Final Project 321.Rmd

Jenn Jung-Hye Hwang

October 30 2022

This code chunk allows R Markdown to knit your output file even if there are coding errors.

knitr::opts\_chunk$set(error = TRUE)

The packages we need available for this program are: mosaic, dplyr, and olsrr.

library(mosaic)

## Registered S3 method overwritten by 'mosaic':  
## method from   
## fortify.SpatialPolygonsDataFrame ggplot2

##   
## The 'mosaic' package masks several functions from core packages in order to add   
## additional features. The original behavior of these functions should not be affected by this.

##   
## Attaching package: 'mosaic'

## The following objects are masked from 'package:dplyr':  
##   
## count, do, tally

## The following object is masked from 'package:Matrix':  
##   
## mean

## The following object is masked from 'package:ggplot2':  
##   
## stat

## The following objects are masked from 'package:stats':  
##   
## binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,  
## quantile, sd, t.test, var

## The following objects are masked from 'package:base':  
##   
## max, mean, min, prod, range, sample, sum

library(dplyr)  
library(olsrr)

##   
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':  
##   
## rivers

favstats(~revenue, data=Movie2)

## Error in favstats(~revenue, data = Movie2): object 'Movie2' not found

# read in data file  
Movie <- read.csv("movies.csv")  
Movie2 <- Movie %>% filter(revenue>0)  
Movie2 <- Movie2 %>% mutate(logrev = log(revenue))

library(readxl)  
library(lubridate)

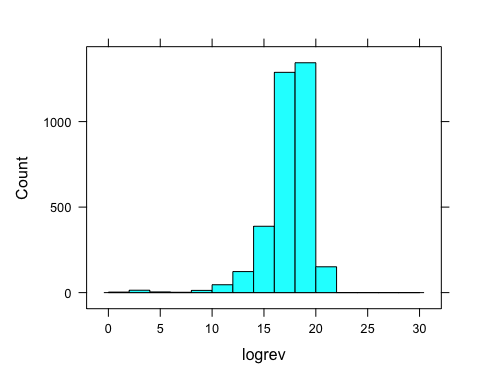
## Loading required package: timechange

##   
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':  
##   
## date, intersect, setdiff, union

library(dplyr)  
Moviex <- read\_excel("tmdb\_5000\_movies.xlsx")  
  
Movie2 <- Moviex %>% filter(revenue>0)  
Movie2 <- Movie2 %>% mutate(logrev = log(revenue))  
  
Movie2 <- Movie2 %>% mutate(year = year(release\_date),   
 month = month(release\_date),  
 day = day(release\_date))

#Histogram   
histogram(~logrev ,data = Movie2, type="count", breaks=seq(0, 30, 2))



# numerical summary of y value revenue   
favstats(~revenue,data = Movie2, na.rm=TRUE)

## min Q1 median Q3 max mean sd n missing  
## 5 15352895 51751835 140165096 2787965087 117031353 183483090 3376 0

FM <- lm(revenue~ runtime + budget + vote\_average + popularity + vote\_count + year + month + day + L1.e + L1.f + G1.d + G1.c + G1.a + G1.h, data = Movie2 )

## Error in eval(predvars, data, env): object 'L1.e' not found

ols\_plot\_resid\_qq(FM, print\_plot = TRUE)

## Error in check\_model(model): object 'FM' not found

ols\_plot\_resid\_stud\_fit(FM, print\_plot = TRUE)

## Error in check\_model(model): object 'FM' not found

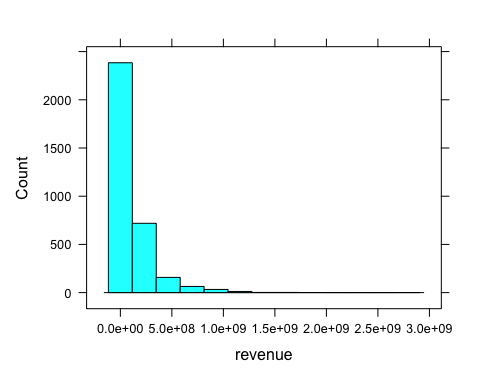
ols\_plot\_resid\_stud\_fit(FM, print\_plot = TRUE)

