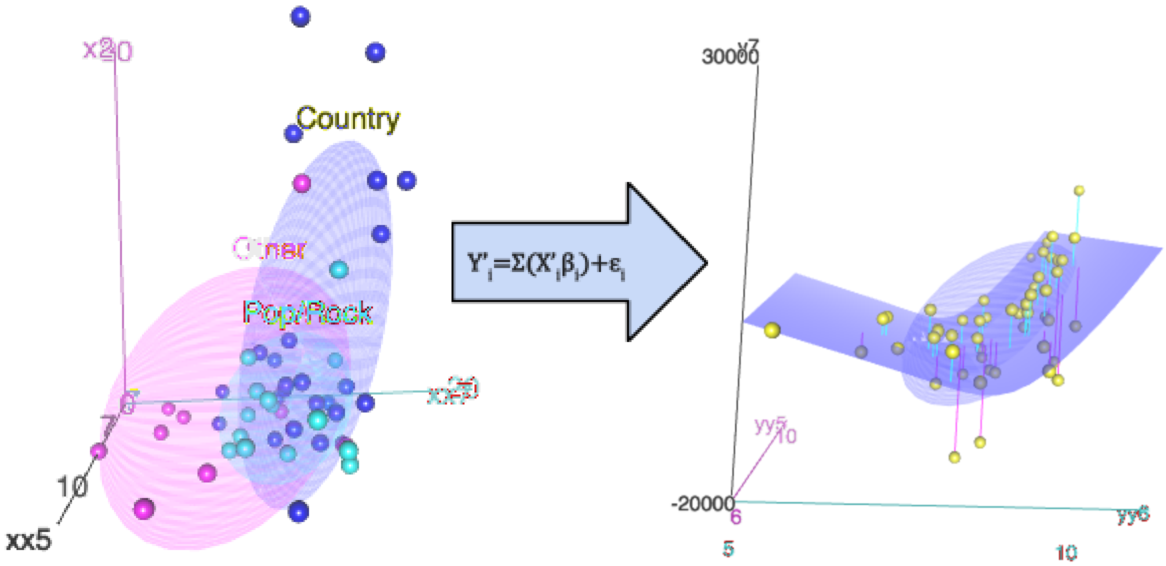
Pollstar Bands:
A Regression Analysis

Projecting Profits, Income, and Bar Sales
for future concerts
from Facebook and Pollstar band data.



Jeffrey Hall

07.06.2021
Emporia Granada Theatre

INTRODUCTION
- Hypothesis
- Predictors
- Regressors
- Procedure
- Sample Data

PROCEDURE
- Transformations
- Variable Selection

Models
- Profits: Model 1
- Profits: Model 2
- Profits: Model 3
- Bar Sales: Model
- Income: Model
  - Confidence:
  - Standardized Coefficients:

OUTLIERS
  - What Went Wrong?
  - What Went Right?

CONCLUSION

CONSIDERATIONS FOR FURTHER INVESTIGATION
  - Square
  - Demographics
  - Time as a Predictor
  - Larger sample
  - Turnpike Troubadours

RETROSPECTIVE COMMENTARY
  - Transformations
  - Income and Co-Bill Shows
  - Larger Sample for Bar Sales
  - Statistical Measures

INTRODUCTION

The Emporia Granada Theatre relies heavily on profits from concerts in order to thrive. In order to bolster and expedite this decision making process, I have been developing regression models that can help to mitigate risk while simultaneously optimizing profit margins. Previously, profit margins and risk were strongly correlated, but with the use of these models we should finally be able to get the best of both worlds.

Hypothesis

Our hypothesis is that by transforming (1) data publicly available on Facebook and (2) the data made available to us by our Pollstar subscription, we might be able to find significant enough correlations to Income, Bar Sales, and Profit to make reasonable projections about future events based on the headlining act.

Predictors

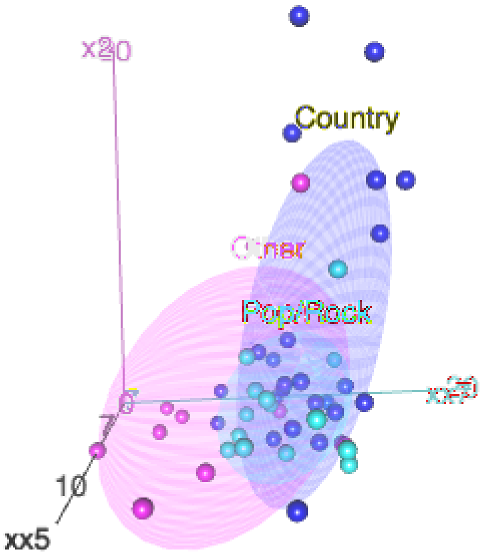
The following predictors were considered.

Facebook:

1. Followers (x_7)

Pollstar:

2. Genre (x_0)
3. Headline Shows (x_1)
4. Co-Bill Shows (x_2)
5. Boxoffice Reports (x_3)
6. Avg. Tickets Sold (x_4)
7. Avg. Gross (x_5)



Regressors

We will regress subsets of those 7 predictors onto the following regressor variables:

1. Bar Sales (y_5)
2. Income (y_6)
3. Profit (y_7)

Procedure

1. Identify appropriate transformations on data to ensure that the normality assumption is met for each variable.
2. Identify harmful correlations between Predictor variables
3. Perform simple linear regression
4. Consider all possible interactions between variables, and use stepwise methods to find the best subset of variables and interactions.
5. State and visualize the resultant models, and analyze their strengths and weaknesses.

