# Analysis of Attendance for GatorNights

**Nathan Morse**
Capstone Project for STA4211
University of Florida
December 3, 2017

## Abstract

The University of Florida's GatorNights program puts on events for students every Friday night. This study analyzed the effect various factors have on attendance by creating a multiple regression model. Ten variables were selected and went into a fairly accurate prediction model. Two of the three most significant factors that affected attendance related to which movie was shown: movies of the comedy genre and movies with high ratings on IMDB.com tend to increase attendance significantly. The budget for printed marketing materials was also important, but comparing this to Facebook marketing was difficult because of the limited data.

## 1. Introduction

Since 2000, the University of Florida has sponsored a weekly event known as GatorNights at the J. Wayne Reitz Union. Every Friday night, any UF student can come for free food, movies, activities, bowling, arts, and often a live musician or comedian. The UF Provost's Office funds the program with $250,000 each year, meaning the average weekly cost is roughly $10,000. Events are planned by members of the Reitz Programming Board, an entirely student-run organization, and executed by the Reitz Union staff.

While student leaders and student affairs professionals work hard to make sure the program is maximizing both its impact to students and its cost-effectiveness for the Provost's funding, there are many unknowns in the decision-making process. For example, is it more effective to advertise with flyers and signs or Facebook ads? Facebook advertising has become increasingly relevant since it can target specific types of people and cuts down on labor and printing costs. However, it may leave out some international students and students who do not care to use social media—two key demographics that may be more likely than others to come to a school-run non-alcoholic event on a Friday night. Again, the answer to this kind of question has been unknown to the decision-makers.

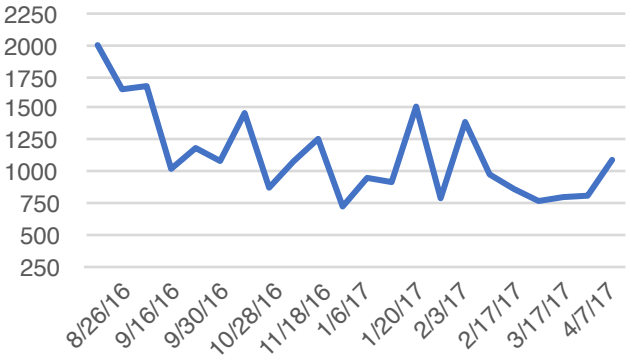
In this study, we analyzed data from the 2016-2017 academic year regarding each of the 22 GatorNights events. 16 potential predictor variables were collected, including budget breakdowns, information on the movie and entertainer, sports games that conflict with the event, and Facebook data. Below are the 16 predictors:

| Name | Variable [range] |
|---|---|
| `sem` | Semester {0=fall, 1=spring} |
| `semwk` | Week in semester [1,14] |
| `budg_print` | Printing budget [$0, $356.50] |
| `budg_ent` | Entertainer budget [$0, $10,200] |
| `mov_imdb` | Movie IMDB rating [5.3, 8.2] |
| `mov_wknd` | Movie opening weekend sales [$4,223,828, $179,750,616] |
| `mov_rel` | Release date [12/25/12, 2/15/17] |
| `mov_rating` | Movie MPAA rating {0=G, 1=PG, 2=PG-13, 3=R} |
| `mov_actadv` | Genre: action/adventure {0,1} |
| `mov_anim` | Genre: animated {0,1} |
| `mov_comedy` | Genre: comedy {0,1} |
| `mov_drama` | Genre: drama {0,1} |
| `mov_scifi` | Genre: science fiction {0,1} |
| `ent_twitter` | Entertainer Twitter followers [0, 302000] |
| `vip` | VIP groups present {0,1} |
| `sports` | Conflicting sports games {0,1} |
| `fb_reach` | Facebook total reach [204, 21210] |
| `fb_paidrch` | Facebook paid reach [0, 10099] |
| `fb_imp` | Facebook post impressions [339, 61440] |

The full dataset is available online at this address: http://www.nmorse.info/gndata/data.csv. The sixteen predictors were compared to attendance counts from each week. Every attendee is required to swipe their

GatorOne ID, so the attendance numbers are exact.[1] Figure 1, below, displays the attendance pattern:

**Figure 1: Attendance**




Attendance generally has a downward trend throughout the year, with a few spikes and a few dips. The lowest attendance was 728, highest attendance was 1,997, and the median was 1,054. The average attendance count was 1,133.

