Broadway data analysis

Julianna Szabo

12/23/2020

Executive summary

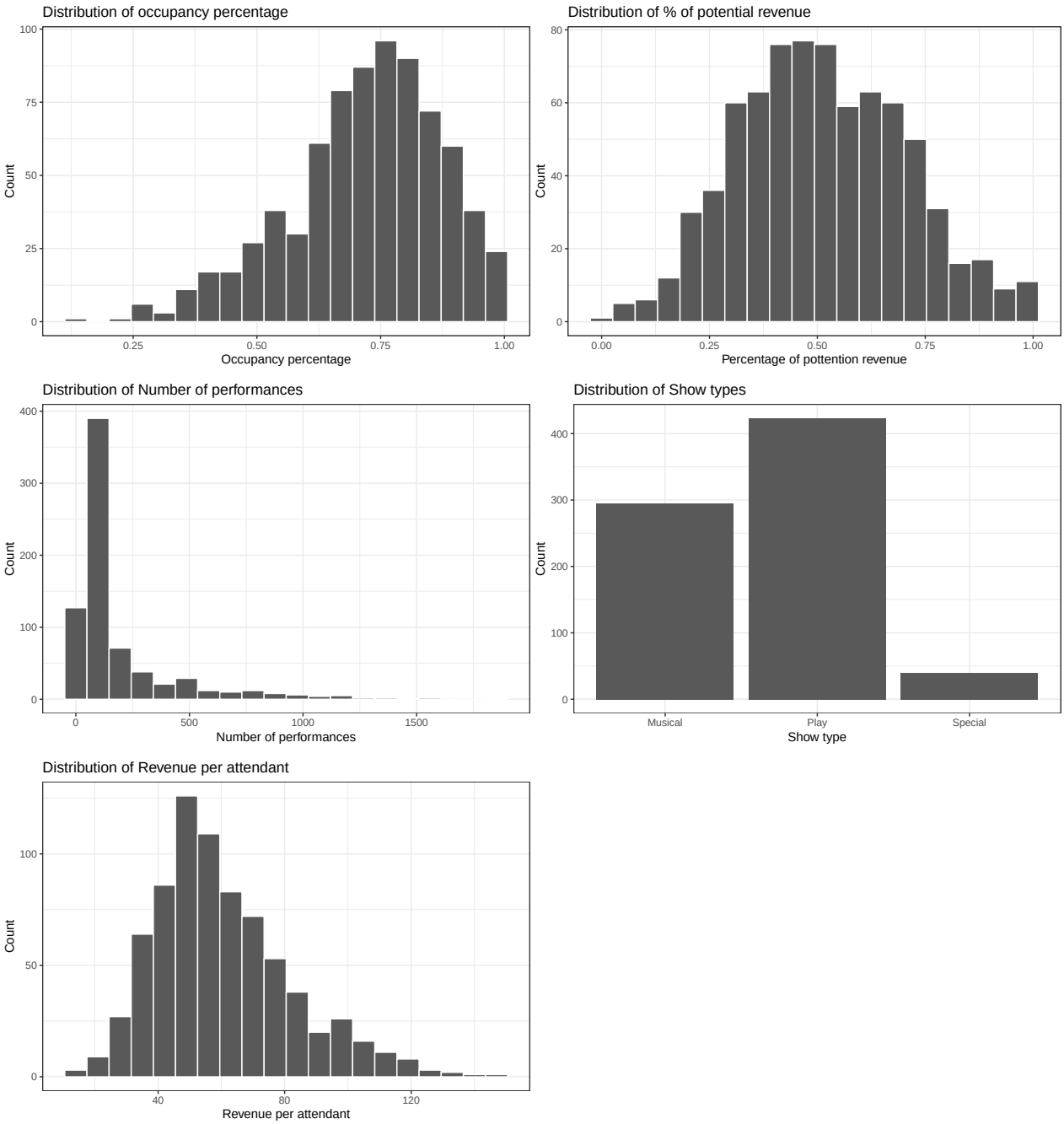
Introduction

This report aims to examine the main factors influencing the revenue per attendant of a Broadway show. Looking at the data, it is predicted that the main influencing variable would be the occupancy percentage, but have also found some other interesting variables that could affect the dependent variable. This project has great benefit for people involved in the theatre industry to see how different elements affect their revenue and possible in the end profit.

The main research question of this report will be:
What are the essential variables that affect the revenue per attendant of Broadway shows?

Data

The data comes from the CORGIS Dataset Project and was originally provided by the Broadway League. You can find a link to the original file in my GitHub folder. The data is a cross sectional time series covering all the shows that have been on Broadway from 1990 to 2016. For this project, the cross sectional aspect is the more important, and therefore, the data was aggregated based on show name. The time series aspect has been discarded but there have been a few adjustments made to control the effects it has on the data.
One of the main ones is using Revenue per attendant as the dependent variable. Since there are shows that ran for over 25 years and some for less than a year it would be unfair to compare the total revenue per show during its runtime. However, since the data includes the total revenue and the total number of attendants over the whole runtime of the show, using the ratio of these two variable gives an easy comparable relevant variable.
Overall the quality of the data is good. It included over 800 observations originally after the aggregation to cross sectional data, which, after cleaning, resulted in 758 complete observations. This shows a very representative sample, that has the potential of generalisation to other cities such as London with a similar theatre scene. There have been a few discrepancies in the data especially with the variables representing percentages. For both Occupancy percentage and Percentage of potential revenue, there were observations with values over 100%, which have been dropped.
Looking at the data more in detail, one can see that there are four explanatory variables to consider for one dependent variable.