Sample Data

I compiled data from 53 events spanning from 2017 to 2020. Since the raw data contains financial information, I have omitted it to respect the privacy of the Emporia Granada Theatre.

PROCEDURE

Transformations

Logarithmic transformations initially seemed ideal for the variables with strictly positive values, since they span across multiple orders of magnitude. Since the domain for Co-Bill Shows includes zero, the slightly modified transformation of $x' = ln(x + 1)$is used instead. More discussion on this can be found in the commentary at the end of this report.

Variable Selection

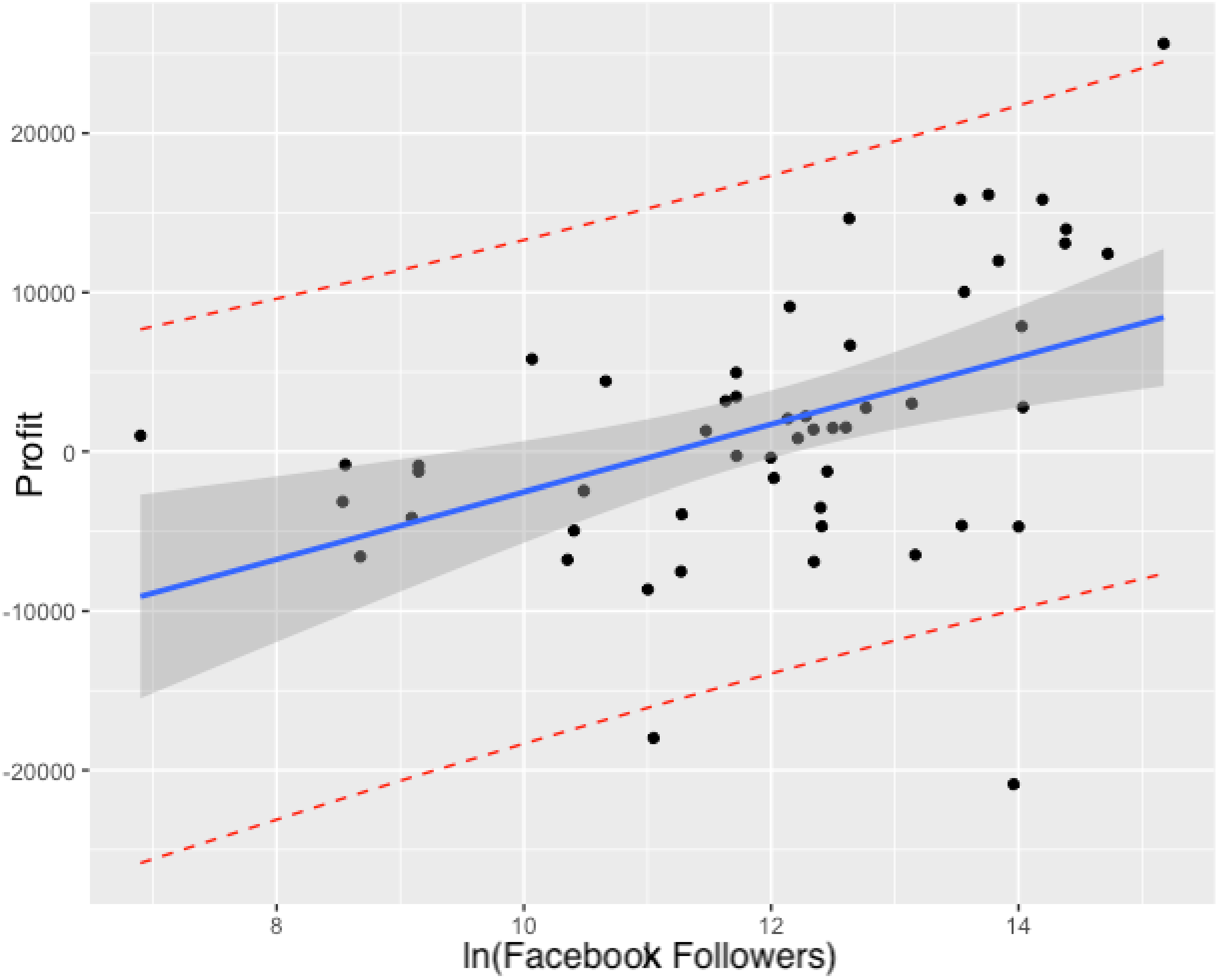
Headline Shows (x_1), Boxoffice Reports (x_3), Avg. Tickets Sold (x_4), and Avg. Gross (x_5) are all *highly* correlated. Although Principal Component Analysis was considered, *elimination* proved to be the best solution to the multicollinearity problem this would pose in the models. Of these four variables, Avg. Gross proved to be the most useful for predicting the regressors under consideration. Surprisingly, Avg. Gross turned out to be mostly independent of Facebook Followers.