## Error in check\_model(model): object 'FM' not found

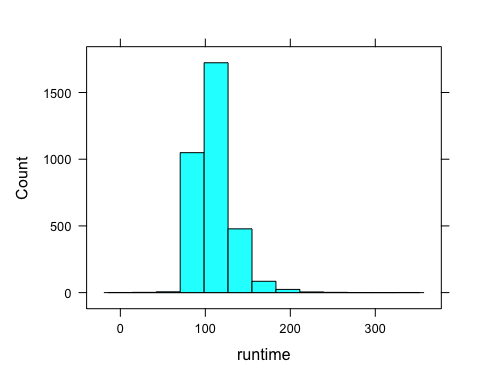
ols\_plot\_cooksd\_chart(FM, print\_plot = TRUE)

## Error in check\_model(model): object 'FM' not found

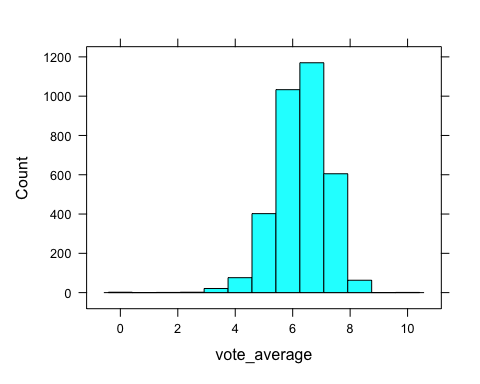
# histogram of y value revenue and qualitative variables   
  
histogram(~revenue, data = Movie2, type="count")



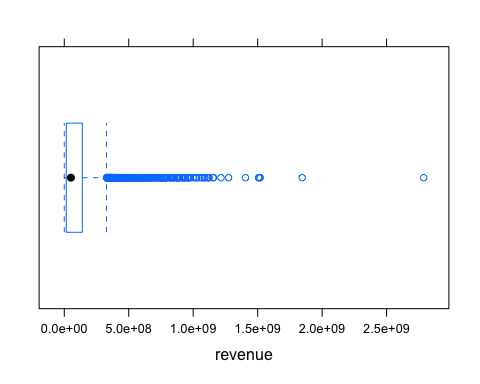
histogram(~runtime, data = Movie2, type="count")



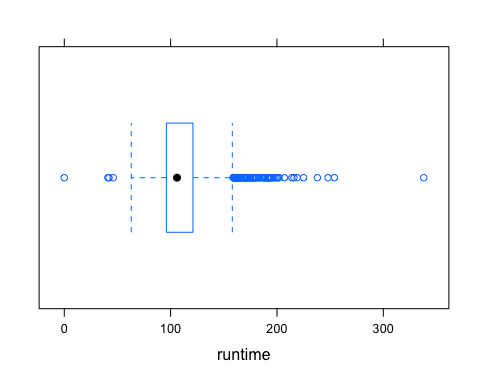
histogram(~vote\_average, data = Movie2, type="count")



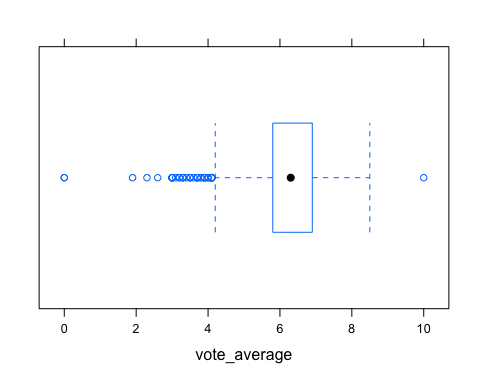
bwplot(~revenue, data = Movie2)



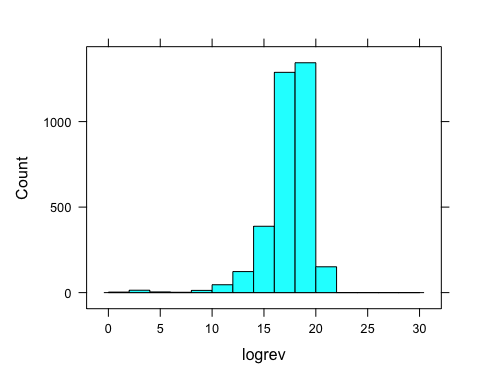
bwplot(~runtime, data = Movie2)