(The decision was made to crop the Distribution for Number of performances to 2000 max to show a clearer distribution of the values between 0 and 1000. This distribution would not be very visible if some of the extreme values would have also been shown. These values will be kept in the distribution for analysis)

variable	type	n	mean	median	min	max	sd
Occupancy percentage	x	758	0.72	0.74	0.15	1.00	0.15
Percentage of possible revenue	x	758	0.51	0.50	0.01	1.00	0.19
Number of performances	x	758	279.96	99.00	0.00	8400.00	721.28
Revenue per Attendant	y	758	60.48	56.10	12.64	145.64	21.80

As shown in the graphs above, there are four quantitative ordered variables (including the dependent variable) and one qualitative nominal variable. All quantitative variables, with the exception of the Number of shows, are distributed somewhat normally with a left or right tail. The summary table of the variables also show the distribution of the values. Further, none of the variables are highly correlated (see Appendix 1), so the analysis can be conducted without eliminating any variable.
After seeing the distribution of the variables, the decision was made to do transformations on some of the quantitative variables. After examining grams with possible log transformations (see Appendix 2), log transformations were applied to the Number of performances and Revenue per attendant. This was decided based on the distribution of the observations, but also due to the interpretation being more cohesive across variables. One additional transformation that was later added was to instead of a log transformation of Number of performances a dummy variable was created where 0 denotes the shows with less than one year (416) of performances while 1 denotes the ones with more than that.

Model

To create a more robust mode, the dataset has been split into train and test sets. The model exploration was done on test dataset, and then model picked was rerun and tested on the test set. I will also be working with a 95% confidence interval.
After examining the different options for models (see Appendix 3 and Model comparison file in Out folder), the best fitting model is the linear model using all four explanatory variables. The formula of this model is shown here:

$$\begin{aligned}\ln(\text{Revenue/attendant}) = \beta_0 + \beta_1 &\text{Occupancy percentage} + \beta_2 \text{ Percentage of possible revenue} \\ &+\beta_3 \text{ Number of performances} + \beta_4 \text{ Plays (Show type)} \\ &+\beta_5 \text{ Specials (Show type)}\end{aligned}$$

	Estimate	Std. Error	t value	Pr(>|t|)	CI Lower	CI Upper	DF
(Intercept)	3.44	0.06	57.88	0.00	3.33	3.56	600
occupancy_percentage	0.23	0.12	1.90	0.06	-0.01	0.46	600
percentage_of_poss_revenue	0.97	0.10	9.76	0.00	0.77	1.16	600
as.factor(num_of_performances_d)1	0.02	0.03	0.78	0.44	-0.04	0.09	600
as.factor(show_type)Play	-0.11	0.02	-4.48	0.00	-0.16	-0.06	600
as.factor(show_type)Special	-0.08	0.08	-1.10	0.27	-0.23	0.07	600

Interpretation of coefficients:

Beta 0: When all explanatory variables are 0, the Revenue per attendant would be ln(3.44) - this is almost meaningless.
Beta 1: the Revenue per attendant increases by approximately 23% on average for every additional percentage of occupancy of the theatre, when all other variables are the same.
Beta 2: the Revenue per attendant increases by approximately 97% on average for every additional percentage of the possible revenue achieved when all other variables are the same.
Beta 3: For shows that run longer than one year, the Revenue per attendant increased by approximated 2% on average, when all other variables are the same.
Beta 4: If a show is a Play instead of a Musical, on average the Revenue per attendant is 11% lower on average, when all other variables are the same.
Beta 5: If a show is a Special instead of a Musical, on average the Revenue per attendant is 8% lower average, when all other variables are the same.

Even with this model, that fits the data the best, it still only explains 42% of the observations. Further,

about half of the betas have a very low p value and can therefore we considered very good approximations for this dataset, while three values (Occupancy percentage, Number of performances, and Special show type) have a high p value and therefore the real slope value will fall outside of the 95% confidence interval. Looking at the previous models, it is clear that the real slope of the Occupancy percentage is closer to 1 (or 100%), while the Number of performances is most likely closer to 0.1 (or 10%).

	Train - Estimates	Test - Estimates
Intercept	3.44	3.41
Occupancy Percentage	0.23	0.23
Percentage of pott. revenue	0.97	1.00
More than one year of performances	0.02	0.02
Show type is Play	-0.11	-0.10
Show type is Special	-0.08	-0.08

The model shown is very robust based on the train and test robustness check run. As can be seen in the table above the model gives almost the same coefficients when it is rerun of the test sample. While the R squared here is only 38% but it is very close to the 41% of the original train model.

Generalization

This project has used data from Broadway in New York City to analyse the different variables affecting the Revenue per attendant in theatre. From the findings, generalisations can be made for overall theatre industry, especially in cities like London or Hamburg which have a similarly large theatre fascination within the city. It can be said with confidence that overall Musicals have a higher Revenue per attendant than Plays, and potentially Specials. Further, larger occupancy and larger percentage of potential profit reached, generally lead to higher Revenue per attendant. However, since the model only covers approximately 40% of the observations, this is not conclusive, since there are many observations that cannot be explained by this model. However, the f statistic shows that the model does fit the data better than one with on independent variables, which gives the model some credibility especially for generalization.

Summary

Overall, this project demonstrates that all four exploratory variables (Occupancy percentage of the theatre, Percentage of potential profit, the total Number of performances, and the Show type) have some impact on Revenue per attendant. The change in Percentage of potential profit has the largest and the Number of shows the smallest effect. While the model only explains 40% of the observations, it shows the general tendencies of revenue per attendant for each variable, however, it also demonstrates that this does not have to be the case.