Models

Profits: Model 1

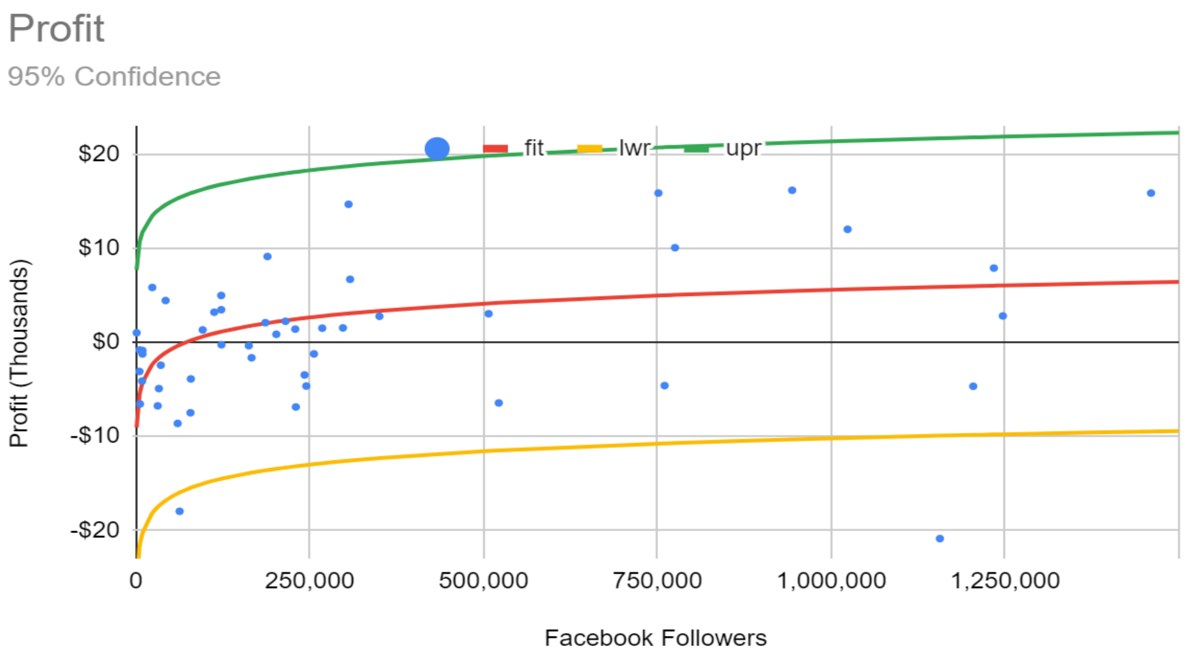
The simplest model projects Profits directly from the natural logarithm of Facebook Followers.

$$\hat{y}_7 = \ - 23677.4 + ln(x_7) * 2115.6$$



This model is found to be highly significant, with a p-value of 0.000005889, and an F-Statistic of 12.88, but it's such a simple model that it only has a predictive r-squared value of 12.5%, meaning that it can only explain about 12% of what's going to happen. This means a relatively poor standard error ($7,714), therefore projections are imprecise.

Transformed back into the original units, *Model 1* looks like this:

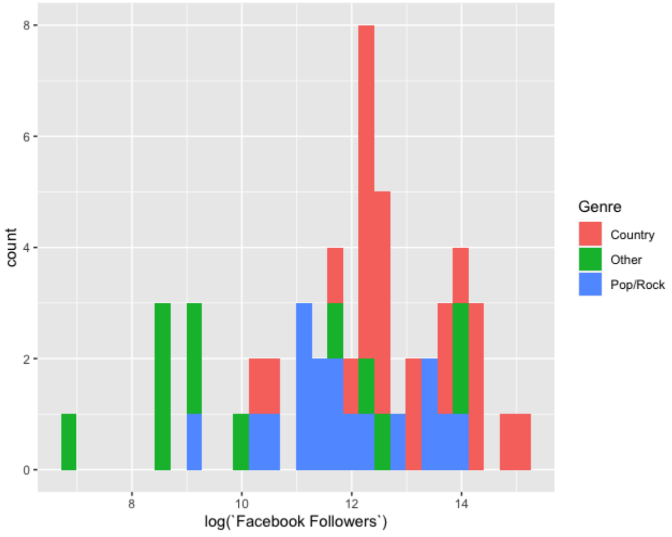


Main Insight

This model isn't great for future predictions, but it does give us one major insight. **The average break-even point lies between 13 and 66 thousand facebook followers.** This means that bands with fewer followers than that will lose us money more often than not. No number of facebook followers in this range will guarantee profits.

Profits: Model 2

Our next model for Profit takes Genre into account, in addition to Facebook Followers. Now, it's important to note that Genre is not evenly distributed across Facebook Followers in our sample data. The histogram shows that the country bands in our data set often have more facebook followers than the Pop/Rock bands, and that other genres typically have less followers than either Country or Pop/Rock.