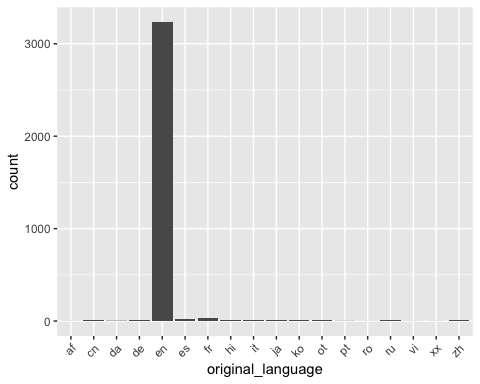
bwplot(~vote\_average, data = Movie2)



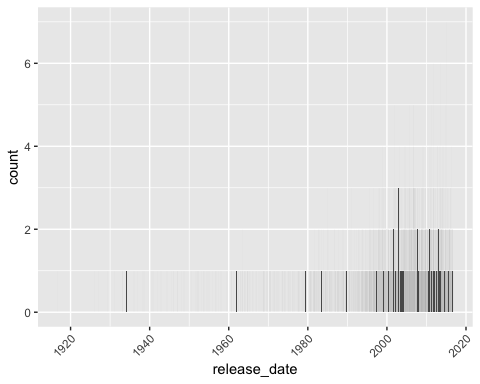
#Histogram   
histogram(~logrev,data = Movie2, type="count", breaks=seq(0, 30, 2))



#histogram of categorical variables   
  
ggplot(Movie2, aes(x = `original\_language`)) +  
 geom\_bar() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))

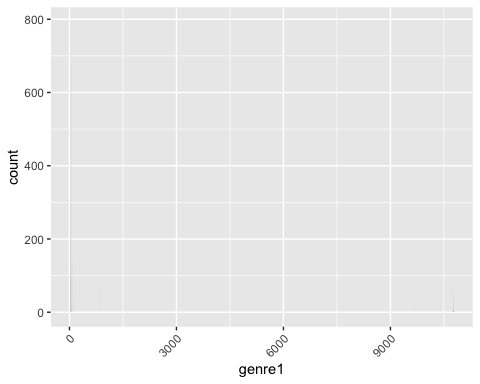


ggplot(Movie2, aes(x = `release\_date`)) +  
 geom\_bar() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))



ggplot(Movie2, aes(x = `genre1`)) +  
 geom\_bar() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))

## Warning: Removed 1 rows containing non-finite values (stat\_count).



Movie2 <- Movie2 %>% mutate(L1.e = ifelse(original\_language == "en", 1, 0),  
 L1.f = ifelse(original\_language == "fr", 1, 0))

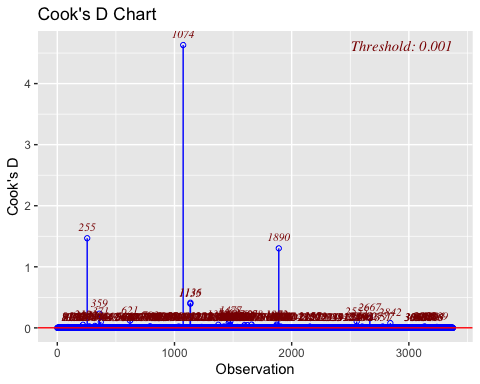
Movie2 <- Movie2 %>% mutate(G1.d = ifelse(genre1 == "Drama", 1, 0),  
 G1.c = ifelse(genre1 == "Comedy", 1, 0),   
 G1.a = ifelse(genre1 == "Action", 1, 0),   
 G1.h = ifelse(genre1 == "Horror", 1, 0))

#interaction   
Movie2 <- Movie2 %>% mutate(I.1 = vote\_average\*runtime)

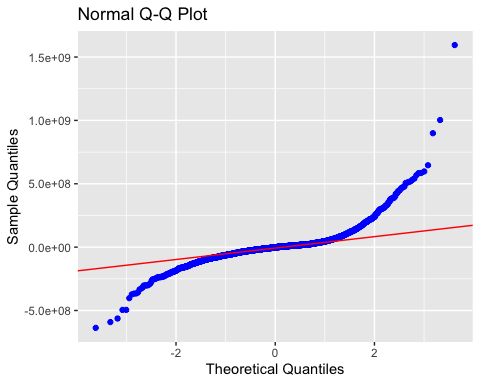
#Step-wise selection gives this reduced model  
RM <- lm(revenue~ budget + vote\_count + popularity, data = Movie2)  
Movie2 <-Movie2[-c(1074,255,1890,1136),]