## 2. Objectives

The ideal outcome of this analysis is a multiple regression model that can accurately predict attendance. That would also show the effect each factor has on attendance and by how much. This can help the program planners make data-driven decisions on what movies to show, when to schedule major entertainment, and how best to advertise. Attendance is not the only measure of success for a program like GatorNights, but the insights this analysis provides can still be very useful for the planners and the students who use this program.

## 3. Methodology

The greatest challenge for this dataset is that there are so many predictors yet so few observations. Generally, there should only be one predictor for every ten observations. However, because GatorNights occurs only once a week, there were only 22 observations with 16 predictors. Following the rule of thumb, there would only be 2 predictors, which would not provide a reliable prediction model. Thus, we proceed with caution and let the variable selection process narrow down the predictors.

***Model Selection.*** To start off, all 16 predictors were placed into a linear model in R. No interactions were found to be significant. Next, automatic stepwise selection procedures were run with the `stepAIC` function, and the backwards elimination procedure returned the model shown below.

| | Coefficients |
|---|---|
| (Intercept) | −1.259e+03 |
| semwk | −2.034e+01 |
| budg_print | 2.507e+00 |
| budg_ent | −4.623e−02 |
| mov_imdb | 2.406e+02 |
| mov_actadv | 3.715e+02 |
| mov_comedy | 3.735e+02 |
| mov_drama | 3.528e+02 |
| ent_twitter | 2.719e−03 |
| vip | 3.319e+02 |
| fb_paidrch | −3.642e−02 |

This model has 10 variables, which is still more than ideal given the number of observations. However, the model has a strong fit, with an adjusted R-squared value of 0.899.

***Diagnostics and remedial measures.*** With the variables selected, diagnostics were run to test that the data meets the assumptions of the model. The model assumes a normal distribution of errors and predicted values, equal variance of error terms, and low multicollinearity. It also must contain no outlying points.

***Influential observations.*** First, Cook's distance was computed for the data to identify outliers. A plot is shown below in Figure 2. Two points had values above 1, which indicates these are influential points. Because this dataset has so few observations, removing only the top one caused the adjusted R-squared value to be higher than removing both, so only the observation with the highest Cook's distance value was removed.

[1] Each attendee may bring one guest, but guests were not counted in these attendance swipes. The number of guests is dependent on the number of actual attendees and is less important to decision makers, so it is fair to exclude them from the model.