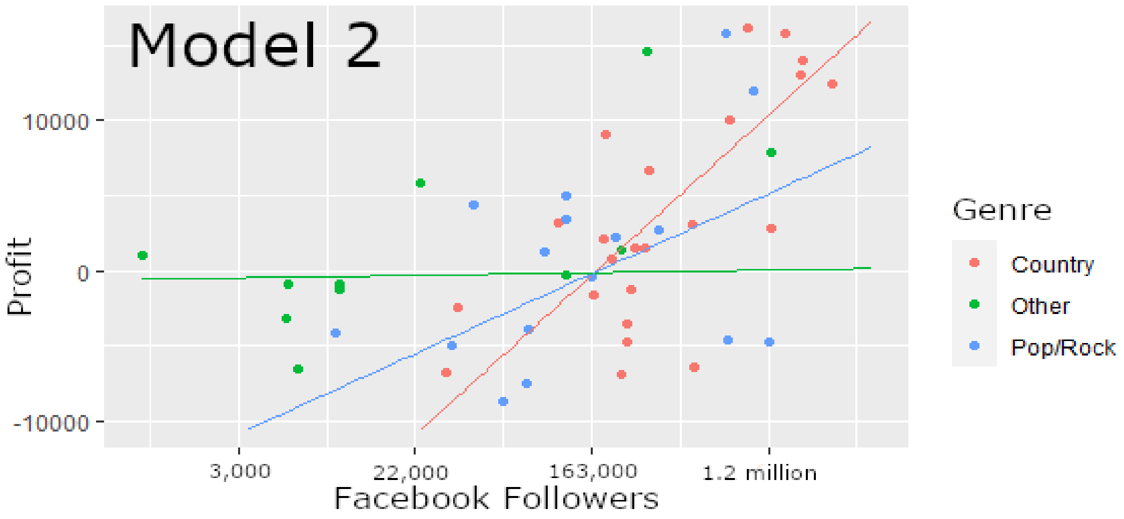


These trends in the data strongly bias the model, but this may be acceptable when we're projecting for bands that follow the same trends.

$$\text{Country: } \hat{y}_7 = \quad - 64054.6 + ln(x_7) * 5314.6$$

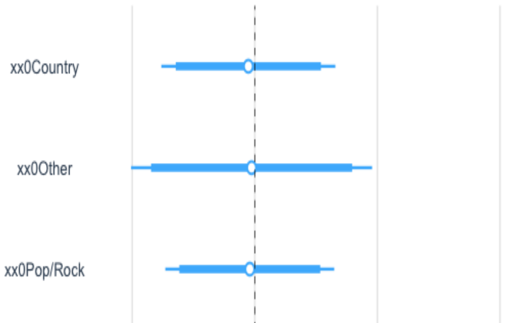
$$\text{Pop/Rock: } \hat{y}_7 = \quad - 32100.3 + ln(x_7) * 2657.3$$

$$\text{Other: } \hat{y}_7 = \quad - 1173.3 + ln(x_7) * 86.9$$



Standardized Coefficients:

Adding Genre to our model reduces our standard error significantly, down to $7,019. This is nearly as good as our standard error will get for Profit models, but just looking at the



graph shows that it's predictions will be terribly inaccurate at predicting things like Country bands with few followers, just for one example. This model only has an F-Statistic of 6.03, so it's not very significant. Furthermore, four of our six coefficients are insignificant.

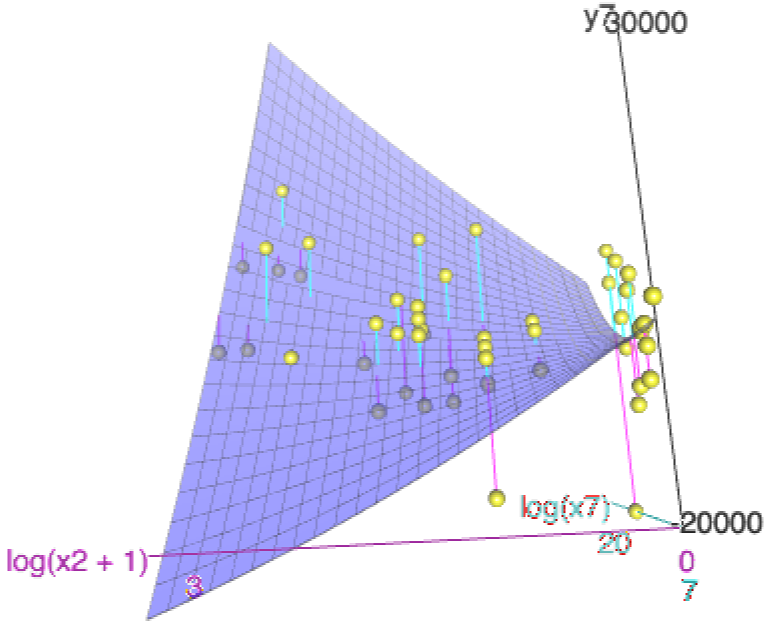
Main Insight

This model shows that **a large number of Facebook Followers is most important for Country Bands**, but may not significantly affect "Other" Genres besides Country and Pop/Rock.

