Problem Set 5

MGSC 310

Prof. Shahryar Doosti

Follow the submission instructions stated in the course website

Question 1, Does Increasing a Movie's Budget Ever Pay Off?

a. We are going to work with the movies dataset again. The Top 5000 movies on IMDB is from the following link: https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset. You can download the data set 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. 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.

```
# install.packages(tidyverse)
library(tidyverse)

# removing missing values of budget and gross
movies <- movies[!is.na(movies$budget),]
movies <- movies[!is.na(movies$gross),]

# removing empty content rating or not rated
movies <- movies[(movies$content_rating != "" & movies$content_rating != "Not Rated"),  ]

# removing movies with budget > 400M
movies <- movies[movies$budget<4e+8,]

# simplifying variables
movies$grossM <- movies$gross/1e+6
movies$budgetM <- movies$budget/1e+6
movies$profitM <- movies$grossM-movies$budgetM

# creating new column `rating_simple` using `fct_lump` (from `tidyverse` package)
# to pick 4 major levels and lump all other levels into "Other".
movies$rating_simple <- fct_lump(movies$content_rating, n = 4)

# creating train and test sets
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. Estimate a linear model using the `lm` command with `grossM` on the left hand side, and `imdb_score` and

`budgetM` on the right-hand side. Be sure to estimate on the training set. Use the summary command to show the summary of your model.

d. Interpret the coefficient on `budgetM`. Holding fixed `imdb_score`, does spending more money on movies seem to have a net positive return on movie gross?

e. Now estimate a linear model using the `lm` command with `grossM` on the left hand side, and `imdb_score`, `budgetM` and the square of `budgetM` as independent variables. Be sure to estimate on the training set. Use the summary command to show the summary of your model.

f. Let's investigate the marginal impact of budget for different levels of budget. Use the margins command to calcuclate the marginal impact of an additional dollar of budget at budget levels of 25, 50, 75, 90, 100, 200, and 300 million. For which levels does it make sense to increase your movie's budget?

Question 2, Movie Residuals and Predicted Values

a. Use the movies data and estimate a model predicting movie gross using `imdb_score`, `budgetM`, the square of `budgetM` and `ratings_simple` as independent variables. Use the relevel command to change the reference category of ratings to "R". Print the summary of this regression table.

b. Interpret the coefficient on a movie rated G.

c. Use the `predict` function to generate the predictions in the test and training set.

d. Generate residuals for test and training. Note that residuals is the difference between true and predicted outcome (`grossM`)

e. Plot the residuals against the predicted values in the test and training sets. Do our errors appear homoskedastic or heteroskedastic?

f. Plot the predicted versus true in the test and training set.

g. Use the function below and the `RMSE` function in the package `caret` to calculate in-sample and out-of-sample RMSE. Is the model overfit? And if so, how do we know?

```
rmse <- function(t, p) {
    sqrt(mean((t - p)^2))
}
```