**Figure 2: Cook's distance**

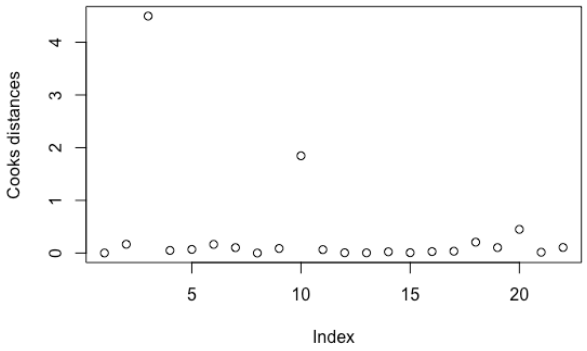



**Figure 3: Normal Q-Q Plot**

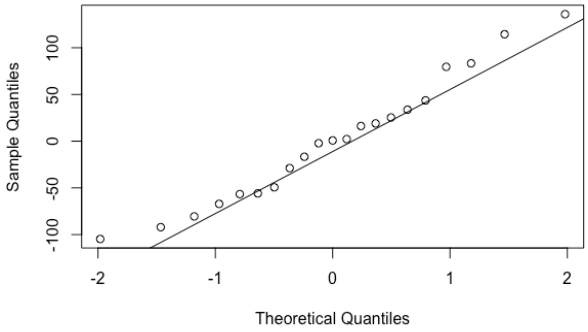



**Figure 4: Residual histogram**

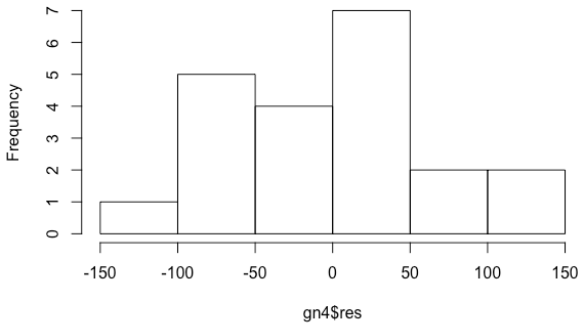



**Figure 5: Bruesch-Pagan test**

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.006079863    Df = 1
p = 0.9378492
```

**Figure 6: Variance Inflation Factors**

| | |
|---|---|
| semwk | 1.647252 |
| budg_print | 1.395176 |
| budg_ent | 6.214564 |
| mov_imdb | 2.571878 |
| mov_actadv | 5.965170 |
| mov_comedy | 1.292679 |
| mov_drama | 6.631336 |
| ent_twitter | 5.379881 |
| vip | 3.377306 |
| fb_paidrch | 2.927300 |

***Normality.*** A Q-Q plot was generated and shown in Figure 3. The points generally follow a straight line. Given this, it is safe to assume the predicted values are normally distributed and the assumption for the regression model is met. This can be tested further by reviewing a histogram of the residuals for normality, shown below in Figure 4. This histogram is roughly normally distributed, so it passes.

***Equal Variance.*** Another assumption to test is equal error variances. For this, a Breusch-Pagan test was run, with output shown in Figure 5. The null hypothesis for this test is that the error terms have constant variance, so a high p-value indicates that the error terms do have equal variances.

***Multicollinearity.*** To ensure variables are not correlated among themselves, variable inflation factors were calculated, shown in Figure 6. Factors significantly higher than 1 indicate a high level of multicollinearity, and there are several with relatively high values. This does not affect the validity of the model for predictions, but it does mean inferences and interpretations of specific variables may not be appropriate.

## 4. Results

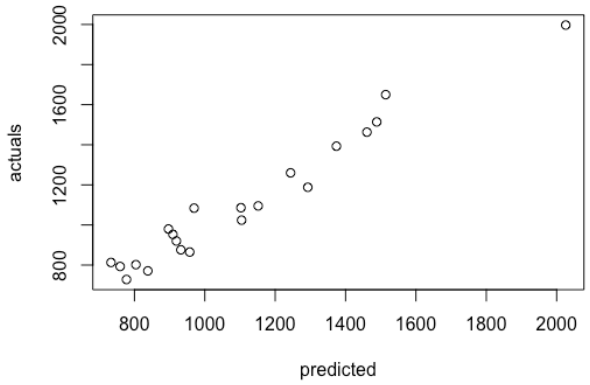
After fitting the model, selecting variables, checking that the model met the assumptions, and performing appropriate remedial measures, a final model was created with an adjusted R-squared value of .9194. The variables and their corresponding coefficients and p-values are displayed below, in order of significance.

| Variable | Coefficient | p-value | |
|---|---|---|---|
| mov_comedy | 3.935e+02 | 1.04e-05 | *** |
| budg_print | 1.974e+00 | 0.000226 | *** |
| mov_imdb | 2.480e+02 | 0.000271 | *** |
| vip | 4.154e+02 | 0.000551 | *** |
| ent_twitter | 5.452e-03 | 0.002680 | ** |
| semwk | -2.367e+01 | 0.006623 | ** |
| fb_paidrch | -7.137e-02 | 0.007811 | ** |
| budg_ent | -6.858e-02 | 0.009712 | ** |
| mov_actadv | 2.959e+02 | 0.017878 | * |
| mov_drama | 2.917e+02 | 0.029375 | * |

In order of importance, the significant variables are: whether or not a comedy movie is shown, the budget for printed materials such as fliers and signs, the IMDB rating of the movie being shown, whether or not a VIP group was present, the number of Twitter followers the entertainer has, the number of Facebook users reached by paid posts until the day before the event, the budget for the entertainer's fee, whether the movie is action/adventure, and whether the movie is a drama.

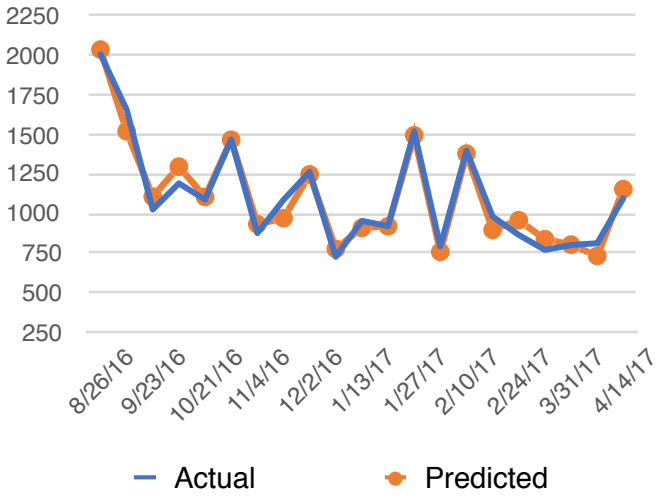
A plot of the predicted vs. actual attendance values is below in Figure 7.

**Figure 7: Predicted vs. Actual values**




The predicted and actual values follow a roughly close line, which is graphical confirmation that the model is fairly accurate. Figure 8, below, shows another graphical display of the predictions and the actual attendance.

**Figure 8: Predicted vs. Actual line graph**




## 5. Discussion

The most significant factor is whether a comedy movie is shown. 394 more people are expected to show up when there is a comedy movie. The movie and the entertainer are the two most visibly advertised components of each GatorNights event, so it makes sense that this has the largest effect on students' decisions to attend or not. Of the movie genre variables, comedy was by far the most important, which indicates that students prefer comedy movies over drama, action, or other genres.