Profits: Model 3

When we swap out *Genre* for *Co-Bill Shows* (transformed), we can get a better model than either of the first two that we have for predicting profits, particularly when we also consider the interaction between *Co-Bill Shows* and *Facebook Followers*.

$$\hat{y}_7 = -4210.2 + ln(x_7) * 289.6 - ln(x_2 + 1) * 24839.3 + ln(x_2 + 1) * ln(x_7) * 2093.3$$



With an F-Statistic of 10 and a p-value of 0.00003. This is the best model for profit that we have. The

Predictive R^2 value of 27% tells us that this model is somewhat useful for making future predictions.

Main Insight

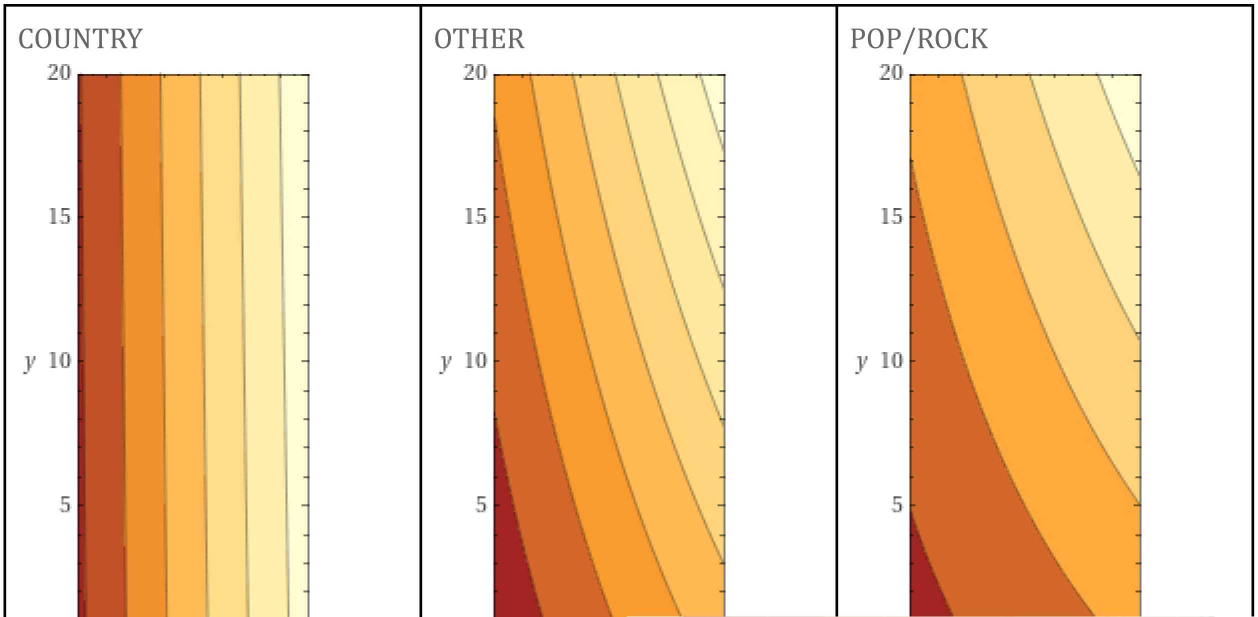
This model is very interesting. While *Model 1* showed us that more Followers increases average Profit, *Model 3* shows us that **Followers only helps when Co-Bill Shows is greater than 0**. It also shows us that the best profits come from bands that have a lot of followers AND co-bill shows.

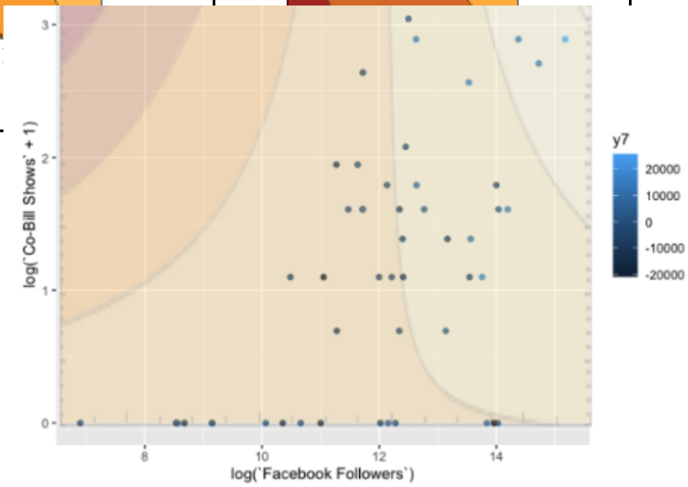
Bar Sales: Model

$$\text{Country: } \hat{y}_5 = 2.933 + ln(x_7) * 0.3537 + x_2 * ln(x_7) * 0.0004189$$

$$\text{Pop/Rock: } \hat{y}_5 = 5.204 + ln(x_7) * 0.1994 + x_2 * ln(x_7) * 0.006973$$

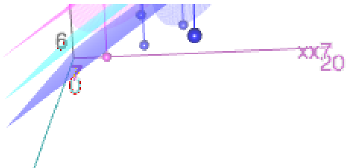
$$\text{Other: } \hat{y}_5 = 6.138 + ln(x_7) * 0.09488 + x_2 * ln(x_7) * 0.005848$$





In the plots above, x=ln(`Facebook Followers`), y=`Co-Bill Shows`, and the heat map indicates the fitted values for Profit. This is our best model yet, ~~with a Predictive R^2 value of 39.6%~~* and a Standard Error of 62.5%.

