Problem Set 4

Noah Estrada-Rand

9/26/2019

Movie Profitability

b)

```
movies <- movies[!is.na(movies$budget),]
movies <- movies[!is.na(movies$gross),]
movies <- movies[movies$budget<4e+8,]
movies$grossM <- movies$gross/1e+6
movies$budgetM <- movies$budget/1e+6
movies$profitM <- movies$grossM-movies$budgetM
movies$cast_total_facebook_likes000s <- movies$cast_total_facebook_likes / 1000
set.seed(2019)
train_indx <- sample(1:nrow(movies), 0.8 * nrow(movies), replace=FALSE)
movies_train <- movies[train_indx, ]
movies_test <- movies[-train_indx, ]
```

c)

Number of rows for train and test sets.

```
nrow(movies_train)
```

```
## [1] 3103
```

```
nrow(movies_test)
```

```
## [1] 776
```

d)

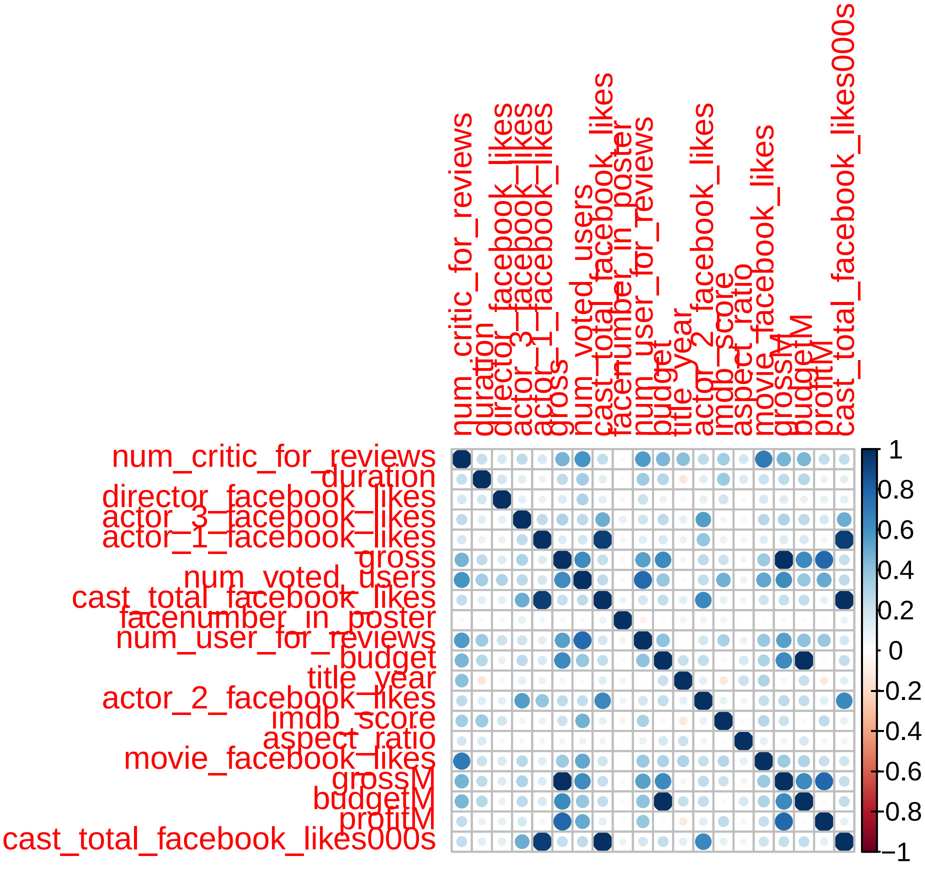
```
nums <- sapply(movies, is.numeric)
cormat <- cor(movies[,nums], use="complete.obs")
print(cormat[,"profitM"])
```

```
##        num_critic_for_reviews                      duration
##                    0.24353361                    0.09423033
##       director_facebook_likes        actor_3_facebook_likes
##                    0.10485194                    0.17831580
##        actor_1_facebook_likes                         gross
##                    0.05850519                    0.78438560
##               num_voted_users     cast_total_facebook_likes
##                    0.50043953                    0.11507040
##          facenumber_in_poster          num_user_for_reviews
##                   -0.02128043                    0.38106102
##                        budget                    title_year
##                    0.02352410                   -0.11615920
##        actor_2_facebook_likes                    imdb_score
##                    0.12969431                    0.25215121
##                    aspect_ratio          movie_facebook_likes
##                     -0.05979073                    0.22941383
##                          grossM                       budgetM
##                      0.78438560                    0.02352410
##                         profitM cast_total_facebook_likes000s
##                      1.00000000                    0.11507040
```

e)

The following is the correlation matrix plot for the movie data.

```
corrplot(cormat)
```



f)

The linear model regressing profit against imdb scores and cast total facebook likes is summaryized below:

```
mod1 <- lm(profitM~imdb_score + cast_total_facebook_likes000s,data = movies_train)
summary(mod1)
```

```
##
## Call:
## lm(formula = profitM ~ imdb_score + cast_total_facebook_likes000s,
##     data = movies_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -384.16  -25.27   -8.76   14.49  495.64
```

```
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -68.64310    5.77272 -11.891  < 2e-16 ***
## imdb_score                    12.01315    0.88830  13.524  < 2e-16 ***
## cast_total_facebook_likes000s   0.33117    0.05769   5.741 1.03e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.06 on 3100 degrees of freedom
## Multiple R-squared:  0.07082,    Adjusted R-squared:  0.07022
## F-statistic: 118.1 on 2 and 3100 DF,  p-value: < 2.2e-16
```

g)

