Visualizing Data for Movie Audience Expectation

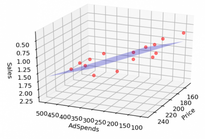
1 Team Members: Jongwon Lee (1)

2 Abstract: This project will visualize data collected for indie film audience expectation done with multiple linear regression. It is a undergraduate level paper that already had been written and got awarded at Sungkyunkwan University in Korea. (I was a member of the writers)

The abstract of the thesis will be provided in English. It collected data and did a multiple regression to predict audience turnout for indie-film that were imported to Korea.

3 Question: Can we visualize the result of multiple linear regression with more than 3 independant variables?

If impossible, we can compare each variable with 3 xyz. For example, like the following with each variable. This is interactive that the viewer can scroll and explore 3 data relation.



4 Datasets and Methods:

I have collected Dataset in 2016 having 300 movies with 5 independent variables. I will be adding more variables (rotton tomato, IMDB) and see if I can improve the regression.

Multiple Linear Regression will be done with scikit-learn. Visualization with seaborn, pandas and etc.

The thesis aimed to explore which variables affect indie film's number of audiences.

Independent Variables:

1. fscreen: first week screen
2. preview: premiere audiences
3. ani: dummy variable (0 or 1), if genre is animation: 1, otherwise: 0
4. naver: https://movie.naver.com/ 's (Korean Portal) user participation point, similar to Tomatometer
5. youtube: views of Korean movie trailer on Yotube

Dependent Variable: Y: fweek: first week audiences

Other Variables:

1. expectation : Y by inserting 5 independent variables to the regression result
2. percent_error : percent_error between fscreen and expectation

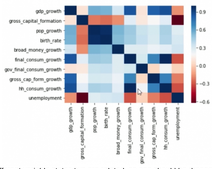
(Y-fweek) = β0 + β1(fscreen) + β2(preview) + β3(ani) + β4(naver) + β5(youtube)

β0 = -7958.917, β1 = 250.855, β2 = 2.934419, β3 = 7606.66, β4 = 9.218283, β5 = 0.099709

There had been rotton tomatoes tomatometer and number of post on naver cafe (closed social network for movie maniacs) before release. Those two variables were deleted after t-test.

5 Objectives: This project will provide a heatmap with seaborn to visualize correlation of data, when we decide to remove variables.

For example, like the following with all variables.



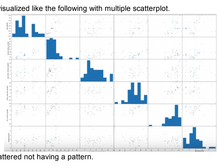
dark blue(1) square with two different variable states two are related, so one should be dropped.

6 Motivation: With just numerical results provided by libraries or Stata, unless the performer is extremely familiar with the regression process or principle, it seems hard to feel what each values mean. For example, when I write a paper with a bunch of statistical results but no visualization, it is hard for the general readers to grasp the idea of the paper. Thesis written by only economist without data experts struggle with this issue.

(I wish I could statistically prove this problem, however it is just common problem in economics department. Professors always try to find programmers or students that can visualize better than what Stata generates. Also, students who read economic journals or thesis suffer with infinite statistics with just sentences without visualization.)

For example, a typical article would keep listing staticis of a certain issue throughout the article. (in this case, election) It only includes a visualization of a linear regression but fails to provide more of other following statistics.

Moreover, correlation can be visualized like the following with multiple scatterplot.

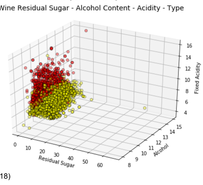


we want each square to be scattered not having a pattern.

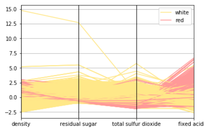
These two charts are from this source: https://www.youtube.com/watch?v=8DhvVs59lI4

Other existing visualization method for multi-variate data are:

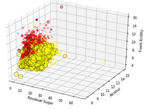
1. combine hue and depth to extend to 4 dimensional (Sarkar, 2018)



2. pararell coordinate (Sarkar, 2018)



3. try to combine as many as possible - hue, depth, size (Sarkar, 2018) but it does not seem to be effective



I would like to incorporate them in this project

Final result of this project would be a webpage that is interactive. Below image is part of project Opioids(Chan, 2019). Viewer can select variables (work place, state) and see the generated result. I want the viewer the input my idependent variable and show expectation (Y).

7 References:

Chan, C. (2019). Opioids: The Branding and Commercialization of Pain. Parsons.

Economist Data Team (2018). Mainstream Election-forecasting Could be Improved by a Popular Academic Approach. Economist.

Lee, J. (2016). Indie Film Audience Expectation (Korean). Sungkyunkwan Univ.

Prettenhofer, P. (2014). Multiple Regression Using Statsmodels. DataRobot.

Sarkar, D. (2018). The Art of Effective Visualization of Multi-dimensional Data. Medium.

In [9]:
```
import pandas as pd
df = pd.read_csv("movie_list.csv")
df
```

Out[9]:

	id	code	title_en	released_on	fscreen	preview	ani	naver	youtube	fweek	expectation	percent_error
0	13	20149629	Begin Again	8/13/14	185	9215	0	1276	94057	133628	131589.732100	1.525330
1	154	20150020	Son of Saul	2/25/16	47	3007	0	381	6111	11292	11492.875580	1.778919
2	107	20147684	Night Train to Lisbon	6/5/14	52	4584	0	431	3097	19679	19301.931170	1.916098
3	33	20166721	Foosball	9/7/16	212	1390	1	976	7075	67462	66008.048240	2.155216
4	247	20149860	Brave Rabbit	8/28/14	38	415	1	40	70	4829	4935.134048	2.197847
...	...	...	...	...	...	...	...	...	...	...	...	...
294	220	20159965	Yowamushi Pedal the Movie	1/14/16	52	679	1	1210	20334	6197	31829.620730	413.629510
295	291	20147723	Twice Born	10/30/14	30	328	0	144	1349	3380	-10622.074000	414.262544
296	294	20154484	Our Last Tango	12/31/15	19	775	0	153	239	3222	-11700.949060	463.157947
297	270	20144881	Journey to the West: Conquering the Demons	2/5/15	68	1575	0	1436	12113	4103	23200.083800	465.441964
298	270	20149626	Vincent	8/14/14	17	126	0	173	285	3764	-15098.152010	502.188410