Main Insight



The biggest takeaway here is that **Bar Sales are independent of Co-Bill Shows for Country**, but not for Pop/Rock or any other genres. Co-Bill Shows are most beneficial to bar sales for Pop/Rock.

*See commentary on page 13.

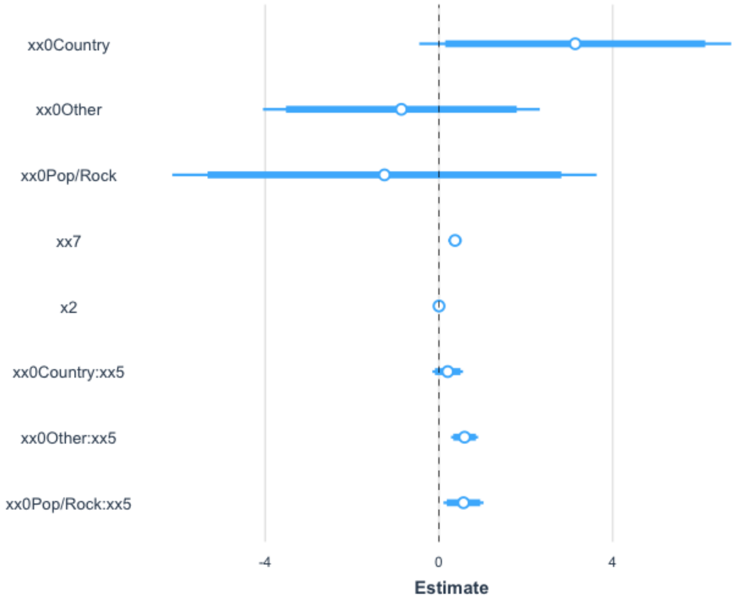
Income: Model

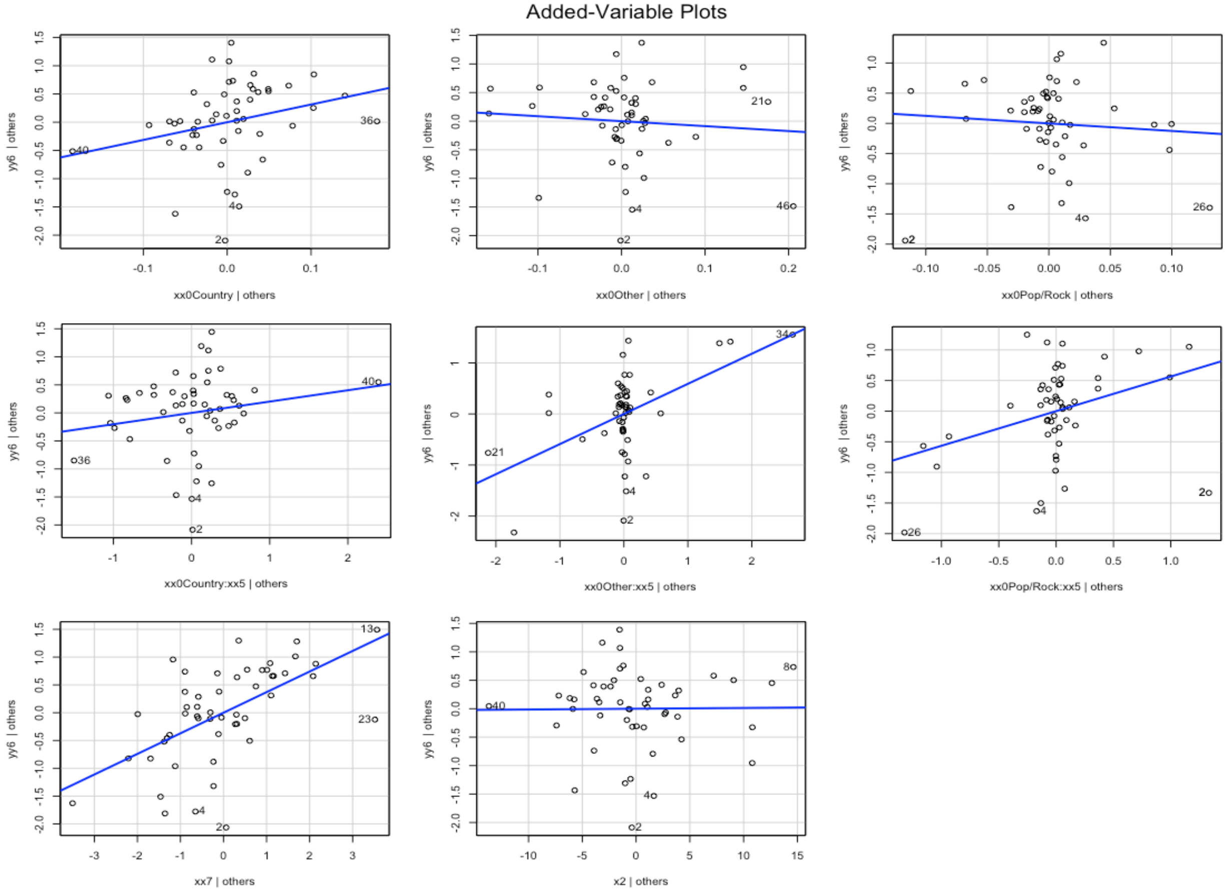
Finally, our last model predicts Income from Genre, Co-Bill Shows, Avg. Gross, and Facebook Followers. With an F-Statistic of 14, **a Predictive R² value of 55%**, and a P-Value of 2.6E-09, this is our most robust and trust-worthy model of the bunch, finally predicting over half of the variation in income for future concerts. With 4 predictor variables onto our 1 regressor, the entire model cannot be visualized at once. The best we can do to visualize it is to provide the added-variable plots, and the confidence intervals for the coefficients.