The next most significant variable, the printing budget, makes sense as well—more advertisements bring in more students. Specifically, every dollar budgeted for printing materials is expected to increase attendance by two people. This could be very useful for planners if they need an easy way to boost attendance after the movie and other components have already been finalized. Spend $50 on printing, and they can expect roughly 100 more people.

Comparing this to Facebook advertising is not easy. The paid post reach, `fb_paidrch`, is a measure of how many Facebook users have seen sponsored posts by GatorNights. GatorNights regularly posts graphics and information about each event, but paid posts allow users who match a target audience to see GatorNights posts even if they do not follow the GatorNights page or do not have any friends who have shared posts. The overall reach for all posts, paid or unpaid, was originally included as `fb_reach` but was taken out of the model during the variable selection process.

The paid reach, then, was the only significant variable relating to Facebook, and it actually has a *negative* coefficient. For every 100 people who see a sponsored post, attendance is expected to be *lower* by 7 people. There are a few things to understand when considering the counterintuitive effect of this variable. First, only four GatorNights events in the dataset used paid advertising, so 17 events had values of 0 for this variable. The low number of events with paid posts decreases the reliability. Second, this variable has higher multicollinearity than other variables, so there may be confounding factors that are

not obvious. For example, perhaps paid posts were only used when attendance was expected to be low anyway. Without those paid posts, it's possible attendance would have been even lower. Given this, it would not be safe to make any definite conclusions about the effect of paid posts on attendance with this limited data.

Likewise, the budget for paying the entertainer has a negative effect on attendance, which is also counterintuitive. One would think that a higher entertainer fee would relate to higher popularity and higher attendance. Again, this variable has a high variable inflation factor, so there may be confounding variables that are unknown to the data in this study. When looking at another measure of entertainer popularity, the number of twitter followers the entertainer has, it indeed has a significant positive effect on attendance.

GatorNights planners are also interested in whether VIP events affect attendance. When a GatorNights event has a VIP group, it means a student organization or residence hall has a social event for its members at GatorNights, giving them access to a special room with refreshments, reserved lanes in the bowling alley, and other perks. The presence of VIP groups is indeed associated with higher attendance, but again its multicollinearity factor limits conclusions that can be drawn from this.

## 6. Conclusion

In general, the most important factors for GatorNights attendance are the movie, entertainer selection, and advertising methods. It's also important to note that the week in the semester, `semwk`, has a significant negative effect on attendance, meaning each week the attendance is expected to decrease by 24 students from the previous week, controlling for all other factors. This confirms what was said earlier about the attendance line graph generally decreasing over time.

This information should be very useful for GatorNights planners. Attendance decreases over time, but there are many ways to increase it: show comedy movies, show movies with high IMDB ratings, print more fliers, partner with student organizations to have VIP events, book entertainers whose Twitter followings indicate they're popular, and so on. This model will help the GatorNights team make data-driven decisions when planning events so that they can maximize attendance and minimize unnecessary costs.