Problem Set 4

MGSC 310

Prof. Shahryar Doosti

Follow the submission instructions stated in the course website

Movie Profitability

a. In this problem set, we will work on a data from kaggle.com (link: https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset) providing information on top 5000 IMDB movies. You can download the data from Canvas as well.

b. Run the code below to filter out films with missing budget and gross and unreasonably large budgets. Also, use the code to create new variables that are simplified versions of the budget variables, and ratings. Also split your data into testing and training sets. Do not forget to use `set.seed(310)` to ensure your results are comparable to mine.

```
# removing missing values of budget and gross
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(310)
train_indx <- sample(1:nrow(movies), 0.8 * nrow(movies), replace=FALSE)
movies_train <- movies[train_indx, ]
movies_test <- movies[-train_indx, ]
```

c. Check the number of rows for testing and training sets.

d. In building a regression model, a good place to start is producing a correlation matrix that shows which variables are positively or negatively correlated with the variable we want to predict. We can only correlate numeric variables so run the code below to produce the correlation matrix. This code does two things: 1) `sapply(movies, isnumeric)` returns the name of variables which are numeric. 2) And, "`use="complete.obs"` in the `cor()` function, removes missing values from correlation matrix. It then prints the correlation coefficient between `profitM` and all the numeric varibles in the data frame. Which variables are most strongly (positivelty or negatively) correlated with profits?

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

e. Use the `corrplot` package to produce a plot of the correlation matrix.

f. Use the training set to regress `profitM` on `imdb_score` and `cast_total_facebook_likes000s` as predictors. Store this model as mod1 and use `summary()` to output the results.

g. What is the estimated impact of cast facebook likes on movie profits? Interpret the estimated effect. Be mindful of the units of variables.

h. What are *p-value* associated with `imdb_score` and `cast_total_facebook_likes000s`? What does *p-value* mean?

i. What does this estimate *p-value* imply about the relationship between `imdb_score` and `profit`? What variables are statistically significant at 95% level of confidence?

j. What is the R^2? What does it mean? What is the *Adjusted* R^2?

k. What is the *F-stat* of the model? In your own words, explain what *F-stat* does.

l. You can access the residuals from the model by `mod1$residuals`. Check the length of the residual vector. Note that the residual vector should have the same length as the train set (for each observation, we can calculate the residual).

m. *Extra Credit* Calculate the R^2 directly from data and model using the equation given in the lecture.