$$\text{Country: } \hat{y}_6 = 23.1e^{x_2/704} x_5^{0.202} x_7^{0.37}$$

$$\text{Pop/Rock: } \hat{y}_6 = 0.286e^{x_2/704} x_5^{0.566} x_7^{0.37}$$

$$\text{Other: } \hat{y}_6 = 0.419e^{x_2/704} x_5^{0.59} x_7^{0.37}$$



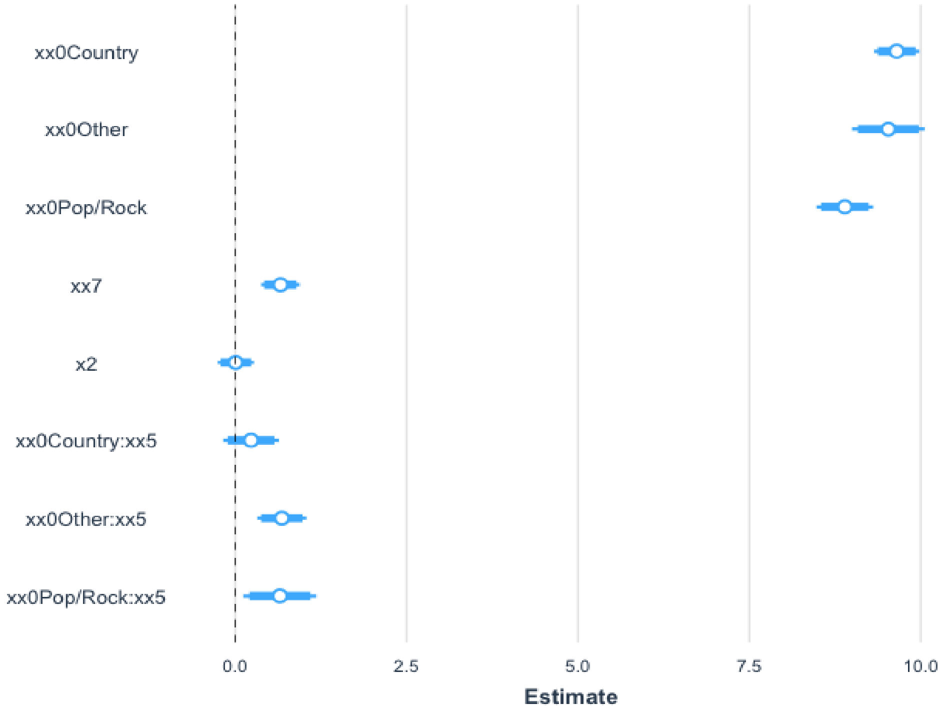


Confidence:

Coefficients:	2.5 %	97.5 %
xx0Country	-0.45	6.73

xx0Other	-8.50	0.49
xx0Pop/Rock	-10.14	1.35
xx7	0.21	0.53
x2	-0.04	0.04
xx0Country:xx5	-0.15	0.55
xx0Other:xx5	0.28	0.90
xx0Pop/Rock:xx5	0.10	1.03

Standardized Coefficients:



Main Insights

1. **Facebook Followers** always help a lot with income.
2. **Avg. Gross** is also always helpful; it's especially helpful for "Other" genres.
3. **Country** acts tend to fare the best in terms of income.
4. **Other** genres could *possibly* compete with Country for income in ways that Pop/Rock never could, but we'd need more data to find out.

OUTLIERS

There were a couple notable outliers in this dataset worth investigating.

What Went Wrong?

The Mac Sabbath show was a complete flop. It was originally going to be a ticketed event, but in the end a few local businesses decided to sponsor it and make it a free show. I don't know how much the sponsors paid us, but I don't think it made up for the fact that we hardly sold any alcohol that night. The band is terrible and gimmicky. Main takeaway: Maybe don't hire gimmick bands (Like Gwar), and probably don't do free shows for hired acts... and if you ever do, DO NOT count on alcohol sales to pay for it. Nearly 500 people came, and the show was still a flop.

What Went Right?

The Turnpike Troubadours show brought in record breaking bar sales a few times over, nearly FOUR standard deviations above the expected value for bar sales. This basically means a 1:6000 chance against that we know what went right at this show based on PollStar and Facebook data alone. Some other factor(s) that we haven't considered yet made this show a huge success. I'd really like to talk to someone who was there and find out more about it.

