Methods and Initial Results
Master of Arts Program in Computational Social Sciences
Wanlin Ji

1 Research Question

How to precisely predict the audience loss percentage of TV series using exogenous factors, including time, rating and commercial factors?
This study will be based on the case of TV series "Blindspot", aired from 9/21/15 - 5/23/16. To identify the influence of I am going to use Nielsen minute-by-minute ratings data for every "Blindspot" episode to compare the effect of Pooling Regression, fixed effect model and random effect model.

2 Data Sources

Our main data used in this research is Nielsen "Blindspot" minute-by-minute ratings data. Data can be accessed as a csv file from Nielsen Datasets at Kilts Center for marketing at University of Chicago Booth School of Business at no cost to University faculty members and students.
This dataset is useful for analyzing each episode in detail to see if there were significant points in a given episode that resulted in losing large chunks of the audience.

2.1 How the data is collected

Electronic and proprietary metering technology is used by Nielsen to carry out their audience measurement, which provided a detailed and minute to minute measurement of ratings and total loss percentage of audience for every episode of "Blindspot". Nielsens U.S. TV families are a random sample that represent a cross-section of representative homes throughout the country.
From these sampling families, Nielsen measures viewing using national and local people meters, which capture information about whats being viewed and when. This electronic data can record the all the variables by the minute.

2.2 Variables

There are nine variables in the "Blindspot" minute-by-minute ratings data. According to the Time variable, all the factors are measured by minute. Variables include:
- Unnamed Column: Arbitrary index generated by the data engineer when exporting data to csv
- Network: The name of the network the show (also known as a telecast) aired on
- Date: The airing date of a particular telecast
- Time: The local time indicating the particular minute for the show

- Program: The name of the telecast
- Length: The duration length of the telecast
- Rating: The rating of the telecast
- Minute In Commercial: Dummy variable for whether there is a commercial during that given minute or not (1 = minutes with commercials)
- Total Loss perc: Percentage of individuals who stopped watching the broadcast in that minute from the total number of viewers in that minute

The major variables for analysis here are Total Loss percentage, which depicts how many percent of audience quit, Minute In Commercial, a dummy variable to measure if there is commercial in that minute, and Rating. Through the inspection, I found the Length, Network, Program are almost the same for all the cases in dataset. We assume Rating stands for the quality of the show in that minute. For our analysis, we also assume that the dependent variable is Total Loss percentage, and the other two are independent variables.

2.3 Exploratory Analysis

After inspection, there is no missing value in the dataset, thus no need for imputation. The five number summary for key variables is here.

Table 1: Descriptive statistics for key variables

Variables	Mean	Min	1st Quartile	Median	3rd Quartile	Max
Index	1043.5	0	521.8	1043.5	1565.2	2087.0
Rating	1.706	0.955	1.372	1.552	1.964	3.831
Minute In Commercial	0.3472	0	0	0	1	1
Loss Percentage	3.417	0	1.776	2.548	3.824	29.634

* Calculated under R. All codes disclosed on the Github.

By comparing the three major variable, commercial, rating and total loss percentage, we found that the scope can really mislead us. The ratings are not so violent as we saw previously. Until now, we have got some solid recognition of our data at hand.
Now let's consider some of the hypothesis. We must pick up a dependent variable that could help us predict. From my perspective, the industry are more concerned with the number of audience, so we should think of the total loss percentage as our primary target here. Note I don't have the number of viewers who join the watching during the show, so it is impossible to analyze the real number of audience, but we can still aim to manage the viewers loss according to the insight from marketing that it is always more expensive to earn a new customer than to prevent losing an old customer. Same goes with viewers.
Next we want to consider the independent variables that could have an effect on the total loss percentage to find the primary drive behind it. We want to assume the rating stands for the quality of narratives, then we can see the total loss percentage an explanatory variable for analyzing total loss percentage. This seems plausible explanation. But we also may pay attention to the sudden fluctuations that comes with the total loss percentage, which is largely corresponding to Minute In Commercial.

Figure 1: Variable tendency in the first episode

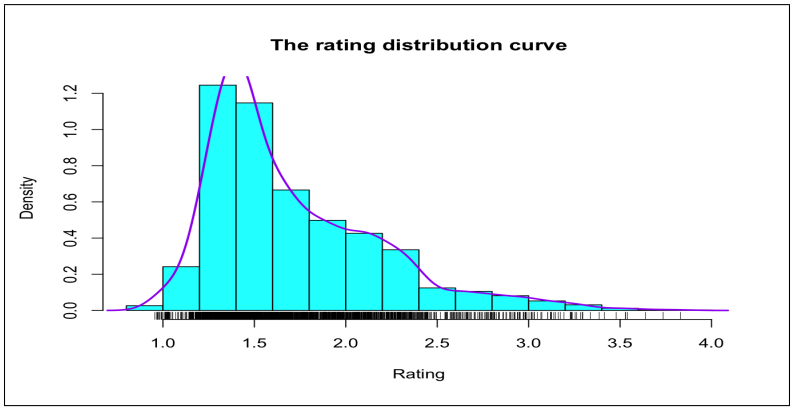
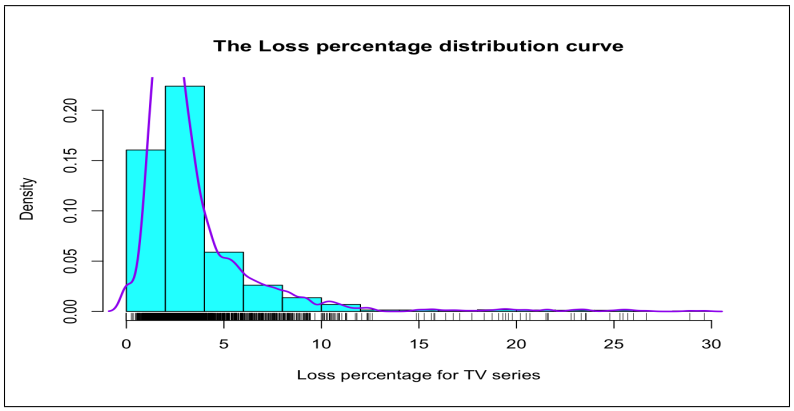