Appendix

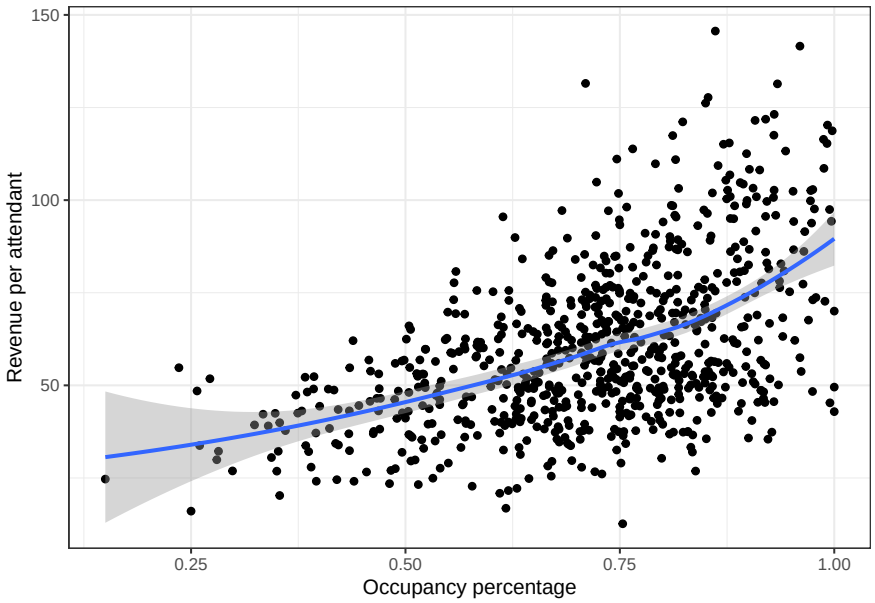
Appendix 1 - Correlation

Var1	Var2	corr_val
ln_occupancy_percentage	occupancy_percentage	0.98
ln_percentage_of_poss_revenue	percentage_of_poss_revenue	0.93
ln_revenue_per_att	revenue_per_att	0.97
occupancy_percentage	ln_occupancy_percentage	0.98
revenue_per_att	ln_revenue_per_att	0.97
percentage_of_poss_revenue	ln_percentage_of_poss_revenue	0.93