#Full Model  
FM <- lm(logrev~ runtime + budget + vote\_average + popularity + vote\_count + year + month + day + L1.e + L1.f + G1.d + G1.c + G1.a + G1.h, data = Movie2 )  
  
#L1.e + L1.f + G1.d + G1.c + G1.a + G1.h  
#

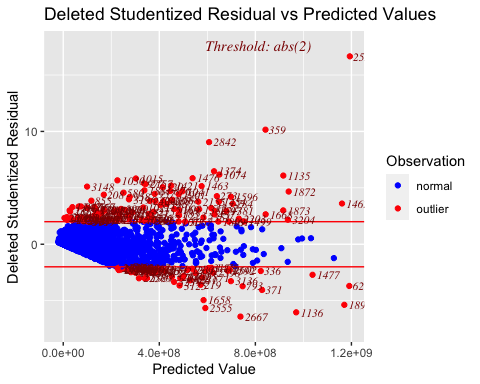
#Cooks d  
ols\_plot\_cooksd\_chart(RM, print\_plot = TRUE)



#Normal Probability Plot  
ols\_plot\_resid\_qq(RM, print\_plot = TRUE)



#Assumption checking   
ols\_plot\_resid\_stud\_fit(RM, print\_plot = TRUE)



#Multiple Partial F-Test  
# To get sum of squares Fit Full model, then   
anova(FM)

## Analysis of Variance Table  
##   
## Response: logrev  
## Df Sum Sq Mean Sq F value Pr(>F)   
## runtime 1 573.6 573.6 179.3216 < 2.2e-16 \*\*\*  
## budget 1 3858.5 3858.5 1206.2459 < 2.2e-16 \*\*\*  
## vote\_average 1 296.0 296.0 92.5263 < 2.2e-16 \*\*\*  
## popularity 1 690.3 690.3 215.7946 < 2.2e-16 \*\*\*  
## vote\_count 1 27.4 27.4 8.5701 0.00344 \*\*   
## year 1 16.1 16.1 5.0379 0.02486 \*   
## month 1 0.4 0.4 0.1244 0.72436   
## day 1 7.1 7.1 2.2212 0.13622   
## L1.e 1 82.0 82.0 25.6462 4.322e-07 \*\*\*  
## L1.f 1 0.3 0.3 0.1018 0.74965   
## Residuals 3360 10747.9 3.2   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Overall F-Test   
# Fit Full model, then   
summary(RM)

##   
## Call:  
## lm(formula = revenue ~ budget + vote\_count + popularity, data = Movie2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -637064422 -37820898 -1907243 23049803 1594571379   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.897e+07 2.444e+06 -7.761 1.11e-14 \*\*\*  
## budget 1.727e+00 4.694e-02 36.803 < 2e-16 \*\*\*  
## vote\_count 6.446e+04 2.030e+03 31.758 < 2e-16 \*\*\*  
## popularity 2.815e+05 7.396e+04 3.806 0.000144 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 100700000 on 3372 degrees of freedom  
## Multiple R-squared: 0.6989, Adjusted R-squared: 0.6987   
## F-statistic: 2609 on 3 and 3372 DF, p-value: < 2.2e-16

# To get sum of squares Fit Reduced model, then   
anova(RM)

## Analysis of Variance Table  
##   
## Response: revenue  
## Df Sum Sq Mean Sq F value Pr(>F)   
## budget 1 5.6989e+19 5.6989e+19 5617.498 < 2.2e-16 \*\*\*  
## vote\_count 1 2.2278e+19 2.2278e+19 2195.926 < 2.2e-16 \*\*\*  
## popularity 1 1.4694e+17 1.4694e+17 14.484 0.0001438 \*\*\*  
## Residuals 3372 3.4209e+19 1.0145e+16   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

1 - pf(333.7, 5, 3369)

## [1] 0