Figure 2: Variable tendency in the first episode



For the rating, it is decreasing with time. For the total loss percentage, it is repeating itself on a episode basis. To carry out data analysis, we need to make sure the sequences are stationary.

3 Generalized Linear Regression Model

3.1 Unit Root Test

We adapted ADF test and Phillips-Perron test on three main variables, and found that all the p values are under 0.01 significant level, there p-values are all smaller than the printed p-value. All the three sequences showed no sign of unit root, and no need for Johansen Co-integration test. It also avoids spurious regression. Let's move on to the next stage.

3.2 Linear Regression Model

Considering there is no co-integration existing in the relationship, the linear regression model is a good fit for the data. Also considering the Minute In Commercial has a

Figure 3: Dependent Variable: Audience loss percent tendency

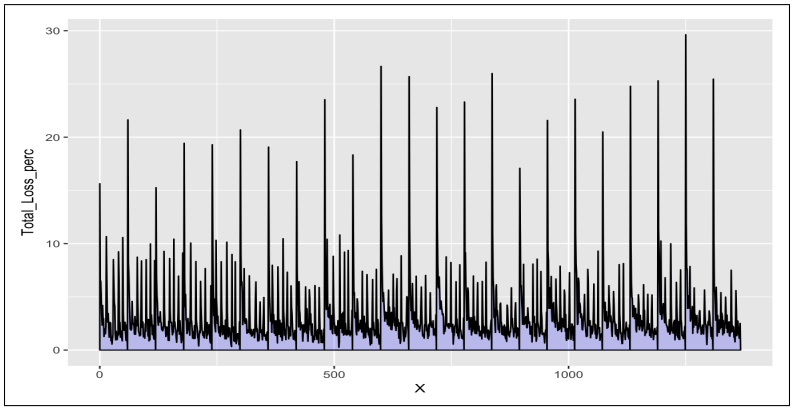
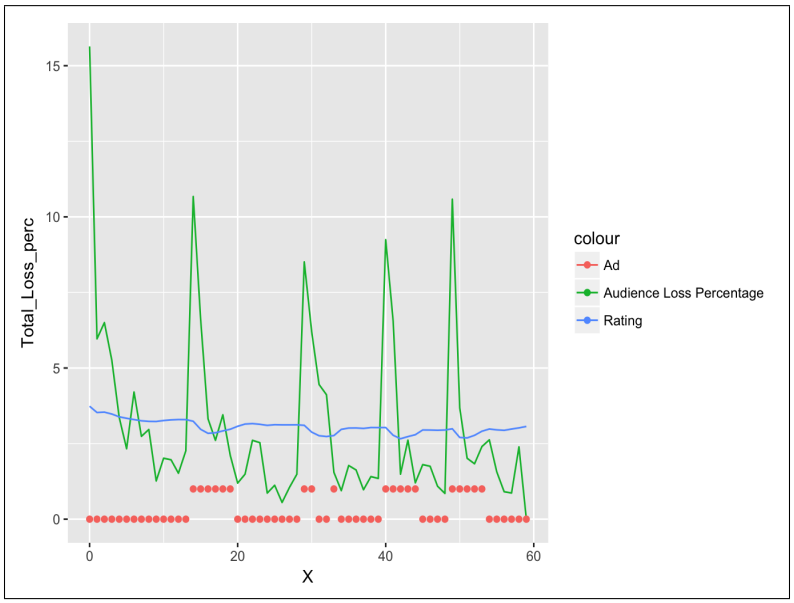


Figure 4: Variable tendency in the first episode



sudden influence on the Y, we need to transform Y in order to smooth the curve and make it easier for analysis. The model we used is in the following:

$$lnY = \alpha + \beta_1 X_1(Rating) + \beta_2 X_2(MinuteInCommercial) + \beta_3 X_3(X) * X_2(MinuteInCommercial)$$

3.3 Initial Result

And after calculation in R, we have the following result:
This result largely meet my expectations from the graph. From the analysis above, we found that except the relatively inactive variable Rating shows a significant level, other variables in our model reports a fairly strong significant level, indicating our model has strong explanation power.

Table 2: ADF Test and PP Test

Variables	Dickey-Fuller	Dickey-Fuller Z(alpha)
Rating	-6.9171^{**}	-238.09^{**}
Minute In Commercial	-13.53^{**}	-538.02^{**}
Loss Percentage	-12.75^{**}	-1574.7^{**}

[*] Calculated under R.

Table 3: Linear regression with intersection term

Coefficient	Value	Std. Error
Intercept	3.967^{**}	2.786
β_1	-0.404^{*}	0.134
β_2	9.530^{***}	5.681
β_3	3.705^{**}	1.342

[*] *, **, *** stands for significance level 0.05, 0.01, 0.005