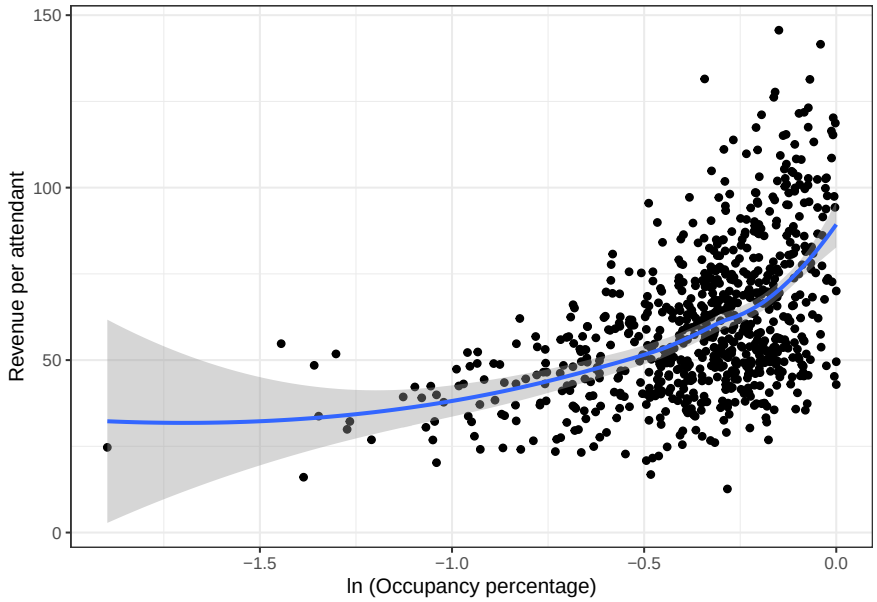
Appendix 2 - Ln transformation

Occupancy percentage

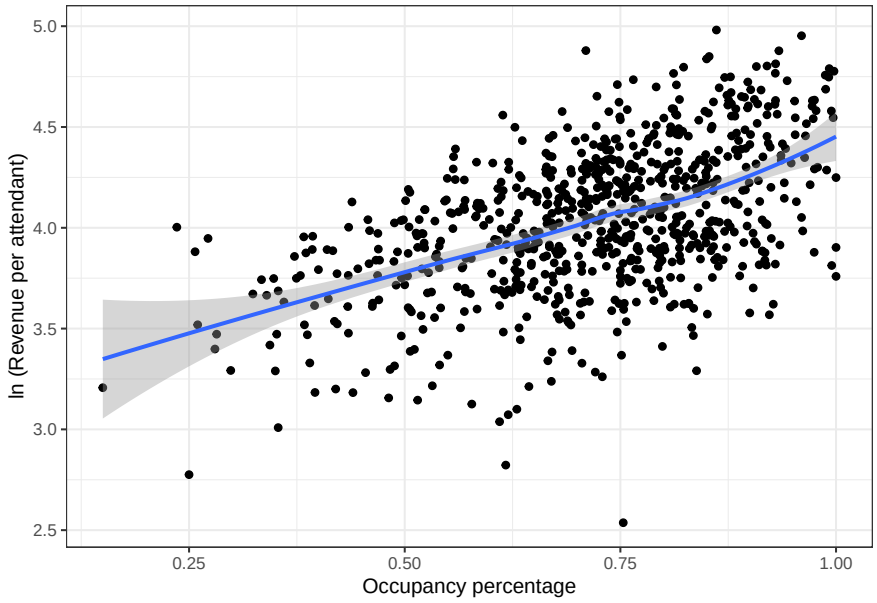
Level - level regression



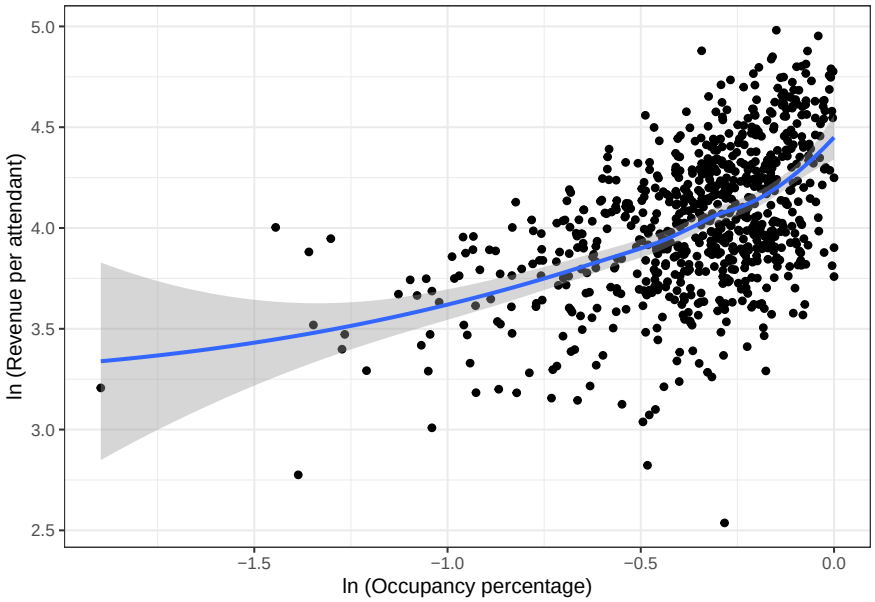
Log - level regression



Level - log regression

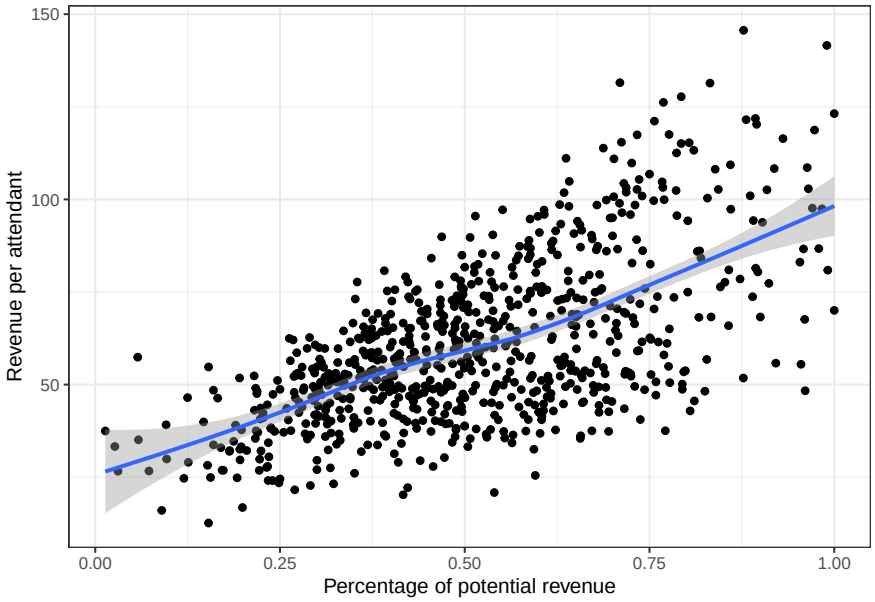


Log - log regression

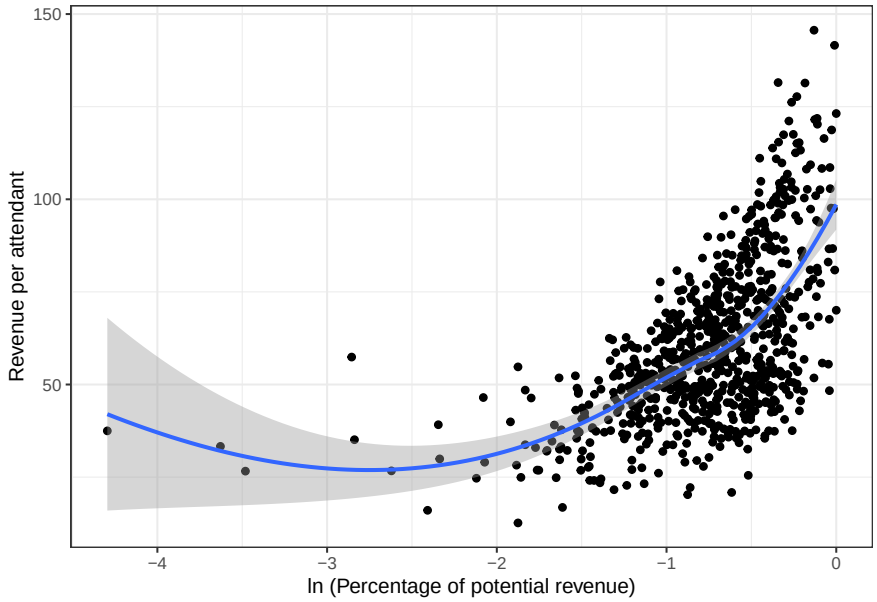


Percentage of potential profit

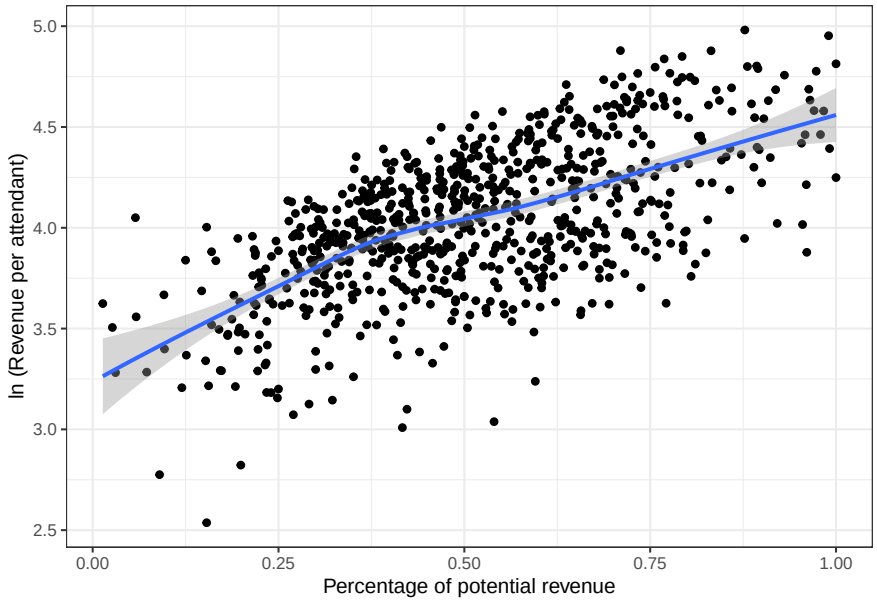
Level - level regression



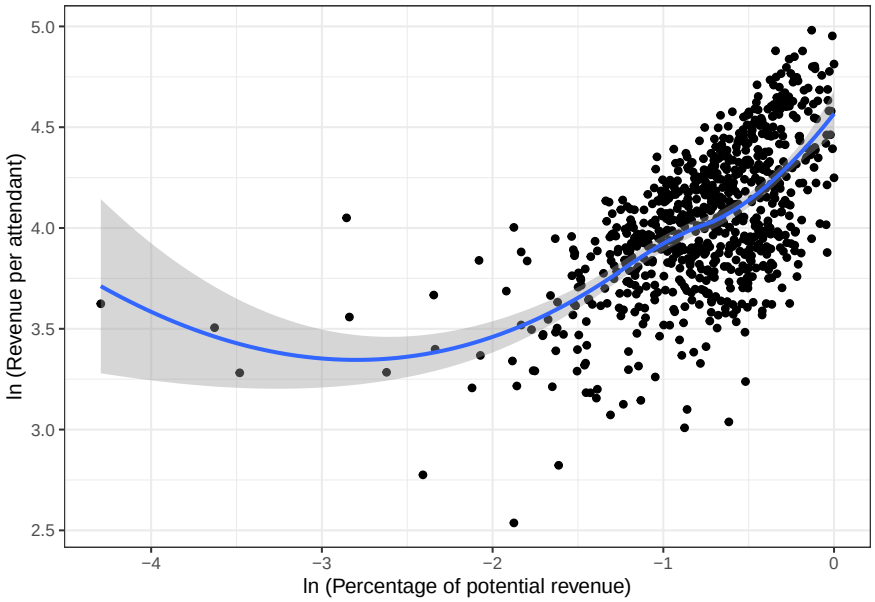
Log - level regression



Level - log regression

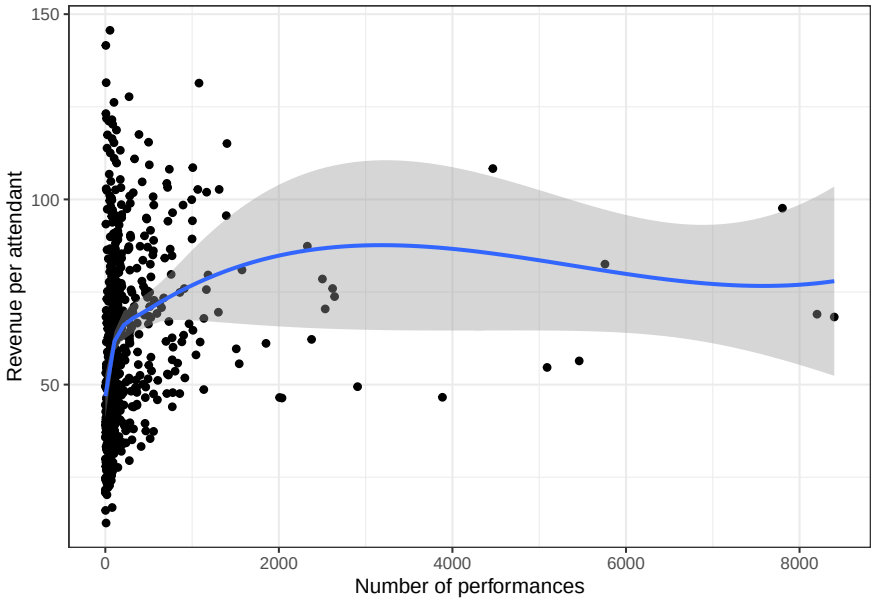


Log - log regression

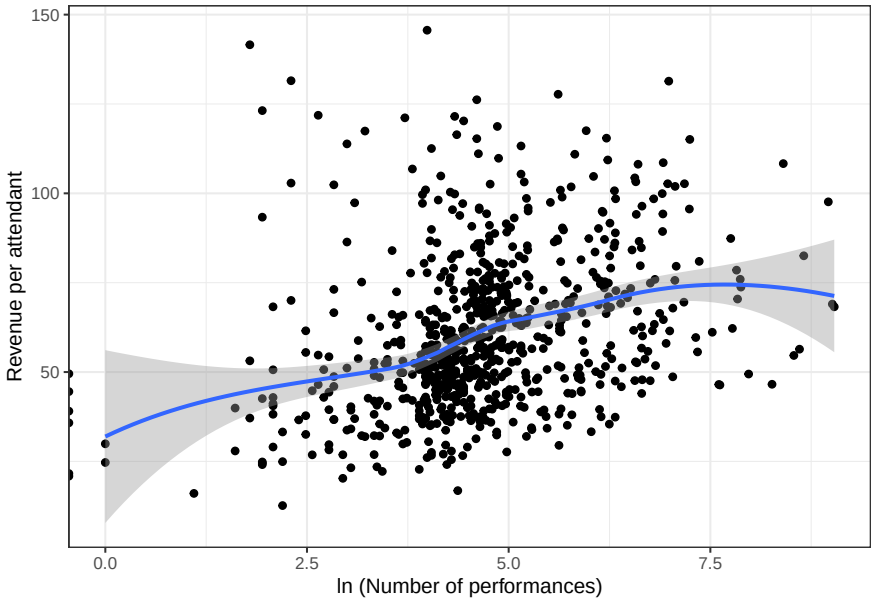


Number of performances

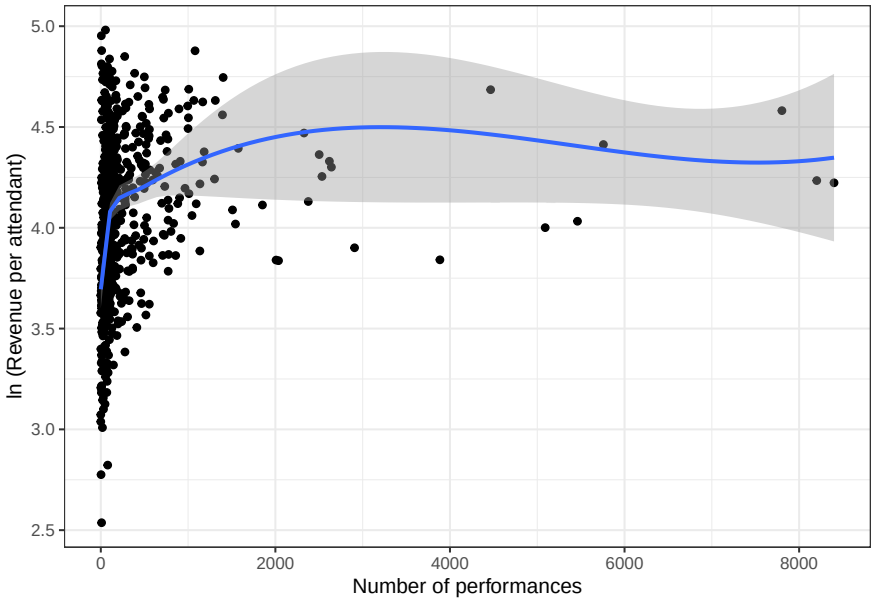
Level - level regression



