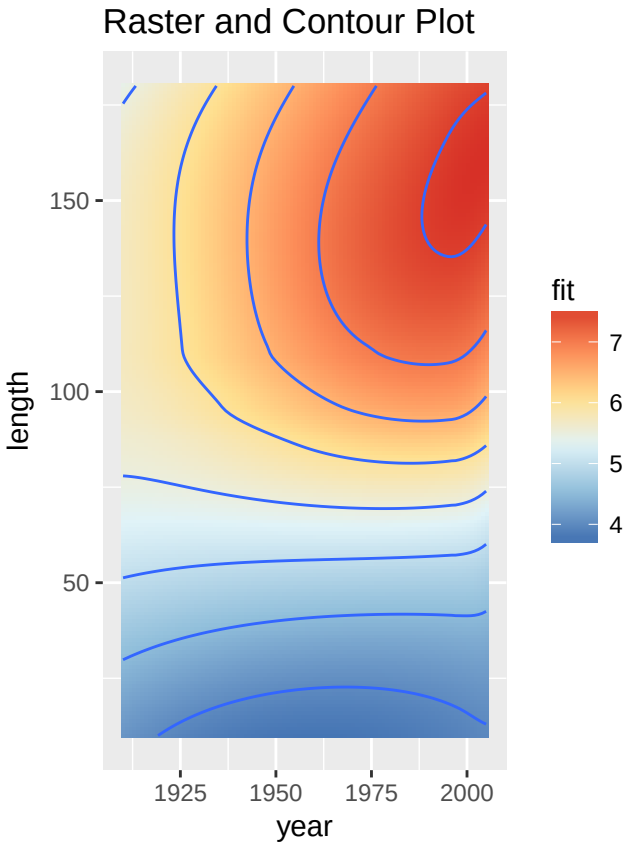
```
movies.grid = expand.grid(length=c(20,85,92,110,180), year=c(1900,1950,1990,2000, 2005))
movies.rlm = rlm(logbudget ~ length+year, data = movies)
mov.predict.rlm = predict(movies.rlm, newdata = movies.grid)

gg = ggplot(data.frame(movies.grid, fit = as.vector(mov.predict.rlm)), aes(x = length+year,
y = fit, color="red")) + geom_line() + facet_grid(~length)
gg + labs(title = "Budget fit conditional on Length (in minutes) RLM")
```



On comparing our robust linear model, linear model, and loess model fits, we find the robust linear model to show reasonably plausible results. Based on length and year - we found that length of fewer than 20 mins to have a lower budget. Similarly, for movies of length between 80 to 180 minutes, the fit showed an average budget and for movies with a length greater than 180 minutes, the budget rose sharply because movies of such length over all the years are rare and demand more budget.

Section 03:

```
movies.grid.3d= expand.grid(length = seq(10, 180, 1), year = seq(1910,
2005, 1))
movies.lo = loess(logbudget ~ length+year, data = movies, psi = psi.bisquare)
mov.predict.3d = predict(movies.lo, newdata = movies.grid.3d)
mov.plot.df = data.frame(movies.grid.3d, fit = as.vector(mov.predict.3d))
ggplot(mov.plot.df, aes(x = year, y = length, z = fit)) + geom_raster(aes(fill = fit)) +
coord_fixed() + scale_fill_distiller(palette = "RdYlBu") + geom_contour() + labs(title="Raster and Cont
```



A raster and contour plot displays the fit of the model better than the set of faceted plots in Question 2. Since contours join together the points that have the same value of the response variable, it gives us a better picture of all the regions having high or low peaks for a different combination of the explanatory variables. Also, because of the interaction in the loess model in question 2, for the faceted plot, split (cut) is made on the explanatory variable, whereas in the raster and contour plot, color is used for the response variable. Finally, the set of faceted plots is based on the grid levels of the length of the movies; which restricts the exploration of the fitted model.Raster and contour plot, on the other hand, provide a continuous relation between the response variable & the explanatory variables; describing the trend and the appropriateness of the fitted model in great depth.