The estimated effect of cast facebook likes is that for every thousand likes increased in cast facebook likes, the profit of the movie incrases by $330,000 dollars.

h)

The pvalue for imdb_score is <2e-16 while the pvalue for cast_total_facebook_likes is <2e-16. Pvalue is the probability of observing the results we got by chance. In this case it is the probability of observing an effect of imdb scores and cast_total_facebook_likes on profit assuming that all other variables are held constant.

i)

The estimate pvalue in this case implies that we can reject the null hypothesis which states that there is no relationship between imdb scores and profit and instead say that imbd score has a statistically significant effect on profit. In this case, both variables are statistically significant at 95% confidence level.

j)

The R^2 is .07082 and the adjusted R^2 is .07022. R^2 indicates how much of the variance in the outcome variable, in this case profit, is explained by the model we have created. In this case, profit is regressed against imdb scores and total Facebook likes in the 1000s, thus R^2 tells us how much of the variation in profit is explained by imdb scores and total cast Facebook likes.

k) The f stat of the model is 118.1. The f stat tells us the significance of all present coefficients. In other words it checks if all coefficients are zero, and if not, then the score goes up. It infomrs us of the significance and existance of variable effects within our linear model.

l)

From the results below, it becomes clear that the amount of residuals is equivalent to the number of rows we have in the train set.

```r
length(mod1$residuals)
```

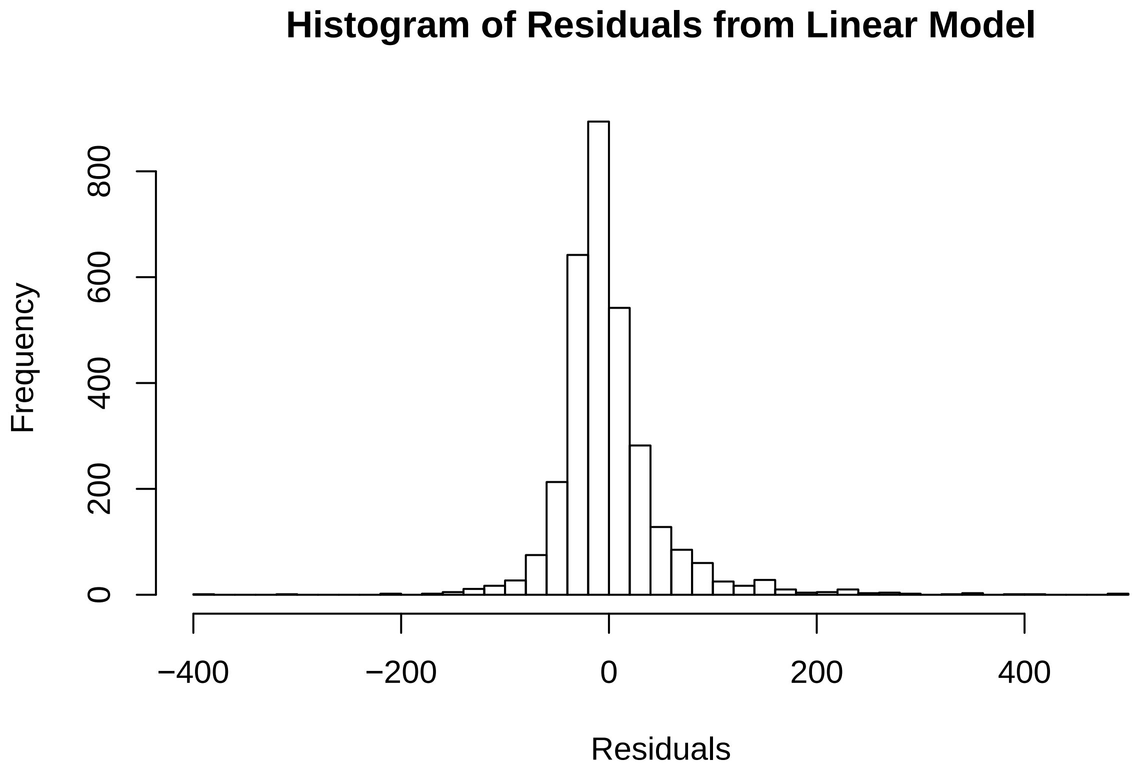
```
## [1] 3103
```

```r
nrow(movies_train)
```

```
## [1] 3103
```

m)

```
hist(mod1$residuals, breaks = 40,
     main = "Histogram of Residuals from Linear Model",
     xlab = "Residuals")
```



This histogram appears to have a normal distrubtion, indicating that our model fits the data well.

Extra Credit n)

The manually calculated R squared value is show below. Steps are split up into total sum of squares and residual sum of squares:

```
tss <- sum((movies_train$profitM - mean(movies$profitM))^2)
rss <- sum((mod1$residuals)^2)
r.squared <- 1-(rss/tss)
print(r.squared)
```

```
## [1] 0.07092309
```