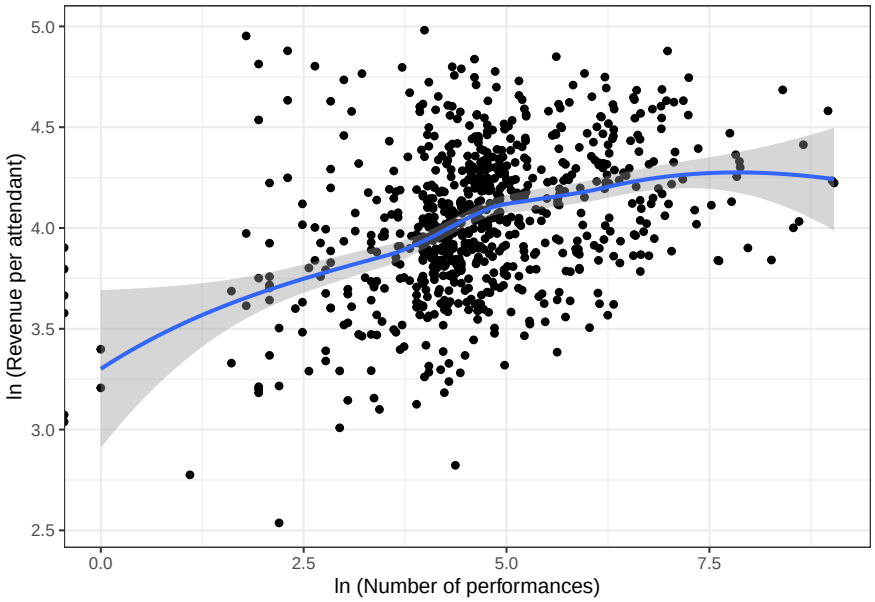
Log - level regression



Level - log regression



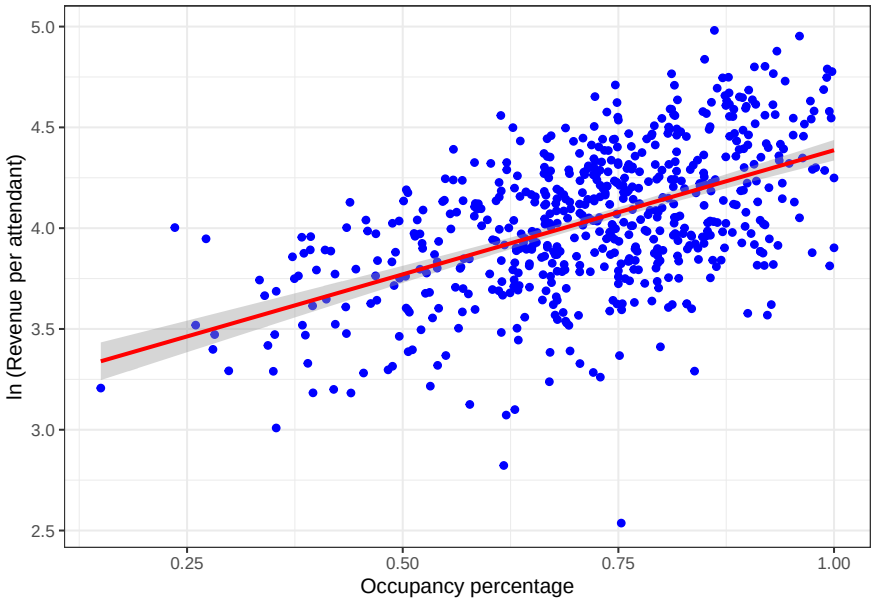
Log - log regression



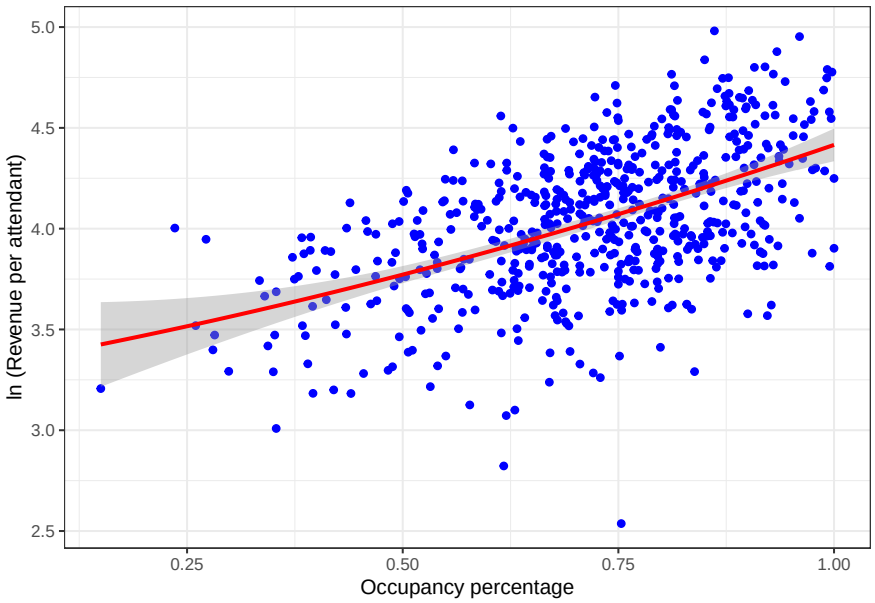
Appendix 3 - Regression modes

To make my regression model more robust, I created a train and test data set

Regression 1 - Simple linear regression

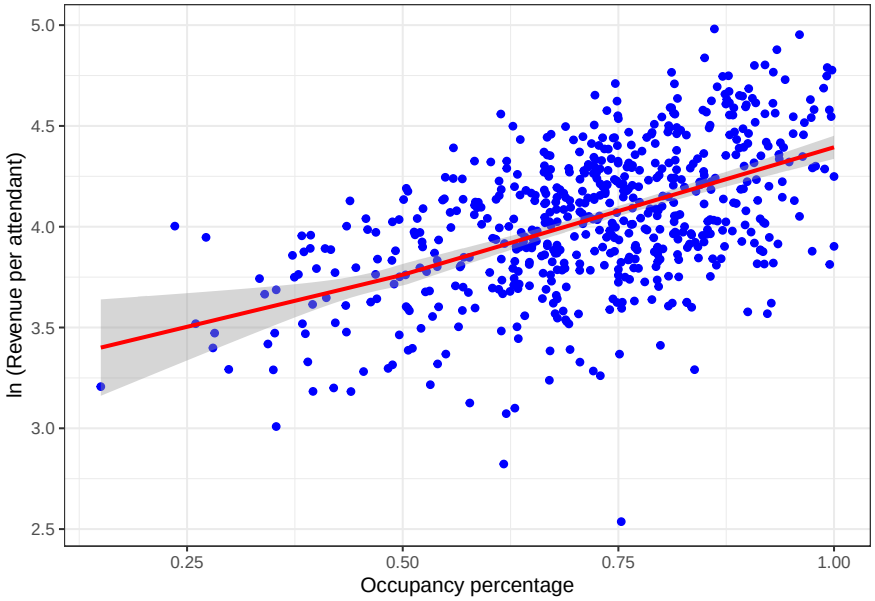


Regression 2 - Quadratic (linear) regression

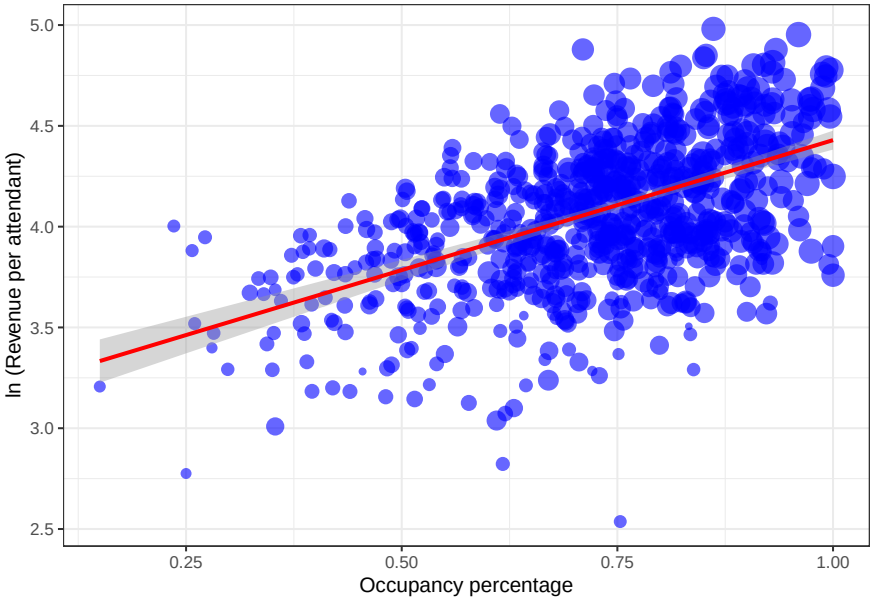


Regressipn 3 - Piecewise linear spline regression

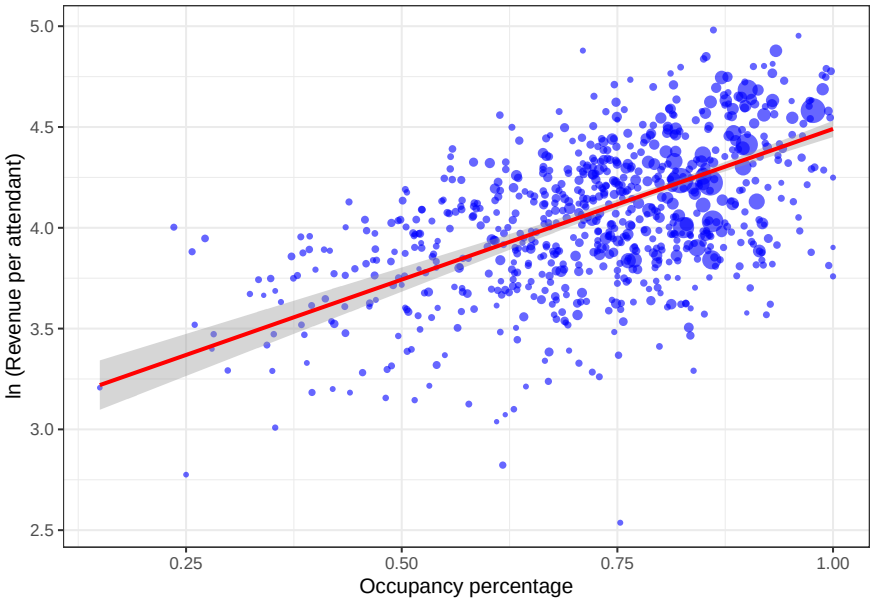
Using 0.5 as a cutof point



Regression 4 - Weighted linear regression, where weights = percentage of total revenue



Regression 5 - Weighted linear regression, where weights = number of performances