CONCLUSION

We looked at three models for Profit, which at best can predict a third of the variation for future profits. We looked at one model for bar sales, and finally we looked at one rather decent model for gross income. This last model is clearly the most useful and the most reliable for making predictions about future events. However, the other models are still insightful, as they can tell us where to look for prospective bands.

- \`Facebook Followers\` is the single best predictor for all three regressors (Income, Bar Sales, and Profit)
- \`Facebook Followers\` helps a lot with income
- \`Facebook Followers\` only helps *profits* when \`Co-Bill Shows\` is large (more than a few).
- \`Co-Bill Shows\` helps with *bar sales*, especially at Pop/Rock shows (Exception: Country)
- \`Avg. Gross\` helps with income, especially for "Other" genres.

CONSIDERATIONS FOR FURTHER INVESTIGATION

Square

The “Bar Sales” data was compiled from Square as daily totals. This could result in an error if (1) there were multiple events on the same day, or if (2) a significant value of non-bar commodities was sold. This can be corrected for by (1) identifying dates in the given data set when more than one event occurred, and adjusting for irrelevant sales according to time-stamp, and (2) grouping Square’s data by category and only considering relevant categories.

Demographics

In order to explain more of the variance in income/profits/bar sales, we would need a greater variety of predictor variables. By the Law of Diminishing Marginal Returns, we can expect that there’s little to gain from getting more or better measures of a band’s popularity. But since no demographic variables were considered, this may offer the best opportunity for improvement in the model. This data hasn’t been collected anywhere, but we could start collecting it with an ID scanner at the bar. We can require this, since it’s a way to validate IDs, and then passively collect Age and Gender information on our clientele.

Time as a Predictor

We could also use the date and time of an event to diversify the predictor variables further. We could use the week of the year, day of the week, and hour of the day as cyclical variables. This may explain more of the variance. We have this data already. I only left it out of these models so we could use the models without choosing a specific time and date. The trade-off here is ease of use versus precision. Including date/time would not help significantly with choosing which bands to hire, but would likely explain more of the variance and therefore could offer better predictions for future events after they’ve been scheduled.

Larger sample

While a sample of 50+ events is quite significant, more data is always better. There are a few more events listed on the Emporia Granada’s PollStar page than were considered here. By looking into those we could surely add a few more events to our sample. I’d also like to talk to Dawn from the Emporia Arts Council (EAC) about data from EAC shows, so we can include more of their events in our sample dataset.

Turnpike Troubadours

We should try to learn the story behind this event and see how many factors we can identify that contributed to this show’s success.

RETROSPECTIVE COMMENTARY

11.10.2021

Shortly after writing this report, I found a few flaws in these models and made some significant improvements. More recently, I found some errors in the statistical measures and made corrections as I was able.

Transformations

The models mentioned in this report rely on logarithmic transformations. This comes close to achieving normally distributed standard errors, but I found that it didn't always quite hit the mark in a statistically significant way. By using transformations of the form y^c (for constant c) and optimizing c to minimize skewness, I was able to get a distribution of standard errors that passed the Shapiro Test for normality.

Income and Co-Bill Shows

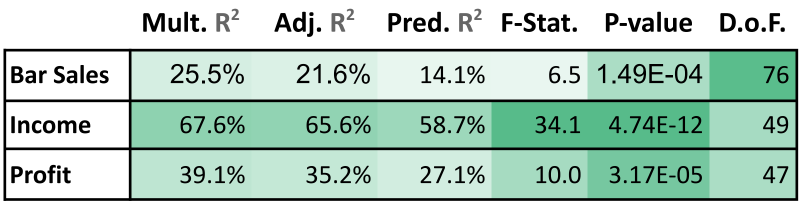
Looking at the graphs for this Income model, it's clear that Co-Bill shows is independent of income once the other variables are controlled for. Removing this variable from the model, and using better transformations as described above resulted in a much better model. The predictive R^2 value increased from 54.7% to 58.7%, and the F-Statistic more than doubled from 14.0 to 34.1.

Larger Sample for Bar Sales

While I never got more data for Profits or Income, I was able to compile a larger sample of bar sales data, using 81 events rather than 53. I also filtered out irrelevant categories from the bar sales to improve accuracy. Redoing the analysis with the new data set -and with better transformations as described above- I was able to produce better models. However, it's difficult to say exactly how much better they were, for reasons described below.

Statistical Measures

I recently realized that some of my R^2 values were artificially high, so I pulled up the old project and dug deeper. For models using Genre as a predictor, a quirk in my code was throwing off the statistical measures for each model. I was able to go back and recreate many of these models and get accurate statistical measures for them. Unfortunately my original formula for this Bar Sales model would not compile (nans produced), and I've forgotten how I originally got around that.



	Mult. R^2	Adj. R^2	Pred. R^2	F-Stat.	P-value	D.o.F.
Bar Sales	25.5%	21.6%	14.1%	6.5	1.49E-04	76
Income	67.6%	65.6%	58.7%	34.1	4.74E-12	49
Profit	39.1%	35.2%	27.1%	10.0	3.17E-05	47

In the end, these were the measures on my best model for each regressor (pictured above), with Income being the most applicable for business intelligence. These models are stated explicitly in this workbook and integrated into the dashboard I've created there.