Model Comparison

Looks like these models with mainly one variable are not a great fit for the data. Therefore, I will include additional variables to try and get a better fit. Further, it looks like the original use of "Number of Performances" has no impact so I will try and create a dummy variable and use that instead. I will use 0 for any show that had less than one year of performances so less than 8*52 (416) and one for those that have had more.

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	3.15 ***	3.30 ***	3.25 ***	3.14 ***	3.00 ***
	(0.06)	(0.17)	(0.17)	(0.06)	(0.11)
occupancy_percentage	1.23 ***	0.76		1.29 ***	1.49 ***
	(0.08)	(0.51)		(0.09)	(0.15)
occupancy_percentage_sq		0.35			
		(0.38)			
lspline(occupancy_percentage, 0.5)1			1.03 **		
			(0.36)		
lspline(occupancy_percentage, 0.5)2			1.27 ***		
			(0.10)		
nobs	606	606	606	606	606
r.squared	0.28	0.28	0.28	0.27	0.30
adj.r.squared	0.28	0.28	0.28	0.27	0.30
statistic	247.95	123.86	123.39	218.79	98.76
p.value	0.00	0.00	0.00	0.00	0.00
df.residual	604.00	603.00	603.00	604.00	604.00
nobs.1	606.00	606.00	606.00	606.00	606.00
se_type	HC2.00	HC2.00	HC2.00	HC2.00	HC2.00

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Additional models

Check if it becomes better if one of the weights are included as variables. I will also be testing model number 9 (which is the best fit) against the test data set

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	3.33 ***	3.19 ***	3.35 ***	3.44 ***	3.41 ***
	(0.06)	(0.06)	(0.06)	(0.06)	(0.13)
occupancy_percentage	0.30 *	1.15 ***	0.27 *	0.23	0.23
	(0.12)	(0.08)	(0.12)	(0.12)	(0.21)
percentage_of_poss_revenue	0.97 ***		0.94 ***	0.97 ***	1.00 ***
	(0.10)		(0.10)	(0.10)	(0.16)
as.factor(num_of_performances_d)1		0.12 ***	0.09 **	0.02	0.02
		(0.03)	(0.03)	(0.03)	(0.07)
as.factor(show_type)Play				-0.11 ***	-0.10
				(0.02)	(0.05)
as.factor(show_type)Special				-0.08	-0.08
				(0.08)	(0.16)
nobs	606	606	606	606	152
r.squared	0.39	0.29	0.40	0.41	0.38
adj.r.squared	0.39	0.29	0.39	0.41	0.36
statistic	180.94	141.00	130.64	85.85	22.08
p.value	0.00	0.00	0.00	0.00	0.00
df.residual	603.00	603.00	602.00	600.00	146.00
nobs.1	606.00	606.00	606.00	606.00	152.00
se_type	HC2.00	HC2.00	HC2.00	HC2.00	HC2.00

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.